IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

An Intelligent Email Spam Detection System Using Machine Learning Algorithms: A Systematic Review

Mehvish Tamboli[1], Prof. Shweta Dale[2], Sakshi Salke[3], Dr. V. S. Ranmalkar[4], Tamanna Shaikh[5] and Dr. P. M. Patil[6]
Department of Computer Applications (MCA)[1-6]
Adsul Technical Campus Chas, Ahilyanagar, Maharashtra, India

Abstract : *Unsolicited bulk email, commonly called spam, remains one of the most persistent nuisances and security risks on the internet, ranging from harmless advertising to large-scale phishing and malware-delivery campaigns. Because manual filtering cannot keep pace with the volume and evolving disguises of modern spam, organizations and mailbox providers rely heavily on artificial intelligence (AI) and machine learning (ML) to separate spam from legitimate ("ham") messages automatically. This paper presents a systematic review of AI- and ML-based email spam detection techniques, following a structured protocol comparable to PRISMA-style reviews. Studies drawn from digital libraries and preprint repositories were screened against explicit inclusion and exclusion criteria, and the resulting literature is organized into four technique families: rule-based and blocklist filtering, classical machine learning built on statistical text representations such as TF-IDF and bag-of-words, deep sequential architectures such as LSTM and CNN, and transformer-based language models such as BERT and its lightweight variants. The review consolidates reported performance figures, commonly used benchmark corpora, and evaluation metrics, and examines recurring challenges around concept drift, adversarial evasion, weak explainability, data-privacy compliance, and class imbalance in real-world mail streams. The analysis points to a steady progression from purely lexical filtering toward statistical classifiers and, more recently, toward transformer-based and hybrid ensemble pipelines that combine contextual embeddings with classical classifiers, which report the strongest accuracies in the surveyed literature. The paper closes by identifying open research gaps, including the need for temporally refreshed and adversarially robust benchmark datasets, lightweight explainable models suitable for real-time deployment, and proposes directions for future work in drift-aware, human-in-the-loop spam filtering.*

Key Words: email spam detection; machine learning; deep learning; natural language processing; phishing; concept drift; BERT; systematic review

1. INTRODUCTION

Electronic mail remains one of the cheapest and most widely used channels for both legitimate communication and unwanted, automated messaging. Industry telemetry from major mailbox providers indicates that a large share of daily email traffic is spam, and the largest providers report blocking well over ninety-nine percent of it before it reaches an inbox, underscoring the scale at which automated filtering already operates [6], [13]. Beyond simple advertising, spam increasingly carries phishing links, malware attachments, and business-email-compromise scams designed to steal credentials or money, which has turned spam filtering from a productivity concern into a frontline cybersecurity control [13], [19].

Copyright to IJARSCT
www.ijarsct.co.in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Early filters relied on manually curated blocklists and keyword or Boolean rules, an approach that is transparent and cheap to run but brittle against spammers who continuously reword or obfuscate their messages [5], [21]. This limitation motivated a shift toward statistical and machine-learning classifiers that learn the distinguishing patterns of spam directly from labelled examples rather than from hand-written rules, followed more recently by deep-learning and transformer-based models capable of capturing sequential and contextual structure in message text [3], [7].

This review surveys the AI/ML techniques proposed for automated email spam detection, covering rule-based filtering, classical supervised learning over statistical text representations, deep sequential architectures for parsing and classification, and transformer-based semantic models. It further consolidates the datasets and evaluation metrics reported across studies and examines recurring technical and ethical challenges. The remainder of the paper is organized as follows: Section II reviews background and related work, Section III describes the review methodology, Section IV surveys the principal technique families, Section V covers datasets and evaluation metrics, Section VI discusses challenges, Section VII offers a comparative discussion, Section VIII outlines future directions, and Section IX concludes.

II. BACKGROUND AND RELATED WORK

Automated spam detection sits at the intersection of text classification and information retrieval. The earliest widely deployed filters, such as blocklists of known spammer domains and Boolean keyword rules, are fast and auditable but cannot recognize paraphrased or intentionally misspelled trigger words, a weakness spammers routinely exploit through techniques such as substituting characters in flagged terms [5], [21].

