An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services

A. Kovač*, I. Dunđer* and S. Seljan*
* Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences, Zagreb, Croatia
akovac@ffzg.hr; ivandunder@gmail.com; sanja.seljan@ffzg.hr

***Abstract* – Phishing attacks have become today one of the most common security breaches performed on different communication channels. Their goal is to direct users to malicious websites or to infect a user's computer in an effort to acquire personal or sensitive data for later misuse. Phishing is often the first step in the process of cybercrime, and in order to be able to recognize potential attacks and adequately protect users, it is necessary to understand the underlying principles of attack strategies. Therefore, applying machine learning for training a system that would recognize phishing messages would be essential for increasing the level of security from cyberattacks. The aim of this paper is to analyze the diverse types of phishing messages and to provide an overview of machine learning techniques used for the detection of phishing (and spam) e-mails, hereby mainly focusing on regression and classification algorithms. In addition to the mentioned techniques, an analysis of datasets that are used for training of systems for detecting phishing attacks (and spam) is presented with regard to their size, language and accuracy scores.**

Keywords - information security; information privacy; machine learning; phishing; attack detection; spam; electronic messages

I. INTRODUCTION

Phishing attacks have become one of the most common security breaches today. They are executed on different communication channels with a goal to acquire personal or sensitive data from users for later misuse. Phishing as a term was coined from three words – "**p**assword", "**h**arvesting" and "**fishing**", broadly defined as the process of password harvesting or password "fishing". Phishing is a common type of social engineering attacks, related to fraudulent actions using various internet services with the main aim to steal not only passwords, but various types of confidential and sensitive data [1] such as login or credit card credentials and other precious information [2].

Even though they share some similarities, phishing attacks are to be distinguished from spam messages. While phishing attacks are characterized by the intention of stealing personal or sensitive data, spam messages are unwanted and tiresome messages that often come in form of marketing materials and advertisement, but without malicious intents.

According to [3], three most common types of data that were misused are credentials (usernames, passwords, pin numbers), personal data (name, e-mail, addresses) and medical data (treatment information, insurance claims etc.).

Phishing messages are often integrated into electronic messages that motivate users to click on a link that takes them further to fake websites that imitate real websites, such as websites of banks or online payment services, and where they are asked to provide bank credentials, account information, personal data etc. Users are potentially under attack even if they only open a suspicious link, whereas the real danger keeps growing if a user starts to enter valuable personal information into entry fields that are prepared specifically for storing data of unsuspecting users.

Attacks can be also carried out through social media platforms or instant messaging applications, which have become one of the most usable media nowadays. Social media represent an almost ideal target, since it is easier for attackers to gather personal account information, private phone numbers, data on daily activities, interests, relationships etc. What makes it even harder for users to detect phishing attacks is the fact that legitimate organizations, such as banks or insurance companies, have their official phone numbers and essential information shown publicly on their websites, so attackers can misuse this kind of information to approach potential victims who are therefore not always able to distinguish if the message is authentic or not.

Knowledge of underlying principles of attack strategies is fundamental for recognizing potential attacks and adequate protection of users. Reference [4] proposed a new phishing anatomy that includes attack phases, attacker types, vulnerabilities, threats, targets, attack mediums, and attacking techniques in order to present the lifecycle of a phishing attack and to increase awareness.

Another group of authors made a distinction between social engineering and technical aspects of phishing attacks, and proposed a taxonomy consisting of phases and classes for each criterion [3].

Training a system that would be able to recognize phishing messages would be essential for increasing the level of security from cyberattacks. There are various approaches that can be employed for analyzing phishing and spam messages, including Natural Language Processing (NLP) techniques, the use of blacklists or whitelists, content evaluation, hybrid methods or the use of machine learning [5], which are at the center of this paper.

Here the authors present an **overview of supervised machine** learning techniques that can used for phishing and spam detection, mainly **focusing on regression and classification**. In addition to these techniques, an analysis of datasets that are nowadays utilized for training of machine learning systems for detecting phishing attacks or spam messages is given with regard to **size** and **language** of datasets, and their **accuracy score** obtained during performance testing. The paper also provides an analysis of different types of phishing messages.

Literature review in this paper is based on research papers and other works that are amongst others included in Web of Science and IEEE databases, and by using Boolean expressions and key words, such as: ("mail" AND "phishing" AND "machine learning") NOT ("website" OR "webpage") NOT ("neural network") NOT ("deep learning"). Extracted papers were analyzed and selected according to applied machine learning algorithms, accuracy results and dataset characteristics.

The organization of the paper is as follows: after the Introduction section, the next section presents information on recent phishing attacks on electronic messaging services. The third section explains the distinct types of phishing, whereas the fourth section gives an overview of machine learning algorithms that are suitable for detecting spam and phishing e-mails. The fifth section deals with the most frequently used machine learning techniques, the size and language of datasets that are used for training of machine learning systems and their corresponding accuracies. The closing section concludes the paper and suggests future research.

