A Late Multi-modal Fusion Model for Detecting Hybrid Spam E-mail

Zhibo Zhang*, Ernesto Damiani, Hussam Hamadi, Chan Yeun, and Fatma Taher

***Abstract*—In recent years, spammers are now trying to obfuscate spam filtering systems by introducing hybrid spam e-mail combining both image and text parts, which is more destructive and complicated compared to e-mails containing text or image only to cyber security. Traditionally, Optical Character Recognition (OCR) technology is used to eliminate the image parts of spam by transforming images into text. Although OCR scanning is a very successful technique for processing text-and-image hybrid spam, it is not an effective solution for dealing with huge quantities due to the Central Processing Unit (CPU) power required and the execution time it takes to scan e-mail files. To address this problem, this paper proposes a late multi-modal fusion model for a text-and-image hybrid spam e-mail filtering system compared to the classical early fusion detection model based on the OCR method. Convolutional Neural Network (CNN) and Continuous Bag of Words were implemented to extract features from image and text parts of hybrid spam respectively, whereas generated features were fed to the sigmoid layer and machine learning based classifiers to determine the e-mail ham or spam. The obtained two classification probability values were fed to a late decision model and the concluding classification decisions were analyzed with text-only classifiers based on the OCR technique in terms of prediction accuracy as well as computational efficiency. The experimental results show that the proposed late fusion model is highly superior to the benchmark in terms of execution time whereas other performance metrics are adequate. These findings reveal the superiorities of using CNN rather than OCR to detect hybrid spam e-mails.**

***Index Terms*—Convolutional neural network, cyber security, hybrid spam e-mail, late fusion, spam filtering**

I. INTRODUCTION

Due to the increased usage of e-mail in day-to-day commercial transactions and general communication, spam, or unwanted commercial mass e-mails, have become a major problem for cyber security in recent years [1]. As for 2021, it translates into an average daily volume of 122.33 billion messages globally, and nearly 85% of all e-mails are spam [2]. Furthermore, the financial loss caused by the spamming attack is staggering, e-mail spam costs businesses $20.5 billion every year [3]. These phenomena highlight the demands to develop real-time spam e-mail detection and classification systems to provide stable and reliable spam e-mail filtering services that fulfill user demands.

Spam e-mail filtering systems have been developed for a few years whereas spam e-mail attackers have adopted several attacking methods that are designed to confuse and degrade the functioning of these filters. Conventionally, spam e-mails are constituted by texts or images only [1]. Furthermore, Zhang *et al.* are focusing on the review of approaches dealing with image spam e-mails [4]. It is noticed that existing spam e-mail filtering systems combatting text or image spam e-mails only are relatively effective in terms of both filtering accuracy and computational efficiency. As a result, spammers devised text-and-image hybrid spam e-mail, examples as shown in Fig. 1 below, a method of combining text and graphics to circumvent the increasing protection of spam e-mail filtering systems. In comparison to image-based only and text-based only spam e-mail, hybrid spam e-mail is more destructive and complicated not only because hybrid spam e-mail contains more harmful materials than spam emails consisting of text or image only but also for the reason that text-and-image hybrid spam e-mails are more challenging to detect and filter [5]. Therefore, it is extremely crucial to develop the intelligent multi-modal fusion spam e-mail detection model aimed at filtering text-and-image hybrid spam e-mail for the sake of cyber security in modern society.

Fig. 1. Examples of hybrid spam e-mail combining text and image.

In response, several anti-spam filtering systems take Optical Character Recognition (OCR) [6] into their consideration to extract and recognize the words embedded in the image parts of hybrid spam e-mails and transform the image parts into text. The extracted text from hybrid spam e-mail will be fed into a pure text-based spam e-mail classifier with the text parts inserted in the original spam e-mail to build an early fusion model, altering the multi-model detection problem to a simple text-based spam e-mail detection task. However, spam e-mail attackers are increasingly attempting to confuse scanners by incorporating more sophisticated graphics and colors, causing the inability of the OCR technique to convert the content in the image to text accurately. Besides, another drawback is that OCR-based hybrid spam e-mail filtering systems consume much bandwidth and might cause the Central Processor Unit (CPU) to overheat [7].

