https://doi.org/10.2352/EI.2026.38.3.MOBMU-330

Multilingual Email Phishing Attacks Detection using Open-Source Intelligence and Machine Learning

Panharith An[1,2], *Rana Shafi*[1,2], *Tionge Mughogho*[1], *Onyango Allan Onyango*[1,2], *Nikola Nachevski*[1,2], *Reiner Creutzburg*[2,3,4]

[1] *Kadir Has University, Department of Administrative Sciences, 34083 Cibali/Fatih, İstanbul, Turkey*
[2] *SRH University of Applied Sciences, Sonnenallee 221, Berlin, Germany*
[3] *German University of Digital Science; Marlene-Dietrich Allee 14, D-14772 Potsdam, Germany*
[4] *Technische Hochschule Brandenburg, Dept. of Informatics & Media, Magdeburger Str. 50, D-14770 Brandenburg, Germany*
Email: mr.anpanharith@gmail.com, newranashafi@gmail.com, tionge362@gmail.com, onyango.allan@stu.khas.edu.tr, niknacevski@gmail.com, reiner.creutzburg@gmail.com, reiner.creutzburg@german-uds.de, reiner.creutzburg@srh.de, creutzburg@th-brandenburg.de
Keywords: Email Phishing, Open-source Intelligence, OSINT, Multilingual, Machine Learning, Classification Algorithms, Artificial Intelligence

Abstract

Email phishing remains a prevalent cyber threat, targeting victims to extract sensitive information or deploy malicious software. This paper explores the integration of open-source intelligence (OSINT) tools and machine learning (ML) models to enhance phishing detection across multilingual datasets. Using Nmap and theHarvester, this study extracted 17 features, including domain names, IP addresses, and open ports, to improve detection accuracy. Multilingual email datasets, including English and Arabic, were analyzed to address limitations of ML models trained predominantly on English-language data. Experiments with five classification algorithms: Decision Tree, Random Forest, Support Vector Machine, XGBoost, and Multinomial Naïve Bayes. It revealed that Random Forest achieved the highest performance, with accuracies of 97.37% on both the English and Arabic datasets. For OSINT-enhanced datasets, the model achieved higher accuracy than baseline models without OSINT features. These findings highlight the potential of combining OSINT tools with advanced ML models to detect phishing emails more effectively across diverse languages and contexts. This study contributes an approach to phishing detection by incorporating OSINT features and evaluating their impact on multilingual datasets, addressing a critical gap in cybersecurity research.

Introduction

Email Phishing is a malicious practice in which perpetrators send fraudulent messages to victims, inducing them to disclose private information or execute malicious software. Phishing has become one of the most common social engineering attacks, targeting users' emails to fraudulently steal confidential and sensitive information. Phishing attacks have grown to be one of the leading paths to bigger attacks, such as ransomware attacks and denial-of-service attacks [1]. Following the COVID-19 pandemic, the world has witnessed an increase in online activities, including online shopping, online banking, and digital banking [2]. This has undoubtedly increased the attack surface and vectors for these phishing attacks [3].

The issue of multilingual phishing canalso not be overlooked, as most ML models rely on phishing indicators in English, thereby limiting their detectability of phishing attacks in other languages. [4] identified eleven parameters deemed essential for effective fraud filters in their investigation of email phishing detection. However, gaps remain in addressing challenges such as the ability to detect phishing emails in multiple languages, the lack of generalizable models that adapt across different-language datasets, and OSINT-enhanced features to improve model accuracy. Open-source Intelligence (OSINT) is a branch of intelligence that gathers and analyzes information exclusively from freely and publicly accessible sources [5].

To address the gaps, this research leverages these parameters to include indicators across multiple languages. As a result, this study primarily focuses on applying machine learning techniques to develop models that detect phishing attacks while accounting for cross-language gaps.

In a multilingual context, this study aims to determine whether models trained on OSINT-enhanced features are more likely to outperform models trained on the original dataset, which has fewer features.

Key contributions of this paper are:

- A comparison study among four different datasets and five different machine learning classification algorithms to detect phishing emails in both English and Arabic.
- A study comparing the performance of ML models on datasets with and without OSINT-driven feature-selection to evaluate its impact on improving detection accuracy.

Research Questions

- What OSINT tools can be used to extract features from our English and Arabic datasets?
- Which ML models can be used to train the datasets in detecting phishing emails?
- How can OSINT-enhanced datasets in both languages be more accurately detected by the same algorithm compared to the original dataset?

Literature Review

Search Process

This study focuses on two distinct topics: OSINT and Machine Learning applied to email phishing detection. For better understanding, a Systematic Literature Review (SLR) was adopted across four databases: Scopus, Web of Science, IEEE Xplore, and Google Scholar. Query strings:

- ("Machine Learning" OR "Artificial Intelligence" OR "ML" OR "AI") AND ("Phishing" OR "Phishing Attacks" OR "Cybersecurity" OR "Cyber Security")
- ("Machine Learning" OR "Artificial Intelligence" OR "ML" OR "AI") AND ("OSINT" OR "Open-source Intelligence" OR "Open source intelligence")
- ("OSINT" OR "Open-source Intelligence" OR "Open source intelligence") AND ("Phishing" OR "Phishing Attacks" OR "Cybersecurity" OR "Cyber Security")

Figure 1 shows the process of our search strategies and results.

