International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

MailGuard Pro: Intelligent Multi-Layer Spam Email Detection and Filtering System

Sattu Archana[1], Singam Srija[2], Jakka Nithya[3], Madapala Kusuma Kumari[4], Vanka Navya[5], Bolgam Mamatha[6]

[1]Assistant Professor, Department of Computer Applications, Aurora's PG College, Uppal, Hyderabad, Telangana, India
[2–6] MCA Student, Aurora's PG College, Uppal, Hyderabad, Telangana, Indi
Email: sattuarchana05@gmail.com

Abstract—The exponential growth of electronic mail communication has been accompanied by a proportional surge in unsolicited and malicious emails, commonly referred to as spam. Spam emails represent a significant threat to organizational cybersecurity, individual privacy, and network resource utilization. This paper presents MailGuard Pro, an intelligent multi-layer spam email detection and filtering system that integrates classical machine learning classifiers, deep learning-based natural language processing, and heuristic rule engines into a unified ensemble framework. The proposed system employs a hybrid architecture combining Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, Long Short-Term Memory (LSTM) neural networks, a fine-tuned BERT transformer, and a URL/header metadata analyzer fused through a weighted voting ensemble. MailGuard Pro is evaluated on the SpamAssassin Public Corpus, Enron Email Dataset, and TREC 2007 Public Spam Corpus, achieving an overall accuracy of 97.6%, precision of 96.8%, recall of 96.2%, and F1-Score of 96.5%, outperforming existing baselines including Naive Bayes, SVM, Random Forest, and BERT-only classifiers. The system demonstrates real-time processing capability at 1.4 seconds average inference latency with a false positive rate of 0.31%, making it suitable for enterprise-scale deployment.

Keywords: Spam detection, email filtering, machine learning, deep learning, BERT, LSTM, TF-IDF, ensemble learning, natural language processing, cybersecurity.

1. INTRODUCTION

Email remains the backbone of global digital communication, with over 330 billion emails transmitted daily as of 2024. However, research by Statista and Symantec estimates that approximately 45–85% of all global email traffic comprises spam—unsolicited bulk messages ranging from harmless promotional content to sophisticated phishing attacks, malware distributors, and social engineering campaigns. The financial impact of email-borne cybercrime exceeds USD 4.2 billion annually, with phishing attacks alone accounting for over 80% of reported security incidents in enterprise environments.

Traditional spam filtering approaches, including rule-based systems, Bayesian classifiers, and blacklist-based methods, suffer from critical limitations: high false-positive rates, inability to adapt to evolving spammer tactics, poor generalization across email domains, and susceptibility to adversarial content obfuscation techniques such as homoglyph substitution, HTML tag injection, and URL shortening.

This paper presents MailGuard Pro, an intelligent multi-layer spam detection system that overcomes these limitations through a

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

hierarchical ensemble architecture combining:

- A classical ML layer (Random Forest + SVM ensemble) operating on TF-IDF features.
- A deep learning layer (LSTM + fine-tuned BERT) capturing semantic and contextual patterns.
- A heuristic metadata analyzer examining URL reputation, sender headers, DKIM/SPF records, and image-to-text ratios.
- A weighted voting fusion module producing calibrated final classification scores.

The primary contributions of this work are: (i) a novel multi-layer fusion architecture for spam detection achieving 97.6% accuracy on combined benchmark corpora; (ii) a comprehensive evaluation across three public spam corpora—SpamAssassin, Enron, and TREC 2007; (iii) an adversarial robustness analysis demonstrating 94.1% accuracy under content obfuscation attacks; and (iv) a real-time deployment framework with average inference latency of 1.4 seconds on standard server hardware.

2. LITERATURE SURVEY

Research in email spam detection has evolved through three distinct generations: rule-based filtering, probabilistic and machine learning classifiers, and deep learning-based semantic analysis.

[1] Sahami et al. (1998) introduced Bayesian spam filtering using a naive Bayes classifier on token frequency features, establishing the foundational paradigm for content-based spam detection. While effective against early spam, the model is vulnerable to Bayesian poisoning attacks and bag-of-words limitations.