II. Recent Phishing Attacks on Electronic Messaging Services

Reference [3] reported that phishing e-mails generally contain links to malicious websites (38%) or contain malicious attachments (38%). Seventy-five percent of organizations around the world experienced some kind of phishing in 2020, whereas 96% of phishing attacks were initiated with the help of e-mails, 3% with fake websites and 1% through voice phishing.

According to [6], social engineering attacks in 2019 were the number one threat for individual users and the number two for organizations. Namely, social engineering often appears as the first phase of a cybercrime [7]. Reference [1] reported an increase of phishing campaigns during the COVID-19 crisis. Reference [8] reported an increase of 76% of phishing attacks in 2020 in comparison to 2019. According to the same source, most of the phishing servers were located in Hong Kong, China.

In many cases the human factor, which includes stress, fear, anxiety, risk-taking, demographic factors and education level, decides on the success of a phishing attack [9]. Due to the fact that during the COVID-19 crisis people spent more time using their computers, the number of cyberattacks has increased 35% [8]. During that period phishing e-mails often impersonated government bodies, medical and health organizations asking for disease-related information, offering testing methods and treatment, financial help for government packages, personal health equipment etc.

Reference [10] reported that from May to August 2021 there was a 7.3% increase of phishing attacks. The increase of phishing campaigns was also reported by [11] between 2019 and 2020. Especially big organizations were targets of phishing campaigns, with a large number of victims.

When analyzing the types of losses, security officers reported the following: 60% of organizations lost data, 52% of organizations had their accounts or other credentials compromised, 47% of organizations were infected by ransomware, 29% by malware, whereas 18% of organizations experienced financial loss [3].

According to [3] and [12], the most targeted sector by phishing attacks was the financial sector (60% more than the next sector), followed by higher education. In 2021, the most targeted industries were retail, manufacturing, food and drinks, research and development, and technology.

According to [13], in 2020 77% of organizations experienced phishing attacks and in 2021 this number raised to 86%. The same source reported that 99% of organizations had formally some sort of security training programs, but only 57% of them actually did provide training, and less than 50% covered the topic of phishing in their programs.

According to [3], countries that experienced the most of phishing attacks in 2020 were the United States, the United Kingdom, Australia, Japan, Spain, France and Germany. Companies that were impersonated most often were Microsoft, ADP, Amazon, Adobe Sign, Zoom and public bodies, which all have frequent communication via e-mail with their customers.

A group of authors performed research on the occurrence of phishing attacks in European countries, and showed interestingly that more educated persons are more susceptible to phishing attacks [14].

III. Types of Phishing Attacks

There are several types of phishing attacks [15] which are performed through various communication channels, and which can be differentiated by content, type of targeted action (stealing data, clicking, computer infection etc.) and type of target user.

E-mail phishing: Phishing via e-mails most often comes in the form of messages that contain one or more fake URL addresses which redirect a user to a trap website. Such websites may be camouflaged and displayed as official websites of financial corporations

[16]. E-mail phishing attacks can contain a message with tampered links that alert or warn a user to perform an action of interest to an attacker (so-called "phisher"), for example, to immediately update sensitive data. Unfortunately, if the victim proceeds, all entered information will fall into the hands of the attacker [15]. Phishing e-mails may also contain fake webforms that ask for sensitive data, links to videos or pictures that present fake news etc.

Instant messaging phishing: In instant messaging applications a threat comes again with a suspicious URL, and here the attacker tries to get a user's sensitive data, especially passwords. The phisher might use voice chats, textual chats or even a combination of both [16]. Instant messaging applications and services are a favorable place to start group conversations with numerous users, and to share malicious links to everyone at once. On the other hand, according to [16] online social networks such as Facebook and Twitter have a "rapid growth of phishing attacks for several reasons: 1) it is easy to impersonate people and create fake profiles, 2) users' willingness to trust and 3) popularity of social networking sites".

Smishing: According to [15], SMS phishing or shortly smishing is a type of "social engineering attack carried out through SMS in order to steal user data including personal information, financial information and credentials", which often results in "money laundering activities". These attacks appear in form of an SMS that is sent to a user's cell phone, and which also contain misleading information or malicious links to harmful websites. If users respond to the instructions given by the attacker or click on the provided link in the SMS, they are automatically redirected to a fake website and run the risk of revealing their valuable data. For example, during holidays attackers take advantage of seasonal discounts and send such links to victims, which are attracted by fake discounts for products that do not actually exist. This is similar to instant messaging phishing; however, it should still be distinguished since these two types of attack are carried out through different services for exchanging messages.

Bulk phishing: Here phishing is attempted massively within the same organization, aiming to imitate communication between the company and the user.

