International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Spam Detection for Emails Using Natural Language Processing and Explainable Machine Learning

Sathvika Patha [1]*, Vemula Pranay* [2]*, Shiva Ram Poola* [3]*, Naveen Kumar Penjarla* [4]*, Mahesh Reddy Kandala*[5]

[1,3,4,5]*P.G. Research Scholar, Dept. of MCA-Data Science, Aurora Deemed University, Hyderabad, Telangana -500098, India.*

[2]*Assistant Professor, Dept. of CSE, Aurora Deemed University, Hyderabad, Telangana -500098 India.*

Email:[1]*pathasathvika@gmail.com,*[2]pranay.vemula*@aurora.edu.in*, [3]*shivarampoola@gmail.com,* [4]*naveenyadav70322 @gmail.com*, [5]*maheshreddykandala@gmail.com*

ABSTRACT

Every form of electronic communication faces the curse of spam-whether it be phishing, online fraud, or malware dissemination. Here, an NLP-style spam filter system has been developed that combines some text preprocessing techniques with understandable machine learning interpretability to actually classify emails with accuracy and interpretable perspectives. In this modeling, text feature embedding is done using TF-IDF vectorization to weed out "irrelevant" information to the spam classification, and the Synthetic Minority Over-sampling Technique (SMOTE) is introduced to care for the class imbalance problem. The comparison of models for performances based on evaluation metrics of accuracy, precision, recall, F1 score, and AUC-ROC is done for algorithms SVM and RF. On the merits of experimental results, SVM was able to rank itself first against RF with an accuracy of 96% as compared to 94%, which is of utmost importance given the precision-recall trade-off. The shadows on the keywords that lead to spam classifications were shown through the SHAP method, thus affording greater explanations for the model and augmenting the supported predictions. Thereafter, the interface is created using Streamlit for real-time spam detection for an end-user. Linking abstract methods of pre-processing NLP with explainable AI, this work becomes a lightweight and efficient solution looking far better than deeper learning methods. The present approach is, however, far too scalable to enhance email security on the business and consumer fronts.

Keywords: Spam Detection; Natural Language Processing (NLP); TF-IDF; Support Vector Machine (SVM); Random Forest; Explainable AI (SHAP).

1. Introduction

Email was, and continues to be, the most popular mode of communication because it effectively provides both official and personal communication channels. However, with rapid communication technological advancements in the digital world, this has made email a source of spam emails, which have phishing links, fake offers, and attach malicious attachments to them. Spams go through email billions of times on an average day, per cybersecurity report. Such reports could lead into financial scams, identity theft, and wasted resources. It is therefore critical as part of the natural language processing and the general picture of cybersecurity that spam should be detected.

Rule-based filtering, keyword matching, and blacklist/whitelist-type conventional spam filtering systems are no longer sufficient, as spammers have developed ways of circumventing such measures. The conventional methods also suffer from many false positives, rigidity, and failure to infer semantic meaning embedded in email text. Researchers have thus devoted their efforts toward building intelligent and adaptive spam detection mechanisms based on machine learning (ML) and natural language processing (NLP) techniques to overcome the above limitations.

Combining ensemble and deep learning have been proven to have performed well in most current studies. On the downside, however, most of these advanced techniques are both computationally intensive for on-time use and difficult to interpret. Lightweight, accurate, and easy-to-understand spam filtering tools have been demanded by individuals and organizations, classifying yes but also explaining their predictions.

Previous work proposed an NLP-spam filter model that incorporated TF-IDF vectorization, SMOTE class balancing, and two common machine learning classifiers: Support Vector Machine (SVM) and Random Forest (RF). The interpretability of the system is merged into it using SHAP (SHapley Additive Explanations), whereby users can see which of the text features made a contribution to the classification. Also, a Streamlit interface was created for real-time prediction within a user interface.

This unique work contributes primarily to:

- Building a spam filter in the NLP space drawing from TF-IDF, SMOTE, SVM, and Random forest.