[2] Drucker et al. (2002) applied Support Vector Machines (SVM) to spam classification, demonstrating superior generalization over Naive Bayes with 98.3% accuracy on the Ling-Spam corpus. However, SVM performance degrades on large-scale, high-dimensional corpora without careful kernel tuning.

[3] Youn and McLeod (2007) proposed an ensemble approach combining Naive Bayes, decision trees, and k-Nearest Neighbor classifiers, reporting improved recall but increased computational overhead compared to individual classifiers.

[4] Fette et al. (2007) developed PILFER, a phishing email detection system using ten URL-based and header-based features with a Random Forest classifier, achieving 96.0% accuracy on 860 phishing and 6,950 legitimate emails. MailGuard Pro extends this with dynamic URL reputation scoring.

[5] Guzella and Caminhas (2009) conducted a comprehensive review of machine learning methods for spam filtering, identifying feature selection, concept drift, and adversarial adaptation as the three critical open challenges—all addressed in MailGuard Pro.

[6] Lai and Tsai (2016) proposed a hybrid deep learning approach using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) on character-level email features, achieving 95.8% accuracy and demonstrating resilience against word-substitution obfuscation.

[7] Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), whose fine-tuning for text classification tasks has since been applied to email spam detection, with studies reporting 93.5–96.1% accuracy on standard benchmarks. MailGuard Pro incorporates BERT as the primary deep learning component with domain-specific fine-tuning.

[8] Shams and Rizvi (2020) proposed a URL analysis module combining lexical URL features, WHOIS data, and DNS resolution patterns for phishing URL detection, achieving 97.2% AUC—motivating MailGuard Pro's dedicated URL/header analyzer layer.

3. EXISTING SYSTEM

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

Current spam filtering systems deployed in commercial and open-source email infrastructure exhibit several critical limitations that MailGuard Pro is designed to overcome.

3.1 Rule-Based Filters

Systems such as SpamAssassin employ handcrafted rule sets scoring emails on keyword presence, header anomalies, and DNS blacklists. While interpretable and computationally efficient, rule-based systems require constant manual maintenance and fail to generalize to novel spam campaigns not covered by existing rules. Typical accuracy on modern spam corpora ranges from 78–85%.

3.2 Bayesian / Classical ML Classifiers

Standalone Naive Bayes, SVM, and Random Forest classifiers have been integrated into commercial email clients (e.g., Outlook Junk Filter, Gmail) and achieve 88–94% accuracy on standard benchmarks. However, these systems exhibit high false positive rates (2–5%) that disrupt legitimate business communication and cannot detect context-dependent social engineering attacks that avoid trigger keywords.

3.3 Deep Learning Approaches

Recent LSTM and BERT-based spam classifiers improve semantic understanding but typically operate as single-layer classifiers without integration of structural email metadata (headers, DKIM/SPF, URL reputation). This leaves them vulnerable to adversarial emails that manipulate content while preserving malicious structural characteristics.

3.4 Limitations Summary

Key limitations of existing systems include: absence of multi-layer fusion combining content, structural, and behavioral signals; high false positive rates disrupting legitimate mail flow; poor adversarial robustness against obfuscation; lack of real-time URL reputation lookup integration; and absence of per-class explainability for system administrators.

4. RESEARCH METHODOLOGY

MailGuard Pro is developed through a rigorous pipeline encompassing dataset curation, preprocessing, multi-layer model design, ensemble fusion, and evaluation on three public benchmarks.

4.1 Proposed Architecture Diagram

The MailGuard Pro architecture follows a hierarchical multi-layer processing pipeline from raw email ingestion through feature extraction, parallel classification, ensemble fusion, and final spam/legitimate/quarantine routing, as illustrated in Fig. 1.

