Taj Al-Ma'rifa Journal مجلة تاج المعرفة
ISSN : 3104-8846, ISSN-L: 3104-8838

On the evaluation of email spam filters in adversary environment

*** Lutfia Milod**
Higher Institute of Sciences and Technology Computer Department
Gharian, Libya
Lutfiaramadan78@gmail.com

[2] abdulmawla Najih
Higher Institute of Sciences and Technology Computer Department
Gharian, Libya
nabdulmawla@gmail.com

[3] Nooreddin Hemidat
Higher Institute of Sciences and Technology Computer Department
Gharian, Libya
Nuri.hamidatt@gmail.com

[4] Rehab. M.Ibrahem
Higher Institute of Sciences and Technology Computer Department
Gharian, Libya
rehab@gmail.com

*Corresponding Author: * Lutfia Milod

Keyword
adversary environment , machine learning, Random forest , ROC curve, spam filters.

Abstract
This paper proposes a novel framework for assessing the robustness of email spam filters against adversarial attacks. We introduce a methodology to quantify vulnerability by simulating attacker strategies that deliberately modify the training and/or test sets to maximize performance degradation. This degradation is measured using the Area Under the Curve (AUC) within the high-sensitivity region (0.9-1.0) of the Receiver Operating Characteristic (ROC) curve**.** Our experimental results demonstrate that Random Forest-based filters exhibit superior robustness compared to existing approaches, representing a significant advancement in adversarial-resistant spam filtering. The proposed evaluation framework provides critical insights into filter vulnerabilities under optimal attack conditions, addressing a key gap in current spam filter assessment methodologies.

Received : 24/03/2026 Accepted : 30/03/2026

DOI: https://doi.org/10.64943/jkc.2026.040204

تقييم مرشحات البريد الإلكتروني المزعج في بيئة معادية

*** لطفية ميلود**
المعهد العالي للعلوم والتقنية غريان
Lutfiaramadan78@gmail.com

[2] عبدالمولى الناجح
المعهد العالي للعلوم والتقنية غريان
nabdulmawla@gmail.com

[3] نورالدين حميدات
المعهد العالي للعلوم والتقنية غريان
Nuri.hamidatt@gmail.com

[4] رجاب محمد إبراهيم
المعهد العالي للعلوم والتقنية غريان
rehab@gmail.com

***الباحث المرسل:** * لطفية ميلود

الكلمة المفتاحية: البيئة المعادية , التعلم الالي , منحني ROC , الغابات العشوائية , مرشحات البريد المزعج.

الملخص

يقدم هذا البحث إطارًا جديدًا لتقييم متانة مرشحات البريد الإلكتروني العشوائي (السخام) في مواجهة الهجمات الخبيثة. نقوم بتقديم منهجية جديدة لقياس حدود القابلية للاختراق من خلال محاكاة استراتيجيات المهاجم التي تقوم بتعديل مجموعة التدريب و/أو مجموعة الاختبار بشكل استراتيجي لتعظيم تدهور الأداء، والذي يتم قياسه باستخدام مقياس AUC في منطقة الحساسية العالية (0.9-1.0) لمنحنى ROC. تظهر نتائجنا التجريبية أن مرشحات Random Forest الغابة العشوائية) تُظهر متانة فائقة مقارنة بالأساليب الحالية، مما يمثل تقدمًا كبيرًا في مجال تصفية البريد العشوائي المقاوم للهجمات .يوفر إطار التقييم المقترح رؤى حاسمة حول نقاط الضعف في المرشحات تحت ظروف الهجوم المثلى، مما يعالج فجوة مهمة في منهجيات تقييم مرشحات البريد العشوائي الحالية.

تاريخ الإستقبال: 2026/03/24 تاريخ القبول: 2026/03/30

DOI: https://doi.org/10.64943/jkc.2026.040204

1.INTRODUCTION

Spam email, also known as unsolicited bulk email (UBE) or junk email, refers to unsolicited messages sent to recipients without their consent. The problem of spam has persisted and evolved over many years. in computing terms, means something unwanted. It has normally been used to refer to unwanted email or Usenet messages, and it is now also being used to refer to unwanted Instant Messenger (IM) and telephone Short Message Service (SMS) messages. Spam email is unwanted, uninvited, and inevitably promotes something for sale. Often the terms junk email, Unsolicited Bulk Email (UBE), or Unsolicited Commercial Email (UCE) are used to refer to spam email. Spam generally promotes Internet – based sales, but it also occasionally promotes telephone- based or other methods of sales too [10].

