# Email Spoofing Detection

SHRUTI GHOSH[1], SPOORTHI R PRAKASH[2], VARSHA BHARDWAJ[3], YASHASVI PRATAP[4]
DR. POOJA NAYAK S [5]

[1234] *Student, Department of Information Science and Engineering, DSATM, Bengaluru-88, Karnataka*

[5]*Faculty Department of Information Science and Engineering, DSATM, Bengaluru-88, Karnataka*

***Abstract—Emails are widely regarded as an essential form of communication in the present age of modernization, serving both personal and professional needs. Information about payments , financial statements, authentication information and card payment reports are just a few examples of the sensitive and personal information that is frequently routed over email. Emails are valuable to hackers. As a matter of fact, information in emails can be utilized for evil. Cyber attackers are always looking for innovative methods to abuse people for their own benefit. It is vital to remain attentive while receiving an email, whether from an anonymous source, someone close to you, or an organisation you are familiar with. Spam email is an unwanted email that is typically distributed in bulk to a large number of recipients. Email Spoofing is a way of impersonation, in which the scammer sends spam emails using a forged email address. The goal of email spoofing is deceiving users into believing that the email is from someone they know or can trust, for example: a colleague, vendor or brand. Exploiting that trust, the spammer asks the recipient to disclose personal information or take some other action. Spoofers use this method of deception because they know a person is more likely to engage with the content of the email if they are familiar with the sender. Spoofing is commonly used by cyber criminals as an aspect of a phishing attempt. Phishing is a method of acquiring data by impersonating an email id and sending an email that appears to be from a reputable source that may fairly request such information. The idea is to trick victims into selecting a link or downloading a file that will install malware on their device. Spoofing is also linked to domain forgery, which involves using an email id that is identical to another email id. Cyber-attacks have risen significantly in the past couple of years. Spoofed emails have caused significant financial and security breaches. This paper's primary goal is to design and examine several machine-learning algorithms and techniques for spotting fraudulent and spoofed emails.***

**Keywords: spoofing, phishing, spamming, domain forgery, malware, cyber attack, fraudulent**

## I. INTRODUCTION

Online users' experiences have changed significantly because of the Internet's quick growth of technologies. Due to this ever-growing and ever-evolving technology, security concerns are becoming more and more overwhelming. An increasingly popular technique to connect and communicate quickly and effectively is through electronic mail or email. Although it is an asset for the majority of computing, business and educational enterprises, it also poses some security risks that, if not adequately addressed, can have fatal consequences both financially and psychologically. Email technology is vulnerable to things like phishing assaults, spamming, email spoofing, the propagation of viruses and malware. One of these attacks that frequently occurs in email messages is email spoofing. Email spoofing is the process of sending emails from a fraudulent sender address to the victim's email address.

Figure 1: *Protection against Email Spoofing*

Sending a fake email or spoofed email is typically done to make someone or something look bad by misleading them or making them work or do something for the gain of the fraudster. Malware can also be propagated into our machine through these spoofed emails. By disguising these spoofed emails to appear to be from a reputable source, spammers and spoofers can use this method to deliver harmful emails to targeted people. This can be done by building a duplicate website of a trustworthy source. Spoofing is an effort to electronically gain sensitive or secret information from users which is often done with the aim of theft and for other malicious purposes. Spoofing often uses a smart device, an internet connection and it focuses on the flaws in numerous sensing devices brought on by users, who are generally viewed as the weakest link in the network security system. Spoofing emails pose a social engineering risk that could result in the theft of sensitive data gathered via unknowing, naive and trusting individuals.

The short messaging service (SMS), emails, messages, Barcode scanners and phone calls are just a few of the forms of communication that can be used for this purpose. Nevertheless, email spoofing efforts are by far the most common among these. As a result, this study concentrates more on email communications and how to detect spoofed emails. Attackers that use spoofing frequently use carefully prepared messages to trick their victims into revealing sensitive data that will be used by the cyber criminal to gain unauthorized access to the victim's account. These emails contain malicious attachments or links that infect the user's computer with malware. By causing the key operating components of the system to be damaged, malware leads to system failure. Furthermore, the spoofed email may mislead the user and direct them to fraudulent websites, resulting in the theft of sensitive information such as bank account data, login passwords, credit card information and so on. The primary motivation of the crooks behind this sort of fraud is financial gain. When the sender convinces the receiver and obtains their personal information via trickery, it is called spoofing. Even though only a small percentage of receivers fall for the fraud, most of the time, the spoofer sends the email to millions of receivers, which has the potential to yield the spoofer a significant financial profit. With the continuing development of technology and increasing email use, there is a greater chance that sensitive information may be compromised by fraudsters and thus increases the importance of having a good email spoofing detection system.