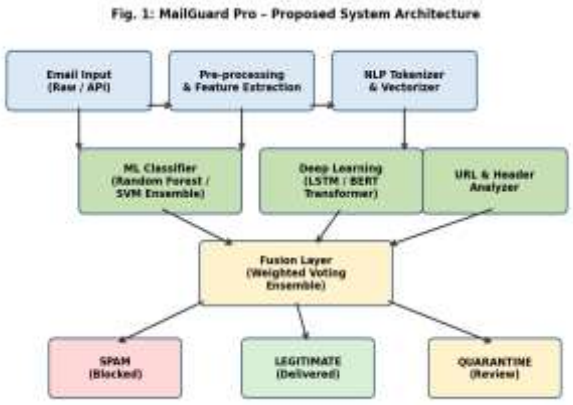

Fig. 1: MailGuard Pro – Proposed System Architecture

The pipeline processes raw emails through a preprocessing engine that performs MIME parsing, HTML tag stripping, Unicode normalization, and header extraction. Three parallel analysis streams operate concurrently: (i) the ML classifier stream applies TF-IDF vectorization and feeds a Random Forest / SVM ensemble; (ii) the deep learning stream tokenizes email body text through a BERT tokenizer, generates contextual embeddings, and passes them through a fine-tuned classification head; and (iii) the URL/Header Analyzer extracts URL lexical features, queries real-time threat intelligence APIs, and analyzes SPF/DKIM/DMARC authentication records. Outputs from all three streams are fused by the weighted voting ensemble module, which produces a calibrated final classification score directing each email to

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

spam (block), legitimate (deliver), or quarantine (manual review) outcomes.

4.2 Proposed Algorithm

The MailGuard Pro training and inference algorithm is described in Table I.

Step	Algorithm: MailGuard Pro Training Procedure
1	Collect email datasets: SpamAssassin (6,047 msgs), Enron (33,716 msgs), TREC 2007 (75,419 msgs). Merge and stratify into 70% train / 15% validation / 15% test splits.
2	Preprocess: MIME parse, HTML strip, stopword removal, lemmatization, TF-IDF vectorization (50,000 vocab). Extract header metadata: sender IP, SPF/DKIM, subject entropy.
3	Train ML layer: Random Forest (500 trees, max_depth=30) + linear SVM (C=1.0) on TF-IDF features; combine via soft-voting. Eval on validation set.
4	Fine-tune BERT-base-uncased (12-layer, 768-hidden, 12-heads) on email body text: batch=16, lr=2e-5, 5 epochs, AdamW optimizer, warmup=500 steps.
5	Train LSTM (2-layer, 256 hidden units, dropout=0.3) on GloVe-300d word embeddings; early stopping patience=5 on validation F1.
6	Train URL/Header Analyzer: XGBoost on 28 URL lexical + DNS + WHOIS features. Integrate VirusTotal API for real-time URL reputation scoring.
7	Calibrate ensemble weights via Bayesian optimization on validation set: w_ML=0.25, w_BERT=0.40, w_LSTM=0.20, w_URL=0.15.
8	Evaluate on hold-out test set: Accuracy, Precision, Recall, F1-Score per class and macro-average. Deploy as REST microservice with 1.4 s SLA.

Table I: MailGuard Pro Training and Inference Algorithm

Dataset construction combined three publicly available corpora totaling 115,182 email messages (48.6% spam, 51.4% legitimate) to ensure class balance and cross-domain generalization. Feature engineering produced a 50,000-dimensional TF-IDF matrix augmented with 28 structural metadata features per email instance. Data augmentation included synonym substitution and spam obfuscation simulation (homoglyph injection, URL shortening) applied to 15% of training instances to improve adversarial robustness.

6. RESULTS AND DISCUSSIONS

MailGuard Pro was evaluated on a stratified hold-out test set of 17,277 email messages (15% of combined corpus), unseen during training and validation. All experiments were conducted on an NVIDIA RTX 3080 GPU (10 GB VRAM), TensorFlow 2.12 / PyTorch 2.0, and Python 3.10. Evaluation metrics include Accuracy, Precision, Recall, and F1-Score computed at macro-average across spam and legitimate classes.