To rule out spam automatically from user's email, the task of spam filter is used. Many works have been used to filter spam. These works used one of two

approaches: the rule based method which generates a set of rules that classify an email as spam or legitimate. The other approach is to use text mining or machine learning approach which considers spam filtering as two-class text classification that classifies a message as spam or legitimate. The basic and common concept of machine learning approaches is that using a classifier to filter out spam and the classifier is learned from training data rather than constructed by hand. Therefore, it can result in better performance. From the machine learning viewpoint, spam filtering based on the textual content of e-mail can be viewed as a special case of text categorization, with the categories being spam or non-spam. [2,7].

1.1MACHINE LEARNING METHODS RELEVANT FOR SPAM FILTERING

Spam filtering is a canonical text classification problem in the domain of machine learning (ML) and natural language processing (NLP). The objective is to train a model to automatically assign an incoming message (email, SMS) to one of two classes: **spam** (unwanted, malicious) or **ham** (legitimate). Modern spam filters are typically built using a supervised learning pipeline, which involves **feature extraction** (converting text into a numerical representation) and **model training & classification** (learning discriminative patterns from the data) [6].

1.2 FEATURE EXTRACTION: FROM TEXT TO NUMERICAL VECTORS

A fundamental requirement for ML models is that input data must be numerical. The process of converting raw text into a feature vector is called vectorization.

- **Bag-of-Words (BoW):**

  The Bag-of-Words (BoW) model represents text as an unordered multiset of words, disregarding grammar and word order, by creating a vocabulary from the training corpus and encoding each document as a vector of word frequencies or presence indicators, though this approach results in a high-dimensional, sparse feature space and fails to capture semantic meaning or contextual relationships between words [9].

- **Term Frequency-Inverse Document Frequency (TF-IDF):**

  Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word to a document within a corpus by calculating the product of its Term Frequency (TF) how often it appears in the

document and its Inverse Document Frequency (IDF) which penalizes words that are common across many documents, thereby emphasizing discriminative terms that are frequent in specific classes but rare globally, such as "FREE" in spam emails, making it a standard feature weighting technique in information retrieval and text mining [14].

N-grams:

An N-gram is a contiguous sequence of N items (e.g., words) from a text, where using unigrams, bigrams, and trigrams as features helps capture local word order and meaningful phrases, such as bigrams like "click below" or "free offer," which are highly indicative of spam and would be overlooked by a unigram model, thereby improving contextual representation in text classification [8].

- **Word Embeddings**

Word Embeddings (e.g., Word2Vec, GloVe) map words into a high-dimensional continuous vector space, positioning semantically similar words—such as "V1@gr@" and "viagra"—closely together, thereby enabling deep learning models to capture nuanced semantic relationships and improve robustness against word obfuscation in tasks like spam filtering, though they are less common in traditional high-throughput filters [12].

1.3. CLASSIFICATION ALGORITHMS

Once emails are converted into feature vectors, they are used to train a classifier.

A. Traditional Supervised Learning Models

These models are renowned for their performance, efficiency, and interpretability, making them staples in production systems.

- **Naïve Bayes (The Classic Choice):**

Naïve Bayes is a historically significant algorithm for spam filtering, famously employed in early systems like CRM114, which applies Bayes' Theorem under the "naïve" assumption that all features (words) are conditionally independent given the class label; while this assumption is often violated in text (e.g., "wire" and "transfer" are not independent), the algorithm remains highly popular due to its computational efficiency, scalability, and surprisingly strong performance on text classification tasks even when this simplification does not hold [6, 1].

- **Logistic Regression:**

  Logistic Regression is a widely-used linear classification model in industry that estimates the probability of a sample belonging to a specific class, offering advantages such as high efficiency, well-calibrated probability estimates, and excellent performance with large, sparse feature spaces, which often allows it to outperform Naïve Bayes on modern spam datasets and makes it both a strong baseline and a core component of many contemporary spam filters [15].