To move beyond rigid rule sets, researchers turned to statistical classifiers trained on term-frequency–inverse-document-frequency (TF-IDF) or bag-of-words representations of message text. Naive Bayes classifiers, in particular, became a long-standing baseline because conditional word-probability models fit naturally with the way spammy vocabulary clusters together, and comparative studies repeatedly find Naive Bayes competitive with, or superior to, support vector machines (SVM), decision trees, and k-nearest neighbours on classic corpora [8], [9], [12]. Ensemble and tree-based methods such as random forests have also been reported to match or exceed Naive Bayes depending on the dataset and feature set used [10], [28].

A separate and more recent strand of work applies sequential deep-learning architectures, including long short-term memory networks (LSTM) and convolutional neural networks (CNN), which can model word order and local context rather than treating a message as an unordered bag of tokens [14]. Building further on this, transformer-based language models, most notably Bidirectional Encoder Representations from Transformers (BERT) and lighter derivatives such as DistilBERT, generate contextual embeddings that capture semantic relationships between words beyond raw lexical overlap, and have been fine-tuned specifically for spam and phishing classification [15], [16], [17]. Alongside these technical advances, a growing body of research studies how spam filters can be deliberately evaded through adversarial word substitutions and how filters trained on older corpora degrade against newer spam styles, concerns that motivate much of the challenges discussion in Section VI [13], [18], [19].

III. SYSTEMATIC REVIEW METHODOLOGY

This review follows a structured protocol modeled on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), adapted from its clinical origins to a computing and engineering literature base, and consisting of research-question formulation, search-string design, database querying, and staged screening against inclusion and exclusion criteria.

A. Research Questions

- **RQ1:** Which AI/ML techniques have been applied to automated email spam and phishing detection?
- **RQ2:** What datasets and evaluation metrics are used to assess these techniques, and what performance levels are reported?
- **RQ3:** What technical and operational challenges are most frequently reported, and how have researchers

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

proposed to mitigate them?

B. Search Strategy and Sources

Candidate studies were located using search strings that paired terms such as “spam detection,” “email classification,” “phishing detection,” and “anti-spam filtering” with technique-related terms including “machine learning,” “natural language processing,” “BERT,” “deep learning,” and “adversarial attack.” Sources spanned peer-reviewed journal articles, indexed conference papers, and preprint repositories, reflecting how quickly this subfield moves.

C. Inclusion and Exclusion Criteria

A study qualified for inclusion if it proposed, evaluated, or critically examined an AI/ML technique for email spam or phishing detection and provided enough methodological detail to characterize the technique and its results. Studies were excluded when they addressed unrelated spam domains without transferable methodology (for example, pure SMS or social-media spam with no email-specific findings), were not available in English, or amounted to promotional or vendor material lacking methodological substance.

D. Screening and Synthesis

Records were screened first by title and abstract and then by full text, removing duplicates and clearly off-topic entries at the first pass. The remaining studies were coded by technique family, dataset, evaluation metric, and reported limitation, and this coded information was synthesized narratively and comparatively across Sections IV through VII, mirroring the qualitative thematic-synthesis approach used in other recent AI-assisted systematic reviews of fast-moving computing subfields [1], [2].

IV. TECHNIQUES AND APPROACHES

A. Rule-Based and Blocklist Filtering

The earliest and still-common layer of spam defence matches sender addresses or domains against blocklists, or filters message content using keyword and Boolean rules. Such systems are cheap to run and easy to audit but cannot recognize semantically equivalent phrasing and are vulnerable to obfuscation, such as spammers deliberately misspelling trigger words to evade a filter's vocabulary [5], [21].

B. Classical Machine Learning with Statistical Text Representations

Much of the literature represents message text as TF-IDF or bag-of-words vectors and feeds it to supervised classifiers such as Naive Bayes, SVM, logistic regression, decision trees, and random forests. Reported accuracies for these pipelines commonly fall in the 83–99 percent range depending on dataset size and preprocessing, with Naive Bayes frequently matching or outperforming more complex classifiers on classic benchmark corpora, and random forest performing comparably in several studies [8]–[10], [12], [28]. This body of work suggests that for well-curated, moderately sized corpora, statistical representations paired with simple classifiers remain a practical and interpretable choice, and continue to serve as the baseline against which newer deep-learning methods are measured [24].

