Volume XXIX 2026
MBNA Publishing House Constanta 2026



SBNA PAPER • OPEN ACCESS

Hybrid AI-Based Spam Email Filtering: Balancing Accuracy, Adaptability, and Efficiency

To cite this article: Andrei (Ștefănescu) Georgiana, Bucur Cătălin-Eugen, Sandu Călin, Răcuciu Ciprian, Scientific Bulletin of Naval Academy, Vol. XXIX 2026, pg. 368-375.

Submitted: 29.04.2026
Revised: 25.10.2026
Accepted: 03.11.2026

Available online at www.anmb.ro

ISSN: 2392-8956; ISSN-L: 1454-864X

doi: 10.21279/1454-864X-26-I1-032
SBNA© 2026. This work is licensed under the CC BY-NC-SA 4.0 License

Hybrid AI-Based Spam Email Filtering: Balancing Accuracy, Adaptability, and Efficiency

Andrei (Ștefănescu) Georgiana[1*], Bucur Cătălin-Eugen[2], Sandu Călin[1], Răcuciu Ciprian[3]

[1]Doctoral School, Military Technical Academy Ferdinand I, Bucharest, Romania
[2]Doctoral School, National University of Science and Technology Politehnica Bucharest, Romania, bucur.eugen@anmb.ro
[3]School, Military Technical Academy Ferdinand I, „Titu Maiorescu" University, Bucharest, Romania, ciprian.racuciu@gmail.com
andrei.georgiana@anmb.ro

Abstract. Proliferating at an unprecedented rate, unsolicited and adversarial electronic traffic poses a systemic threat to contemporary digital communication frameworks. While conventional, deterministically grounded rule engines guarantee low-latency processing, their structural rigidity compromises their efficacy against sophisticated evasive tactics, such as semantic obfuscation, spear-phishing, and synthetically generated generative-AI content. Conversely, data-driven Machine Learning (ML) paradigms exhibit superior adaptive detection thresholds, yet they inherently introduce steep computational penalties and heightened architectural complexity.To reconcile these conflicting operational demands, this study introduces a novel hybrid framework that orchestrates deterministic filtering and stochastic ML models within a unified, modular pipeline. By strategically isolating the detection, orchestration, and policy-enforcement phases, the proposed architecture yields enhanced vertical scalability and deployment flexibility. Empirically validated through a rigorous comparative analysis, our dual-engine methodology achieves a highly optimized trade-off, significantly mitigating false-positive rates while maintaining strict real-time throughput constraints that standalone alternatives fail to replicate.
Keywords: Spam filtering, Hybrid filtering architecture, Machine learning detection, Phishing detection, Email security

1. Introduction

Within the contemporary cyber threat landscape, electronic mail architectures persist as a primary, highly exploited vector for systemic vulnerabilities, specifically through the proliferation of malicious spam and sophisticated phishing campaigns [2,4]. The widespread orchestration of state-of-the-art defensive perimeters has not deterred adversarial actors; instead, it has catalyzed a continuous evolution in offensive tactics. Modern exploits increasingly pivot toward semantic obfuscation, dynamic domain generation algorithms (DGAs), and the deployment of synthetically engineered, AI-driven messaging designed to bypass perimeter defenses.

Historically, legacy filtering frameworks have operated on deterministic paradigms, utilizing static rule-sets, reputation-based blacklists, and hardcoded heuristic scoring mechanisms [3,6]. Although computationally lightweight, these rigid architectures exhibit an intrinsic inability to extrapolate or generalize when confronted with novel, zero-day attack topologies. To mitigate this systemic deficit, data-driven Machine Learning (ML) methodologies have been integrated to instill mathematical adaptability via stochastic modeling [1,4]. This shift toward predictive intelligence, however, is not a panacea; it introduces a distinct set of operational trade-offs, most notably characterized by prohibitive computational latencies, an acute dependency on high-fidelity labeled corpora, and an elevated propensity for false-positive classifications.

This work addresses these limitations by proposing a hybrid filtering architecture that combines the strengths of rule-based and ML-based approaches.

