[1] **Pramod P. Ghogare**
[2] **Husain H. Dawoodi**
[3] **Manoj P. Patil**

A Multi-Dimensional Strategy for Spam Email Classification Leveraging Origin, Text, and Image Features in a Hybrid Model

Journal of Electrical Systems

Abstract: - This article proposes a novel method of integration of an origin, content, and image-based approach for spam email classification. A machine learning classifier takes each extracted feature from the email after a fine-tuned, customized Natural Language Processing (NLP) for spam email classification. A weight is assigned to each feature's classification result and final classification for spam emails is determined by considering all the feature weights. The proposed model demonstrates outstanding classification ability, achieving an impressive accuracy greater than that of the existing personal email provider and reducing substantial false positives compared to individual feature-based classification. The proposed hybrid model excels across accuracy, recall, precision, and f-score, underscoring its comprehensive effectiveness in classification tasks compared to the classification achieved by using each section of an email separately. To provide accurate classification with the least false positives, it is helpful to consider respective features from multiple sections of an email. While previous research focused on the individual section of the email and considered a few features simultaneously, this article proposes a novel approach to classifying spam email by considering significant features from the email's origin, content, and image with integration.

Keywords: Spam, Feature Extraction, Classification, Machine Learning, Natural Language Processing.

I. INTRODUCTION

As spam emails continue to evolve in complexity, it becomes vital for spam filters to adapt and effectively address these challenges. Managing these unwanted emails is crucial to saving time and resources. Spammers employ various tactics to engage users and enhance the effectiveness of spam emails. One such strategy involves the utilization of images to capture the attention of recipients. This presents a challenge for email providers in detecting and blocking spam emails, resulting in some of these deceptive emails infiltrating users' inboxes. While origin-based and content-based techniques are effective for identifying text-based spam, image-based spam poses unique challenges that necessitate more advanced image processing and machine learning methods with integration to existing methods.

II. RELATED WORK

The email header holds significant details about the sender and email source, serving as a valuable resource for effective spam email classification. This section of an email comprises metadata that offers evidence and insights into the email's origin and its journey through the source to destination. The origin-based classifiers utilize email source or network information to classify spam emails. The email sender's information and metadata are cross-referenced with the history of known spam senders. If the incoming email details align with spam-related characteristics, the email is labelled as spam [1]. The features such as IP address, e-mail address, e-mail's subject line, recipients, sender validity, IP address, can be used for classifying spam e-mails [2]. One of the properties of spam email is that major spam e-mails do not have valid sender addresses [3]. When it comes to advantages of origin-based spam email classification, it reduces the cost of computation of email content and does not consume too many system resources [4]. The header feature is language-agnostic and avoids the costly language processing of email content, leading to faster classification [5]. Email classification based on origin depends on sender details like email addresses or domains yet encounters difficulties in spam emails classification. Spammers frequently mask their origins using authentic-looking domains or compromised accounts, which results in misclassifications. The solution is the content-based spam email classification. In their research on spam email classification, Sharma et al. found that content-based filters demonstrate superior effectiveness when compared to origin-based filters [6]. This method checks email content, possibly ignoring spam with seemingly normal origins but malicious content. Additionally, it fights with evolving spam threats as it relies on historical data and known sources, diminishing its effectiveness against zero-day attacks. The content-based approach primarily involves analysing textual content for specific keywords or phrases frequently found in spam emails. The presence of these keywords may indicate the probability of spam email. Many spam emails employ HTML to hide malicious content or tracking elements.

[1, 2, 3] School of Computer Sciences, KBC North Maharashtra University, Jalgaon, Maharashtra, India.
[1*] pramod.ghogare@yahoo.com, [2] hhdawoodi@nmu.ac.in, [3] mppatil@nmu.ac.in

Analysing the HTML structure and content can unveil suspicious elements within the email. Although this method proves effective in detecting obvious and incorrectly covered spam emails, it runs the risk of generating false positives or examining more sophisticated spam emails. As spammers consistently evolve their tactics, it becomes imperative to integrate content-based approaches with other techniques like origin analysis, text analysis, and visual analysis. This combination contributes to the development of a more robust and accurate spam email classification system. Employing a multi-layered approach that incorporates diverse methods can substantially enhance the overall efficiency and effectiveness of spam detection [7].

Until now, we've covered origin and content-based spam classifiers. Now, let's shift focus to the latest and widely utilized type of spam, which is image-based spam emails. Even though creating image spam is tricky, it has advantages over text-based spam. However, numerous spammers have resorted to sending spam emails, incorporating images, to bypass text-based spam filtering software [8]. Spammers can easily change background color, text color, and embedded text in images using software due to which recognizing the information in images is harder than with text. Wu et al. tackled this problem by using three different methods to classify image spam: looking at the text, looking at the visuals, and using both text and visuals together [9].

Krasser et al. presented a way to efficiently and affordably extract important features and classify images. A method that depends on the image format to find important details, making it better at classifying types of images. This way, the time saved and make the classifying work better overall [10]. Chen et al. introduced a framework to find familiar sources of spam emails. It examines the visual features of images depending on their types, the grouping works by comparing how similar these features are, using things like shared topics or IP addresses. By combining they reveal where spam usually comes from. What's interesting was that the suggested method can work with different kinds of features without needing a set limit previously [11]. Das et al. brought together several classifiers to get information from both images and text. The results from these combined classifiers help in making the final decision about whether it's spam or not. It allows for a more thorough analysis and accuracy of spam email classification [12].

In the ever-evolving field of detecting spam emails, researchers are constantly exploring new methods to filter out spam emails. The challenge lies in outsmarting spammers who regularly devise new tactics to bypass filters. Machine learning plays a vital role in this process, training programs to classify between spam and non-spam emails based on extensive datasets. This automation allows for swift and accurate real-time identification of spam. Implementing machine learning reduces errors, strengthens email security, and enhances the user experience [13].