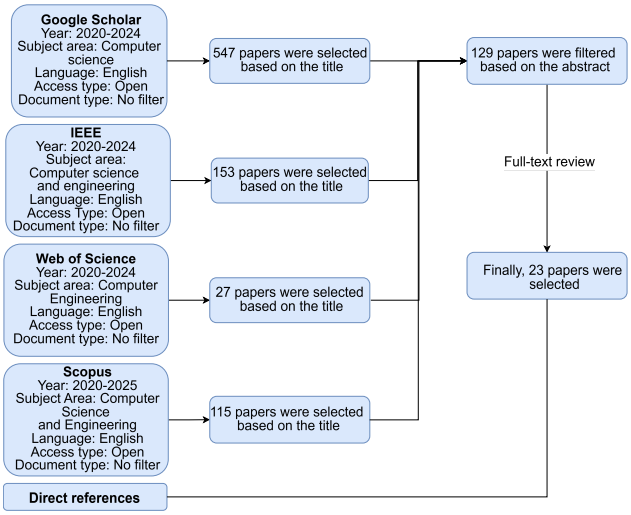


Figure 1. *Search Process Diagram in SLR.*

Table 1 summarizes related works in the domains of OSINT investigation, phishing attacks, and machine learning. While these studies provide valuable insights and advancements in phishing detection, significant gaps remain. These include limitations in multilingual phishing detection, inadequate integration of OSINT features, and challenges in optimizing feature extraction. This study seeks to address these gaps by combining OSINT tools with advanced ML algorithms to improve phishing-detection accuracy across diverse languages and contexts.

Research Method

Data Collection

This study will experiment with different groups of datasets. Also, the source code and datasets are available on our GitHub repository: `https://github.com/panharithan/osint-phishing`. Figure 2 depicts how each group of the dataset was chosen or generated.

- Group 1 (English): English phishing emails from Kaggle [6]

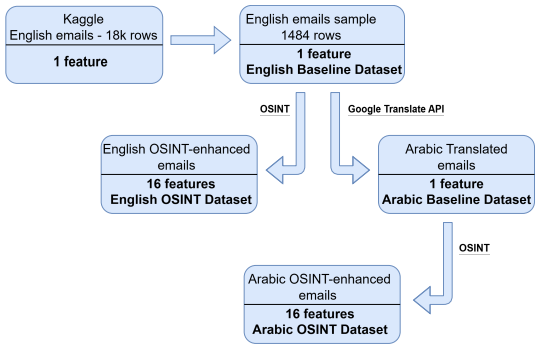


Figure 2. *Choosing and generating the dataset groups.*

- Group 2 (English Baseline): randomly sampling 1,484 rows from group 1 that contain URLs/domains.
- Group 3 (English OSINT): obtained by adding 17 OSINT features to group 2.
- Group 4 (Arabic Baseline): obtained using Google Translate API on Group 2, and verified by a native Arabic speaker (over 90% correctness)
- Group 5 (Arabic OSINT): obtained by adding 17 OSINT features to Group 4.

Choosing the Languages

Arabic was chosen as a second language for this multilingual study because it differs substantially from English, which uses the Roman alphabet. This study aims to compare the performance of ML models across two distinct languages.

The English dataset, published in 2023, originally contained one feature (Email Text) and one label (Email Type: Phishing or Safe). The number of rows was originally 18,651. However, after removing duplicates using SHA-256 hashing, the dataset was reduced to 17,522 non-duplicates (6,544 records are of the phishing type). Given its diversity and size, this 2023-published dataset has also been studied by other researchers, including [7], [8], and [9], highlighting its usefulness and contributions to the scientific community.

This study randomly sampled 1,484 rows from the aforementioned English dataset, which contained valid URLs, domain names, and email addresses. We named these samples the English Baseline dataset, from which we extracted OSINT features and generated the English OSINT dataset.

It is also important to note that, despite the small size of the OSINT dataset groups, this study is based on the principles of random sampling, ensuring that each data point has an equal chance of being selected, thereby preserving the dataset's representativeness. This approach minimizes selection bias and preserves the statistical integrity of the sample, making the findings applicable to the broader dataset of 17,000 rows.

Additionally, random sampling is a widely accepted method in research for managing large datasets effectively while maintaining reliable and valid results.

Afterward, this study translated the English Baseline dataset into Arabic, yielding our two final datasets: the Arabic Baseline dataset and the Arabic OSINT dataset.

Summary of Literature Reviews and Related Works.

