



AI/ML-Powered Phishing Detection: Building an Impenetrable Email Security System

Shravan Kumar Rajaram^{1*}, Siddharth Konkimalla², Manikanth Sarisa³, Hemanth Kumar Gollangi⁴, Chandrakanth Rao Madhavaram¹, Mohit Surender Reddy¹

¹Microsoft Support Escalation Engineer

²Adobe Inc. Sr Network Development Engineer

³Principal Software Engineer, Ally Financial Inc

⁴Department of Computer Science, Southeast Missouri State University.

*Corresponding Author

Shravan Kumar Rajaram

Microsoft Support Escalation Engineer.

Article History

Received: 08.10.2023

Accepted: 19.11.2023

Published: 06.12.2023

Abstract: The most extensive threat type, for the present moment at least, remains phishing attacks that rely on people's susceptibility to trick them into sharing personal data. Such attacks generally consist of fake emails originating from an apparently reliable source, for instance, businesses, banks or government facilities. While normal forms of email filtering prove somewhat useful in the fight against phishing, the techniques are highly unlikely to catch the modern-day complex phishing channels such as zero-day phishing or spear phishing. As a result of this, there is a growing uptake of applying Artificial Intelligence (AI) and Machine Learning (ML) to address email security. They are capable of training on sampled volumes of emails and then using this training to improve the recognition of phishing and non-phishing instances.

Here, we propose to work on a system of phishing detection using AI/ ML, which will be instrumentally crucial in making email security reliable and adaptable. The system employs Random Forest, Support Vector Machines (SVM) as well as Neural Networks to classify the emails based on the features extracted from the subject, body as well as links of the emails. By employing both phishing and innocuous email corpus, we trained and tested these models for the purpose of understanding the viability of phishing identification. The achievements of the study were the improved accuracy of detection when compared to conventional approaches and the further reduction of misrecognition, which improves security in general. It should be noted that by integrating a multi-model approach with learning mechanisms, the proposed system is indeed versatile and strong against advanced phishing threats.

Keywords: Phishing detection, Machine Learning, Email Security, Random Forest, Support Vector Machine, Neural Networks, Feature Extraction.

Cite this article:

Rajaram, S. K., Konkimalla, S., Sarisa, M., Gollangi, H. K., Madhavaram, C. R., Reddy, M. S., (2023). AI/ML-Powered Phishing Detection: Building an Impenetrable Email Security System. *ISAR Journal of Science and Technology*, 1(2), 10-19.

1. Introduction

Phishing is a type of cyber-attack founded in social engineering, which involves scams that impersonate parties in order to create trust from the users and then gain access to their information. Most of these attacks are done through email messages, which will look like they are from a trusted source such as a bank, an online service provider, or a government institution. The aim of most phishing often targets the retrieval of sensitive information such as user names and passwords, credit card details, and identification numbers. [1-4] occasionally, it may also be used to drop malware or ransomware and thus lead to system compromise or result in loss of money.

Email continues to be one of the most popular media for communicating in organizations, making it vulnerable to cybercriminals. Due to the frequency of contact via digital means and the growing complexity of the different layers of phishing, email security has become a major concern for individual users, organizations, and governments across the world. Cybercriminals have changed traditional forms of phishing scams to more sophisticated forms like spear phish, where the scammer directly targets a person or an organization. Such occurs can simply evade traditional security measures totally leading to extensive losses and reputational loss.

This is why it is mandatory that with the changes in phishing tactics, there should be a corresponding change in the systems used to combat the problem. In fact, those preprogrammed rule-based

filters used in conventional email security solutions prove utterly inadequate in the identification and blocking of sophisticated phishing attempts. Such filters tend to be programmed to look for specific attributes of phishing emails, including well-known malicious keywords, other domains or attachments. However, here modern phishing tactics are more diverse and can include, for example, polymorphic emails, where the content and the sender's data change each time, and zero-day vulnerabilities. Phishing attacks are, therefore, likely to bypass these static filters and, hence, the continued need for intelligent adaptive filter systems.

1.1. Problem Statement

The problem is that for rapidly developing phishing campaigns, conventional mail filters seem insufficient, though their effectiveness steadily decreases with time. These systems normally operate on the basis of a set of rules or blacklists in order to signal which possible emails are dangerous. However, this approach does not suit unidentified and new forms of threats since sophisticated phishing has been developed to avoid detection. For instance, the attackers may send emails that are in the name of trusted brands or organizations, though they may be slightly different in the domain name or format of the message that they come in. Furthermore, attackers often set polymorphic phishing that changes the look or the style of the email with every attempt to get by the filters. These emerging threats pose a real problem for rule-based systems that cannot develop detection mechanisms fast enough.

Traditional systems also lack the ability to detect a brand-new phishing attack, one that targets a fresh yet unknown vulnerability that cannot be protected with patches provided by hackers. The main reason is that the first three architectures are rule-based systems and thus could not protect against the attacks since their models do not include new attack patterns. Therefore, those organizations using only these methods stand to be easily compromised by phishing malware.

To fill these gaps, there is, therefore increased call for intelligent and context-aware systems to detect phishing in real time. This is where the AI/ML-based detection systems come into the picture. In this case, AI/ML models have access to very large datasets of past emails, which have been both legitimate and phishing in nature, allowing them to learn the characteristics of phishing attempts. In the long run, it can enhance its capacity to identify different types of phishing techniques and thus is much more reliable than rule-based systems. AI/ML-based solutions have an inherent advantage over other forms of solutions by their capacity to learn from newly received data.