This paper proposes a late multi-modal fusion model for

Manuscript received November 17, 2022; revised December 23, 2022; accepted February 2, 2023.

Zhibo Zhang, Ernesto Damiani, Hussam Hamadi, and Chan Yeun are with Khalifa University, Abu Dhabi, UAE.

Fatma Taher is with Zayed University, Dubai, UAE.

*Correspondence: 100060990@ku.ac.ae (Z.Z.)

DOI: 10.7763/IJCTE.2023.V15.1334

detecting and filtering text-and-image hybrid spam e-mails based on Machine Learning-based classification approaches. Firstly, Convolutional Neural Network (CNN) is deployed for the feature extraction and classification of image parts of hybrid spam emails. Secondly, Natural Language Processing skills including Bag of Words (BoW) and WordVec are used to generate features from the text parts of hybrid spam emails whereas the extracted features were analyzed by some popular Machine Learning-based classifiers encompassing Random Forest (RF), Naïve Bayes (NB) and Support Vector Machine (SVM). Thirdly, a late fusion model is established combining the decisions of both image and text parts of hybrid spam e-mails. Finally, the filtering and classification results on hybrid spam emails of the proposed late fusion model and the early fusion model based on OCR techniques are discussed and measured concerning classification performance and execution time.

Therefore, the main contributions proposed in this paper are listed as follows:

1) Analyzing the necessity of instant and stable services for the classification and filtering of hybrid spam e-mails.
2) Proposing a late multi-modal fusion model for hybrid spam email filtering service systems.
3) Comparing and investigating the proposed late fusion model and conventional OCR-based early fusion model regarding both filtering performance and computational resources.

The rest of this paper is organized as follows: Section II reviews related works focusing on similar topics. Section III introduces the proposed late multi-modal fusion model of text-and-image hybrid spam emails detection. Section IV provides experimentation results and analysis in terms of prediction accuracy as well as execution time and computational resources. Section V concludes this paper and prospects for future work.

II. Related Works

The detection of hybrid spam e-mail is a special case of spam e-mail filtering services. Both image-based and text-based feature extraction and classification techniques are deployed in the application of hybrid filtering systems. Many studies have been conducted to address the issue of text-based or image-based only spam email filtring whereas some researchers contribute to the detection of hybrid spam e-mail in recent years.

For text-based only spam e-mail filtering, many approaches, especially Machine Learning-based methods, have been proposed and indicated good results. Isra *et al.* [8] utilized Deep Neural Network (DNN) model containing a bidirectional Long Short Term Memory layer and compared results with classic classifiers k-nearest neighbors (k-NN) and Naive Bayes (NB). Karim *et al.* [9] proposed an anti-spam system that fully relies on unsupervised methodologies examining only the text extracted from the domain and header information of e-mail considering the requirements of user privacy. Feng *et al.* [10] presented a Support Vector Machine-based Naive Bayes spam email filtering system and achieved a higher spam-detection accuracy and a faster classification speed by eliminating samples located nearby the hyperplane.

For image-based only spam e-mail filtering, Deep Learning-based methods including Convolutional Neural Networks (CNN) have become popular in the field of image spam e-mail detection. Sharmin *et al.* [11] presented an approach to extract image features from image spam e-mails using Convolutional Neural Networks (CNN) and classified generated features using the SVM classifier. Two DCNN models and transfer learning-based pre-trained CNN models were explored for image spam e-mail detection by Sriram *et al.* [12]. A CNN-XGBoost framework consisting of eight layers with samples using data augmentation techniques was built to accomplish image spam detection task by Kim *et al.* [7]. And this research also discussed the high processing cost of Optical Character Recognition (OCR) required for analyzing images. This paper highlighted the weakness of the OCR technique that intelligent spammers can purposefully include unusual text characters into an image, rendering them invisible to standard OCR software but still understandable to human victims.

