EVALUATION OF MACHINES LEARNING ALGORITHMS IN DETECTION OF MALWARE-BASED PHISHING ATTACKS FOR SECURING E-MAIL COMMUNICATION

Kambey L. Kisambu[1] and Dr. Mohamedi Mjahidi[2]

[1]Msc, Cyber Security, Department of Computer Science, University of Dodoma
[2]Lecturer, Department of Computer Science, University of Dodoma, Tanzania

ABSTRACT

Malicious software, commonly known as Malware is one of the most significant threats facing Internet users today. Malware-based phishing attacks are among the major threats to Internet users that are difficult to defend against because they do not appear to be malicious in nature. There were several initiatives in combating phishing attacks but there are many difficulties and obstacles encountered. This study deals with evaluation of machine learning algorithms in detection of malware-based phishing attacks for securing email communication. It deeply evaluate the efficacy of the algorithms when integrated with major open-source security mail filters with different mitigation techniques. The main classifiers used such as SVM, KNN, Logistic Regression and Naïve Bayes were evaluated using performance metrics namely accuracy, precision, recall and f-score. Based on the findings, the study proposed improvement for securing e-mail communication against malware-based phishing using the best performing machine-learning algorithm to keep pace with malware evolution.

KEYWORDS

Malware; Malware Analysis, Malware-based, Phishing attacks, Spams, e-mails, Machine learning, algorithms, mail filters, Detection, Mitigation techniques.

1. INTRODUCTION

In today's computerized world, especially with the spread of smart phones and Internet access, malware is becoming a major concern. Malware is software created and used by cyber-attackers to disrupt computer systems, gain computer access, or gather sensitive user information. Many problems in computer security, such as the distribution of phishing scams, are embedded in the spread of malware and botnets that are widely used in launching those attacks. While Phishing is a cybercrime model where an attacker impersonate a real person or institution by advancing them as an official person or organization between emails or other means of electronic communication [1]. Malware-based phishing attacks are among the major threats to Internet users that are difficult to track down or defend against because they do not appear to be malicious in nature. The attacker usually dispatches malevolent connections or extensions through phishing e-mails that can execute numerous tasks, such as capturing account information from the victim. A typical phishing e-mail is sent to bulk users' accounts and are dispatched to prospective victims' inboxes while consistently occurs with clickable URL links. It intends to attract the recipient into trusting that the email received is from a trusted source [2]. This attracts the recipient to visit

David C. Wyld et al. (Eds): ICAIT, CBIoT, WiMo, CRYPIS, ICDIPV, CAIML, NLCA - 2022
pp. 17-32, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121202

the presented website hyperlink, which connects them to fake or fraudulent websites and eventually extracts personal information.

According to statistics given by the Anti Phishing Working Group (APWG) in the 3rd quarter of 2021, the amount of phishing attacks has multiplied rapidly since the beginning 2020, and APWG observed 260,642 phishing attacks in July 2021, the exorbitant monthly attacks in APWG's reporting history (Activity & Report, 2021). According to APWG, the average wire transfer request in Business E-mail Compromise (BEC) attacks has increased from $48,000 in Q3 to $75,000 in Q4 of 2020, while the software as a service and webmail service were the mass recurring exploited by phishing in the last quarter of 2021, accounting for 29.1% of attacks. As for Tanzania, it has been noted that number of internet users in Tanzania has been increased from 27.9 million to 29.1 million from September 2020 to March 2021 respectively (TCRA, 2021). With these statistics, it shows that there is high rate of internet penetration and number of internet users who are more victims of phishing attacks across the country and the world at large. More recently, some studies such as [2] and [3] showed the number of phishing attacks have increased during the Corona virus pandemic (COVID-19) and the phishers take advantage of COVID-19 to fool their target and users especially from healthcare facilities. Many Corona virus themed spam and scam messages sent by attackers exploited people’s fear of contracting COVID-19 and urgency to look for information related to Corona virus.

Even though there are email filters that use machine learning (ML) techniques and a number of researches related to phishing attacks' detection and mitigation, the interesting thing is that phishing attacks are continuing to evolve every year. Moreover, phishers and malware are becoming more intelligent and evolving through obfuscation. Thus, it was stated that the encounter between security techies and malware innovators is a continuous fight with the convolution of malware alternating as quickly as transformation heightened [5]. Consequently, it is required to keep on researching and enhancing the accuracy of the detection techniques simply because there is no single solution to the phishing problem due to the heterogeneous nature of the attack vector [5].In that aspect, there is a crucial need for evaluating machine learning algorithms for detection of malware-based phishing attacks for securing email communication.

The purpose of this study is to presents an overview about various malware based phishing attacks and various techniques used to protect users in e-mail communication. The study intends to narrow the scope and specifically deal with malware-based phishing attack identification and control techniques using ML algorithms. The study is expected to deeply evaluate the efficacy of the algorithms when integrated with major open-source mail systems’ filters, as e-mail communication is the leading route used by phishers. Additionally, the research will look into the efficacy of ML in exposing phishing attacks from COVID-19 related content as some studies showed that phishing incidents massively increased during the COVID-19 pandemic era.

The remaining sections of this paper are organized as follows: Section 2 provides an overview of some literature reviews including related works; Section 3 describes methodologies used; Section 4 explores machine learning (ML) algorithms, experiment made with results and performance evaluation. In section 5, the paper provide the conclusions of the study and future work.