C. Sequential Deep Learning Models

Because spam and phishing messages often depend on word order, tone, and sentence-level structure rather than isolated keywords, several studies apply LSTM and CNN architectures that can model this sequential context. Comparative experiments report LSTM-based models reaching accuracies around 98 percent on TREC-style corpora, generally ahead of CNN and classical classifiers evaluated on the same data, though at a higher computational cost during training [14], [27]. Attention-augmented LSTM and hierarchical hybrid architectures have further improved recall on some benchmark sets, at the cost of added architectural and tuning complexity [14].

D. Transformer-Based and Embedding Models

Transformer architectures, chiefly BERT and its distilled variants, produce contextual embeddings that capture semantic relationships beyond surface word overlap. Fine-tuned BERT models trained across multiple classic spam corpora have been shown capable of a single, generalizable model performing acceptably on all of them simultaneously, rather than requiring a separate model per dataset [15]. In phishing-specific work, transformer models such as DistilBERT combined with adversarial training have been used to defend against evasion attempts while

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

retaining efficiency suitable for near-real-time filtering [20], and lightweight transformer variants have also been fine-tuned specifically to model the short, socially engineered language typical of business-email-compromise attempts [16]. These gains generally come with higher computational cost and reduced interpretability compared with classical methods [17], [20].

E. Hybrid and Ensemble Approaches

Several of the strongest reported results in this literature come from pipelines that combine transformer-derived contextual embeddings with structural or metadata features and feed the result into a classical or ensemble classifier such as random forest. One large-scale comparative study reports a multimodal architecture combining DistilBERT text representations with structural email features achieving the highest accuracy among all approaches tested, illustrating the broader pattern in this literature of blending deep semantic models with classical or feature-engineered signals rather than relying on either alone [24].

V. DATASETS AND EVALUATION METRICS

The surveyed studies rely heavily on a small set of long-standing public corpora: SpamAssassin, Enron-Spam, LingSpam, the PU-series corpora, GenSpam, and the TREC 2005/2007 spam tracks, several of which date back to the early-to-mid 2000s [22], [29], [33], [35]. Dataset sizes vary widely, from a few thousand messages in LingSpam to tens of thousands in TREC 2007 and Enron-Spam, and most exhibit some degree of class imbalance between spam and ham that studies address through resampling or class-weighted training [24], [35]. Because several of these corpora were compiled two decades ago, they no longer reflect current spammer tactics, a limitation raised repeatedly as a threat to external validity [3], [22], [24].

Evaluation metrics are generally consistent across the technique families: accuracy, precision, recall, and F1 score, often reported alongside a confusion matrix, with macro-averaged variants used when class imbalance is severe. Studies on phishing-specific detection additionally report metrics tuned toward minimizing false negatives, since a missed phishing email carries a materially higher cost than a missed promotional spam message [32], [43]. A smaller subset of studies also reports training or inference latency as a proxy for real-time deployability in high-volume mail systems [24].

VI. CHALLENGES AND CONSIDERATIONS

A. Concept Drift and Dataset Aging

Spammer tactics evolve continuously, so a model trained on an older corpus can degrade sharply when applied to contemporary spam. Empirical studies that split benchmark data into “classic” and “modern” eras confirm that both classical and transformer-based models trained on historical data lose substantial accuracy against newer messages, underscoring concept drift as a first-order threat to long-term filter reliability rather than a theoretical concern [24], [30], [31].

B. Adversarial Evasion

Because spam filters are economically motivated targets, spammers actively probe them for weaknesses. Documented evasion strategies include obfuscating flagged words through misspellings or character substitution, inserting words typical of legitimate mail to dilute spam signals, and, more recently, using generative language models to rewrite spam at the word, sentence, or paragraph level to bypass both classical and deep-learning filters [13], [18], [19], [38]. Studies show that even strong transformer-based classifiers remain vulnerable to these perturbations unless explicitly hardened through adversarial training [13], [20], [40].