1.2. Objectives

This research aims to ensure the development of AI/ML tools for detecting phishing emails with reasonable accuracy. The system shall be able to scan and categorize the received emails according to predefined features that the system will derive from the emails and all the links contained therein. In this case, the email will be subjected to a machine-learning algorithm that identifies patterns of phishing, while the system should rarely flag a genuine account as a phishing site.

The following specific objectives guide this research:

- **Develop an AI/ML-based phishing detection system:** The use of classification algorithms will be employed by

this system at various levels of the ML process for emails. Some of the models to look into include; Random Forest, Support Vector Machines (SVM) and Neural Network.

- **Enhance phishing detection accuracy:** A number of methods that will be employed include the classification of many-phishing and legitimate emails to help the system differentiate between communication that is genuine and communication that is a form of cyber-attack.
- **Minimize false positives:** Another requirement is to minimize the number of false positives, meaning that legitimate messages should not be classified as phishing. This is important to ensure that people have confidence in the system so that email interruptions are a thing of the past.
- **Evaluate the performance of various ML models:** This study will work toward comparing the various models of supervised machine learning on parameters such as Precision, Recall, and F1-Score and accuracy. The result of this evaluation will determine the appropriate model for more efficient phishing detection.

The principal goal of the proposed system would be to achieve these objectives and provide a solution to the weaknesses of current email security systems with robust solutions to current and future phishing threats.

2. Literature Survey

2.1. Phishing Detection Techniques

Phishing detection has been a hot topic in cybersecurity research for a long time because of the continuously evolving threats of correlated phishing. [5-9] These techniques can be broadly categorized into three main approaches: A blacklist-based detection approach, a heuristic-based detection approach, and an AI/ML-based detection approach. The advantages and the disadvantages of each method are presented and discussed in light of the effectiveness of dealing with phishing threats.

2.1.1. Blacklist-based Detection

The blacklist-based method basically consists of the creation and synchronization of a list containing the domains, URL addresses, and email addresses reported to be involved in phishing. When encountering an email, the sender's domain, URLs in the body of the email and all other related parameters are checked with the Black List. If a match is found, then the email is marked as phishing and either rejected or placed in a quarantine inbox. This approach is widely used because it is simple and has a comparatively low time complexity of the calculations made during the optimization.

However, blacklist-based detection suffers from a major limitation: It also cannot identify previously unidentified phishing domains in the current environment. The criminals can make up new domains or work with registered legitimate domains to get around blacklists. Further, phishing emails are being sent with polymorphic or dynamic links, which actually change after a short time and are thus not included in static blacklists. The blacklist-based systems are reactive, which makes them incapable of addressing modern, ever-developing styling of phishing threats.

2.1.2. Heuristic-based Detection

Heuristic-based phishing detection involves the use of a defined set of rules/algorithms to detect possible phishing emails. These rules serve the purpose of emulating major characteristics of phishing emails, including subject lines that are likely to be something like “Urgent Action Required,” senders with irregular email addresses like those of the domain they are not a part of, and keywords like “reset password,” “click here,” and “free gift.” To do this, the system scans through the header of the particular email and then the content of the email in order to know whether the particular email has characteristics that will fit these rules of predefined criteria.

As good as heuristic methods are at identifying a simple form of the phishing attack, they fail horribly when testing out more advanced types of phishing attacks like spear-phishing or business email compromise. Heuristics also include false positives the program may detect scraping emails as phishing due to the occurrence of some keywords or miscellaneous patterns in the rules. Furthermore, these rules can be bypassed very easily because all the attacker needs to do is focus on using social engineering to make the phishing emails appear innocent.

2.1.3. AI/ML-based Detection

Due to the shortcomings of the blacklist and heuristic approaches, the present research introduces Artificial Intelligence (AI) and Machine Learning (ML) for detecting phishing URLs. Static solutions, on the other hand, work with a set of rules which is, however cannot beat the AI/ML models that can learn from millions of actual phishing and normal emails. These models are particularly powerful in using historical data to perform a generalization, thus distinguishing the previously unseen phishing attempts.

A few of the techniques that are employed in the detection of phishing sites are Decision Trees, Random Forests Neural Networks SVM and others. These algorithms are based on the features obtained from emails, including the sender’s address, the content of the message, links and any attachments. While the model is being trained, the model is trained to recognize what a fake phishing email looks like in relation to a genuine, legitimate email containing these features. After training we can be able to detect emails with good probability and distinguish them from phishing even with changing tactics. The strengths of AI/ML-based detection are that it can adapt to new styles of attack dynamically, resulting in much fewer false positives and better detection rates for advanced Birthday Attacks.

2.2. Machine Learning Models for Phishing Detection

Several models have been tested when it comes to the use of Machine Learning in Phishing Detection. As has been seen, each model has its characteristics in regard to the type of data it can handle and the level of complexity. Below are some of the most commonly used machine learning models for phishing detection:

2.2.1. Random Forest

Random Forest is an Ensemble learning method in which training creates multiple decision trees, and the number of instances of each class (phishing or not phishing) given by individual trees is returned. In decision forest computing, every tree in the forest is

trained with different data, and where data split in every node of the tree is founded on the most influential component at that level. Random Forest has wide applicability wherein it approves a diverse number of features. At the same time, the assembly of trees minimizes the possibility of over-learning, making it the best model for naïve determination of phishing.