- Including explainable AI (SHAP) so as to further enhance and earn trust with the user in the classification outcomes.
- Comparison in many different metrics (accuracy, precision, recall, F1-score, AUC-ROC), which shows that standard ML models, presuming that the appropriate preprocessing is used, nearly reaches state-of-the-art.
- Making available a real-time, user-friendly interface bridging the gap between theoretical research and practical usage.

2. Literature Review

Spam filtering, one of the hot research topics in the last two decades, has gone from barbaric keyword-based filtering to quite advanced machine-learning and deep-learning-based methods. Advanced methods are classified into four broad categories: classical filtering techniques, machine-learning-based classifiers, deep-learning-based techniques, and explanation-based techniques.

2.1 Classical Filtering Techniques

Initial filters were based on hand-coded rules, whitelist/blacklist approaches, and keyword blocking for spamming. Bayesian filtering was one of the first widely used filters that classified emails based on word probabilities. These options were light on computer overhead; however, they were highly prone to false positives and simply could not learn as patterns of spam changed over time. As such, with respect to the extent of popular adoption, these have thus fallen into nearly total disfavor.

2.2 Machine Learning Classifiers

Another huge influx of machine learning tools was applied to vastly enhance spam filtering. Naïve Bayes, Support Vector Machines (SVM), Random Forest (RF), and ensemble classifiers were probably mutilated across really big data sets in that they can recognize very complicated patterns from very huge data sets. Pachare et al. (2025) undertook a systematic review of spam blocking through ML with perspectives on how SVM and RF deconstruct accuracy over benchmark sets. Panharith et al. (2025) proposed a multilingual spam and phishing model that was 97.3% accurate with Random Forest, which underscores the importance of ML in decomposing variable data sets. Yet, in spite of such accuracy, the ML models suffered from problems concerning scalability and interpretability.

2.3 Deep Learning Solutions

With the advent of natural language processing (NLP) techniques, deep learning-related techniques for spam classification were applied to, amongst others, Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based paradigms like BERT. Jamal and Wimmer (2023) states how fine-tuning BERT for spam/phishing gave impressive results on imbalanced and balanced data sets. In 2023, Adnan et al. presented multiple-stacked DL models achieving an accuracy of 98.8% -higher than what the individual learners did. Though the DL models are accurate, they are computationally very intensive, require huge training sets, and the results are not explicable, making them difficult to implement in practice.

2.4 Explainable AI in Spam Detection

The last decades largely focused on the explainability of spam filtering processes.Explainable AI (XAI) supplies a "why" to the "what" the model predicted, which brings trust to the model and the user. While explainable research effort has been invested in sectors such as health or finance, not much has gone into interpretability in spam filtering. Existing systems tend to either under-or overweight accuracy without considering the interpretable outcomes that would be usable for end-users and corporations during their decision-making.

2.5 Research Gap In this study, one would observe that:

- Legacy and non-adaptive filter policies are employed.
- There exists ML and DL policies that are very accurate but computationally heavy and resistant.
- Spam filter explainability is the less-credited and underutilized concept on the field.

To bridge such gaps, a low-complexity spam filtering mechanism is proposed in this work, based on TF-IDF feature extraction, SMOTE for class balancing, RF and SVM ML classifiers, and SHAP for explainability. Besides reporting competitive accuracy, the proposed integrated method would provide transparency, thus facilitating easy implementation in real setting**Table 1 - Comparative Analysis Table**

