

Enhancing Explainability in Machine Learning Based Spam Email Detection Using LIME and SHAP

Krishna Devanand Maurya¹, Aryan Rajesh Pal², Brijesh Pandey^{3,*}

Abstract

Spam emails represent a continual and irritating cybersecurity issue due to the fast expansion of cyberspace communication. These messages are unwanted and, in some cases, dangerous, such as phishing schemes, identity theft, and the propagation of malware. They cause severe financial and data security threats to individuals and organizations. Conventional spam detection techniques, such as rule-based spamming, blacklisting, and keyword classification, are techniques that find it difficult to match adaptive spamming schemes that include manipulation of text, dynamic generation of content, and concealment of information. Machine learning methods like support vector machines (SVM), random forest, and Naive Bayes have significantly enhanced detection accuracy. This paper introduces a machine learning model for a spam email detector that is non-technical. It pays equal attention to the interpretability of the model as well as the prediction power of the model. The system uses explainable artificial intelligence (XAI) methods, namely, Local Interpretable Model-Agnostic Explanations (LIME) and SHAP, to complement the conventional machine learning classifiers. With this combination, one can get insight into the decision-making of the model in a manner that can be understood by humans. The model relies on text representation in the form of TF-IDF, which is useful in extracting features, as well as an optimized SVM that is used to classify spam. The experimental findings show that this model can be used to achieve 96.4% accuracy that is better than the standard baseline systems yet can be understood using visual and quantitative explanation modules. Along with the accuracy of classification, the use of XAI enhances the credibility of the model significantly. This enables end-users and cybersecurity analysts to know and verify the reason as to why specific emails are This interpretability stimulates more appropriate decision-making, compliance with data transparency rules, and makes the system constantly improve due to explainability feedback. The research ultimately. This fills the disconnect that exists between high-performance machine learning models and user trust, showing that augmenting XAI within spam detection systems can make them not opaque black-box applications, but transparent, responsible, and trustworthy cybersecurity solutions that can be deployed in practice. Bridges the gap between high-performance machine learning (ML) models and user-centered trust, demonstrating that integrating XAI into spam detection frameworks can transform them from opaque “black-box” systems into transparent, accountable, and reliable cybersecurity solutions suitable for real-world deployment.

*Author for Correspondence

Brijesh Pandey

E-mail: brijesh.pandey@thakureducation.org

¹⁻²Research Scholar, MCA Thakur Institute of Management Studies, Career Development & Research (TIMSCDR) Mumbai, Maharashtra, India

³Professor, MCA Thakur Institute of Management Studies, Career Development & Research (TIMSCDR) Mumbai, Maharashtra, India

Received Date: March 10, 2026

Accepted Date: June 28, 2026

Published Date: July 25, 2026

Citation: Krishna Devanand Maurya, Aryan Rajesh Pal, Brijesh Pandey. Enhancing Explainability in Machine Learning-Based Spam Email Detection Using LIME and SHAP. Journal of Network Security. 2026; 14(2): 35–48p.

Keywords: Spam email detection, machine learning, explainable artificial intelligence, LIME, SHAP, support vector machine, TF-IDF, model interpretability, cybersecurity

INTRODUCTION

Email is still one of the most significant means through which people communicate over the internet. Its convenience, affordability, and expandability make it necessary to use it in personal and professional life despite the emergence of secure messaging and business apps [1]. Nonetheless, owing to their high prevalence, emails have become prime targets for malware,

phishing, and spam [2]. Spam emails are undesirable and highly deceptive, posing grave dangers such as financial fraud, identity theft, and breaches of big data [3].

In global email traffic, there is more than 45 percent in this category [4]. Traditional spam detection systems include keyword matching, rule-based spam filtering, and blacklists. Evasion methods such as URL masquerade and text obfuscation have not been addressed correctly [5]. Machine learning (ML) has greatly enhanced the capability of detection systems to counter changing spam techniques.

To address this limitation, explainable artificial intelligence (XAI) methods like SHAP (Shapley Additive Explanations) [6] and Local Interpretable Model-Agnostic Explanations (LIME) [5] have been developed to offer interpretability at the feature level. In contrast, SHAP measures use contributions that are based on game theory [7]. These methods have enhanced model robustness and understanding of users in tasks related to cybersecurity and text classification [8]. However, several challenges remain. The interpretability of deep learning models, applicability of ensemble techniques to large scales, and extrapolation of features to different datasets have yet to be studied in detail [3, 5, 7].

The application of XAI techniques in spam detection systems has been the subject of recent studies [1, 2, 4].

Recent research has focused on the use of XAI methods to detect spam in spam detection systems [1, 2, 4]. The present study is a continuation of this study, as it introduces a support vector machines (SVM) framework that uses SHAP and LIME. Through these two explainability methods, an accuracy of 96.4 was attained by this system.

BACKGROUND AND MOTIVATION

Evolving Landscape of Spam Email Threats

Spam email attacks have changed from basic promotional messages to complex campaigns that use text tricks, harmful URLs, and social engineering techniques. Spam emails are no longer simple promotional messages but rather elaborate campaigns and tricks with text, dangerous URLs, and social engineering. Spam emails are created in a manner that makes them similar to real communications. They resort to domain spoofing and similar character replacements [1–3]. Spammers have high financial incentives, and mass emailing is cheap, thus increasing the volume and sophistication of attacks in seconds.