Spear phishing: This type of attack is focused on a specific group of users or an organization. Here the attackers need to know more information about specific users or the inner workings of an organization, especially its power structure [15]. A paper emphasizes that "spear phishing attacks require understanding of the organization's context to create effective phishing e-mails" [16]. In that sense, attackers study an organization or their victims for a longer period. Spear phishing attacks are carried out carefully and thoughtfully, and if attackers succeed in stealing sensitive data of a targeted organization, consequences might be catastrophic for an entire population or society as a whole.

Whaling: This is a type of phishing attack where an attacker acts like a senior member of an organization and performs an attack on other employees of the organization – all for the purpose of retrieving sensitive data from employees [17]. According to [15], whaling is a form of spear phishing that targets high-profile employees.

Vishing or voice phishing: This type of phishing is conducted through phone calls or voice messages. The attackers pretend to represent a company and try to access personal information.

Pharming: This type of phishing manipulates website traffic, aiming to steal passwords and usernames. It is mostly used in e-commerce and on (fake) bank websites. In this case a malicious code is installed on the computer or on a server. Through misdirection on websites, attackers can steal credit card information, bank account information or various passwords. Pharming might include the tampering of DNS in order to redirect user from a domain to a malicious website. A DNS server translates names into IP addresses in form of numbers. So, if an attacker (so-called "pharmer") changes these details, the computer will be using a corrupted IP address for accessing a specific website.

Some common examples of phishing attacks include fake documents, such as invoices, and attempt different types of fraud (so-called "scam"), generally through e-mails that ask for account information updates, personal data or financial help. They regularly contain messages on settlements for resolving problems by clicking on links, or messages from a human resource department. They might also come in form of fake alerts on "unusual" computer activities that require immediate action etc. However, regardless of the different fields of application, diverse scenarios and communication channels, the goal is always to steal personal data or sensitive information from users.

According to [18], there are four critical steps during a phishing attack via e-mails:

1. The attacker creates a phishing website that has almost identical visual identity like the legitimate one, so it can attract more users.

2. The attacker creates an e-mail that often includes created phishing websites and sends it to a large number of potential victims. If there is a case of e.g., spear phishing, the pool of potential victims is narrowed down to a specific target group.

3. The victim opens the e-mail and clicks on the link that redirects to a phishing website, where the user is asked to provide sensitive information such as bank credentials.

4. The phisher gains that way valuable information from victims and misuses them.

In a phishing campaign, messages are often sent massively asking users to act by clicking on various links or sharing information, whereas in spear phishing and whaling attacks the focus is on specific organizations or their senior members that have a higher level of responsibility and more access to valuable resources.

Phishing and spear phishing were the most common incidents during the COVID-19 crisis, often appearing as the first step of trying to gain access to a network by using stolen credentials with the help of malware that is automatically installed after a victim has clicked on a

malicious link or opened a harmful attachment. Malicious e-mails are currently reported as the most frequent type of incident [8].

IV. Machine Learning Algorithms

Machine learning approaches differ by methods and techniques. The main approaches are supervised, unsupervised, semi-supervised and reinforcement learning [19]. According to [20], supervised learning implies that a machine learns under guidance, using labeled data and defining output. In unsupervised learning the machine uses unlabeled data and finds a way to "understand" hidden data and structure in order to create an output. In semi-supervised learning the training dataset consists of a small number of labeled examples and a large number of unlabeled examples [21], and therefore stands between supervised and unsupervised learning. In reinforcement learning, the machine uses an established pattern, but the input depends on a specific action and context. Here the machine learns from past experiences and interacts with the environment. The system uses the principle of reward and punishment in order to be trained.

The detection of spam and phishing content is generally researched using the supervised machine learning approach and with focus on two types of problems: classification and regression [22].

Classification is a well-established method for content categorization, as it enables the distinction and labeling of messages according to so-called classes, e.g. phishing or regular messages, spam or ham (i.e. no spam), true or false, users or non-users, but also other types of classes, such as those related to speech recognition, word tagging, natural language processing (NLP) tasks etc. Classification algorithms can be further divided into more fine-grained categories, such as multi-class classifier when there are more than two outcomes. Classification algorithms can be further divided into mainly two categories: linear models (logistic regression, support vector machine) and non-linear models (k-nearest neighbors, naïve Bayes, decision tree, random forest) [20, 23, 24]. In classification, the input values are discrete data, and the output is a discrete value, such as phishing or no-phishing, male or female. The final aim is to find the boundary that separates data from a dataset into different classes (Fig. 1). In a recent paper, seven classification algorithms were used to filter spam e-mails [25].