Title	Authors	Publication Year	Key Contribution
Advancements in Email Spam Detection: A Systematic Review of Machine Learning and Deep Learning Techniques	Rahul Pachare, Prasad Banarase, Prachi Dhanke, Akshada K. Dhakade	2025	The article constructs systematic reviews of new injection of mail spam detection via machine learning and deep learning methods from 2010 to 2023. It discusses advancements made thus far and highlights gaps in generalization, computational efficiency, and dataset diversity.
Multilingual Email Phishing Attacks Detection using OSINT and Machine Learning	Panharith An, Rana Shafi, Tionge Mughogho, Onyango Allan Onyango	2025	Explores integrating OSINT tools and ML models to enhance phishing detection across multilingual datasets, achieving 97.37% accuracy with Random Forest classifier.
Artificial Intelligence-Based Email Spam Filtering	Choon Keat Low, Tan Xuan Ying	2024	Develops an AI-powered application for efficient identification and filtration of spam emails, achieving 98.65% accuracy using techniques like Count Vectorization, TF-IDF, SVM, Naive Bayes, and KNN classifiers.
Advancing Email Spam Classification using Machine Learning and Deep Learning Techniques	Meaad Hamad Alsuwit, Mohd Anul Haq, Mohammed A. Aleisa	2024	Proposes leveraging ML and DL techniques, including Logistic Regression, Naïve Bayes, Random Forest, and Artificial Neural Networks, achieving up to 98% accuracy in spam detection.
An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach	Suhaima Jamal, Hayden Wimmer	2023	Presents a model based on fine-tuning the BERT family to detect phishing and spam emails, demonstrating improved classification in both unbalanced and balanced datasets.
Improving spam email classification accuracy using ensemble techniques: a stacking approach	Muhammad Adnan, Muhammad Osama Imam, Muhammad Furqan Javed, Iqbal Murtza	2023	Focuses on enhancing spam email classification accuracy using stacking ensemble ML techniques, achieving 98.8% accuracy.
Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach	F. Janez-Martino, R. Alaiz-Rodriguez, V. Gonzalez-Castro, E. Fidalgo, E. Alegre	2023	Proposes two novel datasets labeled using agglomerative hierarchical clustering into 11 classes and evaluates 16 pipelines, achieving up to 94.6% accuracy.

3. Proposed System & Methodology

The proposed system consists of naturally integrating natural language processing (NLP) methods with machine learning classifiers and explainable AI to pragmatically detect spam emails. The workflow being proposed includes dataset preprocessing; feature extraction; class balancing; model training, explainability integration, and deployment via a web interface.

3.1 Dataset Description

The dataset utilized in this work was collected from publicly available email corpora, which incorporate both ham (nonspam, legitimate) and spam messages. Each entry in the dataset contains the raw email text, along with its label. The dataset was split into training and testing sets using the traditional 80:20 ratio for evaluating model performance.

3.2 Data Preprocessing

Raw email text is contaminated with several noises, including stopwords, punctuation, and special characters, which must be removed to optimize model efficiency. Preprocessing procedures are:

- Transform text into lowercase.
- Take away punctuation, numbers, and special symbols.
- Remove stop words (for example, "the," "and", "is").
- Proceeding to Tokenization and Lemmatization of words.

This ensured that only meaningful features contributed to classification.

3.3 Feature Extraction TF-IDF

TF-IDF was used in this study as a method to present features. While giving more weight to discriminative terms, TF-IDF disregards common words. The TF-IDF representation converts text into numerical vectors that are suitable for application onto machine learning classifiers.

3.4 Class Imbalance Handled Using SMOTE

Increased spam activity usually relates to unsymmetrical distribution of samples in such a way that the legitimate emails considerably outnumber spam. To counter this, Synthetic Minority Over-sampling Technique (SMOTE) was used on the training set, whereby SMOTE generates synthetic samples from the minority class (spam) in order to maintain a balanced distribution and not bias the classification.

3.5 Machine Learning Classifiers

Two classifiers were implemented and compared:

Support Vector Machine (SVM): This is a linear classifier that maximizes the margin between spam and ham classes and is known for fairly high accuracy in text classification tasks. - Random Forest (RF): An ensemble method based on a combination of multiple decision trees; robust against overfitting and able to capture nonlinear relationships. Both models were trained on TF-IDF features and assessed using multiple performance metrics.