Random Forest models are appropriate to use in classification problems especially with huge data features of emails. The model invented in the decision-tree algorithm combines many decision trees, and this enhances accuracy and less vulnerability compared to individual decision trees. However, Random Forests models may be computationally intensive or expensive especially when working with big data sets.

2.2.2. Support Vector Machines (SVM)

SVM stands for Support Vector Machines, which is a strong classification technique based on the concept that in a high dimensional space, product features can carefully segregate distinct categories by identifying different characteristics of varying degrees. For phishing detection, SVM builds the hyperplane that is far from both phishing and non phishing data in such a manner that the boundary between both classes is clearly seen.

SVM is especially useful when the data does not lie on a hyperplane since SVM can use kernels to transform the features into higher-dimensional space where data are separable by hyperplane. Many studies applying SVM models proved that these types of models are accurate in classifying complex phishing datasets characterized by overlapping patterns. However, when it comes to big data or when choosing the right kernel functions, it becomes rather slow, which decreases scalability to actual real-time detections.

2.2.3. Neural Networks

Deep Learning models of Neural Networks have drawn a lot of attention in the context of phishing detection because of high learning capabilities from a data set. Neural networks are a set of interconnected neurons, the layers of which acquire and extract more abstract features of the given data. The technique used in training the network includes making the network learn patterns from the text content of the emails, the headers, and the links as regards detecting potential phishing emails.

CNN and RNN, for instance, can be even more useful when it comes to analyzing the content of the email content turning new shapes and trying to reveal the detail system to find those relations that are not easily recognizable with other simple methods. The capabilities include large-scale data processing, self-organizing and automatic high level feature extraction. However, it operates well, mostly with high computational complexities and often needs large data sets for optimal performance.

2.3. Email Feature Extraction

Feature extraction is one of the most important phases when it comes to phishing emails as it helps the machine learning model to abstract the most significant features that make an email a phishing one or not. Increased feature extraction improves the accuracy of the model for identifying the phishing attack. Below are some common types of features that are typically extracted from emails for phishing detection:

2.3.1. Header Analysis

Email header is information about the mail content including the sender's address, recipient's address, subject line and other routing information. The header of an email can contain fake features, such as the sender's alleged email address, the displayed sender's name, and the path of the email to reach the receiver's mailbox. For instance, if the sender domain is different from the organizations they claim to represent, then it is likely that it is a phishing email. Palmour also points out that header analysis consists of inspecting "Reply-To" addresses, as well as discrepancies in times or IPs.

2.3.2. Content Analysis

The reality is that content analysis focuses on the main section of the email, searching for links, URLs, attachments, and specific language patterns of a spam message. Some pharmaceutical drugs used in Kamphaeng Phet belong to groups of medicines based on the source of the product, bought through phishing emails that contain links directing the user to fake websites. Some of these URLs may seem similar to actual domain names but have slight typographical errors or slight modifications (like 'g00gle.com' instead of 'google.com'). Moreover, these emails contain certain specific linguistic features that are considered stimuli of urgency and importance, like the words 'act now' and 'your account has been hacked'. Content analysis can also reveal the presence of virus files like .exe files or even a Word document with a macro virus.

2.3.3. Domain Features

The domain of the sender is a factor that helps in the identification of the phishing spree. A lot of attacks involve the creation of domains that look like credible ones with the intention of fooling the recipient. For example, instead of getting an email from "example.com", a phishing email would be from "examp1e.com". By measuring the similarity between the sender domain and trusted domains, machine learning models can be trained to identify the domain spoofing attacks. Furthermore, it is possible to compare the domain to a positive list of allowed domains or a negative list of domains belonging to phishers.

When the header analysis and content analysis are joined with the domain features, the machine learning models are able to create a complete picture of each given email. They thus are able to distinguish between phishing and legitimate with unblemished efficiency. This multi-faceted teaching approach helps make sure that the detection system will be ready for any kind of phishing attacks and new enterprises.

3. Methodology

3.1. Dataset Collection

It is clear from the results that the quality and the variety of the data can significantly affect the accuracy of the established anti-phishing model. [10-12] In this research, a data set of a combination of phishing and legitimate emails was obtained from public data sources, which are normally used in similar research.

3.1.1. Data Source

- **Phishing Emails:** 10000 phishing emails were downloaded from popular, trusted phishing datasets, including PhishTank, a website that hosts a large database of authenticated phishing sites and emails. Such emails consist of different types of phishing schemes concerning people and companies, with texts aimed at pilfering confidential data.
- **Legitimate Emails:** From the Enron Email Dataset, a large publicly available dataset for email analysis, 15000 legitimate emails were used. Enron is a genuine collection of corporate email messages. It is a rich source, presenting various types of emails, including personal, business, and organizational, which is very appropriate for building a competency-based system.

3.1.2. Data Preprocessing

To improve the quality of the collected data and to eliminate any discrepancies in the data, the dataset was preprocessed. The steps involved:

- **Data Cleaning:** Deletion of redundancies and unsuitable data and noise that is present in the database. This helps to balance equally between the number of both phishing and genuine emails.
- **Handling Missing Data:** To address the problem of missing values, including missing email content or email headers, some values were imputed, while all the samples with missing values were removed.
- **Normalization:** The text content of the emails was preprocessed by using lowercase in all text, using words or symbols that were not letters of the alphabet as backslash and using common terms where the words looked to be similar to something else, e.g. all the URLs in the text were replaced with token.url.