2. LITERATURE REVIEW

2.1. Phishing Attacks Categories

Usually, phishers conduct their attacks either by using psychological brainwashing of individuals into revealing their personal information (i.e. deceptive attacks as a form of cracking) or by misleading users into unfolding their private information through hi-tech trickery (i.e., technical methods) by downloading malevolent code into the victim's system [5]. Although phishers prefer deceptive attacks over technical methods, mitigation of technical methods attacks cannot be overlooked. Figure 1 illustrates the types of phishing and techniques used by phishers to conduct a phishing attack whereby malware-based phishing that falls under the technical subterfuge with six (6) sub-attack techniques will be the area of study in this research. The forms of malware-based phishing attacks are described hereunder:

2.1.1. Key Loggers and Screen Loggers

Key Loggers are the type of malware used by phishers to install either through Trojan horse email attachments or through direct download to the user's computer [5]. This software monitors data and record user keystrokes and sends them to the hacker or phisher. Key loggers and screen loggers are specific variation of malware that track keyboard input and send relevant information to a hacker or phisher via the Internet [6]. They can implant themselves into users' computer browsers as small convenient plan of action that run automatically when the browser is started.

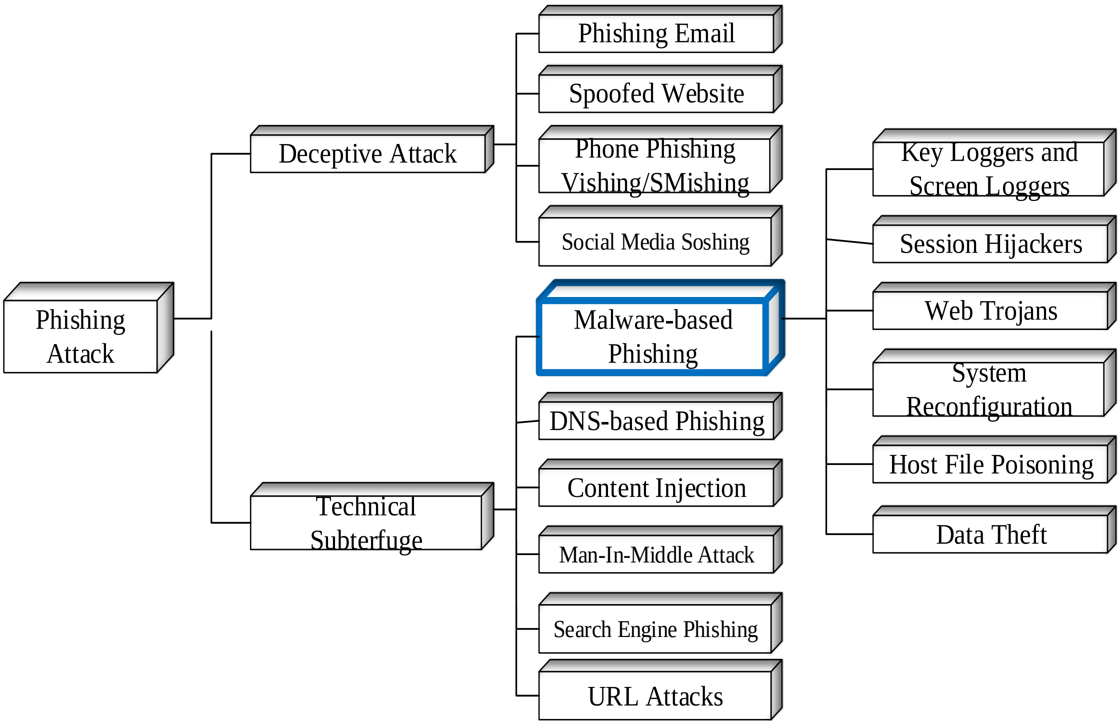


Figure 1. Types and Techniques of Phishing Attacks.
Source:[5]

2.1.2. Session Hijacking

Malware can also be used to hijack a session when a user logs into a system through a web browser to perform a transaction. The infectious software hijacks the user session and performs malicious activity once the user credentials are proved to be correct with the transacting system. In this type of phishing, the attacker observes the user's tasks by planting malevolent software inside a browser component or via network interception. Once the link is fixed, the malicious software controls and perform unwarranted actions, such as transmission of savings, without the user's knowledge [7].

2.1.3. System Reconfiguration Attacks

In this form of phishing attack, the phisher exploits the site on a user’s computer for malevolent activities with the aim of compromising computer information [5]. System design can be altered using different methods, such as altering the operating system and redesigning the user’s Domain Name System (DNS) server address.

2.1.4. Web Trojans

Web Trojans are malicious programs or codes that collect a user’s detailed information, such as credentials, by popping up in a hidden mechanism over the login screen [5]. Phishing attacks often lead users to Web Trojans or clone websites that operate when users are trying to log on [7].These Trojans can capture important information and send them to the phisher. The sites can typically include duplicated icons and may even culminate realistic-looking SSL padlocks and third-party verification services.

2.1.5. Host File Poisoning

This kind of phishing refers to a way to trick a user into going to the phisher’s site by poisoning (changing) the host’s file. When the user types a particular website address in the URL bar, the web address will be translated into a numeric (IP) address before visiting the site [5]. Usually, the attacker modifies this file in order to lead the user to a fraudulent website for phishing purposes.

2.1.6. Data Theft

Data theft in phishing attacks refers to the unauthorized accessing and stealing of confidential information by a business or individual. Data theft can be done by a phishing email that leads to the download of a malicious code to the user’s computer, which in turn steals sensitive information stored on that computer directly [5]. Stolen information such as system passwords, credit card information, social security numbers, and other personal data could be used directly by a phisher or indirectly by selling it for different purposes.

2.2. Malware-based Phishing Attacks Phases

A typical phishing attack includes three phases of phishers that cover several stages. To begin with, mailers send out many deceitful emails (usually through botnets), which redirect users to deceptive websites or download malicious code and install it on their machines as shown in stage 1, 2 and 3. Attackers use obfuscation techniques as the second step to conceal the malevolent texts under various layers of obscurity [8]. Various studies such as from Al-Shira’h & Al-Fawa’reh (2020) showed that constant investigation endure obfuscation and evasion attacks in most cases, while dynamic analysis itself requires a considerable amount of manual inspection for crafting detection patterns from the diversity of malware variants. Specifically, attackers try to prevent static analysis of some features by using obfuscation techniques like obfuscating the host with an IP label for malicious URLs that are statistically identical to benign ones.

