DOI: 10.47743/eufire-2023-1-12

DETECT THE TENTATIVE BEFORE BECOMING REAL: A MACHINE LEARNING APPROACH FOR PHISHING EMAIL DETECTION IN ROMANIAN HEALTHCARE

GEORGE – BOGDAN MERTOIU
Alexandru Ion Cuza University of Iași
Iași, Romania
george.mertoiu@student.uaic.ro

GABRIELA MEȘNIȚĂ
Alexandru Ion Cuza University of Iași
Iași, Romania
gabriela.mesnita@feaa.uaic.ro

Abstract
Phishing attacks pose a significant threat to individuals and organizations, and their accurate and effective detection is crucial to preventing data breaches and financial losses. With the increasing use of email as a communication channel, phishing attacks have become more widespread and sophisticated. Our study addresses the use of machine learning-based models to detect phishing emails by analyzing the text of the message. A characteristic of the study is given by the fact that it uses a dataset composed of private emails in Romanian, obtained from public institutions in the field of health. Since the models were applied to the text, natural language processing techniques specific to the Romanian language were used to extract the features. The results obtained highlighted that some models outperform others in terms of accuracy, underlining the importance of choosing a machine learning approach for phishing detection in a given language. The conclusions of this study can support research for the development of effective phishing detection tools for Romanian-speaking users and organizations.
Keywords: *phishing detection; machine learning; healthcare; Romania.*
JEL Classification: O310.

1. INTRODUCTION

Technologies are always and constantly advancing with the aim of improving some social or economic aspects. At the same time, the digitalization of society has created an increasing dependence on technology. Organizations in the health field, and not only, have greatly benefited from this technological advance. And they rely on a backbone of connected computing infrastructures and mobile medical devices that use patient-oriented technologies. In addition, healthcare professionals rely on electronic medical records, computer-controlled medical devices, complex medical imaging platforms, and a multitude of other tools to support the current standard of care.

Although technological advances are visible and tangible, new vulnerabilities and threats, both physical/psychological and cyber, are periodically identified with an impact in the health field. Thus, the last period was marked by a pandemic threat, which, in addition impose some social and professional changes, such as telework, movement or socialization restrictions, put a high pressure on the digital infrastructure. And these changes, combined with the geopolitical changes that generate a constant dose of mistrust, , have led to the creation of a continuous insecurity condition. The insecurity is the main factor to threats manifestation, some of which are even cybercrimes.

Against this background of insecurity, prevention, mitigation, emergency management and disaster recovery are important responsibilities in most fields, especially in health. And the high degree of dependence of the health system on technology presents a new and important challenge for doctors, experts and decision makers.

Healthcare organizations are currently considered some of the most vulnerable when it comes to cyber threats. According to a report by the Ponemon Institute (Ponemon Institute, 2022), the estimated average cost of a data breach in the healthcare sector is $4.82 million. Furthermore, a report by IBM X-Force (IBM, 2023) found that healthcare organizations were among the top targets for cyberattacks in the period 2020-2022, with a percentage that varied between 6.6% and 5.8%.

One of the main reasons of medical organizations vulnerabilities to cyberthreats is the large amount of sensitive information managed by these, such as patient medical records, which can be sold on the dark web at a high price (Davis, 2021). Additionally, they are often targeted by hackers as they may have weaker security systems than other industries and users may be more prone to phishing attacks.

The healthcare industry faces several challenges when it comes to cybersecurity. One of the main challenges is the complexity of the systems and infrastructure used by healthcare organizations. Systems often rely on a wide range of devices and technologies, including medical devices, electronic health records and mobile devices. This complexity makes it difficult to implement comprehensive cybersecurity measures (HIMSS, 2023).

Another challenge is the lack of cybersecurity professionals. In 2022, a cybersecurity workforce study (ISC2, 2022), found that there is currently a deficit of 3.4 million skilled cybersecurity professionals globally. And all this only turns the organizations in the mentioned field into a continuous target of cyberattacks. In fact, according to a study, the attacks, carried out in 2021, on organizations in the medical field in the USA led to the compromise of more than 40 million patient files (Jercich, 2021).

Practically, with technological development it has become almost impossible to eliminate security risks for modern information systems Given the many challenges posed by new ways of conducting cyberattacks, statistical data analysis, natural language processing and Machine Learning (ML) are increasingly used techniques in cyber security and data privacy challenges. (Ali *et al.,* 2020). Thus, identifying a way to recognize and obtain relevant information for integration into data sets that can be used in empirical analysis or probability theory (Makawana and Jhaveri, 2018) has become a major utility for cybersecurity systems.

The purpose of this study is to identify the optim ML model in the Supervised Learning category, that can be used to detect email messages in the phishing category. The choice to detect phishing attacks is given both by their constantly growing percentage and by the fact that it represents the most common method used for unauthorized access. The novelty of the solution is given by the fact that the dataset used will contain characteristics extracted from the subject and the text of the message, which will be in Romanian.

The methodological approach is quantitative/experimental by exposing the way to apply Machine Learning algorithms, presenting the activities performed and analysing the results.

In this sense, chapter 2 aims to expose the importance of detecting attacks in the phishing category, considering the role they play in supporting other types of attacks, respectively their effects.