Table 1: Sample Data Features

Feature	Description	Type
Sender Email	Email address of the sender	Categorical
Subject	The subject line of the email	Text
Body Text	The main content of the email	Text
Hyperlinks	URLs embedded within the email	Categorical
Attachment Type	File type of any attachments in the email	Categorical

3.2. Feature Extraction

The next step after preprocessing was to choose features that would be used in training the machine learning algorithm to detect phishing. These features fall into three major categories: textual, domain, and behavioral features.

3.2.1. Textual Features

Textual features relate to the content of the analyzed email. These features can assist in distinguishing between phishing mail from non-phishing-mails through analysis of the language employed, which is normally a lure.

- **Tokenization:** Building tokens from the email body and the subject line as the individual words it comprises. More familiar approaches, such as the bag-of-words and term frequency-inverse document frequency (TF-IDF), were applied to the text inputs to transform the textual data into numeric values.
- **Keyword Detection:** Keywords that might be typically used in phishing emails, including “click here”, “urgent”, and “account locked”, were also found. If any of these keywords show up, then it would mean that the user probably has malicious intent.
- **Email Subject Length:** The length of the subject of the email was defined as the phishing campaign consisting of a short and catchy title.

3.3. Behavioral-Based Phishing Detection System

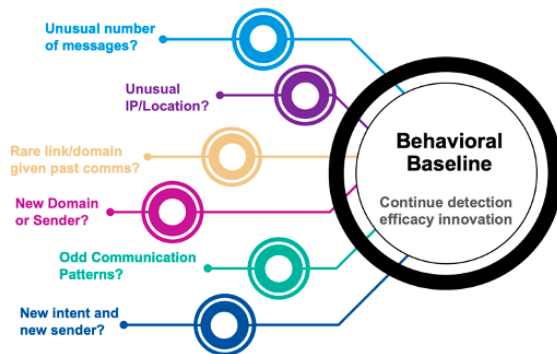


Fig. 1. Behavioral-Based Phishing Detection System

The image illustrates a behaviour-based anti-phishing that is based on the detection of the variation between the new activity and a baseline. [13] This approach employs several behavioral parameters to identify incidents of phishing; as such, it proves to be a strong appendage of the phishing detection framework.

3.3.1. Behavioural Baseline Abnormalities

The left section highlights specific anomalies that might trigger a phishing alert based on deviations from established behaviors:

- **Unusual Number of Messages:** If one starts receiving messages from the sender in an enormous number at one time or over a short period, then it may be a phishing attempt.

3.2.2. Domain Features

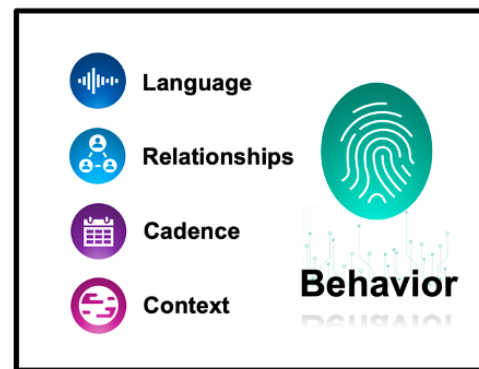
Domain-based features focus on the email’s sender and embedded URLs.

- **Sender Domain Analysis:** Checking if the domain of the sender matches known legitimate domains (e.g., official company email addresses). Suspicious domains (e.g., slight misspellings or variations of popular domains) can signal phishing attempts.
- **URL Features:** Analyzing the URLs embedded in the email, such as detecting if the URL redirects to known phishing sites or checking if the domain has a low reputation score.

3.2.3. Behavioral Features

These features are based on the behavior and patterns associated with emails:

- **Forwarding Patterns:** Messages that have been relayed several times before they get delivered to the target are usually an indication of phishing attacks.
- **Time of Receipt:** It is common for the phisher to send the email at odd hours or during a time when a normal business or organization is not operating.
- **User Engagement Metrics:** Analyzing how recipients interact with emails, such as how quickly they open or click links, can provide insights into phishing attempts.



- **Unusual IP/Location:** If the sender of the emails has another location or IP address, then it can be a suspicious account, and it may also be phishing.
- **Rare Link/Domain:** If the link or domain provided in the email is not familiar to the user in earlier communication with the sender, this could become suspicious.
- **New Domain or Sender:** The nature of a new sender or a new domain, when they are considered collectively with three other features, can be used as a sign of a phishing attack.
- **Odd Communication Patterns:** In other words, any change in the tone, language or mode of communication of the sender may indicate that a particular account has been compromised.

- **New Intent and New Sender:** Another symptom of a phishing attack is a shift in the kind of message, for instance, from a friendly greeting to the need to address issues of utmost importance or new urgent financial needs coupled with the sender's new stranger identity.