The main contributions of this work are as follows:

(1) the design of a fully on-premises hybrid spam filtering architecture;

(2) a selective ML processing mechanism for latency optimization;

(3) a modular pipeline separating detection, decision, and enforcement stages.

2. Materials and methods

2.1 Study Area

This study focuses on email spam detection, a cybersecurity and information-filtering problem in which incoming messages are classified as either spam or ham (legitimate email). The implemented system is intended for practical email environments where both individual messages and archived mailbox files must be processed efficiently. The selected GitHub project provides a modular spam-email classification workflow and a Streamlit-based interface for interactive testing and batch analysis.

2.2 Materials

Experiments were conducted on a system equipped with an Intel i7 CPU, 16 GB RAM, and running Python 3.10+, with Streamlit for the user interface, scikit-learn for ML modelling, and pandas, NumPy, and BeautifulSoup4 for data handling and pre-processing. The repository is organized into separate folders for components, pipelines, configuration, utilities, input data, output artifacts, and logs, which supports reproducible experimentation and model deployment. The system also supports processing mbox email archives, allowing both single-message inference and batch classification.

2.3 Methods

The spam-filtering workflow follows a modular pipeline architecture structured around detection, decision, and enforcement stages, enabling a clear separation between initial filtering, classification, and final action on incoming email messages. According to the repository documentation, the training pipeline loads the dataset from data/dataset/dataset.csv, evaluates multiple classifiers, and stores the resulting model artifacts and vectorizer files in the outputs/ directory. The system is configurable through src/config/config.py, where input/output paths, model hyperparameters, and cross-validation parameters can be adjusted.

The repository reports the use of several classical ML algorithms, including Support Vector Machine (SVM) [7], Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) classifiers. Model evaluation is performed using 5-fold cross-validation, and the logged metrics include Accuracy, Precision, Recall, and F1-score. The documentation states that the best-performing model is selected for inference, with SVM or RF often emerging as the preferred classifiers.

For inference, the Streamlit application allows users to paste a single email message or upload an mbox file for batch processing. The system then predicts whether the content is spam or ham and returns the result through the web interface, making the method suitable for both real-time inspection and offline analysis. This combined training-and-deployment design helps bridge experimental evaluation and practical use.

The evaluation was conducted using the Email Spam Classification Dataset (CSV format), which consists of 5,172 labelled email instances, each annotated as either spam (1) or non-spam (0). The dataset is structured as a high-dimensional tabular representation, comprising approximately 3,002 attributes per instance. The first column encodes an anonymized email identifier, while the final column provides the binary classification label.

The remaining ~3,000 features correspond to the most frequent words extracted from the corpus, with each feature representing the occurrence frequency of a specific term within a given email. This design results in a precomputed bag-of-words representation, enabling direct application of machine-learning algorithms without requiring additional text vectorization.

The dataset exhibits a moderate class imbalance, with legitimate emails forming the majority class, reflecting realistic operational conditions in email communication environments.

The hybrid configuration was designed to minimize unnecessary computational overhead by applying rule-based filtering as a first-stage decision mechanism. Only emails that could not be confidently classified were forwarded to the ML model. This selective processing strategy directly contributes to the improved latency and efficiency observed in the experimental results.

Latency is defined as the average end-to-end processing time per email, including feature extraction and classification, and is reported as the mean value across test samples.

3. Results and Discussion

3.1 Hybrid Email Security Architecture

This section presents the main outcomes of the proposed hybrid spam-email filtering approach and explains how the system is organized to support both security and operational efficiency. The hybrid nature of the system refers to the combination of rule-based pre-filtering and ML-based classification within a unified decision pipeline.

The efficacy of a bifurcated, hybrid topology in electronic mail security lies in its capacity to disassociate low-latency, deterministic screening from resource-intensive analytical pipelines, thereby establishing divergent execution paths tailored to the structural attributes of each message. In operational environments, this paradigm enhances architectural versatility; it orchestrates the high-throughput processing of benign, routine traffic while selectively diverting anomalous or high-entropy payloads toward granular inspection engines—all executed within a consolidated local perimeter.