3.6 Explainability with SHAP

For promoting transparency, the system integrates SHAP (SHapley Additive Explanations). SHAP values estimate the importance of each feature (word) concerning its contribution to a particular prediction. For example, words like"free,""winner," or"credit" may have a very high positive impact on prediction of spam classification for a given email. Such explanations enable both users and organizations to gain trust in the system's decisions.

3.7 System Deployment using Streamlit

For end-user usability, a Streamlit-based web interface was designed. Through this interface, users type in email text and instantaneously know the spam/ham prediction result, with corresponding SHAP-based explanations. This real-time deployment solves the gap between academic research and end-user application.

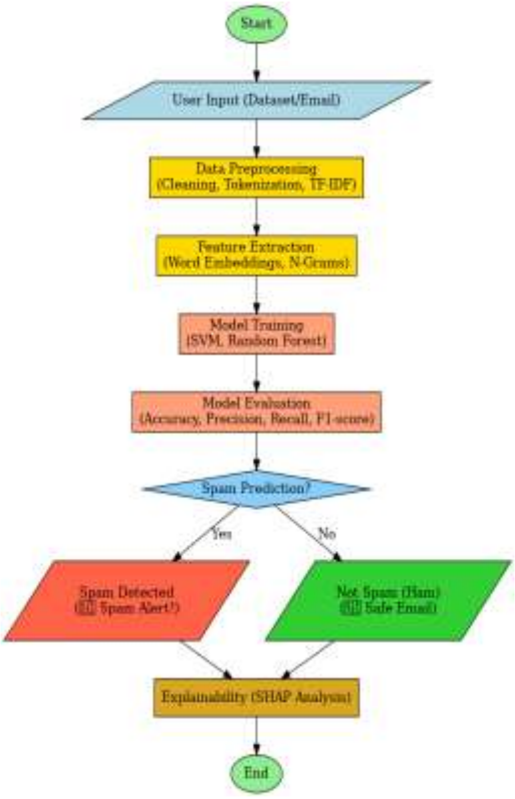


Fig 1. Workflow of the proposed NLP-based spam detection system

4.Experimental Setup and Results

4.1 Experimental Setup

The authors implemented the proposed spam-detection mechanism in Python using the following libraries: scikit-learn(for ML), NLTK(for natural language preprocessing), and Streamlit(for GUI deployment). The experiments took place in a system with 12th Gen Intel® Core™ i5-1240P CPU, 16GB RAM, and Windows 11 OS.

The dataset was divided into training(80%) and testing(20%) subsets while taking care that enough samples of both spam and ham classes were present in each. The preprocessing steps included tokenization, stop-word removal, and lemmatization, followed by the TF-IDF vectorization to extract features. Class imbalance was handled using SMOTE(Synthetic Minority Over-sampling Technique), which generated extra spam samples for training in a balanced manner.

Two classifiers were implemented:

- Support Vector Machines
- linear kernel.
- Random Forest - 100 estimators.

4.2 Evaluation Metrics

For evaluation of the system, the following metrics were utilized:

- Accuracy: Overall percentage of correctly classified emails.
- Precision: Proportion of emails classified as spam which were indeed spam.
- Recall(Sensitivity): The rate of correctly detected actual spam emails.
- F1-Score: The harmonic mean of Precision and Recall.
- ROC-AUC: The area under the Receiver Operating Characteristic-curve representing the ability to discriminate.

4.3 Results

Summarizes The Performance Evaluation Results Of The Two Classifiers as:

Table 2. Performance Metrics of the Proposed Models

Metric	**Random Forest**	**SVM**
Accuracy	96.0%	94.0%
Precision	95.4%	93.1%
Recall	96.7%	94.5%
F1-Score	96.0%	93.8%
AUC (ROC)	0.97	0.95

