Anti-Spoofing in Medical Employee's Email using Machine Learning Uclassify Algorithm

Bander Nasser Almousa, Diaa Mohammed Uliyan
Department of Information and Computer Science(ICS)-College of Computer Science and Engineering, University of Ha'il (UOH), Hail, Kingdom of Saudi Arabia

***Abstract*—Since the advent of COVID-19, healthcare and IT cybersecurity have been an issue. Digital services and foreign labor have increased cyberattacks. July 2021 saw 260,642 phishing emails. 94% of 12 countries' employees experienced epidemic cyberattacks. Phishing attacks steal sensitive data from spam emails or legitimate websites for profit. Phishing spam uses URL, domain, page, and content variables. Simple machine-learning methods stop phishing emails. This study discusses phishing emails and patient data and healthcare employee accounts cybersecurity. This paper covers COVID-19 email and phishing detection. This article examines the message's URL, subject, email, and links. Uclassify classifies content, spam, and languages and automates emails. Semi-supervised machine learning dominates healthcare. The Uclassify algorithm used multinomial Naive Bayesian classifiers. Document class is [0–1]. This article compared Multinomial Naive Bayesian in two experiments with other algorithms. Experiment 1 achieved an MNB accuracy of 96% based on a database from Kaggle Phishing. Experiment 2 showed that the Multinomial Naive Bayesian system accurately predicted URL and hyperlink targets based on PhishTank data. 96.67% of respondents correctly identified URLs, and 91.6% did so for hyperlinks. These two experiments focused on Tokenization, Lemmatization, and Feature Extraction (FE) and contained an internal feature set (IFS) and an external feature set (EFS). MNB is more exact than earlier methods since it uses decimal digits and word frequency. MNB only takes binary inputs. MNB can detect phishing and spoofing.**

***Keywords*—Spoofing; phishing; machine learning; Uclassify algorithm; medical employee's email**

I. Introduction

The Internet's quick services affect the world. Consumers can shop and bank anytime with improved Internet infrastructure. Despite its benefits, internet security and privacy are issues. Phishing, malware distribution, and privacy exposure are all possible with the Internet's anonymity[1]. Emergency and patient care services require healthcare workers to exchange electronic health information with patients. This data is one of the most sensitive. Therefore, special measures must be taken regarding threats to confidentiality, integrity, and availability (CIA)[2]. Healthcare should prepare for cyberattacks. Attackers have used email to target healthcare companies since the COVID-19 outbreak began [3]. The pandemic increased health cybercrime. Digital change encourages "telework" or "work from home," which boosts email efficiency [4]. Remote operators must be trained for safety. More than 3,000 employees in 12 countries have experienced remote working, and 94% have experienced cyberattacks during the pandemic, according to a report by the International Association of IT Asset Managers (IAITAM) [5]. The Anti-Phishing Working Group (APWG) says there were 260,642 phishing attacks in July 2021, the most ever seen in one month. It rises from 44,008 in the first quarter of 2020 to 128,926 in the third quarter [6]. Phishing emails steal any account credentials. Banks and hospitals were attacked the most [7]. Spoofing deceives victims into divulging passwords, usernames, and personal information. Scammers entice recipients to visit fake websites or download viruses [1]. 16% of 2015 FTC complaints were about identity theft. Phishing and social engineering steal data. Phishing online fraud using a fake website and email [8].

URIs hurt or redirect phishing victims. Second, domain length, numbers, spelling, and brand name may impact Phishing—status, domain owner, and age impact URLs. Reputation determines page believability. Content-based domain scanning—hidden, body, meta, and pictures—estimates daily, weekly, or monthly page views, average page visits, internet traffic, domain category, and similar websites. Finally, scanners check page type, user, and registration [9].

A. Problem of Phishing Attack Detection

Study email security. Email security avoids loss, theft, and hacking. Spam, phishing, and malware utilize email to transfer sensitive data and access networks [10] Phishing persists despite email security. Phishing is inevitable. This thesis tests copyright checking without email sender server settings [11]. The goal: Email is often attacked. Inconsistent protocols, roles, and services compromise email security [12]. Spoofing websites steal customer data—most often fake websites. Machine learning detects spam and phishing sites [13].

These research questions are addressed in this paper as follows:

1) Which machine learning methods are used to create phishing detection models?

2) What datasets are used for phishing email detection models?

3) How can modern datasets and resources recognize phishing emails?

4) Which email features use Phishing detection emails?

5) What are the early stages of phishing and the current trend of phishing emails?

To answer these questions or to address these issues, the findings in the research could be:

1) SVM, LR, and DT recognize fake emails, whereas k-means clustering does not [14]. Multinomial Naive Bayesian classifiers power Uclassify ML. Its feature vector classifies input into the most likely class. This algorithm considers all input data irrelevant. Data sets are unaffected by changes.

2) ML-based model training and testing require a dataset. Phishing detection methods were developed utilizing pooled datasets. Producers often update long-lived datasets. 2021 revised Nazario's dataset. 2010 spam email, 2005 phishing corpus, 2006 Enron spam, and 2002 spam assassin datasets are updated routinely [15].

3) Classify creates ten category numerical values after phishing using ML and web services—health. Uclassify web API will categorize questions by feature vector subject for this study. Uclassify contains ML classifiers for sentiment, themes, language, age, gender, and more.[16]. Customer emails initiate phishing. This malicious email links to an attacker-based website. Reassures email recipients:

a) The sender and email address are not from "UMass Amherst it@umass.edu>," despite the assertion.

b) Phishing emails are misspelled. This email's colon should not precede the comma.

c) Phishing emails employ urgency. It spurs action.

d) The message URL is UMass Amherst's webpage. Hovering reveals a different page.

e) UMass Amherst and Microsoft Corporation are impersonated in the letter. Again, bogus if the sender needs to know who they are.

f) The email link leads to a bogus SPIRE-phishing login page at "tantechhold-ings.com." [6].

4) Extracting features from raw data entails retaining the original data set while converting it into numerical portions that can be processed. BoW, IG, and Word2vec are text characteristics in phishing email detection research. The research also utilized latent Dirichlet allocation, part-of-speech tagging, PCA, and LDA [6].

5) Phishing targets psychology and emotion. They employ social engineering and technology. Attackers' personalization, clever phishing, and tactics increase prevention [17].

II. Literature Review

A. Introduction

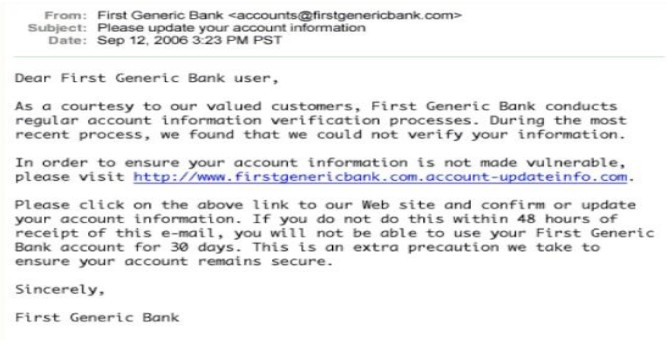
Digital healthcare providers fear cyberattacks. Social engineering targets individuals. This lengthy research shows how cyber threat ignorance impacts healthcare—technical and organizational healthcare cyber security. Table I shows the authors' five questions to classify literature by subject. Healthcare cyber defense was examined first. All personnels are studied [18].