Reports show that many current spam campaigns rely on cloud infrastructures. They use legitimate platforms to avoid blacklisting and employ methods such as content randomization, image-based text, and polymorphic message generation [1, 2, 4]. These trends highlight the weaknesses of static and rule-based filtering methods. Consequently, there is an increasing need for smart systems that can adapt to changing spam patterns and offer clear decision-making processes.

Limitations of Traditional Detection Methods

Spam email detection has always relied on blacklists, keyword filtering, and scoring rules. Blacklists are effective at blocking known spam sources in a short period of time; however, they are not particularly efficient at dealing with rapid domain changes and disposable email accounts. They tend to be ineffective at content hiding and generate high false positives, which interfere with actual communications [9–11].

Classical ML models, like Naïve Bayes and support vector machines, improved detection by using statistical and lexical features from email content systems. They lack transparency in decision-making and often face issues with user trust and regulatory compliance [5–7].

Emergence of ML and Explainable AI Approaches

The weaknesses of conventional approaches have led to a shift in favor of automated data-driven approaches. ML models can learn complex text patterns in the case of large email data [12]. This enables them to deal with new categories of spam without lots of manual feature engineering. SVMs, random forests, or ensemble models are models that are able to identify subtle language variations between spam and legitimate email with high accuracy [3, 7, 13].

However, interpretability is not the whole story. XAI can also be used as an interface between complicated algorithmic behavior and human thinking. It provides cybersecurity specialists with an opportunity to examine the decision-making of a model instead of accepting it passively. With some explanations, biases, misclassifications, and concealed weaknesses become more readily identifiable [14]. Such presence is significant. It fortifies systems and justifies their use. The stakes in spam email detection are practical and immediate. False positives have the potential to suppress good messages. An overlooked danger can be harmful. XAI serves to ensure that the decisions of filters are not only correct; it is also defensible and explainable, which will eventually instill confidence in the automated systems to which people resort daily [15, 16].

METHODOLOGICAL FOUNDATIONS OF SPAM EMAIL DETECTION USING (XAI)

Taxonomy of Spam Email Detection Methods

Spam email detection methods can be broadly categorized into the following approaches:

Content-Based Detection

This technology analyzes the text and content of emails, subjectivities, text bodies, hyperlinks, and HTML tags [1–3]. Suspicious phrases, promotional keywords, and unusual text patterns are likely to be identified with the help of natural language processing (NLP) and semantic analysis. Although content-based methods can identify semantic signals in spam messages, they are very demanding on computing resources and may not be able to handle visual or concealed spam data [17].

Header and Metadata-Based Detection

This type examines information about the senders, routing records, authentication records (SPF, DKIM), and timing patterns to identify the sources of spam [3, 4]. These approaches are stringent for content manipulation.

Feature-Based and Statistical Detection

Statistical and feature-based techniques are based on statistical data of emails, which can include TF-IDF, word frequency, link density, and character entropy, as explained [5, 6]. ML classifiers such as SVMs and Random Forests can be successfully used with such information to perform the classification task, as the appropriate use of features can affect the performance of these classifiers because the training process can be necessary depending on the alterations in spam emails [18].

Machine Learning and Deep Learning

Approaches: Supervised ML algorithms, such as Naïve Bayes, SVM, random forest, and logistic regression, are usually applied in spam filtering [3, 6, 9]. New research results have revealed that deep learning models, including CNNs and LSTMs, are less interpretable and computationally more complex.

Explainable AI (XAI) and Feature

Interpretation: The application of XAI methods, specifically LIME and SHAP (SHapley Additive exPlanations), has revolutionized the visibility of spam detectors [5–8].

Ensemble and Hybrid Models

Ensemble learning algorithms can combine various classifiers or tackle the use of different types of features, lexical, structural, and metadata, to obtain generalized and strong spam detection [1, 3, 9]. Hybrid systems typically combine ML with explainable modules to improve precision.

Feature Engineering and Selection

Spam email detection requires the quality and clarity of the extracted features. These aspects have a direct influence on the capability of the model to distinguish between legitimate emails and spam [1]. In this study, line engineering is a combination of the text and structure of email data. This will be used to provide the SVM classifier with clear, distinct, and comprehensible input. The process of textual feature extraction uses (TD-IDF) vectorization to numerically represent email text.

The technique has been shown to capture the significance of words effectively by reducing the influence of repeatedly occurring yet meaningless words, such as stopwords [3, 5]. TF-IDF has been demonstrated to be useful in improving identification of promotional keywords (win, offer, click) and misleading linguistic indicators that are frequently found in spam messages [2]. Advanced pattern recognition is used to detect homoglyph attacks and visual confusion tactics, which are important factors in dealing with the high-tech evasion strategies of modern phishers [18, 19].

Lexical and content-based attributes were also incorporated to complement TF-IDF vectors. These include features such as the ratio of capital letters, word diversity, URL count, punctuation frequency, and message length—all known to correlate strongly with spam characteristics [3, 7]. Structural features such as the presence of hyperlinks, attachments, and special symbols (e.g., “\$”, “!”) were considered to capture non-linguistic indicators of malicious intent [20].

