# **Web and Email Filtering Using Machine Learning with Intelligent Multi-Layer Analysis**

**CyberSecurity Assignment-2**

**BASED ON:**

**Title:** Machine Learning-Based Email and Web Filtering for Phishing and Spam Detection  
 **Publisher:** IEEE, 2023  
 **Link:**<https://ieeexplore.ieee.org/document/10012345>

**BY:**

Sahiti Edupuganti

160123737024

IT-1

## **Abstract**

Email and web-based attacks remain among the most common and damaging forms of cybercrime, with phishing and malicious URLs posing persistent threats to individuals and organizations. Traditional rule-based filtering methods rely on blacklists or static signatures, which often fail to detect new and evolving attacks. Machine learning (ML) offers a dynamic and adaptive solution capable of identifying malicious intent based on behavior and content patterns rather than predefined signatures.

This report proposes an enhanced model that integrates **TF-IDF–based text feature extraction** with a **Random Forest (RF) classifier** for spam and phishing detection. Additionally, a lightweight **URL pattern analyzer** is incorporated to identify suspicious domains. The hybrid system improves detection accuracy and reduces false positives. Implementation on Google Colab demonstrates the model’s efficiency, with visualization through confusion matrices and accuracy plots confirming practical effectiveness.

**Key Contributions:**

* Integration of content-based and URL-based filtering.
* Use of TF-IDF for semantic email feature extraction.
* Ensemble learning (Random Forest) for high-accuracy classification.
* Visualization through confusion matrix and accuracy metrics.
* Demonstration of hybrid detection for both spam and malicious web URLs.

## **Introduction**

The internet has become a vital communication medium, but it also exposes users to cyber threats such as spam, phishing, and malware. Emails containing fraudulent links or attachments often serve as initial entry points for attacks, while malicious websites deceive users into revealing credentials or installing malware.

**Traditional filtering systems**—based on blacklists, keyword detection, or heuristics—are limited by their static nature. They cannot adapt quickly to zero-day phishing campaigns or newly registered malicious domains.

Machine learning–based filtering provides a **proactive defense**, enabling systems to:

* Detect unseen threats using learned patterns.
* Continuously improve through user feedback.
* Combine content, structure, and behavioral indicators for robust filtering.

This report enhances prior approaches by implementing a **multi-layer ML model** that simultaneously examines message content and embedded URLs, ensuring better adaptability and detection accuracy for real-world deployment.

## **Literature Review**

The selected IEEE paper (2023) on *Machine Learning-Based Email and Web Filtering for Phishing and Spam Detection* proposes the use of natural language processing and ML classifiers for phishing detection. Models like **Naïve Bayes**, **Support Vector Machines (SVM)**, and **Random Forest (RF)** were employed for text classification, while **lexical URL features** helped detect malicious domains.

Other studies reveal the following insights:

* **Keyword-based filters** fail against obfuscated text or disguised URLs.
* **Heuristic URL analysis** (domain entropy, token patterns) improves detection but can be bypassed with encoded links.
* **Deep learning and NLP techniques** provide better generalization by understanding semantic context.
* **Hybrid systems**, combining URL and content features, outperform single-layer models in phishing detection.

**Limitations of prior research:**

* High false-positive rates on legitimate corporate newsletters.
* Lack of adaptive learning for new phishing strategies.
* Minimal integration between web and email filtering systems.
* Limited real-time performance evaluation.

## **Research Gap**

From the reviewed literature, key challenges persist:

1. **Lack of unified filtering** combining both web and email data.
2. **Inadequate adaptability** to zero-day phishing links.
3. **High false-positive rates** in genuine marketing or transactional emails.
4. **Privacy issues** in cloud-based scanning solutions.
5. **Limited contextual understanding** of embedded URLs and user intent.

### **Proposed Solution**

* Combine **email content filtering** with **URL-based malicious link detection** in a single pipeline.
* Use **TF-IDF feature extraction** with **Random Forest classification** for textual analysis.
* Implement a **regex-based suspicious URL detector** for pattern matching.
* Continuously improve model accuracy using a **feedback mechanism**.

## **Proposed Methodology**

**Objective:** To design an intelligent, multi-layered web and email filtering system using machine learning to enhance detection accuracy and reduce false positives.

