Evaluation And Comparison of Machine Learning Models for Ham and Spam Email Classification

Sravya Gaddamanugu

M.Sc. Forensic Science

JAIN Deemed-to-be University

Email ID: sravya.gaddamanugu1176@gmail.com

Guide Name: Dr. Preeti Gupta

Department of Computer Science & Engineering

JAIN Deemed-to-be University

Email ID: preeti.alha@gmail.com

ABSTRACT:

Email is one of the most widely used ways of digital communication nowadays, but with that, junk emails have also become more prevalent. These spam messages are not only annoying they can also be harmful, causing security issues such as phishing or data theft. The goal of this research focuses on detecting and filtering spam emails by using machine learning techniques and algorithms. A dataset of 15,267 emails containing spam and ham. Prior to the model training, the dataset was pre-processed by cleaning the textual words converting it into numerical feature vectors using the TF-IDF technique, enabling the algorithms to effectively interpret and analyse the given data. Then applied five well known machine learning algorithms: logistic, svm, naive bayes, decision tree, and random forest. These models were developed and evaluated using tools such as python, the scikit-learn library and Jupiter notebook. To evaluate the effectiveness of each model, performance metrics such as accuracy, precision, recall, F1-score, ROC curve and confusion matrix were employed. Among the models tested, SVM achieved the highest accuracy of closely followed by Random Forest. The results obtained indicates that machine learning models, specifically SVM performs very accurate at detecting spam and improving the security of email communication systems.

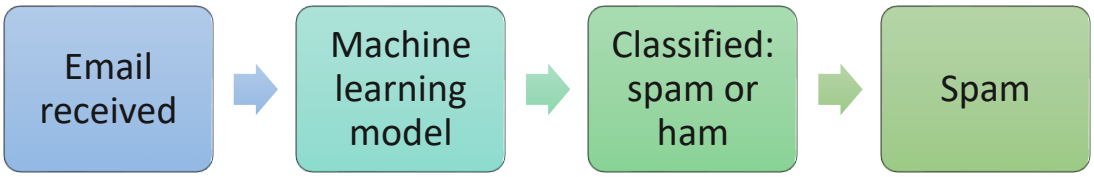
KEYWORDS: Machine learning, Support vector machine, Spam, Accuracy, Ham, Emails

I. INTRODUCTION:

In the modern digital era, email continues to serve as a fundamental tool for career related and everyday communication. However, the growing volume of email correspondence has introduced a persistent issue: spam emails. These unsolicited messages, often disguised as legitimate content, pose serious threats ranging from phishing attacks to malware infiltration.

Spam has evolved from a simple nuisance to a sophisticated tool used by cybercriminals to exploit users through deceptive schemes. As spammers continue to refine their tactics, effective spam detection has become a critical area of research. To address this, machine learning (ML) techniques have gained significant traction in identifying spam by analysing email content, metadata, and linguistic patterns. While deep learning models are being explored for their ability to capture complex patterns, traditional ML algorithms remain valuable due to their simplicity, interpretability, and effectiveness on structured datasets.

This study presents a comprehensive evaluation and comparison of five well-established machine learning models: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest. Using two publicly available datasets—the UCI Spam and ham base dataset and a raw email dataset from Kaggle this research examines the ability of each model to detect spam effectively. The datasets consist of 15,267 emails, with 2 columns and labelled as either ham and spam, providing a balanced foundation for experimentation. Here is the classification:



The primary aim of this research is to compare the accuracy of these models and assess their precision, recall, F1-score, and performance using confusion matrices and ROC curves.

By doing so, we aim to identify the most efficient algorithm for practical spam detection scenarios. The structure of this paper is organized as follows: Section II defines the problem statement, outlining the challenges and limitations currently faced in email spam detection. Section III identifies the research gap by reviewing existing solutions and highlighting areas where improvements are needed. Section IV states the objectives of the study, detailing both the general and specific goals. Section V provides a comprehensive review of the existing literature on spam classification, summarizing key methodologies and findings. Section VI explains the methodology, including a description of the datasets, the preprocessing techniques applied, and the experimental setup. Section VII presents the results of the comparative analysis of different classification models, along with a detailed discussion of their performance. Finally, Section VIII concludes the paper by summarizing the key findings and offering recommendations for future research in the field of email spam detection.

