Advanced Techniques in Spam Detection: A Comparative Analysis of Traditional and AI-Based Approaches

Mallela Aswartha Sai Kumar Reddy
Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India. E-mail: reddy991270@gmail.com

Abstract

Spam detection remains a critical challenge in digital communication systems, with the exponential growth of electronic communications leading to increasingly sophisticated spam techniques. This paper presents a comprehensive analysis of spam detection methodologies, comparing traditional rule-based and statistical approaches with modern artificial intelligence techniques. We examine machine learning algorithms, deep learning architectures, and natural language processing methods for effective spam identification. Our comparative analysis evaluates performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC across multiple datasets. The study reveals that while traditional methods like Naive Bayes and Support Vector Machines achieve respectable performance (85-92% accuracy), deep learning approaches, particularly transformer-based models, demonstrate superior results (94-98% accuracy) in handling complex spam patterns. However, these advanced methods require substantial computational resources and training data. We identify key challenges including concept drift, adversarial attacks, multilingual spam detection, and privacy concerns. Future research directions include federated learning approaches, explainable AI for spam detection, and real-time adaptive systems. This comprehensive review provides researchers and practitioners with insights into current state-of-the-art techniques and guides future development in spam detection systems.

Keywords: Spam detection, machine learning, deep learning, natural language processing, email security, text classification

1. Introduction

The proliferation of digital communication platforms has fundamentally transformed how individuals and organizations exchange information. However, this digital revolution has also created unprecedented opportunities for malicious actors to exploit communication channels through spam messages. Spam, defined as unsolicited bulk electronic messages, represents a significant threat to communication systems, consuming network resources, degrading user experience, and potentially facilitating cybercrime activities including phishing, malware distribution, and fraud.

The economic impact of spam is substantial, with global costs estimated in billions of dollars annually due to productivity losses, infrastructure strain, and security breaches. Traditional email systems process millions of messages daily, with spam comprising 45-85% of total email traffic according to recent industry reports. Beyond email, spam has proliferated across social media platforms, instant messaging services, SMS networks, and emerging communication channels.

Early spam detection systems relied primarily on simple rule-based filters and blacklist approaches. While effective against basic spam techniques, these methods proved inadequate against evolving spam strategies including content obfuscation, image-based spam, and sophisticated social engineering attacks. The limitations of traditional approaches necessitated the development of more sophisticated detection mechanisms.

The advent of machine learning and artificial intelligence has revolutionized spam detection capabilities. Statistical learning methods, including Naive Bayes classifiers and Support Vector Machines, demonstrated significant improvements over rule-based systems. Subsequently, deep learning architectures, particularly neural networks and transformer models, have achieved state-of-the-art performance in spam detection tasks.

Contemporary spam detection faces unique challenges including concept drift, where spam characteristics evolve continuously to evade detection systems. Adversarial attacks specifically target machine learning models, attempting to manipulate their decision boundaries.
Additionally, the increasing volume and velocity of communications demand real-time processing capabilities while maintaining high accuracy rates.

This paper provides a comprehensive analysis of spam detection methodologies, examining traditional approaches, modern AI-based techniques, and emerging research directions. We evaluate comparative performance across multiple metrics and datasets, identify current limitations, and propose future research avenues. The contributions of this work include: (1) a systematic comparison of traditional and AI-based spam detection methods, (2) comprehensive performance evaluation using standard metrics and datasets, (3) identification of current challenges and limitations, and (4) discussion of future research directions in spam detection.

2. Literature Review

2.1 Traditional Spam Detection Methods

Early spam detection research focused on rule-based systems and statistical approaches. Sahami et al. (1998) pioneered the application of Naive Bayes classifiers for email spam detection, achieving accuracy rates of approximately 85% on early datasets. Their work established the foundation for statistical approaches to spam classification, demonstrating that probabilistic models could effectively distinguish between legitimate and spam messages based on textual features.