Companies and organizations are also teaching their staff about security breaches and all of their forms, risks associated with them and how to handle these security breaches. The control activities connected to the loss of confidential data do not exclude the Internet Service Providers (ISP) either. There are a number of techniques to screen and stop dubious emails and to stop sensitive identity theft. People will constantly discover new techniques to lower the hazards involved with spoofing. But, as the crime of stealing sensitive information arises, the fortuity of doing so will also continue to rise. Criminals will constantly come up with fresh schemes to trick people. Although this problem has been addressed by security techniques like Sender Policy Framework (SPF), Sender ID and DKIM (Domain Keys Identified Mail), it is still feasible to simply spoof emails using certain applications. Therefore, it is necessary to create new defences against the email spoofing attack.

Therefore, in this paper, we study the various prevailing anti-spoofing techniques and see which is best for email spoofing detection. In the area of artificial intelligence and machine learning, the system is given the capacity to learn without being specifically designed. Here, the system learns to recognise the spoofed emails by studying various parameters and is trained to have maximum efficiency in detecting the spoofed emails and other emails with malware. This provides increased security to the users against the fraudsters who send such

spoofed emails and allows users to use this mode of communication in a safe and efficient manner without falling prey to email spoofing.

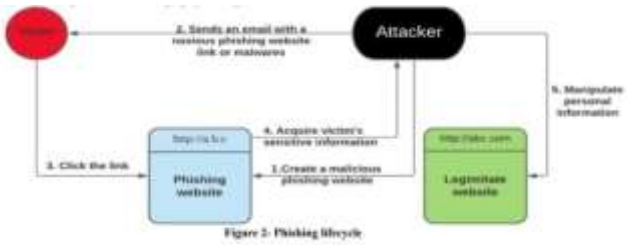


Figure 2: *Email Phishing*

For classification of the trustworthy emails and spoofed emails, we employ supervised machine learning techniques. Based on existing examples, supervised learning algorithms estimate the chances of the unknown data being spoofed, containing malware or other malicious links. Many such algorithms are a subcategory of machine learning algorithms that learn from information available in real time. The risk of losing sensitive information to fraudsters has grown due to the widespread use of emails and with the worldwide technological advancements. The comparison among several machine learning techniques for email spoofing detection and thereby detecting the IP address of the sender is our main priority and the key focus of the paper.

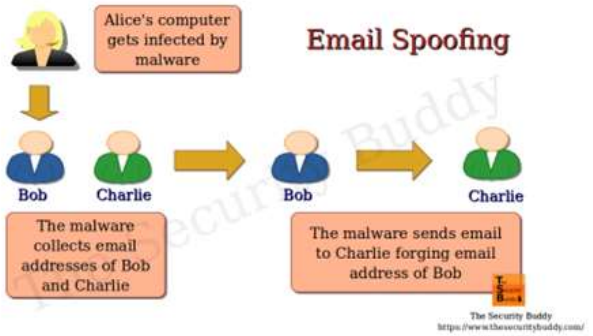


Figure 3: *Email Spoofing Detection*

## II. RELATED WORK

Numerous studies and evaluations that were published in recent years have given researchers vital information that they need to comprehend different methods of email phishing and email spoofing detection. This information includes work samples from the authors deemed most pertinent.

A random forest machine learning classifier was employed by Andronicus et al. [1] to categorise phished emails. He aimed to decrease the number of characteristics required while improving classification accuracy. It was a very effective tool for detecting content-based phishing attacks. He demonstrated a model that was built using feed forward neural networks to classify data taken from email headers and HTML bodies. The results revealed a 98.73 percent classification accuracy. In a different approach, an accuracy of 99.5 percent was reached using a dataset of more than 7000 emails and a range of variables.