Furthermore, phishers create fraudulent websites (regularly organized on compromised computers) that actively induce victims to redirect to attacker website as shown in stage 4. The victim user can also download the Remote Access Trojan (RAT) and when installed in the computer in the network, can spread in the organization network as shown in stage 5 and induce users to provide private details.

Finally, the stolen information is submitted to phisher server (stage 6) and phishers use the stolen confidential information (stage 7) to hack the user's data, such as money. The information circulation is shown in Figure 2.

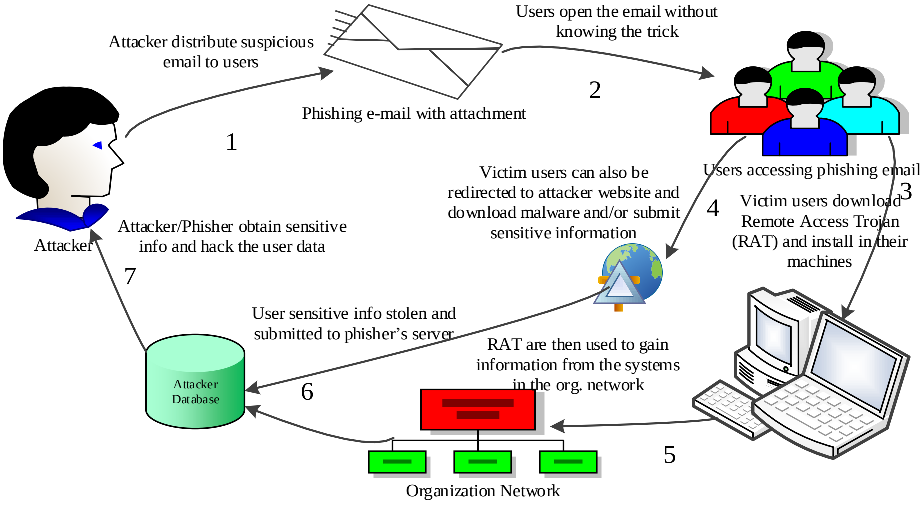

Attacker distribute suspicious email to users
Phishing e-mail with attachment
Users open the email without knowing the trick
2
1
Users accessing phishing email
3
Victim users can also be redirected to attacker website and download malware and/or submit sensitive information
4
Victim users download Remote Access Trojan (RAT) and install in their machines
Attacker/Phisher obtain sensitive info and hack the user data
Attacker
7
User sensitive info stolen and submitted to phisher's server
RAT are then used to gain information from the systems in the org. network
Attacker Database
6
5
Organization Network

Figure 2. Information Flow in the Stages of Malware-Based Phishing Attacks

2.3. Related Works

Several studies have been conducted to address phishing attacks, detection and mitigation techniques. Each of the studies has strengths and weaknesses that could be addressed in future work by new researchers.

Rastenis et al. (2021) analyzed the existing spam and phishing email classification solutions and revealed from multiple papers that all of them are concentrated on the categorization of recognized and malicious email. As most public email datasets almost exclusively collect English emails, they investigated the suitability of automated dataset translation to adapt it to email classification written in other languages. The study focuses on solution for email classification written in only three languages, namely English, Lithuanian, and Russian [9]. The proposed solution in the study with automated translation for dataset augmentation and adaptation for the three languages prove the classification results do not decrease because of the automated translation. The result for English-only text, the accuracy was 90.07% +- 3.17% while for multi-language texts (English, Russian and Lithuanian) it was 89.2% +- 2.14%. The study was not able to demonstrate if the suggested explanation is suitable for other languages such as Swahili and how the email classification performance is affected when adapting feature optimization.

Madhavan et al. (2021) discussed the comparative analysis of disclosing fraudulent emails using various machine learning methodological analysis along with the suggested concepts with consideration of various evaluation metrics such as accuracy, efficiency, error, and evaluation time of the model. The study presented the issues based on several setbacks faced in spam filtering and classification when a particular algorithm is considered, such as evaluation time, cost, and computing resources. The study draws the variation between the strengths, weaknesses, and hindrances of some of the existing techniques that use the machine learning methodologies to identify spam emails [10]. Although the study was not able to demonstrate how efficient the developed algorithm was able to perform at best, a hybrid algorithm was suggested as the best and most feasible solution for spam detection in e-mail communication to overcome the observed challenges. Also since the study focused only on spam detection and classification, there is a need to focus and draw contrast on the strengths, weaknesses, and limitations of malware-based phishing detection using ML algorithms and propose the best mitigation measures.

Ayman El Aassal and Shahryar Baki [11] performed a systematic study and assessment of phishing emails. The study introduced a novel taxonomy of features for phishing emails, websites, and URL detection based on their structure and how the features are processed by the web and email servers. The study proposed a novel phishing identification framework named PhishBenchused to evaluate and compare the existing features for phishing detection. The framework was also intended to act as a ready-to-use platform for security researchers. It was discovered that phishers always change their attack techniques to bypass defense mechanisms. One of the solutions suggested to minimize the attack is by retraining using a more recent dataset, as the experiment showed that slightly helps existing models to detect newer attacks. The researchers mentioned that retraining the model alone is inadequate to deal with the new attacks; unfortunately, they were not able to experiment alternative solutions.

Sanouphab Phomkeona and Okamura [12] proposed a new method to extract features from email and a deep-learning approach to detect zero-day malware spam. They extracted some features from e-mail's header and body parts that included risk words detected, machine translation detected, and other features by using several APIs. They also used four different language email datasets for more diversity and a realistic purpose to build a database of words. The experiment results showed a 78% accuracy rate for zero-day email spam detection and a 92.8% accuracy rate for normal spam. The accuracy rate for features used in the zero-day email spam detection didn't increase much because the spam email dataset used contains only normal spam and not malicious or phishing spam. Thus, there is a need to balance the dataset when conducting this study for malware-based phishing attacks, to include malicious or phishing spam dataset in order to increase accuracy and improve phishing detection and mitigation.

