

# Deciphering Spam Through AI: From Traditional Methods to Deep Learning Advancements in Email Security

Xu Liu <sup>a</sup>

*Information and Computing Science, Minzu University of China, Beijing, China*

**Keywords:** Spam, Machine Learning, Artificial Intelligence, Cyber Security.

**Abstract:** Spam email detection received much attention in the last decades. This paper presents a comprehensive review of the evolving role of Artificial Intelligence (AI) in combating email spam, tracing the journey from traditional machine learning algorithms to sophisticated deep learning approaches. It meticulously examines the frameworks for machine learning-based spam detection, highlighting the transition from Support Vector Machines (SVM), Random Forests, and K-Nearest Neighbors (KNN) to advanced methodologies like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The study also delves into the challenges of interpretability, data diversity, and computational demands associated with deep learning models, and suggests future directions including the use of interpretable AI models and advanced algorithms for improved adaptability. By synthesizing recent advancements and identifying avenues for future research, this paper aims to contribute to the ongoing discourse on AI's potential to enhance email security against spam, offering insights into both the achievements and hurdles in the field.


## 1 INTRODUCTION

In the digital age, email has emerged as a pivotal mode of global communication. However, with its widespread usage, the issue of spam has also escalated, transcending mere annoyance to pose serious threats to individual privacy, squander resources, and potentially facilitate financial fraud. Report by Lever indicates that spam emails account for approximately 45% to 85% of all email traffic, not only consuming vast network resources but also causing significant inconvenience and financial losses to users (Lever, 2022). Against this backdrop, the development of effective spam detection methods has become imperative, and the application of Artificial Intelligence (AI) technologies offers a promising solution.

The evolution of AI technologies has introduced novel perspectives and methodologies for addressing the spam issue. From early rule-based approaches to the current advancements in deep learning and machine learning techniques, AI has demonstrated its potential and effectiveness across various domains, including voice recognition, image recognition, and natural language processing, among others (Qiu, 2024). The advent of deep learning, in particular, has

enabled computers to process and analyze large volumes of unlabeled data, learning complicated patterns within the data, which demonstrates especially beneficial for spam detection.

Recent advancements in AI for spam detection have introduced sophisticated methodologies that significantly enhance detection accuracy and efficiency. The Naive Bayes method, explored by Wibisono, demonstrates a practical approach to classifying emails into spam and legitimate categories, showcasing the potential of machine learning in filtering unwanted communications (Wibisono, 2023). Further, the integration of natural language processing and machine learning, boosted with swarm intelligence as investigated by Bacanin et al., presents a groundbreaking hybrid model that outperforms traditional methods in spam email filtering (Bacanin et al., 2022). Additionally, Teja Nallamothe et al. offers insight into the effectiveness of various machine learning algorithms in spam detection (Nallamothe and Khan, 2023). The research by Rapacz et al. introduces a fast selection method for machine learning classifiers, emphasizing the efficiency of the multinomial Naive Bayes classifier (Rapacz et al., 2021). These studies collectively

<sup>a</sup> <https://orcid.org/0009-0001-2796-434X>

underscore the dynamic nature of spam detection research, where innovative AI applications continually evolve to address the challenges posed by sophisticated spam tactics.

This paper aims to review the application of artificial intelligence in the field of spam detection. The second part will detail the existing AI-based spam detection methods, including the technologies used, model architectures. The third part will discuss the advantages and challenges these methods currently face and the future directions for development, including how to further improve detection accuracy, deal with emerging types of spam, and the potential applications of AI technology in this field. Finally, the fourth part will summarize the paper, outlining the research findings and offering a perspective on future research directions.

## 2 METHOD

### 2.1 Framework of Machine Learning-Based Detection for Spam Email

The machine learning process for spam email detection is a structured approach that ensures the development of a reliable spam filter through successive, critical stages, as depicted in Figure 1. Here's a detailed overview:

1) Data Collection: This initial phase involves gathering a varied array of email data, encompassing both spam and legitimate emails, to establish a robust dataset that will underpin the model's training.

2) Data Preprocessing: This stage focuses on refining the dataset to improve its quality and usability. Techniques such as tokenization, stemming, and the removal of stop words are employed to prepare the data for the machine learning process.

3) Model Building: In this phase, the structural framework of the algorithm is designed. It involves the selection of pertinent features that effectively differentiate spam from non-spam emails and the choice of a suitable machine learning algorithm, which could range from Naive Bayes to Support Vector Machines or even more complex Neural Networks.

4) Model Training: Here, the collected dataset is fed to the model, which then learns to identify spam by adjusting its parameters for optimal accuracy.