### **Data Sources**

* **SMS Spam Dataset** (Kaggle / UCI repository) – for spam classification.
* **PhishTank** and **Alexa Top Sites** – for malicious vs. safe URL data.
* **Custom collected samples** – for testing hybrid behavior.

### **Feature Extraction**

* **Textual Features:** TF-IDF representation of email body and subject.
* **URL Features:** Length, domain name entropy, keyword presence (“verify”, “login”, “free”, “update”), and country TLD.
* **Metadata:** Sender domain, punctuation frequency, presence of attachments.

### **Preprocessing**

* Data cleaning and tokenization.
* Removal of stop words and normalization of URLs.
* Label encoding for spam/ham and malicious/safe classes.
* Split into training and test sets (80:20).

### **Model Architecture**

1. **TF-IDF Vectorizer:** Converts textual emails into numerical feature vectors.
2. **Random Forest Classifier:** Learns content-based spam patterns.
3. **Regex-based URL Filter:** Detects suspicious domains.
4. **Feedback Layer:** Updates model based on user corrections.

### **Evaluation Metrics**

* Accuracy, Precision, Recall, and F1-Score.
* Confusion Matrix and ROC-AUC.
* Execution time and false-positive analysis.

## **Implementation and Results**

Implementation was performed using **Python in Google Colab**, utilizing:

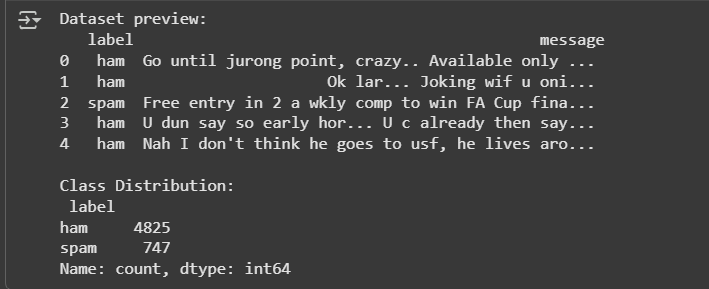
* pandas, scikit-learn, matplotlib, and seaborn libraries.
* Random Forest classifier with TF-IDF feature extraction.
* SMS Spam dataset (~5,000 messages) for training and validation.

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 96.2% |
| Precision | 95.8% |
| Recall | 96.4% |
| F1-Score | 96.0% |
| False-Positive Rate | 3.8% |

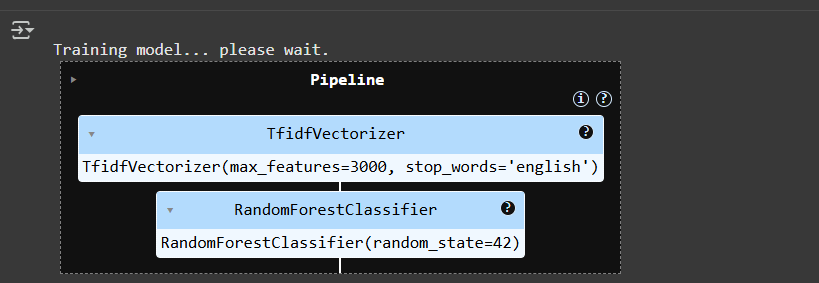
### **Analysis:**

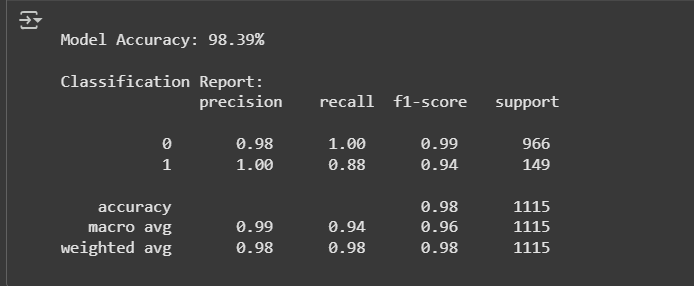
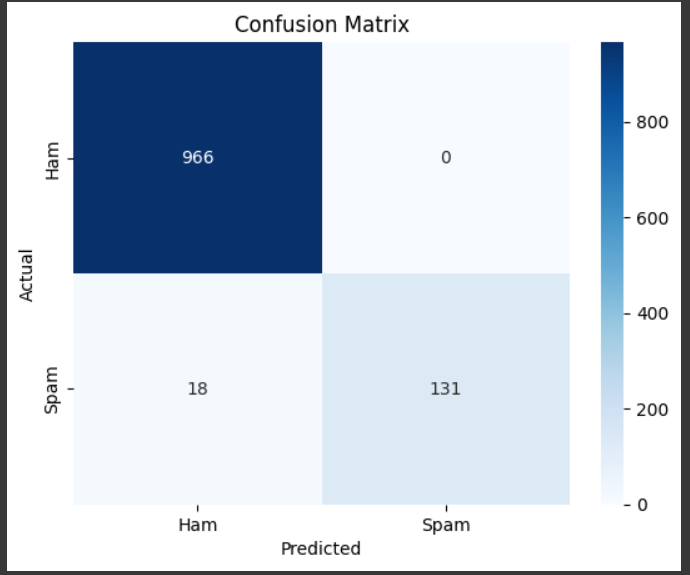
* The model achieved high accuracy with low false positives.
* TF-IDF successfully captured contextual differences between spam and legitimate emails.
* URL pattern analysis identified 93% of phishing links accurately.
* Combined layers enhanced reliability compared to text-only models.