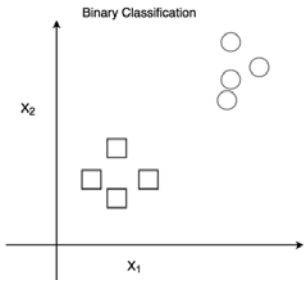

Figure 1. Binary classification

Another problem within supervised machine learning is **regression**. Here the main task is to find correlations between dependent and independent variables, i.e. to find a mapping function that can map an input variable (x) to a continuous output variable (y). It is used to predict continuous quantity variables, such as weather changes, salary, price, age, house prices, market trends etc. In regression analysis, the system is trained by input features and output labels, aiming to establish relationships among variables and to estimate to what extent features affect the output, i.e. the continuous target variable. There are various types of regression algorithms (linear and non-linear), such as simple linear regression, multiple linear regression, support vector regression, decision tree regression, random forest regression etc. In regression, the continuous output variable is a real value (integer or float). The final aim is to find the line of best fit which can predict the most accurate output (Fig. 2).

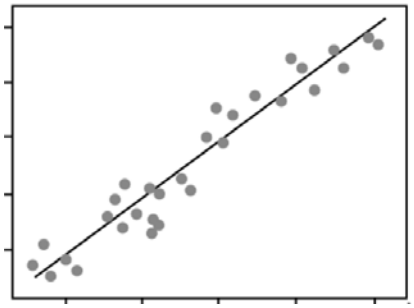

Figure 2. Regression model

What is common to classification and regression is that historical data predict future events [22]. The main difference between classification and regression is the data labeling approach: classification uses discrete labels from a finite set and here the labels are "counted", whereas regression is based on continuous labels with infinite possibilities for prediction, but here they are "measured" [20].

However, detecting phishing attacks is still quite complicated due to many reasons. For instance, strategies for crafting phishing messages are limited only by the imagination of the attacker. Furthermore, the complexity of languages and syntactic rules which vary from one language to another make it difficult for a computer to "understand" and distinguish messages. Some words can have multiple meanings and are context-sensitive, and as such cannot be approached always uniformly. In order to alleviate these problems, extensive data preprocessing and normalization are essential, as this helps to reduce errors, false positives and false negative outcomes.

Algorithms driven by machine learning have proven to be highly effective for detecting phishing attacks [26], however, they are not able to result in 100% accuracy. Nevertheless, according to previous research, they still reach over 90% [27].

Decision Tree: This classification algorithm is one of the most popular for classification and regression [15]. It is part of the supervised machine learning approach. A decision tree (DT) has two nodes – one is for making a decision, and the other is called a leaf node [28]. Nodes for making decisions are used to make many individual decisions and have different branches, whereas leaf nodes are the output of those decisions and do not contain any

further branches [28]. Decision tree is an algorithm in which data is "continuously split according to a certain parameter" [29].

According to [20], decision trees have many advantages including less data preparing and preprocessing, no need for data normalization and scaling, implementation simplicity, and clear structure comprehension through visualization. On the other hand, there are many disadvantages – complexity increases with the increase in labels, as well as the fact that a minor change can lead to a whole different tree structure etc. [20]. As a final disadvantage, especially important for e-mail classification, the author states that the algorithm requires more time to train the data, and that it tends to overload. A work mentions other advantages of the decision tree classifier, such as the ability of selecting discriminatory features, dealing with incomplete and noisy data, but also some disadvantages, such as calculation growth with increase of data and high classification error rate with a small dataset in comparison with the number of classes [30].

Random Forest: random forest (RF) is also a popular supervised machine learning algorithm, heavily employed for solving classification and regression problems. This algorithm contains numerous decision trees on different subsets of the dataset and "takes the average to increase the predictive accuracy of that dataset" [28]. Unlike decision trees, random forest makes predictions depending on each tree and bases the final decision on the maximum votes of predictions [28]. According to [15], random forest is an "ensemble classifier used for classification and regression, whereas decision trees are based on randomly selected sets in a training sample". On the other hand, accuracy of random forests depends on the larger variety of trees [29].

According to [20], advantages of the random forest algorithm are in the automatization of lost values in data and its efficiency in handling large datasets. On the other hand, disadvantages are in the context of more computing and more resources that are needed for efficient results. Random forest requires more time for training as well since it integrates many decision trees. In a research authors used the random forest algorithm to differentiate between phishing and legitimate URL addresses with accuracy of 86% [31].

Naïve Bayes: According to [20], the naïve Bayes (NB) algorithm assumes that the existence of a certain property in the class is not related to the existence of any other property. This algorithm is also used for classification tasks such as text classification and spam detection, and it is considered to be quite straightforward. According to [15], the naïve Bayes classifier is a "generative probabilistic model in machine learning and is based on the Bayes theorem".

The Bayes theorem is a "mathematical probabilistic technique which helps to calculate the conditional probabilities of an event" [20]. It is considered *naïve* as "it assumes that the presence of a feature in a class is independent of the presence of any other feature" [32].