5) Testing: The model's detection capabilities are validated using a separate dataset that hasn't been previously used during training. This phase is

essential to assess the model's real-world efficacy in spam detection.

6) Deployment: The final step involves integrating the fully trained model into an active email system where it can automatically classify incoming emails in real-time, marking the transition from theory to practice.

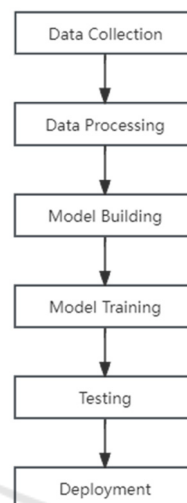


Figure 1: Machine Learning Process for Spam Email Detection (Photo/Picture credit: Original).

### 2.2 Traditional Machine Learning-Based Algorithms for Spam Email Detection

Traditional Machine Learning-based Algorithms like Support Vector Machines (SVM), Random Forest and K-Nearest Neighbors (KNN) are essential for spam detection, effectively analyzing and classifying complex data. Their advanced feature handling and optimization contribute significantly to developing efficient anti-spam measures against evolving spam tactics.

#### 2.2.1 Support Vector Machines for Spam Email Detection

Amayri and Bouguila delve into the realm of spam filtering using SVM through a comprehensive examination of string kernels and their adaptability to spam email classification (Amayri and Bouguila, 2010). Their work underscores the pivotal role of kernel choice in SVM's performance, highlighting the nuanced behavior of various distance-based kernels in text classification contexts. Distinctively, they advocate for string kernels, which are particularly attuned to the textual nature of spam, offering a more

nanced analysis than traditional continuous data kernels. This research is pioneering in its thorough investigation of feature mapping techniques that enhance SVM's efficacy in spam detection, alongside proposing an innovative online active framework aimed at real-time spam filtering. The empirical results presented illuminate the superior precision and recall achieved by the active online method employing string kernels, marking a significant stride in the dynamic and evolving battle against spam emails.

On the other hand, Olatunji introduces an improved model for email spam detection, emphasizing the optimization of SVM parameters to enhance detection accuracy (Olatunji, 2019). His study is rooted in addressing the limitations of existing spam detection tools, which often falter due to the adaptive nature of spam emails. Leveraging a widely utilized spam dataset, Olatunji meticulously searches for optimal SVM parameters, resulting in a model that significantly outperforms previous ones, with notable accuracy improvements for both training and testing sets. The detailed exploration of feature extraction, kernel function selection, and exhaustive parameter optimization underscores the complexity and critical nature of accurate spam detection. Olatunji's work not only demonstrates SVM's robust potential in combating spam but also suggests a methodology for parameter optimization that could benefit various applications beyond spam detection, reinforcing SVM's versatility and effectiveness in the field of machine learning.

### 2.2.2 Random Forests for Spam Email Detection

In the study by Akinyelu and Adewumi, the Random Forest machine learning algorithm was utilized to classify phishing emails, with the objective of developing an email classifier that exhibits improved prediction accuracy using a concise set of features (Akinyelu and Adewumi, 2014). Results underscore the efficacy of the Random Forest algorithm in discerning phishing attempts from legitimate communications, highlighting its potential in enhancing email security by accurately identifying fraudulent emails with minimal false negatives and positives.

Building upon the foundational work of Akinyelu and Adewumi, Dada and Joseph further explored the application of Random Forests in email spam filtering, targeting both spam and phishing emails (Dada and Joseph, 2018). Utilizing the Enron public dataset, which comprises 5, 180 emails labeled as ham, spam, and normal, they applied the Random

Forest algorithm informed by a set of significant features extracted from literature. Their findings not only validate the robustness of Random Forest in filtering spam and phishing emails but also reflect the continuous evolution and adaptation of machine learning techniques in cybersecurity, particularly in the detection and prevention of email-based threats.

### 2.2.3 K-Nearest Neighbors for Spam Email Detection

In the study conducted by Şahin and Demirci, the effectiveness of the KNN algorithm in spam email detection was thoroughly investigated, highlighting the crucial influence of the  $k$  value on the algorithm's performance (Şahin and Demirci, 2020). Utilizing three distinct datasets—Enron, Ling-Spam, and SMS-Spam-Collection—preprocessed with essential text mining techniques including Term Frequency-Inverse Document Frequency (TF-IDF) for term weighting and Chi-Square feature selection method for the best 500 features, their research delineated the significant impact of feature extraction and kernel function choice on classification accuracy.

