AI BASED FEATURE SELECTION WITH UNSUPERVISED LEARNING FOR EFFICIENT SPAM AND PHISHING EMAIL CLASSIFICATION

[1]Krishna Reddy Seelam, [2]M Balaswathi, [3]K Swathi

[1]Associate Professor, Department of ECE, Sree Dattha Institute of Engineering and Science

[2,3]Assistance Professor, Department of ECE, Sree Dattha Institute of Engineering and Science

ABSTRACT

Email has become one of the most important forms of communication. In 2014, there are estimated to be 4.1 billion email accounts worldwide, and about 196 billion emails are sent each day worldwide. Spam is one of the major threats posed to email users. In 2013, 69.6% of all email flows were spam. Links in spam emails may lead to users to websites with malware or phishing schemes, which can access and disrupt the receiver's computer system. These sites can also gather sensitive information from. Additionally, spam costs businesses around $2000 per employee per year due to decreased productivity. Therefore, an effective spam filtering technology is a significant contribution to the sustainability of the cyberspace and to our society. Current spam techniques could be paired with content-based spam filtering methods to increase effectiveness. Content-based methods analyze the content of the email to determine if the email is spam. Therefore, this project employs artificial neural networks to detect SPAM, HAM, and Phishing emails by applying features selection algorithm called PCA (principal component analysis). All existing algorithms detected only SPAM and HAM emails, but proposed algorithm designed to detect 3 different classes called SPAM, HAM, and Phishing. To implement this project, we have combined three different datasets called UCI, CSDMC and SPAM ASSASSIN dataset, where UCI and CSDMC datasets provided SPAM and HAM emails and Spam Assassin dataset provided Phishing emails. All these emails were processed to extract important features used in spam and phishing emails such as JAVA SCRIPTS, HTML tags and other alluring URLS to attract users.

Keywords: malware, phishing, decreased productivity, spam filtering, content-based methods, artificial neural networks, PCA (principal component analysis), UCI, CSDMC, Spam Assassin, SPAM, HAM, phishing emails

1. INTRODUCTION

In today's digital age, email communication remains a ubiquitous and vital tool for personal and professional correspondence. However, this convenience has also made email platforms a prime target for malicious activities, such as spam and phishing attacks. These threats not only inundate inboxes with unwanted messages but also pose severe risks, including financial fraud and data breaches [1]. Traditional methods of spam and phishing email classification have made substantial progress, yet they still face challenges in terms of accuracy, efficiency, and adaptability to evolving tactics employed by cybercriminals. This research embarks on a journey to enhance the effectiveness of email classification in combating spam and phishing threats by leveraging the power of Artificial Intelligence (AI). Specifically, it focuses on the utilization of AI-based feature selection techniques in conjunction with unsupervised learning algorithms. The fusion of these two approaches aims to achieve more efficient and accurate email classification, thereby providing users and email service providers with improved protection against these malicious activities [2].

The significance of this research is underscored by the sheer magnitude of the spam and phishing problem. According to industry reports, a substantial portion of global email traffic consists of spam, and phishing attacks continue to evolve in sophistication, making them harder to detect using traditional rule-based systems. Consequently, there is a growing need for innovative and adaptive solutions that can keep pace with the evolving tactics of cybercriminals. In this introductory overview, we will delve into the key components and objectives of this research [3]. First, we will explore the challenges associated with spam and phishing email classification, highlighting the limitations of current methods. Next, we will introduce the concept of AI-based feature selection and its potential to enhance the efficiency of email classification. Unsupervised learning algorithms, known for their ability to identify patterns in data without the need for labeled examples, will be discussed as a promising approach for addressing these challenges. The synergy between feature selection and unsupervised learning will be explored, demonstrating how this combination can lead to more robust and adaptable email classification systems. Furthermore, the research will delve into the expected outcomes and benefits of this approach, including improved accuracy in distinguishing between legitimate and malicious emails, reduced false positives, and increased efficiency in processing email traffic [4]. Additionally, the ethical considerations surrounding AI-driven email classification, such as privacy and bias, will be acknowledged and discussed.

In summary, this research sets out to harness the capabilities of AI-based feature selection and unsupervised learning to revolutionize spam and phishing email classification. By addressing the shortcomings of existing methods and adapting to the evolving landscape of email threats, this research endeavors to provide a more secure and efficient email experience for users and organizations alike [5].

2. LITERATURE SURVEY

Manjrekar, et al. [6] proposed Cyber Security Using Machine Learning Techniques. Accounting for cyber security where machine learning is used and using ML to enable cyber security are the two main components of combining cyber security and ML. The authors combine cyber security and ML to address two distinct themes in this survey article. By providing ML strategies for cyber security, the purpose of this paper is to give a wide overview of ML methods employed in cyberspace security.