Support Vector Machines (SVM):

Support Vector Machines (SVM) are highly effective for text classification tasks such as spam filtering, as they identify the optimal hyperplane that maximally separates different classes in a high-dimensional space; while they offer advantages like robustness in high-dimensional settings, memory efficiency (relying on support vectors), and resistance to overfitting, they can suffer from high computational cost on large datasets and provide less interpretability than probabilistic models like Logistic Regression [5 , 16].

- **Decision Trees and Random Forests:**

  Decision Trees and their ensemble extension, Random Forests, are relevant for spam filtering as the former learns hierarchical if-then-else rules based on features, offering high interpretability and the ability to model non-linear boundaries with minimal data preprocessing, while the latter combines multiple decorrelated trees to enhance accuracy and reduce overfitting, though this comes at the cost of increased computational complexity and slower prediction times compared to a single tree [3,13].

B. Advanced and Modern Approaches

- **Neural Networks / Deep Learning:**

  Neural networks and deep learning models are highly relevant for spam filtering due to their ability to automatically learn hierarchical feature representations and complex non-linear patterns from text; architectures such as Recurrent Neural Networks (RNNs/LSTMs) excel at capturing long-range dependencies and contextual sequences, while Convolutional Neural Networks (CNNs) effectively identify informative local patterns and key phrases, enabling state-of-the-art performance against sophisticated spam—though these advantages come at the cost of requiring extensive labeled data, significant computational resources, and offering limited model interpretability [11].

1.3. THE ML PIPELINE AND CONTINUOUS LEARNING

A production spam filtering system functions as a dynamic machine learning pipeline that begins with data collection and labeling to build a representative corpus, followed by text preprocessing and feature engineering (e.g., TF-IDF vectorization) to transform emails into feature vectors, then proceeds to model training and evaluation with an emphasis on high precision to minimize false positives before deployment, and finally implements continuous learning through active learning strategies where low-confidence predictions are reviewed by humans and used to periodically retrain the model, thereby creating an adaptive feedback loop that evolves against new spam tactics [6, 15].

2. RELATED WORK

The challenge of spam filtering has been extensively addressed through machine learning, evolving from foundational probabilistic models to modern deep learning approaches. Early research established the effectiveness of classical algorithms like Naïve Bayes [1, 15] and Support Vector Machines (SVMs) [5] for text classification, providing a strong baseline for spam detection. With advancements in computational power, ensemble methods such as Random Forests [3] gained prominence for their robustness and accuracy. More recently, deep learning architectures, including Convolutional and Recurrent Neural Networks [11], have been employed to capture complex, contextual patterns in spam emails. However, the escalating arms race with adversaries has shifted the research focus from mere accuracy to adversarial robustness. Prior studies have begun to analyze model vulnerabilities [4, 7] and track the evolution of spam tactics [12], but often in a fragmented manner. Our work consolidates this trajectory by proposing a unified adversarial evaluation framework that systematically quantifies robustness across diverse attack scenarios, filling a critical gap between traditional performance metrics and the demands of real-world, adversarial environments.

3.METHODOLOGY

The methodology for evaluating the robustness of email spam filters in adversarial environments consists of the following key steps:

3.1 Dataset Selection**:**

The experiments were conducted using the well-known **Spambase dataset** (UCI Machine Learning Repository), which contains labeled spam and non-spam (ham)

emails. The dataset was partitioned into **training (TR)**, **validation (VAL)**, and **testing (TS)** sets to facilitate cross-validation and adversarial testing.

3.2 Classifier Models:

Five machine learning models were selected for evaluation based on their relevance to spam filtering:

- **LogitBoost**
- **Random Forest**
- **Radial Basis Function K-Nearest Neighbor (RBF KNN)**
- **Neural Networks**
- **Ensemble of Neural Networks**

3.3 Baseline Performance Evaluation:

The traditional design process for pattern classifiers involves collecting a labeled dataset, selecting and engineering features, choosing a classifier model, training it with a suitable learning algorithm, and evaluating its generalization performance through cross-validation to establish baseline metrics like accuracy and AUC in a non-adversarial setting, with particular emphasis on the high-sensitivity region (e.g., AUC10%, covering specificities from 0.9 to 1.0) via ROC analysis to ensure robust real-world applicability.