Table I outlines this systematic review's research questions and background [18].

TABLE I. Systematic Review's Research Questions and Background

Background	Research Questions
Social engineering cyberattacks against healthcare workers are the most effective. These assaults use social media data.	RQ1: What are the prevalent social attacks perpetrated against individuals by healthcare organizations?
Data governance, encompassing data security, privacy, and IT infrastructure security, protects organizations against cyberattacks. Several methods prevent DDoS assaults. WannaCry has increased healthcare data leaks.	RQ2: Which policies and governance mechanisms have strengthened healthcare organizations?
Cyber risk assessment improves healthcare. Data breaches are expected as healthcare becomes more complicated. IT security dominates risk assessments. Social engineering requires reassessing healthcare cyber threats from insecure human behavior.	RQ3: How does an organization's cybersecurity risk assessment incorporate human elements in cybersecurity?
Healthcare companies now train on cybersecurity—for example, phishing email training. Given recent examples of using social media data to target healthcare practitioners, raising awareness of this evolving, dangerous environment is vital.	RQ4: How can training raise healthcare workers' cyber threat awareness, and how can we quantify an organization's training and awareness efforts?
Europe needs healthcare infrastructure. Outages may generate national emergencies. ENISA advises on cyber resilience. Multi-nation cyberattacks have had enormous economic and human repercussions.	RQ5: Which national and international organizations offer cyber defense strategies to boost cyber resilience?

Fig. 1 illustrates a phishing email, whereas Fig. 2 describes its essential elements. PhishTank hosts samples and notes.


From: First Generic Bank <accounts@firstgenericbank.com>
Subject: Please update your account information
Date: Sep 12, 2006 3:23 PM PST

Dear First Generic Bank user,

As a courtesy to our valued customers, First Generic Bank conducts regular account information verification processes. During the most recent process, we found that we could not verify your information.

In order to ensure your account information is not made vulnerable, please visit http://www.firstgenericbank.com.account-updateinfo.com.

Please click on the above link to our Web site and confirm or update your account information. If you do not do this within 48 hours of receipt of this e-mail, you will not be able to use your First Generic Bank account for 30 days. This is an extra precaution we take to ensure your account remains secure.

Sincerely,

First Generic Bank

Fig. 1. Phishing email example[8].

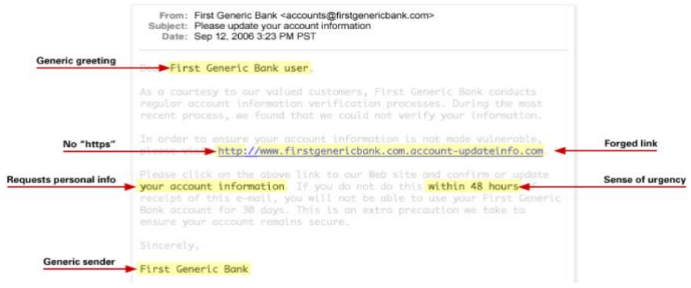


Fig. 2. An example of a phishing email with annotations highlighting its key components [8].

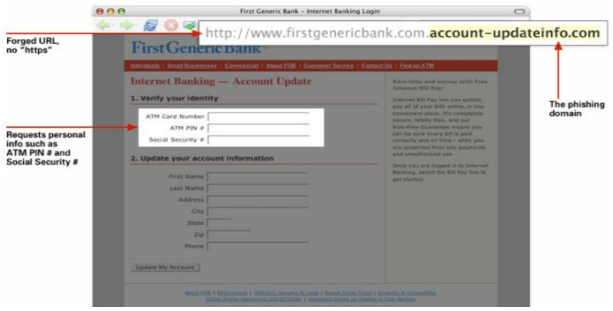


Fig. 3. Phishing website [8].

The phisher's email connects to a bogus website (Fig. 3). This fake website was created to deceive email recipients. The attacker may utilize Fig. 3's remote data input area. Phishers then emailed and messaged AOL subscribers. Phishers asked AOL users for their account numbers. AOL's TOS couldn't track attackers' AIM accounts, worsening the issue. Finally, AOL sent emails and instant messages pushing users to withhold vital information [8].

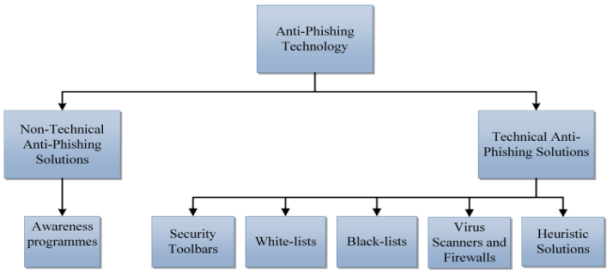


Fig. 4. Phishing prevention technology [8].

Blacklisting and whitelisting virus scanners safeguard internet users (see Fig. 4). Commercial software combats zero-day phishing. Microsoft, Google, and PhishTank blacklists let researchers test solutions. Anti-phishing may block and whitelist phishing URLs on the client's PC or server. Users trust whitelists. Zero-day phishing detection takes accuracy. Whitelists may mistake good websites for phishing. AIWL records user-irritating visual effects and manual white-list maintenance. Multinomial Naive Bayesian classifiers monitor AIWL logins. AIWL whitelists login URLs. The hardest part is picking a trustworthy site from a fraud list. Finally, blacklists and heuristics may whitelist. Avoiding trustworthy websites may expedite phishing detection. URL blacklists stop Phishing. Anti-phishing blacklists URL-verified phishing. Blacklists' low false-positive rate and simplicity make them famous. The research discovered that blacklist software missed 80% of zero-day phishing assaults. Blacklist users enjoy its simplicity and low false positive rate. Blacklists' low false-positive rate and simplicity make them popular [8].

B. Definition of the Spoof

The official definition is masquerade: "A type of threat action whereby an unauthorized entity gains access to a system or performs a malicious act by illegitimately posing as an authorized entity" and Spoof: "Attempt by an unauthorized entity to gain access to a system by posing as an authorized user" and phishing attack is: "A technique for attempting to acquire sensitive data, such as bank account numbers, through a fraudulent solicitation in email or on a website, in which the perpetrator masquerades as a legitimate business or reputable person" [19].

C. Related Works

Technology introduces security vulnerabilities. In the 1960s, "phone hacking" developed access control and encryption. AOL hackers coined "phishing" in 1996 after obtaining private data. Phishing emails verified AOL customers' credentials. Many provided hackers their login details to purchase items. Customers pay millions. eBay, HSBC, and others fight phishing. These techniques identify fake emails and websites. Despite its effectiveness, only some read online phishing prevention literature. ML finds fraud. Dots, domain ages, and links checked URLs for Spoofing or Phishing. Consumer education prevents phishing emails [20]. For five years, they analyzed US public sector data concerns. Only US public sector international concerns are studied. Targeted personnel cause 22% of security breaches, and hackers 45%. Social dishonesty increases. The Email Sender Centre prevents phishing and impersonation as threats rise. Scammers may be caught. Scam emails conceal [21]. COVID-19 phishing was examined. Cyberattacks grow. Backup data, secure remote worker networks, communicate with IT, and educate staff in the attack. COVID-19 and $6 trillion hacking by 2021 ended it. This article covers multifactor login, VPNs, new hacking regulations, and IT-employee communication. Pandemic hackers targeted hospitals. Phishing, ransomware, homework, and government attacks boost hacking threats. Education, VPNs, multifactor authentication, firmware upgrades, and a firm safety policy decrease these hazards. Spoofing needs numerous countermeasures: Fake COVID-19 affects ML [5]. Phishing and impersonation use distinct strategies. Fig. 5 shows Internet and phone; email, IM, and social media work well. Theft motivates these assaults. It accurately detects phishing emails, URLs, IPs, and images [5].