In chapter 3, we will consider the identification of an optimal solution for the application of ML to detect phishing attacks through email messages, by applying a methodology that requires a combination of data collection and pre-processing, selection of features, respectively training and evaluation of some models.

2. CYBERATTACKS, THREATS THAT CONTINUE TO EXCEED VIRTUAL BOUNDARIES

Cyberattacks are those activities carried out by malicious individuals or groups, through the use of computer systems, which aim to disrupt, damage or gain unauthorized access to a target system or network. These can take many forms, including malware/ransomware infections, phishing scams, denial-of-service attacks, and more.

According to a report issued by ENISA, in 2022 (ENISA, 2022), the main threats identified at European level, to cyber security, were the following:

1. Ransomware – a threat in which certain situations are encrypted and an organization's data is exfiltrated, and a payment is requested to restore access.

2. Malware – threat represented by software or firmware designed specifically to perform unauthorized activities that negatively impact the confidentiality, integrity, or availability of a system.

3. Social Engineering threats – threats that exploit the weak points of the human psyche and everyday habits rather than the technical vulnerabilities of information systems.

4. Threats against data – threats that allow the exfiltration of important and protected data, being on an upward trend, because one of the main targets for attackers is to access sensitive data for negative reasons like ransom, defamation, extortion, disinformation, etc.

5. Threats against availability and integrity – Availability and integrity are the target of a multitude of threats and attacks, among which Denial of Service (DoS) and Web Attacks stand out.

6. Disinformation / Misinformation – Disinformation campaigns by spreading false or partially false information are on the rise. They are supported by the increased use of social media platforms and online media, as well as the increase in people's online presence. They are used in hybrid attacks to reduce the overall perception of trust, a major enabler of cybersecurity.

7. Supply-chain attacks – Threats representing a combination of at least two attacks. The first attack is on a digital service/product provider which is then used to attack the targeted target(s) in order to gain access to internal resources.

Additionally, the SOPHOS report for the year 2022 (Sophos, 2022) reinforces the fact that among the main threats remain Ransomware/Malware and e-mail related threats, but draws attention to the fact that there is an upward trend aimed at IoT and AI.

According to different studies in the area of cyber security, a cyberattack is carried out in several stages. Their number varies between 5 (Goedegebure, 2017), 6 (PaloAlto, 2023) and 7 (Lockheed Martin, 2023) and even 14 (Jackson, 2022). In this sense it can be generalized that an attack consists of the main important steps: Recognition, Scanning, Gaining Access, Maintaining Access and Concealing unauthorized presence.

An analysis of these stages reveals that they begin with the Recognition stage, during which the available information about the target system is collected. Later, the information obtained is used to initiate the actual attack in the next stage, where we encounter various types of attacks, the most prevalent being through phishing campaigns.

Phishing, a tool for exploiting human vulnerability

Phishing is a type of cyberattack in which hackers or cybercriminals try to trick people into revealing sensitive information, such as personal data, login credentials or financial information. Phishing attacks can be conducted via email, text messages, social media or other communication channels.

Phishing attacks often involve creating fake websites or login pages that mimic legitimate websites or services. For example, a phishing email may appear to come from a bank or social media platform and contain a link to a fake login page that looks like the real thing. When the victim enters their credentials, the information is sent to the attacker, who can then use it for malicious purposes (FTC, 2022).

Phishing attacks can also involve the use of social engineering techniques, such as creating a sense of urgency or fear in the victim, to increase the likelihood that they will divulge sensitive information (Porter, 2021). For example, a phishing email may claim that the victim's account has been compromised and that they must change their password immediately or risk losing access to their account.

Phishing has become one of the most common cyber threats overall, with 81% of organizations affected in 2022 (Jones, 2023), and healthcare organizations are no exception. Phishing can range from mass email campaigns designed to get recipients to reveal their passwords or access malicious applications delivered in the form of seemingly legitimate documents, to highly targeted campaigns designed to get invoice payments false.

A 2022 study (Ell and Gallucci, 2022) found that 26% of organizations experienced a "significant" increase in the number of email threats in 2021, and of these, 88% were victims of ransomware. Also, at least one business email has been compromised in 92% of organizations, and 93% have experienced data breaches due to negligence or compromised employee credentials. According to other studies, during the peak period of the COVID-19 pandemic, phishing attacks increased by 220% (Warburton, 2020) and as a result of the Russian-Ukrainian conflict, an intensification of phishing attacks targeting organizations was identified from NATO member countries (Huntley, 2023). Moreover, in 2022, 48.63% of all emails globally were spam, an increase of 3 percent compared to 2021, when only 45.56% of emails sent were spam (Kulikova *et al.,* 2023).

Traditional phishing detection approaches rely on predefined rules and heuristics. These approaches are effective, but they are not scalable and cannot detect new and complex attacks. Machine Learning (ML) has become an innovative way with high possibilities for detecting phishing attacks. ML models can learn from data and detect previously unseen attacks, making them an effective solution to the problem.

The key to successfully detecting phishing is to identify tentative attacks before they become real. ML models can identify tentative attacks based on features such as email sender, subject line, body content, and URL.