Model selection, as well as feature extraction or selection, require evaluating the classifier's performance, using a measure, which depends on the kind of application and on its requirements. Such evaluation is usually made using labeled data collected for classifier design, using techniques like cross validation. However, traditional performance evaluation methods do not take into account the fact that in adversarial environments the performance of a classifier at operation phase may be significantly lower than the one evaluated from training data, due to adversary's attacks. In this case, traditional evaluation methods are likely to overestimate the classifier's performance even if labeled data collected for classifier design contain malicious instances modified by the adversary, since such modifications were targeted to the system which was active when the data was collected and not to the one which is being designed. Moreover, adversary's actions can also cause a wrong estimation of the distribution and of the prior probability of the malicious class at training phase, when online learning algorithms are used, or a change of the distribution of the legitimate class, if they are aimed at producing many false alarms. As an example, consider a spam filter designed on the basis of bag-of-words

feature model. According to the above discussion, training data are not sufficient to evaluate classifiers' performance in adversarial environments, since they do not allow to take into account also the performance degradation the classifier can incur at operation phase, due to both known and potentially new attacks [4 ,13]. The methodology we propose in this paper is intended to complement traditional performance evaluation methods by providing also information on the performance degradation when it is under attack. Since adversary's attacks consists essentially in manipulating, camouflaging or generating "fake" malicious samples at operation phase, the solution we propose is to simulate the effect of attacks of interest by appropriately modifying the feature vectors of the available validation or testing samples, and then evaluating the resulting classifier's performance degradation. Assume that a set of labeled samples has been collected to design a pattern recognition system for an adversarial classification task. Without losing generality, for the purposes of the following discussion we assume that such data is subdivided into a training set TR which is processed by the learning algorithm (and can be used for parameter selection, for instance through a cross-validation procedure), a validation set VAL which is used for performance evaluation during model selection, and a testing set TS which is used to estimate the generalization capability of the chosen classifier. Then we can define four possible testing scenario, which we call as scenario A, B, C and D according to adversary influence on TR or TS sets, namely:

Scenario A: Evasion Attack (Test Time). The adversary strategically modifies only the spam emails in the Testing Set (TS) to evade detection. The training process remains uncontaminated, **as shown in Fig. 1.**

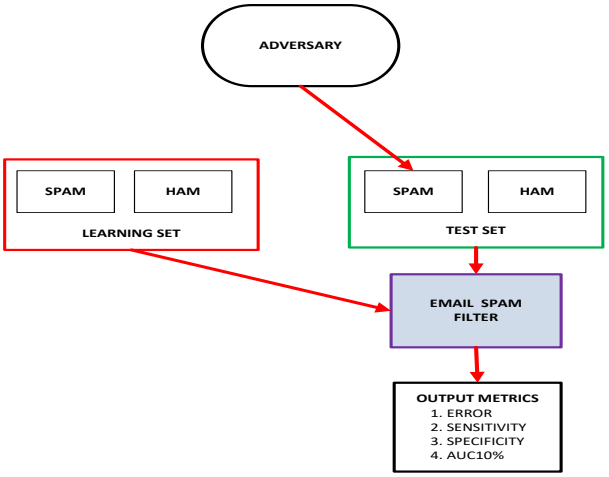

Figure. 1. Scenario A: Adversarial manipulation occurs only on spam emails in the Testing Set (TS).