Al-Ghamdi, et al. [7] proposed Digital Forensics and Machine Learning to Fraudulent Email Prediction. Criminal activity needs to be combated through digital forensics. Unfortunately, cyber events are becoming significantly challenging, and human capabilities are limited. Using the SeFACED dataset, this research proposes a content base, E-mail multi-classification, into four different classes: Normal, Fraudulent, Threatening, and Suspicious, using four primary Machine Learning algorithms, namely Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF).

Abdul Samad, et al. [8] proposed Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection. This work presents the findings of an experimental study that attempted to enhance the performance of machine learning models to obtain improved accuracy for the two phishing datasets that are used the most commonly. Three distinct types of tuning factors are utilized, including data balancing, hyper-parameter optimization and feature selection.

Khadidos, et al. [9] proposed Binary Hunter–Prey Optimization with Machine Learning—Based Cybersecurity Solution in an Internet of Things Environment. this work presents a binary Hunter–prey optimization with a machine learning-based phishing attack detection (BHPO-MLPAD) method in the IoT environment. The BHPO-MLPAD technique can find phishing attacks through feature selection and classification. In the presented BHPO-MLPAD technique, the BHPO algorithm primarily chooses an optimal subset of features. The cascaded forward neural network (CFNN) model is employed for phishing attack detection.

Saraswathi, et al. [10] proposed an integrated machine learning (ML) framework for fraudulent website detection to solve this problem. Artificial neural networks (ANN), support vector machine (SVM), random forests (RF), and K-nearest neighbor (K-NN) are algorithms to detect phishing websites accurately. Some URLs can be used to classify them as appropriate or phishing. Data from publicly available phishing websites can be collected from the UCIrvine ML repository for training and testing.

3. PROPOSED SYSTEM

Phishing email dataset

Phishing is the fraudulent attempt to obtain sensitive information or data, such as usernames, passwords and credit card details, by disguising oneself as a trustworthy entity in an electronic communication. Typically carried out by email spoofing, instant messaging, and text messaging, phishing often directs users to enter personal information at a fake website which matches the look and feel of the legitimate site. Phishing is an example of social engineering techniques used to deceive users. Users are lured by communications purporting to be from trusted parties such as social web sites, auction sites, banks, colleagues/executives, online payment processors or IT administrators. In this challenge our goal is to train a classifier that will detect phishing emails.

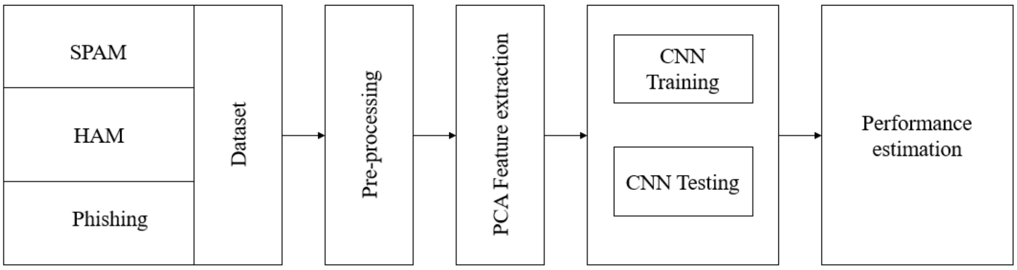


Fig. 1: Block diagram of proposed system.

Data fields

- index - Email Id
- Subject - The Email's headline
- Content - The Email's internal message
- Content-Type - The Email's content format

CNN Classifier

According to the facts, training and testing of CNN involves in allowing every source data via a succession of convolution layers by a kernel or filter, rectified linear unit (ReLU), max pooling, fully connected layer and utilize SoftMax layer with classification layer to categorize the objects with probabilistic values ranging from.

Convolution layer is the primary layer to extract the features from a source image and maintains the relationship between pixels by learning the features of image by employing tiny blocks of source data. It's a mathematical function which considers two inputs like source image $I(x, y, d)$ where x and y denotes the spatial coordinates i.e., number of rows and columns. d is denoted as dimension of an image (here d=3 since the source image is RGB) and a filter or kernel with similar size of input image and can be denoted as $F(k_x, k_y, d)$..