The key to successfully detecting a cyberattack is to identify tentative attacks like phishing before they become real. Machine Learning (ML) models can be successfully used to identify phishing attacks based on features such as email sender, subject line, body content, and URL.

3. ML-BASED SOLUTION FOR IDENTIFYING PHISHING EMAIL MESSAGES

Phishing detection can be approached using various ML techniques, including sentiment analysis, content analysis, and sender and URL analysis.

In general, detection of phishing emails can be done by two main types of methods: black/white list and ML. The first uses a predefined list of phishing or legitimate addresses that are compared to indicators in the received email, such as the sender's email address or IP address. Depending on the degree of data matching and the list used, the phishing email is rejected before reaching the email server (Khonji, Iraqi and Jones, 2013; Gupta, Arachchilage and Psannis, 2018). In general, the blacklisting method has a low false positive rate. However, the method depends on recipients reporting phishing emails (Fang *et al.,* 2019). On the same note, automatic whitelisting is also up to the user to create a collection of legitimate addresses. Whitelisting can be used to prevent phishing emails, but it is not effective enough to detect all phishing attacks due to the high percentage of false negatives (Jain and Gupta, 2019). Regardless of the list type, they do not provide security against zero-day attacks, as the details of new email addresses or URLs cannot be known.

Alternatively, ML can streamline the automated detection of phishing emails through various methods. In this sense, two methods for improving the classifier have been proposed in studies in the field (Toolan and Carthy, 2010):

1) testing and evaluating multiple ML models;
2) improving the classifier by focusing on feature selection from the dataset.

This article proposes a solution for detecting phishing e-mail messages targeting medical organizations in Romania. In the development of the solution, the Python programming language was used together with specific ML libraries (pandas, scikit-learn, numpy etc.) and the following stages were completed:

1. Data collection and preparation: In this step, the data is obtained and prepared for use by the ML models.
2. Feature extraction: This step consists in extracting the relevant features for running the ML models. During this stage, necessary actions such as dimensionality reduction or feature scaling are carried out
3. Model selection and training: Involves choosing an ML model and applying it to previously extracted features.
4. Model evaluation: After the training stage, the model is tested on a separate data set and evaluated in terms of performance.

3.1. Data collection and preparation

Currently, for the training and evaluation of ML models, datasets containing features extracted from email messages from the phishing category are available. But the messages come from various sectors of activity and are in English, which is why we considered the possibility that the phishing messages aimed at the health field in Romania present certain specific characteristics. Thus, 292 raw phishing e-mails collected between 2020 and 2023 from 3 healthcare organizations were processed in a dataset. These represent only the messages that were not detected by the primary security solutions implemented by the respective institutions, in their own email servers. For the legitimate part of the data set, 208 messages collected from the 3 health organizations in Romania were used. The data were made available for research purposes only, not being public.

Therefore, in total, the dataset contains 440 emails, and of these, 47% are legitimate. To go through all the email files of type ".eml" and extract the features, the Python library, "email" was used. From all the available data, only those containing the subject and the message were extracted. Thus, those that contained details on email addresses or details on attached files were discarded because they are not relevant for the following reasons:

- An email address may contain truthful details as a sign of the existence of spoofing tools.
- If the same email address is used, it will be blacklisted at some point.
- The details of the attached files are relevant only by knowing the hash, an aspect that would require downloading it and implicitly the possibility of system vulnerability.

The dataset consists of features grouped into a text variable and represents word lists specific to each message and a label variable, for Supervised Learning models, indicating whether the email is phishing („Phish") or legitimate („NonPhish").

3.2. Features extraction

The features represent a list of the most frequent terms, extracted from phishing and legitimate emails, the selection of which has been widely applied in the data-mining literature. A series of steps were followed to construct the phishing dataset, described below.

The first step consisted of grouping the messages into two categories: phishing and legitimate. Each category was parsed and only the information contained in the subject and body of the email was extracted. Later, from the extracted data, tags and html elements, URLs, were removed, because they can be used by other phishing detection models that consider their examination, respectively "stop words". "Stop words" represent a list of words that are filtered out before or after processing natural language (text) data because they are meaningless (Rajaraman and Ullman, 2011). For this action, the Python library, NLTK's stop words corpus, was used, because it contains "stop words" in Romanian as well.

Finally, we used TF-IDF (Term Frequency Inverse Document Frequency) from scikit-learn library for Python. This tool finds the most frequent terms that appear in the corpus. It calculates the number of times a word appears in a document multiplied by a (monotonic) function of the inverse of the number of documents in which the word appears. A higher weight is given to the terms that appear often in a document and do not appear in many documents (Berry, 2004).