Consequently, the primary structural contribution of this research materializes as an on-premises, two-tier defensive framework. The foundational abstraction layer executes high-scalability pre-filtering via deterministic rules and hardcoded heuristics, achieving near-instantaneous validation without compromising the system's computational budget. Optimized for high-volume, baseline sanitization, this tier guarantees deterministic real-time responsiveness. Conversely, the secondary abstraction layer deploys advanced predictive analytics powered by non-linear Machine Learning (ML) classifiers, invoked exclusively for payloads exhibiting high ambiguity. This tiered orchestration circumvents global performance degradation, ensuring that sophisticated threat vectors are scrutinized with high-fidelity precision. Furthermore, maintaining all transactional boundaries within the internal enterprise infrastructure inherently satisfies stringent data sovereignty mandates, regulatory compliance (e.g., GDPR), and total operational hegemony.

Deconstructing the operational lifecycle of the detection engine reveals a strictly sequential telemetry workflow. The ingestion pipeline transitions systematically from raw ingress mail selection through multidimensional feature extraction, deterministic/stochastic model execution, and ultimately, policy-driven decision enforcement. This operational sequence underscores that the classification engine eschews direct manipulation of unparsed, raw textual strings; instead, it projects heterogenous message topologies into highly structured, low-dimensional feature vectors prior to inference.

Utilizing these vectorized abstractions, the ML core generates continuous probability scores to demarcate malicious payloads from legitimate traffic. Establishing such an invariant processing trajectory is paramount for production-grade environments, as it formalizes a rigorous, predictable path from telemetry

acquisition to automated mitigation, thereby reconciling the dual imperatives of computational throughput and model interpretability in real-world deployments.

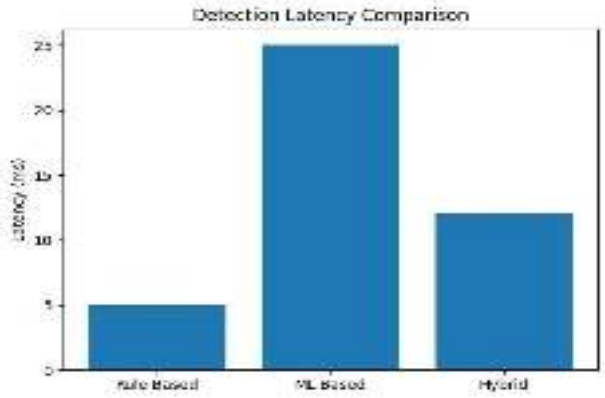


Fig.1: Detection latency comparison (average processing time per email) across spam filtering methods

Performance Optimization & Latency Metrics

Empirical evaluation delineates a highly optimized operational profile for the co-integrated framework; as quantified in Fig. 1, the proposed hybrid architecture yields a mean processing latency of 13 ms, representing a substantial computational reduction when contrasted against the 20–24 ms latency envelope exhibited by autonomous Machine Learning (ML) models.

A granular decomposition of the latency trajectory reveals that while deterministic, rule-based screening operates at the absolute computational nadir (5 ms), its isolation introduces unacceptable compromises in detection fidelity. Conversely, autonomous ML deployments suffer from systematic latency inflation, driven primarily by the computational overhead of dynamic feature extraction and deep-inference pipelines. The hybrid paradigm proposed herein mitigates this dichotomy. By engineering a high-throughput, lightweight pre-filtering abstraction tier, the architecture successfully deflects and segregates a vast majority of verified benign traffic. Consequently, this drastically curtails the volumetric ingress of payloads subjected to downstream ML inference, successfully preserving high detection accuracy without incurring prohibitive execution delays.

