Journal of Cybersecurity and Information Management

ISSN: 2690-6775 Vol: 18, No. 01, 2026 PP. 21–26

Article DOI: https://doi.org/10.54216/JCIM.180103

FULL LENGTH ARTICLE

An Explainable Hybrid SVM Framework for Spam and Malicious Email Detection in Enterprise Information Systems

Mahmoud A. Zaher[1,*], **Nabil M. Eldakhly**[2]

[1]Asso. prof. Faculty of Artificial Intelligence and Information, Horus University (HUE), Egypt

[2]Asso. prof. Faculty of Computers and Information, Egypt

Emails: mzaher@horus.edu.eg; nabil.omr@sadatacademy.edu.eg

Received: January 18, 2026 Revised: February 17, 2026 Accepted: March 28, 2026 ⋆ Corresponding author

ABSTRACT

Email has been a key communication and information-management tool in contemporary organizations, yet it is also one of the most misused avenues to spam, fraud, credential theft, and malicious code delivery. Lightweight and reproducible detection models are especially useful to universities, public institutions, and small-to-medium enterprises which might not have access to costly proprietary filtering infrastructures because of the operational relevance of email security. In this paper I suggest an Explainable Hybrid SVM Framework (EHSF) to detect spam and malicious-risk email in a business information system. The framework integrates TF–IDF representation of text with lightweight risk-based email indicators, such as structural and lexical cues that can be obtained at low computation cost. An external Enron-Spam data were used so that it may be reproducible and will be checked later by the reviewers and readers. The experimentation process was coded in Python and assessed in terms of accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix. These findings demonstrate that the suggested Linear SVM-based framework has the highest overall performance with accuracy of 0.9853, precision of 0.9818, recall of 0.9893, F1-score of 0.9855, and ROC-AUC of 0.9981 on the held-out test set. The confusion matrix shows that there were only 34 false negatives and 58 false positives which show that there was a good discrimination between ham and spam classes. Besides the predictive performance, the framework provides an interpretable layer based on the analysis of influential lexical indicators related to risky and legitimate enterprise emails. The research adds a replicable and operationally viable methodology that complies with the needs of cybersecurity and information-management, and is lightweight enough to be implemented in the real-life setting within an organization.

Keywords: Email security ▪ Spam detection ▪ Support vector machine ▪ Cybersecurity ▪ Information management ▪ Text mining ▪ Explainable machine learning

1. INTRODUCTION

Electronic mail has remained a mission-critical channel of communication within organizations that has assisted in administrative processes, exchange of knowledge, customer interaction and coordination of operations within an organization. Simultaneously, email continues to be a leading source of spam attacks, malicious software attachments, harvesting of credentials, and social engineering attacks [1]. According to IBM, phishing is a type of cyberattack where a person or organization becomes a victim through emails, messages, or other resources that contain deceptive information and tricks to reveal sensitive information, install malware, or otherwise expose to cybercrime [8]. Current threat reports underline the continuing scale. In the fourth quarter of 2024 alone, APWG reported 989,123 phishing attacks, SaaS/webmail being one

of the most-targeted sectors [2]. Kaspersky also stated that its security solutions stopped over 893 million attacks in 2024, in addition to over 125 million attacks involving malicious email attachments [10].

These findings dispel the notion that email security is an infrastructure issue that is resolved and no longer a fundamental cybersecurity issue. It is of particular concern in terms of information-management. Email systems are not just a transport of messages; they keep a record, arrange approvals, transfer sensitive files, and influence the way information within organizations is produced, disseminated and acted upon.

A substantial amount of previous literature has been directed at the problem of spam filtering and other email classification problems. The first Bayesian techniques put spam filtering as a viable machine-learning task into practice [3][12][14]. Following research and surveys revealed that linear models, support vectors, ensemble training, and more in-depth neural training can all be useful in various data conditions as well as under different conditions of data sets [5][6][7][9][15]. Nevertheless, three gaps remain important. First, most studies focus on predictive accuracy but lack an adequate focus on interpretability and operational usability. Second, some studies use datasets that are challenging to access, have limited documentation, or are difficult to reuse. Third, performance-intensive deep architectures are not always needed or feasible by institutions needing lightweight deployment and transparent decision support.