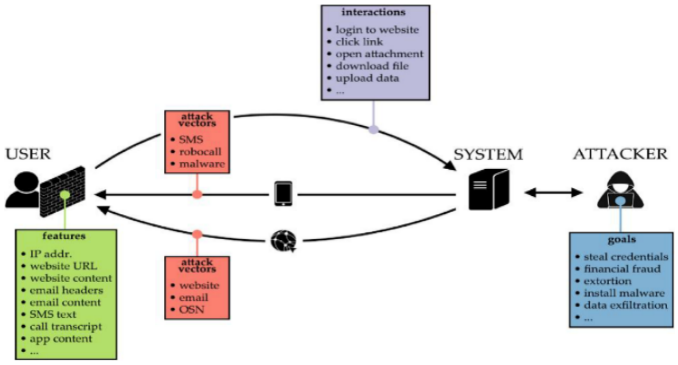


Fig. 5. Phishing attack [5].

SMTP is the Internet protocol for transmitting email. Fig. 6 depicts the three essential email message submission processes [22].

1) The sender's MUA sends the message to the service provider's MSA through STMP or HTTP/HTTPS (MSA).

2) The recipient's email provider receives the message via SMTP from the sender's MTA.

3) The user receives the message over HTTP/HTTPS, POP3, or IMAP via the Mail Delivery Agent (MDA) (IMAP) [22].

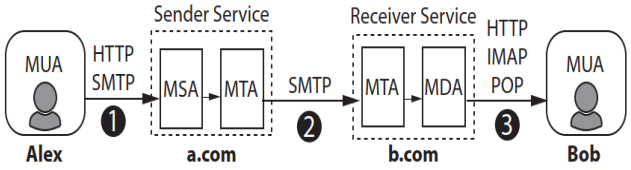


Fig. 6. Alex and bob's email transmission [22].

III. Research Methodology

4.5 billion people use the Internet. Email is trusted online. Fraudulent emails include malware or unsafe URLs. Even with better filtering, spam emails' constantly shifting content makes them hard to distinguish. Corporate email, commercial antispam services, and end-user training filter spam, yet this deadly trap caught non-experts. It trains SVM, Multinomial Naive Bayesian, CNN, and LSTM spam email detectors. This article offers different ML models sans spam email datasets. CNN and LSTM utilize Model Loss and ROC-AUC; MNB and SVM need precision, recall, and f-measure. Finally, all models are compared for great accuracy and DL and ML model evaluation parameters. English and spam detection enhanced [23]. Phishing detection has been improved. ML algorithms shine. ML algorithms will be studied with an accuracy of 94.4%. Content-based filters identify fake emails [24]. ML-based spam classification. Uclassify's own naive polynomial Bayesian classifier. Classification limits are calculated using the document category probability 0-1 [25]. The proposed solution considers the detection aspects of phishing attacks, URLs, and domains. ML will fight phishing attacks Fig. 7 depicts the Phases of a typical Spear-phishing attack [26].

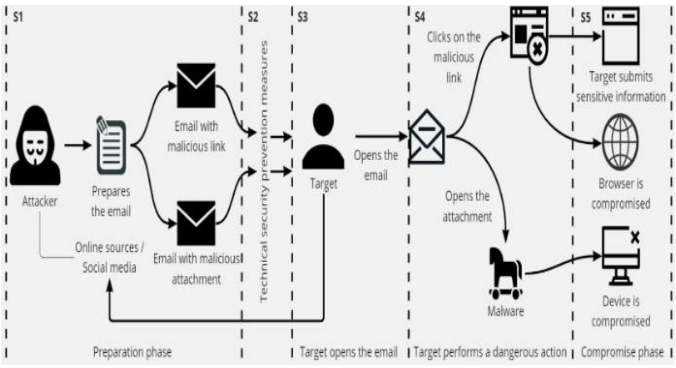


Fig. 7. Phases of a typical Spear-phishing attack.

An attacker gathers as much sensitive target information as possible during preparation (S1). Leak social media and databases. Knowing the victim eases persuasion. Spam-detecting email infrastructure demands technological skills—clean, succinct, balanced email content (S2). The attacker then wants S3 opened. Email subjects matter. Secure email. The reader gets balanced (S2). The attacker then wants S3 unlocked. Email subjects matter. Secure email. Email motivates. Certain emails may create confidence. Avoid downloading attachments (S4). Preventing assaults requires security measures. Antivirus and OS updates protect. S5 attacks browsers, devices, and credentials [26]. Table II outlines our suggestions. "Use of @ symbol," "right-click blocked," and "hiding suspicious links" are three of the seventeen traits we are eliminating. RFC engineers and computer scientists accept @, ", and # in URLs. Browsers no longer allow status bar changes to turn off right-clicking and hide suspicious links. Invalid characteristics. 54-character URLs are questionable. Most bogus URLs are under 54 characters [32]. It impairs judgment. We verified the URL length. Chrome and IE allow 2083-character URLs. Thus, a URL with 1000–1750 characters is suspicious, but phishing with more.

TABLE II. Features of the Website and its Description [27]

#	Feature	Description
1	**IP domain names**	Using IP address in domain part is phishy because attacker is trying to disguise name with numbers.
2	**Long URL**	Long URLs disguise questionable keywords. URLs over 1750 are phishy.
3	**'-' symbol**	The domain's "-" indicates legitimacy. Example: Pay-Pal.com
4	**Sub-domain(s) in URL**	Phishing URLs have several subdomains.
5	**Use of HTTPS**	HTTPS secures URLs.
6	**Request URL**	The URL domain should load all text and graphics.
7	**URL of anchor**	All <a> tags should have domain-matching links.
8	**Server form handler"SFH."**	User data may be transferred. Checking Server Form Handler prevents this.
9	**Abnormal URL**	The URL's WHOIS data confirms its identity—no phishing website.
10	**Redirect page**	Links sometimes redirect users to other pages. Phishy redirects exceed four.
11	**Using pop-up Window**	Pop-up password entry is unethical. Pop-up sites are phishing.
12	**DNS record**	Phishy URLs lack DNS records.
13	**Website Traffic**	Website traffic visits. Websites without traffic records are fishy.
14	**Age of domain**	Domains age from registration. Phishing site registered under one year.

D. Research Design

Businesses and people email. Spammers make money. Bio-inspired machine learning identifies fake emails in this study. A study analyses how to use varied datasets for successful results. Genetic and Particle Swarm Optimization improved classifier performance. Features or automatic parameter selection may assess spam categorization algorithms. Comparing recommended and basic models will determine whether parameter changes enhance them [28].

E. Dataset

The detection approach is crucial to the proposed system, but the datasets used by the authors to test and train their algorithms affect their credibility. Website detection datasets match email detection datasets, showing no concerns. Spam and viruses sometimes are utilized. However, the papers discuss fraudulent email detection. These publications are harmful (the spam dataset contains spam email URLs). Online datasets for fraud detection algorithms are available—Table III lists prevalent phishing and ham datasets.