Reference	Method	Dataset	Descriptions	Limitations
A Feature Extraction Approach for the Detection of Phishing Websites Using Machine Learning [10].	DT, RF, Multilayer perceptrons, XGBoost, SVM, KNN, and NB.	Phishing URLs. The source is not mentioned.	ML techniques were analyzed and implemented to detect phishing attacks with less time complexity, accurately. The MLPs technique provided the best precision score and the highest NDCG score.	The feature extraction algorithms needed to be more optimized, and new features needed to be chosen carefully to increase accuracy. The dataset was tested with basic classification algorithms, which yielded lower accuracy compared to other algorithms.
Phishing Attacks Facilitated by Open-Source Intelligence [11].	OSINT (Maltego)	Gathering public email addresses from social media platforms.	Phishing attacks were launched using OSINT to assess individuals' responses to spoofed emails. 5 of the 20 participants fell victim to phishing attacks.	The sample size of 20 participants, which may potentially be non-diverse, could affect the generalizability of the findings and fail to account for variability in user behavior across different demographics.
Multi-Language Spam/Phishing Classification by Email Body Text: Toward Automated Security Incident Investigation [12].	SVM, RF, DT, NB, LR, and KNN.	Nazario (phishing emails), SpamAssassin (spam emails), and a combined dataset from Vilnius Gediminas Technical University.	Multiple ML models were evaluated for their accuracy in identifying multilingual phishing emails. The support vector machine achieved the highest accuracy (84.0% ± 1.6%); however, it was among the slowest solutions.	Deeper spam/phishing email classification performance analysis could be executed to increase performance by adapting feature optimization (including header and formatting-related features, etc.) and evaluating deep-learning solution suitability for this task.
A Comparison of Machine Learning Algorithms for Multilingual Phishing Detection [13].	XML Roberta, LR, SVM, RF, GPT-4, and GPT-3.	English, French, and Russian spam emails from Enron.	An accuracy comparison of ML models in detecting multilingual phishing emails. XLM-Roberta achieves the highest accuracy among the tested models.	The French and Russian datasets used were translated from English; therefore, they may not fully capture linguistic nuances and cultural context, potentially leading to false positives or negatives.
Automation of the Information Collection Process by OSINT Methods for Penetration Testing During Information Security Audit [14].	Programmatic data collection with OSINT.	Extracting emails, phone numbers, organization addresses, etc., from the Internet.	The use of OSINT methods to automate data collection from open sources for penetration testing.	Software developed does not describe the level of sub-pages it could dive into. No clear relation between software development and pen-testing or information security audit exists.
Machine Learning Techniques for Detecting Phishing URL Attacks [15].	Neural Networks (NN), Naïve Bayes, and Adaboost.	Over 11k Websites from Kaggle.	The confusion matrices of three ML models in detecting phishing URL attacks were studied, and their accuracies were observed. Neural Networks (NN), Naïve Bayes, and Adaboost were studied, and the results indicated that the accuracies achieved were 90.23%, 92.97%, and 95.43%, respectively.	The model is indifferent as to whether the website's URL is active or contains an error. Short links, sensitive phrases, and phishing URLs that do not replicate other websites will be misclassified by the system.
An application for predicting phishing attacks: A case of implementing a support vector machine learning model [16].	SVM	Website Phishing URLs from the work of its reference study.	A GUI was also created for the user to report any email that turns out to be a phishing email. The outcome of the model shows that the polynomial function performed better with 84.5% accuracy, while the radial basis function had an accuracy score of 82.6%.	The SVM prediction model is based on a dataset of approximately 1,400 records, which may have influenced the model's accuracy.
Detecting Phishing Domains Using Machine Learning [17].	ANN, SVM, DTs, and RF algorithms.	URLs and UCI phishing domains dataset.	Developed and compared four models for investigating the efficiency of using machine learning to detect phishing domains. The findings show that the model based on the random forest technique is the most accurate of the other four techniques and outperforms other solutions in the literature.	Future work includes examining more machine learning algorithm techniques for phishing domains.
Phishing Website Detection through Multi-Model Analysis of HTML Content [18].	MLP and NLP models	Phishing URLs from OpenPhish.	MLP model and NLP models were used to create an advanced phishing detection model that focuses on HTML content. The standalone MLP model achieved an accuracy of 89.92%. The NLP-1 and NLP-2 models achieved accuracies of 93.84% and 96.76%, respectively. The MultiText-LP model, a fusion of NLP-1, NLP-2, and MLP models, achieved an accuracy of 97.18%, illustrating the synergistic effect of combining both approaches.	Using two pre-trained and one MLP model simultaneously requires a powerful GPU like A4000. Standard GPUs may struggle with their size, making training and deployment less efficient, particularly in resource-constrained environments.

Preprocessing

The next step is to preprocess the dataset to prepare it for machine learning, ensuring it is structured, reliable, and ready for training, thereby improving model performance. This includes cleaning to remove noise and duplicates, handling missing values through imputation or default categories, and encoding categorical features using one-hot or label encoding.

Balancing the dataset

Balancing the dataset is essential to prevent bias in the machine learning model and ensure fair representation of all classes during training. After balancing the dataset using undersampling, 504 emails remain: 252 phishing and 252 safe.

OSINT Feature Collection

This study developed a Python program to automate the extraction of domain names and URLs from raw output, utilizing two OSINT tools: Nmap and theHarvester. Figure 3 depicts the process of OSINT feature extraction using Python.

The Python program is based on the following commands, which produce 17 features shown in Table 2:

- `nmap -Pn -T4 --max-retries 3 [domain name]` [19], [20]
- `python theHarvester.py -d [domain name] -l 500 -b all` [21]

