Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 3, June 2025, pp. 2361~2369
ISSN: 2302-9285, DOI: 10.11591/eei.v14i3.9198

Detecting spam using Harris Hawks optimizer as a feature selection algorithm

Mosleh M. Abualhaj, Ahmad Adel Abu-Shareha, Sumaya Nabil Alkhatib, Qusai Y. Shambour, Adeeb M. Alsaaidah
Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan

Article Info

Article history:

Received Aug 25, 2024
Revised Dec 7, 2024
Accepted Dec 25, 2024

Keywords:

Feature selection
Harris Hawks algorithm
ISCX-URL2016 dataset
Machine learning
Spam

ABSTRACT

The Harris Hawks optimization (HHO) was used in this study to enhance spam identification. Only the features with a high influence on spam detection have been selected using the HHO metaheuristic technique. The HHO technique's assessment of the selected features was conducted using the ISCX-URL2016 dataset. The ISCX-URL2016 dataset has 72 features, but the HHO technique reduces that to just 10 features. Extra tree (ET), extreme gradient boosting (XGBoost), and support vector machine (SVM) techniques are used to complete the classification assignment. 99.81% accuracy is attained by the ET, 99.60% by XGBoost, and 98.74% by SVM. As we can see, with the ET, XGBoost, and k-nearest neighbor (KNN) techniques, the HHO technique achieves accuracy above 98%. Nonetheless, the ET technique outperforms the XGBoost and KNN techniques. ET outperforms other methods due to its robust ensemble approach, which benefits from the diverse and relevant feature subset selected by HHO. HHO's effective reduction of noisy or redundant features enhances ET's ability to generalize and avoid overfitting, making it a highly efficient combination for spam detection. Thus, it looks promising to combat spam emails by combining the ET technique for classification with the HHO technique for feature selection.

Corresponding Author:

Mosleh M. Abualhaj
Faculty of Information Technology, Al-Ahliyya Amman University
Amman 19111, Jordan
Email: m.abualhaj@ammanu.edu.jo

1. INTRODUCTION

Email is one of the most widely used and effective methods of communicating over the internet, as well as being able to share information or messages [1]. Between 2019 and 2021, the number of people having access to email increased from 3.9 billion to 4.1 billion worldwide. In addition, it is anticipated that this number will increase to 4.6 billion by the year 2025. Spam emails have also increased at a rapid rate, which is not surprising given the relevance and widespread use of email. As of the 13th of August 2024, the USA was the nation that had the most spam emails sent within a single day across the entire world, with approximately eight billion [2]. Spam is undesirable emails that contain various information, such as adverts, offers, and links to suspicious websites. Additionally, spam email is an efficient carrier of malware that infects computers with viruses. By sending unsolicited commercial emails, spammers intend to conduct email fraud. As a result, it is necessary to have spam emails separated from other emails [3], [4].

Nowadays, machine learning (ML) algorithms are wildly used to alleviate spam emails. ML algorithms are fed by large amounts of spam and non-spam emails to separate the spam and non-spam emails that are received. ML algorithms are divided into several types, including supervised learning algorithms, which are used with labeled data such as spam emails [5]-[7]. Several supervised learning algorithms can be used to classify spam from non-spam emails, including extra tree (ET), extreme gradient boosting (XGBoost), and support vector machine (SVM) [7]-[9]. The supervised learning algorithms use spam and non-spam email data to construct a supervised learning model that alleviates spam. The model incorporates several criteria to distinguish between spam and non-spam, taking into account the attributes of the email. The supervised learning model's performance would improve in proportion to the degree of accuracy of the email attributes. The email data contains a substantial amount of information. At the same time, as specific attributes are essential for identifying spam from non-spam, others are either less significant or irrelevant. ML, using so-called feature selection algorithms, can assist in determining the vital attributes that distinguish incoming emails as spam or non-spam. This study will utilize the Harris Hawks optimization (HHO) algorithm to identify the primary attributes that distinguish spam and non-spam emails [10]-[13].

The proposed ML model, which uses the HHO algorithm for feature selection along with various classifiers, offers significant advantages over traditional methods. By optimizing feature selection, it reduces the number of features needed and improves accuracy, all while lowering computational costs. Unlike static methods, this approach adapts to changing spam patterns using advanced optimization techniques. When paired with powerful classifiers like ET, which are highly effective with complex data, this model greatly boosts spam detection efficiency and scalability, making it better suited to tackle today’s spam challenges.