Non-machine learning methods often depend on preset rules, heuristics, or fixed filtering methods. While these approaches can provide some spam filtering, they are usually less flexible when dealing with new and evolving spamming methods. Keeping up with the dynamic nature of spam emails can be challenging for non-machine learning methods, leading to higher rates of false positives or false negatives. Machine learning methods excel in adaptability and learning capability from new data, allowing them to identify patterns and features not explicitly defined by human-crafted rules. This adaptability makes machine learning highly effective in spam email classification, enabling them to recognize intricate and subtle patterns indicative of spam.

Machine learning involves training machines to observe, understand, and represent information about numerical phenomena, detecting irregularities in provided data. Various machine learning algorithms include the artificial immune system, C4.5, Maximum Entropy Model, Support Vector Machine (SVM), Memory-based learning, Bayesian methods, genetic algorithms, clustering techniques, Artificial Neural Network (ANN), Perceptron, Naïve Bayes (NB), and Random Forest (RF) [14]. In spam email detection, machine learning methods outperform non-machine learning techniques by efficiently learning patterns and characteristics from large datasets. These algorithms can automatically generalize and accurately predict new, unseen emails. Artificial intelligence (AI) techniques, particularly computer vision algorithms, have demonstrated remarkable success in various fields, excelling in image synthesis, motion recognition, emotion identification, image summarization, and object detection [15].

In machine learning, classifier selection is crucial for achieving greater accuracy and minimizing misclassification. In the context of extensive datasets for spam email classification, RF has proven effective,

delivering improved results with lower error rates and higher precision compared to alternative methods [16]. The effectiveness of spam filters depends on specific factors like implementation, training dataset size and quality, features used, and ongoing updates to filtering mechanisms. Researchers consistently work on enhancing and combining filtering techniques to stay ahead of evolving spam tactics, ensuring a safer email experience for users. Various factors, including algorithm choice, data size, quality, preprocessing, feature selection, and decision criteria, influence machine learning techniques for spam email filtering. Youn et al. evaluated algorithms like Neural Network, Support Vector Machine, Naive Bayes, and J48 for spam filtering, discovering that even a small number of elements can be useful for classification, while including all features may negatively impact performance. The quality of the training data is crucial for the classifier's effectiveness, requiring inclusion of significant terms and possible distributions within the class for an ideal dataset [17]. When Urmi et al. employed machine learning to examine SMS spam detection, they found that the RF classifier works best for classifying spam emails as well as SMS spam. Out of the five machine learning classifiers that the authors evaluated, RF yielded the best accuracy [18].

In spam email classification, data preparation is important which involves preprocessing. Various methods, including stopping, stemming, lemmatization, normalization, HTML tag removal, punctuation mark removal, and special characters' removal, are used for text and spam email classification. HaCohen-Kerner et al. conducted comprehensive text classification experiments using different preprocessing methods on four datasets. They evaluated the impact of methods like rectifying misspelled words, translating them to lowercase, eliminating HTML objects, removing punctuation marks, eliminating stop words, and eliminating duplicated declining characters [19].

Now, let's discuss the benefits of integrating multiple methods for spam email classification. Analysing both image and text properties is crucial since spammers often use images to spread spam. This study proposes a method categorizing spam emails based on origin, content, and images features. Combining algorithms proves valuable in improving both false positive and detection rates in email classification. Using the combination approach reduced false positives by about 3%, and the detection rate improved by approximately 4% compared to the best individual classifiers. This underscores the potential of combining algorithms for better overall performance in email classification tasks [20].

Byun et al. introduced a framework that combines a text-based anti-spam filter with an image-based filter, enhancing spam detection. Evaluating it on TREC 2005 and 2007 spam corpora, they observed reduced discrepancies between training and test data, indicating better generalization. Integrating image-based filters with text-based ones improved the overall filtering process, offering a more robust solution to combat spam in emails containing both text and images [21]. Bansod et al. applied a neural network to classify both text-based and image-based spam emails. They used various techniques, including blacklisting, whitelisting, word extraction from images, stemming, stop word removal, term frequency calculation, and weight measures for text-based email classification. For emails with text in images, they used ANN to extract the text and conducted classification similarly to text-based emails. This combined approach aimed for efficient spam email classification, regardless of content format [22]. Similarly, it was found that blending origin-based filters with content-based filters improves the efficacy of the filtration technique, rendering it more adaptable and capable of addressing evolving spamming patterns [23].

When Kumar et al. combined the Bayes classification method with techniques like text parsing, word tokenization, and stop words removal to extract features from text and images, they discovered that the suggested hybrid method performed better than the SVM and Neural Network (NN) methods. This demonstrates how well the hybrid technique works to distinguish between phishing and authentic emails [24].

Gao et al. introduced a model that integrates a hybrid machine learning approach, commencing with a thorough text preprocessing phase involving tasks like word tokenization, stop word removal, and feature vector creation. A spam detection classifier was done using a hybrid model incorporating three popular machine learning techniques: SVM, ANN, and RF. Experimental findings show that each individual model performs well in spam detection. Nonetheless, the hybrid model outperforms them in terms of effectiveness, underscoring the superiority of this combined approach [25]. Zhang et al. addressed the challenge of detecting spam emails containing both text and

image components, showing notable improvements in execution time compared to traditional OCR-based methods but without a significant enhancement in accuracy [26].

Employing a hybrid method for classifying emails as spam is advantageous as it allows for adaptation to emerging spamming techniques. With technological advancements, spammers employ diverse tactics, including incorporating images into spam emails. However, many studies solely focus on analysing the text content, origin, or image content of emails, inspecting the role of all these features together in spam classification. This creates a gap in understanding how spammers leverage images to propagate spam.