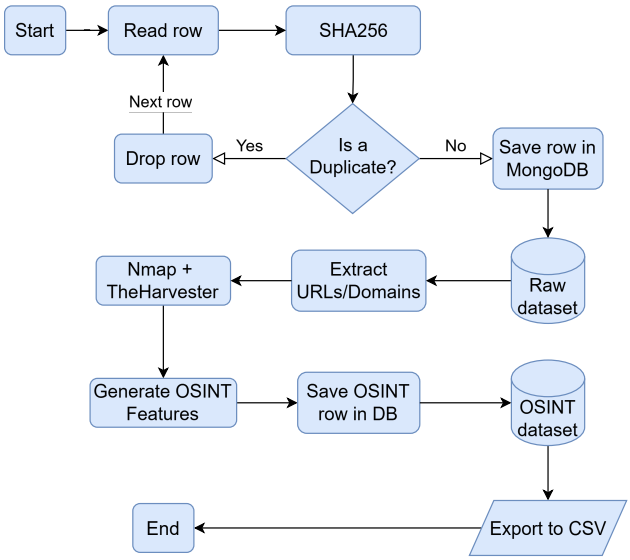


Figure 3. *Flowchart of OSINT Data Collection & Feature Extraction*

System Model in the Experiment

This study conducted 20 experiments across four datasets, using five ML classification algorithms: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Multinomial Naive Bayes (Multinomial NB), and XGBoost. These algorithms were selected for this experiment for their proven effectiveness in classification tasks and their widespread use in phishing detection and OSINT-related studies. Their prior use in the studies cited in Table 1 provides a baseline for comparing and validating this study's methodology and results.

We used Grid Search to optimize the hyperparameters of our models. Doing so ensures that each classification model achieves

Feature	Description	Example
hostname	Domain name retrieved from the email server.	`www.oreilly.com`
host_up	Host device reachability (1 = up, 0 = down).	`1, 0` (Sum will be calculated)
alternate_ip_count	Count of alternative IP addresses.	`2, 6` (Sum will be calculated)
ip_address	Main IP mapped to the domain name.	`34.169.83.167, 23.52.38.113`
common_web_ports_open	Open web server ports (0 = closed, 1 = open).	`0, 1` (Sum will be calculated)
open_ports_count	Total open TCP/UDP ports.	`4, 8` (Sum will be calculated)
filtered_ports_count	Count of filtered TCP/UDP ports.	`2, 1` (Sum will be calculated)
open_ports	Total listening ports on the device.	`80, 443`
rdns_record	Reverse DNS record for the IP.	`example.com`
https_supported	HTTPS support (0 = no, 1 = yes).	`1, 0` (Sum will be calculated for each URL)
services	Running services on open ports.	`HTTPS, FTP, mysql, ssh, pop3`
host_found	Host record detected.	`1, 0` (Sum will be calculated)
interesting_url	URL contains suspicious content.	`3, 10` (Sum will be calculated)
asn_found	Autonomous system number identified for the IP.	`0, 3, 11` (Sum will be calculated)
ip_found	Email contains an IP address.	`1, 3` (Sum will be calculated)
latency	Response delay during the scan (in seconds).	`0.037, 0.098, 0.100` (Sum will be calculated)
scan_duration	Total time taken to complete the scan (in seconds).	`43.64, 55.99` (Sum will be calculated)

Features and Descriptions from OSINT

the best performance in this comparative study.

Figure 4 summarizes the training process of ML models for email phishing detection in this experiment.

Results

Experimental Setting

Our experiments were conducted in VSCode, using a Jupyter kernel. We utilized four Python libraries: Scikit-learn, Pandas, NumPy, and Matplotlib. The latter was used to visualize the dataset size before and after balancing, as well as the confusion matrices for each test. Tables 3 and 4 list the hyperparameters used for training the models via Grid Search.

Machine Learning Models' Hyperparameters after finetuning English Dataset

Model	Hyperparameter	Baseline	OSINT
Decision Tree	criterion	"entropy"	"entropy"
	max_depth	None	None
	min_samples_split	2	2
	min_samples_leaf	1	4
Random Forest	n_estimators	400	400
	max_depth	None	40
	min_samples_split	2	5
	min_samples_leaf	2	1
SVM (linear)	C	10	10
XGBoost	n_estimators	200	200
	colsample_bytree	1.0	0.8
	learning_rate	0.05	0.05
	max_depth	8	8
	subsample	1.0	0.8
MultinomialNB	alpha	0.5	1.0

Machine Learning Models' Hyperparameters after finetuning for Arabic Dataset

Model	Hyperparameter	Baseline	OSINT
Decision Tree	criterion	"gini"	"entropy"
	max_depth	10	None
	min_samples_split	2	2
	min_samples_leaf	1	4
Random Forest	n_estimators	200	400
	max_depth	None	40
	min_samples_split	5	2
	min_samples_leaf	2	2
SVM (linear)	C	5	4
XGBoost	n_estimators	300	600
	colsample_bytree	0.8	0.8
	learning_rate	0.05	0.05
	max_depth	4	4
	subsample	1.0	0.8
MultinomialNB	alpha	2.0	1.0

Experimental Results

To evaluate the performance of our models, we focused on key metrics such as accuracy, F1 score, recall, and precision. Table 5 provides the experimental results obtained after training each model on both the English baseline dataset and the English OSINT dataset.

Experimental Results on English dataset before and after adding OSINT features (in %)

Classifier	Dataset	Accuracy	F1 Score	Precision	Recall
DT	Baseline	90.13	89.80	92.96	86.84
	OSINT	90.13	89.80	92.96	86.84
RF	Baseline	91.45	91.61	89.87	93.42
	OSINT	95.39	95.36	96.00	94.74
SVM	Baseline	98.03	98.04	97.40	98.68
	OSINT	96.71	96.69	97.33	96.05
XGBoost	Baseline	92.76	92.90	91.14	94.74
	OSINT	92.11	92.21	91.03	93.42
MNB	Baseline	98.68	98.68	98.68	98.68
	OSINT	98.68	98.68	98.68	98.68

Starting with RF, which demonstrated a notable increase in all 4 metrics, improving its accuracy from 91.45% to 95.39%. DT and Multinomial NB showed no change, and SVM and XGBoost experienced slight performance decreases, while maintaining high accuracies of 96.71% and 92.11%, respectively. Additionally, Multinomial NB was the best-performing model, achieving an accuracy of 98.68%.