3.3. Model selection and training

After preparing the data set, the following algorithms from the Supervised Learning category were applied to it: Logistic Regression, Multinomial Naive Bayes, Random Forest, Decision Tree Classifier, Support Vector Classifier, K-Neighbors Classifier, MLP Classifier, Neural Networks, Gradient Boosting and Add Boost.

a) *Logistic Regression*: It is a linear model used for binary classification tasks. It works by estimating the probability of an email being phishing based on a set of input features. It is simple to implement, interpretable, and can handle large datasets (Cunningham, Cord and Delany, 2008).

b) *Multinomial Naive Bayes*: Represents a probabilistic model, used often for text classification tasks. It works by estimating the probability of an email being phishing or legitimate based on the occurrence of words or phrases in the email. It is simple to implement, computationally efficient, and can handle large datasets (Cunningham, Cord and Delany, 2008).

c) *Random Forest*: It is an ensemble model used for classification and regression tasks. It builds a large number of decision trees and combine their predictions to obtain a more accurate and robust final prediction. It is less prone to overfitting than decision trees and can handle high-dimensional data (Nasteski, 2017).

d) *Decision Tree Classifier*: It is a tree-based model that is used for classification tasks. It works by partitioning the feature space into smaller regions based on the input features and making decisions based on the majority class in each region. It is interpretable and can handle both categorical and numerical data (Cunningham, Cord and Delany, 2008).

e) *Support Vector Classifier*: It is a model that is used for binary classification tasks. It works by finding a hyperplane that maximally separates the two classes in the feature space. It is effective in high-dimensional spaces and can handle nonlinear decision boundaries (Nasteski, 2017).

f) *K-Neighbors Classifier*: It is a lazy learning model that is used for classification tasks. It works by finding the k nearest neighbors to a new data point in the feature space and making a prediction based on the majority class of the neighbors. It is simple to implement and can handle both categorical and numerical data (Nasteski, 2017).

g) *MLP Classifier*: It is a neural network model that is used for classification tasks. It works by learning a nonlinear function that maps the input features to the output classes. It is effective in high-dimensional spaces and can handle nonlinear decision boundaries (Windeatt, 2008).

h) *Neural Networks*: It is a deep learning model that is used for both classification and regression tasks. It works by learning a hierarchy of nonlinear features from the input data and making predictions based on the learned features. It is effective in high-dimensional spaces and can handle complex data (Freeman and Skapura, 1991).

i) *Gradient Boosting*: It is an ensemble model that is used for classification and regression tasks. It works by building an ensemble of weak prediction models and combining their predictions to form a more accurate and robust final prediction. It is less prone to overfitting than other ensemble models and can handle high-dimensional data.

j) *Ada Boost*: It is an ensemble model like Gradient Boosting and works by building an ensemble of weak prediction models and adjusting their weights to focus on misclassified examples. It is less prone to overfitting than other ensemble models and can handle complex data (Bahad and Saxena, 2020).

The algorithms were applied on 5 randomly generated training sets. The results are presented in (Table 1) together with the value of the confusion matrix indicators (TN-True Negative; FP-False Positive; FN- False Negative; TP- True Positive;). It should be noted that the results are related to the confusion matrix indicators and represent the following:

1. *Precision*: This metric measures the proportion of true positive predictions out of all positive predictions. It is useful when the cost of false positives is high. The formula for calculating Precision is: TP / (TP + FP).

2. *Recall*: This metric measures the proportion of true positive predictions out of all actual positives. It is useful when the cost of false negatives is high. The formula for calculating Recall is: TP / (TP + FN).

3. *F1 Score*: This is the harmonic mean of precision and recall, and provides a balanced measure between the two metrics. The formula for calculating F1 Score is: 2 * (Precision * Recall) / (Precision + Recall).

4. *Accuracy*: This metric measures the proportion of correctly classified instances out of all instances. It is a simple and commonly used metric, but can be misleading if the data is imbalanced. The formula for calculating Accuracy is: (TP + TN) / (TP + TN + FP + FN).

Table 1. The results of applying ML models

Model	Training set	Precision	Recall	F1-score	TN	FP	FN	TP	Accuracy
Logistic Regression	1	0.89	0.78	0.83	63	6	14	49	0.8485
	2	0.81	0.92	0.86	55	14	5	58	0.8561
	3	0.77	0.95	0.85	51	18	3	60	0.8409
	4	0.75	0.95	0.84	49	20	3	60	0.8258

Model	Training set	Precision	Recall	F1-score	TN	FP	FN	TP	Accuracy
	5	0.89	0.78	0.83	63	6	14	49	0.8485
Multinomial Naive Bayes	1	0.86	0.87	0.87	60	9	8	55	0.8712
	2	0.78	0.95	0.86	52	17	3	60	0.8485
	3	0.8	0.95	0.87	54	15	3	60	0.8636
	4	0.73	0.95	0.83	47	22	3	60	0.8106
	5	0.86	0.87	0.87	60	9	8	55	0.8712
Random Forest	1	0.89	0.81	0.85	63	6	12	51	0.8636
	2	0.81	0.95	0.88	55	14	3	60	0.8712
	3	0.76	0.97	0.85	50	19	2	61	0.8409
	4	0.74	0.95	0.83	48	21	3	60	0.8182
	5	0.88	0.79	0.83	62	7	13	50	0.8485
Decision Tree Classifier	1	0.75	0.79	0.77	52	17	13	50	0.7727
	2	0.82	0.78	0.8	58	11	14	49	0.8106
	3	0.86	0.79	0.83	61	8	13	50	0.8409
	4	0.82	0.79	0.81	58	11	13	50	0.8182
	5	0.76	0.89	0.82	51	18	7	56	0.8106
Support Vector Classifier	1	0.91	0.79	0.85	64	5	13	50	0.8636
	2	0.88	0.89	0.88	61	8	7	56	0.8864
	3	0.77	0.95	0.85	51	18	3	60	0.8409
	4	0.77	0.94	0.84	51	18	4	59	0.8333
	5	0.94	0.81	0.87	66	3	12	51	0.8864
K-Neighbors Classifier	1	0.83	0.87	0.85	58	11	8	55	0.8561
	2	0.78	0.89	0.83	53	16	7	56	0.8258
	3	0.79	0.84	0.82	55	14	10	53	0.8182
	4	0.78	0.89	0.83	53	16	7	56	0.8258
	5	0.82	0.89	0.85	57	12	7	56	0.8561
MLP Classifier	1	0.86	0.87	0.87	60	9	8	55	0.8712
	2	0.83	0.87	0.85	58	11	8	55	0.8561
	3	0.84	0.89	0.86	58	11	7	56	0.8636
	4	0.8	0.89	0.84	55	14	7	56	0.8409
	5	0.87	0.84	0.85	61	8	10	53	0.8636

