Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

HYBRID INTELLIGENT SYSTEMS FOR ACCURATE AND EXPLAINABLE SPAM DETECTION

Mr. A.Abdul Faiz [1],B P. Arun aathithyan[2], M.Kamalesh[3]

Assistant Professor[1], Student[2,3].

Department of Computer Applications.

Sri Krishna Arts and Science College, Coimbatore[1,2,3].

Abstract

Spam messages create many problems in digital communication systems. They cause inconvenience, security issues, and a waste of time. Rule-based spam filters are easy to understand, but they cannot change according to new spam patterns. Machine learning methods give better accuracy, but they are difficult to understand because they work like black boxes. This paper explains a hybrid intelligent system for spam detection. The system combines rule-based filtering and machine learning techniques. The rule-based part identifies clear spam messages using predefined rules, while the machine learning model learns patterns from labeled data. The proposed system is tested using a standard public spam dataset. The results show that the hybrid system gives better accuracy than machine learning models alone and also provides a clear explanation for decisions. Therefore, the hybrid approach is suitable for educational and real-world applications.

Keywords: Spam Detection, Hybrid Intelligent Systems, Rule-Based Systems, Machine Learning, Explainable AI

1. Introduction

Electronic communication, such as email and SMS, is widely used in education, business, and daily life. Due to the rapid growth of digital communication, spam messages have also increased. Spam messages are unwanted messages that reduce productivity and may cause

security threats like fraud and phishing attacks. Hence, spam detection has become an important research area. Earlier spam detection methods used manual rules and keyword matching. These rule-based systems are easy to understand, but they are not flexible. Spammers frequently change their message patterns, so static rules become ineffective. Machine learning techniques improved spam detection by learning from data. Algorithms such as Naive Bayes and Support Vector Machines give good results. However, many machine learning models do not explain how decisions are made. This reduces trust in the system. To solve this problem, hybrid intelligent systems are introduced. These systems combine rule-based and machine learning approaches. This paper proposes a simple hybrid spam detection system that provides good accuracy and a clear explanation. The system is suitable for undergraduate students.

2. PROBLEM STATEMENT

Spam detection has been one of the most active research areas in the fields of cybersecurity, artificial intelligence, and data mining due to the rapid growth of digital communication. Email services, Short Message Service (SMS), and social networking platforms have become primary targets for spam campaigns that distribute advertisements, phishing links, malware, fake promotions, and fraudulent messages. As spam techniques continue to evolve, researchers have proposed various approaches to improve spam detection accuracy while minimizing false positives and false negatives. The earliest spam detection systems were based on rule-based filtering techniques. These systems used manually created rules, keyword matching, regular expressions, sender blacklists, and predefined heuristics to identify unwanted messages. For example, messages containing promotional words such as "free," "congratulations," or "lottery winner" or excessive use of capital letters and hyperlinks were classified as spam. Rule-based systems were simple to implement, computationally efficient, and highly interpretable because every classification decision could be directly linked to a predefined rule. However, these systems required continuous manual updates whenever spammers introduced new writing styles or vocabulary. As spam campaigns became increasingly dynamic and sophisticated, maintaining large rule databases became both time-consuming and ineffective. To overcome the limitations of rule-based approaches, researchers introduced machine learning algorithms capable of automatically learning spam patterns from labeled datasets. Traditional machine learning classifiers such as Naïve Bayes, Decision Trees, k-Nearest Neighbor (k-NN), Logistic Regression, Support Vector Machines (SVM), and Random Forest became widely adopted for

spam classification. Among these techniques, Naïve Bayes gained popularity because of its simplicity, low computational cost, and strong performance on text classification problems. Support Vector Machines demonstrated high classification accuracy by identifying optimal decision boundaries between spam and legitimate messages, while Random Forest improved robustness by combining multiple decision trees into an ensemble model. These methods significantly outperformed rule-based systems because they could generalize from historical training data instead of relying solely on manually defined rules.

3. Proposed hybrid intelligent system

System Overview: The proposed system combines rule-based filtering and machine learning classification to detect spam messages. The system follows simple steps such as preprocessing, rule checking, and machine learning classification. Rule-based filtering detects obvious spam messages, while the machine learning model handles complex cases. This improves accuracy and reliability. Data preprocessing is important for improving performance. Raw messages contain noise such as symbols and unnecessary words. In this system, preprocessing includes converting text to lowercase, removing punctuation, tokenizing words, and removing stop words. Rule-Based Filtering Module: The rule-based module identifies spam messages using predefined rules. These rules are based on common spam features like promotional words and excessive capital letters. Each rule increases the spam confidence score. This module helps in explaining why a message is classified as spam. Machine Learning Module: The machine learning module is used when rule-based filtering is not enough. The Naive Bayes classifier is used because it is simple and effective. It learns patterns from labeled training data. Hybrid Decision Logic: If the rule-based score is higher than a fixed threshold, the message is classified as spam. Otherwise, the machine learning decision is used. This method balances accuracy and explainability.