This paper introduces a method for analysing multiple sections of emails, including sender/origin, content, and image features, with the aim of improving the classification of spam emails exhibiting diverse anomaly patterns. The primary objective is to minimize the misclassification of non-spam emails. The novelty of the research lies in leveraging origin, content, and image features, with testing conducted on the latest email definitions sourced from personal Yahoo mailbox, thus addressing the limitation of earlier studies relying on publicly available datasets. The primary goal of the proposed hybrid framework is to achieve a balanced and effective solution, addressing both efficiency and accuracy in spam email classification.

The rest of the article is organized as follows, section III describes the proposed framework for hybrid spam email classification using origin, content, and image features. Section IV gives details of the data set used for the experiment. Section V discusses the results of the experiments. Section VI concludes the article.

III. PROPOSED FRAMEWORK

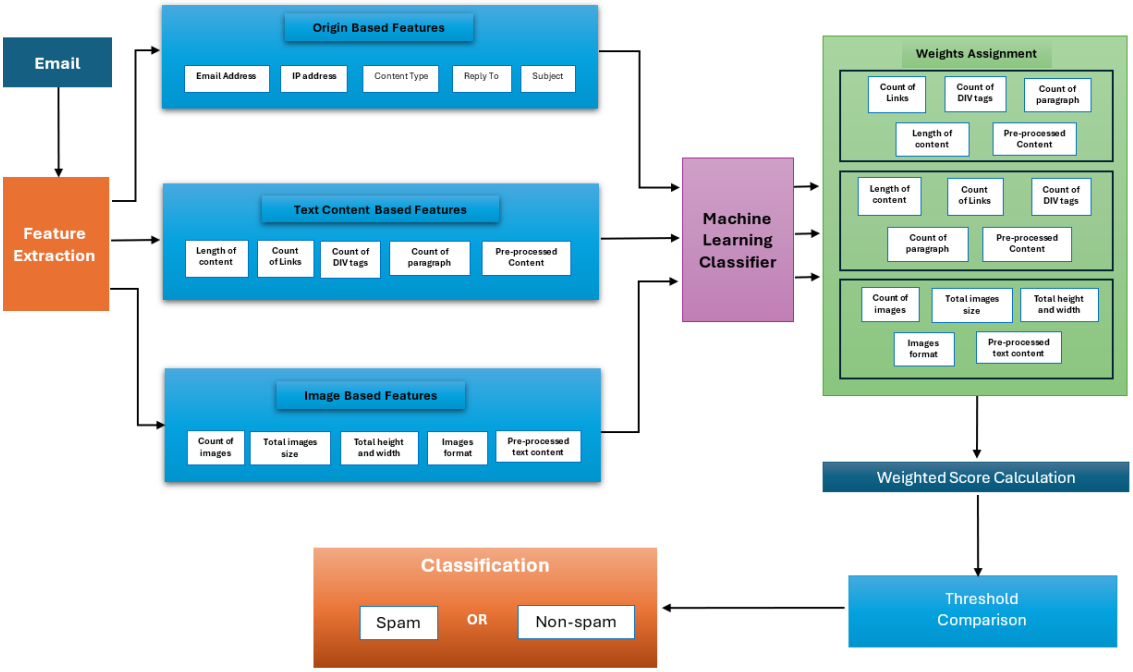

Figure 1 Proposed Spam Email Classification Hybrid Framework

The proposed framework operates cohesively through three discernible phases: origin, content, and image-based classification, collectively contributing to the holistic determination of classification spam emails, refer to Fig. 1. The origin-based classification uses email source information embedded within the email header. Transitioning to the subsequent phase, encompassing content-based classification, it bifurcates into two sub-phases: content-based and image-based classification. Text-based classification involves the extraction of email content features, while image-based classification employs image features.

In the conclusive phase, results from the origin, content and image classifiers synergistically inform the ultimate decision, considering the outcomes of the sub-models. To formulate this ultimate decision, weights are assigned to each feature's output, and an aggregate is computed based on these assigned weights. Subsequently, this aggregate

undergoes a comparison against a predefined threshold. If the aggregate surpasses the threshold, the email is classified as spam; otherwise, it is categorized as non-spam.

Distinguishing itself from prior approaches, the proposed framework uniquely addresses three pivotal aspects of an email for classification: origin, content, and image. This novelty from traditional methods, which often employ multiple classifiers or not concentrate on features from origin, content and images simultaneously, underscores the framework's comprehensiveness. Unlike preceding studies reliant on limited datasets, proposed framework prioritizes the utilization of contemporary emails from personal accounts for training and testing, recognizing the imperative need to adapt to spamming techniques. While previous experiments combined content and image-based classifiers using merged datasets from disparate sources, proposed framework employs complete emails, encompassing origin, content, and image features within a singular dataset for a more exhaustive analysis. Moreover, the proposed framework seeks to achieve a balanced and effective solution by addressing both efficiency and accuracy considerations.

IV. DATASET

Finding up-to-date, comprehensive datasets with both text and images is a big difficulty in spam email classification, since it makes testing and training models on the most recent spam email forms very challenging. When it comes to emails with images, which are increasingly being utilized by spammers, existing datasets sometimes lack diversity. Further complicating the gaining of fresh data are worries about data privacy, which forces previous researchers to depend on a small number of old or poorly sampled spam emails datasets. Classification frameworks' capacity to classify current spam emails may be hampered by these constraints.