Moving on to Table 6, which highlights the experimental results obtained using our translated Arabic datasets.

Experimental Results on Arabic dataset before and after adding OSINT features (in %)

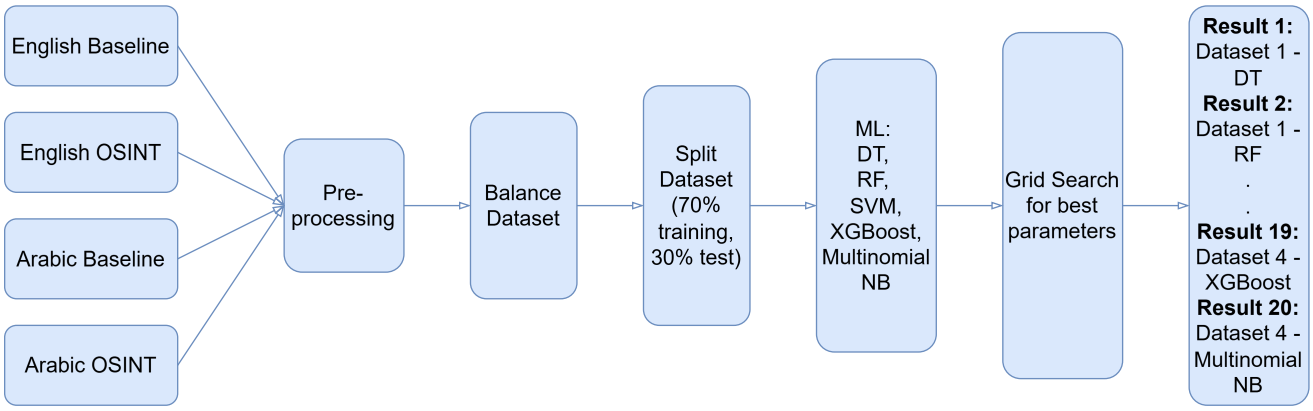
Classifier	Dataset	Accuracy	F1 Score	Precision	Recall
DT	Baseline	87.50	87.74	86.08	89.47
	OSINT	78.29	78.71	77.22	80.26
RF	Baseline	72.37	63.16	94.74	47.37
	OSINT	82.89	85.39	74.51	100
SVM	Baseline	94.08	94.34	90.36	98.68
	OSINT	94.08	94.19	92.41	96.05
XGBoost	Baseline	90.79	91.36	86.05	97.37
	OSINT	91.45	91.50	90.91	92.11
MNB	Baseline	91.45	91.82	87.95	96.05
	OSINT	94.74	94.74	94.74	94.74

As shown in the table above, DT experienced a decline in performance across all 4 metrics, decreasing its accuracy from 87.50% to 78.29%. However, OSINT features significantly enhanced RF's recall from 47.37% to 100%, and overall performance, despite a drop in precision, also yielded notable improvements for Multinomial NB and XGBoost. SVM maintained the same level of accuracy, improving its precision and recall while slightly degrading its F1 score.

Table 7 shows the accuracy gain or loss (Δ acc) that each classifier experiences after OSINT features are added, allowing direct comparison of each model's performance on the English and Arabic datasets.

As observed in the table, RF benefits most from OSINT feature inclusion in both languages, followed by Multinomial NB, whose accuracy remains constant on the English dataset and increases by 3.29 percentage points on the Arabic dataset.

Below are confusion matrices for both RF and Multinomial



width=

Figure 4. *Diagram of model-training pipelines for email-phishing detection.*

Accuracy Difference for each classifier across the Arabic and English datasets

Model	English Δacc	Arabic Δacc
DT	+0.0	-9.21
RF	+3.94	+10.52
SVM	-1.32	+0.0
XGBoost	-0.65	+0.66
MultinomialNB	+0.0	+3.29

NB models trained on the English and Arabic datasets, respectively.

Fig.6 shows that RF trained on the Arabic OSINT data produced zero false negatives, which means that the model successfully identified every single actual phishing email in the test set. However, this comes at the cost of precision, which drops to 74.51% due to 26 false positives.

Fig.7, the confusion matrix for Multinomial NB trained on the English OSINT dataset, demonstrates near-perfect, balanced performance. With only one false positive and one false negative, the model correctly classified the vast majority of instances, achieving an accuracy of 98.68%. The confusion matrices in Fig.5 and Fig.8 yield almost similar results.