TABLE III. PRESENTS THE FEATURES OF THE DATASET [6]

#	Dataset feature	Description
1	**Dataset source**	The generally utilized data sources of legitimate and phishing websites, along with the approaches that grip every head, are mentioned; however, the insufficient understanding regarding the methodologies used in collecting and preserving every source results in no concord with all regarding the quality of various origins.
2	**Dataset size**	The evaluation dataset size differs between various approaches. As seen, the reliable outcome depends on the size of the dataset; the bigger the better
3	**Dataset redundancy**	There is not sufficient information in the literature regarding dataset redundancy. Although numerous presentations and overly between various sources of datasets, particularly of phishing websites, can be seen
4	**Dataset timeliness**	Although if a similar source of data and size of the dataset is utilized in two plans, their phishing website's information might not be the same. The phishing blacklist supplier generally amends their data hourly, because phishing websites last for short terms.
5	**Ratio of legitimate to phishing websites**	The ratio of legitimate to phishing examples displays the level at which experiments portray an actual world distribution (≈100/1)
6	**Training set to testing set ratio**	The extensibility of the approach is seen in the ratio of training to testing examples.

This study uses the 2020 Akashsurya156-revised Kaggle Datasets Phishing Email Collection. 22 traits, 21 predictors, and one target variable define email authenticity. Each feature improves model learning. The most deceptive electronic communication phrases were found—525,754 emails in CSV. Spam and fake emails are rare, skewing the classification dataset. Campaigners tested fake and real emails. These two groups balance actual and fraudulent emails in training models [29]. Fig. 8 shows information used for spoofing URL recognition [30].

Fig. 8. Information used for spoofing URL recognition.

Phishing and URL data are needed for ML model training. A phishing crawler's program that gathers phishing URLs from the PhishTank website only records the web address if the site is active. Phish Search finds phony URLs that change. BeautifulSoup pulls page code. Using "identifiers" and queries, we may identify the request's absolute phishing URL. 10,000 phishing and 10,000 non-phishing URLs were crawled from the dataset. The training uses 8,000 URLs and evaluation 2,000. Table IV summarizes the study data. Fig. 8 depicts data distribution [30].

TABLE IV. DATA ON PHISHING AND LEGITIMATE URLS [30]

URL	Actual Data	Used Data	Train	Test
Benign URL	17058	10000	8000	2000
Phishing URL	19653	10000	8000	2000

F. Detection Method of Phishing Attack Through Email

Semi-supervised ML identifies bogus emails and impersonation. Fig. 12 depicts categorization. Semi-supervised learning employs small labeled and unlabeled examples. Binary classification identifies phishing emails. The automatic categorization is increasing ML-detected phishing. ML vectors describe emails and web pages. Fake emails and sites are 1. Label 0 denotes a valid email or page. Semi-supervised teaching detects and evaluates phishing emails and web pages. Fig. 11 shows the planned steps. Detect spoofing, phishing, and impersonation emails. These safeguard data and prevent fraud—these secure critical data. Establish email components.

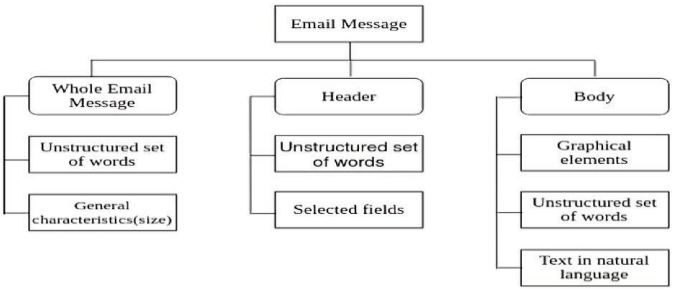


Fig. 9. Taxonomy for email messages [31].

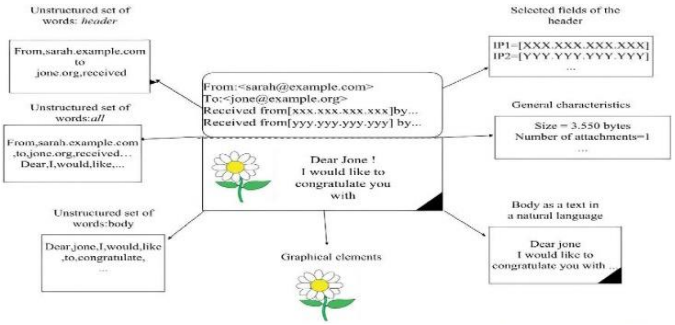


Fig. 10. The format of an email message [31].

From, To, Subject, and Message-ID—the message's content—make up the email header. Fig. 9 and 10 exhibit email taxonomy and arrangement. Emails to scam websites. Email URLs trick consumers into providing critical information. Many studies have revealed email spoofing characteristics and methods to identify fake emails. Primary, latent topic, and dynamic Markov chain features. Researchers found anti-phishing techniques. Machine learning, deep learning, word embedding, and NLP are discussed [31].

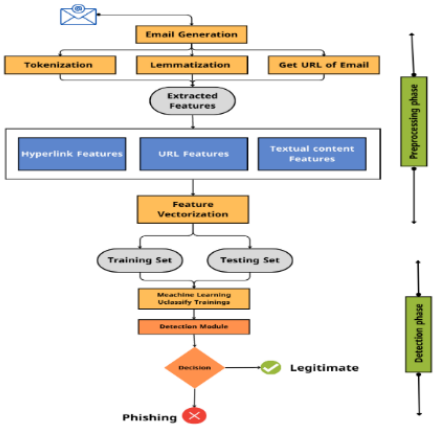


Fig. 11. The general structure of the proposed methodology.

1) Tokenization: Computationally splitting a text into words or tokens [32]. Thus, it trims the input text word by word. Next, filtering removes unnecessary words from token results and finds the root term of each filtered word. Tagging then finds the root form of a preceding word or stem word. Finally, analyzing texts finds word relationships [33].

2) Lemmatization: Removing inflectional suffixes from words returns the base form. Thus, it turns words into their roots. This removes the inflectional ending and restores the root. It uses lexical and morphological analysis to find a word's root form, unlike stemming. The word "better" means "excellent." [32, 34].

3) Feature Extraction (FE): Lexical components, conceptual frameworks, and linguistic phrases make text categorization difficult. Thus, learning from a large dataset takes a lot of work. Therefore, it's computationally expensive. Extraneous and redundant features also impair categorization algorithm accuracy and effectiveness. Use just the key elements to simplify and accurately classify data with little duplication [35].

Our classification has two main goals: 1. Extracting multi-view characteristics from each email that classifiers can manage; 2—applying a disagreement-based semi-supervised learning method to automatically identify and use unlabeled data. Fig. 12 displays our high-level email categorization model. Plan, train and classify. Initialization sorts incoming emails by our standard so an ML classifier can handle them (each email has unique attributes). Research and public spam datasets examined subject length, message size, attachment size, and word count. These studies describe email. 14 criteria were used to categorize emails using the variables above. Email techniques like route tracing and content recording may gather and compute the information. Creating two attribute sets—an internal feature set (IFS) and an external feature set (EFS)—from standard features. IFS contains email content properties, whereas EFS includes routing and forwarding [36]. Table V shows the email categorization model's multi-view data and disagreement-based semi-supervised learning architecture:

- Internal Feature Set (IFS): Email body features include subject length, message size, number and kind of attachments, wings, words in subject and message, and several embedded photos [36].
- External Feature Set (EFS): Unlike IFS, EFS affects email routing, forwarding, confirmations, replies, importance, frequency of sending and receiving emails, and sender name length [36].