Several works have been proposed for spam detection. Esquivel *et al.* [14] developed a process to filter spam emails based on the reputation of the IP address. End-hosts, authentic servers, and spammers are the three categories used to classify email senders. Next, an empirical analysis is performed on each of the three groups based on the internet edge to determine the impact of applying the IP reputation approach. Finally, the existing list of IP reputations is regularly updated in accordance with the method developed for creating individualized IP reputation lists. The findings demonstrated that the developed method can distinguish as much as 90% of spam from non-spam emails.

Xu *et al.* [15] developed a solution to the problem of spam emails. However, the developed solution treats spam based on images rather than spam based on text. The first step in achieving this goal involves converting the images into a Base64-encoded string. Next, the Base64 string that was produced is segmented into groups. Next, the n-gram approach is used to tokenize and extract attributes from each image, subsequently representing it as a vector with binary features. Finally, the SVM is used to distinguish between spam and non-spam images. Several assessment criteria, such as accuracy, recall, precision, and f1-score, are utilized to assess the developed solution. Compared to the existing feature extraction solution, the results demonstrated that the developed solution has obtained a performance that is significantly superior for image-based spam classification.

Debnath and Kar [16] developed email spam detection solutions using ML and deep learning approaches. The goal of these approaches is to differentiate spam from non-spam emails accurately. The email dataset from Enron has been utilized, and deep learning models have been constructed to identify and categorize new forms of spam email using long short-term memory (LSTM) and bidirectional encoder representations from transformers (BERT) methodologies. A natural language processing approach was utilized to assess and prepare data for the text of the email. The results are compared to the previous solutions for identifying spam in emails. The developed deep learning approach achieved the highest accuracy of 99.14% when using BERT, 98.34% when using BiLSTM, and 97.15% when using LSTM.

Our study aims to address these challenges by using a novel feature selection method that is the HHO. This method is designed to select the most relevant features for spam detection, thereby reducing the feature space and enhancing the performance of ML classifiers. By solving the issues of inefficient feature selection and model performance degradation due to irrelevant features, our work offers a significant improvement over existing approaches in spam email detection. This combination of HHO and advanced classifiers not only enhances detection accuracy but also reduces the model’s computational demands, making it more feasible for real-world applications.

2. METHOD

This section presents the suggeted method for identifying spam emails. First, the ISCX-URL2016 dataset will be discussed. Then, the HHO algorithm used for feature selection will be presented. Finally, the algorithms used for the classification process will be elaborated.

2.1. ISCX-URL2016 dataset

The evaluation process in this work utilized the spam instances from the ISCX-URL2016 dataset. Upon completion of the cleaning process for the ISCX-URL2016 dataset, there are now 14,479 instances and 72 features remaining. The instances are categorized into benign and spam instances. There are 6,698 spam instances and 7,780 benign instances [17], [18]. Since the number of instances is evenly distributed, there is no requirement to apply oversampling or undersampling techniques to the dataset. However, the 72 features in the ISCX-URL2016 dataset contain values distributed over wide ranges, as shown in Table 1. These values should be scaled to the same ranges to avoid bias in the ML classification algorithms [17], [18]. The min-max scaling algorithm is used in this work to scale the values of the features in the ISCX-URL2016 dataset, as shown in Table 2 [7], [19]. In addition, a significant number of the 72 features may have minimal influence in distinguishing the instance as either spam or benign. The HHO algorithm will be utilized to apply feature selection on the dataset, explicitly targeting the identification of features that substantially impact spam detection.