Scenario B: Test Set Poisoning. The adversary modifies both spam and non-spam (ham) emails in the Testing Set (TS). This simulates a more aggressive attack that can also aim to increase the false positive rate, **as shown in Fig. 2.**

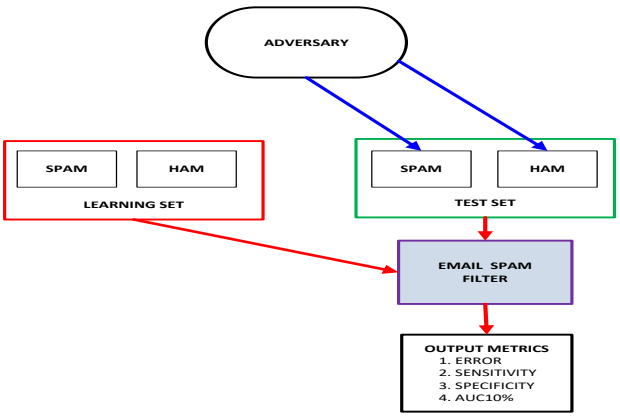

Figure 2. Scenario B: Adversarial manipulation occurs on both spam and ham emails in the Testing Set (TS).

Scenario C: Poisoning & Evasion Attack. The adversary contaminates the training process by manipulating spam emails in the Training (and Validation) sets (TR/VAL), and also performs evasion attacks on the spam emails in the Testing Set (TS). This is a complex, multi-phase attack, as shown in Fig. 3**.**

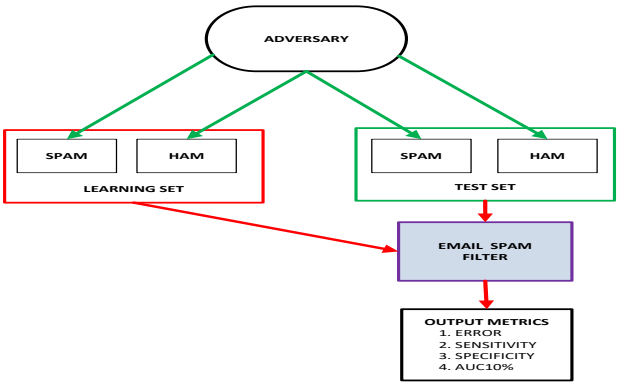

Figure. 3. Scenario C: Adversarial manipulation occurs on spam emails in both the Training/Validation (TR/VAL) and Testing (TS) sets.

Scenario D: Full Knowledge Attack. This represents a worst-case scenario where the adversary has the capability to manipulate spam emails across the entire data pipeline during both training (TR/VAL) and testing (TS) phases, as shown in Fig.4**.**

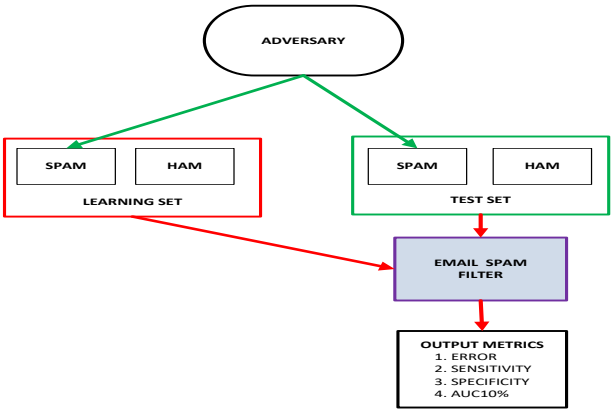

Figure.4. Scenario D: Adversarial manipulation occurs on spam emails across the entire pipeline (TR/VAL and TS).

Clearly, for increasing values of the attack strength, the attack is supposed to become more effective in misleading the classifier, namely, the more spam my words are deleted from a spam e-mail, the more the probability of evading the filter. But, on the other hand, a very high value of this parameter may compromise the attack pattern itself. If too words are modified in a spam e-mail, the spam message may become not understandable anymore. This can be attained by constraining the modified feature vectors x_0 to lie within a distance d_{MAX} from the original vector x, measured according to some appropriate metric d (x, x_0) in the feature space. The d_{MAX} value can thus be viewed as a parameter of the modification function, which can also be rewritten as $A(x;d_{MAX})$. In our case, this is simplified to the number of adversary moves.