TABLE V. TWO FEATURE SETS FOR EMAIL CATEGORIZATION MODELS[36]

	Internal Feature Set (IFS)	External Feature Set (EFS)
1	subject length	the number of receipts
2	message size	the number of replies
3	total attachments	the level of importance
4	type of attachments	email frequency
5	size of attachments	frequency of email receipt
6	words in the subject	sender name length
7	message word count	
8	embedded images	

Uclassify ML detects phishing emails: Fig. 12 shows semi-supervised ML. Classical ML improves predictions. The main methods include unsupervised, semi-supervised, and reinforcement learning. Scientists forecast data through supervised, semi-supervised, or reinforcement learning. Semi-supervised ML partly marked inputs. ML handles real-world learning issues. Semi-supervised ML employs unlabeled primary data and little human input. Labeled datasets are harder to obtain, costly, and may need topic expertise. Thus, fewer are preferable. Unlabeled datasets are cheaper and more straightforward.

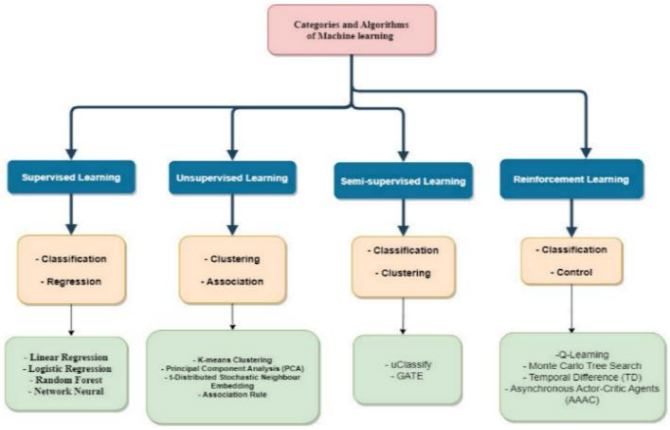


Fig. 12. ML phishing detection methods.

Semi-supervised ML trains unsupervised algorithms. Unsupervised ML uncovers input dataset structures. Supervised ML approaches use the best-unlabeled data predictions on new data. Unlabeled data re-rank or re-evaluate labeled data. Semi-supervised ML assumes smoothness, cluster, or manifold for unlabeled training data. Hybrid learning architectures integrate "discriminative" and

"generative" learning. HDNs integrate architectures. Action banks increase human activity recognition. Semi-supervised ML rules healthcare. It processes audio and data. Image and voice analysis improves [37].

1) Uclassify detection method: We chose the Uclassify algorithm, a free web service for ML that facilitates creating and using text classifiers. Local Classification Server runs classification engine. Technically, our classification engine runs on the Amazon cloud Specification [38]:

- Windows service-based.
- Sockets to XML.
- Multiple uses
- Accurate but forgetful
- Fast parallel request processing
- Transactional integrity
- 0-1 outcomes

2) Architecture: The classification server operates as a Microsoft Windows service, functioning through XML calls transmitted via sockets, facilitating seamless integration with diverse Operating Systems. The system also provides XML-formatted responses. The application programming interface (API) exhibits high similarity to a freely accessible web API. Thread-safe classifiers (Readers/Writers lock) allow dynamic input class and training data. Low-resource C++ servers manage massive data sets. Important classifiers categorize requests quickly. We batch-processed 2.4 KB of blog entries using five primary classifiers on a contemporary PC for speed. 3,600,000 messages/hour! One core. Many threads handle queries. It is unbreakable. No paging file, repair errors. Transactional conduct avoids classifier writing errors. We need to remember classwork. Multinomial Hybrid complementary MNB, class normalisation, and exceptional smoothing improve Multinomial Naïve Bayesian classifiers. Sort [0–1]. Set limitations—90%+ spam. CPU classifies. Removes text and class-untrained classifier spam. Retraining the spam message on the spam class may solve it. Spam groups. Fig. 13 shows the pseudocode of Uclassify algorithm.

```
<?xml version="1.0" encoding="UTF-8" ?>
<uclassify
xmlns="http://api.uclassify.com/1/ResponseSchema"
version="1.01">
  <status success="true" statusCode="2000"/>
  <readCalls>
  <classify id="call_1">
    <classification textCoverage="1">
      <class className="negative" p="0.628401"/>
      <class className="positive" p="0.371599"/>
    </classification>
  </classify>
  </readCalls>
</uclassify>
```

Fig. 13. Pseudo code of Uclassify.

a) Multinomial naïve bayes: The Naïve Bayes model addresses classification predicaments by applying probability methodologies. Equation-1 represents the Naïve Bayes algorithm in this article [28].

$$P(\text{ Class } \mid \text{ WORD }) = \frac{(P(\text{ WORD} \mid \text{Class }) \times P(\text{ Class }))}{(P(\text{ WORD }))} \quad (1)$$

The present study involves the identification of WORD, which is a set of (word $_1$, word $_2$, .. word $_n$) extracted from an uploaded email. The variable' Class' indicates the email's classification into 'Spam' or 'Ham.' The algorithm computes the likelihood of a given class based on the bag of words supplied by the program. The expression P(Class | WORD) represents the posterior probability, while P(WORD | Class) denotes the likelihood, and P(Class) signifies the prior probability, as stated in reference. Assuming that the variable 'Class' represents the category of 'Spam,' one could rephrase the equation to identify spam emails based on the provided text. It can be subsequently expressed as a more concise equation, denoted as equation (2) [28].

$$P(\text{ Class } \mid \text{ WORD }) = \frac{\Gamma_{i=1}^{n} P(\text{ word\_i i} \mid \text{ Spam }) \times P(\text{ Spam })}{P(\text{ word\_1, word\_2 ,... word\_n })} \quad (2)$$

Multinomial, Gaussian, and Bernoulli are the three varieties of Nave Bayes algorithms. The Multinomial Naive Bayesian algorithm has been chosen to identify spam emails because it is text-related and outperforms the Gaussian and Bernoulli distributions [28]. Multinomial Nave Bayes (MNB) is a classifier that employs Multinomial Distribution for each feature, focusing on term frequency. Equation (3) represents the Multinomial Naive Bayesian model.

$$P(p \mid n) \propto P(p) \prod_{1 \le k \le nd} P(t_k \mid p) \quad (3)$$

The representation of the number of tokens as "nd" and the calculation of $P(t_k \mid p)$ where "n" denotes the number of emails, are the subject matters under consideration.

$$P(t_k \mid p) = \frac{(\text{count}(t_k \mid p)+1)}{(\text{count}(t_p)+\mid V \mid)} \quad (4)$$

Equations (3) and (4) denote the conditional probability for MNB as $P(t_k \mid p)$. The variable $P(p)$ stands for the prior probability, while t_k stands for the presence of spam terms in an email. The algorithm designates 1 and |V| as the smoothing constants. The Scikit-learn library was utilized to load the MNB module to test the algorithm. The parameters of this model are discretionary. Without any specified values, the default settings for the Alpha parameter are '1.0', for the Fit Prior parameter are 'True,' and for the Class Prior parameter are 'None.'