Table 1. Sample of the ISCX-URL2016 spam dataset

#	Feature	Min value	Max value
1	LongestPathTokenLength	0	1393
2	Query_LetterCount	-1	1173
3	Extension_LetterCount	-1	1179
4	URL_Letter_Count	15	1202
5	LongestVariableValue	-1	1385
6	LongestPathTokenLength	0	1393

Table 2. Sample of the ISCX-URL2016 before and after scaling

#	Before scaling	After scaling
1	17, 2, 2, 6	0.012274368, 0, 0, 0.363636364
2	0, 8, 2, 2	0, 0.545454545, 0, 0
3	0, 3.5, 2, 2	0, 0.136363636, 0, 0
4	0, 4.5, 2, 2	0, 0.227272727, 0, 0
5	3, 0, 2.6666667, 3	0.333333333, 0, 0.060606064, 0.333333333

2.2. Feature selection

Recently, researchers proposed the HHO algorithm as a novel swarm optimization algorithm. Figure 1 shows the HHO algorithm hunting behavior. It aims to simulate Harris Hawks' searching and hunting activities as they pursue rabbits. Hawks are among the most intelligent birds in nature, as demonstrated by their reported advancements in eating activities. Three primary phases mathematically represent their hunting activities for rabbits in swarms: the exploration phase, the transition from exploration to exploitation, and the actual exploitation phase. The initial phase, the exploration phase, involves imitating the hunting behavior of Harris' hawks as they search for prey. The second phase is a transitional phase that imitates the actions of Harris hawks, who execute various moves based on the prey's energy level when trying to escape. Using the knowledge from the previous phase, the exploitation phase conducts a localized search to improve the quality of current solutions [10], [13]. HHO was chosen for its effective balance between exploration and exploitation, allowing it to find optimal feature subsets by avoiding local optima. Its fast convergence and adaptability make it well-suited for high-dimensional datasets like those used in spam detection, outperforming other algorithms in complex optimization tasks.

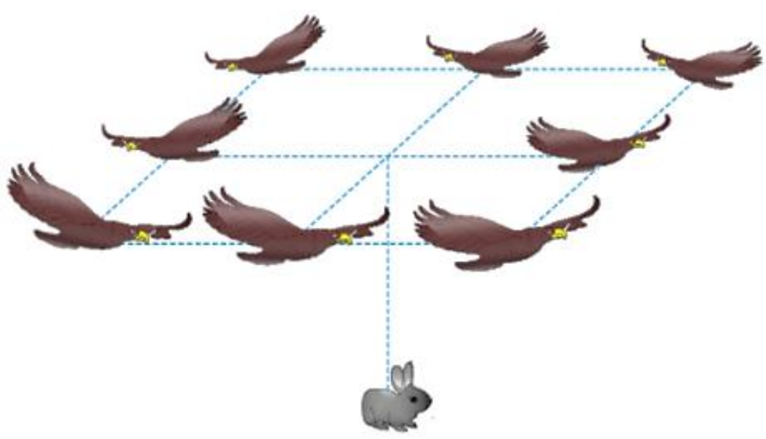

Figure 1. HHO hunting behavior [10]

The HHO algorithm has been used to analyze the spam features in the ISCX-URL2016 dataset to identify their importance in detecting spam emails. After examining the 72 features in the ISCX-URL2016 dataset, the HHO algorithm found a sub-dataset of ten key features to identify spam emails. The resulted sub-dataset conatins the following features: NumberRate_Domain, tld, pathurlRatio, pathDomainRatio, CharacterContinuityRate, ArgUrlRatio, Filename_LetterCount, NumberofDotsinURL, dld_getArg, NumberRate_DirectoryName.

2.3. Classification algorithms

One of the most popular supervised learning algorithms is the SVM. It is primarily utilized for classification problems in an ML context. The SVM algorithm aims to construct an optimal decision boundary that effectively separates an n-dimensional space into distinct classes, enabling efficient classification of new data points. The decision boundary that optimally separates the data points is called the hyperplane in SVM. The SVM algorithm contains several hyperparameters, the values of which vary depending on the problem to be handled [20]. These hyperparameters can be assigned using several methods, including the random search (RS) method, which will be used in this work. Table 3 shows the values of the SVM algorithm hyperparameters using the RS method.

Table 3. Hyperparameters of the SVM algorithm

#	Hyperparameter	Description	Value
1	c	Controls regularization	1.5
2	kernel	Determines the type of decision boundary	rbf
3	gamma	Defines how far the influence of a single training example reaches	0.01
4	degree	The degree of the polynomial kernel function	3
5	coef0	The independent term in the kernel function	0.0

XGBoost is a library of gradient-boosting algorithms tuned for use with contemporary data science tools and issues. Its key advantages include being highly scalable, parallelizable, and executing quickly, as well as typically outperforming other algorithms. XGBoost additionally uses a more regularized model formalization to control overfitting, which improves its performance. The XGBoost algorithm contains several hyperparameters, the values of which vary depending on the problem to be handled [21]. Table 4 shows the values of the XGBoost algorithm hyperparameters using the RS method.