Gibert et al., (2020) presented a methodological review of malware identification and classification perspectives using machine learning. Different studies were reviewed, compared and examined as maintained by various factors including input features, classification algorithm, characteristics of the dataset, and the objective task. There were four main contributions, including a detailed explanation of the methods and features in a traditional machine learning workflow and literature on malware detection through deep learning. The other main contribution was a discussion on research issues and challenges faced by researchers, with emphasis placed on the problem of concept drift and the challenges of adversarial learning, among others. The study insisted on an endless battle between security analysts and malware developers due to the complications of malware development as quickly as innovation grows [4]. This study emphasizes that, there is a need to add effort in this never-ending battle of mitigating the attacks, specifically malware-based phishing attacks, for securing email communication.

Alkhalil et al. (2021) investigated problems presented by phishing and proposed a new anatomy that describes the complete life cycle of phishing attacks. The anatomy provides a wider outlook for phishing attacks with an accurate definition covering end-to-end mechanisms. The proposed new anatomy of phishing involves attacker types, attack phases, vulnerabilities, targets, threats, attack media, and attacking techniques that when combined could help in developing a holistic anti-phishing system. The study highlighted that there is no single solution for mitigating phishing attacks due to the heterogeneous nature of the attack vector but there was no any experimental setup, which prompted to conduct this research study. The study insisted on the importance of developing efficient anti-phishing techniques that prevent users from being exposed to the attack as an essential step in mitigating the attacks by detecting and/or blocking them. With regard to the stated significance, it is vital to evaluate machine learning algorithms in detection of malware-based phishing to assist in developing an efficient anti-phishing solution.

Azeez and Ajayi (2019) carried out a comparative analysis of three famous machine learning algorithms (Decision Tree, Nave Bayes and Logistics Regression Model) for verification of compromised, suspicious and fake URLs sent by spammers and phishers. The analysis determined the best of all the algorithms based on the metrics such as F-Measure, Precision, and Recall used for evaluation. The result obtained based on the confusion matrix measurement shows that the Decision Tree algorithm achieves the highest values for the three metrics and provides an efficient and credible means of maximizing detection of compromised and malicious URLs. The study cautioned on inconsistencies noticed in various researchers' findings that made corresponding results not dependable based on the values obtained and conclusions drawn from them but it was not able to provide the way forward. The authors of the study proposed that, two or more supervised machine learning algorithms can be hybridized, making one effective and more efficient algorithm for fake URL verification but were not able to implement [13]. The study aimed to design a system to detect suspicious links in e-mails and notify users instead of blocking them. The study also used only three ML algorithms to draw conclusion but some popular algorithms such as SVM could have been used for comparative analysis.

Rafatet al. (2021) showed that text pre-processing methods nullify the detection of malicious content in an obscure communication framework based on their study and experiment. They used the Spamassassin corpus as a mail filter with and without text pre-processing and examined it using machine learning (ML) and deep learning (DL) algorithms to classify it as spam e-mails. The study proposed a DL-based approach that consistently outperforms standard ML models in detecting malicious content. Although the results showed the power of DL algorithms over the standard ML in filtering spam, the effects were unsatisfactory for detecting encrypted communication for both forms of algorithms [14]. The study need to be linked with the evaluation of machine learning algorithms in detecting malware-based phishing attacks.

Sameena Naaz (2021) conducted a detection of phishing study on the Internet of Things (IoT) using a machine learning approach. The ML algorithms that include random forest classifier, support vector machine, and logistic regression have been applied to the IoT dataset for the detection of phishing attacks. The results of the study have been compared with previous studies that were carried out on the same dataset as well as on different dataset from MillerSmiles archive, PhishTank archive and Googleâ€™s. Although the study was limited to feature selection and feature extraction, as well as observation for some false alarm rates, it was found that Random Forest works better in terms of accuracy and error rate. There was a suggestion for improvement to use other methods and approaches for feature selection and feature extraction as well as the implementation of hybrid ML algorithms that improve accuracy and minimize false alarm rates [15]. However, the study was not able to mention and simulate the other methods for feature selection and feature extraction. This study will therefore focus on evaluating machine-learning algorithms in detection of malware-based phishing using different approaches for feature selection and feature extraction to improve accuracy and reduce false alarm rates.

3. METHODOLOGIES

The research methods and steps used in this study include literature review, data collection, dataset creation, practical experimentation, and integrating the ML Model with the spam filter as shown in figure 5. The steps begin with a systematic literature review that covers various studies, related works, and features for machine learning models to provide context for the topic. This is followed by data collection and then a section on dataset creation is discussed, because in order to proceed with the classifier training and testing, a dataset must be in place. Data processing, including pre-processing, classifier evaluation and results is examined. Based on the best performing machine learning algorithm, the ML model will be improved and integrated with spam filter to round up the study. In order to accomplish this study, the emulation experiment was conducted using an environment comprising of a virtual server with Python libraries installed and mail server components such as Dovecot, Postfix, Amavis, Spam Assasin, and Webmail.