Model	Training set	Precision	Recall	F1-score	TN	FP	FN	TP	Accuracy
Neural Networks	1	0.87	0.87	0.87	61	8	8	55	0.8788
	2	0.82	0.87	0.85	57	12	8	55	0.8485
	3	0.81	0.9	0.86	56	13	6	57	0.8561
	4	0.81	0.89	0.85	56	13	7	56	0.8485
	5	0.87	0.87	0.87	61	8	8	55	0.8788
Gradient Boosting	1	0.82	0.81	0.82	58	11	12	51	0.8258
	2	0.87	0.83	0.85	61	8	11	52	0.8561
	3	0.87	0.87	0.87	61	8	8	55	0.8788
	4	0.84	0.83	0.83	59	10	11	52	0.8409
	5	0.83	0.83	0.83	58	11	11	52	0.8333
Ada Boosting	1	0.88	0.83	0.85	62	7	11	52	0.8636
	2	0.8	0.83	0.81	56	13	11	52	0.8182
	3	0.84	0.81	0.82	59	10	12	51	0.8333
	4	0.85	0.81	0.83	60	9	12	51	0.8409
	5	0.92	0.86	0.89	64	5	9	54	0.8939

Source: self-representation

3.4. Model evaluation

For the evaluation of the effectiveness of a classifier are commonly used two methods: the *holdout* method and the *k-fold cross-validation* method (Rithchie, 2018).

In a holdout split, the classifier is evaluated by computing the error rate. The error rate is calculated by comparing the predicted labels to the true labels in the validation set. The error rate is typically defined as the number of incorrect predictions divided by the total number of predictions.

In the k-fold cross-validation method the data is split into k-folds. Each fold is a subset of the data that will be used as the validation set for one iteration of the cross-validation process. The classifier is evaluated by the average performance of model across all k-folds

In this study were used both methods, for the evaluation. For the tested models, 70% of the data was used for training (308 rows), while 30% for testing (132 rows).

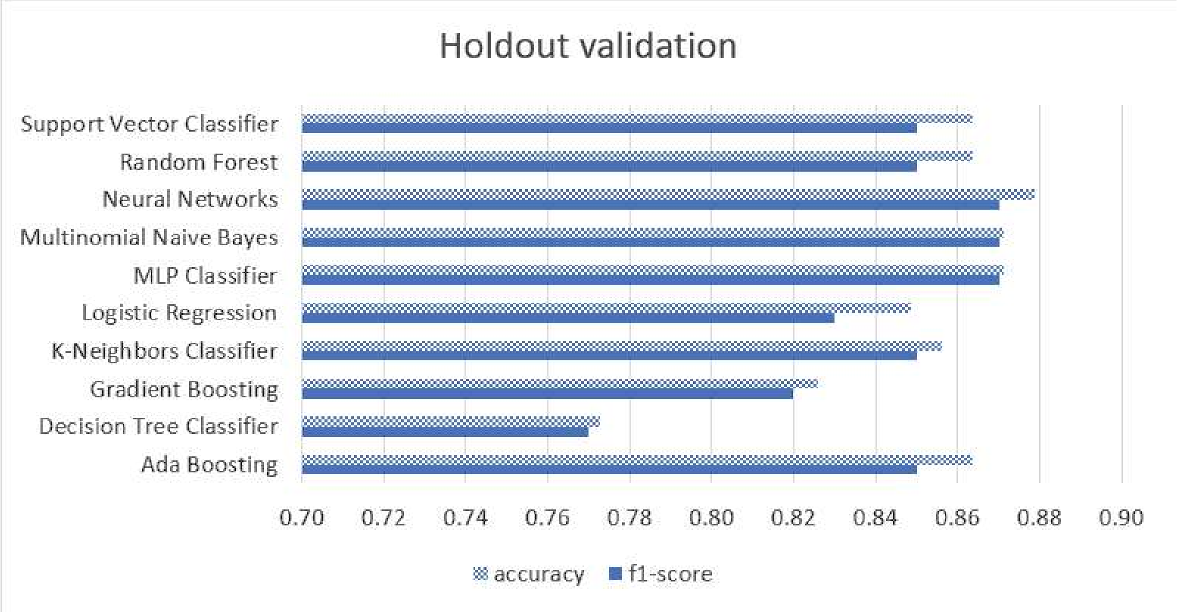
Moreover, the cross-validation method was used, dividing the data set into 5 groups. The model is trained 5 times, with different test set, the results obtained for each set being those in (Table 1).