C. Explainability and Transparency

As filtering pipelines move from transparent keyword rules to opaque deep neural and transformer models, it becomes harder for administrators or flagged senders to understand why a message was blocked. Recent work has begun pairing transformer classifiers with explanation methods to surface which tokens or features drove a phishing classification,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

though such explainability techniques remain less mature for email-specific tasks than for other text-classification domains [20], [43].

D. Data Privacy and Class Imbalance

Email content routinely contains personally identifiable information, so training and storing spam-classification datasets must be handled in line with applicable data-protection obligations, particularly when messages are retained for retraining production filters [3]. Separately, most public corpora exhibit skewed spam-to-ham ratios, which can bias naive classifiers toward the majority class; studies mitigate this through oversampling, class weighting, and threshold tuning rather than accuracy alone as the primary evaluation criterion [24], [29].

VII. COMPARATIVE DISCUSSION

Table I summarizes the technique families covered in this review along with their commonly reported strengths and limitations. Taken together, the surveyed studies point to a clear trajectory: purely lexical and rule-based filters have largely given way to statistical machine-learning classifiers, which are increasingly paired with, or replaced by, transformer-derived semantic embeddings, and the strongest reported results in the literature typically come from hybrid pipelines that combine both rather than any single method on its own [14], [15], [24].

A second observation is that reported performance figures should be interpreted cautiously: studies differ substantially in dataset age, class balance, and evaluation protocol, and a model reported as achieving very high accuracy on a decades-old, well-curated corpus may not generalize to the more heterogeneous and adversarial conditions of a live mail stream [22], [24], [30]. Finally, the technical and security-oriented strands of this literature have largely developed on separate tracks: accuracy-focused papers emphasize benchmark performance, while security-focused papers emphasize robustness to evasion; only a minority of surveyed studies evaluate both together, a gap this review identifies as significant [13], [24], [40].

TABLE I. Summary of AI/ML technique families for email spam detection.

Technique Category	Representative Methods	Reported Strengths	Reported Limitations
Rule-based / blocklist	Boolean keyword matching, sender/domain blocklists	Fast, simple, easy to audit	Misses paraphrasing/obfuscation; easily evaded
Classical ML + TF-IDF	Naive Bayes, SVM, Random Forest, kNN, logistic regression	Competitive accuracy (often 83–99%), cheap, interpretable	Sparse features miss semantic and contextual overlap
Sequential deep learning	LSTM, CNN, attention-augmented hybrids	Captures word order/context; up to ~98% accuracy	Higher training cost; needs larger labelled corpora
Transformer-based / LLM	BERT, DistilBERT, fine-tuned phishing-specific variants	Strong semantic understanding; generalizes across corpora	Costly to run; opaque; vulnerable without adversarial training
Hybrid / ensemble	Transformer embeddings + Random Forest / structural features	Highest reported accuracies (~99%+) in several studies	Added pipeline complexity and maintenance overhead

VIII. FUTURE RESEARCH DIRECTIONS

The gaps identified in this review point to several directions for further work. First, the field needs refreshed, temporally diverse benchmark datasets that reflect current spammer tactics, since continued reliance on decades-old corpora such as SpamAssassin and early Enron-Spam releases limits how confidently reported accuracies transfer to production settings [22], [24], [30]. Second, adversarial robustness should be evaluated alongside accuracy as a standard reporting requirement, given documented evasion techniques ranging from simple word obfuscation to large-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

language-model-generated spam rewrites [13], [18], [19]. Third, lightweight explainability methods suited to real-time filtering pipelines are needed so administrators and flagged senders can understand classification outcomes without materially slowing message delivery [20], [43]. Fourth, drift-aware and continual-learning approaches deserve further empirical study as a way to keep deployed filters current without full retraining cycles [24], [31]. Finally, as generative language models are used both to craft more convincing spam and, defensively, to power more capable classifiers, further research should examine this dual-use dynamic and its long-term implications for filter design [19], [20].

IX. CONCLUSION

This review examined AI and ML techniques for automated email spam detection, spanning rule-based filtering, classical machine learning over statistical text representations, sequential deep-learning models, and transformer-based semantic classifiers. The evidence points to a steady shift toward semantic, embedding-based, and hybrid ensemble methods, which report the strongest performance in the literature, while classical TF-IDF-based classifiers remain a competitive and considerably cheaper baseline. At the same time, technical progress has not resolved, and in some respects has complicated, longstanding concerns about dataset aging, adversarial evasion, limited explainability, and privacy-conscious data handling. Closing these gaps will likely require temporally refreshed and adversarially audited benchmark datasets, explainability methods suited to real-time filtering, and continued attention to the evolving, dual-use relationship between generative AI and spam itself.