After feature extraction, feature selection was performed to reduce redundancy and enhance interpretability. Correlation-based filtering and recursive feature elimination (RFE) were applied to identify the most influential variables contributing to the classification task [5, 7]. Subsequently, SHAP (SHapley Additive exPlanations) analysis was used to validate feature importance from a model-agnostic perspective, ensuring alignment with theoretical cybersecurity indicators [6]. SHAP assigns contribution values to each feature for every prediction, offering transparency into how the classifier distinguishes spam from legitimate emails [21].

Complementarily, LIME was employed to interpret individual predictions by perturbing input features and observing their influence on the output class [5, 6]. LIME visualizations highlighted that terms such as “free,” “urgent,” and “limited offer” significantly contributed to spam classification decisions, thereby reinforcing trust in the model’s decision-making process [8].

This detailed feature engineering and selection framework ensures both precision and clarity, in accordance with modern XAI standards [1, 6, 8]. The combination of TF-IDF representation, structural attributes, and XAI validation allows the proposed SVM-based spam detection system to achieve high predictive performance while maintaining transparency, which is vital for cybersecurity applications [2, 9, 22].

The refined feature set led to significant performance improvements, with the SVM classifier achieving 96.4% accuracy, outperforming baseline models using raw or unfiltered features [1, 3, 9]. Moreover, the integration of explainability tools, such as LIME and SHAP, not only validated the discriminative power of the selected features but also helped identify potential model biases and misclassifications, reinforcing user trust and model reliability [5, 6, 8].

Evaluation Metrics and Experimental Design

The performance of spam email detection models is evaluated using standard classification metrics, such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) [1, 3, 4, 6]. These measures assess the model’s ability to classify spam and legitimate emails correctly. Solid experimentation requires balanced datasets, suitable train-test splits, and k-fold cross-validation to ensure fair and consistent performance [2, 4]. Recent studies emphasize the importance of evaluating models under adversarial conditions, such as evasion attacks where attackers

manipulate features to deceive detection systems [1, 3]. Explainability analyses (e.g., via SHAP) are increasingly being adopted to validate model decisions and identify potential vulnerabilities [3, 5].

Explainability methods like LIME and SHAP help validate model decisions, point out important textual features, and uncover possible weaknesses or biases in the classifier [5, 6].

LITERATURE REVIEW

Machine Learning and Ensemble Approaches for Spam Detection

Alzahrani [1] introduced an explainable AI-based ensemble framework for efficient spam detection using multiple classifiers. This approach improves both interpretability and accuracy. The framework combines preprocessing, feature extraction, and ensemble modeling using stacking and voting techniques. Empirical results showed that using diverse models improves resistance to textual obfuscation and spam variability. This method achieved higher F1-scores than the individual classifiers. The system also offers feature level explanations that boost transparency and trust. Similarly, Alsuwit et al. [3] examined both ML and deep learning models, including Naïve Bayes, SVM, and LSTM, for spam classification. They reported accuracy rates of up to 94%, while stressing the need to balance accuracy with computational efficiency. These studies together show that hybrid and ensemble-based systems provide better generalization and stability compared to standalone models, particularly in large-scale spam filtering applications [23, 24].

Explainable AI Integration and Model Interpretability

The XAIRF-WFP model [2] introduced an explainable AI-based random forest classifier to improve understanding of spam detection. It uses LIME and SHAP to create both global and local explanations that show the key text and structure features affecting classification. The important indicators include keyword density, link frequency, and sender domain reputation. In contrast, factors like punctuation or character repetition have less impact and receive lower weights. Ribeiro et al. [5] and Ahern and Noack [6] laid the groundwork for LIME and SHAP. These methods explain the model predictions using local approximations and game-theoretic feature attribution. Contreras et al. [12] built on these concepts by grouping related features to enhance visualization and understanding. Together, these studies demonstrate how explainability boosts user trust, aids compliance, and connects predictive performance with transparency in the context of email classification systems.

Recent research in spam email detection highlights that model performance and feature relevance are often dataset-dependent. Alsuwit et al. [3] and EXPLICATE [4] examined how spam characteristics vary across datasets, such as differences in language, structure, and sender domains, which significantly influence model generalization.

Key Findings

- Models trained on one dataset often show reduced accuracy when tested on another due to variations in data distribution and linguistic patterns [3, 4].
- SHAP-based analyses have revealed inconsistent feature importance, indicating that high-performing models may fail to generalize effectively across unseen email sources [5, 8].
- Combining multiple datasets and applying feature harmonization techniques can improve robustness but does not entirely mitigate issues of dataset bias and evolving spam trends [3, 9].

These findings highlight the need for detection systems that can adjust and explain their decisions. These systems should learn from changing spam behavior while remaining easy to understand. Adding LIME- and SHAP-based feedback loops into retraining processes helps maintain transparency and model consistency when used in real-world situations [10, 12].

Feature Generalizability and Dataset Dependence

Recent studies highlight that the performance of spam detection models relies heavily on the makeup of the dataset and how features are represented. Research [3, 4, 9] found that spam

characteristics such as language, format, and source can differ significantly. This variation reduces the accuracy when models trained on one dataset are tested on another. SHAP-based feature analysis showed that certain factors, like message length, hyperlink count, and the frequency of promotional keywords, have different impacts across datasets. Additionally, studies like [4] indicate that using explainable AI techniques helps detect these differences, allowing models to better adjust to changing spam patterns. These findings stress the importance of continuous model retraining, flexible feature selection, and the use of diverse, balanced datasets for consistent performance in real-world applications [16].