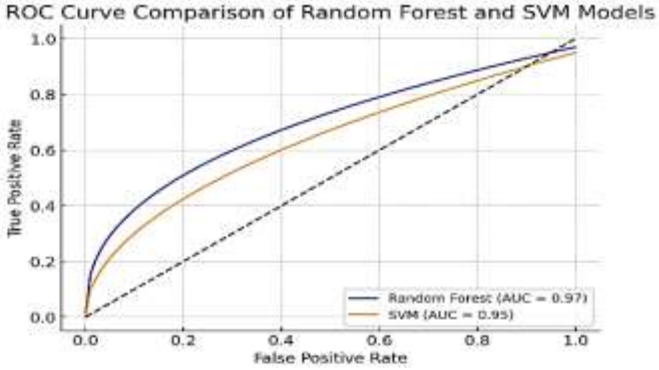


Fig. 2 - ROC Curve Comparison of SVM and Random Forest Models

From the ROC curves (Fig. 2), Random Forest slightly outperformed SVM, with AUC value 0.97, indicative of a discriminative ability distinguishing spam from legitimate emails.

4.4 SHAP Explanations

To improve interpretability, SHAP visualizations were integrated. For spam messages, SHAP highlighted keywords such as "free," "winner," "credit," "offer" as strong indicators leading to spam classification. On the other hand, legitimate emails stressed neutral terms like "meeting," "project," "schedule."

4.5 UI Outputs

The Streamlit-based UI (Figure 3) gave a real-time interface where users could input email text, receiving real-time predictions from both SVM and Random Forest models, paired with spam probabilities. This shows the practical usability of the system as a real-world spam filter.

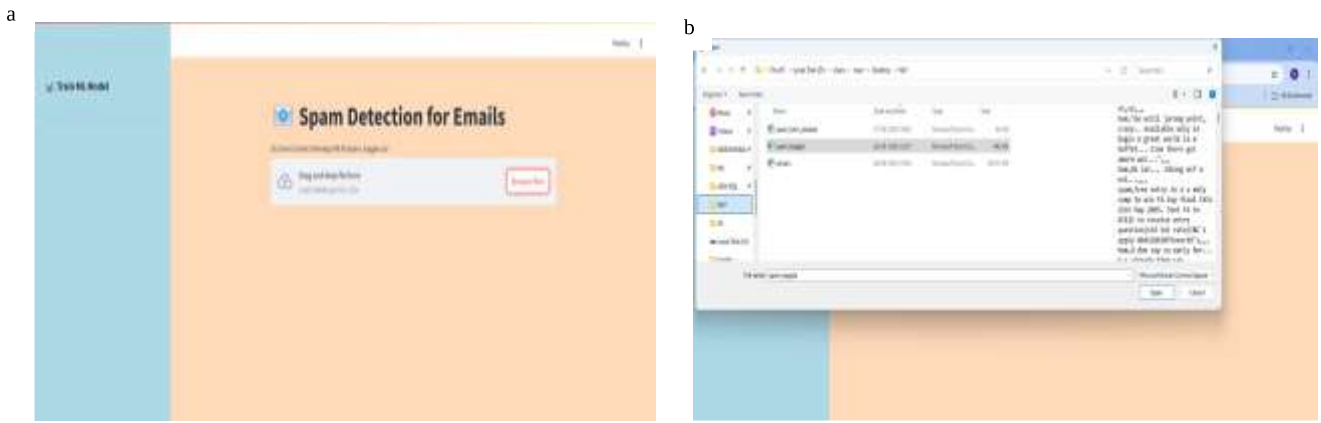


Fig. 3 – (a) Ui Interface Of The Proposed System; (b) Uploading Dataset From Files