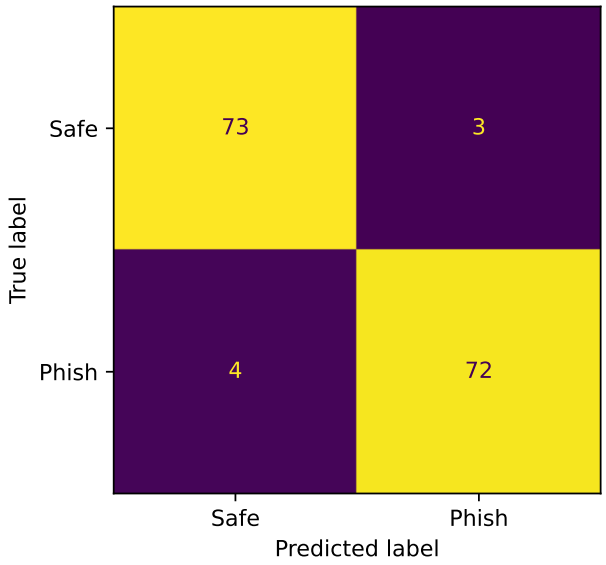


Figure 5. *Confusion matrix of RF trained on English OSINT dataset*

Discussion

Our experimental results reveal a nuanced impact of integrating OSINT features into our datasets on model performance for email phishing detection. While the addition of OSINT features generally enhanced the performance of several models, the extent of this improvement varied across classifiers and languages.

Specifically, on the English OSINT dataset, Multinomial NB achieved the highest accuracy of 98.68%. RF also performed strongly on this dataset, achieving an accuracy of 95.39%. However, SVM and XGBoost exhibited slight performance decreases, whereas DT showed no change.

In contrast, DT was the only model that showed a decline in performance after training on the Arabic OSINT dataset. This decline is likely due to the model's sensitivity to highly correlated features, as is the case in our dataset. Many OSINT attributes, such as IP addresses and open-port counts, are interdependent.

Another clear pattern evident from our results is that the utility of OSINT features is not only model-dependent but also language-dependent. Some models improved with OSINT features, while others degraded in performance depending on the specific classifier and the linguistic dataset being processed.

This paper [10] used a dataset of 10000 phishing URLs and, similar to our experiment, trained the models with a 70:30 training-to-testing split. It used 50 features. Their results show that XGBoost demonstrates the best performance.

In this paper [22], the database used contains 11215 records and 21 features, and was tested using RF, achieving an accuracy of 100%.

Rishikesh et al. [23] used a dataset of 36711 URLs from

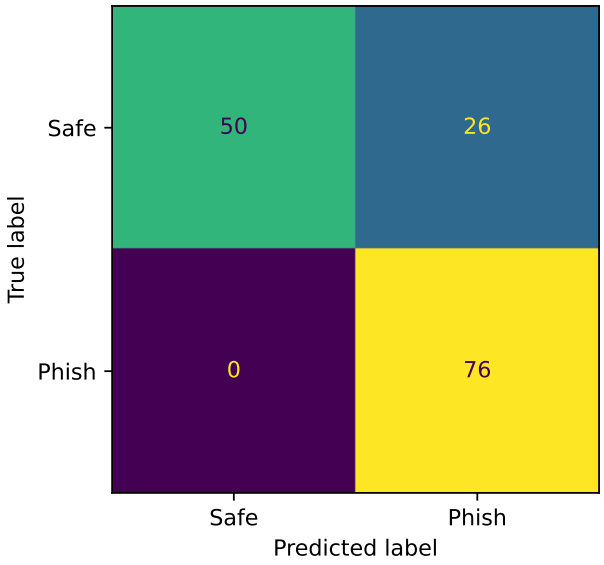


Figure 6. *Confusion matrix of RF trained on Arabic OSINT dataset*

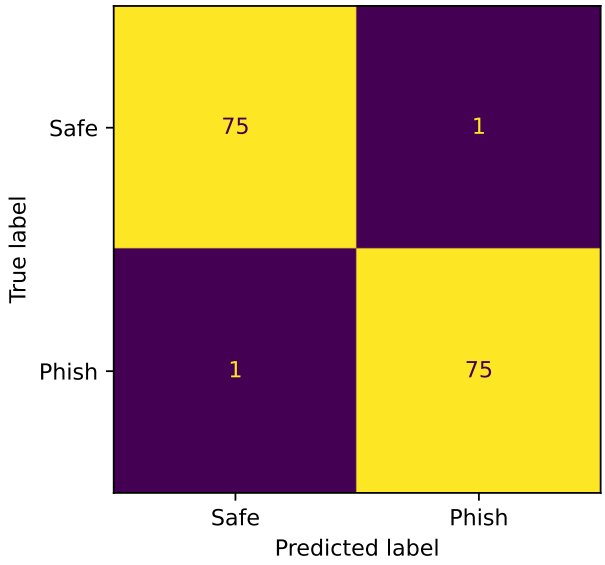


Figure 7. *Confusion matrix of MultinomialNB trained on English OSINT dataset*

which they extracted 16 features. Their results after splitting the dataset into a 70:30 ratio show higher performance for RF, which achieved an accuracy of 96.84%.

Our results align with prior research, indicating that RF exhibits a positive impact of OSINT feature integration across both English and Arabic datasets. As presented in Table 7, RF's accuracy improved by +3.94% in the English dataset and a remarkable +10.52% in the Arabic dataset. Furthermore, Multinomial NB maintained a high accuracy across both datasets (baseline and OSINT-enhanced) and both languages (English and Arabic). However, it is important to note that the datasets used in many other studies were considerably larger than ours, which could explain the observed performance differences.

Our primary objective was to demonstrate that incorporating OSINT features into a dataset improves model accuracy and overall performance in phishing detection. The experimental results supported our hypothesis, although in a few cases, models exhibited a slight decrease in accuracy or no change.