To address these challenges, we employed a personal email dataset sourced from a mailbox to evaluate the proposed spam classification method. This dataset, which includes both text-based and image-containing emails, provides valuable insights into the real-world complexities of spam email detection. Its authenticity makes it especially useful for testing, as it reflects the practical difficulties encountered in email classification. The inclusion of emails with images—a feature lacking in many public datasets—adds depth to the analysis, making the results more meaningful, particularly when compared to other available email datasets and the proposed framework's performance.

The dataset is intentionally designed to encompass a wide variety of personal communications, from informal exchanges to event coordination and notifications. This variety images the complex dynamics of modern email interactions and provides a more representative sample for evaluating spam classification models. Additionally, the dataset includes key metadata such as email body, subject, sender and recipient details, and timestamps, which are crucial for thorough analysis.

The unavailability of a standardized dataset has led to the development of a novel, and scalable dataset that strengthens proposed spam classification framework and contributes significantly to the broader research community.

Table 1 Dataset Distribution

Type of Email	**Training**	**Testing**	**Total**
Spam	6240	1560	7800
Ham	7744	1936	9680
	13984	3496	17480

A. Origin-based sub-model:

The sub-model relies on key sender information from the email header, such as the email address, IP address, content type, reply-to, and subject line, to extract characteristics suggestive of spam. A detailed examination of these components helps find potential spam indicators, including illogical letter sequences, additional digits, or generic titles, and assesses how similar they are to commonly used spam addresses.

To save storage space and processing time, pre-processing email subject lines is done. Next, data transformation makes sure that the features are in a format that is appropriate for machine learning classifiers. For origin-based classification, each feature's data frame and email class (spam or non-spam) are used. The sender-related features are examined by the RF machine learning classifier, which looks for trends that suggest spammy activity. With the use of a labelled dataset, the model is trained to learn the connections between spam classification and sender information. The results are then saved for use in further stages of integration.

Table 2 Origin Features

Origin Features
Sender's email address
IP address
Content type
Reply-to
Subject

B. Content-based sub-model:

This sub-model within the framework relies on the content extracted from the email's body, encompassing steps such as text content extraction, preprocessing, natural language processing, transformation, and classification. The illustrated classification mechanism in Fig.1 operates sequentially as follows.

The email body is used by the content-based spam filter to extract elements including words, characters, HTML tags, special symbols, and punctuation. Pre-processing is also included for improved classification. Shorter emails with more connections to external websites, plenty of DIV components (which denote HTML format), and fewer paragraphs with more images are characteristics of spam emails. Email content is pre-processed to extract high-quality data, including reduction, transformation, and cleaning of the data, to improve decision-making. To improve processing speed and storage requirements, this phase entails deleting special symbols, converting text to lowercase, and removing punctuation and empty letters. It has been demonstrated that pre-processed text performs better in classification tasks, guaranteeing that emails are converted into a format that is appropriate for efficient analysis.

NLP is employed to understand, analyse, and interpret textual data, extracting meaningful features for accurate classification. Pre-processed text undergoes lemmatization to identify word lemmas based on their meanings, enhancing machine learning classifiers' accuracy. The data transformation process formats feature in a suitable manner for machine learning classifiers, considering indicators like the number of links, HTML tags, and content length, which are significant for identifying spam. Automated spam emails often contain many HTML elements and longer texts, while manually typed emails are shorter with fewer HTML tags. During classification, features are standardized into vectors used by the RF algorithm to categorize emails as spam or not. The outcome is expressed as weights, with 0.05 or 0.10 assigned for spam indicators and 0.0 for non-spam, utilized in the final integration step.

Table 3 Feature for Content-Based Classification

Content Features
Length of content
Count of Links
Count of DIV tags
Count of paragraph
Pre-processed Content

C. Image-based sub-model:

The sub-model illustrated in Fig. 1 operates through key stages like image extraction, text and feature extraction from images, pre-processing, transformation, and classification. Modern spamming methods embed images as URLs within emails to maintain a small size and manage images efficiently. These images load dynamically when

the user opens the email, undergoing text extraction. Image tags in the email help determine the total number of images. Features such as size, height, and width of each image are recorded and detailed in Table 4.

Table 4 Features for Image-Based Classification

Image Features
Total count of images
The total size of the image
Total height and width of all images
All images format
Pre-processed text in all images

The feature for the total number of images in an email is crucial since non-spam emails usually consist mostly of text and attachments rather than inline images. Spammers use images within the email body to bypass filters, often splitting a single image into multiple ones to evade detection. The larger and more colourful images are typically used by spammers compared to non-spam email senders, justifying the inclusion of image size as a feature.

To determine the total height and width of all images, the width and height of each image are summed. These sums are crucial for classification. The image format also matters, as spammers and legitimate senders often use different software. Text detection techniques identify regions within images containing text, with Optical Character Recognition (OCR) algorithms converting these visual elements into text data. Any recognized text errors are corrected using methods like spell-checking and context analysis. Lastly, lemmatization is applied to the text extracted from images to refine and standardize it for better classification.

$$W_{total} = \sum_{i=1}^{n} w_i \quad (1)$$

$$\mathrm{H}_{total} = \sum_{i=1}^{n} h_i \quad (2)$$

The features extracted from images in email bodies are organized into a feature vector for classification. Using a labelled dataset, the classifier is trained with the RF algorithm to differentiate legitimate emails from spam. The results from the image-based classifier are stored in the same vector used by the content-based classifier. This combined vector is crucial for the final integration stage, where outcomes from all features are compared to determine if an email is spam.

D. Integration of Models

The primary innovation of the proposed framework is the integration that follows.

The classifier results for each feature's vector from the individual sub-models have been consolidated into a single feature vector. If the received email is classified as spam, a weight of 0.10 is assigned to each feature outcome, while non-spam emails have a weight of 0.00. During the integration phase, the results from the origin-based, content-based, and image-based sub-models are all considered and assigned weights.