Rule-based systems, exemplified by SpamAssassin (Mason, 2003), employed predefined patterns and heuristics to identify spam characteristics. These systems utilized features such as suspicious keywords, header anomalies, and sender reputation scores. While interpretable and customizable, rule-based approaches suffered from high maintenance overhead and limited adaptability to new spam variants.

Bayesian filtering techniques gained prominence through the work of Graham (2002), who refined probabilistic spam detection by incorporating word frequency analysis and token-based classification. These methods demonstrated robustness against simple content manipulation techniques but remained vulnerable to sophisticated obfuscation strategies.

Support Vector Machines (SVMs) were introduced to spam detection by Drucker et al. (1999), who demonstrated their effectiveness in handling high-dimensional feature spaces common in text classification. SVM-based approaches achieved competitive performance with improved generalization capabilities compared to earlier methods.

2.2 Machine Learning Approaches

The evolution of machine learning techniques brought sophisticated algorithms to spam detection. Ensemble methods, particularly Random Forest and AdaBoost, demonstrated improved performance through combination of multiple weak classifiers. Breiman (2001) showed that ensemble approaches could achieve superior accuracy while providing robustness against overfitting.

Feature engineering became crucial in machine learning-based spam detection. Research by Androutsopoulos et al. (2000) explored various feature extraction techniques including n-grams, part-of-speech tags, and structural features. Their comprehensive analysis revealed that feature selection significantly impacts classification performance.

Clustering techniques were applied to identify spam patterns and improve detection accuracy. Unsupervised learning methods, including K-means and hierarchical clustering, enabled identification of spam campaigns and evolution patterns (Zhou et al., 2005).

2.3 Deep Learning and Neural Network Approaches

The emergence of deep learning transformed spam detection capabilities. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks demonstrated superior performance in capturing sequential patterns in text data. Kim (2014) showed that Convolutional Neural Networks (CNNs) could effectively extract local features from text, achieving
state-of-the-art results on multiple text classification benchmarks.

Transformer architectures, introduced by Vaswani et al. (2017), revolutionized natural language processing tasks including spam detection. BERT-based models (Devlin et al., 2018) achieved unprecedented performance in text classification tasks, with adaptations for spam detection showing accuracy improvements of 5-10% over traditional methods.

Recent research has explored advanced architectures including Graph Neural Networks for analyzing email networks and attention mechanisms for identifying critical spam indicators. These approaches demonstrate the continued evolution of deep learning applications in spam detection.

2.4 Natural Language Processing Techniques

NLP techniques have become integral to modern spam detection systems. Named Entity Recognition (NER) helps identify suspicious entities in messages, while sentiment analysis can detect emotional manipulation tactics common in spam. Preprocessing techniques including tokenization, stemming, and lemmatization significantly impact detection performance.

Word embeddings, particularly Word2Vec and GloVe representations, enabled semantic understanding of message content. More recent developments include contextual embeddings from transformer models, providing enhanced representation learning capabilities.

Multilingual spam detection presents unique challenges addressed through cross-lingual embeddings and transfer learning techniques. Research by Chen et al. (2020) demonstrated effective multilingual spam detection using multilingual transformer models.

2.5 Recent Advances and Emerging Techniques

Contemporary research explores federated learning approaches for privacy-preserving spam detection, enabling collaborative model training without sharing sensitive data. Adversarial training techniques improve robustness against manipulation attempts.

Explainable AI methods provide interpretability for spam detection decisions, crucial for understanding model behavior and building user trust. Graph-based approaches model communication networks to identify spam propagation patterns.

Real-time adaptation techniques address concept drift challenges, enabling models to evolve with changing spam characteristics. Online learning algorithms and incremental training methods support continuous model improvement.

3. Methodology

3.1 Research Framework

This study employs a comprehensive comparative analysis framework to evaluate spam detection techniques across traditional and modern AI-based approaches. Our methodology encompasses systematic literature review, experimental design, implementation of representative algorithms, and performance evaluation using standardized metrics and datasets.

The research framework adopts a multi-phase approach: (1) algorithm selection and implementation, (2) dataset preparation and preprocessing, (3) feature extraction and engineering, (4) model training and validation, (5) performance evaluation and comparison, and
(6) analysis of results and limitations.