On the other hand, however, several studies highlight the OCR-based image spam filtering systems designed to scan and read the text and analyze images. Prashant *et al.* [13] proposed a Support Vector Machine (SVM) with Gaussian kernel-based classifier for the detection of image spam on textual features converted using the OCR technique. Based on keyword patterns using the OCR technique, a fax spam detection framework was proposed in [14]. Estqlal *et al.* [15] proposed a hybrid method based on combined feature vectors from text regions and features of the image spam to address the noise in OCR recognization. However, few of the previous studies take the computational resources the OCR technique will consume into consideration.

For text-and-image hybrid spam e-mail filtering, Yang *et al.* [16] presented a new model called multi-modal architecture based on model fusion combining a Convolutional Neural Network (CNN) model and a Long Short-Term Memory (LSTM) model to filter hybrid spam e-mail. Two-hybrid multi-modal architectures by forging the image and text classifiers were proposed to analyze the whole content of e-mails by processing it through independent classifiers using Convolutional Neural Networks [17].

Despite the development accomplished in the area of filtering spam e-mail involving both text and image parts, most studies focus on the improvement of classification accuracy whereas few papers consider computational efficiency. This paper aims to build a late fusion model of hybrid spam e-mail and measures it in terms of filtering precision and execution time compared with the early fusion model using OCR techniques.

III. Methodology

The proposed late fusion hybrid spam e-mail filtering training frameworks consist of mainly three parts: text-part classifier, image-part classifier, and decision fusion model. Decisions made by the text-part classifier and image-part classifier will be fed into the late decision fusion model.

A. Text Classification Model

The text data separated from the original hybrid spam e-mails should be pre-processed before being fed to the Machine Learning based classifiers. Therefore, Natural Language Processing techniques including Stopwords Removal, Lemmatization, and Term Frequency-Inverse Document Frequency (TF-IDF) are deployed to process the text data.

Firstly, the most frequently used words in English, called stop words, are removed for the sake of efficiency. Secondly, Lemmatization is utilized to combine a word's many inflected forms so that they may be analyzed as a single item. For instance, "Include," "includes," and "included" are all examples of words that might be rendered as "include". After that, the weight of a term t_j in document d_i is calculated as the following Eq. (1):

$$w_{ji} = tf_{ji} \times idf_{ji} = tf_{ji} \times \log\left(\frac{N}{df_j}\right) \tag{1}$$

Here, tf_{ji} denotes the times that the term t_j appearing in document d_i, N is the total number of e-mail documents in the e-mail set, whereas df_j represents the number of e-mail documents containing the term t_j. Therefore, each e-mail document could as $d_i = (w_{1i}, w_{2i}, \ldots, w_{Ti})$, whereas T denotes the number of the feature set. Different from the traditional Bag of Words technique that gets rid of word order, the Continuous Bag of Words method takes vector embeddings of n-words before and after the target word.

After the stages of pre-processing and feature generation, the text-based e-mail data are split into two sets: the training set and the testing set. The training set will be used to train the following Machine Learning-based classifiers and the testing set will be utilized to test the performance of the trained classifiers. In this study, the following three Machine Learning techniques are employed:

1) Naïve Bayes (NB) classifier

A Naive Bayes classifier is a simple probabilistic classifier that is based on the Bayes theorem and is based on independent assumptions. The probability model of the Naive Bayes classifier can be shown as the following Eq. (2):

$$P(C/X) = \frac{P(X/C)P(C)}{P(X)} \tag{2}$$

In the above Eq. (2), X denotes a set of function vectors, C stands for a class variable with multiple outcomes, *P(C/X)* denotes the likelihood of something happening in the future, *P(X/C) P(C)* stands for prior likelihood, and *P(X)* denotes the proof based on function variables. In this research, the Laplace Smoothing parameter α is set to 1.0 to solve the zero-probability problem in Naïve Bayes classification. In the classification, some word frequency is zero in some testing samples and the Laplace Smoothing parameter avoids this situation.

2) Support Vector Machine (SVM) classifier

The Support Vector Machine (SVM) is another supervised machine learning algorithm. When having a tiny quantity of labeled data, SVM is the fastest and most reliable classification model. A hyperplane is used in the SVM model to segregate positive and negative values (spam and ham) from the dataset. In this research, LinearSVC is utilized for the classification as the regularization parameters can be set only in LinearSVC. L1 is set as the regularization parameter, hinge is set as the loss function, and the maximum iteration number is set as 1000.