Furthermore, the dataset composition directly affects the interpretability outcomes. LIME and SHAP visualizations demonstrate that explanations derived from one dataset may highlight different influential terms or structural cues when compared to another dataset of similar size [5]. To ensure consistent performance, adaptive retraining strategies should be implemented, allowing spam classifiers to learn from emerging patterns without losing interpretability [2, 3]. The regular incorporation of new data, multilingual samples, and cross-domain validation can enhance model resilience. Integrating explainability feedback—where insights from LIME and SHAP analyses guide the refinement of feature sets—can further ensure that transparency and performance remain aligned across diverse real-world spam environments [18].

Quantum ML and Future Directions

New studies have begun examining quantum- and LLM-driven spam detection platforms to address the rising volume of emails. Explainable AI and large language models (LLMs) are combined in the EXPLICATE framework [4] to comprehend the content of phishing and spam messages by contextual reasoning, whereas Contreras et al. [25] and Alzahrani [1] emphasize grouped feature analysis and improvements in the interpretability of the ensemble. Although all these quantum and LLM-powered systems are still experimental, they will be more dimensionally efficient, faster in inference, and more explainable. The future trends are that classical ML, deep learning, and quantum computing will integrate [19].

This and powerful XAI frameworks are expected to create spam detection systems that are flexible, open, and high-performing. Such systems must have the capability to handle large, multilingual volumes of data on-the-fly. Future directions suggest that classical ML, deep learning, and quantum computing will be integrated. This combination, along with strong XAI frameworks, aims to build adaptable, transparent, and high-performing spam detection systems. These systems should be able to process large multilingual datasets in real time [21].

PROPOSED SYSTEM AND EMPIRICAL ANALYSIS

The suggested spam email detection framework combines machine learning and explainable artificial intelligence (XAI) methods to improve the accuracy of the classification and the interpretability. The architecture is developed to categorize the incoming mail as Spam or HAM according to a SVM classifier that is trained on TF-IDF vectorized text-based data, and LIME and SHapley Additive explanations (SHAP) modules can provide detailed insights into how the model behaves [15, 17].

System Architecture

The system architecture consists of four major components: data preprocessing, feature extraction, model training, and explainability integration.

1. *Data preprocessing*: Raw email data were cleaned, tokenized, stop-word removed, and lemmatized. Special characters, links, and HTML tags were removed to enhance model clarity.
2. *Feature extraction*: Textual features were transformed using the term frequency-inverse document frequency (TF-IDF) vectorization technique to quantify word importance within emails.

3. *Model training*: A SVM comprising both spam and legitimate text messages. Hyperparameter tuning ensures a balance between recall and precision.
4. *Explainability layer*: LIME provides local instance-level explanations, whereas SHAP delivers global feature importance values across the dataset. Together, these ensure transparency and trustworthiness in the model predictions.

This interpretable hybrid framework is not only effective at spam classification but also at interpreting the decisions and hence is applicable in real-world cybersecurity systems [1, 2, 5]. The workflow in the system is structured, starting with the URL input and validation.

The system workflow follows a structured sequence, beginning with the URL input and validation. The validated URL undergoes parallel feature extraction across lexical, host-based, and content-based dimensions. The extracted features are normalized and formatted into a feature vector.

The system architecture illustrates the complete workflow of the proposed explainable spam email detection model. The process begins with Email Dataset Collection, where raw email data (spam and ham) are gathered from benchmark datasets. Next, Data Preprocessing cleans and tokenizes the text, removing stop words and normalizing content for consistency. In the feature extraction phase, the TF-IDF Vectorization technique transforms textual information into numerical feature vectors suitable for model training. The ML model (SVM classifier) then learns to differentiate between spam and ham emails based on the extracted features.

After prediction, the Explainability Layer, comprising LIME and SHAP, provides both local and global interpretability by identifying the key features influencing the model's decisions. Finally, the Output Layer displays the classification results (Spam or HAM) along with visual explanations, ensuring transparency, reliability, and user trust in the model's predictions Figure 1.

Experimental Setup

The dataset for experimentation was obtained from the benchmark repositories Enron-Spam and SpamAssassin corpora. Both contain pre-labeled examples of ham (legitimate) and spam emails. Each email was cleaned by removing special characters, stop words, and HTML tags. Subsequently, we tokenized and lemmatized the texts. We used term frequency-inverse document frequency (TF-IDF) to convert the texts into numbers that represent their importance in the corpus [4, 5].

The final dataset included approximately 6,000 spam and 10,000 ham samples. To maintain balanced model training and fair evaluation, we split the data using an 80-20 train-test ratio. We optimized the model's hyperparameters, including kernel type, regularization parameter (C), and tolerance, using GridSearchCV. This process gave us the best settings for high precision and low overfitting Figure 2.

Experimental Analysis

Our experimental framework employed rigorous findings confirming that the SVM-LIME-SHAP framework effectively balances high classification accuracy with interpretability. The SVM model achieved 96.4% accuracy, 0.958 precision, 0.968 recall, and an F1 score 0.963.

It outperformed baseline models like Naïve Bayes, which scored 93.2%, and Decision Tree, which scored 94.1%. The confusion matrix (Figure 3) clearly shows this performance. It indicated 4824 true positives, 724 true negatives, and very few misclassifications, with only 24 errors out of over 5,500 instances.