The pseudocode for Multinomial Naïve Bayes is introduced as a spam classifier as shown in Fig. 14. We have two datasets: Tr is defined as Training dataset and Te is defined as Testing dataset. The $\hat{P}(t_k \mid p)$ is a predicting variable also identified as the conditional probability.

Initialise Input Variables;
N ← No. of samples;
X ← Datapoints;
y ← Target Inputs;

For $i = 0; i <$ Tr $X; i++$ do
If (i, y) = Spam then
Learn i = Spam;
Else
Learn i = Ham;
For t in testSize // Test sizes = 20, 25, 30 and 40
Do
For K in CV do
X_test and y_test = testing size;
X_train and y_train = training size;
For i = 0; i < TeX; i ++ do
Calculate $\hat{P}(t_k \mid p)$
Calculate the Accuracy;
Return t_k ;

Fig. 14. Algorithm-multinomial Naïve Bayes.

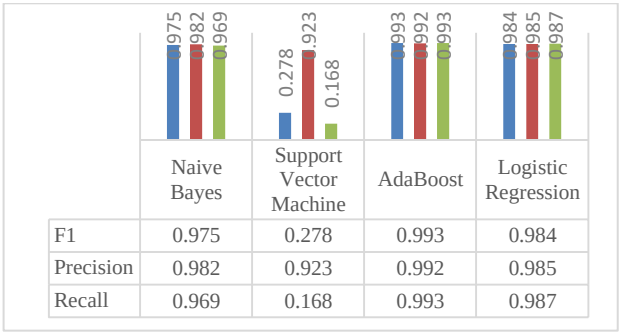
G. Experimental Results

The findings of the experiments, Multinomial Naive Bayesian (MNB), is the algorithm that outperformed all others. The various types of datasets, the Enron, Spam Assassin, and Ling-Spam datasets, provided greater depth by removing certain features from the emails, thereby allowing the optimization techniques more search space. However, the numerical datasets could have been more extensive, despite effectively enhancing the precision of some split sets. Considering the distinct datasets, the Spam Assassin dataset performed exceptionally well with MNB Nave Bayes Algorithm. The MNB performed better after being automatically tailored by bioinspired algorithms, and because it utilizes feature vectors, it performs exceptionally well with text-based datasets. [28]. Using our variables, they created a dataset to test our email categorization model (Table V). 7133 emails from two well-known companies were randomly marked and unlabeled. Three institution security professionals labeled the confidential dataset, leaving 2,300 instances unlabeled. Discord SSL employs MNB, IBK, and J48 with a 0.75 "majority vote" threshold. They tested voting strategy categorization accuracy after 60 and 100 rounds. They were comparing votes. "Majority vote" overcomes "best opinions." They reached our single-view EM semi-supervised learning method to determine how multiple-view data influences email categorization. Table V emphasizes that the EM semi-supervised learning method should train using a unique display dataset with all attributes. Our technique with multiple displays improves classification accuracy over single displays. Our practice also improves classification accuracy after numerous training rounds. Our approach changes after 60 repetitions [36]. Researchers encounter phishing websites. Blacklisting reduces hazards. Non-blacklisted websites cannot detect fraud—ML outcomes. Improve traffic, search engines, and third parties. In this work, machine learning using webpage URLs and email feature extraction—client-side, no-external services—identifies fraudulent websites rapidly and accurately. URLs link to the content [39].

H. Research Findings

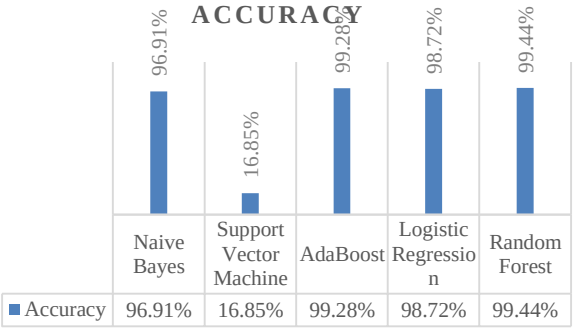
1) Experiment 1: Categorization algorithm implementation outcomes. It compares feature analysis algorithms to 525,754 Kaggle phishing emails—2020 Kaggle Datasets Phishing Email Collection updated by Akashsurya156. The collection includes fake and real emails. Fig. 15 shows that 90% of phishing emails were categorized correctly. All Random Forest metrics are 99.4%. Like Random Forest, AdaBoost offers 99% accuracy. Multinomial Naive Bayes logistic regression achieves F1 scores and recalls accuracy > 98%. SVM accuracy was 92%. Despite its poor recall and F1 score, it must outperform other classifiers [29].

According to the research, Multinomial Naive Bayes, AdaBoost, Logistic Regression, and Random Forest correctly recognized Phishing and authentic emails. MNB has 96.91% accuracy. In Fig. 16, the Random Forest classifier had the best accuracy, whereas the other classifiers did not vary. The Support Vector Machine algorithm's lowest accuracy rate was 16.85%, making it unsuitable for unbalanced dataset categorization [29].



	Naive Bayes	Support Vector Machine	AdaBoost	Logistic Regression
F1	0.975	0.278	0.993	0.984
Precision	0.982	0.923	0.992	0.985
Recall	0.969	0.168	0.993	0.987

Fig. 15. Compares models based on their F1 score, precision, and recall [29].



	Naive Bayes	Support Vector Machine	AdaBoost	Logistic Regression	Random Forest
Accuracy	96.91%	16.85%	99.28%	98.72%	99.44%

Fig. 16. Accuracy of state-of-art methods compared with Naïve Bayes [29].

According to the data below in Fig. 17, Random Forest is the most widely used classification model for identifying phishing emails. Compared to the other classifiers used in the study, it shows the highest true positive rates and the lowest false positive rates, thus establishing its status as the most reliable classifier [29].

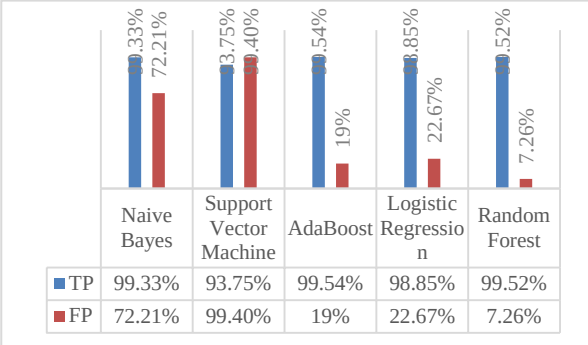


Fig. 17. Compares models based on their accuracy [29].

2) Experiment 2: ML-based bogus website detection. PhishTank analyses were actual and fake websites. First, the URL method features train and assess the ML model and compare its predicted output to the actual task. Precision, recall, accuracy, and F1-score support the conclusions. Fig. 18 presents metrics for the three models. SVM has 98.05 percent accuracy and KNN 95.67 percent. SVM had 98.24% specificity and KNN 94.40%. SVM recall 97.86%, KNN 96.99%. SVM precision values are 98.25%—F1 score. The KNN, NB, and SVM models provide 95.65, 96.66, and 98.05 percent F1 scores, respectively [30].