Naïve Bayes is a very efficient machine learning algorithm with fast predictions based on the probability of the data [20, 33]. This algorithm counts the frequency (the numbers of words) and combination of values (words) in a dataset [34]. Finally, it is used in text classification such as e-mail classification, as well as problems with numerous classes [20].

In text classification tasks such as e-mail classification, there are two possible outcomes: a message is spam or no spam. Each e-mail dataset (spam and no spam) has its own term frequency, and based on that it is possible to detect spam or no spam messages according to model probabilities for new e-mails, i.e. messages that have not been seen previously in the dataset for model training [34]. However, in this context it is important to emphasize once more that spam e-mails do not have to be phishing e-mails.

In a research authors used a term frequency and inverse document frequency (TF-IDF) matrix for feature extraction and the naïve Bayes algorithm [35]. The authors obtained a true positive score of 91% for detecting spam messages. Reference [34] used the naïve Bayes classifier to identify unwanted e-mails using a dataset from Kaggle and obtained an accuracy higher than 99%. Another research used, amongst other methods, the naïve Bayes classifier to distinguish desirable (ham) from potentially harmful (spam) messages, and here the multinomial naïve Bayes classifier achieved the best score [36].

Support Vector Machine: Unlike the above-mentioned algorithms, support vector machine (SVM) is part of linear models as well as logistic regression. According to [20], support vector machines have many advantages which primarily relate to the functioning of semi-structured and unstructured data. It can also work with very complex data. In terms of disadvantages, they are the same as in decision trees – it takes more time to train the model for a larger dataset. According to [37], support vector machine is one of the "foremost usually used classifier in phishing e-mail detection". It has become exceedingly popular in the data mining community due to very efficient generalization performance and the ability to manipulate high-dimensional data. This algorithm has shown to be highly successful for document classification, especially when spam detection is approached as a binary classification problem [38-40]. A study used SVM for malware and phishing website detection with the help of discriminative features such as textual properties, link structures, webpage contents, DNS information and network traffic [41].

K-Nearest Neighbor: This algorithm is used for both regression and classification, but it is more commonly used in classification problems [20]. The k-nearest neighbor (KNN) algorithm is easy to understand and apply. It is based on homogeneous data, which means that it learns the pattern available within.

One of the advantages is that it is cheap and flexible, but what still does not make it ideal is that it requires a large dataset [20]. The "k" is a parameter that counts the nearest neighbor that will be included in the voting and

decision-making process, hence, the name of the algorithm itself. Its main advantage is that it is easy to implement, and it supports multi-class datasets. However, unrelated class characteristics may affect the accuracy of the model [20]. An author proposed an algorithm with high practical value that fuses KNN and SVM in order to overcome disadvantages of unbalanced sample data in KNN and the time-consuming model training in the SVM classifier [42]. Reference [32] performed research on phishing detection, and concluded that KNN outperformed other classifiers with an accuracy of 95% in detecting phishing websites, whereas the naïve Bayes classifier performed well for detecting authentic websites.

A paper used several approaches for phishing detection, especially focusing on machine learning. Here the authors emphasize that machine learning algorithms automatically assign weights of each feature by "programmatic statistical calculations", and this enables algorithms to ameliorate the flexibility issue of manually tuned rules. The authors conducted a research based on more than forty bag-of-words features that are extracted from e-mails using TF-IDF [43].

V. Datasets and Detection Accuracy

Machine learning algorithms generally provide high levels of accuracy when it comes to phishing or spam detection. Accuracy is here understood as the percentage of datasets being correctly categorized [44]. Results show values over 90% in [27], then 95% in [32, 46], 99% as in [34, 45] etc.

TABLE I. Overview of Datasets and Detection Accuracy

Ref.	*Algorithm*	*Accuracy (%)*	*Total dataset*	*Training dataset*	*Testing dataset*
[44]	NB	99.46	Spam: 2222 Legitimate: 3778	Spam: 63% Legitimate: 63%	Spam: 37% Legitimate: 37%
	SVM	96.20			
	KNN	96.20			
[47]	NB	83.5	Spam: 250 Legitimate: 250	Spam: 60% Legitimate: 60%	Spam: 40% Legitimate: 40%
[48]	SVM	97.6	Spam: 680 Legitimate: 4532	Spam: 82% Legitimate: 95%	Spam: 18% Legitimate: 5%
	DT	82.6			
[49]	RF	98.72	Spam: 481 Legitimate: 2412	90%	10%
	NB	94.94			
	SVM	98.42			
[45]	SVM	99.87	Phishing: 414 Legitimate: 1191	*NA*	*NA*
	RF	99.87			
	NB	99.81			
[50]	SVM	93.00	Phishing: 4000 Legitimate: 4000	80%	20%
	RF	91.00			
	NB	91.00			
[46]	SVM	95.00	Phishing: 500 Legitimate: 500	60%	40%

NA = not available; dataset size = number of e-mails

Table I presents results of experimental works, applied algorithms, the performance of the detection model shown as an accuracy score, and statistics of datasets that were used to train the machine learning model and to test (evaluate) its efficiency. All datasets contained mostly e-mails in English (with some non-English expressions).