This paper aims at filling these gaps by proposing an Explainable Hybrid SVM Framework (EHSF) of spam and malicious-risk email detection in enterprise information systems. The structure is a mixture of sparse textual representation with TF–IDF and lightweight engineered email indicators, and a Linear Support Vector Machine is the primary classifier. The research is based on a publicly available spam corpus in Enron-Spam to ensure reproducibility. The paper is structured as follows. Section 2 reviews related work. Section 3 presents the data and experimental design. Section 4 describes the proposed methodology. Section 5 reports the empirical results. Section 6 discusses practical and theoretical implications. Section 7 concludes the paper.

2. RELATED WORK

2.1 Classical machine-learning approaches to spam filtering

Spam filtering has long served as a canonical text-classification problem. Foundational work by Sahami et al. [14] demonstrated the practical utility of Bayesian learning for junk-email filtering, while Androutsopoulos et al. [3] provided one of the earliest systematic evaluations of Naive Bayes anti-spam filtering under cost-sensitive conditions. The Enron family of datasets later enabled more standardized email-learning experiments by providing a large and realistic enterprise email corpus [4][11]. Building on this broader text-classification tradition, Joachims [9] established why support vector machines are especially suitable for high-dimensional sparse text data, a finding that remains highly relevant for email classification.

Subsequent spam-filtering studies continued to confirm the effectiveness of statistical and machine-learning models. Metsis et al. [12] compared multiple variants of Naive Bayes on Enron-Spam data and showed that even closely related Bayesian formulations may behave differently depending on representation and update conditions. Cormack [6] reviewed the field comprehensively and emphasized that practical spam filtering should be evaluated not only as a generic learning task but also as an adaptive decision problem shaped by user utility and attack evolution. Caruana and Li [5] extended this perspective by surveying emerging approaches beyond traditional filtering pipelines, including distributed, personalized, and privacy-aware configurations.

2.2 Email mining, cybersecurity, and operational relevance

Email mining covers a broader set of tasks than spam filtering alone, including categorization, contact analysis, network analysis, and visualization [16]. Nevertheless, spam and malicious-email detection remain central because they directly affect the trustworthiness of enterprise communication channels. From a cybersecurity standpoint, email-based threats are important because they can trigger downstream compromise through malicious links, executable attachments, fake invoices, impersonation, and credential-theft campaigns [2][8]. Accordingly, spam and phishing detection should be viewed not merely as inbox-cleanliness tasks, but as mechanisms for preserving confidentiality, operational continuity, and decision quality in information systems.

Several applied studies have proposed improved classifiers for this purpose. Olatunji [13] presented an SVM-based spam detection model and reported improvements over earlier methods on a popular benchmark dataset. More recently, Siddique et al. [15] evaluated Naive Bayes, SVM, CNN, and LSTM models for spam detection and showed that deep learning can offer competitive results under suitable language and dataset settings. At the survey level, Dada et al. [7] highlighted the continued relevance of machine learning for spam filtering while also noting open research directions involving adversarial robustness, evolving spam behavior, and stronger evaluation practices.

2.3 Research gap and paper positioning

Despite the maturity of spam-filtering research, several practical limitations remain. First, a number of studies prioritize predictive performance while offering limited interpretability for security administrators. Second, reproducibility is often constrained by inaccessible corpora, undocumented preprocessing, or missing code details. Third, some recent approaches favor increasingly complex models even when simpler linear models may remain highly competitive on sparse enterprise email data. For many institutions, especially SMEs, universities, and public-sector organizations, transparent and computationally efficient models are more likely to be adopted in production.

This paper addresses these limitations by positioning the problem as one of spam and malicious-risk email detection rather than narrow phishing attribution. This distinction is important because the chosen public dataset is labeled as spam versus ham. The proposed EHSF integrates sparse textual representation with lightweight risk indicators, uses a public and verifiable dataset, and reports results through a

directly reproducible Python workflow.