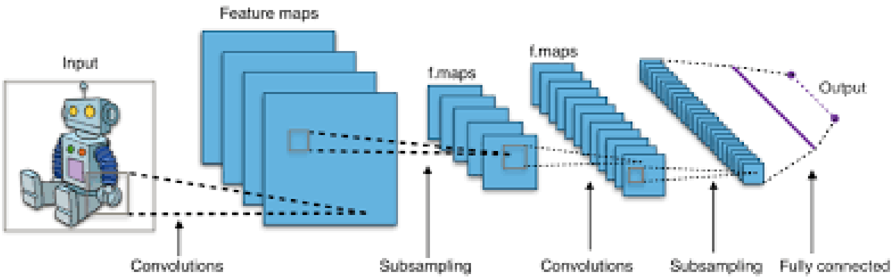
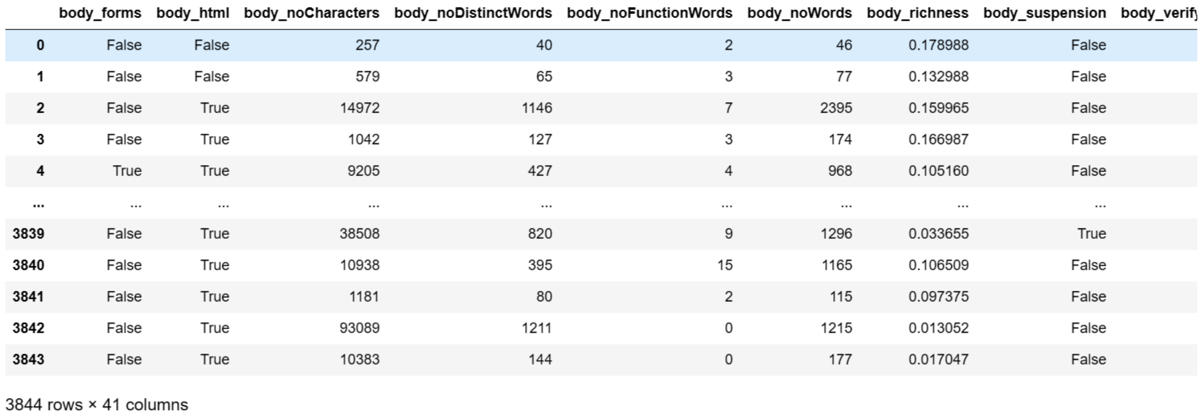


Fig. 2: Representation of convolution layer process.

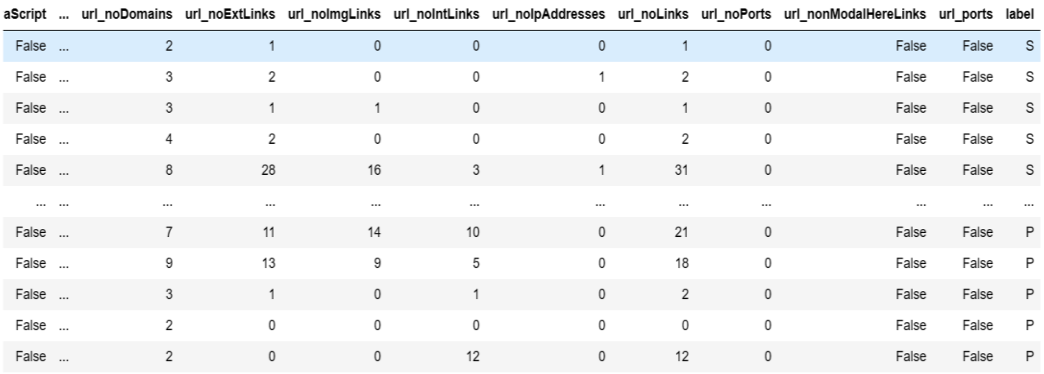
4. RESULTS AND DISCUSSION

Figure 3 shows that the sample dataset



	body_forms	body_html	body_noCharacters	body_noDistinctWords	body_noFunctionWords	body_noWords	body_richness	body_suspension	body_verif
0	False	False	257	40	2	46	0.178988	False	
1	False	False	579	65	3	77	0.132988	False	
2	False	True	14972	1146	7	2395	0.159965	False	
3	False	True	1042	127	3	174	0.166987	False	
4	True	True	9205	427	4	968	0.105160	False	
...	...	...	...	...	...	...	...	...	
3839	False	True	38508	820	9	1296	0.033655	True	
3840	False	True	10938	395	15	1165	0.106509	False	
3841	False	True	1181	80	2	115	0.097375	False	
3842	False	True	93089	1211	0	1215	0.013052	False	
3843	False	True	10383	144	0	177	0.017047	False	