The accuracy of algorithms for detecting phishing or spam messages in almost all cases exceeds 90%. The most accurate algorithms were support vector machine (SVM) and random forest (RF) with a score of 99.87%, closely followed by naïve Bayes (NB), as shown in [45]. Interestingly, these scores were obtained after training with relatively small datasets (414 phishing and 1191 legitimate e-mails).

In the research performed by [50] the RF algorithm obtained an accuracy score of 91%. In a different paper the best score was obtained by the naïve Bayes algorithm, followed by SVM and KNN (over 96%) [44].

Specifically, when detecting spam messages which are similar to phishing messages the best result was achieved by the naïve Bayes algorithm with an accuracy of 99.46% [44].

When it comes to the size of the dataset, it varied among different studies. Smaller datasets consisted of 500 e-mails as in [47], 500 phishing and 500 legitimate as in [46], or 414 phishing and 1191 legitimate e-mails as in [45]. Accuracy scores in the aforementioned research ranged from 83.50% up to 99.87%. The worst result was achieved when using the dataset with the least data [47] with an accuracy of only 83.50%.

A larger dataset was used by [50] containing 4000 phishing and 4000 legitimate e-mails. Here the authors obtained a score of 93% for SVM, and 91% for NB and RF. Another larger dataset was used in [44] consisting of 2222 spam and 3778 legitimate messages, and scored 99.46% for NB and 96.20% for SVM and KNN.

When exploring the ratios of sizes of training and testing datasets, they ranged from 60:40 [46] to 90:5 [48].

Machine learning features also varied from research to research. For instance, [44] used the "e-mail subject", "from field" and "body" from messages as the main source of information. Reference [47] used a feature set consisting of alphanumeric words, language, grammatical or spelling errors, inappropriate words, i.e. words that are related to advertisement of some products or services etc. In [45], the features that were extracted were links, tags (such as the <script> tag) and words. In [46], eighteen header-based, URL-based, script-based and psychological features were used. In [48], the dataset has been filtered using DNSBL (domain name system blacklist) and anti-spam filtering applications, which helped in creating appropriate datasets. In [49], the dataset was preprocessed by removing HTML tags, separation tokens and duplicate e-mails so that only the data of the body and the subject were kept.

VI. Conclusion

The aim of this paper was to present an overview of machine learning algorithms that can be used for phishing (and spam) detection, focusing primarily on classification and regression. Results show that machine learning algorithms achieved high accuracy and scored mainly above 90%. The most accurate models were based on

SVM and RF with a score of 99.87%, while in some other research they gained lower results. Similar scores were also obtained by other algorithms, such as NB, KNN and DT. The analyzed supervised machine learning algorithms overall offer an immense potential for phishing and spam detection. However, they are still not achieving perfect accuracy.

The size of the explored datasets varied from smaller (400-500 e-mails) to larger ones (2000-4000). The ratio of the training and testing datasets also varied, sometimes even significantly.

The paper also provides an analysis of different types of phishing messages. Analyzed sources confirmed that phishing e-mails represent a considerable security threat for individuals and organizations which experienced loss of data, compromised accounts and credentials, computer infections and financial loss. Therefore, education and raising awareness is an essential factor in order to fight data breach and to increase security. It is necessary to provide security training programs for all types of users, regardless of age, education level, employment or industry.

For future work, the authors of this paper plan to conduct a more detailed survey of recent literature, and a more fine-grained analysis of available datasets. Also, the authors plan to acquire data for the Croatian language, to assess various machine learning algorithms for their efficiency, and to compare experimental results with accuracy scores that were obtained by other authors. This would also include devising guidelines for feature extraction and selection in machine learning models for the Croatian language.

REFERENCES

[1] Australian Cyber Security Centre, "ACSC Annual Cyber Threat Report: July 2019 to June 2020", https://www.cyber.gov.au/sites/default/files/2020-09/ACSC-Annual-Cyber-Threat-Report-2019-20.pdf

[2] M. Dadkhah, S. Shamshirband, and A. Wahab, "A hybrid approach for phishing web site detection", The Electronic Library, vol. 34, no. 6, pp. 927–944, 2016.

[3] M. Rosenthal, "Must-Know Phishing Statistics: Updated 2022", Tessian, https://www.tessian.com/blog/phishing-statistics-2020/

[4] Z. Alkhalil, C. Hewage, L. F. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy", Frontiers in Computer Science, vol. 3, article 563060, 2021.