Table 5 Sample of the integrated feature vector

Feature (x)
Sender Email Address
Sender IP Address
Subject Line
In-Reply-To
Content-Type
Length of actual content
Count of Links
Count of DIV tags
Count of paragraph
Pre-processed Content

Total count of images
The total size of the image
Total height and width of all images
All images format
Pre-processed text in all images

After categorizing each feature, a weighted vector is constructed for classification of spam emails. The formula utilized for the overall decision is articulated as follows:

$$f(x) = \sum_{i=1}^{15} W_i \cdot x_i \quad (3)$$

Equation (3) signifies a linear combination of input variables x_i with their corresponding weights W_i. This linear function is utilized to decide regarding the classification of spam emails, considering each feature and its associated weight.

In this context:

x represents the vector of features $x_1, x_2, x_3, x_4, \ldots\ldots, x_{15}$.

W represents the weights of each feature respectively $W_1, W_2, W_3, W_4, \ldots\ldots, W_{15}$.

In instances where the classifier identifies an email as spam, a weight of 0.05 or 0.10 is assigned to the respective feature. On the other hand, a weight of 0.0 is given if the email is not categorized as spam.

To find out the overall classification outcome, the summation of all weights within this vector is computed. Should the cumulative weight meet or surpass a predefined threshold, set at 0.70 in this instance, the email is classified as spam; otherwise, it is categorized as non-spam. This methodical approach ensures a thorough evaluation, accounting for the weighted contributions of diverse features to render a robust decision regarding the spam classification of the email.

Table 6 presents an illustration of both individual and integrated feature classification, along with an analysis of the advantages and disadvantages of different classification methods.

Table 6 Sample results of the proposed framework

Feature	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**
Email Actual Class	Non-spam	Non-spam	Spam	Spam
Sender Email Address	0.00	0.00	0.10	0.00
Sender IP Address	0.00	0.10	0.00	0.00
Subject Line	0.00	0.00	0.10	0.10
In-Reply-To	0.00	0.00	0.00	0.00
Content-Type	0.10	0.10	0.10	0.10
Total	0.10	0.20	0.20	0.20
Origin-based classification (A)	Non-spam	Spam	Spam	Non-spam
Length of actual content	0.00	0.00	0.00	0.00
Count of Links	0.00	0.10	0.00	0.00
Count of DIV tags	0.00	0.10	0.10	0.10
Count of paragraph	0.00	0.00	0.00	0.10
Pre-processed Content	0.10	0.10	0.10	0.10
Total	0.10	0.30	0.20	0.30
Content-based classification (B)	Non-spam	Spam	Non-spam	Spam
Total count of images	0.10	0.00	0.10	0.10
The total size of the image	0.00	0.00	0.00	0.10
Total height and width of all images	0.10	0.00	0.00	0.10
All images format	0.00	0.00	0.10	0.00
Pre-processed text in all images	0.10	0.10	0.10	0.10
Total	0.30	0.10	0.30	0.40
Image-based Classification (C)	Spam	Non-spam	Spam	Spam

Classification based on A, B and C (D)	Spam	Spam	Spam	Spam
Total of all weights (A + B + C)	0.50	0.60	0.80	0.90
Hybrid Method Classification	Non-spam	Non-spam	Spam	Spam

The above Table 6 presents a complete assessment of email features to classify as spam or non-spam. Each feature undergoes careful evaluation across four sample emails, with their actual classifications serving as references. Features such as Sender Email Address, Sender IP Address, Subject Line, In-Reply-To, Content-Type, text length, number of links, DIV tags, paragraphs, and pre-processed material are scrutinized, each assigned a weight indicating its significance in the email samples after classification. The spam probability is assessed by combining values for each sample, calculating the summation of weights for each feature.

The table adopts an inclusive approach, integrating origin, content, and image-based features into the model. The final classification decision is based on the total weights, allowing for a balanced assessment that leverages the strengths of all features.

The hybrid classification, where each characteristic from the origin, content, and image section contributes to the final classification, is represented by a logistic regression model. The likelihood that an email is spam is estimated by the logistic regression model in the following way:

$$\mathrm{P}(spam|O,C,I) = \frac{1}{1+e^{-(\beta_0+\beta_o O+\beta_c C+\beta_I I)}} \qquad (4)$$

where,

$\mathrm{P}(spam|O,C,I)$ is the predicted probability that an email is spam given its origin (O), content (C), and image (I) features,

β_0 is the intercept (bias term)

$\beta_o, \beta_c, and\ \beta_I$ are the regression coefficients for the origin, content, and image feature sets respectively.
O is the sum of the origin-based features (e.g., sender IP address, sender email address),
C is the sum of the content-based features (e.g., count of links, spammy keywords),
I is the sum of the image-based features (e.g., total number of images, size, format).

The logistic regression equation becomes:

$$\log\left(\frac{P(spam|O,C,I)}{1-P(spam|O,C,I)}\right) = \beta_0 + \beta_o O + \beta_c C + \beta_I I \qquad (5)$$

This equation represents the log-odds of the email being classified as spam based on the combined contributions of the origin, content, and image features.

Now, let's see how each feature contributes to the classification and why combining them improves accuracy.

The term $\beta_o O$ captures the contribution of origin-based features. These features might capture spam emails like patterns such as subject line, IP address or sender email address. However, if origin-based features alone are insufficient, relying only on O would lead to false negatives.

The term $\beta_c C$ represents the contribution of content-based features. These features might capture spam-like patterns such as the use of specific phrases, links, or HTML tags. But when relying just on content features, certain emails with legitimate material might still be mistakenly tagged as spam, leading to false positives.

The term $\beta_I I$ captures the contribution of image-based features. Spam emails often include images with embedded text to bypass content-based filters. Image analysis can identify such patterns.