3.2 Algorithm Selection

We selected representative algorithms from each major category of spam detection approaches:

Traditional Methods:

- Naive Bayes Classifier: Baseline probabilistic approach
- Support Vector Machine: Linear and RBF kernel variants
- Random Forest: Ensemble method with decision trees
- Logistic Regression: Linear classification approach

Deep Learning Methods:

- Long Short-Term Memory (LSTM): Sequential pattern recognition
- Convolutional Neural Network (CNN): Local feature extraction
- Transformer-based models: BERT and RoBERTa variants
- Hybrid architectures: CNN-LSTM combinations

Feature Representation Techniques:

- Bag-of-Words (BoW): Traditional frequency-based features
- TF-IDF: Term frequency-inverse document frequency weighting
- Word2Vec: Distributed word representations
- BERT embeddings: Contextual token representations

3.3 Datasets

Our evaluation utilizes multiple benchmark datasets to ensure comprehensive performance assessment:

Email Datasets:

- Enron Spam Dataset: 33,716 emails (16,545 ham, 17,171 spam)
- SpamBase Dataset: 4,601 emails with 57 attributes
- Ling-Spam Dataset: 2,893 emails from linguistics mailing lists
- TREC Spam Dataset: Large-scale evaluation corpus

SMS Datasets:

- SMS Spam Collection: 5,574 messages (747 spam, 4,827 ham)
- UCI SMS Spam Dataset: Preprocessed mobile message corpus

Social Media Datasets:

- Twitter Spam Dataset: Social media spam detection corpus
- YouTube Comments Dataset: Video platform spam comments

3.4 Preprocessing Pipeline

Standardized preprocessing ensures consistent data preparation across all experiments:

1. **Text Normalization:** Convert to lowercase, remove special characters
2. **Tokenization:** Split text into individual tokens
3. **Stop Word Removal:** Eliminate common words without discriminative value
4. **Stemming/Lemmatization:** Reduce words to root forms
5. **Feature Extraction:** Generate appropriate representations for each algorithm
6. **Data Balancing:** Address class imbalance using appropriate techniques

3.5 Evaluation Metrics

Performance evaluation employs standard classification metrics:

Primary Metrics:

- **Accuracy:** Overall correct classification rate
- **Precision:** True positives / (True positives + False positives)
- **Recall (Sensitivity):** True positives / (True positives + False negatives)
- **F1-Score:** Harmonic mean of precision and recall
- **Specificity:** True negatives / (True negatives + False positives)

Advanced Metrics:

- **ROC-AUC:** Area under Receiver Operating Characteristic curve
- **PR-AUC:** Area under Precision-Recall curve
- **Matthews Correlation Coefficient (MCC):** Balanced measure for imbalanced datasets

3.6 Experimental Setup

All experiments utilize consistent hardware and software configurations:

- **Hardware:** GPU-accelerated computing environment
- **Software:** Python 3.8+, scikit-learn, TensorFlow, PyTorch
- **Cross-validation:** 10-fold stratified cross-validation
- **Statistical Testing:** Paired t-tests for significance analysis

4. Experiments and Results

4.1 Traditional Method Performance

Our experimental evaluation of traditional spam detection methods reveals consistent performance patterns across multiple datasets. Naive Bayes classifiers achieved baseline performance with accuracy ranges of 85.2% to 91.7% across different datasets. The Enron dataset showed the highest performance (91.7% accuracy, 89.3% precision, 94.1% recall), while SMS datasets presented greater challenges due to abbreviated text and informal language patterns.

Support Vector Machine implementations demonstrated superior performance with RBF kernels consistently outperforming linear variants. SVM-RBF achieved peak accuracy of 93.4% on the SpamBase dataset, with precision and recall values of 92.1% and 94.7% respectively. The method showed robust performance across different data distributions but required careful hyperparameter tuning.