To distinguish between legitimate and phishing or spoofed emails, Gilchan Park [2] looked for dependable features. In contrast to genuine emails, phishing or spoofed emails have comparable sentence structures and don't distinguish between the subjects and objects of the target verbs. Emilin Shyni [3] published a method that combines natural language processing, machine learning techniques and also image processing. They employ 61

distinct features in all. Using a multi-classifier, they were able to achieve a classification accuracy of over 96 percent. In the most recent research, a lot of scholars have concentrated their attention on the structural features of phishing or spoofed emails. In order to create a real-time email spoofing detection system, Chandrasekaran [4] built on the intrinsic structural characteristics of phishing emails and merged them with the frequency distribution of key functional phrases. They believe that the repeated distribution of the same message to many recipients, the usage of hyperlinks and identity theft are crucial elements of the design of phishing emails. They employed these structural elements and got good results. However, they only had 400 emails to work with and didn't look at the structural characteristics of the benign samples.

To detect phishing emails from genuine emails, Fette [5] also anticipated that phishing emails will reveal more structure and content information. They think that emails' HTML content is a crucial component. The number of domain names, the amount of JavaScript codes, the number of dots in the URL, and other statistics were counted. A 96% prediction accuracy was attained. They used unbalanced samples, including 860 phishing samples and 6,950 genuine samples, based on the data set they chose. Research on the characteristics of phishing and spoofed samples was constrained by the quantity of samples available. In addition, there was no discussion of the variations between the data sets. HTML content was used in a phishing website detection study.

Another essential aspect of material is structure. Aburrous [6] identified phishing pages based on their URLs, domain names, scripts and page layout. Moghimi [7] focused on the influence of page resources and resource access techniques on categorization when recognising phishing sites in online banking. Das A [8] used recursive neural networks to construct semantic emails utilising email templates, harmful content and adaptations based on earlier research. However, they also encountered grammatical errors or strange sequences that were clear to a person. Chen [9] employed a two-stage random natural language generation system to synthesise emails depending on factors including sender style, subject structure, and email structure. It is impossible to properly control how the subject of the email matches to the content because the synthesis process is arbitrary.

In order to identify the attacker using email headers, A. Hamid I.R. et al. [10] proposed a mixed-selection approach based on the fusion of both behavior-based and content-based criteria. In order to identify website phishing, Aburrous, M. et al. [11] created a fuzzy logic-based model employing fuzzy data mining techniques. These results showed the importance of the URL and Domain Identity in email spoofing detection.They came to the conclusion that machine learning-based solutions would play a significant role and act as a future road map. Additionally, Said Salloum et al. [14] discussed many contemporary approaches to spoofed email detection created using machine learning (ML) and natural language processing (NLP) techniques, as well as their shortcomings.

## III. EVALUATION

IP (Internet Protocol) is used to send packets across the internet. The IP addresses of the source and destination are included in the IP header that is present in the email. These two parameters are essential for ensuring that packets are authenticated and transmitted to the correct place. The source IP address is in charge of verifying and responding to this. Spoofing is a technique for assuming another person's identity. During spoofing, the originating IP address is changed to a different IP address. As a result, the receiver accepts these spoofed emails since they have a different IP Address that may be mistaken for a trustworthy one. For example, a 192.180.18.17 IP address attacker is attempting to send a packet to a 192.180.10.09 IP address website. The attacker is aware of a reliable IP address, let's say 192.180.18.19.Before sending the packets the attacker will change its IP Address to the trusted IP Address 192.180.18.20. When the receiver sees that the source of the packets is a trusted party, they identify it as a trusted source and accept the packets.

Few methods to defend against email spoofing attacks are:

[1] Penetration tests used to aim at higher network security.

[2] Hop Count Filter used to prevent spoofed IP Traffic.

[3] Man-In-Middle Attacks based on Address Resolution protocol.

[4] For Open Network Access Systems, stop address spoofing and unauthorised access.

*A. Penetration Tests that aim at a higher Network Security*

In penetration tests, a tester looks for all potential entry points that an attacker could use to attack the system and if any vulnerabilities are found, they will be identified. A penetration test has four stages: preparation, discovery, exploitation, and reporting. The tester chooses what needs to be tested at the planning stage. The tester gathers as much data about the network as they can during the discovery phase. The information acquired during the discovery phase is utilized to exploit the network and to identify vulnerabilities during the exploitation phase. The tester then creates an impenetrable system and a report on this. It is necessary to use this strategy continuously.