3.3.2. Behavioral Detection Factors

The right section complements the left by showing how different factors contribute to understanding email behavior:

- **Language:** Review the context of the email for any irregularities; in particular, read the language the author has specifically used to send an email.
- **Relationships:** Assessing the already set relation between the two users, the sender and the recipient. For example, phishing attempts are usually received from unknown or even unwanted addresses.
- **Cadence:** Considering when and how often I need to send an email. An increase in the frequency of messages could be another sign of abuse, says Pam.
- **Context:** Assessing the content of the email according to previous correspondence to determine if the email corresponds to an ongoing conversation or contains issues unrelated to previous discourses.

3.3. Model Development

The gathered preprocessed data, along with the extracted features, were employed to train different machine learning algorithms. Such models were chosen because it has been found that they are particularly useful when tackling multi-class classification problems, with data entering the models in a variety of formats.

3.4. Phishing Detection System's Workflow

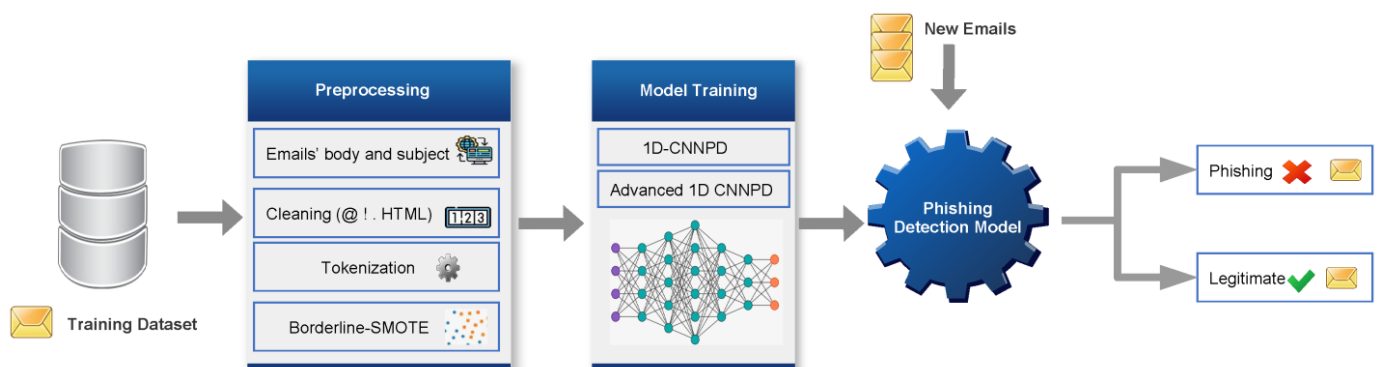


Fig. 2. Phishing Detection System's Workflow

The overall information on the work of the phishing [15] detection system is illustrated in the image as follows; the figure can be elaborated on as follows:

3.4.1. Training Dataset (Input)

The system starts with a training set, which is a set of labelled emails, both phishing and non-phishing. This data is utilized for the construction and training of the machine learning model for the identification of phishing.

3.3.1. Random Forest Classifier

Random Forest is a flexible method derived from learning the decision trees. It creates several models from different subsets of the data and combines their results to raise a prediction of whether the given email is a phishing one or a genuine one. This method is very accurate and less sensitive to noise if compared to the other approaches to image analysis. Moreover, every tree in the forest decides the classification and the result is the ones that have the most votes.

3.3.2. Support Vector Machine (SVM)

SVM, in fact, draws hyperplanes in a high-dimension space in order to differentiate phishing messages from other legitimate messages. It is quite suitable for email classification because it is also effective with both linear and nonlinear data. SVM, on the other hand, uses kernel functions to transform features into a higher orthogonal space in order to increase the regions of the decision boundary. In this regard, the current study employed the radial basis function (RBF) kernel, which addresses non-linearity in the contextual features of emails.

3.3.3. Neural Networks (Multi-Layer Perceptron)

An MLP is a feed-forward neural network, and this type of ANN is made up of layers of neurons. The network learns the complex relationship between the input attributes and the target (phishing or legitimate). The network created by Matplotlib noted here has the input layer, one and/or more hidden layers, and the output layer. The hidden layers apply activation functions such as ReLU (Rectified Linear Unit) to map the patterns in the data set.

- **Source:** Mails either from mail datasets such as Enron corruption, PhishTank or normal mails that are actually composed by the user.
- **Content:** The data set, therefore, incorporates features such as the body of the email, subject area, the meta data incorporated as well as the details of the sender that are relevant in the classification exercise.

3.4.2. Preprocessing Stage

Even before training the model, emails go through various steps of data cleaning and preparatory stage to make the content in the right format for the model.

- **Email Body and Subject Extraction:** To this end, the text only or the body of the email, including the subject lines, will be extracted and analyzed.
- **Cleaning (Removing noise):** These include removing special characters/ symbols like @,!, and HTML tags on the text in order to have a cleaner text data set. It helps to attenuate noises in the dataset as a result.
- **Tokenization:** The material of the email is preprocessed, so it is converted into pieces that are analyzable, such as words or phrases.
- **Borderline-SMOTE (Synthetic Minority Over-sampling Technique):** This procedure is applied in order to mitigate the problem of class imbalance of the dataset with respect to phishing and legitimate emails. The Borderline SMOTE comes up with synthetic samples for the minority class (phishing emails) by synthesizing data around the decision-making border.

3.4.3. Model Training

Preprocessing, standing for cleaning and preparing data for analysis, is followed by training of the model itself, which seems to classify between phishing and legitimate emails.