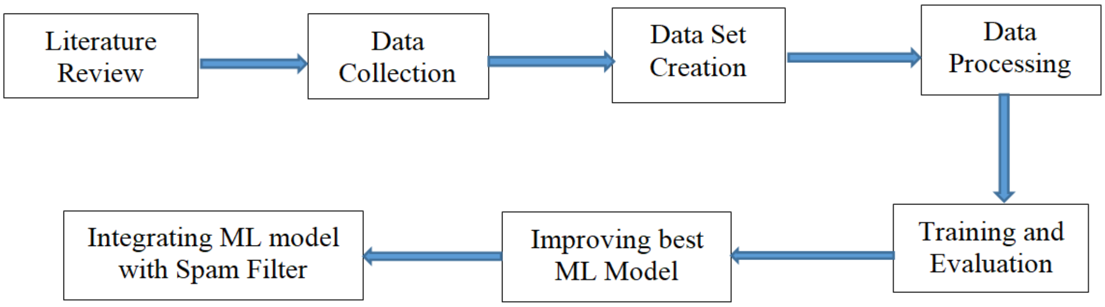


Figure 3. Research Methods Adopted

3.1 Fighting Spams and Phishing Approaches

One of the approaches most commonly used in fighting spammers is the email security filters, which use filtering techniques (spam filters). This technique is based on analyzing the message content (header and body) and other information, which can help to identify the legitimacy of the messages before they reach the user's mailbox. After identifying messages that contain scams, the action that follows depends on the settings that are applied by the mail filter itself. Some filters mostly utilize a mail server settings and usually take a separate measures of deleting the message, putting it in quarantine or labeling it as spam. However, the most appropriate method of detecting malware and spam is by using ML because they have some characteristics that are learned by the machine with the help of previously collected data in the ML algorithm [16].

Figure 4 shows the main steps taken in spam and scam mail filtering using machine learning technique. When the message is received, the initial course of action in the process is to extract the words from the message body (tokenization). This is followed by the subsequent step, which is to modify the words to their base form (lemmatization, e.g., "extracting" to "extract"). Tokenization is therefore the process of making larger words into smaller words and put into appropriate data type while lemmatization is the process of converting a word into its natural base form [17]. Also, the stop-words removal takes place by eliminating words that transpire frequently in many texts (e.g., "the," "you," "and," "to," "a," and "for") [18]. The conventional features that are usually used in spam filtering are Term Frequency with Inverse Document Frequency (TFIDF) but there were studies such as Malero, (2014) that presented alternative approach of Relative Frequency with Power Transformation (RFPT) coupled with lemmatization technique and it considerably showed improvements over TFIDF [19]. Finally, the presentation changes the messages in a format that a machine learning algorithm can use for classification.

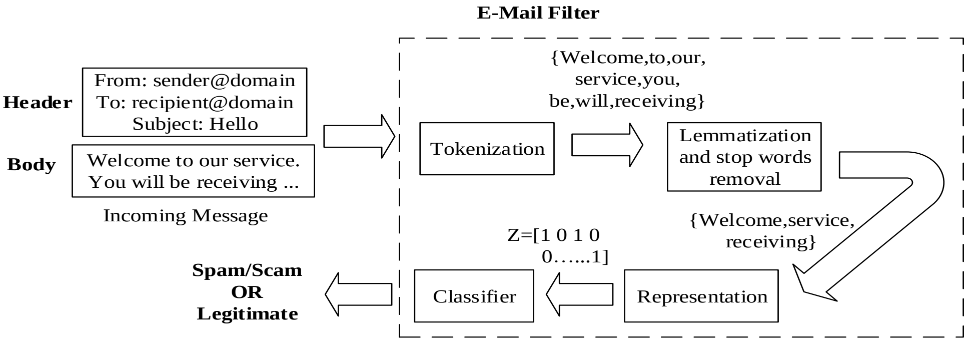


Figure 4. E-mail filtering Process

3.2. Tools Used

Scikit-Learn (SKLearn) is an environment that is incorporated with Python programming language and it is widely used in machine learning experiments. The library offers a wide range of supervised algorithms that will be suitable for this study. The library offers high-level implementation to train with the 'Fit' methods and 'predict' from an estimator (Classifier).

3.3. Machine Learning Algorithms Used

The classification techniques used in mail filtering can be grouped as content-based filtering techniques, case-based spam filtering methods, rule-based spam filtering techniques, previous likeness-based spam filtering techniques, and adaptive spam filtering techniques [20]. Various studies such as [20], [21], and [9] revealed that the most popular ML algorithms used in text classification are the Nave Bayes Classifier (NBC), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). This subsection explain each of the ML models that will be implemented to achieve the aim of this study.

3.3.1. Logistic Regression

Logistic Regression is a classification algorithm which is based on the probability concept and its cost function lies between 0 and 1. In this algorithm, the sigmoid function is used to model the data as shown in the function $g(z) = 1/(1 + e^{-z})$.

3.3.2. Naïve Bayes (NB)

Naïve Bayes model is used to resolve classification problems by using probability techniques defined by the following formula:-

$$P((\text{Phish OR Ham})|\text{WORD}) = \frac{P(\text{WORD}|(\text{Phish OR Ham})|) \text{ X } P((\text{Phish OR Ham})|)}{P(\text{WORD})}$$

There are three types of Naïve Bayes algorithms, which are Multinomial, Gaussian and Bernoulli. Multinomial Naïve Bayes (MNB), algorithm that uses Multinomial Distribution for each given feature, focusing on term frequency, has been selected to perform the spam email identification because it is text related and outperforms Gaussian and Bernoulli as per various studies.

To test this algorithm, MNB module was loaded from the Scikit-learn library. The parameters for this model are optional. If none is specified, the default values are: Alpha value set to `1.0', Fit Prior is set to `True' and Class Prior is set to `None'.

3.3.3. Support Vector Machine (SVM)

This algorithm plots each node from a dataset within a dimensional plane and through classification technique the cluster of data is separated by a hyper plane into their respective groups and is defined as:- H = VX +c

where c is a constant and V is the vector.

The Stochastic Gradient Descent (SGD) classifier, which is the linear model was loaded from scikit-learn library. SGD is the optimized version of SVM algorithm and it provide more accurate results than SVM itself [22]. Also there is a disadvantage of working with SVM algorithm since it cannot handle a large dataset, whereas SGD provides efficiency and other tuning options.