3) Random Forest (RF) classifier

Both classification and regression may be done with the Random Forest technique. The method predicts classes by employing numerous decision trees, each of which predicts a categorization class. The model evaluates this to assign the highest number of predicted classes as the assigned prediction. In this research, the maximum feature number of the Random Forest classifier is set as $\sqrt{N}$ and the maximum depth of the Random Forest classifier is set as 50 to control the depth of the sub-trees.

B. Image Classification Model

In this study, a CNN model is designed for the classification of the image parts of spam e-mails. For the hyperparameters used in the proposed CNN model, the learning rate is set to 0.0001, the optimizer algorithm is chosen as RMSprop, the epoch is set to 30, and the batch size is 20. Before being fed into the CNN model, the image files are decoded from JPG content to RGB grids of pixels 256×256 and converted into floating-point tensors after that. Then, the pixel values between 0 and 255 are rescaled to [0, 1], and the input shape is transformed into (128×128×3). The derived CNN model has four convolution layers of filter sizes 32, 64, 128, and 128 respectively. Finally, using a sigmoid activation function, a dense layer of a single neuron is utilized. The detailed structure of the CNN model is represented with layer details in Table I.

TABLE I: DETAILED ARCHITECTURE OF CNN MODEL

Layer Type	Output Shape	Parameter Number
Conv2D (32, (3,3))	(None, 126, 126, 32)	896
MaxPooling2D ((2,2))	(None, 63, 63, 32)	0
Conv2D (64, (3,3))	(None, 61, 61, 64)	18,496
MaxPooling2D ((2,2))	(None, 30, 30, 64)	0
Conv2D (128, (3,3))	(None, 28, 28, 128)	73,856
MaxPooling2D ((2,2))	(None, 14, 14, 128)	0
Conv2D (128, (5,5))	(None, 10, 10, 128)	409,728
MaxPooling2D ((2,2))	(None, 5, 5, 128)	0
Flatten	(None, 3200, 1, 1)	0
Dense (512)	(None, 512)	1,638,912
Dense (1)	(None, 1)	512

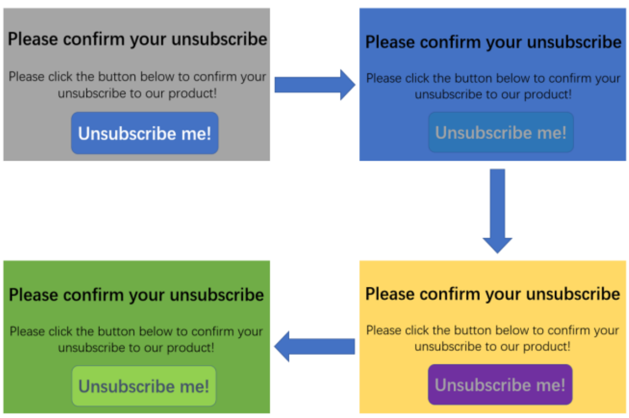

Fig. 2. The feature generation process of the processed CNN model.

As shown in Fig. 2, the feature extraction process of the CNN model on a sample image spam e-mail is visualized. From the figure, it is clear that with increasing layer depth, the images retrieved by the layer become increasingly abstract, meaning that higher layers convey less information about the current specific input and more information about the targets which are used to classify the image e-mail ham or spam.

C. *The Late Fusion Model*

In the late fusion model architecture as shown in Fig. 3, the confidence values determining the hybrid e-mail spam or ham are obtained by taking the probability of the text-based classifiers and image classifiers into consideration.