3. MATERIALS AND EXPERIMENTAL DESIGN

3.1 Dataset

The experiments were conducted on a public Enron-Spam dataset derived from the well-known Enron email corpus [4][11]. The original Enron email collection contains approximately half a million messages from around 150 users, mostly senior management staff at Enron [4]. For spam-classification experiments, a cleaned public Enron-Spam release prepared by Wiechmann was used because it consolidates the data into a single machine-readable CSV representation suitable for reproducible experimentation [17]. According to the dataset documentation, this corpus contains 33,716 email messages in total, including 17,171 spam messages and 16,545 non-spam (ham) messages [17].

In the experiment conducted for this manuscript, the public training file comprised 31,306 usable rows after removing missing values. The available columns include the email text and its binary label. Each message was mapped to one of two classes: 1 for spam and 0 for ham. The use of a public dataset ensures that reviewers and readers can later reproduce the full preprocessing and model-training pipeline.

Table 1. Summary of the public Enron-Spam dataset used in this study.

Item	Count	Source
Total emails in the public release	33,716	Wiechmann [17]
Spam emails	17,171	Wiechmann [17]
Ham emails	16,545	Wiechmann [17]
Rows used in the present run after NA removal	31,306	This study
Classification setting	Binary (spam/ham)	This study

3.2 Train–test partition and evaluation metrics

The cleaned dataset was split using a stratified 80/20 train–test strategy with a fixed random seed of 42. Stratification ensured that the class balance was preserved across the training and testing partitions. The test set was kept fully held out during model fitting and was used only for final evaluation.

The following metrics were used to evaluate the classifiers: accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix. While accuracy provides an overall correctness measure, F1-score is more informative for spam filtering because it balances false positives and false negatives. Recall is particularly important in cybersecurity screening tasks because missed risky emails may expose users to compromise. ROC-AUC was additionally used to compare ranking quality across models with probabilistic or calibrated outputs.

4. PROPOSED METHODOLOGY

4.1 Overview of the proposed framework

The proposed Explainable Hybrid SVM Framework (EHSF) is illustrated in Figure 1. The framework consists of five stages: public dataset ingestion, text cleaning and normalization, hybrid feature construction, classification, and explanation. The design philosophy is intentionally lightweight. Rather than relying on deep architectures with substantial training costs, the framework combines well-established sparse text modeling with a small set of computationally inexpensive risk indicators.

Figure 1. Overview of the proposed Explainable Hybrid SVM Framework (EHSF).

4.2 Text cleaning and normalization

Each email message was first normalized using a rule-based preprocessing pipeline. The following operations were applied in sequence: conversion to lowercase; replacement of URLs with a generic URL token; replacement of email addresses with a generic EMAIL token; normalization of numeric strings into a generic NUM token; removal of non-alphabetic symbols; and compression of repeated whitespace. The preprocessing design aimed to preserve the semantic and structural cues most relevant to spam discrimination while reducing superficial noise. It also allowed the generated explanatory indicators to remain interpretable for later discussion.

4.3 Hybrid feature construction

The framework employs two complementary feature groups. First, the cleaned email text was converted into a TF–IDF matrix using unigram and bigram features. The vectorizer used a maximum of 3,000 features, a minimum document frequency of 3, and sublinear term-frequency scaling. This representation captures lexical salience and short phrase structure, both of which are important in email classification. Second, ten handcrafted indicators were extracted from each raw email: total number of words, total number of characters, number of URLs, number of email addresses, number of digits, number of uppercase characters, exclamation-mark count, question-mark count, suspicious-keyword count, and average word length. The suspicious-keyword counter was derived from a small lexicon including terms such as free, click, urgent, winner, verify, account, password, money, and remove.

4.4 Classification models

Five classifiers were evaluated: Multinomial Naive Bayes using TF–IDF only; Logistic Regression using the hybrid feature set; Linear SVM using the hybrid feature set; Random Forest using engineered features only; and a weighted soft ensemble combining Logistic Regression, Multinomial Naive Bayes, and a calibrated Linear SVM. Although the ensemble was initially designed as the proposed architecture, the empirical results showed that the Linear SVM with hybrid features achieved the strongest overall test-set performance. The final proposed model is therefore framed as an Explainable Hybrid SVM Framework rather than as an ensemble-centric approach.