*B. Hop Count Filter Used To Prevent Spoofed IP Traffic*

Hop Count can be used to find the packets with spoofed IP Address. The hop count method is used to check how many hops the packet should travel in order to reach the destination. The hop count value is determined by examining the routing architecture and therefore, it cannot be faked. Also, this method helps us to know whether the packet is a fake or not. The IP header is unable to reveal the hop count information. Indirectly, the hop count value is kept in the Time to Live (TTL) field. This method compares the hop count value of the packets that arrive at their destination, to the stored hop count value. The packets are thrown away or discarded if the values don't match.

```
for each packet:
        extract the final TTL Tf and the IP address S;
        infer the initial TTL Ti;
        compute the hop-count Hc = Ti ;
        if (Hc -Hs)
                the packet is spoofed;
        else
                the packet is legitimate;
```

Figure 4: *Hop-Count Inspection Algorithm*

*C. Man-In-Middle Attacks and Address Resolution Protocol*

The IP address is changed to the relevant MAC address using the Address Resolution Protocol (ARP). A man-in-the-middle attack, which can be used to change the ARP cache, is always a possibility. Along with the ARP cache, it also keeps a table that lists all the hosts that are still up and running. The basic premise of this type of attack is that if one node, let's say node A, is aware of the precise IP/MAC address mapping of another node, let's say node B, node A will store this information for a long time in a file like: Node B is running; nobody is around. This will be used in the midst of a conflict. The IP address, MAC address and time value of each user will therefore be stored by each gateway. The retention time of the communication is represented by this time value. This time value is typically set to 60 minutes. ARP will request a new MAC address after 60 minutes. One of two things may occur if there is a discrepancy between the stored value and the value obtained in the ARP response:

*1) The host stored in the table doesn't exist anymore*

*2) A man-in-middle attack took place*

There are two possibilities and the ARP will distinguish between them by sending multiple (say 50), unicast requests having an average delay of 10 msec at random intervals to the host in the table. It will then maintain the information coming from the table and if the ARP replies at least once, then it will reject it, which therefore indicates that it is under attack. If during the ARP unicast request, the ARP reply was received, then it will assume that the host was not active and will then update the database accordingly.

Therefore, by studying this table, we can arrive at one of the two conclusions. We either understand that the host does not exist anymore and update the contents of the table to reflect this information, or understand that a man-in-middle attack has taken place. Now that we know that an attack has taken place, the necessary measures to protect against these emails are taken and the users will be safe from the attacks and they will not lose any sensitive information to the attackers and fraudsters.

*D. Prevent Unauthorized Access And Address Spoofing For Open Network Access Systems*

There is a suggested system for granting access control. This system consists of a frame filter, a set of LAN switches, a user authentication server and a server that assigns IP addresses. Using the ports on the LAN switches, the user PC's are linked. User PC's can only join the network using the switch's ports. The system is configured to only allow access to the server that assigns IP addresses when a PC joins the network using the switch. So, when a PC and switch are connected, the server that assigns IP address, will assign the PC with a unique IP address.

An authorization to connect to the host's authentication server is then granted to the PC. The filter copies all of the data following the completion of the authentication. The PC is then permitted to establish a connection with the host's authentication server. Before allowing the PC to connect to an external network, the filter copies all of the host's information to it when the authentication is finished. As a result, each time a packet enters the system, the filter inspects its contents before delivering it to its destination. So the filter checks the source MAC address, destination MAC address, source IP address and destination IP address. This cross-checking ensures that every potential issue is identified. If spoofing is detected, it will simply reject the packet and thus protects users against spoofed emails.

*E. Machine Learning Using Logistic Regression*

In Machine Learning, Logistic Regression is one of the best algorithms which is used in classification problems. A prediction is done by this algorithm which helps in determining the class of an instance. In email spoofing detection, a data set can be used which will be used to train the project model and this algorithm will help in classifying the messages: whether they are spam or ham messages. This data set will comprise several messages that will comprise a huge set of messages along with their category. The logistic regression model gives us a probability value between 0 and 1. This probability value can in turn determine the class of the instance. This is used to detect the spoofed emails and classify the set of received emails into spoofed and non-spoofed emails. This protects the user against the malware attacks done via spoofed emails and is the most efficient way for email spoofing detection.