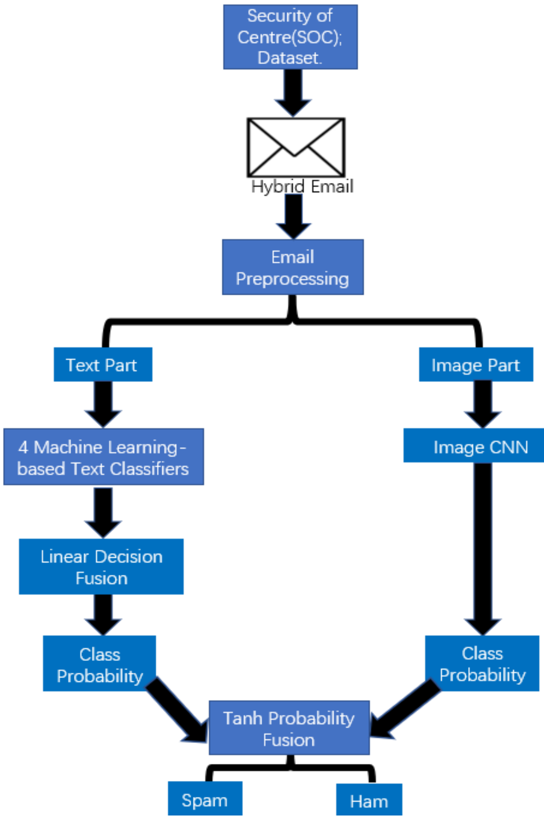

Fig. 3. The proposed late fusion model architecture.

As discussed in the above sections, this model employed three Machine Learning-based classifiers for the classification of the text parts whereas one CNN model was deployed for the classification of the image parts.

However, for the fusion of the text and image, this approach is not appliable because even if the text part contained in the hybrid spam e-mail is purely harmless, this e-mail would still be regarded as a spam e-mail in the context of human cognition if the image part inserted in the hybrid spam e-mail is judged as spam image and vice versa. Therefore, for the decision fusion of the text-based spam e-mail classifier and the image-based spam e-mail classifier, exponential fusion which is a variation function based on the Tanh function rather than linear fusion is proposed for the fusion model, defined as the following Eq. (3):

$$P_{sum} = \begin{cases} P'_{sum} = \frac{2}{1+e^{-2P_{text}}} + \frac{2}{1+e^{-2P_{image}}} - 2, & if\ P'_{sum} < 1 \\ 1 & if\ P'_{sum} \geq 1 \end{cases} \quad (3)$$

In the above equation, P_{text} represents the final probability value of the text-based classifier obtained in Eq. (3) whereas P_{image} demonstrates the confidence value of the image-based classifier. The fusing probability value P_{sum} is ranged from [0, 1] by setting all P'_{sum} greater than 1 to be 1. If $P_{sum} > 0.5$, the hybrid e-mail is regarded as a spam e-mail; otherwise, it is a ham e-mail.

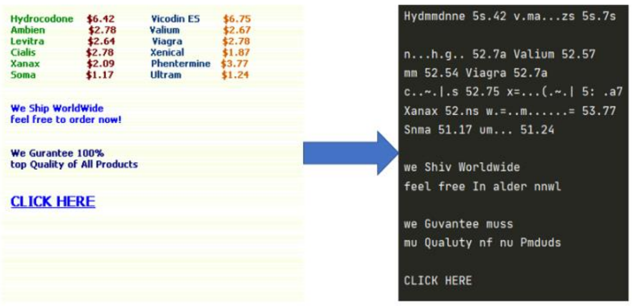

Fig. 4. An example of the OCR technique.

D. *OCR-Based Early Fusion Model*

In the late fusion model architecture, OCR techniques are employed for extracting the textual content embedded into an essentially blank image, so that the traditional text-based spam e-mail filters can be applied afterward. Unsurprisingly, the effectiveness of the early fusion method is highly dependent on the accuracy of the OCR technique deployed. Fig. 4 shows an example of transforming a spam image into pure text content using the Tesseract OCR Engine originally developed by Hewlett-Packard. One disadvantage of the OCR system is that it faults recognition, which can influence negatively the performance of OCR text extraction especially when spam e-mail attackers disguise the content of an image. Moreover, another issue in implementing the OCR method is the high computational complexity that the OCR technique needs.

IV. Experiment Results and Analysis

This paper simulated the performance of the models proposed in Section III in the environment of Python 3.8. The experiment is carried out in the operating system of Windows 10, with 4 cores CPU, 8.00 GB RAM, and 4G GPU.