[5] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs", Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[6] Positive Technologies, "Cybersecurity Threatscape: Q3 2019", https://www.ptsecurity.com/ww-en/analytics/cybersecurity-threatscape-2019-q3/?sphrase_id=70070, 2019.

[7] J. Rastenis, S. Ramanauskaitė, J. Janulevičius, A. Čenys, A. Slotkienė, and K. Pakrijauskas, "E-mail-Based Phishing Attack Taxonomy", Applied Sciences, vol. 10, no. 7, 2363, 2020.

[8] APCERT, "Annual Report 2020", https://www.apcert.org/documents/pdf/APCERT_Annual_Report_2020.pdf

[9] H. Abroshan, J. Devos, G. Poels, and E. Laermans, "COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts During the Pandemic", IEEE Access, vol. 9, pp. 121916–121929, 2021.

[10] ESET, "Enjoy Safer Technology. Threat Report 2021: T2 2021", https://www.welivesecurity.com/wp-content/uploads/2021/09/eset_threat_report_t22021.pdf

[11] IBM, "X-Force Threat Intelligence Index 2022", https://www.ibm.com/downloads/cas/ADLMYLAZ

[12] CISCO, "Cybersecurity threat trends: phishing, crypto top the list", https://umbrella.cisco.com/info/2021-cyber-security-threat-trends-phishing-crypto-top-the-list

[13] Proofpoint, "Threat Report: 2022 State of the Phish - An In-Depth Exploration of User Awareness, Vulnerability and Resilience", https://www.proofpoint.com/us/resources/threat-reports/state-of-phish

[14] M. Pejić Bach, T. Kamenjarska, and B. Žmuk, "Targets of phishing attacks: The bigger fish to fry", Procedia Computer Science, 2022, in press.

[15] P. Kalaharsha, and B. Mehtre, "Detecting Phishing Sites - An Overview", arXiv:2103.12739 [cs.CR], p. 13, 2001.

[16] A. Aleroud, and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey", Computers & Security, vol. 68, pp. 160–196, 2017.

[17] A. Arshad, A. U. Rehman, S. Javaid, T. M. Ali, J. A. Sheikh, and M. Azeem, "A Systematic Literature Review on Phishing and Anti-Phishing Techniques", Pakistan Journal of Engineering and Technology (PakJET), vol. 4, no. 1, pp. 163–168, 2021.

[18] M. K. P. Madushanka, and A. L. Hanees, "Phishing E-Mail Filtering Mechanism Using Heuristic Technique", Proceedings of the 5th Annual Science Research Sessions-2016, Sri Lanka, pp. 261–271, 2016.

[19] R. Saravanan, and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification", Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 945–949, 2018.

[20] M. Gupta, and S. D. Pandya "A Comparative Study on Supervised Machine Learning Algorithm", International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 10, no. 1, pp. 1023–1028, 2022.

[21] O. Chapelle, B. Schölkopf, and A. Zien, Eds., Semi-Supervised Learning / Adaptive Computation and Machine Learning series. London, United Kingdom: MIT Press, 2006.

[22] V. Nasteski, "An overview of the supervised machine learning methods", DOI: 10.20544/HORIZONS.B.04.1.17.P05, UDC: 004.85.021:519.718, 2017.

[23] N. Dutta, U. Subramaniam, and S. Padmanaban, "Mathematical models of classification algorithm of Machine learning", QScience Proceedings of the International Meeting on Advanced Technologies in Energy and Electrical Engineering, vol. 3, 2018.

[24] R. Caruana, and A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms", Proceedings of the 23rd international conference on Machine learning - ICML '06, pp. 161–168, 2006.

[25] P. S. Teja, C. Amith, K. Deepika, and K. S. Raju, "Prediction of Spam Email using Machine Learning Classification Algorithms", International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 9, no. 6, 2021.

[26] M. Ahsan, R. Gomes, and A. Denton, "SMOTE Implementation on Phishing Data to Enhance Cybersecurity", Proceedings of the 2018 IEEE International Conference on Electro/Information Technology (EIT), pp. 0531–0536, 2018.

[27] I. Skula, "Automated detection techniques of phishing", https://www.researchgate.net/publication/352523092_Automated_detection_techniques_of_phishing, p. 8, 2021.

[28] K. Nikhil, D. S. Rajesh, and D. Raghavan, "Phishing Website Detection Using ML", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), vol. 7, no. 4, pp. 194–198, 2021.

[29] H. Kumar, A. Prasad, N. Rane, N. Tamane, and A. Yeole, "Dr. Phish: Phishing Website Detector", International Advanced Research Journal in Science, Engineering and Technology (IARJSET), vol. 8, no. 1, pp. 176–182, 2021.