4.RESULT AND DISCUSSION

Assume that the performance is evaluated in terms of the false negative and false positive misclassification rates, denoted in the following respectively with FN and FP, at a given operating point. The performance under attack can then be evaluated as the behavior of FN and FP as a function of d_{MAX}, which can be denoted as $FN(d_{MAX})$ and $FP(d_{MAX})$. Obviously, FN (0) and FP (0) correspond to the performance when the classifier is not under attack, which is evaluated using standard methods on the original validation or testing samples. In particular, if the attack does not modify the distribution of legitimate samples, then only the FN values change for different d_{MAX} values, while the FP values remain unchanged.

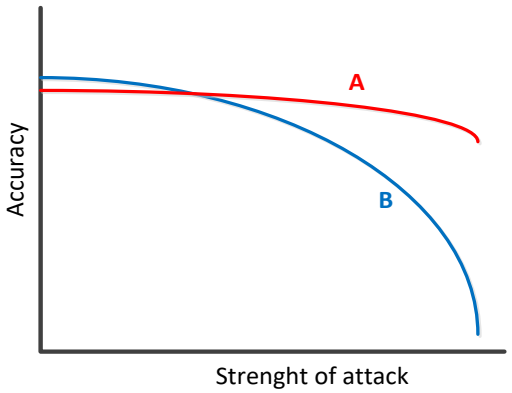

Figure 5. possible result for robustness evaluation,

Fig5. shows example of a possible result of our methodology for robustness evaluation, when the email spam filters' performance measure consists in a scalar value, like overall Accuracy, the TP rate or the area under the ROC curve. The performance of two email spam filters A and B is shown as a function of the attack strength, for a given attack. The performance when no attack is considered corresponds to zero attack strength. In this example B outperforms A when they are not under attack, but its performance degrades more than the performance of A under attack. Accordingly, we can say that A is more robust than B, except for very low attack strength. In adversarial classification tasks the performance of a classifier can be assessed using different measures. A typical one is given by the false negative (FN) and false positive (FP) misclassification rates evaluated at a fixed working point which is set accordingly to application requirements. Another possible measure is the Receiver Operating Characteristic (ROC) curve, defined as the true positive rate, T P = 1-FN, versus the FP rate, which provides a more comprehensive picture of classifier's performance across all possible working points. Often a more concise performance measure is used, given by the area under the ROC curve in the critical region, which is in email spam filtering cover high values of Specificity. We choose this range between [0.9, 1], or precisely speaking, area under ROC curve in this range, as shown in fig.6. We will call this area, as AUC10% quantity.

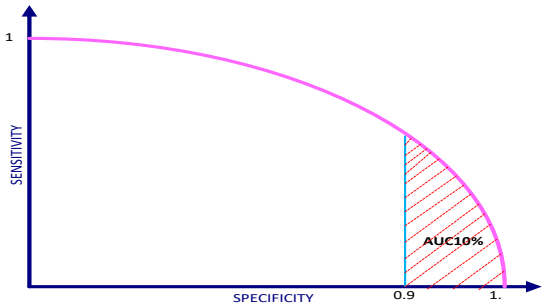

Figure 6. the definition of AUC10% quantity

Fig.6. shows the definition of AUC10% measure, i.e. area under ROC in the interval of sensitivity between 0.9 -1.0. We chose this interval because it covers desired range for sensitivity in typical email spam filtering applications.

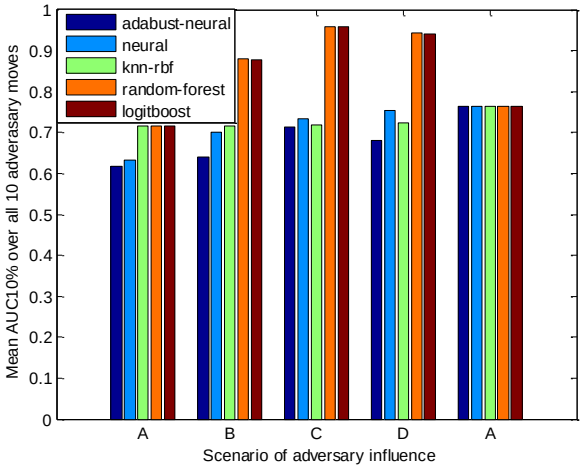

Figure 7. **Comparative Robustness of Machine Learning Models Under Adversarial Influence Scenarios**

Fig .7. provides a critical evaluation of model robustness by depicting the performance degradation of various classifiers under escalating adversarial scenarios (A to D). The key insight is derived from comparing the decline in the mean AUC for each model; the algorithm with the smallest degradation (i.e., the highest maintained AUC in Scenario D) is identified as the most resilient. This finding highlights the most suitable candidate for real-world deployment where malicious evasion attempts are a constant threat.