A. *Data Set*

This paper implements four e-mail datasets for the experiments: two text-based e-mail datasets and two image-based e-mail datasets. The Enron corpus [18] and Ling-Spam Dataset [19] are used as the text-based spam e-mail datasets, whereas SpamArchive Image Spam Dataset [20] and Princeton Spam Image Benchmark Dataset [21] are the sources for the image-based spam e-mail datasets. To show the unstable performance of the OCR techniques, the words-missing text dataset and the words-substitution text dataset are created based on the original text-based spam e-mail datasets. For instance, the word "security" can be changed to "s3cur1ty" in the words-substitution text dataset whereas the words-missing text dataset is self-evident. Below Table II shows the specifics of the datasets utilized in the experiments.

TABLE II: DATASETS USED IN SIMULATION

Type	Original Datasets	Number
Pure Text	Enron Corpus; Ling-Spam Corpus	Ham: 2800 Spam: 2800
Words-substitution Text	Enron Corpus; Ling-Spam Corpus	Ham: 2800 Spam: 2800
Words-missing Text	Enron Corpus; Ling-Spam Corpus	Ham: 2800 Spam: 2800
Pure Image	SpamArchive Image; Princeton Spam Image Benchmark	Ham: 2800 Spam: 4400
Mixed Dataset	SpamArchive Image; Ling-Spam Corpus	Ham: 150 Spam: 150

For the Mixed Dataset, every hybrid e-mail is composed of one image-based e-mail and one text-based e-mail. The size of the Mixed Dataset is relatively small the reason that the Mixed Dataset is used only for testing rather than training.

B. Statistical Metrics

In this section, the execution time is deployed to measure the computational complexity of the proposed models. From the architecture of the early fusion and late fusion models shown in Fig. 2, the execution time of the two models can be defined as the following Eqs. (5), (6) respectively where the execution time for text-classifier, image-classifier, and OCR execution time could be shown directly in the Python environment and N is the number of images:

$$T_{early-fusion} = T_{text-classifier} + N \times T_{OCR} \tag{5}$$

$$T_{late-fusion} = \text{Max}(\, T_{text\ classifier},\, T_{image\ classifier}) \tag{6}$$

In terms of evaluating the performance of the filtering models, the confusion matrix is used where *FP*, *FN*, *TP*, and *TN* are defined as follows. Based on these, the statistical metrics including accuracy, recall, precision, and F1-score are defined as the following Eq. (7)–(10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F1 - Score = \frac{2 \times (Precison \times Recall)}{Precison + Recall} \tag{10}$$

C. Results and Discussion

As shown in Fig. 5, the execution time of the late fusion model and the early fusion model is compared. It is noted from the figure that the execution time that the Tesseract OCR needs to process one image is about 2.2 seconds and this execution time will increase linearly as the number of images to be processed expands. Moreover, although the text-based classifier is faster than the image-based classifier as the amount of images increases, the execution time needed for the early fusion model is still much higher than the late fusion model as the OCR requires time to convert the images to texts. Besides, as the increase of the number of image files, the superiority of the proposed late fusion model is increasingly obvious in terms of execution time. The reasons behind this are that the proposed late fusion model would spend fixed execution time for the training processes of the CNN model and text-based Machine Learning models whereas the benchmark early fusion model would spend most execution time on implementing the OCR techniques. The great differences between the proposed late fusion methods and the late fusion models using the OCR technique are because of the great differences between the OCR technique and CNN model in terms of execution time in the testing stage. CNN models consume much execution in the training stage rather than the testing stage and need to be trained only once. Other than that, the noise of OCR recognization introduced in [15] can be avoided by the proposed late fusion model as well. Rather than deploying keyword patterns in the OCR technique [14], the CNN classifier employed in the proposed late fusion model can classify the image parts of hybrid spam e-mails more efficiently.