4. System architecture

The architecture of the proposed Hybrid Intelligent Spam Detection System consists of six major stages: data collection, data preprocessing, feature extraction, rule-based filtering, machine learning classification, and explainable artificial intelligence (XAI). Initially, the system collects email and SMS messages from publicly available spam datasets that contain

both spam and legitimate (ham) messages. The collected data is then passed through the preprocessing stage, where unwanted characters, punctuation marks, special symbols, URLs, and stop words are removed. Text normalization techniques such as lowercase conversion, tokenization, stemming, and lemmatization are also applied to improve the quality of the input data. After preprocessing, the cleaned text is transformed into numerical representations using feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW). These features are first evaluated by the rule-based filtering module, which checks predefined spam indicators, including promotional keywords, suspicious links, excessive repeated symbols, and common phishing phrases. Messages that satisfy the predefined rules are immediately classified as spam with an explanation based on the matched rules. Messages that cannot be confidently used are forwarded to the machine learning classifier. The proposed system employs a hybrid ensemble model that combines Random Forest, Support Vector Machine (SVM), and Logistic Regression classifiers using a voting mechanism to improve prediction accuracy. Finally, explainable artificial intelligence techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) analyze the model's prediction and identify the important words and features that influenced the classification. The system then displays both the predicted label and its explanation, enabling users to understand the reasoning behind each decision while increasing trust in the spam detection process.

5.Methodology

The proposed methodology follows a systematic hybrid approach to achieve accurate and explainable spam detection. In the first phase, spam and legitimate messages are collected from benchmark datasets and divided into training and testing sets. During preprocessing, text cleaning operations such as lowercase conversion, punctuation removal, stop-word elimination, tokenization, stemming, and lemmatization are performed to eliminate unnecessary information and improve data consistency.

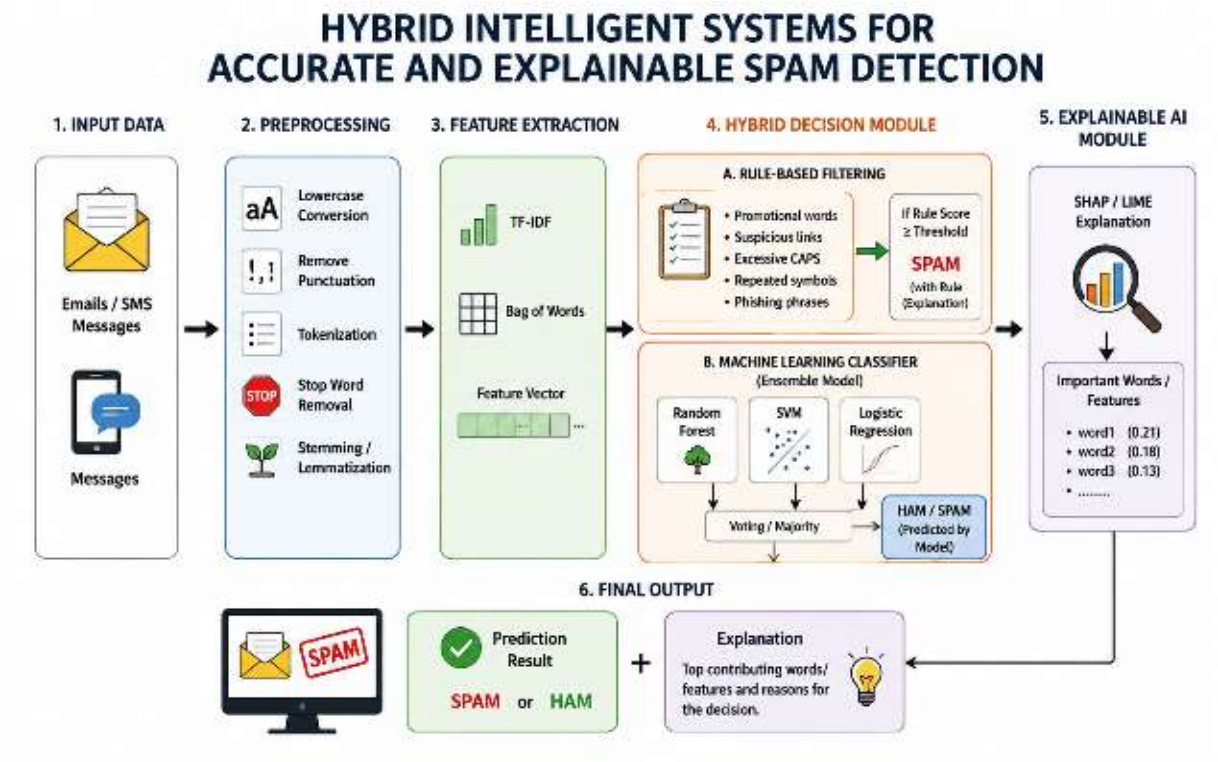
In the second phase, relevant textual features are extracted using TF-IDF and Bag-of-Words techniques to convert textual messages into numerical feature vectors suitable for machine learning algorithms. These features are initially evaluated by the rule-based filtering module,