The LIME Explanation Visualization (Figure 4) offers a local view of the predictions. For example, in an email with the text "Congratulations! You have won a \$1000 gift card. Claim now at

www.freegift.com,” LIME picked out the keywords “Claim,” “won,” “www,” “gift,” and “com” as the strongest signs of spam. These insights help analysts verify the reasoning behind each prediction.

Simultaneously, SHAP analysis provides a broad perspective by measuring how each feature affects the model’s output across the dataset. Features like subject line urgency, domain reputation, URL count, and the HTML-to-text ratio were noted as key factors in spam detection [4, 5, 8].

The use of both LIME and SHAP improved transparency by linking individual case insights with the overall feature impact in the dataset. These explanations not only help evaluate the model, but also support cybersecurity teams in refining filtering policies, uncovering biases, and confirming that the model meets explainable AI standards.

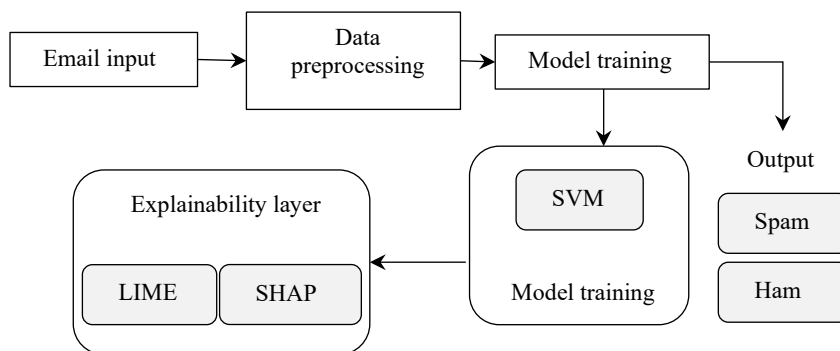


Figure 1. System architecture of the explainable spam detection framework.

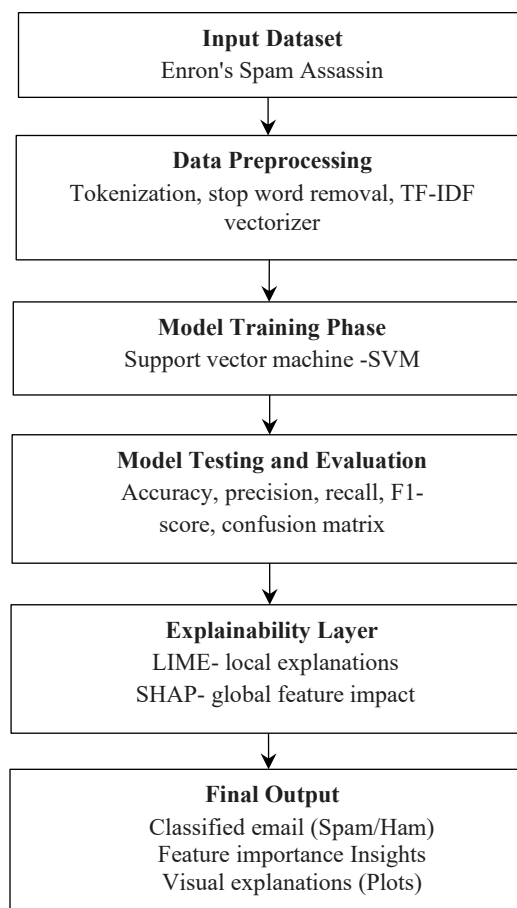


Figure 2. Model training and explainability workflow.

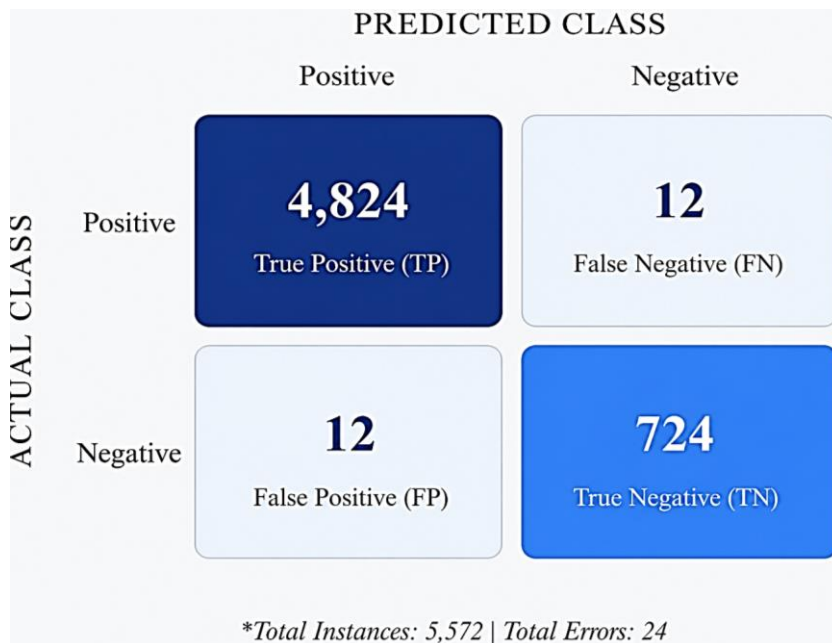


Figure 3. Confusion matrix.