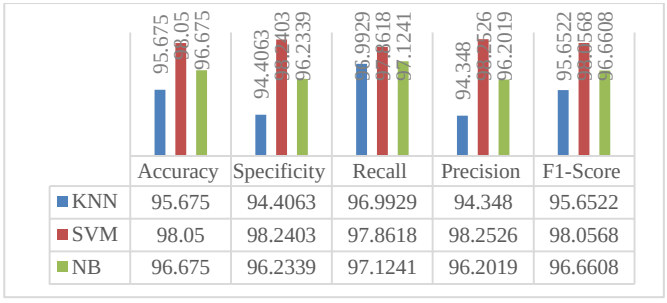


Fig. 18. Efficiency metrics of the ML model utilizing URL-based FE data [30].

ML model testing uses the Hyperlink method—ML model predictions vs. outcomes. Fig. 19 compares the three models' metrics. SVM has 94.55% accuracy and KNN 89.87%. KNN 90.7% and SVM 93.76% specificity. SVM recalls 95.36, KNN 88.87%. SVM's 93.64% accuracy is good—final F1-score analysis. KNN and NB have 89.72% and 91.41% F1 scores, respectively, while SVM has 94.49%.

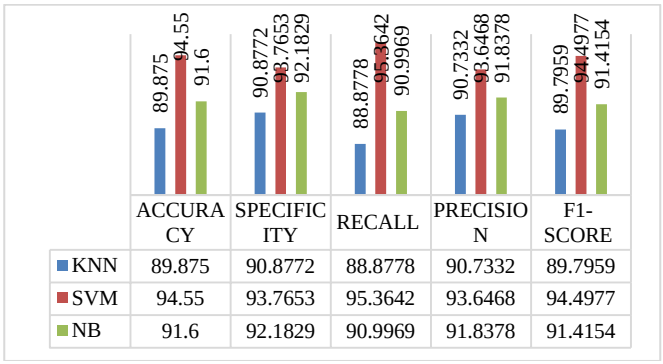


Fig. 19. Compares models based on their true and false positive rates[30].

I. Findings Analysis

SVMs classify email spam best. Fig. 20 compares the Multinomial Naive Bayes Classifier and SVM with different-sized training emails. (Fig. 20). Most training data categorizes spam. 95.5% of emails are SVM. MNB ignores email word order. Support Vector Machines find the best hyperplane to discriminate classes from predictions, unlike MNB Classifiers [40].

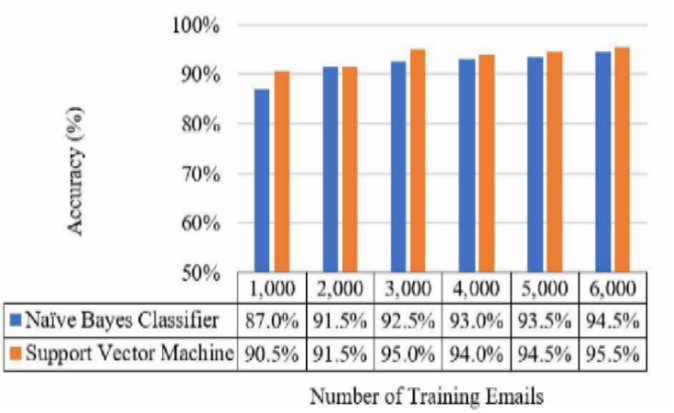


Fig. 20. A comparison of the accuracy of the Multinomial Naive Bayes classifier and the Support Vector Machine [40].

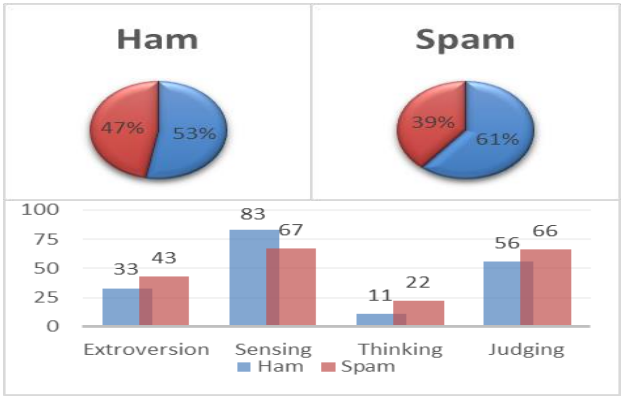


Fig. 21. Dataset personality recognition and sentiment analysis descriptive experiment [24].

This experiment uses CSDMC 2010 Spam corpus. TREC 2007 verifies. Descriptive evaluation Sentiment analyzers and personality recognition algorithms reveal in Fig. 20. BLR and DMNBtext for InfoGainAttributeEval attribute selection [24]. Everyone understands email spam. "Is this email genuine or spam?" Non-spam is shorter and snappier. Personality identification separates messages. The personality analysis creates a new dataset. Communication personality data concludes. Text type affects personality model dimensions. Emails differ. Spam and ham are different. Selected sentiment classifiers analyze message polarity. Derived polarity, like personality, adds three new datasets (one for each classifier). Fig. 21 shows spam positive and ham negative.

J. Discussion

Organized emails involve complex multi-view data production and email classification. The IFs and EFs evaluate email structure and word sequences, while numerous two-

feature dataset creation methods and unlabeled data selection have shown good classification accuracy stability. Unlabeled samples were examined, which reduced classification accuracy. It is algorithmic. Well-designed biometric authentication and malware detection classifiers attract deep learning projects [36]. Experiment 1 focused on improving security measures for email phishing avoidance, as humans often struggle to detect potential hazards. Machine learning (ML) was utilized to successfully catch spam emails. However, since phishing emails are relatively rare, this skewed the email data, leading to the need for a more realistic model training. The detection model employed several algorithms, including Random Forest, Multinomial Naive Bayes, Support Vector Machine (SVM), AdaBoost, and Logistic Regression. Among these, Random Forest achieved the best classification rate for phishing emails at 99%. AdaBoost, Multinomial Naive Bayes, and Logistic Regression achieved a precision rate of 96%. On the other hand, SVM could only identify 16.85% of phishing emails [29]. In Experiment 2, the focus was on phishing attacks, where cybercriminals create fake websites to deceive users and obtain passwords and financial data. Phishers copy legitimate websites, making it challenging to differentiate between real and fraudulent sites. The effectiveness of anti-phishing measures diminishes in such cases. To address this issue, an ML model was developed using URL- and hyperlink-based feature extraction, which involved examining strings to identify patterns. Two feature extraction methods (URL-based and hyperlink-based) were employed to process the raw input data. Experimentally, the URL-based feature extraction in combination with SVM exhibited the highest accuracy (98.05%), specificity (98.24%), recall (97.86%), precision (94.34%), and F1-score (95.65%) [30].

IV. Conclusion

The Uclassify algorithm is one of the best machine learning algorithms to detect phishing and emphasize internal and external features; Uclassify can classify the message as phishing or not phishing. Based on the results, we determined that Uclassify outperforms some existing algorithms. It can detect phishing or spoofing. This algorithm was evaluated on three databases: Kaggle Phishing, Email Collection and PhishTank.