3.3.4. Decision Tree (DT) Classifier

The Decision Tree model is based on the predictive method and it creates a category which is further distributed into sub-categories or sub trees. The algorithm usually runs until the user has terminated or the program has reached its end decision. Similar to MNB and SGD, DT algorithm was loaded from the Scikit-learn library and it is executed on the default parameters which are `Gini' for Criterion and `best' for Splitter.

$$\text{Gini: } G_i = 1 - \sum_{K=1}^{n} P_{(i;k)2}$$

3.3.5. Random Forest Classifier

Random Forest (RF) algorithm can be used for both classification and regression whereby the algorithm predicts the classes by using multiple decision tree, where each tree predicts the classification class. This module was loaded from Scikit-learn library and it is based on the depth of the tree and number of DT to be produced. The termination criteria is usually considered as the more the depth and number of trees the more the computational time required for the algorithm.

3.3.6. K - Nearest Neighbor (KNN)

KNN algorithm calculates Euclidian distance and ranks the samples according to the distance between the neighbors. It makes use of the concept of similarity that helps to classify spams based upon the distance between the new mail that is to be classified and mails in the training set.

3.3.7. Multilayer Perceptron (MLP)

The MLP is a feed-forward Artificial Neural Network (ANN), which is a supervised method that includes non-linear hidden layers between the input and the output layer. The algorithm works with the linear activation function on a training dataset set by default known as Hyperbolic Tan.

$f(\bullet) : R^m R^o \rightarrow$

where ‘m’ is the input (spam words in this case) and ‘o’ is number of outputs from the function.

3.4. Datasets, Model Training and Testing Phase

As discussed through this paper, supervised learning methods were used and the model was trained with known data and tested with unknown data to predict the accuracy and other algorithms performance measures. K-Fold cross validation method was applied to acquire the reliable results although the method have disadvantages such as having a chance that the testing data could be all spam or scam emails, or the training set could include the majority of spam and scam emails. The weakness was resolved by Stratified K-fold cross validation, which separates the data while making sure to have a good range of Spam/Scam and Ham into the distributed set [22]. The parameter tuning was lastly conducted with the Scikit-Learn to improve the accuracy.

In case of datasets, the study accessed the publicly available datasets and included each email as an individual text file since the text files were string based. A list of the few spam and phishing email datasets from the public repository that were used in this study are:

(i) Ling-Spam dataset

The datasets are divided into 10 parts from the `bare' distribution that includes individual emails as a text files. This data is typical primary data since it is not pre-processed, and it includes numbers, alphabets and characters. Each part of the data was trained and tested.

(ii) Spam Assassin dataset

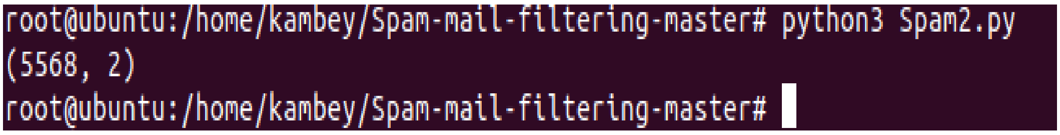
The dataset is more advanced with email text files and header information such as source or From address, IP address, return path, message ID and delivery information.

(iii) Enron Dataset

Enron dataset includes 6 separate datasets that contain 3000-4000 individual emails as text files. The dataset includes numbers, alphabets and characters.

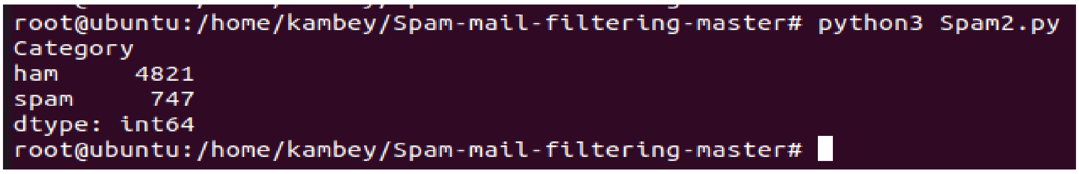
(iv) Kaggle Dataset

The dataset have header and body information and the source dataset is raw as it is not pre-processed. The dataset used here contains 5568 instances with 5568 rows and 2 columns labelled as ‘Category’ and ‘Message’ respectively as shown in figure 5 and figure 6.



```
root@ubuntu:/home/kambey/Spam-mail-filtering-master# python3 Spam2.py
(5568, 2)
root@ubuntu:/home/kambey/Spam-mail-filtering-master#
```

Figure 5. Kaggle dataset display



```
root@ubuntu:/home/kambey/Spam-mail-filtering-master# python3 Spam2.py
Category
ham      4821
spam      747
dtype: int64
root@ubuntu:/home/kambey/Spam-mail-filtering-master#
```

Figure 6. Kaggle dataset classification

The Python code snippet used to show dataset classification is:-

```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import neighbors
data = pd.read_csv ('spamham.csv')
data1 = data.copy()
print(data1.groupby('Category').size())
```

Table 1 presents the dataset comprising of Spam/Scam and Ham with spam/Scam rate shown.

Table 1. Datasets

Dataset Name	**Repository URL**	**Spam/Scam +Ham=Total**	**Rate of Spam/Ham**	**Published Year**
Ling-Spam	http://www.aueb.gr/users/ion/ data/lingspam	591 + 2304 =2895	20%	2000
SpamAssassin	https://spamassassin.apache.org /old/publiccorpus/	1918 + 4379 =6297	30%	2002
Enron dataset	http://www2.aueb.gr/users/ion/ data/enron-spam/	18564 + 18261=36825	50%	2006
Kaggle dataset	www.kaggle.com	747 + 4821 =5568	13%	2012

4. Results and Evaluation

Machine Learning algorithms play a crucial role when it comes to spam and phishing classification. Seven (7) major machine learning algorithms that are used in spam classification were discussed and experimented in this paper. The algorithms that were discussed are evaluated for their performances measure using Python Scikit-Learn tool based on the performance metrics.

4.1. Performance Metrics

A. Confusion Matrix

Though confusion matrix by itself is not a metric for performance evaluation, its components are important for the evaluation of algorithms. As the name suggests, it produces the result in the matrix form and has TP, TN, FP and FN values.

