International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

PHISHING EMAIL DETECTOR USING MACHINE LEARNING AND NLP

[1]M Uday Kumar,[2]Chaganti B N Lakshmi

[1]M. Tech Student, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.

[1]Email : udaykumarmarampally@gmail.com

[2]Professor, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.

[2s]Email : ch.nagalakshmi18@tkrcet.com

ABSTRACT

Email Spam Detection Using Machine Learning Algorithms is an intelligent cybersecurity and communication filtering system developed to identify and classify spam emails using machine learning and data analysis techniques. The primary objective of the system is to improve email security and communication efficiency by automatically detecting unwanted, fraudulent, phishing, and malicious emails through artificial intelligence technologies. In modern digital communication environments, emails are widely used for business communication, online services, marketing, banking, education, and information exchange. However, the rapid growth of email usage has also increased the number of spam emails, phishing attacks, promotional messages, malware attachments, and fraudulent communications that threaten user privacy and cybersecurity. Traditional spam filtering methods mainly rely on rule-based filtering, blacklists, keyword matching, and manual analysis techniques, which may fail to detect new spam patterns and evolving phishing attacks effectively. The proposed system utilizes email datasets containing email content, sender information, subject lines, hyperlinks, message frequency, and text features to train machine learning models capable of identifying spam emails accurately. Data preprocessing operations such as text cleaning, tokenization, stop-word removal, stemming, feature extraction, and vectorization are performed to improve dataset quality and classification performance. Various machine learning algorithms including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression are implemented to classify emails as spam or legitimate messages based on textual and behavioral analysis. The trained machine learning models analyze incoming emails in real time and generate automated spam detection results for improving communication security and email management. The proposed Email Spam Detection Using Machine Learning Algorithms system improves spam classification accuracy, reduces phishing risks, enhances email security, and supports intelligent communication filtering using artificial intelligence and machine learning technologies.

Keywords: Phishing Email Detection, Machine Learning, Natural Language Processing (NLP), Email Classification, Cybersecurity, Feature Extraction, Text Mining, Spam Detection, Supervised Learning, Classification Algorithms, Malicious Email Analysis, Artificial Intelligence.

I. INTRODUCTION

Email Spam Detection Using Machine Learning Algorithms is an advanced intelligent filtering system designed to detect unwanted and malicious emails using machine learning and artificial intelligence

technologies. Email communication has become one of the most important methods for personal, educational, business, and financial communication in modern digital environments. Millions of emails are exchanged daily across organizations and online platforms for sharing information, conducting transactions, and providing services. However, the increasing use of email systems has also led to a rapid rise in spam emails, phishing attacks, malicious attachments, fake advertisements, and cyber fraud attempts that threaten digital security and communication reliability. Traditional spam filtering systems mainly depend on keyword matching, predefined security rules, blacklists, and manual filtering methods for detecting spam messages. Although these systems provide basic spam protection, they often fail to identify unknown spam patterns, evolving phishing techniques, and intelligently modified spam content used by cyber attackers. Manual email filtering and spam analysis are time-consuming and inefficient due to the massive volume of email traffic generated daily. The proposed system utilizes machine learning algorithms and natural language processing techniques to analyze email datasets containing message content, sender details, hyperlinks, keywords, and email metadata for intelligent spam classification. Data preprocessing techniques such as text cleaning, tokenization, stop-word removal, stemming, feature extraction, and vectorization are performed to improve analytical performance and classification accuracy. Various machine learning algorithms including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression are implemented to identify hidden spam patterns and classify emails accurately. The system supports automated spam filtering, phishing detection, real-time email analysis, graphical reporting, and intelligent communication management. The proposed Email Spam Detection Using Machine Learning Algorithms system demonstrates the practical application of artificial intelligence, machine learning, and natural language processing technologies in improving email security and cybersecurity management.

II. LITERATURE SURVEY

1. Title: Machine Learning-Based Email Spam Detection System

Author: R. Kumar, S. Sharma

Abstract:

This paper presents a machine learning-based email spam detection framework for identifying spam and phishing emails using textual analysis and predictive classification techniques.