In the case of the evaluation by the holdout method, the results obtained are those from (Table 2), the graphic representation being in (Figure 1). According to them, the highest percentage of accuracy and implicit detection of phishing email messages is obtained by Neural Networks (87.88%), followed by MLP Classifier and Multinomial Naïve Bayes, both with a percentage of 87.12%. Also, as expected, Neural Networks has lower error rate (12.12%)

Table 2. Holdout evaluation

Model	Precision	Recall	F1-score	Accuracy	Err rate
Ada Boosting	0.8800	0.8300	0.8500	0.8636	0.1364
Decision Tree Classifier	0.7500	0.7900	0.7700	0.7727	0.2273
Gradient Boosting	0.8200	0.8100	0.8200	0.8258	0.1742
K-Neighbors Classifier	0.8300	0.8700	0.8500	0.8561	0.1439
Logistic Regression	0.8900	0.7800	0.8300	0.8485	0.1515
MLP Classifier	0.8600	0.8700	0.8700	0.8712	0.1288
Multinomial Naive Bayes	0.8600	0.8700	0.8700	0.8712	0.1288
Neural Networks	0.8700	0.8700	0.8700	0.8788	0.1212
Random Forest	0.8900	0.8100	0.8500	0.8636	0.1364
Support Vector Classifier	0.9100	0.7900	0.8500	0.8636	0.1364

Source: self-representation



Source: self-representation

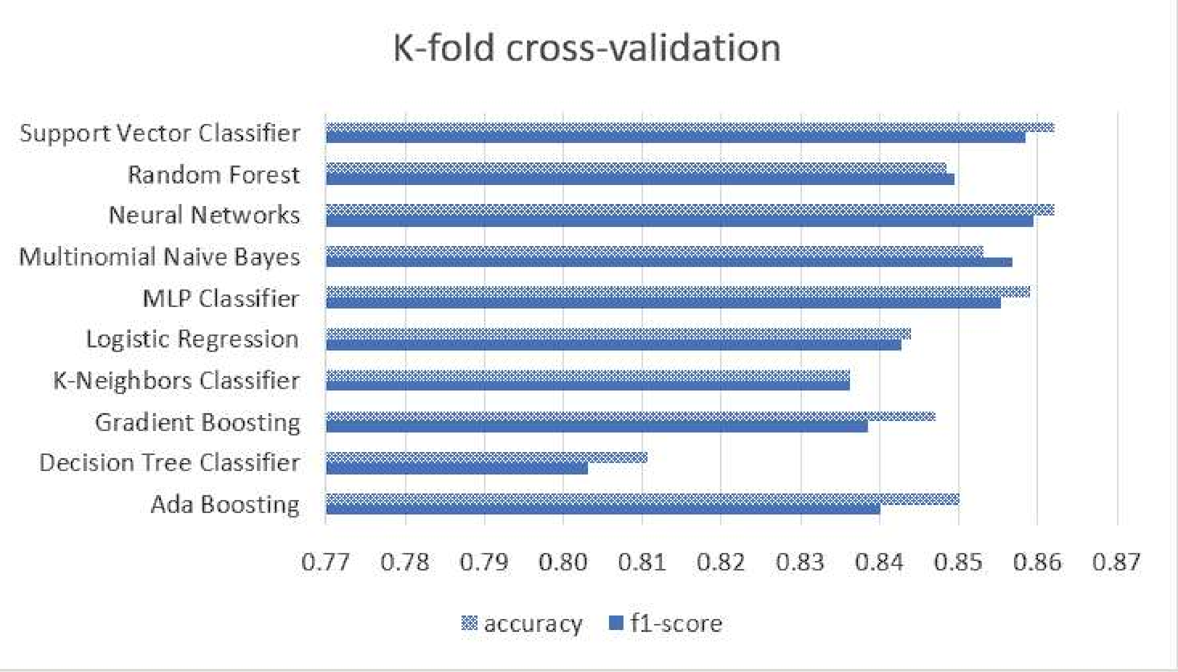
Figure 1. Graphical representation of accuracy and f1-score for holdout evaluation

In the case of the evaluation by the k-fold method, the results obtained are those in (Table 3), the graphic representation being in (Figure 2). According to them, the highest percentage of accuracy and implicit detection of phishing email messages is obtained by Suport Vector Classifier and Neural Networks, both with 86.21%.