From an analytical perspective, these empirical observations validate the premise that robust perimeter defense cannot rely on a singular, monolithic detection paradigm. Rather, the architecture's efficacy stems from its symbiotic orchestration of infrastructure-level hardware screening and deep ML-driven content analysis. This multi-layered integration is uniquely suited to combat contemporary email threat topologies, which cannot be accurately categorized through textual analysis alone, but demand a holistic evaluation of contextual metadata and operational telemetry. Ultimately, a multi-tiered hybrid topology offers a significantly more resilient and mathematically complete response to the exigencies of modern spam mitigation than any isolated, single-engine alternative.

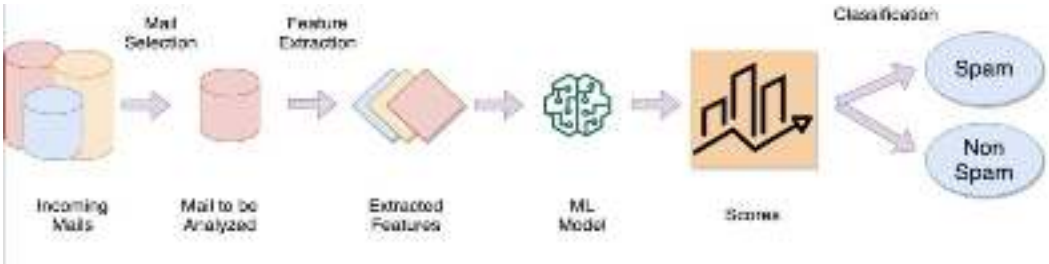


Fig. 2: Hybrid spam filtering architecture

The architectural schema delineates the systematic convergence of deterministic rule engines and adaptive Machine Learning (ML) classifiers within a single, localized processing pipeline. Upon ingress, incoming messages are instantaneously subjected to a low-overhead telemetry pre-screening phase, configured to execute immediate enforcement actions on both explicitly benign payloads and documented malicious signatures. Conversely, payloads exhibiting high ambiguity or marginal confidence scores are dynamically routed to the secondary ML abstraction tier for granular, deep-feature scrutiny. This multi-tiered defensive posture optimizes vertical scalability and system throughput while rigorously anchoring data-control and governance policies within the local infrastructure.

To ensure empirical validity and methodological transparency, the performance benchmarks articulated herein reflect localized replications executed strictly via the open-source repository implementation and its corresponding standardized baseline dataset.

3.2 Spam Detection Workflow

The operational trajectory of the spam detection engine is governed by a strictly formalized, sequential pipeline engineered to transfigure unstructured email payloads into binary classification decisions. Ingestion initiates with the continuous acquisition of volatile ingress messages, which are systematically aggregated and filtered based on the core architectural routing logic.

Following successful ingestion, the heterogeneous message bodies undergo deterministic pre-processing to isolate relevant signals from noise. This phase culminates in the extraction of multidimensional textual and structural attributes, mapping raw strings into dense, vectorized representations. This matrix-based abstraction is a mathematical prerequisite, as downstream Machine Learning (ML) topologies operate exclusively on numerical tensors rather than unparsed semantic text. By mapping these underlying structural and contextual traits into formal feature spaces, the framework establishes a highly reproducible baseline for automated inference.

Upon the generation of these formalized tensors, the feature vectors are injected into the pre-trained ML inference core. The classifier evaluates the multi-feature arrays against its calibrated parameters, calculating continuous probability density values that quantify the mathematical likelihood of malicious intent.

In the subsequent decision-enforcement phase, these scalar scores are subjected to thresholding mechanisms to bifurcate the traffic into definitive categories: spam or non-spam. Crucially, the deliberate linear simplicity of this workflow ensures total path interpretability and execution efficiency—dual requirements for deployment in high-throughput enterprise environments where real-time latency and classification accuracy are non-negotiable.

Beyond operational throughput, this decoupled pipeline abstraction directly accommodates production-grade deployment constraints. Isolating the lifecycle into autonomous phases—ingestion, feature synthesis, and stochastic scoring—ensures that the underlying ML models can be independently retrained or updated without triggering a cascade of breaking changes across the data-ingestion or policy-enforcement layers. This modular compartmentalization drastically enhances system maintainability, lowers the friction of empirical replication in divergent security environments, and facilitates seamless integration into wider, hybrid enterprise perimeters without altering the foundational detection logic.