Conclusion

Our experimental study tests the hypothesis that a model trained on an OSINT feature-enhanced dataset is more likely to outperform a model relying solely on baseline features for email phishing detection. The machine learning classifiers implemented are DT, RF, SVM, XGBoost, and Multinomial NB. They were trained using our English and Arabic datasets, and the results obtained were evaluated using the following metrics: accuracy, F1 score, precision, recall, and confusion matrix. Most models improved their accuracy after training on the OSINT-enhanced datasets, in both English and Arabic, with Multinomial NB achieving the highest accuracy of 98.68%.

Notably, Multinomial NB achieved the highest accuracy of 98.68%, followed by RF, reaching 95.39% on the English OSINT dataset. However, certain models, such as DT on the Arabic dataset and, to a lesser extent, SVM and XGBoost on the English dataset, suffered decreases or no change in accuracy, highlighting the model and language-dependent utility of OSINT features.

For future research, expanding the dataset size and diversifying the range of phishing and safe emails would be crucial to enhance the generalizability of our findings. Another direction is to explore deep learning transformers such as XLM-Roberta and GPT-3, which could be particularly effective and well-suited to our multilingual datasets.

Acknowledgments

The European Union partially supported this work through ERASMUS MUNDUS, Project CyberMACS (Project No. 101082683, `https://cybermacs.eu`).

References

[1] A. Orunsolu, A. S. Sodiya, and A. T. Akinwale, "A predictive model for phishing detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 232–247, 2019.

[2] R. De, N. Pandey, and A. Pal, "Impact of digital surge during covid-19 pandemic: a viewpoint on research and practice," *International Journal of Information Management*, vol. 55, no. 102171, 2020.

[3] E. S. Shombot, G. Dusserre, R. Bestak, and N. B. Ahmed, "An application for predicting phishing attacks: A case of implementing a support vector machine learning model," *Cyber Security and Applications*, vol. 2, p. 100036, 2024.

[4] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*, 2007, pp. 649–656.

[5] J. Nordine, "OSINT framework," 2019. [Online]. Available: https://osintframework.com

[6] S. Chakraborty, "Phishing email detection," 2023. [Online].

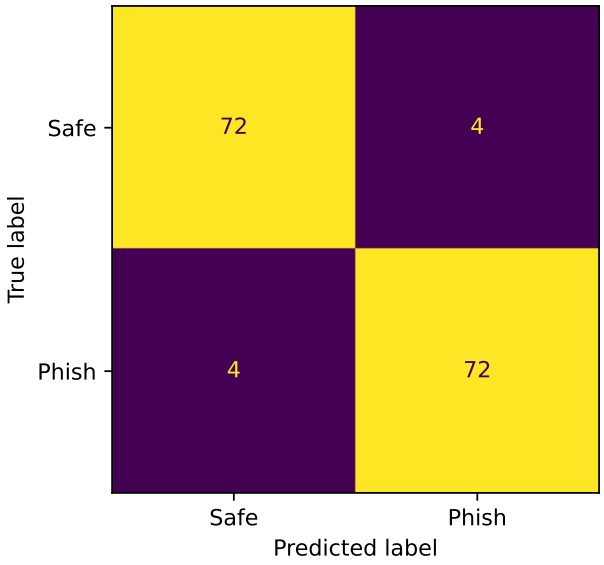


Figure 8. *Confusion matrix of MultinomialNB trained on Arabic OSINT dataset*

Available: https://www.kaggle.com/dsv/6090437

[7] A. S, P. R. Nishant, S. Baitha, and K. D. Kumar, "An ensemble classification model for phishing mail detection," *Procedia Computer Science*, vol. 233, pp. 970–978, 2024, 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705092400646X

[8] O. C. Çetlenbik, R. Gürfidan, and A. Süzen, "Classification of phishing attacks using the roberta model," 03 2024.

[9] L. Zhou, A. Gaurav, V. Arya, R. Attar, S. Bansal, and A. Alhomoud, "Enhancing phishing detection in semantic web systems using optimized deep learning models," *International Journal on Semantic Web and Information Systems*, vol. 20, pp. 1–13, 01 2024.

[10] S. C. G. et al., "A feature extraction approach for the detection of phishing websites using machine learning," *Journal of circuits, systems, and computers*, 08 2023.

[11] U. Maryam, "Phishing attacks facilitated by open-source intelligence," 10 2023.

[12] J. Rastenis, S. Ramanauskaitė, I. Suzdalev, K. Tunaitytė, J. Janulevičius, and A. Čenys, "Multi-language spam/phishing classification by email body text: Toward automated security incident investigation," *Electronics*, vol. 10, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/6/668

[13] D. Staples, S. Hakak, and P. Cook, "A comparison of machine learning algorithms for multilingual phishing detection," 08 2023, pp. 1–6.

[14] A. O. Bryushinin, A. V. Dushkin, and M. A. Melshiyan, "Automation of the information collection process by osint methods for penetration testing during information security audit," in *2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 2022, pp. 242–246.

[15] D. Mosa, M. Shams, A. Abohany, E.-S. El-kenawy, and M. Thabet, "Machine learning techniques for detecting phishing url attacks," *Computers, Materials & Continua*, vol. 75, pp. 1271–1290, 01 2023.