II. PROBLEM STATEMENT:

Despite the availability of numerous spam filtering tools, many existing systems are limited in their ability to detect novel and sophisticated spam techniques. Traditional filters often fail to recognize subtle variations in spam content, leading to misclassification and reduced reliability (E. Prasannakumar, 2016). Furthermore, certain machine learning models used in prior studies either lack generalizability or are not adequately evaluated across diverse datasets. This raises concerns about their applicability (Geetha Gowri, 2022) in real word scenarios where spam messages vary widely in structure, language and tactics.

III. RESEARCH GAP:

A review of existing literature reveals that many prior studies have used relatively small and less datasets for spam email classification, which restricts the robustness and generalization of their findings (Md. Faisal, 2018). Additionally, some researchers have employed deep learning models that, while powerful may not always be practically due to their high and computational requirements, need for large datasets and longer training times (Isra'a Abdul Nabi, 2021) (Preeti Durgapal, 2021). Moreover, limited attention has been given to the comparative analysis of traditional machine learning models using well-preprocesses and sufficiently large datasets (Md. Faisal, 2018). There is a lack of research focusing on optimizing classical algorithms using the effective text preprocessing and feature extraction methods to enhance classification performance while maintaining (U. Saranya, 2021) computational efficiency. This study addresses these gaps by evaluating multiple traditional machine learning algorithms on a large and more diverse email dataset. It emphasizes the importance of systematic data preprocessing and including cleaning, tokenization, lemmatization and vectorization. To improve the accuracy, precision, and efficiency of spam detection by adopting this approach, the study aims to provide a practical and scalable solution for real-world spam classification tasks.

IV. OBJECTIVES OF THE STUDY:

The primary aim of this research is to enhance the accuracy and reliability of spam email classification by systematically analysing the performance of multiple traditional machine learning algorithms. To achieve this, the study is guided by the following specific features:

1. To assess and compare the performance of five widely-used machine learning models namely Naïve Bayes, Logistic Regression, Decision Tree, Random Tree, Support Vector Machine (SVM) in classifying spam and ham (non-spam) emails. This analysis helps to identify which model is most effective in distinguishing spam messages in diverse and realistic dataset.

2. To implement a series of robust text preprocessing techniques such as text normalization, removal of stop words, tokenization, lemmatization and feature extraction using methods like Term Frequency- Inverse Document Frequency (TF-IDF). The objective is to analyse how each preprocessing step contributes to improving the overall model performance and classification accuracy.

3. To identify an ideal machine learning model that strikes an ideal equilibrium between accessibility, computational speed, and precision in classification. This is particularly relevant for real world applications where resources could be limited and open decision making is required, such email security and digital forensics.

4. To provide practical insights and recommendations for the development of more reliable email filtering tools and forensic analysis systems. By leveraging the findings from this study, the goal is to contribute to the design of advanced, data driven spam detection solutions that are both scalable and adaptable to evolving spam strategies.

V. REVIEW OF LITERATURE:

Spam email detection has been an ongoing challenge in the field of cybersecurity, with researchers consistently exploring various machine learning (ML) and deep learning (DL) models to improve the accuracy of spam filters. Several studies have contributed to this area by evaluating the performance of different algorithms and feature extraction techniques. A foundational understanding of (Geetha Gowri, 2022) ml principle includes supervised learning approaches, model section and algorithms is crucial and important in developing effective of spam filters.

In the study "Email Based Spam Detection" by (Md. Faisal, 2018) and colleagues, the Random Forest algorithm was found to perform best, achieving an accuracy of 96.25%. The authors highlighted the advantage of ensemble learning techniques, which combine multiple weak learners to form a strong classifier, improving overall performance. However, one limitation of this study was the relatively small dataset used, which may not fully capture the diversity of real-world spam emails.