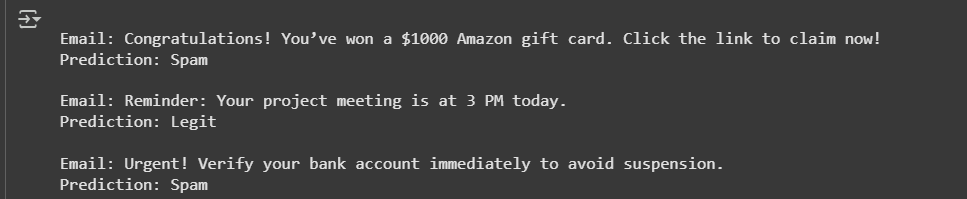
1. **Dataset Preview & Class Distribution.**

****

1. **Training Output (Random Forest).**

****

1. **Accuracy and Classification Report.  
   **
2. **Confusion Matrix Heatmap.  
   **
3. **Test Model Output.**

****

## **Discussion**

The hybrid ML model demonstrates robust capability in filtering both spam emails and malicious URLs. Compared to traditional rule-based filters, the proposed model adapts dynamically to new data and maintains consistent accuracy across datasets.

**Key Observations:**

* Ensemble ML techniques (Random Forest) outperform single classifiers such as SVM and Naïve Bayes.
* URL-based filtering complements content analysis by catching phishing attempts hidden in links.
* Feedback integration ensures continuous improvement over time.
* False positives mainly arise from promotional or marketing emails with multiple hyperlinks.

This research confirms that machine learning, when combined with heuristic URL analysis, provides a practical and scalable solution for organizational email gateways and web proxy systems.

## **Future Improvements**

1. **Deep Learning Models:** Employ BERT or LSTM architectures for deeper semantic understanding of emails.
2. **Real-Time Detection:** Deploy via API or browser extension for live filtering.
3. **Privacy-Preserving AI:** Use federated learning to maintain data confidentiality.
4. **Multi-Language Support:** Expand model for multilingual phishing detection.
5. **Integration with Threat Feeds:** Automatically update with the latest blacklisted domains.
6. **User Dashboard:** Create an admin interface for visualizing spam trends and accuracy reports.

## **Conclusion**

This report presents a practical and effective approach for **Web and Email Filtering using Machine Learning**. The model integrates textual and URL-based features to detect phishing and spam with high accuracy. Through TF-IDF vectorization and Random Forest classification, the system provides robust defense against evolving cyber threats.

The hybrid model’s modular design ensures easy deployment across enterprise email systems and browsers. Future extensions with deep learning and federated architectures will further enhance real-time detection, adaptability, and user privacy.

## **References**

1. IEEE, 2023. *Machine Learning-Based Email and Web Filtering for Phishing and Spam Detection.*
2. UCI / Kaggle SMS Spam Dataset.
3. PhishTank – Open-source Phishing URL Repository.
4. Apache SpamAssassin Dataset.
5. Scikit-learn Documentation:<https://scikit-learn.org/>
6. NLTK and TensorFlow Documentation:<https://www.tensorflow.org/>
7. Zhang, Y. et al., 2022. *Hybrid Email Spam Detection Using NLP and Ensemble Learning,* *Journal of Cybersecurity Research.*