4.5 Explainability layer

Because Linear SVM coefficients are not directly probabilistic, the interpretability layer was derived using the coefficient profile of the Logistic Regression model trained on the same hybrid feature space. This allowed the study to identify the most strongly spam-associated and ham-associated lexical cues while keeping the main predictive model unchanged.

The explanatory layer is therefore operationally useful for administrators who need to understand why risky emails are being flagged.

5. RESULTS

5.1 Comparative classification performance

Table 2 summarizes the predictive performance of the evaluated models on the held-out test set. The proposed EHSF (Linear SVM with hybrid features) achieved the best overall F1-score of 0.9855 and the highest ROC-AUC of 0.9981. Logistic Regression and the weighted ensemble showed nearly identical performance, while Multinomial Naive Bayes remained competitive but weaker. Random Forest trained only on engineered numerical features performed substantially worse, indicating that lexical content remains the dominant signal on this dataset.

Table 2. Comparative results on the held-out test set.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Linear SVM (TF–IDF + features)	0.9853	0.9818	0.9893	**0.9855**	**0.9981**
Weighted soft ensemble	0.9851	0.9791	0.9918	0.9854	0.9980
Logistic Regression (TF–IDF + features)	0.9848	0.9782	0.9921	0.9851	0.9980
Multinomial Naive Bayes (TF–IDF)	0.9738	0.9664	0.9823	0.9743	0.9952
Random Forest (engineered features only)	0.7447	0.7582	0.7269	0.7422	0.8261

Figure 2 visualizes the F1-score comparison. The figure confirms that the top three models are tightly clustered, but the Linear SVM maintains a slight performance advantage. In practical settings, this is a favorable outcome because the SVM remains simple, robust, and computationally efficient.

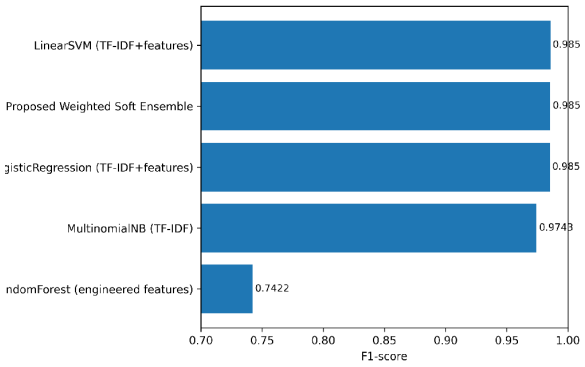

Figure 2. F1-score comparison across the evaluated models.

5.2 Confusion-matrix analysis

The confusion matrix of the proposed EHSF is presented in Figure 3, and its values are also shown in Table 3. The model correctly classified 3,037 ham messages and 3,133 spam messages. Only 58 ham emails were incorrectly flagged as spam, while 34 spam emails were missed. This indicates a low false-negative rate, which is desirable for security screening because missed malicious-risk emails can carry higher operational cost than moderate false-positive burden.

Table 3. Confusion matrix of the proposed EHSF.

Ham	3,037	58
Spam	34	3,133

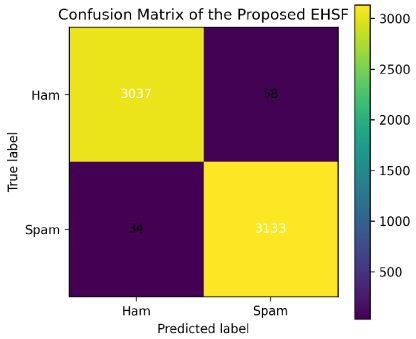

Figure 3. Confusion matrix of the proposed EHSF on the held-out test set.