(E. Prasannakumar, 2016), in their paper "Email Spam Detection Using Machine Learning Algorithms", explored various machine learning models, including Naïve Bayes and SVM, and emphasized the importance of TF-IDF for feature extraction. They found that Naïve Bayes and SVM showed promising results but noted that their models were limited by the quality of the feature extraction process. The study did not delve into the impact of more advanced preprocessing techniques, which could potentially improve classification accuracy.

Another notable study, "Spam Review Detection Using Machine Learning" by (Preeti Durgapal, 2021), extended machine learning techniques to identify spam product reviews. In this research, Random Forest was identified as the top-performing model due to its robustness against overfitting. However, the dataset used in this study was centred on product reviews, which may not directly translate to the domain of email spam detection, indicating a potential gap in research when applying these techniques to email datasets.

A more recent exploration of deep learning for spam detection was presented by (Isra'a Abdul Nabi, 2021) (U. Saranya, 2021) in their paper "Email Spam Detection Using Deep Learning". This study compared various deep learning models, including BERT and BiLSTM, and found that BERT outperformed the other models in terms of accuracy, achieving an impressive 98.67%. While deep learning showed great promise, the authors acknowledged the high computational cost associated with these models, which may limit their applicability in real-time systems or for practical use in email security tools. In this (E. Prasannakumar, 2016) evaluated a range of traditional machine learning models such as Naïve Bayes, Decision Tree, and Logistic Regression. They found that Naïve Bayes was computationally efficient, making it suitable for real-time applications. (U. Saranya, 2021). However, the model struggled with complex or multilingual spam emails, a limitation that could impact its real-world usability.

Recent research in spam email detection has extensively focused on the application of machine learning (ML) techniques to improve the accuracy and efficiency of spam filters. (Suryawanshi, Goswami, & Patil, 2019) conducted a comparative study analysing the performance of various ML and ensemble classifiers. Their findings emphasized that ensemble methods, such as Random Forests and boosting algorithms, tend to outperform single classifiers in detecting spam emails. Similarly, (Karim, Azam, Shanmugam, Krishnan, & Alazab, 2019) provided a comprehensive survey on intelligent spam detection methods, highlighting the strengths and challenges of ML, hybrid, and deep learning techniques, and suggesting the need for models that can adapt to evolving spam strategies. (Agarwal & Kumar, 2018) proposed an integrated approach using Naïve Bayes combined with Particle Swarm Optimization (PSO) for feature selection, resulting in improved classification performance compared to standalone methods. (Harisinghaney, Dixit, Gupta, & Arora, 2014) explored both text and image-based spam detection, using algorithms like KNN and Naïve Bayes, and pointed out the additional complexities introduced by multimedia spam. (Mohamad & Selamat, 2015) focused on optimizing the feature selection process through

hybrid techniques, showing that carefully selected features significantly enhance ML model accuracy. compared multiple machine learning methods, concluding that Naïve Bayes offered the best trade-off between speed and accuracy for spam detection tasks. Additionally, (Ameen & Kaya, 2018) applied deep learning models to spam detection in online social networks, demonstrating that although deep learning models such as CNNs can provide higher accuracy, they often require greater computational resources, limiting their practicality in real-time applications. Overall, these studies collectively highlight the importance of selecting appropriate algorithms, optimizing feature sets, and balancing performance with computational efficiency to develop robust spam email detection systems.