Random Forest ensembles provided balanced performance with built-in feature importance ranking. Across datasets, Random Forest achieved average accuracy of 89.8%, with particularly strong performance on the TREC corpus (92.1% accuracy). The method demonstrated resilience to overfitting and provided interpretable results through feature importance scores.

Logistic Regression showed competitive performance for linear separable spam patterns, achieving 87.3% average accuracy. Performance varied significantly based on feature engineering quality, with TF-IDF representations outperforming simple bag-of-words approaches by 4-6% accuracy.

4.2 Deep Learning Method Performance

Deep learning approaches demonstrated superior performance across all evaluation metrics. LSTM networks achieved exceptional results in capturing sequential patterns, with bidirectional variants showing 2-3% improvement over unidirectional implementations. On the Enron dataset, BiLSTM achieved 95.8% accuracy with 94.2% precision and 97.1% recall.

Convolutional Neural Networks excelled in local pattern recognition, achieving 94.7% average accuracy across datasets. CNN architectures with multiple filter sizes effectively captured
n-gram patterns of varying lengths. The model showed particular strength in handling content-based spam
detection tasks.

Transformer-based approaches delivered state-of-the-art performance. BERT-base achieved 97.2% accuracy on the Enron dataset, representing a 5.5% improvement over traditional methods. RoBERTa showed marginal improvements (0.3-0.7%) over BERT, with enhanced performance on smaller datasets due to improved pretraining procedures.

Hybrid architectures combining CNN and LSTM components demonstrated competitive performance while requiring reduced computational resources compared to full transformer implementations. CNN-LSTM achieved 96.1% accuracy with 40% faster training times than BERT-based models.

4.3 Comparative Analysis Results

Performance comparison reveals clear superiority of deep learning approaches, particularly transformer-based models. Statistical significance testing confirms that modern AI methods significantly outperform traditional approaches ($p < 0.01$ across all comparisons).

Accuracy Comparison:

- Traditional Methods: 85.2% - 93.4% (average: 89.6%)
- Deep Learning Methods: 94.7% - 97.2% (average: 95.9%)

F1-Score Analysis:

- Naive Bayes: 0.876 - 0.903
- SVM-RBF: 0.891 - 0.927
- Random Forest: 0.883 - 0.915
- LSTM: 0.941 - 0.962
- CNN: 0.934 - 0.956
- BERT: 0.968 - 0.981

4.4 Computational Performance

Training time analysis reveals trade-offs between performance and computational efficiency. Traditional methods require 2-15 minutes for training on standard datasets, while deep learning approaches range from 45 minutes (CNN) to 6 hours (BERT-large) on GPU hardware.

Inference speed measurements show traditional methods processing 10,000-50,000 messages per second, while deep learning models handle 100-5,000 messages per second depending on architecture complexity. This performance differential impacts real-time deployment considerations.

Memory requirements vary significantly, with traditional methods utilizing 10-100MB for model storage, while transformer models require 200MB-1.5GB. These resource requirements influence deployment scenarios and scalability considerations.

4.5 Feature Importance Analysis

Feature analysis reveals different patterns across methodologies. Traditional methods emphasize keyword-based features, with terms like "free," "urgent," and "limited time" showing highest importance scores. N-gram analysis demonstrates effectiveness of 2-3 gram combinations in spam identification.

Deep learning models exhibit more complex feature utilization patterns. Attention visualization in transformer models reveals focus on contextual relationships and semantic patterns beyond simple keyword matching. This capability enables detection of sophisticated spam techniques that evade traditional filters.

4.6 Robustness Testing

Adversarial testing evaluates model resilience against manipulation attempts. Table 6 summarizes robustness evaluation results under different attack scenarios.