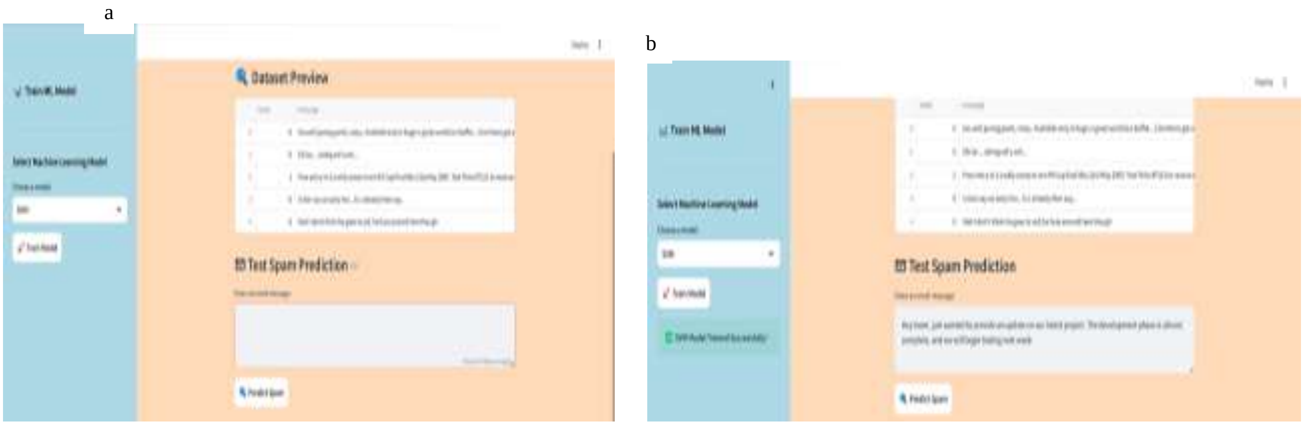
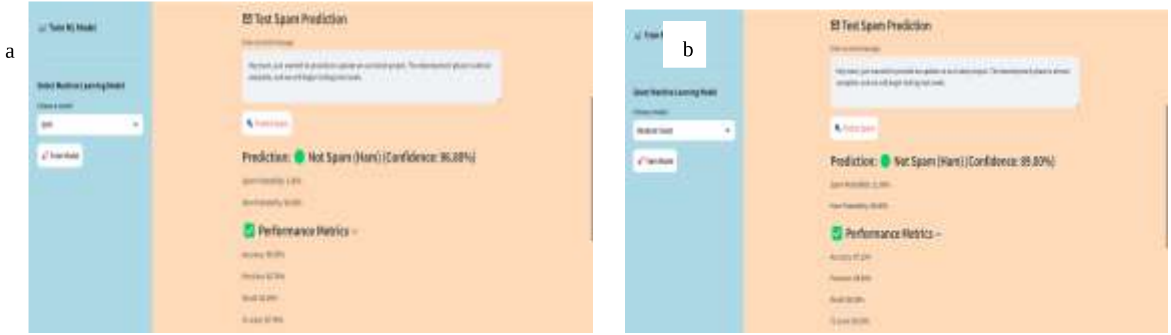


Fig. 4 – (a) Dataset Preview In Streamlit ; (b) Model Selection, Training, and Text Input



Prediction: ● Not Spam (Ham) | Confidence: 96.88%

Prediction: ● Not Spam (Ham) | Confidence: 89.80%

Fig. 5 – (a), (b) Showcasing model predictions and corresponding evaluation measures: SVM vs Random Forest.

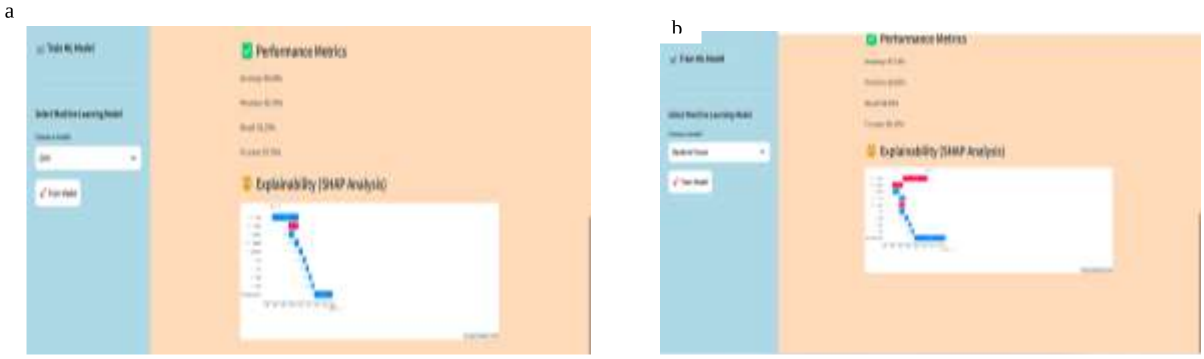


Fig. 6 – (a), (b) Model Explainability through SHAP Analysis and Word Cloud Visualization of Email Data (SVM vs Random Forest)