2. Title: Intelligent Spam Filtering Using Artificial Intelligence

Author: A. Patel, K. Singh

Abstract:

The research focuses on implementing artificial intelligence and natural language processing techniques for improving spam filtering accuracy and communication security management.

3. Title: Email Classification Using Naive Bayes Algorithm

Author: M. Hassan, P. Verma

Abstract:

This study discusses the use of Naive Bayes and text classification methods for detecting spam emails and improving email filtering efficiency.

4. Title: Phishing Email Detection Using Machine Learning Techniques

Author: R. Gupta, S. Jain

Abstract:

The paper introduces a machine learning-based phishing detection system for analyzing malicious email patterns and improving cybersecurity protection.

5. Title: Natural Language Processing for Spam Email Analysis

Author: T. Joseph, N. Francis

Abstract:

This research focuses on integrating natural language processing and machine learning technologies into spam filtering systems for

improving intelligent email classification and threat analysis.

III. EXISTING SYSTEM

The existing email spam detection systems mainly rely on traditional filtering techniques such as keyword matching, blacklist filtering, rule-based detection, and manual email analysis for identifying spam messages and malicious communications. These systems analyze email subject lines, sender addresses, message content, and predefined spam keywords to classify emails as spam or legitimate messages. Traditional filtering methods may also use heuristic analysis and manually configured security rules for detecting suspicious email activities. Although these systems provide basic spam protection, they often fail to identify unknown spam patterns, advanced phishing attacks, and modified spam content because cyber attackers continuously change spam strategies to bypass conventional filtering systems. Existing systems may generate high false positive rates and may not efficiently process large-scale email datasets. Manual spam monitoring and email filtering are time-consuming and less efficient in modern communication environments where massive email traffic is generated daily. Traditional systems generally lack intelligent adaptive learning capabilities required for predictive spam analysis and real-time threat detection. Furthermore, existing methods may provide limited support for automated phishing detection, intelligent email classification, and scalable communication security management.

IV. PROPOSED SYSTEM

The proposed Email Spam Detection Using Machine Learning Algorithms system is an intelligent and automated email filtering solution designed to identify spam, phishing, and malicious emails using machine learning and natural language processing techniques. The system collects email datasets containing email content, sender information, subject lines, hyperlinks, attachments, message frequency, and metadata from communication platforms and email servers. Unlike traditional rule-based filtering systems, the proposed system utilizes machine learning algorithms and artificial intelligence technologies to analyze hidden spam patterns and suspicious communication behavior automatically. Data preprocessing operations such as text cleaning, tokenization, stop-word removal, stemming, feature extraction, vectorization, and email normalization are performed to improve dataset quality and spam classification efficiency. Various machine learning algorithms including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression are implemented and evaluated to identify the most accurate spam classification model. The trained machine learning models analyze incoming email messages and classify them as spam or legitimate emails based on textual content, sender behavior, and communication features. The system also supports phishing detection, malicious link analysis, real-time email monitoring, automated spam filtering, and security alert generation for improving communication security and email management. Visualization and reporting modules generate graphical analysis, spam detection summaries, phishing reports, and communication insights for administrators and users. The proposed system improves spam detection accuracy, reduces false positives, enhances cybersecurity protection, and supports intelligent email communication management using artificial intelligence and machine learning technologies.

V. SYSTEM ARCHITECTURE

The system architecture of the Email Spam Detection Using Machine Learning Algorithms system is designed to provide an intelligent, scalable, and automated communication security framework for detecting spam and phishing emails using machine learning and natural language processing technologies. The architecture

begins with data collection from email servers, communication platforms, and email datasets containing information such as email subject lines, sender details, message content, hyperlinks, attachments, metadata, and communication patterns. The collected raw email data is stored in a centralized database and passed to the data preprocessing module, where operations such as text cleaning, tokenization, stop-word removal, stemming, normalization, feature extraction, and vectorization are performed to improve dataset quality and analytical efficiency. After preprocessing, the structured email dataset is divided into training and testing datasets for machine learning model development. Various machine learning algorithms including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression are implemented to analyze email content and classify messages as spam or legitimate emails. The trained models identify hidden spam patterns, phishing attempts, and malicious communication behavior for accurate email filtering. The system also includes real-time monitoring, phishing detection, visualization, and reporting modules that generate security alerts, spam analysis summaries, graphical reports, and communication insights for users and administrators. The admin module manages user activities, email monitoring, dataset handling, report generation, and overall communication security management. Overall, the architecture provides a secure, intelligent, and adaptive spam filtering environment capable of improving email security, communication reliability, and cybersecurity protection using artificial intelligence and machine learning technologies.