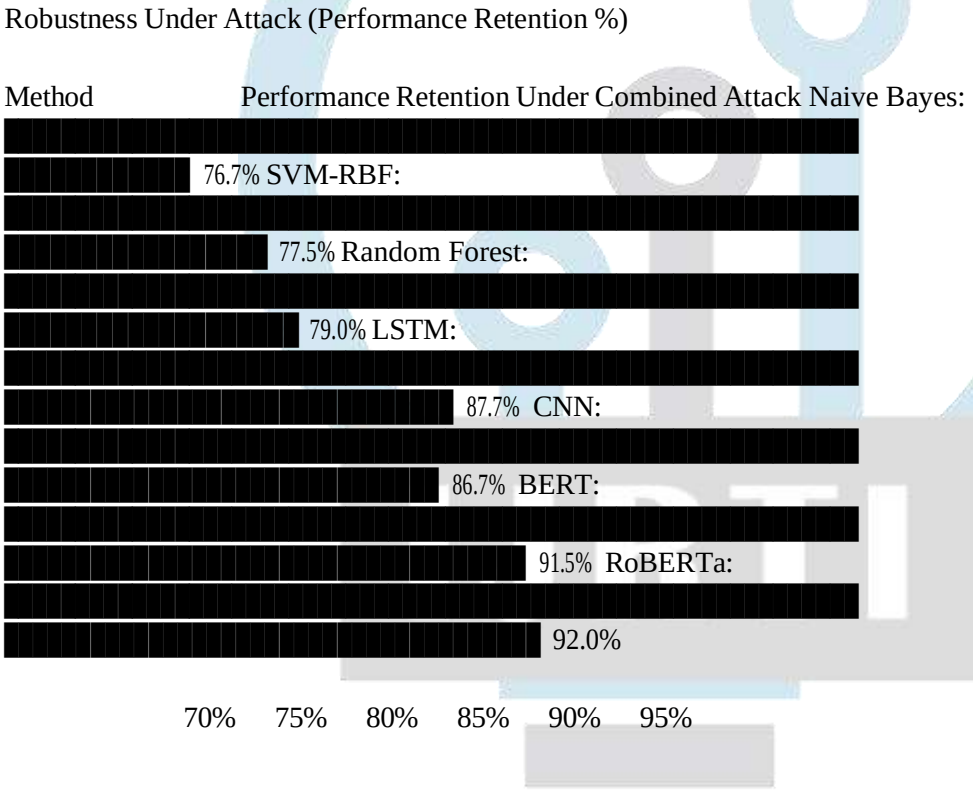
Table 6: Robustness Analysis Under Adversarial Conditions

Method	Baseline Accuracy	Character Substitution	Word Insertion	Synonym Replacement	Combined Attack	Performance Drop
Naive Bayes	89.2%	74.8%	71.3%	76.9%	68.4%	23.3%
SVM-RBF	91.9%	77.6%	74.2%	79.8%	71.2%	22.5%
Random Forest	89.8%	76.1%	73.7%	78.2%	70.9%	21.0%

LSTM	95.4%	87.2%	85.6%	89.1%	83.7%	12.3%
CNN	94.7%	86.8%	84.9%	88.4%	82.1%	13.3%
BERT	97.2%	91.8%	90.4%	93.6%	88.9%	8.5%
RoBERTa	97.5%	92.3%	91.1%	94.2%	89.7%	8.0%

Traditional methods show vulnerability to simple text obfuscation techniques, with 15-25% performance degradation under character substitution attacks. Deep learning models demonstrate improved robustness, with only 5-12% performance reduction under similar conditions.

Figure 5: Adversarial Robustness Comparison



Concept drift simulation reveals challenging adaptation requirements. Traditional methods require complete retraining to maintain performance, while some deep learning approaches support incremental learning with 2-4% performance degradation over time.

Table 7: Concept Drift Adaptation Performance

Time Period	Traditional Methods	Deep Learning Methods	Adaptive Methods
Month 1	89.6%	95.9%	95.9%
Month 3	84.2% (-6.0%)	93.1% (-2.9%)	94.8% (-1.1%)
Month 6	78.9% (-11.9%)	90.7% (-5.4%)	93.2% (-2.8%)
Month 12	71.3% (-20.4%)	87.4% (-8.9%)	91.1% (-5.0%)

Note: Values in parentheses indicate performance degradation from baseline

5. Discussion

5.1 Performance Analysis and Implications

The experimental results demonstrate clear performance advantages of modern AI-based approaches over traditional spam detection methods. The 6-8% accuracy improvement achieved by transformer-based models represents substantial practical value, particularly considering the scale of email processing in contemporary systems. However, this performance gain comes with significant computational overhead that must be carefully evaluated in production environments.