Identified Research Gap

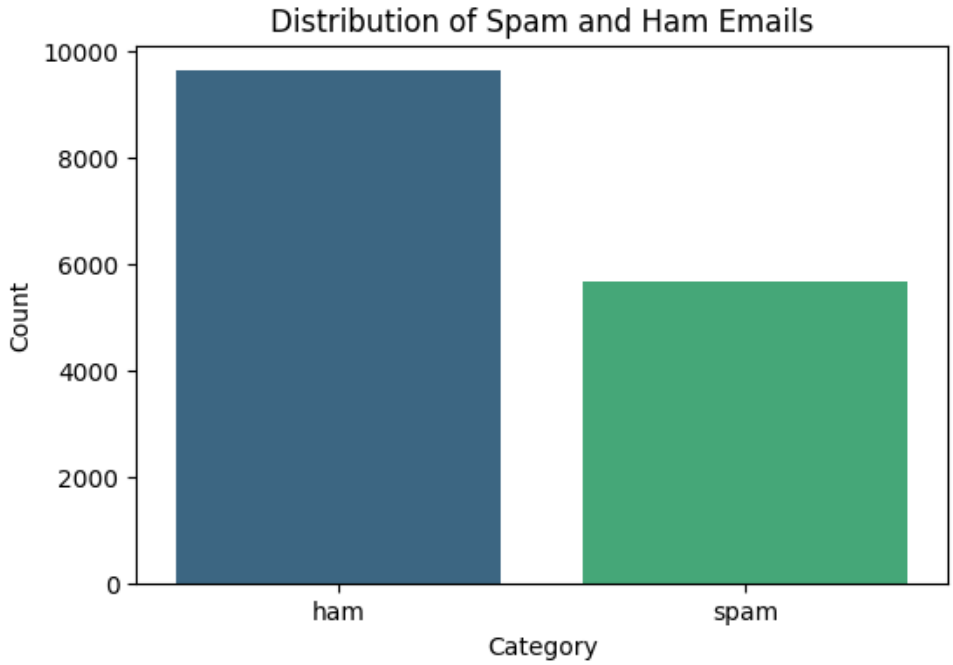
While numerous studies have advanced spam email detection, several limitations remain, particularly in the context of real-world applications. Many studies utilized small or non-email datasets, limiting the generalizability of their findings to practical email classification scenarios. Additionally, some research focused on deep learning models, which, although effective, often require substantial computational resources, making them impractical for integration into forensic systems or lightweight applications. Furthermore, many studies did not emphasize comprehensive preprocessing steps, which are critical for improving model accuracy and reliability. Additionally, there is a lack of focus on model interpretability and scalability, which are essential for deploying models in real-world, dynamic environments like digital forensics.

Contribution of the Present Study

To address these gaps, the present study proposes a fresh perspective by applying multiple traditional machine learning models Logistic Regression, SVM, Naïve Bayes, Decision Tree, and Random Forest on a combined and large-scale email dataset sourced from UCI and Kaggle. Advanced preprocessing techniques like cleaning, tokenization, lemmatization, and feature extraction methods (TF-IDF) are employed to optimize model performance. This study not only evaluates model accuracy but also focuses on achieving a balance between efficiency, accuracy, f1-score and interpretability, making it more relevant for digital forensic tools and real-world spam monitoring systems.

VI. METHODOLOGY:

Dataset Overview: This research draws upon two reputable and publicly accessible email datasets: one from Kaggle, containing raw email data, and the other from the UCI Machine Learning Repository. After merging both sources, the final dataset consisted of 15,267 email records. Among them, 9,616 were categorized as ham (legitimate messages), while 5,651 were marked as spam (unsolicited or promotional content).