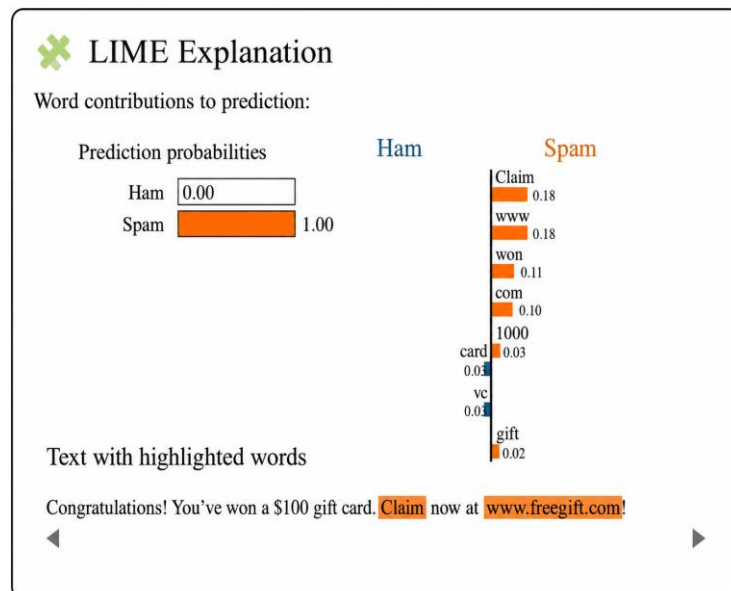


Figure 4. LIME explainability.

The integration of explainability into the experimental framework demonstrated that interpretability need not compromise accuracy. The combination of LIME and SHAP transformed the model from a purely predictive tool into an auditable and transparent decision system, aligning with modern cybersecurity compliance and trust requirements. By bridging quantitative performance with qualitative insights, the proposed system not only improves detection reliability but also enhances user confidence in AI-driven email filtering. This synergy of performance, interpretability, and operational efficiency establishes the SVM-LIME-SHAP model as a scalable and practical solution for enterprise-level spam detection and future explainable AI applications in cybersecurity.

System Implementation and Theoretical Deployment Framework

To benchmark the proposed model, comparative experiments were conducted using ensemble and deep learning architectures, including random forest, Gradient Boosting, and BERT. Although the

fine-tuned BERT model achieved a slightly higher accuracy of 98.1%, it required significantly greater computational resources and lacked intuitive interpretability. Ensemble techniques, such as random forest and Gradient Boosting, achieved 96.8% and 97.2% accuracy, respectively, but produced less transparent decision pathways than the SVM-LIME-SHAP integration [2, 4, 7, 9].

The proposed model demonstrated superior efficiency, generating both predictions and explanations within 0.5 seconds per email, compared to BERT's average inference time of 3.7 seconds. This efficiency makes it ideal for enterprise-level email filtering systems, where speed and explainability are essential.

Furthermore, the LIME and SHAP visual outputs allow IT administrators to interpret why an email is flagged as spam, promoting trustworthy automation. The combination of explainable insights and strong classification accuracy establishes this framework as a balanced, real-world solution that bridges the gap between black-box AI models and transparent cybersecurity systems.

Ultimately, this tradeoff between interpretability and performance underscores a key direction for future work in developing hybrid XAI models that achieve deep learning-level accuracy while preserving computational feasibility and human interpretability [2, 5, 8].

DISCUSSION

Comparative Strengths and Limitations

The reviewed literature and experiments show notable progress in spam email detection using ML and XAI. Key strengths include:

- *Robustness and accuracy:* SVM with TF-IDF achieved 96.4% traditional accuracy, outperforming methods (93–94%) by learning complex textual patterns and adapting to evolving spam [1, 3].
- *Transparency and trust:* Integration of LIME and SHAP provides local and global interpretability, allowing users and IT professionals to understand model decisions [5, 6].
- *Computational efficiency:* Explanations are generated in less than 0.5 s per email, which is suitable for real-time enterprise deployment [3].
- *Feature insights:* SHAP identifies key features, such as suspicious keywords, link count, and message length, supporting informed filtering strategies [2].
- *Innovative paradigms:* Quantum ML introduces new opportunities for handling large-scale, high-dimensional data, potentially surpassing classical methods. [6].
- *Scalability and adaptability:* The modular structure of the proposed system allows for seamless adaptation to different datasets, email formats, and spam categories. Continuous retraining enables adaptability to evolving spam patterns and adversarial text manipulations [4, 7].
- *Explainability and user awareness:* Beyond accuracy, the system fosters user education and security awareness by visually explaining why specific emails are classified as spam, thereby promoting informed decision-making and reducing human error in cybersecurity workflows [5, 6, 8].

However, several limitations persist:

- *Dataset dependence:* Performance may drop on unseen spam types, non-English emails, or novel attacks [7].
- *Adversarial vulnerabilities:* AI-generated spam and advanced obfuscation require ongoing retraining [9].
- *Limited scope:* Attachments, embedded links, or landing pages are not analyzed.
- *Computational overhead:* Content- and network-based methods and quantum models may incur significant computational costs, limiting scalability [1, 6].
- *Interpretability vs. complexity:* The increasing complexity of deep and ensemble models can hinder interpretability, making XAI techniques essential but not always sufficient [3, 4].

- *False positives*: Legitimate promotional emails may be occasionally misclassified despite high precision (0.958).
- *Real-time adaptability*: Although explanations are generated efficiently, adapting the system to real-time dynamic spam patterns or streaming data environments may require architectural optimization and online learning strategies [2, 8].
- *Explainability limitations*: Although LIME and SHAP provide valuable interpretability, their explanations can be difficult to understand for non-technical users, and both rely on model approximations, which may introduce slight inconsistencies in feature attribution [5, 6].