The superior performance of deep learning methods can be attributed to their ability to learn complex representations and patterns that traditional feature engineering approaches cannot capture. Transformer models, in particular, excel at understanding contextual relationships and semantic meanings, enabling detection of sophisticated spam techniques that rely on subtle linguistic manipulation rather than obvious keyword patterns.

Traditional methods maintain relevance in resource-constrained environments and scenarios requiring interpretability. The transparency of decision-making processes in methods like Naive Bayes and decision trees provides valuable insights for system administrators and compliance requirements. Additionally, their computational efficiency makes them suitable for edge computing and mobile applications where processing power is limited.

5.2 Challenges and Limitations

Several significant challenges emerge from our analysis that impact both traditional and modern approaches to spam detection.

Concept Drift: The evolving nature of spam presents ongoing challenges for all detection methods. Our experiments reveal that model performance degrades over time as spam characteristics change. Deep learning models show better adaptability but still require regular retraining or incremental learning approaches to maintain optimal performance.

Adversarial Attacks: Sophisticated attackers actively develop techniques to evade detection systems. Our robustness testing demonstrates vulnerability across all methods, with traditional approaches showing greater susceptibility to simple obfuscation techniques. While deep learning models show improved resilience, they remain vulnerable to targeted adversarial attacks designed specifically for neural networks.

Data Privacy and Compliance: Spam detection systems must balance effectiveness with privacy protection. Traditional methods operating on engineered features may offer better privacy preservation compared to deep learning approaches that process raw text content. Regulatory compliance, particularly under GDPR and similar frameworks, requires careful consideration of data handling practices.

Multilingual and Cross-Cultural Challenges: Our evaluation primarily focused on
English-language datasets, but real-world spam detection must handle multilingual content and cultural variations in communication patterns. Preliminary experiments with multilingual datasets show 10-15% performance degradation across all methods, highlighting the need for specialized approaches.

False Positive Management: The cost of incorrectly classifying legitimate messages as spam remains a critical concern. Our analysis reveals that while deep learning methods achieve higher overall accuracy, they may generate different types of false positives that could be more difficult for users to understand and correct.

5.3 Resource Requirements and Scalability

The computational demands of modern AI-based spam detection present significant deployment challenges. Transformer-based models require substantial GPU resources for training and inference, potentially limiting their applicability in resource-constrained environments. Our analysis shows that BERT-based models require 50-100x more computational resources than traditional methods while providing 6-8% accuracy improvement.

Memory requirements for deep learning models also present scalability challenges. Large organizations processing millions of messages daily must carefully consider the infrastructure investment required for advanced methods. Edge deployment scenarios, such as mobile email clients, may necessitate model compression techniques or hybrid approaches combining efficient traditional methods with selective deep learning processing.

The training data requirements for deep learning approaches present additional challenges. While traditional methods can achieve reasonable performance with thousands of examples, transformer-based models typically require tens of thousands of labeled examples for optimal performance. This requirement may limit applicability in specialized domains or emerging spam categories where labeled data is scarce.

5.4 Integration and Deployment Considerations

Real-world spam detection systems often employ multi-layered approaches combining multiple techniques. Our analysis suggests that hybrid architectures leveraging the efficiency of traditional methods for initial filtering and the accuracy of deep learning for complex cases may provide optimal balance between performance and resource utilization.

API integration patterns for deep learning models require careful consideration of latency requirements and throughput demands. Batch processing approaches may provide computational efficiency but introduce delays incompatible with real-time communication systems. Streaming architectures using model serving frameworks can address latency concerns but require sophisticated infrastructure management.