- **1D-CNNPD (1-Dimensional Convolutional Neural Network for Phishing Detection):** It has been observed that using a 1D Convolutional Neural Network (CNN), the phishing detection model is trained. CNNs are particularly important in processing sequential data, such as text, and are, therefore, highly suitable for use in phishing detection. The 1D-CNN analyses the features of the emails while at the same time looking for patterns.
- **Advanced 1D CNNPD:** For better and more accurate models and better detection, a new, improved version of CNN may be affected, say through the integration of deeper layers or other forms of optimization of the current CNN.

3.4.4. Major Model type done is the Phishing Detection Model (Classification)

Once trained, Phishing Detection Model is now capable of labeling arriving mail either as phishing or otherwise legitimate.

- **New Emails (Testing Data):** The unlabeled emails are fed through the developed phishing detection model to categorize the emails.
- **Phishing (Flagged with a red “X”):** This means that emails referred to as phishing are categorized.
- **Legitimate (Green checkmark):** Any email that is considered as being genuine is automatically passed through without triggering any caution.

For the purpose of phishing detection, the overall system architecture of the system was made more modular and multilayered. Every layer performs a distinct task; data input is

processed, features are extracted, and finally, the data is classified. The architectures make sure that incoming and outgoing messages undergo the evaluation for phishing in a disciplined manner so that the phishing threats are detected as they come, without much of a lag.

3.5. System Architecture



Fig. 3. System Architecture

- **Input (Email):** The system imports emails from the server into the program.
- **Preprocessing:** First, the emails go through preprocessing where noise is eliminated, and the text is standardized.
- **Feature Extraction:** Headings, body, links, and attachments constitute each email and are used for subsequent analysis in the current step.
- **Model Classification:** The extracted features are then used by trained machine learning models that distinguish between phishing and a legitimate email.
- **Output (Phishing/Legitimate):** At the end of the modelling process, the received email is classified as phishing or legitimate. Phishing emails are either deleted and quarantined or marked as suspicious and possibly containing malware.

The architecture guarantees that the system is adaptable, fast, and can manage a great number of emails while offering maximal detection precision. It is also highly extendible to accommodate changes, as rules of phishing evolve with time in the architecture.

4. Results and Discussion

4.1 Model Evaluation Metrics

To measure the performance of each machine learning model used in phishing detection, we evaluated them using four common metrics: Precision, Recall, F1 score, and amount predicted correctly. These metrics were chosen due to the fact that they allow for a comprehensive evaluation of circumstances under which the given model successfully classified the letters while maximizing the accuracy of the results which classify a letter either as a phishing or legitimate one.

- **Precision:** It shows the ratio of the actual quantities of the phishing emails actually received by the filter to all the emails it flags as a phishing email. When accurate, much fewer genuine messages end up classified as phishing.
- **Recall:** The ratio of “got caught” bots regarding all actual phishing emails. High recall means that the system is not

failing to filter many phishing emails or letting many through, if you will.

- **F1-score:** The average of the relative precision and the relative recall. This metric quantification breaks the tradeoff between precision and recall as a metric.

- **Accuracy:** In this case, the percentage of general accuracy of identified emails, including both phishing and genuine ones.

A split 80:20 of the dataset was created in order to train the model and then test it. The training set was employed to train the machine learning models, and the testing set was employed to determine how well the models perform on unseen data.

Table 2: Model Performance

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.96	0.95	0.96	96.4%
Support Vector Machine (SVM)	0.94	0.92	0.93	93.5%
Neural Networks	0.98	0.97	0.98	97.8%

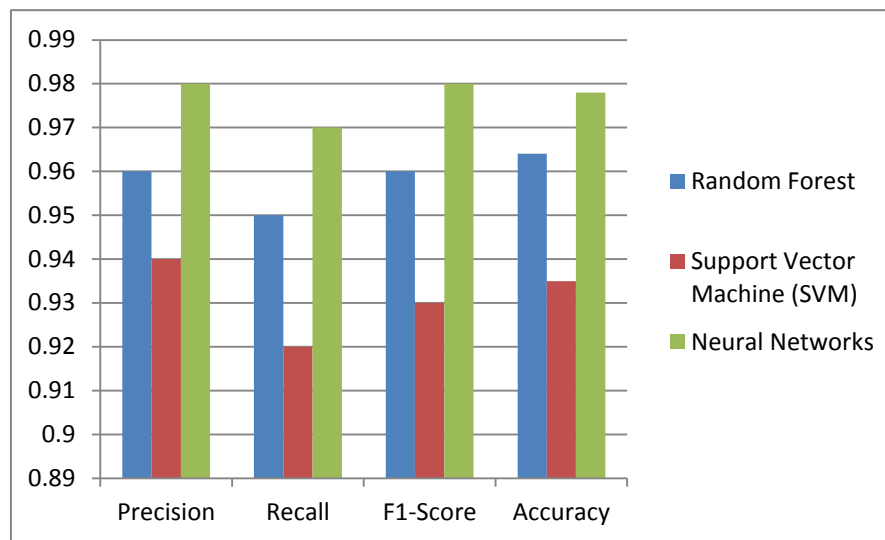


Fig. 4. Graphical Represented Model Performance

4.2. Discussion

All of the models assessed here were able to classify phishing and legitimate emails with fairly good accuracy, but some distinct features and behaviors distinguish the models.