Table 4. Hyperparameters of the XGBoost algorithm

#	Hyperparameter	Description	Value
1	n_estimators	Number of trees in the model	150
2	learning_rate	Controls the weight of each tree	0.05
3	max_depth	Maximum depth of the tree	6
4	min_child_weight	Minimum sum of instance weight needed in a child	3
5	subsample	Fraction of samples used for training each tree	0.8

The ETs ensemble method derives from the original decision tree (DT) algorithm. The classical DT algorithm divides a learning set into binary homogeneous subsets by applying the "if-then" rule at each internal node of the tree. The majority class will label the terminal node. Ultimately, the tree produces multiple class prediction rules to construct a predictive model. The initial categorization of individual trees may excessively fit the training data. The ETs ensemble method significantly improves DT performance. The ETs-based ensemble method utilizes randomization to produce a more robust prediction. The ETs method constructs each tree using the entire training set, incorporating a different test node at each step. Random selection of a single characteristic determines the optimal split, leading to diverse and uncorrelated trees. The ET algorithm contains several hyperparameters, the values of which vary depending on the problem to be handled [22]. Table 5 shows the values of the ET algorithm hyperparameters using the RS method.

Table 5. Hyperparameters of the ET algorithm

#	Hyperparameter	Description	Value
1	n_estimators	Number of trees in the forest	200
2	max_depth	Maximum depth of each tree	15
3	min_samples_split	Minimum number of samples needed to split a node	10
4	min_samples_leaf	Minimum number of samples required at a leaf node	2
5	max_features	Fraction of features considered for splitting at each node	0.8

3. RESULTS AND DISCUSSION

The experiments have been conducted on a laptop with the following hardware and software specs: Acer Aspire E5-575G model, Intel Core i7-7500U CPU (4M cache, 3.50 GHz speed, 2 Threads, and 4 Cores), 8 GB RAM, Ubuntu 24.04 LTS, and Python programming language. The confusion matrix serves as a vital tool in the evaluation of spam detection models, particularly those leveraging the sub-dataset selected by the HHO algorithm. It provides a comprehensive summary of prediction results, allowing for the visualization of how well the classification model performs in distinguishing between spam and non-spam emails.

Several metrics derived from the confusion matrix (Figure 2) provide varied perspectives on the performance of spam detection classifiers. One of the primary metrics is accuracy (1), which measures the overall effectiveness of the spam classification across all classes. This metric gives a broad view of how well the model is performing but does not account for the distribution of classes. Another critical metric is recall, calculated to determine the proportion of actual spam emails that are accurately identified by the model. recall (2) is particularly important in scenarios where the cost of missing positive instances is high, such as in the context of phishing attacks. Additionally, precision (2) serves as an essential metric that gauges the proportion of emails predicted as spam that are genuinely spam. This metric holds significant weight in situations where the consequences of false positives are severe, potentially leading to the loss of important emails [23]-[25].

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{3}$$

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 2. Confusion matrix

Figure 3 shows the accuracy of the ET, XGBoost, and SVM algorithms. The ET algorithm achieved an accuracy of 99.81%, the XGBoost algorithm achieved an accuracy of 99.60%, and the SVM algorithm achieved an accuracy of 98.74%. The ET attained higher accuracy than XGBoost by 0.21% and attained higher accuracy than SVM by 1.07%. Therefore, the ET algorithm comes in first place among the three algorithm, which means that ET algorithm is the most effective in correctly predicting the spam emails.