V. Future Works

Email saves photos, links, directions, and metadata. Email integration may enhance performance. K-nearest neighbors and neural networks will be compared to the current system [40]. Experiment 1 conclusions may help future researchers pick a classified [29]. In future work, we intend to add new features to detect malware-containing fraudulent websites. Our approach could not detect malware affixed to fraudulent web pages. Today, blockchain technology is more prevalent and is an ideal target for phishing attacks, such as blockchain-based phishing schemes. Blockchain is an open and distributed ledger that can effectively register transactions between receiving and sending parties, demonstrably and continuously, making it popular among investors. Detecting phishing schemes in the blockchain environment thus requires additional research and development. Detecting phishing attacks on mobile devices is also an essential topic in this field, as the prevalence of smartphones has made them a common target for phishing attacks [39]. Our model classifies emails well. Many unsolved questions may affect email classification. Email settings affect IFs and EFs' word order and email series. Unknown data selection intrigues. Unlabeled datasets may be "weak" if randomly selected. It may impair classification precision [36].

Acknowledgment

This research has been accomplished at the University of Ha'il -Saudi Arabia.

References

[1] Priestman, W., et al., Phishing in healthcare organisations: Threats, mitigation and approaches. BMJ health & care informatics, 2019. 26(1).

[2] Yeng, P., B. Yang, and E. Snekkenes. Observational measures for effective profiling of healthcare staffs' security practices. in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). 2019. IEEE.

[3] Sendelj, R. and I. Ognjanovic, Cybersecurity Challenges in Healthcare, in Achievements, Milestones and Challenges in Biomedical and Health Informatics. 2022, IOS Press. p. 190-202.

[4] Georgiadou, A., et al. Hospitals' cybersecurity culture during the COVID-19 crisis. in Healthcare. 2021. MDPI.

[5] Al-Qahtani, A.F. and S. Cresci, The COVID-19 scamdemic: A survey of phishing attacks and their countermeasures during COVID-19. IET Information Security, 2022. 16(5): p. 324-345.

[6] Salloum, S., et al., A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access, 2022.

[7] Gangavarapu, T., C. Jaidhar, and B. Chanduka, Applicability of machine learning in spam and phishing email filtering: review and approaches. Artificial Intelligence Review, 2020. 53: p. 5019-5081.

[8] Smadi, S.M., Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. 2017: University of Northumbria at Newcastle (United Kingdom).

[9] Jupin, J.A., et al., Review of the machine learning methods in the classification of phishing attack. Bulletin of Electrical Engineering and Informatics, 2019. 8(4): p. 1545-1555.

[10] Akanksha, K., et al., Email Security. Journal of Image Processing and Intelligent Remote Sensing (JIPIRS) ISSN 2815-0953, 2022. 2(06): p. 23-31.

[11] Finker, C., Mail Authorship Verification and Phishing Recognizing with Machine Learning on iOS. 2020, University of Applied Sciences.

[12] Shen, K., et al. Weak Links in Authentication Chains: A Large-scale Analysis of Email Sender Spoofing Attacks. in USENIX Security Symposium. 2021.

[13] Shahrivari, V., M.M. Darabi, and M. Izadi, Phishing detection using machine learning techniques. arXiv preprint arXiv:2009.11116, 2020.

[14] Ali, G., M. Ally Dida, and A. Elikana Sam, Two-factor authentication scheme for mobile money: A review of threat models and countermeasures. Future Internet, 2020. 12(10): p. 160.

[15] Atlam, H.F. and O. Oluwatimilehin, Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review. Electronics, 2023. 12(1): p. 42.

[16] Khan, R. and M.A. Islam. Quantification of PIR protocols privacy. in 2017 International Conference on Communication, Computing and Digital Systems (C-CODE). 2017. IEEE.

[17] Sharma, T., Evolving Phishing Email Prevention Techniques: A Survey to Pin Down Effective Phishing Study Design Concepts. 2021.

[18] Nifakos, S., et al., Influence of human factors on cyber security within healthcare organisations: A systematic review. Sensors, 2021. 21(15): p. 5119.

[19] Shirey, R., Internet security glossary, version 2. 2007.

[20] Alwanain, M.I., An Evaluation of User Awareness for the Detection of Phishing Emails. International Journal of Advanced Computer Science and Applications, 2019. 10(10).

[21] Md, A.Q., et al., Efficient Dynamic Phishing Safeguard System Using Neural Boost Phishing Protection. Electronics, 2022. 11(19): p. 3133.

[22] Hu, H. and G. Wang. End-to-End Measurements of Email Spoofing Attacks. in USENIX Security Symposium. 2018.

[23] Siddique, Z.B., et al., Machine learning-based detection of spam emails. Scientific Programming, 2021. 2021: p. 1-11.

[24] Ezpeleta, E., et al., Novel email spam detection method using sentiment analysis and personality recognition. Logic Journal of the IGPL, 2020. 28(1): p. 83-94.

[25] Diaz Jr, M.O., A Domain-Specific Evaluation of the Performance of Selected Web-based Sentiment Analysis Platforms. International Journal of Software Engineering and Computer Systems, 2023. 9(1): p. 01-09.

[26] Eftimie, S., R. Moinescu, and C. Răcuciu, Spear-phishing susceptibility stemming from personality traits. IEEE Access, 2022. 10: p. 73548-73561.

[27] Patil, S. and S. Dhage. A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework. in 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). 2019. IEEE.

[28] Gibson, S., et al., Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms. IEEE Access, 2020. 8: p. 187914-187932.

[29] Livara, A. and R. Hernandez. An Empirical Analysis of Machine Learning Techniques in Phishing E-mail detection. in 2022 International Conference for Advancement in Technology (ICONAT). 2022. IEEE.

[30] Penta, U.B., B. Panda, and S.S. Gantayat, MACHINE LEARNING MODEL FOR IDENTIFYING PHISHING WEBSITES. Journal of Data Acquisition and Processing, 2023. 38(1): p. 2455.

[31] Somesha, M. and A.R. Pais, Classification of Phishing Email Using Word Embedding and Machine Learning Techniques. Journal of Cyber Security and Mobility, 2022: p. 279–320-279–320.

[32] Kadam, S., et al. Word embedding based multinomial naive bayes algorithm for spam filtering. in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). 2018. IEEE.

[33] Octaviani, N.L., et al. Comparison of multinomial naïve bayes classifier, support vector machine, and recurrent neural network to classify email spams. in 2020 International Seminar on Application for Technology of Information and Communication (iSemantic). 2020. IEEE.

[34] Ruskanda, F.Z., Study on the effect of preprocessing methods for spam email detection. Indonesia Journal on Computing (Indo-JC), 2019. 4(1): p. 109-118.

[35] Prosun, P.R.K., K.S. Alam, and S. Bhowmik. Improved Spam Email Filtering Architecture Using Several Feature Extraction Techniques. in Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021. 2022. Springer.

[36] Li, W., et al., Design of multi-view based email classification for IoT systems via semi-supervised learning. Journal of Network and Computer Applications, 2019. 128: p. 56-63.

[37] Taye, M.M., Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers, 2023. 12(5): p. 91.

[38] Jon Kågström, f., Roger Karlsson, Emil Ingridsson. uClassify. 2008-2023; Available from: https://www.uclassify.com/.

[39] Aljofey, A., et al., An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 2022. 12(1): p. 8842.

[40] Ma, T.M., K. Yamamori, and A. Thida. A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification. in 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). 2020. IEEE.