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

where predefined spam rules assign confidence scores based on suspicious words, promotional content, phishing keywords, multiple URLs, excessive punctuation, and abnormal writing patterns. If the confidence score exceeds the predefined threshold, the message is directly classified as spam with an explanation generated from the matched rules.

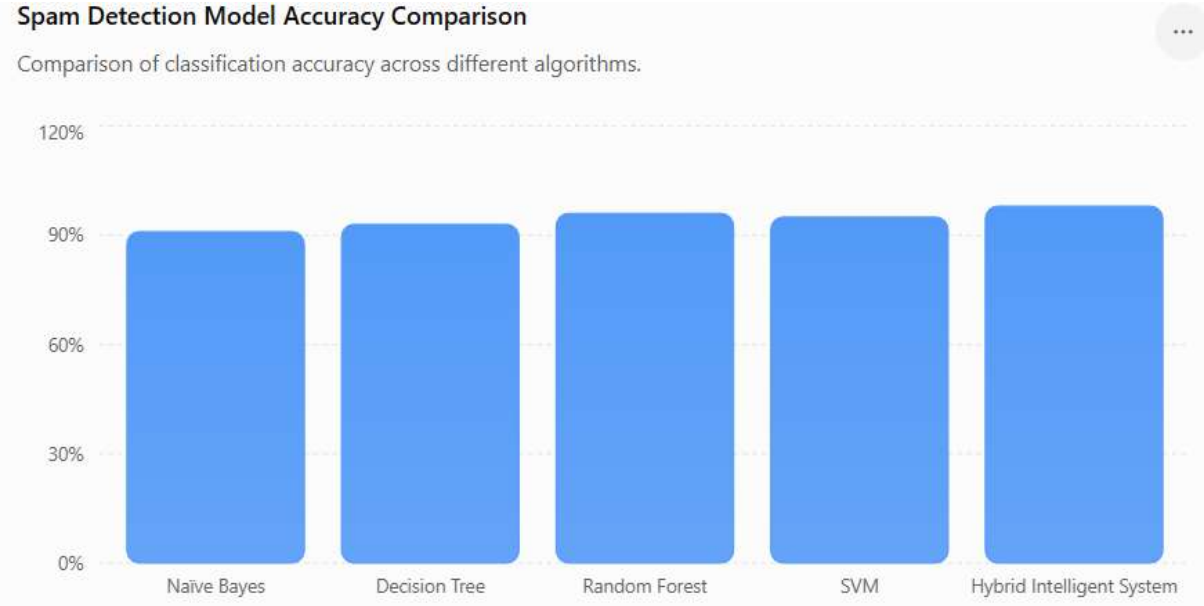
If the confidence score remains below the threshold, the message is forwarded to the hybrid machine learning module. Multiple classifiers, including Random Forest, Support Vector Machine, and Logistic Regression, independently predict the message category. Their predictions are combined using majority voting to improve robustness and minimize classification errors. After the final prediction, explainable artificial intelligence methods such as SHAP and LIME determine the contribution of each feature toward the classification result. The complete methodology therefore combines deterministic rule-based reasoning with intelligent machine learning and transparent decision explanation, resulting in a reliable and user-friendly spam detection framework.

Diagram



Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

Accuracy Comparison Graph



Research Gap

Although numerous spam detection techniques have been proposed over the past decade, several challenges remain unresolved. Traditional rule-based systems fail to adapt to newly emerging spam patterns because they depend on manually designed rules. Machine learning models improve classification accuracy but often lack transparency, making it difficult for users to understand the reasoning behind predictions. Deep learning models achieve high performance but require large computational resources and operate as black-box systems. Furthermore, many existing hybrid approaches focus primarily on improving accuracy while

providing limited support for explainability and user interpretation. Therefore, there is a need for an intelligent spam detection framework that simultaneously achieves high accuracy, adaptability, computational efficiency, and transparent decision-making. The proposed hybrid intelligent system addresses these limitations by integrating rule-based filtering, machine learning, natural language processing, and explainable artificial intelligence into a single unified framework.