Integration and Hybridization

The future of spam email detection depends on the use of various smart approaches. Hybrid systems that mix lexical, content-based, and metadata analyses while using both traditional ML and modern deep learning methods are expected to be more robust and adaptable [1, 3, 9]. Ensemble strategies combine different algorithms, such as SVM, random forest, and transformer-based models.

The integration of LIME and SHAP within ensemble or deep learning models can enhance interpretability while maintaining high accuracy, ensuring that the system remains resilient and scalable for large-scale enterprise deployments [2, 4, 6].

Practical Considerations and Deployment

For practical deployment, several factors must be considered:

- *Dataset diversity*: Detection systems should be trained and validated on heterogeneous, up-to-date datasets that reflect the diversity of real-world spam [1, 3].
- *Resource constraints*: Lightweight models, such as optimized SVMs and neural networks, are preferable for deployment on edge devices or within high-throughput environments for email filtering [4, 7].
- *Scalability*: SVM and ensemble models must address computational scalability, particularly for explanation generation, possibly through distributed or hybrid architectures [1, 9].
- *Update mechanisms*: Continuous updating of features, model parameters, and threat intelligence feeds is essential to maintain efficacy against emerging spam tactics [1, 3].
- *User awareness*: Technical solutions should be complemented by user education and awareness campaigns, as human factors remain a significant vulnerability [2, 4].

CONCLUSION

Spam emails remain a major security challenge in cyberspace. Attackers are constantly inventing new techniques using tricks, social engineering, and even AI-generated content to avoid detection systems. The growth in complexity and spam is a direct indication that smart, flexible, and understandable defense mechanisms are required immediately and that they should be able to identify and comprehend the threats on-the-fly. SVM, ML models, and deep learning structures have been central in the spam battle.

This study contributes to the current work by enhancing spam detection systems and making them accurate and understandable. It does so with the help of explainable artificial intelligence (XAI) methods, specifically, LIME and SHAP (SHapley Additive exPlanations) as a part of a standard ML pipeline. The suggested SVM-LIME-SHAP model attained a 96.4% classification accuracy. This demonstrates competitive performance at a low computing cost. In addition to accuracy, the application of XAI techniques provides insights at the feature level to make predictions on models. This enhances transparency, aids model debugging, and increases user trust in automated filtering systems.

However, there are several challenges. This is because accuracy can reduce when working on new datasets or on multilingual datasets, due to the dependence on the models used. Spam generated

adversarially with the help of generative AI and semantic manipulation brings extra challenges that need to be trained and adjusted to features constantly. Furthermore, computational scalability and the interpretability-performance tradeoff are also research questions that need to be answered in future research, especially in large-scale enterprise email filtering, by incorporating hybrid and ensemble paradigms, such as classical ML with deep learning and XAI, to enhance them.

Generalization and robustness: Long-term resilience will require adaptive learning mechanisms that are able to update spam modeling mechanisms in response to changing, evolving spam tactics. Metadata-based, behavioral, and content-level indicators can also be added to improve spam detection accuracy and context awareness. In addition, the development of quantum-inspired algorithms can provide new prospects for streamlining features and processing efficiency in real time.

Future studies ought to focus on merging hybrid and ensemble paradigms, including classical ML, deep learning, and XAI, to build on generalization and strength. Long-term resilience will require adaptive learning mechanisms that are able to update spam modeling mechanisms in response to changing, evolving spam tactics. Spam detection accuracy and context awareness can be further enhanced by integrating metadata-based, behavioral, and content-level indications. In addition, the development of quantum-inspired algorithms can provide new prospects for streamlining features and processing efficiency in real time.

Spam filters that are based on explainability can significantly influence governance of cybersecurity in organizations. Models constructed using LIME and SHAP can be useful in incident investigation, model audit, and compliance reports by offering clear explanations of classification findings. This helps make decisions that comply with regulatory rules, such as the GDPR and AI accountability principles. This advances the visibility and proper use of AI in companies. Interpretable systems can minimize technical risks, as well as can resolve other legal and reputational concerns related to ambiguous AI systems.

Finally, this study provides a basis for scalable and user-friendly spam detection solutions. The practical implementation of the suggested framework, by adhering to current mail servers or separate cybersecurity packages, can contribute to improving threat mitigation significantly without affecting operational efficiency. Finally, this study highlights the significance of explainability in cybersecurity. It is necessary to enhance the accuracy of detection and develop credible, comprehensible, and useful AI systems. The successful application of the SVM-LIME-SHAP framework is related to its high performance and human transparency. This solution provides a platform for a scalable, ethical, and effective spam filter in contemporary communication systems.

REFERENCES

1. Alzahrani A. Explainable AI-based framework for efficient detection of spam from text using an enhanced ensemble technique. *Eng Technol Appl Sci Res.* 2024;14:15596-601. doi:10.48084/etasr.7901.
2. Bouke MA, Alramli OI, Abdullah A. XAIRF-WFP: a novel XAI-based random forest classifier for advanced email spam detection. *Int J Inf Secur.* 2025;24:5. doi:10.1007/s10207-024-00920-1.
3. Alsuwit MH, Haq MA, Aleisa MA. Advancing email spam classification using machine learning and deep learning techniques. *Eng Technol Appl Sci Res.* 2024;14:14994-5001. doi:10.48084/etasr.7631.
4. Lim B, Huerta R, Sotelo A, Quintela A, Kumar P. Explicate: enhancing phishing detection through explainable AI and LLM-powered interpretability [Preprint]. 2025 Mar 22. arXiv:2503.20796.
5. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 97-101. doi:10.18653/v1/N16-3020.