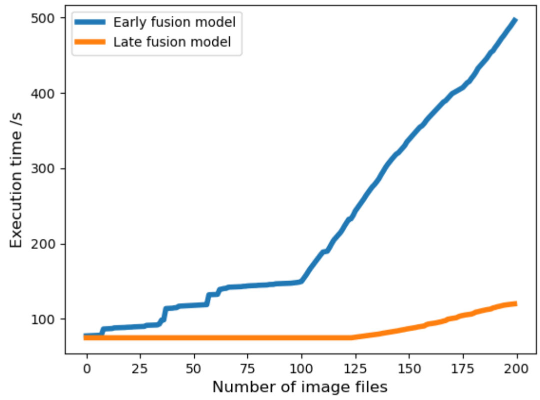

Fig. 5. Execution time for the late fusion and early fusion model.

As shown in Table III, although the words-missing text dataset does not influence the performance of the pure text classifier so greatly, the results for implementing the words-substitution text dataset show a tendency to degenerate compared to the original text dataset-based classification. This phenomenon highlights the fact that a weak OCR performance affects negatively the classification of text-based spam e-mail, demonstrating the deficiency of the OCR-based early fusion benchmark. On the other hand, the performance decay of the proposed late fusion model is in an acceptable range. From Table III, it can also be concluded that the words-changing text datasets and words-missing text datasets, which are simulating the behaviors of the defects of the OCR techniques, can reduce the performance of the text classifier based on the three machine learning techniques in terms of different performance metrics including accuracy, recall, precision, and F1-score. And for the puree image datasets, the utilized CNN-based image classifier showed acceptable results compared with previous research including [11, 12] with 98.9% recall. On the other hand, the image classifier achieved similar performance using the CNN model only rather than using the combination of the Machine Learning model XGBoost and CNN model [7]. For the mixed dataset with both hybrid spam emails, the proposed late fusion classifier achieved 97.4% recall compared with 97.1% recall in [16] and 96.4% recall in [17] whereas the proposed late fusion model utilized no OCR techniques and therefore cost less computing resources. Besides, the noise introduced by the OCR technique in [14, 15] can also be eliminated by the proposed late fusion model.

TABLE III: COMPARISON OF EXPERIMENTAL RESULTS

Dataset	Classifier	Accuracy	Recall	Precision	F1-Score
Pure Text	Text Classifier	96.2%	95.4%	96.8%	96.1%
Words-changing Text	Text Classifier	87.3%	89.4%	86.5%	87.9%
Words-missing Text	Text Classifier	92.4%	91.2%	94.3%	92.7%
Pure Image	Image Classifier	91.2%	98.9%	86.7%	92.4%
Mixed Dataset	Late Fusion	92.3%	97.4%	86.5%	91.6%

V. CONCLUSION

To avoid the hybrid spam e-mail evading from the conventional text-based spam e-mail filters and provide a robotic and real-time filtering service in the cyber security area, this paper introduces a late multi-modal fusion model for a text-and-image hybrid spam e-mail filtering system. Compared to the OCR technique-based early fusion benchmark model, the proposed late fusion model is highly superior to the benchmark in terms of the execution time whereas other performance metrics are adequate. Although the proposed methods show similar results in terms of accuracy with previous related works using OCR techniques, the developed results achieved great superiority in terms of execution time to methods using OCR techniques. In future work, more sophisticated hybrid spam e-mail models, for instance, the punctured e-mails in which images are inserted in the interval of the texts, will be taken into consideration. Besides, although measuring execution time is forthright for providing services, it is not a decent way to evaluate the computational complexity for reason that different instructions executing in the same period vary a lot in terms of CPU resource consumption. Therefore, another issue that will be solved in the future is the cost of the CPU resources that the late fusion model would consume.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Zhibo Zhang and Ernesto Damiani conducted the research; Zhibo Zhang, Hussam Hamadi, and Fatma Taher analyzed the data; Zhibo Zhang and Chan Yeun wrote the paper. All authors had approved the final version.

ACKNOWLEDGMENT

Zhibo Zhang thanks the help of all authors of this work and the Center for Cyber-Physical Systems (C2PS) of Khalifa University.