By combining origin, content, and image features, it captures a broader range of spam characteristics that might be missed by individual features. The logistic regression model uses the interaction between these feature sets to improve classification.

The interaction term in a logistic regression model that includes origin, content, and image features can be represented as:

$$\log\left(\frac{P(\text{spam}|O,C,I)}{1-P(\text{spam}|O,C,I)}\right) = \beta_0 + \beta_o O + \beta_c C + \beta_I I + \gamma OC^{OC} + \gamma OI^{OI} + \gamma CI^{CI} \quad (6)$$

Where:

$\gamma OC, \gamma OI, \gamma CI$ are the coefficients representing the interaction between origin, content, and image features.

These interaction terms show how one type of feature influences the effectiveness of others. For example, even if an email passes origin checks, suspicious content (C) or images (I) can still classify it as spam. Conversely, compassionate content might still identify the email as spam if it has a problematic origin or suspicious images.

To show how the combination of origin-based, content-based, and image-based features improves accuracy compared to using them individually, consider the following:

When only origin-based features are used, the model is limited by the patterns detectable from the sender's details. As shown in the table, the origin-based classification A misclassifies Sample 4. When only content-based features are used, the model can miss emails that evade content filters by using sophisticated obfuscation techniques or relying on images. For example, Sample 3 is misclassified by content-based features alone. When only image-based features are used, the model can miss text-based spam that doesn't use images. In the table, Sample 2 is misclassified by image-based features alone.

The hybrid model combines these features into a single score, resulting in improved accuracy, as demonstrated by the correct classification in Samples 1, 2, 3, and 4. mathematically, this is expressed by:

$$H = \alpha O + \beta C + \gamma I \quad (7)$$

$$\text{P}(spam|O,C,I) = \frac{1}{1+e^{-(\beta_0+\beta_o O+\beta_c C+\beta_I I)}} \quad (8)$$

By capturing complementing data from many sources, the feature combination lowers the number of false positives and false negatives, increasing accuracy. The hybrid classification outperforms any single feature set thanks to the useful signals that each feature category origin, content, and image.

To quantify how much the hybrid model improves accuracy, let’s introduce accuracy for each classifier:

A_O = accuracy of the origin-based classifier,
A_C = accuracy of the content-based classifier,
A_I = accuracy of the image-based classifier.

$$A_H = \alpha A_O + \beta A_C + \gamma A_I \quad (9)$$

Where:
α, β, γ represent the contributions of the origin, content, and image features respectively to the overall classification accuracy, interaction terms further contribute by addressing the dependencies between these feature sets.

This model will typically show an improved accuracy A_H compared to the individual accuracies A_O, A_C, and A_I because it leverages complementary information that cannot be captured by any single feature set.

Spam classification accuracy is enhanced by merging origin-based, content-based, and image-based features, as shown by the logistic regression regressive model. To reduce misclassifications, interaction terms between these attributes are essential for capturing dependencies. When compared to employing separate features, the hybrid technique performs better in classification since it catches the entire spectrum of spam features.

Table 7 Confusion Matrix

	Spam	**Non-spam**
Spam	Spam is classified as spam. (True-Positive)	Spam is classified as non-spam. (False-Negative)
Non-spam	Non-spam is classified as spam. (False-Positive)	Non-spam is classified as non-spam. (True-Negative)

The performance of the proposed framework is evaluated as per following metrics.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN}) \cdot 100 \quad (10)$$

$$\text{Recall} = (\text{TP})/(\text{TP} + \text{FN}) \quad (11)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (12)$$

$$\text{F} - \text{Score} = 2 \cdot (\text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall}) \quad (13)$$

V. Results and Discussion

The proposed model, with an impressive accuracy of 98.34%, has demonstrated exceptional proficiency in accurately classifying samples, making it the preferred choice when accuracy is the primary evaluation criterion. In contrast, the image feature sub-model performed the least well of the four, with an accuracy of 82.94%. Nonetheless, this accuracy is noteworthy, highlighting the critical role played by the image-based sub-model in certain scenarios. The origin-based sub model had an accuracy of 94.14%, while the content-based sub model had 97.03%. This implies that the content of an email is a driving factor in determining whether it is legitimate or spam email. Table 6 shows the classification results for both the individual and combined methods of spam email classification. The proposed model is lowering false results by merging the results from all sections, as those were on the higher side for individual classification, as seen by the false positive count mentioned in Fig. 2. The content-based model yields even the second-best result in false positive and false negative; however, it is minimized with the aid of origin-based and image-based features, which is the novel aspect of the proposed framework.

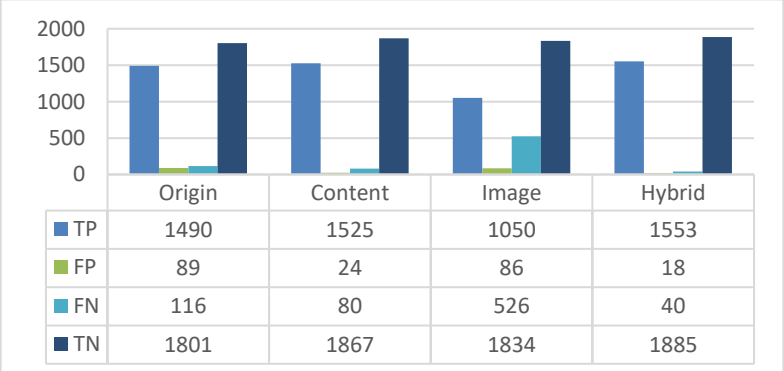

Fig. 2. Comparison of Classification Results of All Sub-Model and Hybrid Model

In evaluating the performance of the proposed hybrid model for the classification, precision, recall, and the F-Score have also been employed, offering valuable insights into its effectiveness, particularly in dealing with imbalanced datasets or different trade-offs between false positives and false negatives. Precision, gauging the model's ability to correctly identify positive instances among its predictions, assumes significance when the cost of false positives is considerable. Fig. 2 highlights the image-based sub-model's lowest precision, indicative of its limitations. However, the precision achieved by each sub-model contributes to the hybrid model's overall precision. Interestingly, the hybrid model's higher precision, compared to individual sub-models, underscores the correctness of its positive predictions, particularly valuable in contexts where false positives have substantial consequences. Despite occasional inaccuracies in outcomes produced by the content-based sub-model, the hybrid model capitalizes on the strengths of its constituent sub-models, thereby enhancing overall precision and recall.