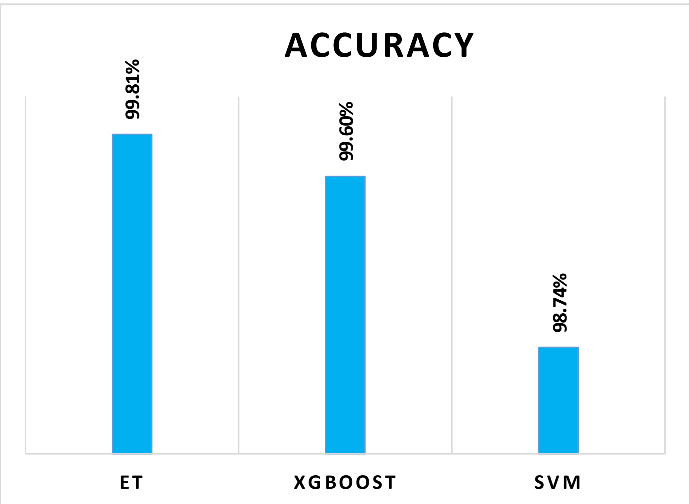

Figure 3. Accuracy of the HHO algorithm

Figure 4 shows the recall of the ET, XGBoost, and SVM algorithms. The ET algorithm achieved a recall of 99.81%, the XGBoost algorithm achieved a recall of 99.60%, and the SVM algorithm achieved a recall of 98.74%. The ET attained higher recall than XGBoost by 0.21% and attained higher recall than SVM by 1.07%. Therefore, the ET algorithm comes in first place among the three algorithm, which means that ET algorithm is the most effective in correctly predicting the legitimate emails.

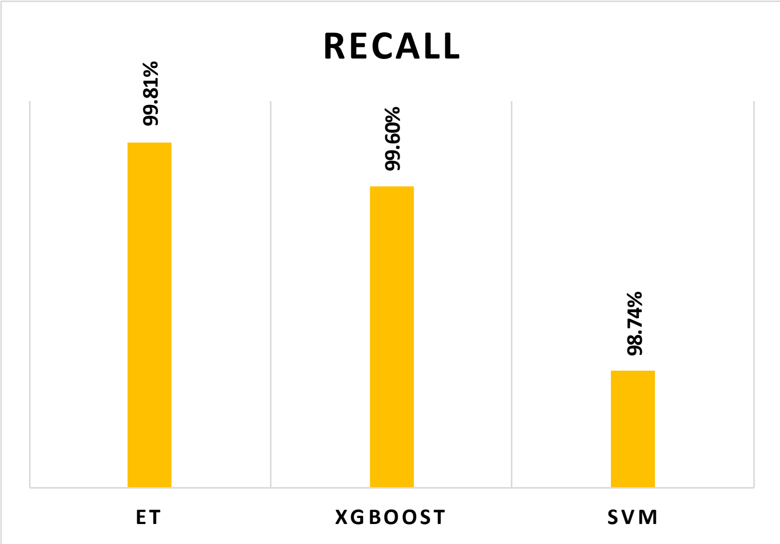

Figure 4. Recall of the HHO algorithm

Figure 5 shows the precision of the ET, XGBoost, and SVM algorithms. The ET algorithm achieved a precision of 99.81%, the XGBoost algorithm achieved a precision of 99.60%, and the SVM algorithm achieved a precision of 98.76%. The ET attained higher precision than XGBoost by 0.21% and attained higher precision than SVM by 1.05%. Therefore, the ET algorithm comes in first place among the three algorithm, which means that ET algorithm makes fewer false positive errors compared to XGBoost and SVM.