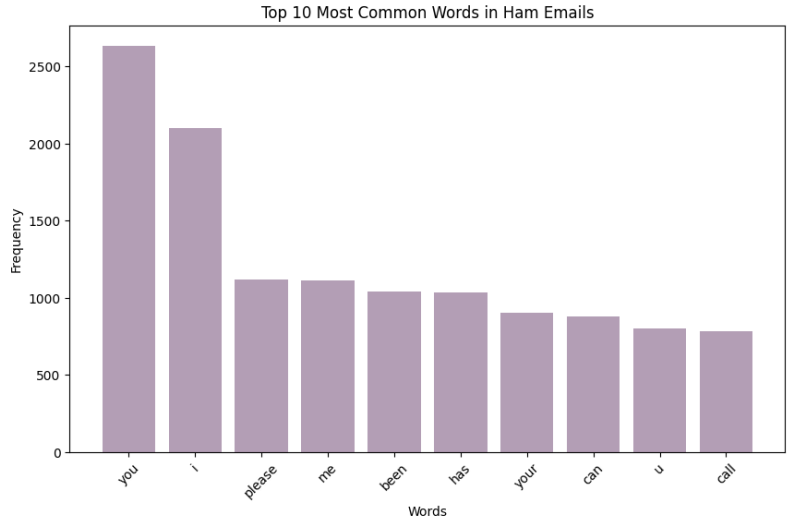
To maintain the natural distribution of spam and ham messages, stratified sampling was employed. This ensured a balanced representation of both classes during model training and testing. The entire dataset was partitioned into two sets: 80% of the data (12,213 samples) was assigned for training purposes, while the remaining 20% (3,054 samples) was held back for testing.

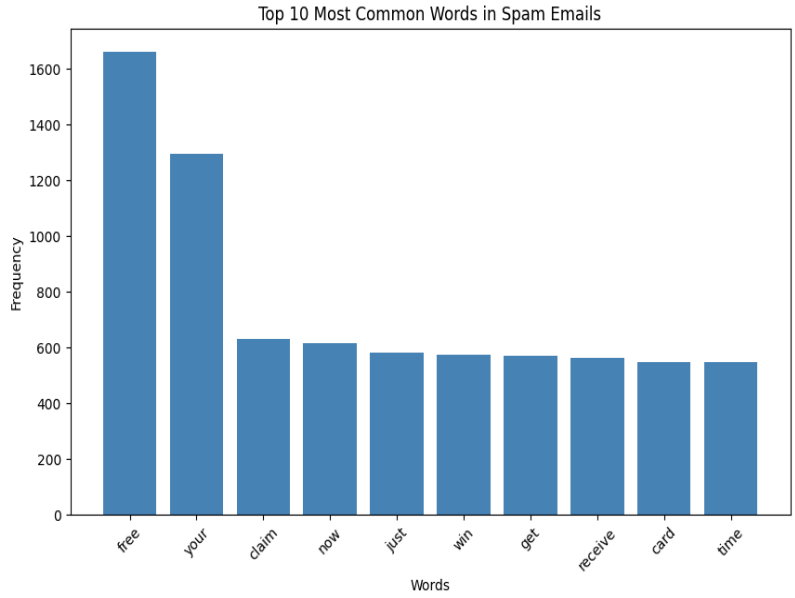
Data Preparation Steps:

Prior to model training, the raw email content underwent several transformation steps to make it suitable for analysis:

- Text Normalization: Each email was cleaned by removing symbols, numbers, HTML tags, and punctuation. The text was then converted to lowercase to eliminate case sensitivity issues.
- Tokenization: The cleaned text was broken into smaller units called tokens (individual words), which helped in analysing text at the word level.
- Filtering Stop Words: Common English words such as “is,” “the,” and “and,” which often carry limited significance in classification, were removed to reduce data noise and improve model focus.
- Lemmatization: Words were reduced to their root forms so that variations (e.g., *"Walking"*, *"walks"*) would be treated uniformly (as *"walk"*). This step helped in standardizing the vocabulary used by the model.
- Word Frequency analysis: After the text was cleaned and standardized, the most frequently used words in both spam and ham emails were analysed. This step helped identify key terms common in spam messages like “free,” “win,” and “claim compared to ham emails, which included more personal or polite language such

as “you,” “please,” and “call.” These insights provided a clearer understanding of the linguistic differences between the two categories and supported feature selection for model training.





Feature Representation:

To convert textual data into numerical inputs for the machine learning models, the TF-IDF (Term Frequency–Inverse Document Frequency) technique was used. This method assigns weight to each word based on its frequency within a specific email and how rare it is across the entire dataset. As a result, commonly used but less informative words are down-weighted, while more significant, email-specific terms carry higher values.