Given the critical importance of accurately identifying spam emails, especially with recall playing a central role, the hybrid approach proves valuable. While the pursuit of high accuracy remains crucial, neglecting recall can lead to significant consequences, particularly in situations where missing positive cases (false negatives) are undesirable. This approach effectively reduces the entry of spam emails into the inbox by extracting text from images and utilizing it for spam email classification.

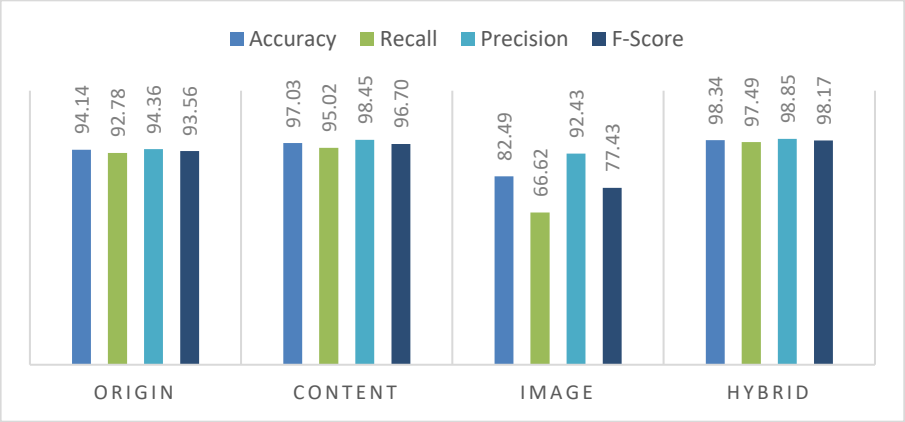

Fig. 3. Classification Results

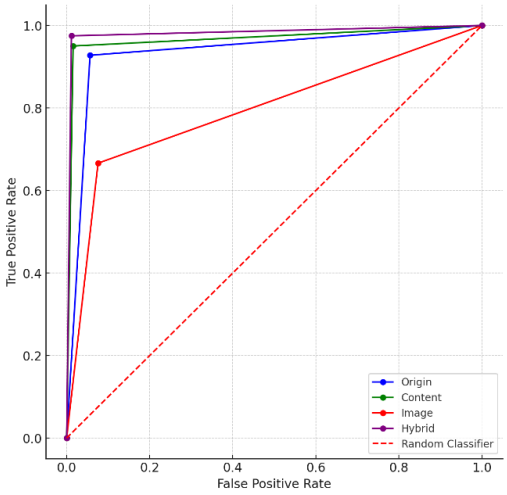

Figure 4 Comparison Origin, Content, Image and Hybrid Results

The utilization of integrated features encompassing origin, content, and image offers a substantial advantage in improving spam email classification compared to relying solely on individual features. Integration provides a more comprehensive and holistic view of the email's origin, content, and image, allowing for deeper analysis and a better understanding of its nature. By combining information about the sender's origin, content, and images, the classification system gains a multi-dimensional perspective that is far more robust in detecting spam from legitimate emails. Fig. 3 shows variation in image-based features classification but when the results of image-based features integrate with origin and content features to aggregate it aids the aggregation and ultimately final

classification. Fig. 4 represents the importance of hybrid method with the impact on the area covered as compared to individual section's classification.

One standout aspect of the proposed framework is its enhanced handling of false positives and false negatives, vital for spam email classification. It effectively minimizes both types of errors, leading to higher precision and recall. Additionally, the framework incorporates a continuous learning and adaptation mechanism, allowing it to evolve and improve over time as it encounters new data and email trends.

Table 8 Comparison of proposed model and service provider

	Personal Email Service Provider	**Proposed Framework**
Accuracy	89.79	98.34
Recall	78.83	97.49
Precision	97.85	98.85
F-Score	87.33	98.17

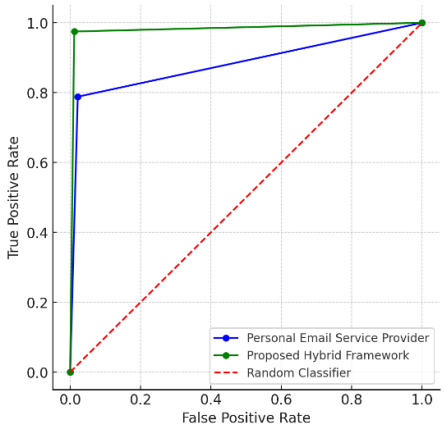

Figure 5 ROC Curve for Proposed Framework and Service Provider Results

The results of the proposed hybrid framework are compared to those of a traditional personal email service provider. Notably, the proposed hybrid framework demonstrates superior performance across all metrics, showcasing its potential to drastically enhance recall, accuracy, and f-score compared to the email service provider as shown in Fig. 5 and Table 8. In conclusion, the results affirm that the proposed framework yields a substantial improvement in the performance of spam email classification. Its adaptability to evolving email landscapes further underscores its status as a state-of-the-art solution for spam email classification, guaranteeing high precision, recall, accuracy, and f-score. Ultimately, this advancement enhances user experience and improves email security.