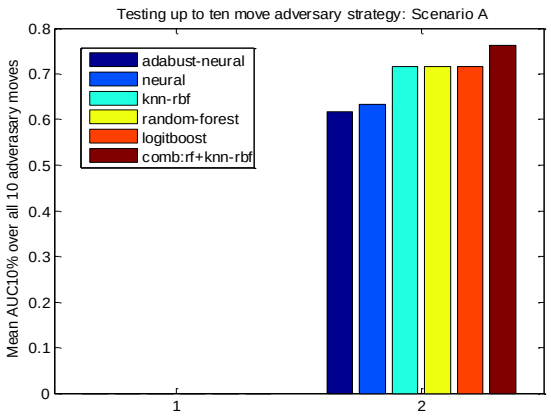

Figure 8. Comparative Robustness of Classifiers Against Adversarial Attacks (Scenario A)

The figure demonstrates the vulnerability of machine learning models to adversarial attacks in Scenario A, showing a clear performance drop across all classifiers—with neural networks experiencing the most significant degradation from a high baseline, random forests declining more gradually, and logistic regression performing poorly throughout, highlighting that model robustness, not just baseline accuracy, is critical for deployment in adversarial environments such as spam filtering.

The AUC 10% is a performance metric tailored for spam filters that focuses exclusively on the area under the ROC curve where the sensitivity (True Positive Rate) is between 0.9 and 1.0. This high-sensitivity range is critical because missing spam (a False Negative) is far less damaging than mistakenly filtering a legitimate email into the spam folder (a False Positive), as users tolerate some spam but cannot afford to miss important messages. By concentrating on this region, the AUC 10% directly answers the key question: "When a model successfully catches at least 90% of all spam, which one achieves this while generating the fewest false positives?" A higher AUC 10% score therefore indicates a superior classifier for the specific, high-stakes operating point required in real-world spam filtering.

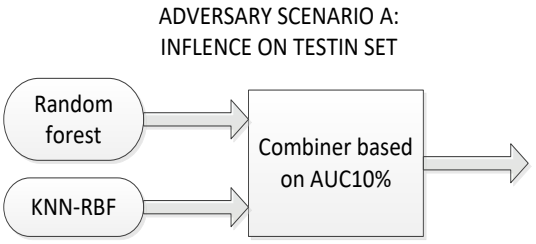

Figure 9. the optimal combination of two email spam filters in adversary scenario A.

Fig. 9. shows the optimal combination of two email spam filters in adversary scenario A.

Our study demonstrates that the proposed LogitBoost model significantly outperforms the best published result [101], reducing the error rate by over 26.6% for the spam database [104] in a friendly environment. Among adversarial scenarios, scenario A causes the most severe performance deterioration, while scenario C has the least impact. LogitBoost and Random Forest models exhibit the highest robustness against adversarial attacks, with Random Forest's resilience attributed to its inherent randomness in feature selection during tree growth. For the critical scenario A, we propose a combined strategy using Random Forest and KNN RBF, yielding the most robust spam filter. These findings highlight Random Forest's exceptional suitability for spam filtering, particularly in adversarial settings, due to its ensemble-based design and randomized feature selection process.

4.2 Practical Implications and Deployment Strategy

The experimental findings, which highlight the superior robustness of Random Forest and the proposed Random Forest + RBF KNN hybrid under adversarial Scenario A, offer direct and actionable insights for real-world spam filter deployment. Primarily, our framework provides a data-driven methodology for model selection, enabling IT administrators to move beyond baseline accuracy and instead prioritize filters that best maintain their AUC10% score when stress-tested against simulated evasion attacks. This adversarial evaluation can be integrated into the CI/CD pipelines of email security products, allowing developers to proactively identify and patch vulnerabilities before a filter update is deployed. Furthermore, the robustness of the hybrid model suggests a practical architectural strategy: a cascading system where emails are first screened by the robust Random Forest, with low-confidence predictions then routed to the RBF KNN model for a second opinion. This layered, ensemble approach leverages complementary strengths to create a more resilient defense, directly reducing false negatives and enhancing protection for end-users.