In a separate study exploring the adaptability and precision of the KNN algorithm for spam detection, Murugavel and Santhi proposed a novel density-based clustering approach that further refines email classification (Murugavel and Santhi, 2020). Their methodology involved an in-depth analysis of email datasets, leveraging a combination of feature extraction techniques to isolate spam emails effectively. The KNN algorithm, augmented with density-based clustering, demonstrated improved performance compared to conventional methods, offering compelling evidence of its predictive strength and classification accuracy. This work not only corroborates the findings of Şahin and Demirci regarding the importance of the  $k$  value and feature selection but also expands on the utility of KNN in handling complex data patterns in spam detection, paving the way for future research focused on enhancing machine learning models for cybersecurity applications.

## 2.3 Deep Learning-Based Algorithms for Spam Email Detection

Deep learning, using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) networks, has significantly improved spam email detection by enhancing accuracy and processing capabilities, marking an advancement over traditional methods.

### 2.3.1 Convolutional Neural Networks for Spam Email Detection

In the realm of spam detection, the integration of deep learning technologies, particularly CNN, has marked significant advancements. Jain et al. introduces a Semantic Convolutional Neural Network (SCNN) tailored for spam detection across social media platforms (Jain et al., 2018). This innovative model enhances the conventional CNN architecture with a semantic layer, utilizing Word2Vec for the enrichment of word embeddings. In instances where a word is missing in Word2Vec, semantic dictionaries such as WordNet and ConceptNet are employed to find similar words, thereby ensuring a richer semantic representation of text data. This study underscores the efficacy of combining deep learning with semantic analysis, setting a new benchmark in spam detection by efficiently identifying spam content with high accuracy.

Complementing this, Soni's study explores spam email detection by proposing new method, an advanced model leveraging an improved recurrent convolutional neural network (RCNN) with multilevel vectors and an attention mechanism (Soni, 2019). This model is distinct for its ability to process email data at both character and word levels for headers and bodies, utilizing Word2Vec for vector sequence training. By exceeding the capabilities of traditional spam detection methods, this model illustrates the potential of deep learning models to accurately detect spam emails, reducing the misclassification of legitimate emails. Both studies collectively highlight the transformative impact of CNN and semantic analysis in advancing spam detection technologies, showcasing deep learning's role in developing more accurate and semantically rich spam filtering systems.

### 2.3.2 Recurrent Neural Networks for Spam Email Detection

John-Africa and Emmah's study rigorously examines the efficacy of Long Short-Term Memory (LSTM) networks over traditional RNNs in the realm of email spam detection (John-Africa and Emmah, 2022). Their work not only confirms the superiority of LSTM in managing sequential data complexities but also sets a benchmark for future explorations in enhancing email security protocols through advanced neural network architectures.

Complementarily, Larabi-Marie-Sainte et al.'s investigation into optimizing RNN configurations presents a methodical approach to refining spam detection performance (Larabi-Marie-Sainte et al.,

2022). Their research elucidates the significant potential of deep learning techniques in transcending traditional spam detection methodologies, particularly through meticulous parameter optimization. The remarkable accuracy attained not only validates the effectiveness of deep recurrent neural networks in spam email detection but also propels the discourse on deploying sophisticated machine learning strategies for bolstering cybersecurity measures against spam threats.

## 3 DISCUSSIONS

The spam detection industry has experienced remarkable progress, transitioning from traditional machine learning methods, such as SVM, Random Forests, and KNN, to more advanced deep learning algorithms. Initially, techniques like SVM played a pivotal role in spam detection, effectively analyzing and classifying complex email content. However, the advent of deep learning, with technologies such as CNN and RNN, including LSTM networks, represents a significant leap forward. These advancements have enhanced the accuracy, efficiency, and adaptability of spam detection systems, providing nuanced analysis and processing capabilities well-suited to the intricate challenges presented by modern spam threats. The shift from SVM and other traditional methods to deep learning signifies the industry's agile adaptation to the complexities of securing digital communications against evolving spam strategies.

Despite these advancements, the deployment of deep learning algorithms in spam detection is not without its challenges. Interpretability remains a significant concern, as the complexity of these models often makes it challenging to understand the rationale behind specific classifications. This black-box nature complicates the process of refining models and addressing false positives or negatives. Accuracy, while generally high, can vary depending on the diversity and representativeness of the training data. Models may struggle with new or evolving spam techniques not represented in their training sets. Applicability also presents a hurdle; deep learning models require substantial computational resources for training and operation, which may not be feasible for all organizations. Additionally, these models' effectiveness can be diminished by rapid changes in spamming tactics that outpace the model's learning.