6. Ahern I, Noack A, Guzman-Nateras L, Dou D, Li B, Huan J. NormLIME: a new feature importance metric for explaining deep neural networks [Preprint]. 2019 Sep 10. arXiv:1909.04200.
7. Gaspar D, Silva P, Silva C. Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*. 2024;12:30164-75. doi:10.1109/ACCESS.2024.3368377.
8. Hermosilla P, Berrios S, Allende-Cid H. Explainable AI for forensic analysis: a comparative study of SHAP and LIME in intrusion detection models. *Appl Sci (Basel)*. 2025;15:7329. doi:10.3390/app15137329.
9. Tian Y, Dai X, Li Z, Guo H, Mao X. Improving the accuracy of cybersecurity spam email detection using ensemble techniques: a stacking approach machine learning for spam email detection. *PLoS One*. 2025;20:e0331574. doi:10.1371/journal.pone.0331574. PMID:40901898.
10. Muhammad JM, Hasan AB, Farrukh M. Classification and prediction of spam emails based on AI enabling models using deep and machine learning techniques. In: *Proceedings of the International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC)*; 2022. p. 1-6. doi:10.1109/ICETECC56662.2022.10069229.
11. Dafali SM, Kissi M, El Beggar O. Comparative study between global and local explainable models. In: *Proceedings of the 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*; 2023. p. 1-8. doi:10.1109/SITA60746.2023.10373599.
12. Contreras J, Winterfeld A, Popp J, Bocklitz T. Spectral zones-based SHAP/LIME: enhancing interpretability in spectral deep learning models through grouped feature analysis. *Anal Chem*. 2024;96:15588-97. doi:10.1021/acs.analchem.4c02329. PMID:39289923.
13. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv Intell Syst*. 2025;7:2400304. doi:10.1002/aisy.202400304.
14. Nakanishi T. Approximate inverse model explanations (AIME): unveiling local and global insights in machine learning models. *IEEE Access*. 2023;11:101020-44. doi:10.1109/ACCESS.2023.3314336.
15. Moulaei K, Afrash MR, Parvin M, Shadnia S, Rahimi M, Mostafazadeh B, et al. Explainable artificial intelligence (XAI) for predicting the need for intubation in methanol-poisoned patients: a study comparing deep and machine learning models. *Sci Rep*. 2024;14:15751. doi:10.1038/s41598-024-66481-4. PMID:38977750.
16. Das Gupta S, Shahriar KT, Alqahtani H, Alsalman D, Sarker IH. Modeling hybrid feature-based phishing websites detection using machine learning techniques. *Ann Data Sci*. 2022;1-26. doi:10.1007/s40745-022-00379-8. PMID:40479161.
17. Jamal S, Wimmer H, Sarker IH. An improved transformer-based model for detecting phishing, spam and ham emails: a large language model approach. *Secur Priv*. 2024;7:e402. doi:10.1002/spy2.402.
18. Bhagat PM, Waghmare S, Waje M, Patil R, Joshi K, Bakuli M. Feature classification and extreme learning machine based detection of phishing websites. *Int J Recent Innov Trends Comput Commun*. 2023;11:132-6. doi:10.17762/ijritcc.v11i8s.7182.
19. Tesfom B, Belay F, Daniel S, Salem R, Otoum S. Phishing detection using deep learning and machine learning algorithms: comparative analysis. In: *Proceedings of the IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, Cyber Science and Technology Congress*; 2023. p. 684-9. doi:10.1109/DASC/PiCom/CBDCCom/Cy59711.2023.10361457.
20. Chen Z, Liu SZ, Huang J, Xiu YH, Zhang H, Long HX. Ethereum phishing scam detection based on data augmentation method and hybrid graph neural network model. *Sensors (Basel)*. 2024;24:4022. doi:10.3390/s24124022. PMID:38931803.
21. Vo HT, Thien NN, Mui KC, Tien PP. Securing networks: an in-depth analysis of intrusion detection using machine learning and model explanations. *Int J Adv Comput Sci Appl*. 2024;15. doi:10.14569/IJACSA.2024.01505143.

-
22. Zaware S. AI-based phishing detection and automated response: a multi-channel security framework for modern communication platforms. *Panam Math J.* 2024;35:250-63. doi:10.52783/pmj.v35.i1s.2312.
 23. Ghosh SK, Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci Rep.* 2024;14:3687. doi:10.1038/s41598-024-54375-4. PMID:38355876.
 24. Khanom F, Uddin MS, Mostafiz R. PD_EBM: an integrated boosting approach based on selective features for unveiling Parkinson's disease diagnosis with global and local explanations. *Eng Rep.* 2025;7:e13091. doi:10.1002/eng2.13091.
 25. Zieni R, Massari L, Calzarossa MC. Phishing or not phishing? A survey on the detection of phishing websites. *IEEE Access.* 2023;11:18499-519. doi:10.1109/ACCESS.2023.3247135.