3844 rows × 41 columns



aScript	...	url_noDomains	url_noExtLinks	url_noImgLinks	url_noIntLinks	url_noIpAddresses	url_noLinks	url_noPorts	url_nonModalHereLinks	url_ports	label
False	...	2	1	0	0	0	1	0	False	False	S
False	...	3	2	0	0	1	2	0	False	False	S
False	...	3	1	1	0	0	1	0	False	False	S
False	...	4	2	0	0	0	2	0	False	False	S
False	...	8	28	16	3	1	31	0	False	False	S
...	...	...	...	...	...	...	...	...	...	...	...
False	...	7	11	14	10	0	21	0	False	False	P
False	...	9	13	9	5	0	18	0	False	False	P
False	...	3	1	0	1	0	2	0	False	False	P
False	...	2	0	0	0	0	0	0	False	False	P
False	...	2	0	0	12	0	12	0	False	False	P

Figure 3. Sample dataset

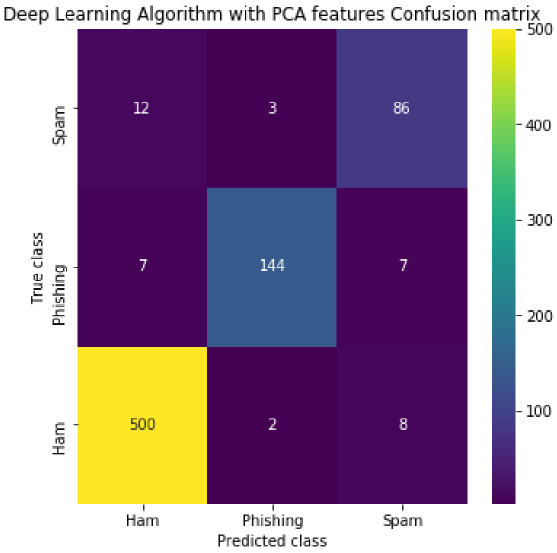


Figure 4. Proposed Confusion matrix.

Figure 4 shows that that Deep Learning with PCA confusion matrix with good TP rate

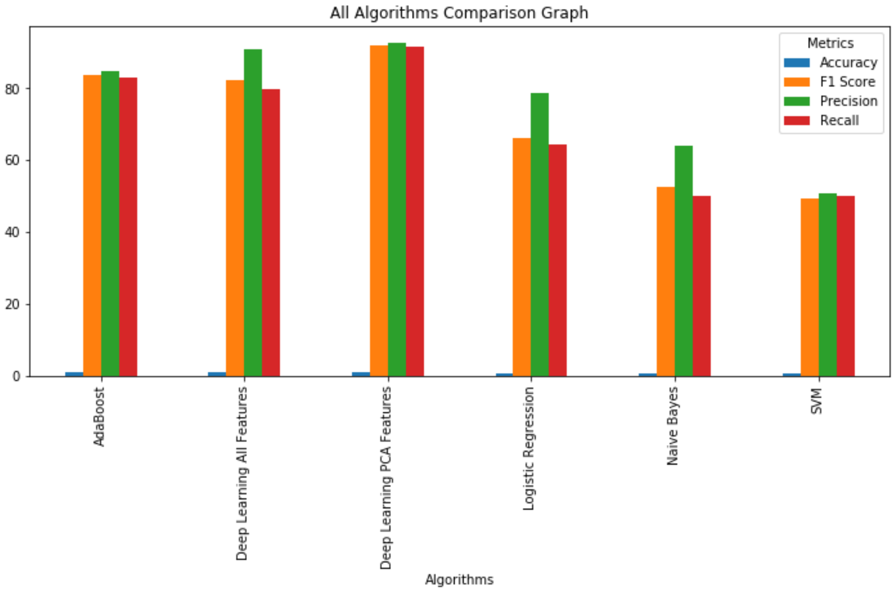
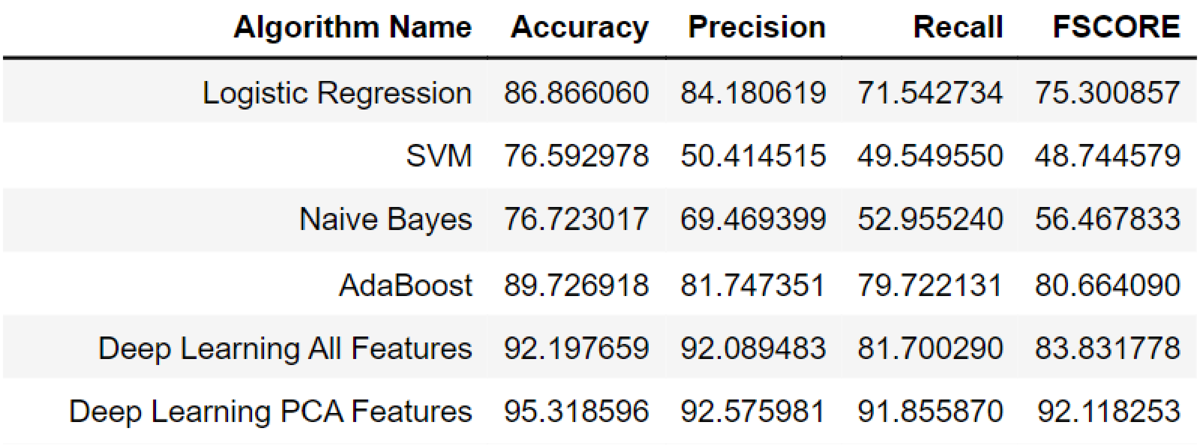