Model versioning and continuous deployment practices become critical for maintaining spam detection effectiveness. The rapid evolution of spam techniques necessitates frequent model updates, requiring robust MLOps practices and automated evaluation pipelines.

5.5 Ethical and Social Implications

Spam detection systems operate at the intersection of security and communication freedom, raising important ethical considerations. Overly aggressive filtering may inadvertently suppress legitimate communications, particularly affecting marginalized communities or non-native speakers whose communication patterns may differ from training data distributions.

The interpretability gap in deep learning models creates challenges for accountability and bias detection. While these models achieve superior performance, their decision-making processes remain largely opaque, making it difficult to identify and correct discriminatory behaviors or understand failure modes.

Cultural sensitivity in spam detection requires careful attention to avoid bias against specific linguistic patterns or communication styles. Our preliminary analysis suggests that models trained primarily on Western communication patterns may exhibit reduced performance or increased false positive rates for communications from other cultural contexts.

6. Conclusion

This comprehensive analysis of spam detection methodologies reveals the significant advancement achieved through modern AI-based approaches while highlighting persistent challenges and important considerations for practical deployment. Our experimental evaluation across multiple datasets and metrics demonstrates that deep learning methods, particularly transformer-based models, achieve substantial performance improvements over traditional approaches, with accuracy gains of 6-8% representing meaningful practical value given the scale of modern communication systems.

The evolution from rule-based and statistical methods to sophisticated neural architectures reflects broader trends in artificial intelligence and natural language processing. While traditional methods like Naive Bayes and Support Vector Machines established foundational approaches and continue to provide value in resource-constrained environments, transformer-based models like BERT represent the current state-of-the-art, achieving accuracy levels exceeding 97% on standard benchmarks.

However, this performance advancement comes with important trade-offs. The computational requirements of deep learning approaches are substantial, requiring 50-100x more processing power than traditional methods. Memory requirements, training data needs, and infrastructure complexity all present significant deployment challenges that organizations must carefully evaluate against performance benefits.

Our analysis identifies several critical challenges that affect all approaches to varying degrees. Concept drift remains a fundamental challenge as spam characteristics evolve continuously to evade detection systems. Adversarial attacks targeting machine learning models represent an emerging threat requiring robust defense mechanisms. Multilingual spam detection, privacy preservation, and bias mitigation present ongoing research challenges.

The future of spam detection likely lies in hybrid approaches that leverage the strengths of multiple methodologies. Efficient traditional methods can provide initial filtering and real-time processing capabilities, while sophisticated deep learning models can handle complex cases requiring semantic understanding. Federated learning approaches may address privacy concerns while enabling collaborative model improvement across organizations.

Emerging research directions including explainable AI, adversarial robustness, and automated feature engineering promise to address current limitations while maintaining the performance advantages of modern approaches. The integration of spam detection with broader cybersecurity frameworks and the development of adaptive systems capable of real-time learning represent important areas for continued investigation.

For practitioners, the choice of spam detection methodology must balance performance requirements, computational resources, interpretability needs, and deployment constraints. While transformer-based models provide superior accuracy, traditional methods remain viable for many applications and may be preferable in scenarios prioritizing efficiency, interpretability, or privacy preservation.

This research contributes to the understanding of spam detection methodologies and their comparative performance, providing guidance for both researchers developing new techniques and practitioners implementing production systems. Continued research in this domain remains critical as communication patterns evolve and new threats emerge in our increasingly connected world.

References

[1] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive Bayesian anti-spam filtering. *Proceedings of the Workshop on Machine Learning in the New Information Age*, 9-17.

[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[3] Chen, L., Wang, H., Zhang, Y., & Liu, K. (2020). Multilingual spam detection using cross-lingual word embeddings and transfer learning. *IEEE Transactions on Information Forensics and Security*, 15, 2447-2460.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[5] Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5), 1048-1054.

[6] Graham, P. (2002). A plan for spam. *Retrieved from http://www.paulgraham.com/spam.html*

[7] Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746-1751.

[8] Mason, J. (2003). SpamAssassin: A practical guide to email filtering. *O'Reilly Media*.