VI. Conclusion

This article classifies emails as either spam or non-spam using a unique and novel approach. By combining origin, content, and image-based feature results into a hybrid model, the suggested unique approach calculates and applies weights. Using feature results and a decision-making process based on the email's origin, text, and images, a classification of spam emails is reached. The findings demonstrate that, in comparison to earlier techniques, the novel strategy described here yields a comprehensive and accurate result with the fewest false positives. To further improve the general accuracy of spam email classification, this work can be expanded in the future to include email detection for non-text images and email attachments.

Acknowledgment
None.

VII. References

[1] N. Agrawal and S. Singh, "Origin (Dynamic Blacklisting) Based Spammer Detection and Spam Mail Filtering Approch," *International Conference on Digital Information Processing, Data Mining, and Wireless Communications*, pp. 99-104, 6-8 july 2016.

[2] P. Liu and T.-S. Moh, "Content Based Spam E-mail Filtering," in *International Conference on Collaboration Technologies and Systems*, 2016.

[3] C.-C. Wang, "Sender and Receiver Addresses as Cues for Anti-spam Filtering," *Journal of Research and Practice in Information Technology*, vol. 36, no. 1, p. 3–7, 2004.

[4] J.-J. Sheu, "An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization," *International Journal of Network Security*, vol. 9, no. 1, pp. 34-43, 2009.

[5] T. Krause, R. Uetz and T. Kretschmann, "Recognizing Email Spam from Meta Data Only," in *IEEE Conference on Communications and Network Security (CNS)*, 2019.

[6] A. Sharma, Manisha, Dr.Manisha and R. Jain, "A Survey on Spam Detection Techniques," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 12, December 2014.

[7] E. P. Sanz, O. M. G. M. Hidalgo and J. C. C. P. Rez, "Email Spam Filtering," in *Advances in Computers*, vol. 74, ScienceDirect, 2008, pp. 45-114.

[8] A. Annadatha and M. Stamp, "Image spam analysis and detection," *Journal of Computer Virology and Hacking Techniques*, vol. 14, p. 39–52, 2018.

[9] C.-T. Wu, K.-T. Cheng, Q. Zhu and Y.-L. Wu, "Using Visual Features for Anti-spam Filtering," 2005.

[10] S. Krasser, Y. Tang, J. Gould, D. Alperovitch and P. Judge, "Identifying Image Spam based on Header and File Properties using C4.5 Decision Trees and Support Vector Machine Learning," United States Military Academy, 2007.

[11] W.-B. Chen and C. Zhang, "Image Spam Clustering – An Unsupervised Approach," in *ACM*, Beijing, China, 2009.

[12] M. Das, A. Bhomick, J. Y. Singh and V. Prasad, "A Modular Approach towards Image Spam Filtering," in *International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, 2014.

[13] H. Sohrab, A. Abtahee, I. Kashem, M. M. Hoque and I. H. Sarker, "Crime Prediction Using Spatio-Temporal Data," in *International Conference on Computing Science, Communication and Security*, Singapore, 2020.

[14] E. Blanzieri and A. Bryl, "A Survey of Learning-Based Techniques of Email Spam Filtering," 2008.

[15] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, 2022.

[16] S. Jukić, J. Azemović, D. Kečo and J. Kevric, "Comparison of Machine Learning Techniques In Spam E-Mail Classification," *Southeast Europe Journal of Soft Computing*, vol. 4, no. 1, pp. 32-36, 2015.

[17] S. Youn and D. McLeod, "A Comparative Study for Email Classification," in *Advances and Innovations in Systems, Computing Sciences and Software Engineering* , 2007.

[18] A. S. Urmi, M. T. Ahmed, M. Rahman and A. T. Islam, "A Proposal of Systematic SMS Spam Detection Model Using Supervised Machine Learning Classifiers," Singapore, 2022.

[19] Y. HaCohen-Kerner, D. Miller and Y. Yigal, "The Influence of Preprocessing on Text Classification Using A Bag-of-words Representation," *PLOS ONE*, 2020.

[20] S. Hershkop and S. J. Stolfo, "Combining email models for false positive reduction," in *Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA, 2005.

[21] B. &. L. C.-H. &. W. S. &. I. D. &. P. C. Byun, "An Anti-spam Filter Combination Framework for Text-and-Image Emails through Incremental Learning," 2009.

[22] R. Bansod, R. S. Mangrulkar and V. G. Bhujade, "Text and Image based Spam Email Classification using an ANN Model- an Approach," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 5, pp. 115-118, May 2015.

[23] A. U. Surwade, M. P. Patil and S. R. Kolhe, "Effective and Adaptive Technological Solution to block Spam E-mails," in *International Conference on Advances in Human Machine Interaction*, Doddaballapur, Bangalore, India, 2016.

[24] A. Kumar, J. M. Chatterjee and V. G. Díaz, "A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing," *International Journal of Electrical and Computer Engineering (IJECE)*, pp. 486-493, 2020.

[25] Y. Gao, J. Song, J. Gao, N. Suo, A. Ren, J. Wang and K. Zhang, "Research on Spam Detection with a Hybrid Machine Learning Model," *3D Imaging—Multidimensional Signal Processing and Deep Learning. Smart Innovation, Systems and Technologies*, vol. 349, p. 227–235, 2023.

[26] Z. Zhang, E. Damiani, H. Hamadi, C. Yeun and F. Taher, "A Late Multi-modal Fusion Model for Detecting Hybrid Spam E-mail," *International Journal of Computer Theory and Engineering*, vol. 15, no. 2, 2023.