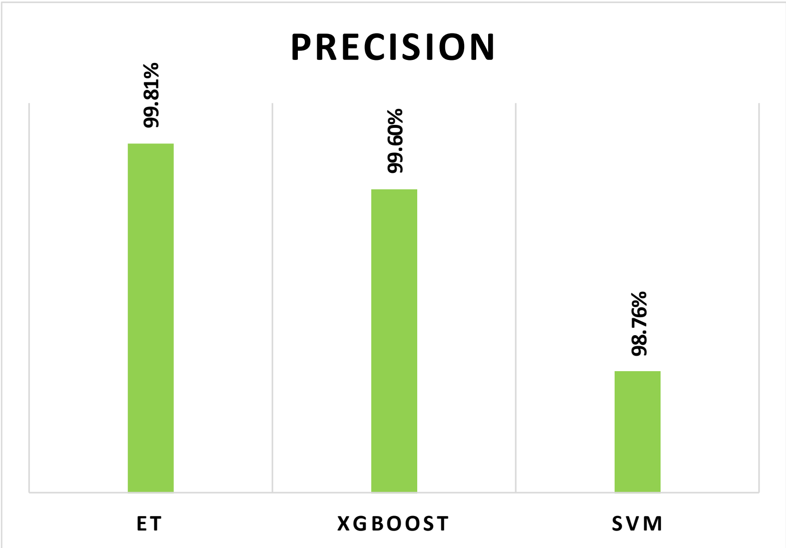

Figure 5. Precision of the HHO algorithm

4. CONCLUSION

Companies must confront the risk of spam emails. Attackers utilize spam emails to propagate a range of attacks. This work uses ML techniques to reduce the propagation of spam emails. The HHO technique is first employed to determine the crucial features that differentiate spam emails from legitimate ones. The HHO technique has condensed the 72 features of the ISCX-URL2016 spam dataset to a mere ten features. The performance of the subset of features selected by HHO has been evaluated using three renowned ML techniques: ET, XGBoost, and SVM techniques. These three ML techniques have been tailored to address the specific requirements of the spam detection problem. The ET, XGBoost, and SVM techniques have achieved an impressive accuracy of 99.81%, 99.60%, and 98.74%, respectively, in accurately identifying spam emails. The ET technique has shown superior performance compared to the XGBoost and SVM techniques, with improvements of 0.21% and 1.07% respectively. The findings suggest that the HHO technique attains the best result with the ET technique, indicating that combining both the HHO and ET techniques can potentially improve the detection of spam emails.

The results of this study present a strong method for detecting spam, which can be valuable for businesses and cybersecurity efforts in protecting against phishing and malware. By using the proposed model, organizations can greatly decrease the amount of spam that ends up in employees' inboxes, leading to enhanced productivity and better protection of sensitive data. Moreover, the model’s high accuracy and efficiency make it ideal for real-time use, allowing companies to respond swiftly to changing spam threats. Future research might look into incorporating deep learning methods like recurrent neural networks or transformers to boost the model's capacity to recognize intricate spam patterns. Furthermore, creating real-time spam detection systems that can learn and adapt to emerging spam trends could significantly improve the model's performance in constantly changing environments.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mosleh M. Abualhaj	✓	✓			✓				✓	✓		✓		
Ahmad Adel Abu-Shareha		✓		✓		✓				✓				
Sumaya Nabil Alkhatib			✓		✓		✓	✓		✓	✓			
Qusai Y. Shambour				✓		✓		✓	✓		✓			
Adeeb M. Alsaaidah					✓	✓			✓					

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in [Canadian Institute for Cybersecurity] at https://www.unb.ca/cic/datasets/url-2016.html [doi: 10.1007/978-3-319-46298-1_30], reference number [18].

REFERENCES

[1] J. Wei, X. Chen, J. Wang, X. Hu, and J. Ma, “Enabling (End-to-End) Encrypted Cloud Emails With Practical Forward Secrecy,” *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2318–2332, Jul. 2022, doi:

10.1109/TDSC.2021.3055495.

[2] A. Petrosyan, “Daily number of spam emails sent worldwide as of August 2024, by country,” Statista, Aug. 13, 2024, [Online]. Available: https://www.statista.com/statistics/1270488/spam-emails-sent-daily-by-country/. (Accessed: Aug. 20, 2024).

[3] G. Kambourakis, G. D. Gil, and I. Sanchez, “What Email Servers Can Tell to Johnny: An Empirical Study of Provider-to-Provider Email Security,” *IEEE Access*, vol. 8, pp. 130066–130081, 2020, doi: 10.1109/ACCESS.2020.3009122.

[4] M. M. Abualhaj, Q. Y. Shambour, A. Alsaaidah, A. Abu-Shareha, S. Al-Khatib, and M. O. Hiari, “Enhancing Spam Detection Using Hybrid of Harris Hawks and Firefly Optimization Algorithms,” *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 901–911, Sep. 2024, doi: 10.47738/jads.v5i3.279.

[5] A. AlMahmoud, E. Damiani, H. Otrok, and Y. Al-Hammadi, “Spamdoop: A privacy-preserving big data platform for collaborative spam detection,” *IEEE Transactions on Big Data*, vol. 5, no. 3, pp. 293–304, Sep. 2019, doi: 10.1109/TBDATA.2017.2716409.

[6] W. Z. Khan, M. K. Khan, F. T. Bin Muhaya, M. Y. Aalsalem, and H. C. Chao, “A Comprehensive Study of Email Spam Botnet Detection,” *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2271–2295, 2015, doi: 10.1109/COMST.2015.2459015.