Table 3. K-fold cross validation evaluation

Model:	Precision	Recall	F1-score	Accuracy
Ada Boosting	0.8553	0.8254	0.8401	0.8500
Decision Tree Classifier	0.7969	0.8095	0.8031	0.8106
Gradient Boosting	0.8452	0.8317	0.8384	0.8470
K-Neighbors Classifier	0.8000	0.8762	0.8364	0.8364
Logistic Regression	0.8118	0.8762	0.8427	0.8439
MLP Classifier	0.8384	0.8730	0.8554	0.8591
Multinomial Naive Bayes	0.8011	0.9206	0.8567	0.8530
Neural Networks	0.8373	0.8825	0.8594	0.8621
Random Forest	0.8080	0.8952	0.8494	0.8485
Support Vector Classifier	0.8415	0.8762	0.8585	0.8621

Source: self-representation



Source: self-representation

Figure 2. Graphical representation of accuracy and f1-score for k-fold cross-validation evaluation

Based on achieved results, the Neural Networks model can be proposed, as a possible solution for detecting phishing e-mail messages written in Romanian and aimed at health organizations. However, the accuracy percentage obtained, below 90%, reveals the need to continue the study by extracting other characteristics, namely the identification of terms specific to this type of message, in order to increase the robustness of the cyber security solution.

4. CONCLUSIONS

Email phishing attacks are one of the fastest growing cybercrimes, targeting both organizations and individuals. Also, according to estimates, annual losses are in the order of billions of dollars. The methods by which these attacks are carried out change rapidly, also depending on the level of knowledge of the attacker. Thus, to compete with human intelligence, a perfect or at least perfectible solution is given by the application of ML models.

The main goal of this study is to propose a possible solution for increasing the level of performance and accuracy of the classification and detection of phishing e-mails written in Romanian. Within it, the results of the classification models for the detection of phishing e-mails in Romanian were examined, using natural language processing tools and supervised learning models.

The solution highlights the importance of examining text features in the email message as it represents a new research direction in email phishing detection.

The neural networks model that achieved the highest percentage of accuracy reveals that in the analysis of texts should be approached models from the category artificial neural networks. Also, given that the research is an early one, it will be necessary to approach some models of deep learning to identify phishing attacks in the text of the messages written in Romanian. There is a possibility that this research is among the first, to our knowledge, that examined and compared several models for detecting phishing in an e-mail with content in Romanian. This will lead to more investigations into detecting phishing in text, whether in emails, social media messages or even malicious websites.

Moreover, the proposed solution is an early one and can be improved by diversifying the features, in order to increase the accuracy for detecting phishing attacks. Thus, the aim is to increase the efficiency in detecting unknown or zero-day attacks.

Future research directions will aim to improve the solution by applying the models to different types of characteristics extracted from the message text, respectively the identification of hybrid models with a high degree of accuracy in identifying this type of attack.

References

1) Ali, R., Ali, A., Iqbal, F., Khattak, A. M. and Aleem, S. (2020). A Systematic Review of Artificial In-telli-gence and Machine Learning Techniques for Cyber Security. *Communications in Computer and Information Science*. 1210 CCIS. pp. 584-593. https://doi.org/10.1007/978-981-15-7530-3_44

2) Bahad, P. and Saxena, P. (2020). Study of adaboost and gradient boosting algorithms for predictive analytics. In: *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019* (pp. 235-244). Singapore: Springer.

3) Berry, M. W. (2004). *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer.

4) Cunningham, P., Cord, M. and Delany, S. J. (2008). Supervised Learning. In: Cord, M. and Cunningham, P., eds., *Machine Learning Techniques for Multimedia. Cognitive Technologies*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-75171-7_2.

5) Davis, J. (2021). Dark Web Analysis: Healthcare Risks Tied to Database Leaks, Credentials. *Health IT Security*. [online] Avaibale at: https:// healthitsecurity.com/ news/dark-web-analysis-healthcare-risks-tied-to-database-leaks-credentials [Accessed 01.04-13.04.2023].

6) Ell, M. and Gallucci, R. (2022). *Cyber Security Breaches Survey 2022*. [online] Available at: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2022/cyber-security-breaches-survey-2022 [Accessed 01.04-13.04.2023].

7) ENISA (2022). *Threat Landscape – report 2022*. [online] Available at: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022?v2=1 [Accessed 01.04-13.04.2023].

8) Fang, Y., Zhang, C., Huang, C., Liu, L. and Yang, Y. (2019). Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access*, 7, 56329-56340. doi:10.1109/ACCESS.2019.2913705.

9) Freeman, J. A. and Skapura, D. M. (1991). *Neural networks: algorithms, applications, and programming techniques.* Addison Wesley Longman Publishing Co., Inc.

10) FTC, Federal Trade Commission (2022). *How to Recognize and Avoid Phishing Scams.* [online] Available at: https://consumer.ftc.gov/articles/how-recognize-and-avoid-phishing-scams [Accessed 01.04-13.04.2023].

11) Goedegebure, C. (2017). *5 Phases of hacking*. [online] Available at: https://www.coengoedegebure.com/5-phases-of-hacking/ [Accessed 01.04-13.04.2023].

12) Gupta, B. B., Arachchilage, N. A. G. and Psannis, K. E. (2018). Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication System*, 67, pp. 247–267. https://doi.org/10.1007/s11235-017-0334-z.