## IV. CONCLUSION

In the world of cybersecurity, spoofed email protection is now a crucial problem. Early and effective detection of email spoofing is now essential due to the exponential rise in the occurrences of spoofed email over the past several years. As a result, many methods were proposed in this sector to address the problems. With advancements in internet technology and the ensuing revolution in online user engagement, security concerns have grown increasingly serious. In this study, multiple email spoofing detection methods are reviewed. These include study of Penetration test, Hop count Filtering, Man-In-Middle Attack, Detecting address spoofing and preventing unauthorized access and using Logistic Regression for email spoofing detection. It was observed from comparing these techniques that Machine Learning using Logistic Regression is the most efficient and accurate technique for email spoofing detection. Using this method, the possibility of detecting spoofed emails

increases, the accuracy of detection of spoofed emails is high and thus, users are capable of efficient email spoofing detection.

## *REFERENCES*


[1] Yongdong Wu, Zhigang Zhao, Ying Qiu, and Feng Bao, "Blocking Foxy Phishing Emails with Historical Information," IEEE ICC 2010

[2] Zhiyun Qian, Z. Morley Mao, Yinglian Xie, Fang Yu, ''Investigation of Triangular Spamming: a Stealthy and Efficient Spamming Technique,'' 2010 IEEE Symposium on Security and Privacy.

[3] D. Mooloo and T.P. Fowdur, "An SSL-Based Client-Oriented Anti-Spoofing Email Application," 2013 IEEE

[4] GHADA AL-RAWASHDEH , RABIEI MAMAT, AND NOOR HAFHIZAH BINTI ABD RAHIM, "Hybrid Water Cycle Optimization Algorithm With Simulated Annealing for Spam E-mail Detection," 2019 IEEE

[5] Markus Schneider, Haya Shulman, Adi Sidis, Ravid Sidis and Michael Waidner, "Diving into Email Bomb Attack," 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)

[6] Kenya Dan, Naoya Kitagawa, Shuji Sakuraba, Nariyoshi Yamai, "Spam Domain Detection Method Using Active DNS Data and E-mail Reception Log," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)

[7] Aparna Jayan and Dija S, "Detection of Spoofed Mails," 2015 IEEE

[8] Andronicus A. Akinyelu and Aderemi O. Adewumi, "Classification of Phishing Email Using Random Forest Machine Learning Technique," Hindawi Publishing Corporation, Journal of Applied Mathematics, Volume 2014, Article ID 425731, 6 pages: http://dx.doi.org/10.1155/2014/425731

[9] Sharon Abraham and Dr. Sabu George, "A Review of Phishing Email Detection based on Different Machine Learning Methods," International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Published by, www.ijert.org, ICCIDT - 2022 Conference Proceedings

[10] Pooja Suresh Kadam, Prof Mansi Kambli, "Analysis different algorithms used for detection of email phishing," International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 04 Issue: 07 | July -2020

[11] Kulwinder Kaur and Dr. Mukesh Kumar, "Spam Detection using KNN, Back Propagation and Recurrent Neural Network," International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, IJERTV4IS090492: www.ijert.org, Vol. 4 Issue 09, September-2015

[12] D. Mooloo and T.P. Fowdur, "An SSL-Based Client-Oriented Anti-Spoofing Email Application ," 2013 IEEE

[13] Dhruv Rathee and Suman Mann, "Detection of E-Mail Phishing Attacks - using Machine Learning and Deep Learning," International Journal of Computer Applications (0975 – 8887), Volume 183 – No. 47, January 2022

[14] Reem K. Alqurashi, Ohoud S. Al-harthi and Sabah M Alzahrani, "Detection of IP Spoofing Attack," International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2736-2741 International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.10.2020.2736-2741

[15] Cheng Jin Haining Wang Kang G. Shin, "Hop-Count Filtering: An Effective Defense Against Spoofed DDoS Traffic"

[16] Gaoqing Yu , Wenqing Fan ,Wei Huang , Jing An, "An Explainable Method of Phishing Emails Generation and Its Application in Machine Learning," 2020 IEEE 4th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC 2020)

[17] Rabab Alayham Abbas Helmi, Chua Shang Ren, Arshad Jamal and Muhammad Irsyad Abdullah, "Email Anti-Phishing Detection Application," 2019 IEEE

[18] Sikha Bagui , Robert Jamie White, Debarghya Nandi and Subhash Bagui, "Classifying Phishing Email Using Mach"

[19] Xue Li, Dongmei Zhang and Bin Wu, "Detection method of phishing email based on persuasion principle," 2020 IEEE 4th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC 2020)

[20] Dr. Reshma Banu, Mr. Anand M, Akshatha Kamath C, Ashika S H S and Ujwala Harshitha S N, "Detecting Phishing Attacks Using Natural Language Processing And Machine Learning," Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019), IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8