	Actual Phishing	**Actual Ham**
Predictive Phishing	True Positive (TP)	False Positive (FP)
Predict Ham	False Negative (FN)	True Negative (TN)

where:-

- ✓ TP indicates True Positive (correct prediction of positive case),
- ✓ TN indicates True Negative (correct prediction of negative case),
- ✓ FP indicates false positive (incorrect prediction of positive case) and
- ✓ FN indicates False Negative (incorrect prediction of negative case).

B. Classification Accuracy

The classification accuracy metric tells us that how many instances are correctly classified out of the total classified instances

Accuracy = (TP + TN)/(TP + TN + FP + FN)

C. *Precision*

Precision indicates the number of correct prediction of positives (TP) divided by correct prediction of positives and incorrect prediction of positives. This indicates that when a model predicts positive, the precision ensures that the items are correctly labeled as positive. Hence a high precision value shows that the algorithm has returned a relevant result.

Precision =TP / (TP + FP)

D. Recall

Precision indicates the number of correct prediction of positives (TP) divided by correct prediction of positives (TP) and incorrect prediction of negative (FN). This indicates that when a model predicts positive, the precision ensures that the items are correctly labeled as positive. Hence a high precision value shows that the algorithm has returned a relevant result.

Recall = TP /(TP +FN)

Recall finds out the ratio between true positive and the sum of true positive and false negative. This will be helpful when the cost of false negative is high.

E. F1 Score

F1 score is calculated by combining precision and recall to evaluate the overall accuracy of the algorithm. Hence a low false positive and low false negative value gives a good model which has predicted the result accurately. F1score is calculated using the following formula
F1 = 2 X (precision X recall) / (precision + recall).

4.1. Performance Evaluation

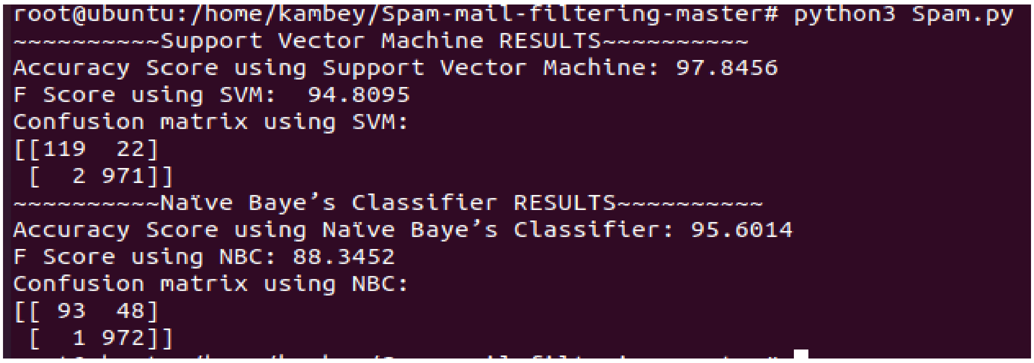
The performance measures of the machine learning algorithms from the datasets presented in Table 2 of this study were simulated and analyzed. The experiment was conducted using the four (4) datasets and the average was taken. Stratified K-Fold Cross Validation (SKFCV) was applied to all the machine learning models to ensure high accuracy since the more the training data, the better accuracy the testing data provides.

The dataset were therefore split into 80:20 for training and test dataset respectively. The results obtained from the algorithms were tabulated in Table 3 below for comparison and it showed all algorithms provided 90% and above accuracy for spam/scam email detection except Random Forest classifier. Amongst the seven (7) algorithms, RF has performed poorly and SVM with optimized version (SGD) is the highest performing algorithm along with MNB that came second. The F-Score that is a measure of a model’s accuracy on a dataset, for SVM is 94.81% which indicating almost perfect precision and recall.

Table 2. Performance Measures for the Machine Learning Algorithms used

ML Algorithm	**Evaluation Metrics**			
	Accuracy (%)	**Precision**	**Recall**	**F1 Score (%)**
Logistic Regression (LR)	91.76	0.54	0.95	68.86
Decision Tree (DT)	92.37	0.57	0.96	71.53
Random Forest (RF)	89.7	0.53	0.97	68.55
Multinomial Naïve Bayes (MNB)	95.6	0.66	0.99	88.35
K - Nearest Neighbor (KNN)	93.24	0.63	0.98	76.7
Support Vector Machine (SVM)	97.85	0.84	0.98	94.81
Multilayer Perceptron (MLP)	94.29	0.67	0.99	79.92

When considering the best two performing algorithms, the confusion matrix, accuracy and F-score measures is shown in Figure 7 using python SKLearn for reference.



```
root@ubuntu:/home/kambey/Spam-mail-filtering-master# python3 Spam.py
~~~~~~~~~~Support Vector Machine RESULTS~~~~~~~~~~
Accuracy Score using Support Vector Machine: 97.8456
F Score using SVM:  94.8095
Confusion matrix using SVM:
[[119  22]
 [  2 971]]
~~~~~~~~~~Naïve Baye's Classifier RESULTS~~~~~~~~~~
Accuracy Score using Naïve Baye's Classifier: 95.6014
F Score using NBC: 88.3452
Confusion matrix using NBC:
[[ 93  48]
 [  1 972]]
```

Figure 7. Performance Metrics for SVM and NB Classifier

5. Conclusions and Future Work

This paper presents a systematic evaluation of machine learning algorithms in detecting malware-based phishing attacks. Through this study, seven machines learning algorithms were used for datasets from four different sources and the averages were calculated. This assisted in selecting the best performing algorithm based on the features considered in detecting a phishing e-mail, Also it helps develop hybrid algorithms through a combination of algorithms as their peer review is made. It is clear from the results that Support vector machines (SVM) outperforms other algorithms including closest rival Naïve Bayes (Multinomial) in detection of spam and phishing mails. Even though it is a small difference compared with MNB that also does a decent job, the better machine should always be used in solving problems such as filtering spam and malware-based phishing mails from ham mails.