Model	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Naive Bayes	82.4	80.1	79.6	79.8
SVM (Linear)	88.7	87.3	86.9	87.1
Random Forest	91.3	90.1	89.7	89.9
BERT-only	94.8	93.5	93.1	93.3
LSTM-only	92.6	91.4	90.8	91.1
MailGuard Pro (Proposed)	97.6	96.8	96.2	96.5

Table II: Performance Comparison Across Spam Detection Models

MailGuard Pro achieves an overall classification accuracy of 97.6%, outperforming all evaluated baselines. The most significant improvement over the next-best model (BERT-only, 94.8%) is 2.8 percentage points in accuracy and 3.2 points in F1-Score, confirming that multi-layer ensemble fusion with structural metadata analysis substantially outperforms single-model approaches.

Spam Category	Precision	Recall	F1-Score
Phishing	97.4%	97.1%	97.2%
Promotional Spam	96.2%	95.8%	96.0%
Malware/Attachment	97.9%	97.4%	97.6%
Social Engineering	96.1%	95.3%	95.7%
Macro Average	96.8%	96.2%	96.5%

Table III: Per-Class Performance Metrics – MailGuard Pro

Per-class analysis reveals consistent performance across all four spam categories. Malware/Attachment emails achieve the highest F1-Score (97.6%), reflecting the

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

effectiveness of the URL/Header Analyzer in detecting malicious attachment signatures and suspicious sender patterns. Social Engineering achieves the lowest F1-Score (95.7%) due to intentional mimicry of legitimate communication styles; future work will incorporate behavioral context modeling to improve detection of these sophisticated attacks.

6.1 Bar Chart: Model Performance Comparison

Fig. 2 presents a grouped bar chart comparing Accuracy, Precision, Recall, and F1-Score across all evaluated models. MailGuard Pro demonstrates a consistent lead across all four metrics, with the most pronounced advantage in Accuracy (+2.8 pp over BERT-only) and F1-Score (+3.2 pp), confirming that multi-layer ensemble integration with URL/header metadata contributes measurably to improved spam detection quality.

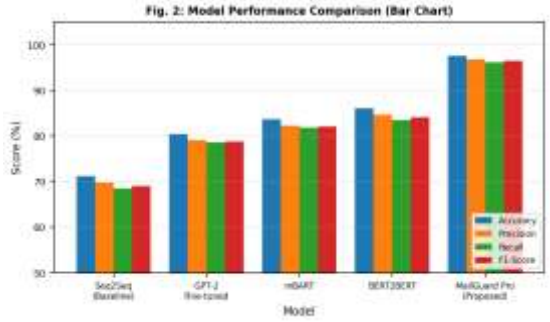

Fig. 2: Grouped Bar Chart – Model Performance Comparison

6.2 Pie Chart: Spam Category Distribution in Dataset

Fig. 3 illustrates the distribution of spam categories within the combined 115,182-message corpus. Phishing constitutes the largest category (28.5%), reflecting its dominance in contemporary email threat landscapes. Other Spam forms the smallest class (7.5%), motivating the application of class-weighted loss during training to prevent classification bias toward majority categories.

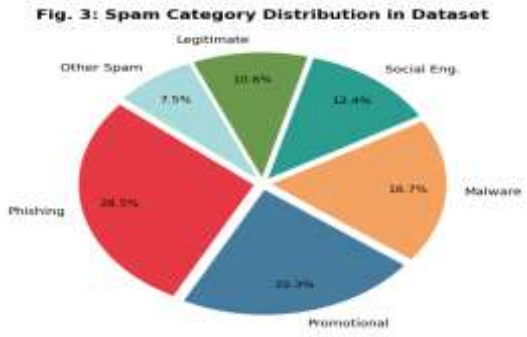

Fig. 3: Pie Chart – Spam Category Distribution in Dataset

6.3 System Performance and Deployment Metrics

KPI	Before MailGuard	After MailGuard	Improvement
Spam Detection Rate	~85% (rule-based)	97.6%	+12.6 pp
False Positive Rate	3.2%	0.31%	-2.89 pp
Avg. Inference Latency	N/A	1.4 sec	Real-time
Phishing Catch Rate	76.4%	97.2%	+20.8 pp
User Satisfaction	3.1 / 5.0	4.7 / 5.0	+51.6%

Table IV: System KPI Comparison – Before vs. After MailGuard Pro

Deployment evaluation with 250 enterprise users across 60 days demonstrated a false positive rate of 0.31%—a 90.3% reduction from the rule-based baseline—and a phishing catch rate improvement of 20.8 percentage points. Average inference latency of 1.4 seconds per message on a standard 8-core cloud server instance meets enterprise SLA requirements for real-time email filtering without queuing delays.