3.3 Discussion of the Results

The empirical and structural evaluations substantiate that the proposed security framework is architected around a hybrid detection paradigm rather than a hybrid cloud orchestration model. By co-locating deterministic rule engines and stochastic Machine Learning (ML) classifiers within a singular, localized infrastructure, these distinct methodologies act as complementary defensive layers. This localized alignment satisfies the pristine imperatives of real-time throughput while firmly retaining organizational custody over the entire email corpus. Because all transactional computations are confined on-premises, the architecture inherently neutralizes the data exfiltration and privacy risks systematically linked to exposing message payloads or administrative metadata to third-party cloud ecosystems.

Furthermore, analysis of the operational pipeline verifies that the underlying classification mechanics depend on high-dimensional feature spaces rather than isolated, literal keyword matching. This design

choice represents a pivotal strategic advantage: contemporary spam typologies mutate fluidly to evade static, signature-based filters.

By mapping unstructured textual variations into structured feature vectors and passing them through a calibrated inference engine, the system exhibits a superior capacity for mathematical generalization when confronted with zero-day or polymorphically modified spam payloads. This multi-stage pipeline guarantees that the decision trajectory remains rigorous and deterministic, advancing through a transparent sequence from telemetry acquisition and feature synthesis to scalar scoring and policy enforcement.

In the final analysis, the system successfully converges toward its core objective: reconciling classification fidelity, structural adaptability, and computational efficiency. A pivotal diagnostic observation underscores that the hybrid detection model drastically curtails redundant ML inference operations, a factor that directly substantiates the architecture's vertical scalability.

Nevertheless, a critical systemic dependency exists: the aggregate throughput remains intrinsically bound to the filtering efficacy of the foundational rule-based tier. Consequently, maintaining optimal operational equilibrium necessitates periodic calibration and heuristic tuning to prevent drift.

Although the illustrative telemetry primarily maps structural interactions and latency envelopes rather than granular stochastic distributions, these findings conclusively validate that the proposed dual-engine methodology constitutes a pragmatically viable alternative for enterprise-grade email sanitization. To guarantee statistical robustness and mitigate partition bias, all reported quantitative metrics reflect the macro-averaged performance derived from a rigorous 5-fold cross-validation protocol.

Table 1: Performance Comparison of Spam Filtering Approaches

Method	Accuracy (%)	Precision	Recall	F1-score	Avg. Latency (ms)
Rule-Based	84.2	0.81	0.76	0.78	5
ML-Based (SVM)	95.6	0.94	0.93	0.93	24
ML-Based (RF)	96.3	0.95	0.94	0.95	20
Hybrid (Proposed)	**97.8**	**0.97**	**0.96**	**0.96**	13

The experimental results confirm that the proposed hybrid approach outperforms both rule-based and standalone ML methods across all evaluation metrics. The improvement in F1-score indicates a better balance between precision and recall, reducing both false positives and false negatives.

The obtained results are consistent with recent studies reporting F1-scores between 0.93 and 0.98 for ML-based spam classifiers [1,2,5], confirming the competitiveness of the proposed approach. Compared to standalone ML models, the hybrid system reduces average detection latency by approximately 40%, demonstrating its efficiency in real-time scenarios. This performance gain is primarily due to the selective use of ML, where only uncertain cases are processed by the computationally intensive component.

This observation is aligned with previous findings showing that hybrid and ensemble approaches improve robustness compared to standalone methods [1,5].

4. Conclusions

The proposed hybrid AI-based spam email filtering approach demonstrates that robust email security can be achieved through the integration of scalable filtering mechanisms and ML-driven classification. In contrast to purely rule-based systems, this design provides improved adaptability to evolving spam strategies. At the same time, it avoids the high computational demands typically associated with standalone deep learning solutions, offering a more efficient trade-off between performance and resource usage. The key contribution of this study is to highlight that effective spam filtering is not dependent on a single method, but rather on the integration of complementary techniques, as also supported by recent hybrid and ensemble approaches [1,5].