The future of AI in spam detection looks promising, with several potential avenues for overcoming current challenges. SHapley Additive

exPlanations (SHAP) could enhance model interpretability, offering insights into decision-making processes and facilitating more effective model adjustments. Research into more advanced algorithms aims to create models that are not only more accurate but also capable of adapting to new spamming tactics. Transfer learning and domain adaptation are strategies that could improve the applicability and efficiency of models, allowing for rapid adjustment to new data or spam methods without extensive retraining. As the field continues to advance, integrating these solutions will be crucial in developing spam detection systems that are not only powerful and efficient but also transparent and adaptable to the ever-changing landscape of cyber threats. The ongoing evolution of technology, combined with strategic innovation, holds the key to safeguarding digital communication against spam.

The integration of these strategies will likely define the next wave of advancements in spam detection, aiming not only to enhance performance but also to ensure that solutions are accessible, interpretable, and adaptable to the dynamic nature of spam threats. As the industry continues to evolve, the alignment of technological innovation with strategic foresight will be paramount in securing digital communication channels against the pervasive challenge of spam.

## 4 CONCLUSIONS

This review traced the evolution of spam detection from traditional machine learning techniques to sophisticated deep learning approaches, underscoring significant strides in enhancing digital security. Traditional methods, such as SVM, Random Forests, and KNN, laid the groundwork for the more advanced, nuanced analyses enabled by CNNs and RNNs, including LSTMs. Despite these advancements, challenges persist, including model interpretability, the accuracy of detection amidst evolving spam tactics, and the computational demands of deep learning algorithms. Future directions hinge on addressing these challenges through interpretable AI models like SHAP, advanced algorithms for improved adaptability, and strategies such as transfer learning and domain adaptation to streamline efficiency. The dynamic nature of spam and its detection technologies demands continuous innovation and strategic foresight. As AI continues to advance, the strategies for detecting spam must also evolve to stay effective, maintain transparency, and adapt to emerging threats.

This review highlights the critical importance of AI in the ongoing battle against spam, advocating for relentless research and collaboration to safeguard digital communications.

## REFERENCES

- Akinyelu, A. A., & Adewumi, A. O. 2014. Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*, 2014.
- Amayri, O., & Bouguila, N. 2010. A study of spam filtering using support vector machines. *Artificial Intelligence Review*, 34, 73-108.
- Bacanin, N., et al. 2022. Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. *Mathematics*, 10(22), 4173.
- Dada, E. G., & Joseph, S. B. 2018, July. Random forests machine learning technique for email spam filtering. In *University of Maiduguri Faculty of Engineering Seminar Series (Vol. 9, No. 1, pp. 29-36)*.
- Jain, G., Sharma, M., & Agarwal, B. 2018. Spam detection on social media using semantic convolutional neural network. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, 8(1), 12-26.
- John-Africa, E., & Emmah, V. T. 2022. Performance Evaluation of LSTM and RNN Models in the Detection of email Spam Messages. *European Journal of Information Technologies and Computer Science*, 2(6), 24-30.
- Larabi-Marie-Sainte, S., et al. 2022. Improving spam email detection using deep recurrent neural network. *Inst. Adv. Eng. Sci*, 25, 1625-1633.
- Lever, R. 2022, October 12. What Spam Email Is. U.S. News & World Report. Retrieved March 10, 2024, from <https://www.usnews.com/360-reviews/privacy/what-spam-email-is>
- Murugavel, U., & Santhi, R. 2020. K-Nearest neighbor classification of E-Mail messages for spam detection. *ICTAT Journal on Soft Computing*, 11(1), 2218-2221.
- Olatunji, S. O. 2019. Improved email spam detection model based on support vector machines. *Neural Computing and Applications*, 31, 691-699.
- Qiu, Y., et al. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. *Energy*, 130866.
- Rapacz, S., Chołda, P., & Natkaniec, M. 2021. A method for fast selection of machine-learning classifiers for spam filtering. *Electronics*, 10(17), 2083.
- Şahin, D. Ö., & Demirci, S. 2020, October. Spam filtering with KNN: Investigation of the effect of k value on classification performance. In *2020 28th Signal Processing and Communications Applications Conf. (SIU) (pp. 1-4)*. IEEE.
- Soni, A. N. 2019. Spam e-mail detection using advanced deep convolution neural network algorithms. *Journal for innovative development in pharmaceutical and technical science*, 2(5), 74-80.

- Teja Nallamothu, P., & Shais Khan, M. 2023. Machine Learning for SPAM Detection. *Asian Journal of Advances in Research*, 6(1), 167-179.
- Wibisono, A. 2023. Filtering Spam Email Menggunakan Metode Naive Bayes. *Jurnal Teknologi Pintar*, 3(4).