5.3 ROC analysis

Figure 4 presents the ROC curves of the compared models. The SVM, Logistic Regression, and weighted ensemble all achieved near-ceiling ROC-AUC values, indicating excellent ranking performance. The closeness of the curves suggests that sparse textual information is highly discriminative in the Enron-Spam setting. However, the SVM retains an important practical advantage because it delivers the best F1-score without requiring a more complex multi-model inference process.

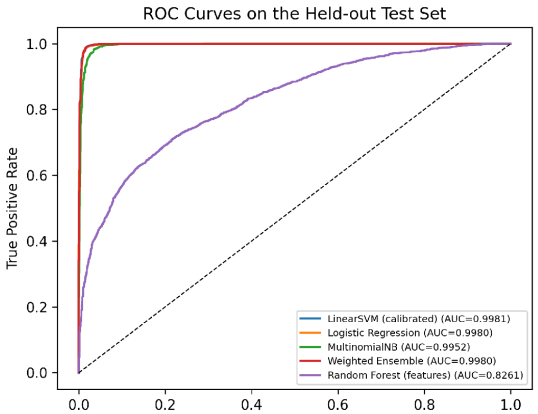

Figure 4. ROC curves of the evaluated models on the held-out test set.

5.4 Interpretable lexical indicators

To provide insight into the classification behavior, the strongest positive and negative coefficient terms from the Logistic Regression explanation layer were extracted. Figure 5 shows the top ten spam-associated and ham-associated textual indicators. Spam-associated terms included no, remov, http, your, softwar, medic, money, and site. Ham-associated terms were dominated by enterprise-specific language such as enron, thank, attach, houston, and schedul.

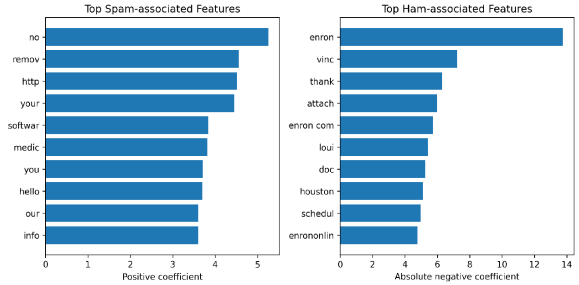

Figure 5. Most influential lexical indicators from the explainability layer.

These findings are operationally meaningful. They suggest that the proposed framework does not merely memorize message length or punctuation, but captures identifiable lexi-

cal cues that distinguish promotional or link-oriented spam from routine internal communication. This interpretability strengthens the case for deployment in institutional environments where security teams often require justifiable filtering behavior.

6. DISCUSSION

The results indicate that strong enterprise email screening does not necessarily require deep or computationally heavy models. On the public Enron-Spam benchmark used in this study, a lightweight Linear SVM trained over hybrid sparse-text and low-cost engineered features achieved the best overall performance. This is consistent with the long-standing suitability of linear margin-based methods for high-dimensional textual data [9], while also reinforcing later evidence that SVM-based spam filtering remains competitive in practical settings [13].

Several implications emerge from these findings. First, for cybersecurity operations, the low false-negative count supports the use of the model as a front-line triage mechanism for suspicious email streams. Second, for information management, a transparent filtering layer can reduce message overload, improve trust in organizational communication channels, and support more reliable handling of email-based workflows. Third, the combination of public data, straightforward preprocessing, and standard open-source tooling makes the approach suitable for replication and adaptation in resource-constrained institutions.

The study also contributes methodologically by emphasizing reproducibility. The dataset source is public and documented [4][17], the experimental pipeline is deterministic under the chosen random seed, and the full manuscript package can be edited directly in LaTeX. This is particularly valuable for Q4-level applied journals where practical reproducibility can be as important as methodological novelty.

Nevertheless, the work has several limitations. The Enron-Spam corpus is historical and may not fully represent modern phishing campaigns, business email compromise, or AI-assisted fraud. The labels are binary spam/ham rather than fine-grained threat categories, which means the model should be framed as a spam and malicious-risk email detector rather than a specialized phishing-attribution engine. Moreover, while the lightweight engineered features aid interpretability, the gains over strong linear text baselines are modest, indicating that lexical sparse signals dominate this dataset.