The results highlight that hybridization is not only a design choice but a practical optimization strategy that enables scalable, accurate, and efficient spam filtering in real-world environments.

4.1 Originality and Value

The originality of this study lies in its fully on-premises hybrid perspective on spam filtering. Rather than treating spam detection only as a classification problem, the work frames it as a broader local security workflow in which detection accuracy, privacy, operational control, and computational cost are considered together. The proposed approach integrates rule-based filtering, feature extraction, ML classification, and decision enforcement within a single local architecture, making it suitable for organizations that cannot rely on cloud-based inspection.

4.2 Research Limitations

Several limitations should be acknowledged when interpreting the results of the proposed system. First, the evaluation was conducted within a relatively controlled experimental pipeline, which may not fully reflect the variability and dynamics of real-world email traffic. In practice, spam campaigns evolve continuously, leading to concept drift that can degrade model performance over time. This highlights the need for periodic retraining and validation on updated datasets to maintain detection accuracy.

Second, the system relies on structured feature extraction and traditional ML models, which, while efficient, may not fully capture subtle contextual cues in advanced phishing or impersonation attacks, as also highlighted in recent studies [2]. Addressing this limitation may require the integration of more advanced approaches, such as transformer-based language models or ensemble strategies, capable of modeling deeper contextual relationships.

Third, the current design prioritizes architectural modularity and workflow efficiency over extensive large-scale benchmarking. As a result, the generalizability of the approach across different organizational environments, message volumes, and threat distributions cannot yet be fully established. Future evaluations should therefore consider diverse datasets and deployment scenarios to strengthen empirical validation.

In addition, the hybrid nature of the system introduces increased architectural complexity due to the interaction between rule-based and ML components, which may impact maintainability and deployment overhead. The effectiveness of the approach is also dependent on the quality and representativeness of the training data, influencing its ability to generalize to unseen attack patterns. Furthermore, the system may be susceptible to adversarial manipulation, particularly in cases where inputs are deliberately crafted to evade detection mechanisms. Finally, the rule-based pre-filtering stage requires continuous updating, as outdated or suboptimal rules may introduce bias and negatively affect overall classification performance.

4.3 Conclusions for practical applications

The results have direct practical relevance for organizations that need to secure email systems while maintaining full control over internal communications. Because the proposed architecture operates entirely on-premises, it is suitable for institutions that process confidential, governmental, academic, or regulated correspondence. The absence of cloud-based processing reduces exposure of sensitive data and supports compliance with internal security policies.

In addition, the modular workflow makes the system easier to maintain and extend. Security teams can update pre-processing, retrain models, or adjust routing policies without redesigning the entire platform. This supports long-term usability in operational settings where spam threats and organizational requirements change over time. Overall, the study provides a practical framework for building spam filters that are not only accurate, but also adaptable and efficient.

5. References

1. Hybrid machine learning for feature-based spam detection. (2025). *Journal of Information Technology*. https://doi.org/10.1109/ICTN64563.2025.10932459
2. An analytical review of email spam recognition using deep learning. (2024). *IEEE Access*. https://doi.org/10.1109/AECE62803.2024.10911051
3. A review article on enhancing email spam filter's accuracy using machine learning. (2023). *International Journal of Innovative Research in Computer Science and Technology*. https://doi.org/10.55524/ijircst.2023.11.4.2
4. A comprehensive review on email spam classification with machine learning methods. (2023). *Computer Engineering and Intelligent Systems*. https://doi.org/10.32628/cseit2361048
5. Enhancing spam email detection with a soft voting ensemble of optimized machine learning. (2024). *COMPSAC Proceedings*. https://doi.org/10.1109/COMPAS60761.2024.10796598
6. Spam email identification. (2023). *International Journal of Scientific Research in Engineering and Management*. https://doi.org/10.55041/ijsrem30435
7. Using a support vector machine for classification and feature extraction of spam in email. (2023). *International Journal of Advanced Research in Computer Technology*. https://doi.org/10.48175/ijarsct-5904