Figure 5 Graphical Performance comparison of various models**.**

Figure 5 is Graphical Performance comparison of various models, in that Deep Learning with PCA features have highest accuracy, precision, recall and F1 score



Algorithm Name	Accuracy	Precision	Recall	FSCORE
Logistic Regression	86.866060	84.180619	71.542734	75.300857
SVM	76.592978	50.414515	49.549550	48.744579
Naive Bayes	76.723017	69.469399	52.955240	56.467833
AdaBoost	89.726918	81.747351	79.722131	80.664090
Deep Learning All Features	92.197659	92.089483	81.700290	83.831778
Deep Learning PCA Features	95.318596	92.575981	91.855870	92.118253

Figure 6: Numerical Performance comparison of various models.

Figure 6 shows that Deep learning with PCA features have a highest accuracy 95.3185, precision 92.575, recall 91.8558 and F-score is 92.1182

5. CONCLUSION

This project employed AI to detect SPAM, HAM, and Phishing emails by applying features selection algorithm called PCA (principal component analysis). All existing algorithms detected only SPAM and HAM emails, but this proposed algorithm designed to detect three different classes called SPAM, HAM, and Phishing. Here, this project combined three different datasets called UCI, CSDMC and SPAM ASSASSIN dataset, where UCI and CSDMC datasets provided SPAM and HAM emails and Spam Assassin dataset provided Phishing emails. All these emails were processed to extract important features used in spam and phishing emails such as JAVA SCRIPTS, HTML tags and other alluring URLS to attract users. In addition, the results obtained using various machine learning models are compared with proposed deep learning with PCA feature selection and proven it has superior classification performance.

REFERENCES

[1] Saeed, M. M., Saeed, R. A., Abdelhaq, M., Alsaqour, R., Hasan, M. K., & Mokhtar, R. A. (2023). Anomaly Detection in 6G Networks Using Machine Learning Methods. Electronics, 12(15), 3300.

[2] Sharma, Banisha, Paurav Goel, and Jaspreet Kaur Grewal. "Advances and Challenges in Cryptography using Artificial Intelligence." 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, 2023.

[3] Greco, Francesco, Giuseppe Desolda, and Andrea Esposito. "Explaining phishing attacks: An XAI approach to enhance user awareness and trust." Proc. of the Italian Conference on CyberSecurity (ITASEC '23). 2023.

[4] Ghaleb Al-Mekhlafi, Z., et al. "Phishing websites detection by using optimized stacking ensemble model." Computer Systems Science and Engineering 41.1 (2022): 109-125.

[5] Manjramkar, Manisha A., and Kalpana C. Jondhale. "Cyber Security Using Machine Learning Techniques." International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022). Atlantis Press, 2023.

[6] Al-Ghamdi, Norah, and Tahani Alsubait. "Digital Forensics and Machine Learning to Fraudulent Email Prediction." 2022 Fifth National Conference of Saudi Computers Colleges (NCCC). IEEE, 2022.

[7] Abdul Samad, Saleem Raja, et al. "Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection." Electronics 12.7 (2023): 1642.

[8] Khadidos, Adil O., et al. "Binary Hunter–Prey Optimization with Machine Learning—Based Cybersecurity Solution on Internet of Things Environment." Sensors 23.16 (2023): 7207.

[9] Saraswathi, P., J. V. Anchitaalagammai, and R. Kavitha. "A System Review on Fraudulent Website Detection Using Machine Learning Technique." SN Computer Science 4.6 (2023): 702.

[10] Muneer, Salman Muneer, Muhammad Bux Alvi, and Amina Farrakh. "Cyber Security Event Detection Using Machine Learning Technique." International Journal of Computational and Innovative Sciences 2.2 (2023): 42-46.