[7] M. M. Abualhaj, A. A. Abu-Shareha, M. O. Hiari, Y. Alrabanah, M. Al-Zyoud, and M. A. Alsharaiah, “A Paradigm for DoS Attack Disclosure using Machine Learning Techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 192–200, 2022, doi: 10.14569/IJACSA.2022.0130325.

[8] K. M. K. Raghunath, V. V. Kumar, M. Venkatesan, K. K. Singh, T. R. Mahesh, and A. Singh, “XGBoost Regression Classifier (XRC) Model for Cyber Attack Detection and Classification Using Inception V4,” *Journal of Web Engineering*, vol. 21, no. 4, pp. 1295–1322, Apr. 2022, doi: 10.13052/jwe1540-9589.21413.

[9] A. Sonny, A. Kumar, and L. R. Cenkeramaddi, “Carry Object Detection Utilizing mmWave Radar Sensors and Ensemble-Based Extra Tree Classifiers on the Edge Computing Systems,” *IEEE Sensors Journal*, vol. 23, no. 17, pp. 20137–20149, Sep. 2023, doi: 10.1109/JSEN.2023.3295574.

[10] Z. Yu, X. Shi, J. Zhou, X. Chen, and X. Qiu, “Effective assessment of blast-induced ground vibration using an optimized random forest model based on a harris hawks optimization algorithm,” *Applied Sciences (Switzerland)*, vol. 10, no. 4, pp. 1–17, Feb. 2020, doi: 10.3390/app10041403.

[11] Q. Y. Shambour, M. M. Al-Zyoud, A. H. Hussein, and Q. M. Kharma, “A doctor recommender system based on collaborative and content filtering,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 884–893, Feb. 2023, doi: 10.11591/ijece.v13i1.pp884-893.

[12] A. Al Saaidah *et al*., “Enhancing malware detection performance: leveraging K-Nearest Neighbors with Firefly Optimization Algorithm,” *Multimedia Tools and Applications*, pp. 1–24, Mar. 2024, doi: 10.1007/s11042-024-18914-5.

[13] Y. Xu, Y. Guo, A. K. Jumani, and S. F. A. Khatib, “Application of ecological ideas in indoor environmental art design based on hybrid conformal prediction algorithm framework,” *Environmental Impact Assessment Review*, vol. 86, p. 106494, Jan. 2021, doi: 10.1016/j.eiar.2020.106494.

[14] H. Esquivel, A. Akella, and T. Mori, “On the effectiveness of IP reputation for spam filtering,” in *2010 2nd International Conference on COMmunication Systems and NETworks, COMSNETS 2010*, IEEE, Jan. 2010, pp. 1–10, doi: 10.1109/COMSNETS.2010.5431981.

[15] C. Xu, Y. Chen, and K. Chiew, “An approach to image spam filtering based on Base64 encoding and N-gram feature extraction,” in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2010, pp. 171–177, doi: 10.1109/ICTAI.2010.31.

[16] K. Debnath and N. Kar, “Email Spam Detection using Deep Learning Approach,” in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, IEEE, May 2022, pp. 37–41, doi: 10.1109/COM-IT-CON54601.2022.9850588.

[17] R. Aloufi and A. R. Alharbi, “K-means and Principal Components Analysis Approach For Clustering Malicious URLs,” in *2023 3rd International Conference on Computing and Information Technology, ICCIT 2023*, IEEE, Sep. 2023, pp. 359–364, doi: 10.1109/ICCIT58132.2023.10273923.

[18] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, “Detecting malicious URLs using lexical analysis,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9955 LNCS, 2016, pp. 467–482, doi: 10.1007/978-3-319-46298-1_30.

[19] A. A. Laghari, A. K. Jumani, R. A. Laghari, and H. Nawaz, “Unmanned aerial vehicles: A review,” *Cognitive Robotics*, vol. 3, pp. 8–22, 2023, doi: 10.1016/j.cogr.2022.12.004.

[20] K. P. Lin and M. S. Chen, “On the design and analysis of the privacy-preserving SVM classifier,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 11, pp. 1704–1717, Nov. 2011, doi: 10.1109/TKDE.2010.193.