[30] F. Cardoni, and E. A. M. Cappella, "A comparison of decision tree and Naïve Bayes Classifier as spam filtering algorithms", https://www.researchgate.net/publication/322791442_A_COMPARISON_OF_DECISION_TREE_NAIVE_BAYES_CLASSIFIERS_AS_SPAM_FILTERING_ALGORITHMS, 2018.

[31] S. S. Ravindra, S. J. Sanjay, S. N. A. Gulzar, and K. Pallavi, "Phishing Website Detection Based on URL", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), vol. 7, no. 3, pp. 589–594, 2021.

[32] N. Bhoj, R. Bawari, A. Tripathi, and N. Sahai, "Naive and Neighbour Approach for Phishing Detection", Proceedings of the 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), India, pp. 171–175, 2021.

[33] K. Prasanthi, T. Deepika, S. Anudeep, and M. Koushik, "An Efficient Email Spam Detection using Support Vector Machine", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 2, 2019.

[34] M. Jaiswal, S. Das, and K. Khushboo, "Detecting spam e-mails using stop word TF-IDF and stemming algorithm with Naïve Bayes classifier on the multicore GPU", International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 4, pp. 3168–3175, 2021.

[35] J. Shobana, and D. Kanchana, "An Efficient Spam SMS Analysis Model based on Multinomial Naïve Bayes model Using Passive Aggressive Algorithm", Journal of Physics: Conference Series, p. 7, 2007.

[36] S. Rapacz, P. Chołda, and M. Natkaniec, "A Method for Fast Selection of Machine-Learning Classifiers for Spam Filtering". Electronics, vol. 10, no. 17, 2021.

[37] M. Alauthman, A. Almomani, M. Alweshah, W. Omoush, and K. Alieyan, "Machine Learning for Phishing Detection and Mitigation", in Machine Learning for Computer and Cyber Security: Principles, Algorithms, and Practices, B. B. Gupta, M. Sheng, Eds. CRC Press, 1st Edition, p. 27, 2019.

[38] G. Rios, and H. Zha, "Exploring Support Vector Machines and Random Forests for Spam Detection", Proceedings of the First Conference on Email and Anti-Spam (CEAS 2004), USA, p. 6, 2004.

[39] Z. S. Torabi, M. H. Nadimi-Shahraki, and A. Nabiollahi, "Efficient Support Vector Machines for Spam Detection: A Survey", International Journal of Computer Science and Information Security, vol. 13, no. 1, 2015.

[40] L. U. Oghenekaro, and A. T. Benson, "Text Categorization Model Based on Linear Support Vector Machine", American Academic Scientific Research Journal for Engineering, Technology, and Sciences, vol. 85, no. 1, 2022.

[41] K. Rashmi, and G. M Bhandari, "Support Vector Machine Based Malware and Phishing Website Detection", International Journal of Computing and Technology (IJCAT), vol. 3, no. 5, pp. 295–300, 2016.

[42] G. Zengle, "A fusion algorithm model based on KNN-SVM to classify and recognize spam", Journal of Physics: Conference Series, p. 7, 2021.

[43] G. Park, and J. Rayz, "Ontological Detection of Phishing Emails", Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2858–2863, 2018.

[44] W. A. Awad, and S. M. ELseuofi, "Machine Learning methods for E-mail Classification", International Journal of Computer Applications, vol. 16, no. 1, pp. 39–45, 2011.

[45] S. Rawal, B. Rawal, A. Shaheen, and S. Malik, "Phishing Detection in E-mails using Machine Learning", International Journal of Applied Information Systems, vol. 12, no. 7, pp. 21–24, 2017.

[46] Z. Yang, C. Qiao, W. Kan, and J. Qiu, "Phishing Email Detection Based on Hybrid Features". IOP Conference Series: Earth and Environmental Science 252, DOI:10.1088/1755-1315/252/4/042051, pp. 1–10, 2019.

[47] P. Sharma, and U. Bhardwaj, "Machine Learning based Spam E-Mail Detection", International Journal of Intelligent Engineering and Systems, vol. 11, no. 3, pp. 1–10, 2018.

[48] A. Yüksel, Ş. F. Çankaya, and I. Üncü, "Design of a Machine Learning Based Predictive Analytics System for Spam Problem", Acta Physica Polonica A, vol. 132, no. 3, pp. 500–504, 2017.

[49] I. Santos, C. Laorden, B. Sanz, and P. Bringas, "Enhanced Topic-based Vector Space Model for semantics-aware spam filtering", Expert Systems with Applications, vol. 39, no. 1, pp. 437–444, 2012.

[50] A. ALmomani, T.-C. Wan, A. Manasrah, A. Altaher, M. Baklizi, and S. Ramadass, "An enhanced online phishing e-mail detection framework based on "Evolving connectionist system"". International Journal of Innovative Computing, Information and Control (IJICIC), vol. 9, no. 3, pp. 1065–1086, 2013.