As observed from all the models of classification in the field of machine learning, every method that is considered has its pros and cons. In the experiment of this study, the two best performing algorithms took a considerable computational time than the other algorithms although the time depends on the depth of a dataset and the classification. Consequently, for an efficient algorithm to be developed that performs at best even when any parameters like evaluation time, acquaintance cost and the memory of allocation, other parameters should be considered.

Therefore, hybrid algorithms seems to be the best and feasible solution for Spam and phishing detection in e-mails. In order to achieve the best detection performance in organization mail systems, it is often better to have enough training samples with balanced distributions for both malware-based phishing and benign files.

The future work that can be performed in fighting phishing attacks involves enhancing the model with more evaluation parameters for effective spam and malware-based phishing filtering.

REFERENCES

[1] M. Baykara and Z. Z. Gürel, "Detection of phishing attacks," *6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding*, vol. 2018-Janua, no. August, pp. 1–5, 2018, doi: 10.1109/ISDFS.2018.8355389.

[2] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN–LSTM model for detecting phishing URLs," *Neural Comput. Appl.*, vol. 0123456789, 2021, doi: 10.1007/s00521-021-06401-z.

[3] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 139–154, 2021, doi: 10.1007/s11235-020-00733-2.

[4] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *J. Netw. Comput. Appl.*, vol. 153, no. November 2019, p. 102526, 2020, doi: 10.1016/j.jnca.2019.102526.

[5] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Front. Comput. Sci.*, vol. 3, no. March, pp. 1–23, 2021, doi: 10.3389/fcomp.2021.563060.

[6] G. RamT and R. Kumar NSR, "Analysis of Phishing in Networks," *Int. J. Sci. Eng. Res.*, vol. 4, no. 9, pp. 196–201, 2013, [Online]. Available: http://www.ijser.org

[7] S. Sagar, Shivani, and V. D. Chakravarty, "Phishing attacks and defences," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 894–897, 2019.

[8] M. Aldwairi, M. Hasan, and Z. Balbahaith, "Detection of drive-by download attacks using machine learning approach," *Int. J. Inf. Secur. Priv.*, vol. 11, no. 4, pp. 16–28, 2017, doi: 10.4018/IJISP.2017100102.

[9] J. Rastenis, S. Ramanauskaitė, I. Suzdalev, K. Tunaitytė, J. Janulevičius, and A. Čenys, "Multi-language spam/phishing classification by email body text: Toward automated security incident investigation," *Electron.*, vol. 10, no. 6, pp. 1–10, 2021, doi: 10.3390/electronics10060668.

[10] M. V. Madhavan, S. Pande, P. Umekar, T. Mahore, and D. Kalyankar, "Comparative analysis of detection of email spam with the aid of machine learning approaches," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012113.

[11] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020, doi: 10.1109/ACCESS.2020.2969780.

[12] S. Phomkeona and K. Okamura, "Zero-day malicious email investigation and detection using features with deep-learning approach," *J. Inf. Process.*, vol. 28, pp. 222–229, 2020, doi: 10.2197/ipsjjip.28.222.

[13] N. A. Azeez and A. A. Ajayi, "Performance evaluation of machine learning techniques for identifying forged and phony uniform resource locators (URLs)," *Niger. J. Technol. Dev.*, vol. 16, no. 4, pp. 155–169, 2019, doi: 10.4314/NJTD.V16I4.2.

[14] K. F. Rafat, Q. Xin, A. R. Javed, Z. Jalil, and R. Zeeshan, "Evading obscure communication from spam emails," vol. 19, no. November, pp. 1926–1943, 2021.

[15] S. Naaz, "Detection of phishing in internet of things using machine learning approach," *Int. J. Digit. Crime Forensics*, vol. 13, no. 2, pp. 1–15, 2021, doi: 10.4018/IJDCF.2021030101.

[16] S. C. Et. al., "A Survey on Machine Learning Approach to Detect Malware," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 2, pp. 2309–2314, 2021, doi: 10.17762/turcomat.v12i2.1961.

[17] T. D. S, S. Nithya, S. P. G, and E. Pugazhendi, "Email Spam Detection and Data Optimization using NLP Techniques," vol. 10, no. 08, pp. 38–49, 2021.

[18] U. De Barcelona and V. Carvalho, "EMAIL FRAUD CLASSIFIER USING MACHINE LEARNING," 2020.

[19] A. Malero, “Applying feature transformation using Relative Frequency with Power Transformation and Lemmatization in automatic Spam Filtering,” vol. 2, no. 10, pp. 21–27, 2014.

[20] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[21] S. Nandhini and D. J. Marseline, “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection,” Feb. 2020. doi: 10.1109/ic-ETITE47903.2020.312.

[22] S. Gibson, B. Issac, and S. Member, “Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms,” vol. 8, 2020, doi: 10.1109/ACCESS.2020.3030751.

AUTHORS

Kambey L. Kisambu is currently pursuing Msc in Cyber Security at the University of Dodoma, Tanzania. He earned his Bsc. Degree in Telecommunications Engineering in 2010 at the University of Dar es Salaam. His current Research interests include but are not limited to Cyber Security and Forensics, Ethical Hacking, Malware Analysis, Machine Learning and secure e-mail communication.

He has worked on several independent projects such as system development, System review, ICT security and vulnerability assessment.

He has been working in ICT industry and has full work experience for more than 10 years. He can be reached at kambeylk@gmail.com

Dr. Mohamedi Mjahidi is currently a Lecturer in the Department of Computer Engineering and Applications at the College of Informatics and Virtual Education of the University of Dodoma in Tanzania. He earned his PhD degree from the School of Natural and Applied Science at GAZİ University, Turkey in 2020. His research and teaching interest areas include but are not limited to Artificial Intelligence and Machine Learning.

He has published various research papers and he can be reached at mmjahidi@gmail.com