Model Implementation:

A comparative evaluation of five different machine learning algorithms was carried out to determine which model best classifies emails as spam or ham:

1. Logistic Regression: A linear classifier that calculates the probability of an email being spam by modelling the relationship between the features and output labels.
2. Support Vector Machine (SVM): An advanced model that identifies the optimal boundary in the feature space to separate spam from non-spam emails with maximum margin.
3. Naive Bayes: A classification technique based on probabilistic reasoning, which assumes independence between words and applies Bayes' theorem for predictions.
4. Decision Tree: A rule-based model that creates a tree-like structure where decisions are made by splitting the dataset according to feature thresholds.
5. Random Forest: A robust ensemble method that builds multiple decision trees and aggregates their predictions, thus enhancing accuracy and reducing the risk of overfitting.

Performance Assessment:

Each model's performance was examined using the following metrics:

Accuracy: shows how many emails the model got right overall; out of all the emails it checked.

Precision: tells us how many of the emails marked as spam were actually spam, helping avoid mistakes were normal emails get flagged.

Recall: looks at how well the model finds all the real spam emails, making sure it doesn't miss any

F1 Score: is a mix of precision and recall, and its especially helpful when there are more of one type of email than the other

Confusion Matrix: This visual summary shows how many emails were correctly and incorrectly predicted as spam or ham, helping to analyse specific error types.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): This metric evaluates the model's discrimination capability by plotting the true positive rate against the false positive rate. A higher AUC implies stronger model performance in distinguishing between classes.

VII. RESULTS AND DISCUSSION

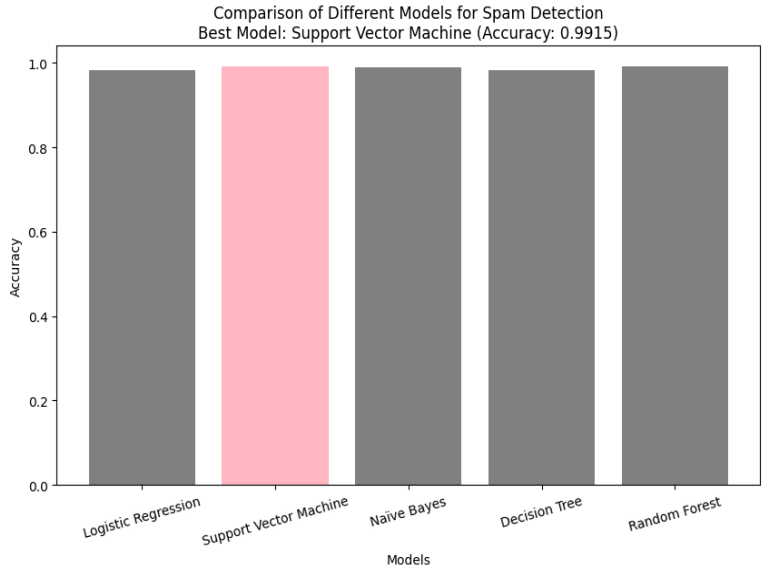
In the context of classifying emails, a set of machine learning techniques was applied and systematically analysed. The algorithms explored in this study included logistic regression, support vector machine, naive bayes, decision tree and random forests. Each model's effectiveness was measured using key evaluation metrics namely accuracy, precision, recall F1 score as well as a detailed examination of their roc curve and confusion matrix are shown below:

TABLE 1: comparison of models

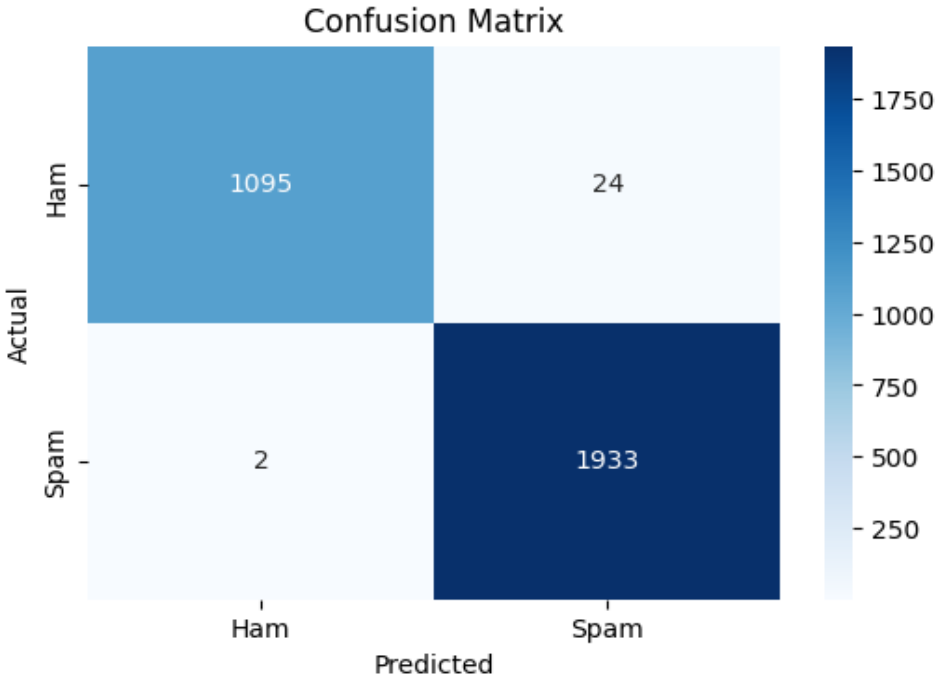
Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Logistic Regression	98.36	97.96	99.48	98.72
Support Vector machine	99.15	98.77	99.90	99.33
Nave Bayes	98.89	99.02	99.22	99.12
Decision Tree	98.20	98.21	98.97	98.58
Random Forest	99.08	98.57	100.0	99.28

Graphs:

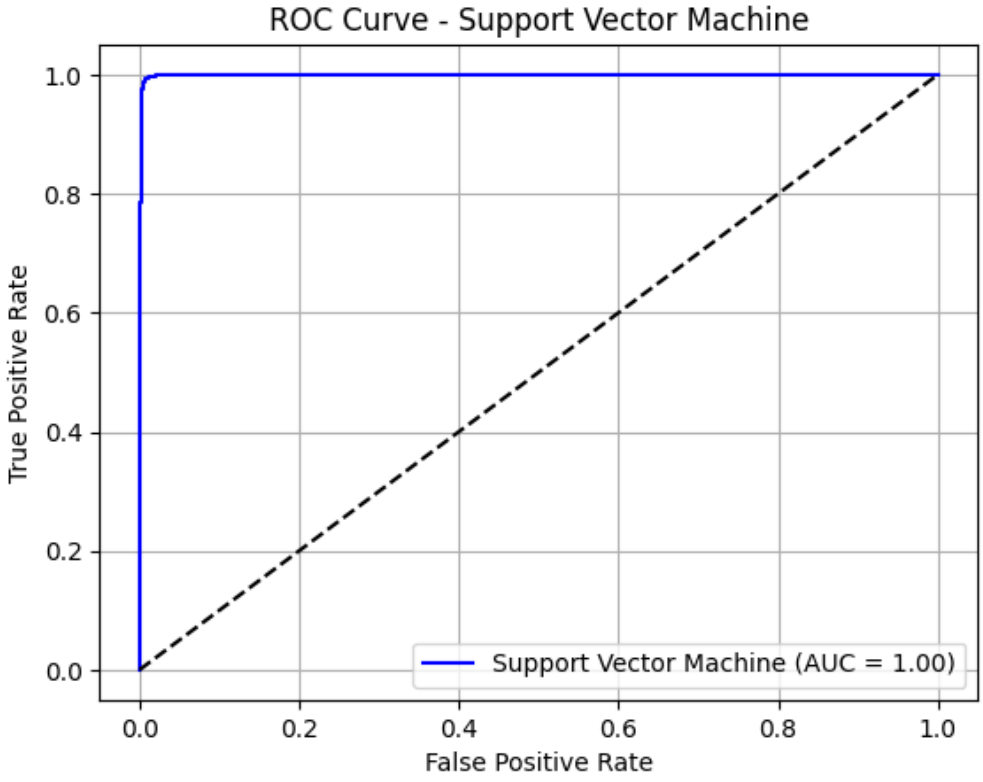
1. Bar Chart: Displaying the Accuracy of each model.