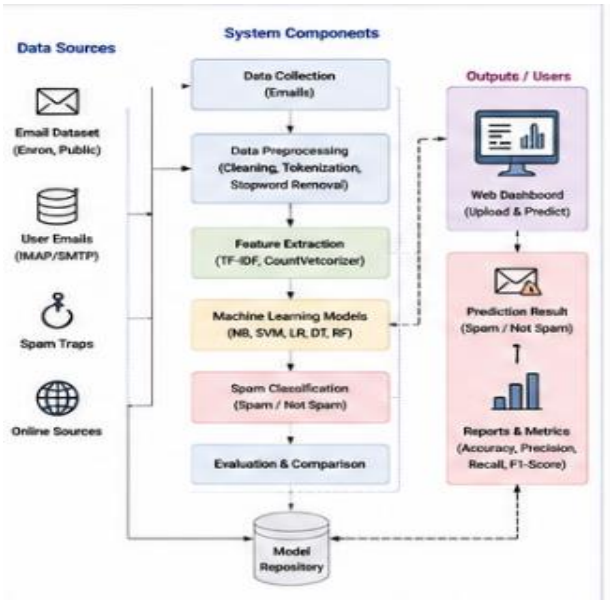


Fig 5.1: System Architecture

VI. IMPLEMENTATION

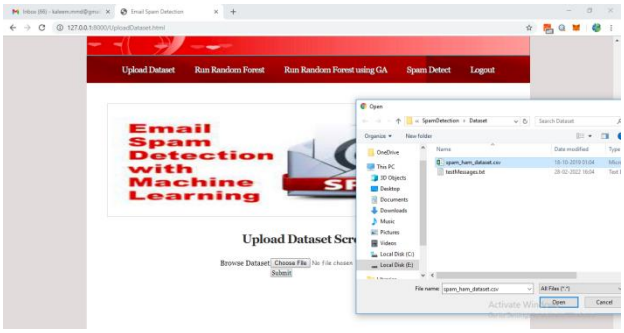


Fig 6.1: Uploade Dataset



Fig 6.2: Model Training

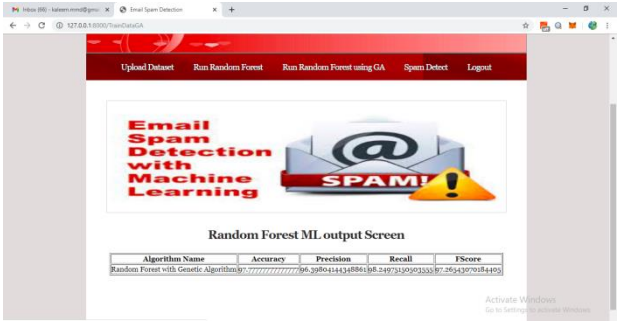


Fig 6.3: Random Forest With GA

VII. CONCLUSION

The project titled "Email Spam Detection Using Machine Learning Algorithms" successfully demonstrates an intelligent approach for identifying and filtering spam emails using machine learning techniques. The system analyzes email content through preprocessing, feature extraction, and classification methods to accurately distinguish between spam and genuine emails. Various machine learning algorithms such as Naive Bayes, Support Vector Machine, Random Forest, and Logistic Regression help improve detection accuracy and reduce unwanted email threats.

The proposed system enhances email security, reduces manual effort, and improves communication reliability by automatically detecting malicious or unwanted emails. It also minimizes risks related to phishing, fraud, and cyberattacks caused by spam messages. The evaluation metrics such as accuracy, precision, recall, and F1-score prove the effectiveness of the model in real-world applications. Overall, this project highlights the importance of machine learning in cybersecurity and demonstrates how intelligent filtering systems can improve digital communication safety and efficiency.