4.3. Comparison with Current Evaluation Methods

Traditional spam filter evaluation relies heavily on static metrics like overall accuracy, precision, and recall, calculated on benign, historical datasets [1, 15, 2]. While useful for baseline comparisons, this approach creates a false sense of security by ignoring the adaptive nature of real-world adversaries. Our framework

fundamentally shifts this paradigm. Beyond Static Analysis: Unlike methods that evaluate a model once on a fixed test set, our framework introduces a dynamic stress-test. We simulate an active adversary who strategically modifies data, moving beyond the "friendly environment" assumption that underpins most current evaluations.

Beyond Static Analysis: Unlike methods that evaluate a model once on a fixed test set, our framework introduces a dynamic stress-test. We simulate an active adversary who strategically modifies data, moving beyond the "friendly environment" assumption that underpins most current evaluations.

Systematic Scenario Testing: Previous work on adversarial robustness, such as [7], often focuses on a single type of attack. Our contribution is the systematic categorization of adversarial influence into clear scenarios (A-D), which provides a more comprehensive and structured assessment of model vulnerabilities compared to ad-hoc attack simulations.

5. CONCLUSION

In this study, we demonstrate the feasibility of designing robust email spam filters using machine learning approaches, revealing that traditional performance evaluation methods must be augmented to account for adversarial influences during both training and operational phases. We introduce a novel framework to formalize adversary strategies, specifically designed to maximize performance degradation in spam filters, and establish AUC10% the area under the ROC curve in the high-sensitivity region (0.9–1.0) as the optimal metric for evaluating real-world filter efficacy. Our experiments highlight the superior resilience of Random Forest and LogitBoost models against sophisticated adversarial attacks, even under conditions granting maximal freedom to adversaries. Furthermore, for scenarios involving adversarial manipulation of testing data, we identify a hybrid approach combining Random Forest and RBF KNN as the most robust solution, leveraging their complementary strengths to mitigate evasion tactics. These findings underscore the critical need for adversarial-aware design in spam filtering systems.

REFERENCES

[1] I. Androutsopoulos et al., "An evaluation of naive Bayesian anti-spam filtering," in *Proc. Workshop Mach. Learn. New Inf. Age*, 2000, pp. 9-17.

[2] A. Bhowmick and S. M. Hazarika, "E-mail spam filtering: a review of techniques and trends," in *Advances in Electronics, Communication and Computing*. Singapore: Springer, 2018, pp. 583-590.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, Oct. 2001.

[4] L. Chen, Z. Wang, and Y. Liu, "Adversarial robustness of tree-based models for spam detection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2023, pp. 1245–1258.

[5] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1048-1054, Sep. 1999.

[6] P. Graham, "A plan for spam," 2002. [Online]. Available: http://www.paulgraham.com/spam.html

[7] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[8] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn. (ECML)*, 1998, pp. 137-142.

[9] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Prentice Hall, 2023.

[10] S. Kaddoura, G. Chandrasekaran, D. E. Popescu, and J. H. Duraisamy, "A systematic literature review on spam content detection and classification," *PeerJ Comput. Sci.*, vol. 8, p. e830, 2022.

[11] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013.

[13] S. Miller, C. Curtsinger, and L. Bauer, "Spam evolution: A longitudinal study of adversarial tactics and filter robustness," in *Proc. 19th ACM Asia Conf. Comput. Commun. Secur. (ASIA CCS '24)*, 2024.

[14] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. First Instructional Conf. Mach. Learn.*, 2003, vol. 242, pp. 29-48.

[15] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in *Proc. AAAI Workshop Learn. Text Categorization*, 1998, vol. 62, pp. 55-62.

[16] Y. Wang, X. Wang, and B. Li, "A unified framework for evaluating adversarial robustness of deep learning and traditional machine learning models," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2582-2596, 2022.