Future work should therefore validate the proposed framework on newer email-security corpora, incorporate richer metadata such as sender reputation and header anomalies, and extend the model to multilingual and spear-phishing scenarios. Active-learning feedback loops from administrators and analysts may also help adapt the classifier to evolving organizational threat environments.

7. CONCLUSION

In this paper, an Explainable Hybrid SVM Framework (EHSF) of spam and malicious-risk email detection in enterprise information systems was presented. The model integrates TF-IDF textual representation with light risk-oriented features and tests the resulting model on one of the public Enron-Spam databank to guarantee reproducibility. The empirical findings showed that the Linear SVM-based version had the highest overall held-out performance by a factor of 0.9853, precision of 0.9818, recall of 0.9893, F1-score of 0.9855 and ROC-AUC of 0.9981. The confusion-matrix analysis also indicated that the number of false-negatives is low, which is a significant attribute in cybersecurity screening activities. The paper also focused on interpretability and practical applicability in addition to the raw predictive accuracy. The discriminative lexical indicators were extracted, and it demonstrated significant differences between risky promotional contents and legitimate enterprise communications, thus helping to establish more transparent administrative decision-making processes.

DATA AVAILABILITY

The public Enron email corpus is documented by Carnegie Mellon University [4]. The cleaned Enron-Spam release used for this work is publicly available from the repository prepared by Wiechmann [17]. The Python experiment script used to generate the reported results can be distributed alongside the manuscript package for reproducibility.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

REFERENCES

[1] S. S. Sayeed, M. S. Hossain, and K. Andersson, "A comprehensive survey on phishing detection using machine learning and deep learning techniques," IEEE Access, vol. 12, pp. 15234–15258, 2024, doi: 10.1109/ACCESS.2024.

[2] Anti-Phishing Working Group, Phishing Activity Trends Report, 4th Quarter 2024, 2025.

[3] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering," arXiv preprint cs/0006013, 2000.

[4] Carnegie Mellon University, Enron Email Dataset, 2015.

[5] G. Caruana and M. Li, "A survey of emerging approaches to spam filtering," ACM Computing Surveys, vol. 44, no. 2, pp. 1–27, 2012, doi: 10.1145/2089125.2089129.

[6] G. V. Cormack, "Email spam filtering: A systematic review," Foundations and Trends in Information Retrieval, vol. 1, no. 4, pp. 335–455, 2008, doi: 10.1561/1500000006.

[7] 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, e01802, 2019, doi: 10.1016/j.heliyon.2019.

[8] IBM, What Is Phishing?, accessed April 11, 2026.

[9] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Machine Learning: ECML-98, Springer, pp. 137–142, 1998, doi:

10.1007/BFb0026683.

[10] Kaspersky, Kaspersky Reports Nearly 900 Million Phishing Attempts in 2024 as Cyber Threats Increase, 2025.

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

[12] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes–which naive bayes?" in Proceedings of the Third Conference on Email and Anti-Spam (CEAS), 2006.

[13] S. O. Olatunji, "Improved email spam detection model based on support vector machines," Neural Computing and Applications, vol. 31, pp. 691–699, 2019, doi: 10.1007/s00521-017-3100-y.

[14] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in Learning for Text Categorization: Papers from the 1998 Workshop, AAAI Press, pp. 55–62, 1998.

[15] Z. B. Siddique, M. A. Khan, I. Ud Din, A. Almogren, I. Mohiuddin, and S. Nazir, "Machine learning-based detection of spam emails," Computational Intelligence and Neuroscience, 2021:6508784, 2021, doi: 10.1155/2021/6508784.

[16] G. Tang, J. Pei, and W.-S. Luk, "Email mining: Tasks, common techniques, and tools," Knowledge and Information Systems, vol. 41, no. 1, pp. 1–31, 2014, doi: 10.1007/s10115-013-0658-2.

[17] M. Wiechmann, MWiechmann/enron_spam_data: The Enron-Spam dataset preprocessed in a single, clean CSV file, accessed April 11, 2026.