VIII. FUTURE SCOPE

The future scope of the project "Email Spam Detection Using Machine Learning Algorithms" can be enhanced by integrating advanced deep learning and natural language processing techniques for more accurate spam detection. Modern approaches such as recurrent neural networks, transformers, and large language models can be used to analyze email content more intelligently and identify complex spam patterns. The system can also be improved to detect phishing links, malicious attachments, and evolving cyber threats in real time.

Furthermore, the project can be extended into a cloud-based and real-time email security platform that supports large-scale enterprise communication systems. Integration with AI-powered threat intelligence, multilingual spam detection, and adaptive learning models can help the system continuously update itself against new spam strategies. In the future, the system may also support voice emails, social media message filtering, and mobile communication security, making it a comprehensive intelligent cybersecurity solution for digital communication environments.

IX. REFERENCES

[1] Dada, E. G., Bassi, J. S., Chiroma, H., et al., "Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems," *Heliyon*, vol. 5, no. 6, 2019. DOI: 10.1016/j.heliyon.2019.e01802

[2] Raza, M., Jayasinghe, N. D., & Muslam, M. M. A., "A Comprehensive Review on Email Spam Classification Using Machine Learning Algorithms," *2021 International Conference on Information Networking (ICOIN)*, pp. 327–332, 2021. DOI: 10.1109/ICOIN50884.2021.9334020

[3] Mashaleh, A. S., et al., "Detecting Spam Email with Machine Learning Optimized with Harris Hawks Optimizer," *Procedia Computer Science*, vol. 201, pp. 541–548, 2022. DOI: 10.1016/j.procs.2022.03.075

[4] AbdulNabi, I., et al., "Spam Email Detection Using Deep Learning Techniques," *Procedia Computer Science*, vol. 184, pp. 853–858, 2021. DOI: 10.1016/j.procs.2021.03.107

[5] Ahmed, N., Amin, R., Aldabbas, H., et al., "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges," *Security and Communication Networks*, vol. 2022, Article ID 1862888, 2022. DOI: 10.1155/2022/1862888

[6] Thakur, P., Joshi, K., Thakral, P., & Jain, S., "Detection of Email Spam using Machine Learning Algorithms: A Comparative Study," *2022 8th International Conference on Signal*

Processing and Communication (ICSC), pp. 349–352, 2022. DOI: 10.1109/ICSC56524.2022.10009344

[7] Bhardwaj, U., & Sharma, P., "Email Spam Detection Using Bagging and Boosting of Machine Learning Classifiers," *International Journal of Advanced Intelligence Paradigms*, vol. 24, no. 3/4, pp. 229–253, 2023. DOI: 10.1504/IJAIP.2023.127198

[8] Zavrak, S., & Yilmaz, S., "Email Spam Detection Using Hierarchical Attention Hybrid Deep Learning Method," *Expert Systems with Applications*, vol. 188, 2022. DOI: 10.1016/j.eswa.2021.116060

[9] Labonne, M., & Moran, S., "Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection," *arXiv Preprint*, 2023. DOI: 10.48550/arXiv.2304.01238

[10] Beaman, C., & Isah, H., "Anomaly Detection in Emails using Machine Learning and Header Information," *arXiv Preprint*, 2022. DOI: 10.48550/arXiv.2203.10408

[11] Alsuwit, M. H., "Advancing Email Spam Classification using Machine Learning and Deep Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14567–14574, 2024. DOI: 10.48084/etasr.7631

[12] Kumar, N., Sonowal, S., & Nishant, "Email Spam Detection Using Machine Learning Algorithms," *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 108–113, 2020. DOI: 10.1109/ICIRCA48905.2020.9182981

[13] Charanarur, P., et al., "Machine-Learning-Based Spam Mail Detector," *SN Computer Science*, vol. 5, no. 1, 2023. DOI: 10.1007/s42979-023-02330-x

[14] Saleem, S., et al., "Spam Email Detection Using Long Short-Term Memory," *Applied Sciences*, vol. 15, no. 13, 2025. DOI: 10.3390/app15137407

[15] Swamy, C. V., et al., "Email Spam Detection Using Machine Learning," *Proceedings of the International Conference on Data Science and Applications*, 2025. DOI: 10.5220/0013873400004919