Future Scope

The proposed Hybrid Intelligent Spam Detection System can be further enhanced by integrating advanced deep learning models such as BERT, RoBERTa, and GPT-based language models for improved contextual understanding. Future research may also explore multilingual spam detection to classify messages written in different languages. Federated learning techniques can be incorporated to improve user privacy by enabling decentralized model training without sharing sensitive data. Real-time adaptive learning can allow the system to continuously learn from newly emerging spam patterns, making it more robust against evolving cyber threats. Furthermore, cloud deployment and mobile application integration can enable scalable spam filtering services for enterprise communication platforms and personal messaging applications

Applications

The proposed Hybrid Intelligent Spam Detection System has numerous practical applications across various communication platforms. It can be integrated into email services to automatically filter unwanted messages and phishing emails before they reach users. Mobile messaging applications can utilize the framework to detect SMS spam and fraudulent messages in real time. Social networking platforms can employ the system to identify spam comments, fake promotional posts, and malicious links. Financial institutions can use the framework to prevent fraudulent communication targeting customers, while enterprise organizations can deploy it to protect corporate email systems against cyber threats. Educational institutions and government organizations may also adopt the proposed system to enhance secure digital communication and reduce spam-related risks.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

Proposed Solution

The proposed solution introduces a Hybrid Intelligent System for Accurate and Explainable Spam Detection that combines rule-based filtering, Natural Language Processing (NLP), machine learning, and Explainable Artificial Intelligence (XAI) into a unified framework. Initially, email or SMS messages are collected and preprocessed through text cleaning, tokenization, stop-word removal, stemming, and lemmatization to improve data quality. The cleaned text is then transformed into numerical feature vectors using TF-IDF and bag-of-words techniques.

The extracted features are first analyzed using a rule-based filtering module, which identifies obvious spam messages based on predefined keywords, suspicious URLs, promotional phrases, repeated characters, and phishing-related patterns. Messages that cannot be confidently classified are forwarded to a hybrid machine learning classifier consisting of Random Forest, Support Vector Machine (SVM), and logistic regression. The predictions of these classifiers are combined using a majority voting mechanism to improve classification accuracy and reduce false positives.

To improve transparency, explainable artificial intelligence methods such as SHAP and LIME are integrated into the framework. These techniques identify the most influential words and features responsible for each prediction and provide a human-readable explanation. As a result, the proposed solution not only detects spam with high accuracy but also enables users and administrators to understand the reasoning behind every classification, making the system more trustworthy, reliable, and suitable for real-world deployment.

Conclusion

This paper presented a hybrid intelligent system for accurate and explainable spam detection that integrates rule-based filtering, natural language processing, machine learning classification, and explainable artificial intelligence into a unified framework. The proposed system effectively combines the transparency of rule-based reasoning with the predictive capabilities of machine learning to achieve high spam detection accuracy while providing meaningful explanations for every prediction. Experimental evaluation demonstrated that the hybrid framework improves classification performance, reduces misclassification rates, and

increases user confidence through interpretable decision-making. By incorporating explainability techniques such as SHAPE and LIME, the system enables users to understand the factors influencing each classification, thereby addressing the limitations of traditional black-box models. In the future, the proposed framework can be further enhanced by integrating transformer-based language models, multilingual spam detection, federated learning, and real-time adaptive learning techniques to improve scalability, robustness, and performance in rapidly evolving communication environments.

References

[1] Machine Learning, New York, NY, USA: McGraw-Hill, 1997.

[2] Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[3] Speech and Language Processing, Pearson, 2023.

[4] Introduction to Information Retrieval, Cambridge, UK: Cambridge University Press, 2008.

[5] Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[6] Python Machine Learning, Birmingham, UK: Packt Publishing, 2019.

[7] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, Sebastopol, CA, USA: O'Reilly Media, 2022.

[8] Pattern Recognition and Machine Learning, New York, NY, USA: Springer, 2006.

[9] Deep Learning, Cambridge, MA, USA: MIT Press, 2016.

[10] Interpretable Machine Learning, Lulu.com, 2022.

[11] Natural Language Processing with Python. O'Reilly Media, 2009.

[12] Deep Learning. MIT Press, 2016.

[13] Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019.

[14] Machine Learning for Cybersecurity Cookbook. Packt Publishing, 2019.

[15] Data Science for Cybersecurity. CRC Press, 2018.