4.2.1. Random Forest

- **Performance:** The accuracy of the Random Forest classifier that was obtained was 96.4%, the precision was 0.96, and the recall was 0.95. F1-score was estimated at 0.96, which suggests equal accuracy and recalling ability of the classifier.
- **Strengths:** The strength of Random Forest is in an approach used for building the model called ensemble learning. Through ensemble, decision trees can work with high-dimensional and noisy data by taking the average of the prediction of multiple decision trees. This enables it to work by making sound predictions even when the features have some interactions.
- **Limitations:** The main drawback of the proposed method was its training time. Based on the last conv layer of the Random Forest model, it was observed that it is best at building several decision trees, and hence, it took the longest time to train among the developed models. With the exception of the above

shortcoming, it fared well on the test data and successfully classified the phishing emails.

4.2.2. Support Vector Machine (SVM)

- **Performance:** In terms of accuracy, SVM had 93.5%, Precision: 0.94, and Recall: 0.92. The F1-score was 0.93, which means that the classification performance of the method presented in this study is quite stable.
- **Strengths:** SVM was able to classify phishing emails in situations that had a linear or near-linear decision space. The method works well, especially if a clear line can easily be drawn between malicious and genuine emails.
- **Limitations:** SVM had certain issues with nonlinear features that are present in the data set, so it ranked lower than Random Forest as well as Neural Networks. In situations where attackers employed better subterfuge like using the sender name, the domain name of the sender or deploying dynamic content, then SVM was not very effective at discriminating between phishing and normal emails, hence the high false negative rates.

4.2.3. Neural Networks

- **Performance:** Neural Networks gave better results than both the Random Forest and SVM; the accuracy was 97.8%, the

precision of 0.98, and a recall score of 0.97. F1-score was the highest of all models at 0.98.

- **Strengths:** The fact that the Neural Network model is capable of programming nonlinear relationships within large data sets played a big part in the enhanced accuracy. It has a deep learning infrastructure that makes the algorithm able to extract features of high levels from data to capture complex phishing attempts. This model also worked well in cases where all other models did not work, that is when it is used in detecting phishing emails with obscured URLs and emails with rich embedded language.
- **Limitations:** One major vice of Neural Networks is that it is computationally very intensive. Training deep neural networks is computationally intensive and time-consuming, particularly when data is huge. In addition, the results of Neural Networks are usually overfitted more often than not; however, in this study, the technique of dropout was employed to control this aspect.

4.3. Comparative Analysis

From the comparative evaluation, it was established that Neural Networks gave the highest average on all the evaluation criteria compared to the other models. The results showed that our model achieved an overall accuracy of 97.8%, with the best measure for precision and recall. This high performance shows that Neural Networks perform very well in the task of phishing detection, particularly where the characteristic of phishing attempts is more elaborate or advanced.

- **Random Forest:** Despite the Random Forest model proving useful and performing generally quite optimally, it was not as time or space-efficient as Neural Networks.
- **SVM:** The training done with the help of SVM was considerably good, but the performance of the model was not good on the twisted data structure and evaluated less than both Random Forest and Neural Networks in terms of accuracy as well as recall.
- **Neural Networks:** This model performed very well, and specifically, it had the fewest false negatives, which is important for the detection of phishing emails.

4.4. Practical Implications

The results also show the impact of combining AI/ML-driven systems to strengthen email security against phishing threats permanently. A high number of True Positives, True Negatives, and F1-score prove the high effectiveness of the chosen approach in tackling most types of phishing attacks, starting from the conventional ones and up to APT, which implements the spear-phishing emails.

- **Zero-Day Threats:** Most of the time, standard detection system falls short of preventing zero-day attacks – such as newly developed phishing attack that exploits unaddressed security loopholes. AI/ML-based systems, particularly Neural Networks, can identify these threats since they are designed to learn from the data itself, not from the dictionaries of threats.
- **Reduced False Positives:** A major problem that phishing detection systems face is the false-negative rate, which occurs

when the system defines a legitimate mail as a phishing one. Based on the high accuracy of the Neural Network model, the authors argue that the false positives are easily distinguishable from the phishing ones and, therefore, will not overwhelm the users through the long list of alerts to check.

- **Scalability:** Thanks to the constant increase in the traffic of incoming emails, including within enterprises, it is critical for the AI/ML models to be scalable. Although the Neural Networks require higher computational time, they are scalable in terms of performance with big data and depict an enhanced ability to increase detection rates.

In conclusion, this research validates that the applications of AI/ML incline, particularly Neural Networks, have a more accurate and efficient means of detecting phishing procedures. Based on these findings, there is an endorsement for the utilization of an email security solution powered by artificial intelligence to protect users from new generations of sophisticated phishing threats.

5. Conclusion

This paper thus presented a generic framework for an AI/ML-based phishing detection system that, in the following subsections, incorporates machine learning algorithms that distinguish between phishing and real emails. The purpose of the current study was to create an improved, more accurate anti-phishing tool that would enable detection of phishing emails without raising False Positive rates. In the current study, they have adopted an extensive assessment of many algorithms of the machine for different models such as Random Forests, SVM, and Neural Networks, which clearly shows that AI/ML-based techniques are far superior to rule-based approaches where most of the time the traditional techniques are unable to track the ever-evolving dynamic nature of phishing techniques.