[21] Y. Jiang, G. Tong, H. Yin, and N. Xiong, “A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters,” *IEEE Access*, vol. 7, pp. 118310–118321, 2019, doi: 10.1109/ACCESS.2019.2936454.

[22] M. M. Abualhaj, S. Al-Khatib, M. O. Hiari, and Q. Y. Shambour, “Enhancing Spam Detection Using Hybrid of Harris Hawks and Firefly Optimization Algorithms,” *Journal of Soft Computing and Data Mining*, vol. 5, no. 2, pp. 161-174, Dec. 2024.

[23] H. Al-Mimi, N. A. Hamad, M. M. Abualhaj, M. S. Daoud, A. Al-Dahoud, and M. Rasmi, “A n Enhanced Intrusion Detection System for Protecting HTTP Services from Attacks,” *International Journal of Advances in Soft Computing and its Applications*, vol. 15, no. 2, pp. 67–84, 2023, doi: 10.15849/IJASCA.230720.05.

[24] Z. Chkirbene *et al*., “A Weighted Machine Learning-Based Attacks Classification to Alleviating Class Imbalance,” *IEEE Systems Journal*, vol. 15, no. 4, pp. 4780–4791, Dec. 2021, doi: 10.1109/JSYST.2020.3033423.

[25] J. Ding, X. H. Hu, and V. Gudivada, “A Machine Learning Based Framework for Verification and Validation of Massive Scale Image Data,” *IEEE Transactions on Big Data*, vol. 7, no. 2, pp. 451–467, Jun. 2021, doi: 10.1109/TBDATA.2017.2680460.

BIOGRAPHIES OF AUTHORS

Prof. Mosleh M. Abualhaj is a senior lecturer in Al-Ahliyya Amman University. He received his first degree in Computer Science from Philadelphia University, Jordan, in 2004, master degree in Computer Information System from the Arab Academy for Banking and Financial Sciences, Jordan in 2007, and Ph.D. in Multimedia Networks Protocols from Universiti Sains Malaysia in 2011. His research area of interest includes VoIP, congestion control, cybersecurity data mining, and optimization. He can be contacted at email: m.abualhaj@ammanu.edu.jo.

Dr. Ahmad Adel Abu-Shareha received his first degree in Computer Science from Al Al-Bayt University, Jordan, 2004, Master degree from Universiti Sains Malaysia (USM), Malaysia, 2006, and Ph.D. degree from USM, Malaysia, 2012. His research focuses on data mining, artificial intelligent, and multimedia security. He investigated many machine learning algorithms and employed artificial intelligent in variety of fields, such as network, medical information process, knowledge construction, and extraction. He can be contacted at email: a.abushareha@ammanu.edu.jo.

Ms. Sumaya Nabil Alkhatib is a senior lecturer in Al-Ahliyya Amman University. She received his first degree in Computer Science from Baghdad University, Iraq, in June 1994 and master degree in Computer Information System from the Arab Academy for Banking and Financial Sciences, Jordan in February. Her research area of interest includes VoIP, multimedia networking, and congestion control. She can be contacted at email: sumayakh@ammanu.edu.jo.

Prof. Qusai Y. Shambour received the B.Sc. degree in Computer Science from Yarmouk University, Jordan, in 2001, the M.S. degree in computer networks from University of Western Sydney, Australia, in 2003, and the Ph.D. degree in software engineering from the University of Technology Sydney, Australia, in 2012. Currently, he is a Professor at the Department of Software Engineering, Al-Ahliyya Amman University, Jordan. His research interests include information filtering, recommender systems, VoIP, machine learning, and data science. He can be contacted at email: q.shambour@ammanu.edu.jo.

Dr. Adeeb M. Alsaaidah received the Bachelor's degree in Computer Engineering from the Faculty of Engineering, ALBalqa Applied University, the Master's degree in Networking and Computer Security from NYIT University, and the Ph.D. degree in Computer Network from USIM, Malaysia. He is currently an Assistant Professor in Network and Cybersecurity Department at Al-Ahliyya Amman University (AAU). His research interests include network performance, multimedia networks, network quality of service (QoS), the IoT, network modeling and simulation, network security, and cloud security. He can be contacted at email: a.alsaaidah@ammanu.edu.jo.