13) HIMSS (2023). Healthcare Information and Management Systems Society (HIMSS). *Cybersecurity in Healthcare*. [online] Available at: https://www. himss.org/resources/cybersecurity-healthcare [Accessed 01.04-13.04.2023].

14) Huntley, S. (2023). *Fog of war: how the Ukraine conflict transformed the cyber threat landscape*. [online] Available at: https://services.google.com/fh/ files/blogs/google_fog_of_war_research_report.pdf [Accessed 01.04-13.04.2023].

15) IBM (2023). *X-Force Threat Intelligence Index 2023*. [online] Available at: https://www.ibm.com/reports/threat-intelligence [Accessed 01.04-13.04.2023].

16) ISC2 (2022). *Cybersecurity Workforce Study*. [online] Available at: https://www.isc2.org/Research/Workforce-Study [Accessed 01.04-13.04.2023].

17) Jackson, C. (2022). *What is a cyber attack? The 14 stages of a cyber attack*. [online] Available at: https://acloudguru.com/blog/engineering/what-is-a-cyber-attack-the-14-stages-of-a-cyber-attack [Accessed 01.04-13.04.2023].

18) Jain, A. K. and Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing,* 10(5), pp. 2015–2028. doi:10.1007/s12652-018-0798-z.

19) Jercich, K. (2021). *The biggest healthcare data breaches of 2021*. [online] Available at: https://www.healthcareitnews.com/news/biggest-healthcare-data-breaches-2021 [Accessed 01.04-13.04.2023].

20) Jones, C. (2023). *50 Phishing Stats You Should Know In 2023*. [online] Available at: https://expertinsights.com/insights/50-phishing-stats-you-should-know/ [Accessed 01.04-13.04.2023].

21) Khonji, M., Iraqi, Y. and Jones, A. (2013). Phishing detection: a literature survey. *EEE Communications Surveys & Tutorials,* 15(4), pp. 2091-2121, Fourth Quarter 2013, doi: 10.1109/SURV.2013.032213.00009.

22) Kulikova, T., Dedenok, R., Svistunova, O., Kovtun, A. and Shimko, I. (2023). *Spam and phishing in 2022*. [online] Available at: https://securelist.com/spam-phishing-scam-report-2022/108692/ [Accessed 01.04-13.04.2023].

23) Lockheed Martin (2023). *The Cyber Kill Chain: A Lockheed Martin Overview*. [online] Available at: https://www.lockheedmartin.com/en-us/capabilities/ cyber/ cyber-kill-chain.html [Accessed 01.04-13.04.2023].

24) Makawana, P. R. and Jhaveri, R. H. (2018). A Bibliometric Analysis of Recent Research on Machine Learning for Cyber Security. In: Hu, YC., Tiwari, S., Mishra, K. and Trivedi, M., eds., *Intelligent Communication and Computational Technologies*. *Lecture Notes in Networks and Systems*, 19, pp. 213-226. Singapore: Springer. https://doi.org/10.1007/978-981-10-5523-2_20.

25) Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons*, 4, pp. 51-62. DOI: 10.20544/HORIZONS.B.04.1.17.P05.

26) PaloAlto (2023). *How to Break the Cyber Attack Lifecycle.* [online] Available at: https://www.paloaltonetworks.com/cyberpedia/how-to-break-the-cyber-attack-lifecycle [Accessed 01.04-13.04.2023].

27) Ponemon Institute (2022). *Cost of a Data Breach Report 2022*. [online] Available at:https://www.ibm.com/reports/data-breach [Accessed 01.04-13.04.2023].

28) Porter, K. (2021). *What is phishing? How to recognize and avoid phishing scams*. [online] Available at: https://us.norton.com/blog/online-scams/what-is-phishing [Accessed 01.04-13.04.2023].

29) Rajaraman, A. and Ullman, J. (2011). Data Mining. In: *Mining of Massive Datasets*, pp. 1-17. Cambridge: Cambridge University Press. doi:10.1017/ CBO9781139058452.002.

30) Rithchie, N. (2018). *Evaluating a Classification Model*. [online] Available at: https://www.ritchieng.com/machine-learning-evaluate-classification-model/ [Accessed 01.04-13.04.2023].

31) Sophos (2022). *2022 Threat Report, Interrelated threats target an interdependent world*. [online] Available at: https://assets.sophos.com/X24WTUEQ/at/ b739xqx5jg5w9w7p2bpzxg/sophos-2022-threat-report.pdf [Accessed 01.04-13.04.2023].

32) Toolan, F. and Carthy, J. (2010). Feature selection for Spam and Phishing detection. *2010 eCrime Researchers Summit*, Dallas, TX, USA, 2010, pp. 1-12. Doi: 10.1109/ecrime.2010.5706696.

33) Warburton, D. (2020). *F5 Labs 2020 Phishing and Fraud Report*. [online] Available at: https://www.f5.com/labs/articles/threat-intelligence/2020-phishing-and-fraud-report [Accessed 01.04-13.04.2023].

34) Windeatt, T. (2008). Ensemble MLP Classifier Design. In: Jain, L.C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G. A., Balas, V. E. and Abeynayake, C., eds, *Computational Intelligence Paradigms. Studies in Computational Intelligence*, 137. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79474-5_6.