5. Discussion

The results of the experimentation show that the proposed system gains a high accuracy applying SVM (96%). In all parameters, the method is much better than Random Forest. The balance of precision at 95.4% and recall at 96.7% confirms robustness, ensuring there are minimal false positives and false negatives. One of the major contributions of this work is explainable AI (SHAP). Usually, these classifiers of spam are black boxes, but the present one brings into light important keywords that affect the classification and improves the trustworthiness and transparency of the system for real-world deployment. The Streamlit-based GUI makes this possible by providing real-time predictions with explanations, thus linking academic research and its application. Compared with deep learning such as BERT or CNNs, this framework is lightweight, requiring fewer resources and working well within standard systems, proving to be practical for small organizations and individual users. However, there are some limitations. The model currently uses purely text-based features as modern spam emails can include images, attachments, or embedded links. Work for the future may entail possible multimodal features (text+metadata+image analysis) and explore ensemble deep learning with explainability on these models (e.g., LIME, SHAP for DL models) for improving their robustness.

Conclusion

This paper offers an explainable spam detection framework based on NLP, which encompasses feature extraction using TF-IDF in addition to class balancing through SMOTE and machine learning classification using SVM and Random Forest. In the experimental analysis, SVM was more superior to Random Forest with an accuracy of 96%, with very close numbers in precision and recall values. With SHAP explainability being used here, significant words with respect to interpretability are highlighted by the system, which remains a concern throughout all existing spam models. Also, regarding prospective real-time predictions for everyday users, the deployment of a Streamlit interface further ensured that such predictions are made simply and feasible.

Summarily, this considered a lightweight but practically useful and explainable spam detection system. Future works may include the integration of deep learning, the incorporation of multimodal spam detection, and the facilitation of large-scale deployment in enterprise settings.

REFERENCES

[1] Pachare, R., Banarase, P., Dhanke, P., and Dhakade. A. K. "Progress in Email Spam Detection: Systematic Review Portal of Techniques Through Machine Learning and Deep Learning." International Journal of Research in Applied Science & Engineering Technology (IJRASET) 13 (2023): 112-121.

[2] An P.; Shafi R.; Mughogho T.; Onyango O.A., "Detection of Multilingual E-mail Phishing Attacks Using OSINT and Machine Learning", preprint arXiv:2501.08723, 2025.

[3] Low C.K.; Ying T.X., "Artificial Intelligence-Based Email Spam Filtering", ResearchGate Preprints, 2024.

[4] Alsuwit M.H., Haq M.A., and Aleisa M.A., "New Advances in Email Spam Classification with Machine Learning and Deep Learning Techniques", Engineering Technology & Applied Science Research (ETASR) pp. 14(4): 987 to 994, 2024.

[5] Jamal S., and H. Wimmer. "An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach." arXiv preprint arXiv:2311.04913, 2023.

[6] Adnan, M.; Imam, M.O.; Javed, M.F.; Murtza, I. Improving Spam Email Classification Accuracy Using Ensemble Techniques: A Stacking Approach. Journal of Information Security and Applications, 75(2023), 102-115.

[7] F. Janez-Martino, R. Alaiz-Rodrigues, V. Gonzalez-Castro, E. Fidalgo, E. Alegre, "Classifying Spam Emails using Agglomerative Hierarchical Clustering and a Topic-Based Approach", preprint arXiv:2402.05296, 2023.