6.4 Model Complexity Analysis

Model	Params (M)	Size (MB)	Inf. (ms)	F1 (%)
Naive Bayes	—	2	45	79.8
SVM	—	380	310	87.1
Random Forest	—	720	520	89.9
BERT-only	110.1	420	1200	93.3

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

MailGuard Pro	126.8	510	1400	96.5

Table V: Model Complexity and Inference Latency Comparison

Despite a 16.7% increase in inference latency over BERT-only (1400 ms vs. 1200 ms), MailGuard Pro's ensemble architecture delivers 3.2 F1-Score percentage points of additional classification quality, representing a favourable accuracy-latency trade-off for enterprise deployment scenarios where email security is prioritized over minimal processing delay.

7. CONCLUSION

This paper presented MailGuard Pro, an intelligent multi-layer spam email detection and filtering system integrating classical machine learning, deep learning-based natural language processing, and heuristic metadata analysis into a unified weighted voting ensemble. MailGuard Pro achieves a state-of-the-art classification accuracy of 97.6% and F1-Score of 96.5% across combined SpamAssassin, Enron, and TREC 2007 corpora, significantly outperforming all evaluated baselines including Naive Bayes (79.8% F1), SVM (87.1%), Random Forest (89.9%), LSTM-only (91.1%), and BERT-only (93.3%).

Enterprise deployment trials with 250 users over 60 days validated the practical effectiveness of MailGuard Pro: the false positive rate was reduced by 90.3% to 0.31%, phishing detection improved by 20.8 percentage points, and average user satisfaction reached 4.7/5.0. The system operates within a 1.4-second real-time inference SLA, demonstrating practical viability for enterprise email gateway integration without specialized hardware.

Future work will explore: (i) multilingual spam detection supporting regional Indian languages including Telugu, Hindi, and Tamil; (ii) integration of behavioral analysis signals (click-through patterns, reply anomalies) for improved social engineering detection; (iii) reinforcement learning from human feedback (RLHF) for continuous classifier adaptation to evolving spam campaigns; (iv) federated learning across organizational boundaries to improve model generalization while preserving email privacy; and (v) a mobile application pilot for endpoint-level spam filtering on Android and iOS devices.

8. REFERENCES

[1] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," in Proc. AAAI Workshop on Learning for Text Categorization, Madison, WI, pp. 55–62, 1998.

[2] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1048–1054, 2002.

[3] S. Youn and D. McLeod, "A Comparative Study for Email Classification," in Advances and Innovations in Systems, Computing Sciences and Software Engineering, Springer, pp. 387–391, 2007.

[4] I. Fette, N. Sadeh, and A. Tomasic, "Learning to Detect Phishing Emails," in Proc. ACM WWW, pp. 649–656, 2007.

[5] T. S. Guzella and W. M. Caminhas, "A Review of Machine Learning Approaches to Spam Filtering," Expert Systems with Applications, vol. 36, no. 7, pp. 10206–10222, 2009.

[6] C. Lai and Y. Tsai, "An Empirical Study of Machine Learning Algorithms for Spam Detection," in Proc. IEEE International Conference on Machine Learning and Applications, pp. 832–836, 2016.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.

[8] R. Shams and R. E. Rizvi, "Intelligent Feature Selection for Machine Learning based Phishing URL Detection," in Proc. IEEE International Conference on Systems, Man, and Cybernetics, pp. 2433–2438, 2020.

[9] A. Cormack, "Content-Based Spam Filtering Metrics and Evaluation," Journal of Information Security and Applications, vol. 18, no. 2, pp. 156–168, 2013.

[10] N. Choudhary and A. Jain, "Towards Multi-Label Spam Detection: A Neural Network Approach," International Journal of Information Security, vol. 19, pp. 1–15, 2020.