[16] E. S. Shombot, G. Dusserre, R. Bestak, and N. B. Ahmed, "An application for predicting phishing attacks: A case of implementing a support vector machine learning model," *Cyber Security and Applications*, vol. 2, p. 100036, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S277291842400002X

[17] S. Alnemari and M. Alshammari, "Detecting phishing domains using machine learning," *Applied Sciences*, vol. 13, no. 8, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/8/4649

[18] F. Çolhak Furkan et al., "Phishing website detection through multi-model analysis of html content," *arXiv (Cornell University)*, 01 2024.

[19] "Finding an organization's ip addresses — nmap network scanning," nmap.org. [Online]. Available: https://nmap.org/book/host-discovery-find-ips.html

[20] "Timing and performance — nmap network scanning," Nmap.org. [Online]. Available: https://nmap.org/book/man-performance.html

[21] S. Mhatre, F. Schwarz, K. Schwarz, and R. Creutzburg, "Osint-based email investigation," *Electronic Imaging*, vol. 36, pp. 328–17, 01 2024.

[22] J. Mehanović D., Kevrić, "Phishing website detection using machine learning classifiers optimized by feature selection," *International Information and Engineering Technology Association*, 2020.

[23] R. Mahajan and I. Siddavatam, "Phishing website detection using machine learning algorithms," *International Journal of Computer Applications*, vol. 181, pp. 45–47, 10 2018.

Author Biography

Panharith An earned his Telecommunication & Electronic BEng in 2018 and spent five years as a full-stack software engineer in Digital Transformation in business, banking, and public sectors in Cambodia. He is currently pursuing the ERASMUS Mundus Joint Master's Degree in Applied Cybersecurity (CyberMACS).

Rana Shafi holds a Bachelor's degree in Mathematics and Computer Science. She is currently pursuing an ERASMUS Mundus joint Master's degree in Applied Cybersecurity in Turkey. Her research interests include Artificial Intelligence, Deep Learning, Computer Networking, and Network Security.

Tionge Mughogho holds a Bachelor's degree in Computer Systems and Security. With two years of professional experience in software development and cybersecurity — spanning the national CERT of Malawi and the banking industry — they have honed their technical skills in critical areas. Tionge is currently pursuing a Master's degree in Advanced Cybersecurity through the ERASMUS Mundus scholarship at Kadir Has University in Istanbul, Turkey, furthering their expertise in the field.

Onyango Allan Onyango is currently a Cyber Defense intern at Kraken Technologies. He received his Master's degree in Pure Mathematics (2023). He is presently pursuing an ERASMUS Mundus joint Master's degree in Applied Cybersecurity at Kadir

Has University, Turkey and SRH University of Applied Sciences, Germany. He has 4 years of professional experience in Networks Security. His research interests include TDA, Network Security, Computer Networks, Operating Systems, and Cybersecurity.

Nikola Nachevski is a master's student in Applied Cybersecurity within the CyberMACS ERASMUS Mundus Joint Master's Degree program at SRH University Berlin and Kadir Has University. He holds a bachelor's degree in Software Engineering from Ss. Cyril and Methodius University in Skopje. His research interests include digital forensics, document and printer forensics, privacy in IoT systems, and the application of machine learning in cybersecurity.

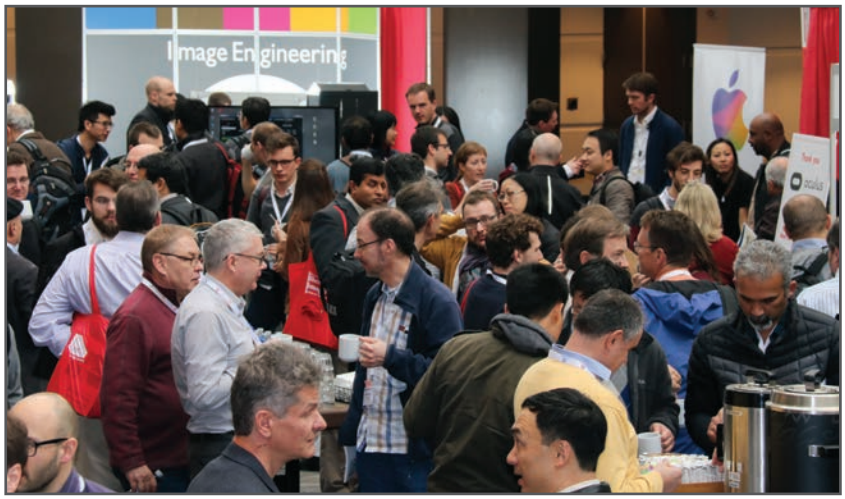
Reiner Creutzburg is a Retired Professor of Applied Computer Science at the Technische Hochschule Brandenburg in Brandenburg, Germany. Since 2019, he has been a Professor of IT Security at the SRH Berlin University of Applied Sciences, Berlin School of Technology. In 2025, he was appointed as a Senior Professor of Cybersecurity at the newly founded German University of Digital Science in Potsdam, Germany.

He is a member of the IEEE and SPIE, and has served as chairman of the Multimedia on Mobile Devices (MOBMU) Conference at the Electronic Imaging conferences since 2005. In 2019, he was elected a member of the Leibniz Society of Sciences to Berlin e.V. His research interests include Cybersecurity, Digital Forensics, Open-Source Intelligence (OSINT), Multimedia Signal Processing, e-learning, Parallel Memory architecture, and Modern Digital Media and Imaging Applications.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org