However, the Neural Network model revealed the highest performance with an accuracy of 97.8% and was considered the best among all other models in terms of evaluation metrics like Precision, Recall, and F1-score. The reason for this superior performance can be credited to the fact that it can model nonlinear relationships in the data where many nonlinear relationships are camouflaged and hard to identify even by highly trained individuals; the feature vector corresponding to a phishing attempt often includes the use of zero-day tactics. Random and SVM models also had reasonably good accuracy but failed to meet the desired accuracy as there were other complex intricates that could not be handled other than Neural Networks.

This research has found that phishing detection systems based on AI/ML can improve the effectiveness of email security by minimizing the threats connected with phishing attacks, including identity theft, data leakage, and financial fraud. Furthermore, the system's ability to respond to new types of phishing attacks without frequent updates by the operator makes the system an excellent tool to use for organizations and people.

5.1. Future Work

Elaborately, the current study reveals that AI and ML models hold an impressive promise to improve the structure of phishing detection; the findings also indicate that NN has achieved a considerably higher detection rate than other models because of its potential to learn the data pattern better than other models.

However, the system demonstrated acceptable accuracy, and improvements could be considered in further works, for example, the use of LSTM or BERT models. Due to its capability of dealing with sequential data, LSTM can correctly analyze the flow of content in an email as well as its context. In contrast, BERT, with its enhanced NLP, might help in detecting more state-of-art phishing messages that mimic legitimate ones. These models could further advance the cause of phishing detection beyond its current state, especially in a situation where sophisticated or more nuanced phishing scams are afoot.

Furthermore, the incorporation of NLP features, including sentiment analysis, POS tagging and semantic analysis, can help to improve phishing detection as these features will allow the system to grasp the contexts of the content behind the emails. Another potential area for enhancing the performance is real-time phishing detection, that could become useful if users were notified of threats as soon as possible. Possibly by embedding learning features and also cooperating with threat intelligence providers, the phishing detection system could remain aware of fresh hazards and refine pin pointing strategies to tend to new and superior types of phishing threats to user accounts.

References

1. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning-based phishing detection from URLs. *Expert Systems with Applications*, 117, 345-357.
2. Srinivasa Rao, R., & Pais, A. R. (2017, April). Detecting phishing websites using automation of human behavior. In *Proceedings of the 3rd ACM workshop on cyber-physical system security* (pp. 33-42).
3. Marchal, S., François, J., State, R., & Engel, T. (2014). PhishStorm: Detecting phishing with streaming analytics. *IEEE Transactions on Network and Service Management*, 11(4), 458-471.
4. Abdelhamid, N., Thabtah, F., & Abdel-Jaber, H. (2017, July). Phishing detection: A recent intelligent machine learning comparison based on models content and features. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 72-77). IEEE.
5. Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials*, 15(4), 2091-2121.
6. Ghafir, I., & Prenosil, V. (2015, April). Blacklist-based malicious ip traffic detection. In *2015 Global Conference on Communication Technologies (GCCT)* (pp. 229-233). IEEE.
7. Satoh, A., Nakamura, Y., Nobayashi, D., Sasai, K., Kitagata, G., & Ikenaga, T. (2019). Clustering malicious DNS queries for blacklist-based detection. *IEICE TRANSACTIONS on Information and Systems*, 102(7), 1404-1407.
8. Mendonça, L., & Santos, H. (2012, October). Botnets: a heuristic-based detection framework. In *Proceedings of the Fifth International Conference on Security of Information and Networks* (pp. 33-40).
9. Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. *arXiv preprint arXiv:2009.11116*.
10. Basnet, R., Mukkamala, S., & Sung, A. H. (2008). Detection of phishing attacks: A machine learning approach. In *Soft computing applications in industry* (pp. 373-383). Berlin, Heidelberg: Springer Berlin Heidelberg.
11. Zareapoor, M., & Seeja, K. R. (2015). Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business*, 7(2), 60.
12. Carmona-Cejudo, J. M., Baena-García, M., del Campo-Avila, J., & Morales-Bueno, R. (2011, April). Feature extraction for multi-label learning in the domain of email classification. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 30-36). IEEE.
13. AI and ML Email Threat Detection, Spambrella, online. <https://www.spambrella.com/ai-and-ml-for-email-threat-detection/>
14. Maleki, N. (2019). A behavioral based detection approach for business email compromises.
15. Gandotra, E., & Gupta, D. (2021). An efficient approach for phishing detection using machine learning. *Multimedia security: algorithm development, analysis and applications*, 239-253.
16. Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, 68, 160-196.
17. Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2009). An evaluation of machine learning-based methods for detection of phishing sites. In *Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part I 15* (pp. 539-546). Springer Berlin Heidelberg.
18. Koval, A., Karlsson, S., Mansouri, S. S., Kanellakis, C., Tevetzidis, I., Haluska, J., ... & Nikolakopoulos, G. (2022). Dataset collection from a SubT environment. *Robotics and Autonomous Systems*, 155, 104168.
19. Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2022). A predictive model for phishing detection. *Journal of King Saud University-Computer and Information Sciences*, 34(2), 232-247.
20. Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1), 41-51.
21. Laclavík, M., Dlugolinský, Š., Šeleng, M., Kvassay, M., Gatíal, E., Balogh, Z., & Hluchý, L. (2011). Email analysis and information extraction for enterprise benefit. *Computing and informatics*, 30(1), 57-87.