REFERENCES

[1] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, p. e01802, 2019. doi: 10.1016/j.heliyon.2019.e01802

[2] What's on the other side of your inbox - 20 SPAM statistics for 2022. *Dataprot*. [Online]. Available: https://dataprot.net/statistics/spam-statistics/

[3] M. Bassiouni, M. Ali, and E. A. El-Dahshan, "Ham and spam e-mails classification using machine learning techniques," *Journal of Applied Security Research*, vol. 13, no. 3, pp. 315–331, Jul. 2018.

[4] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence applications in cyber security: State-of-the-art in research," *IEEE Access*, vol. 10, pp. 93104–93139, 2022.

[5] G. Hnini, J. Riffi, M. A. Mahraz, A. Yahyaouy, and H. Tairi, "MMPC-RF: A deep multimodal feature-level fusion architecture for hybrid spam e-mail detection," *Applied Sciences*, vol. 11, no. 24, Jan. 2021.

[6] C. Xu, K. Chiew, Y. Chen, and J. Liu, "Fusion of text and image features: A new approach to image spam filtering," *Advances in Intelligent and Soft Computing*, vol. 124, Jan. 2011. doi: 10.1007/978-3-642-25658-5_15

[7] B. Kim, S. Abuadbba, and H. Kim, "DeepCapture: Image spam detection using deep learning and data augmentation," in *Proc. ACISP*, 2020.

[8] I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," *Procedia Computer Science*, vol. 184, pp. 853–858, Jan. 2021. doi: 10.1016/j.procs.2021.03.107

[9] A. Karim, S. Azam, B. Shanmugam, and K. Kannoorpatti, "Efficient clustering of emails into spam and ham: The foundational study of a comprehensive unsupervised framework," *IEEE Access*, vol. 8, pp. 154759–154788, 2020. doi: 10.1109/ACCESS.2020.3017082

[10] W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, "A support vector machine based naive Bayes algorithm for spam filtering," in *Proc. 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)*, 2016, pp. 1–8. doi: 10.1109/PCCC.2016.7820655

[11] T. Sharmin, F. D. Troia, K. Potika, and M. Stamp, "Convolutional neural networks for image spam detection," *Information Security Journal: A Global Perspective*, vol. 29, no. 3, pp. 103–117, May 2020.

[12] S. Srinivasan *et al.*, "Deep convolutional neural network based image spam classification," in *Proc. 2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, Mar. 2020, pp. 112–117.

[13] P. Kumar and M. Biswas, "SVM with Gaussian kernel-based image spam detection on textual features," in *Proc. 2017 3rd International Conference on Computational Intelligence Communication Technology (CICT)*, 2017, pp. 1–6. doi: 10.1109/CIACT.2017.7977283

[14] J. Kim, H. Kim, and J.-H. Lee, "Analysis and comparison of fax spam detection algorithms," in *Proc. the 11th International Conference on Ubiquitous Information Management and Communication*, New York, NY, USA, 2017, pp. 1–4. doi: 10.1145/3022227.3022284

[15] E. H. Dhah, M. A. Naser, and S. A. Ali, "Spam email image classification based on text and image features," in *Proc. 2019 First International Conference of Computer and Applied Sciences (CAS)*, 2019, pp. 148–153. doi: 10.1109/CAS47993.2019.9075725

[16] H. Yang, Q. Liu, S. Zhou, and Y. Luo, "A spam filtering method based on multi-modal fusion," *Applied Sciences*, vol. 9, no. 6, Jan. 2019. doi: 10.3390/app9061152.

[17] S. Seth and S. Biswas, "Multimodal spam classification using deep learning techniques," in *Proc. 2017 13th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2017, pp. 346–349.

[18] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Proc. Machine Learning: ECML 2004*, Berlin, Heidelberg, 2004, pp. 217–226. doi: 10.1007/978-3-540-30115-8_22

[19] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists," *Information Retrieval*, vol. 6, no. 1, pp. 49–73, Jan. 2003. doi: 10.1023/A:1022948414856

[20] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded into images," *J. Mach. Learn. Res.*, vol. 7, pp. 2699–2720, 2006.

[21] Princeton spam image benchmark. (2007). Princeton University Computer Science Department. [Online]. Available: https://www.cs.princeton.edu/cass/spam/