REFERENCES

1. "Advancements in Email Spam Detection: A Systematic Review of Machine Learning and Deep Learning Techniques," Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), 2025.
2. "ReviewGenie: a novel automated system for systematic reviews—an exploratory study in speech and language disorders," PMC, 2025.
3. F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: Analysis of spammer strategies and the dataset shift problem," Artificial Intelligence Review, 56(2), pp. 1145–1173, 2023.
4. "Spam Detection in Emails Using Machine Learning Techniques: A Review," Int. J. Comput. Inf. Technol., 2024/2025.
5. 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, 5(6), e01802, 2019.
6. M. A. Aleisa et al., "Advancing Email Spam Classification using Machine Learning and Deep Learning Techniques," Eng. Technol. Appl. Sci. Res., 14(4), pp. 14994–15001, 2024.
7. "A Systematic Review of Deep Learning Techniques for Phishing Email Detection," Electronics (MDPI), 13(19), 3823, 2024.
8. D. Sculley and G. Wachman, "Effectiveness and Limitations of Statistical Spam Filters," arXiv:0910.2540, 2009.
9. "Machine Learning Approaches for Modeling Spammer Behavior," arXiv:1008.4249, 2010.
10. "Comparison of Algorithms on Machine Learning For Spam Email Classification," preprint, 2022.
11. "Efficient Spam Email Classification using Machine Learning Algorithms," preprint, 2023.
12. P. U. Anitha, C. V. Guru Rao, and Suresh Babu, "Email Spam Classification using Neighbor Probability based Naïve Bayes Algorithm," in Proc. 7th Int. Conf. Communication Systems and Network Technologies (CSNT), 2017.
13. T. Jaya, R. Kanyaharini, and B. Navaneesh, "Appropriate detection of ham and spam emails using machine learning algorithm," in Proc. Int. Conf. Advances in Computing, Communication and Applied Informatics (ACCAI), IEEE, 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

14. S. Youn and D. McLeod, "A Comparative Study for Email Classification," University of Southern California, tech. report.
15. "Email Spam Detection Using Hierarchical Attention Hybrid Deep Learning Method," arXiv:2204.07390, 2022.
16. V. S. Tida and S. Hsu, "Universal Spam Detection using Transfer Learning of BERT Model," arXiv:2202.03480, 2022.
17. "CATBERT: Context-Aware Tiny BERT for Detecting Social Engineering Emails," arXiv:2010.03484, 2020.
18. "An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach," arXiv:2311.04913, 2023.
19. B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, "Adversarial Feature Selection against Evasion Attacks," arXiv:2005.12154 (orig. IEEE Trans. Cybernetics, 2014).
20. "Adversarial Attacks Against Machine Learning Based Spam Filters," IEEE Conf. paper, Johns Hopkins Univ. Information Security Institute.
21. E. Hotoğlu, S. Sen, and B. Can, "A Comprehensive Analysis of Adversarial Attacks against Spam Filters," arXiv:2505.03831, 2025.
22. S. Bhowmick and S. M. Hazarika, "Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends," arXiv:1606.01042, 2016.
23. "Classification of Spam Emails through Hierarchical Clustering and Supervised Learning," arXiv:2005.08773, 2020.
24. "Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics: A Study of Concept Drift," arXiv:0812.1014, 2008.
25. "Temporal robustness of classical and transformer-based email spam classifiers under concept drift," Electronics (MDPI), 14(19), 3855, 2025.
26. "PhishSigma++: Malicious Email Detection with Typed Entity Relations," arXiv:2605.11619, 2026.
27. "A Comparison of Traditional Machine Learning Algorithms and LSTM-Based Deep Learning Models for Email Sentiment Analysis," arXiv:2605.03440, 2026.
28. "Explainable Transformer-Based Email Phishing Classification with Adversarial Robustness," arXiv:2511.12085, 2025.