2. Confusion Matrix: Illustrated the true positives, true negatives, false positives, and false negatives for all model emails.



3. ROC Curve: For each model, showing the Receiver Operating Characteristic curve and Area Under the Curve (AUC) for performance evaluation.



Comparative Discussion of Models:

This analysis compared five machine learning models to see how well they can spot spam emails. The Support Vector Machine (SVM) stood out with the best performance, reaching 99.15% accuracy and a recall of 99.90%. This shows it was highly effective at identifying spam without missing many. The Random Forest model also did a great job, with 99.08% accuracy and a strong F1-score of 99.28%, meaning it managed a good mix of finding spam and avoiding errors. Logistic Regression and Naïve Bayes also scored high in accuracy 98.36% and 98.89% but had slightly lower recall rates, so they might overlook a few more spam messages. The Decision Tree model performed reasonably, with 98.20% accuracy, but it had lower precision and the weakest F1-score at 98.58%. Overall, SVM proved to be the most dependable model, especially when it's important not to miss any spam emails.

Justification of Best Model:

SVM stands out as the most effective model for identifying spam emails, as it performs best in terms of accuracy, precision, and recall. The high recall is especially crucial because it reduces the chances of missing spam emails, making it more reliable for filtering.

VIII. CONCLUSION AND FUTURE WORK:

This study evaluated five different machine learning algorithms for classifying spam emails. Among these, Support Vector Machine (SVM) achieved the highest accuracy and recall, making it the most effective in distinguishing between spam and legitimate messages. Its performance was particularly strong in reducing false negatives, a crucial factor in reliable spam detection. Random Forest also delivered solid results, especially in balancing precision and recall, but SVM maintained a slight edge in overall performance. The findings confirm that machine learning models offer significant improvements over traditional rule-based systems, especially when dealing with large and diverse email datasets. However, there is still room to enhance the system’s efficiency and usability. Future work could focus on incorporating additional features such as sender reputation, frequency of similar messages, or time-based patterns to improve the model’s decision-making ability. Another practical step would be developing a user-friendly web interface where users can quickly check if an email is spam by uploading its content. Embedding the model into live email systems for real-time filtering would also increase its practical value. Instead of exploring deep models like LSTM or BERT, future efforts could investigate lightweight models for faster processing, making the solution suitable for deployment on devices with limited computing power. Additionally, integrating explainable AI methods could help users and developers understand how classification decisions are made, increasing trust and transparency.

APPLICATIONS IN FORENSIC SCIENCE

The machine learning models developed for spam email classification hold valuable potential within the domain of forensic science, particularly in digital crime investigations. These models can be integrated into forensic software to streamline the identification and sorting of suspicious emails. By automatically flagging emails that exhibit characteristics of phishing, scams, or embedded malware, they enable investigators to focus more quickly on high-risk communications, reducing manual workload and enhancing the accuracy of digital evidence analysis. In large-scale investigations involving corporate breaches, identity theft, or financial fraud, these models can efficiently sift through vast email archives to isolate messages that may serve as critical evidence. Furthermore, their application in forensic email review systems allows for the separation of benign communication from

messages that may conceal malicious intent or traceable links to cybercriminal activity. This not only speeds up the investigative process but also minimizes the risk of human oversight when handling extensive datasets. Beyond criminal investigations, these classification tools can support cybersecurity audits, incident response teams, and legal e-discovery by offering a precise and automated method of analysing digital correspondence. As cybercrime becomes more sophisticated, the integration of such intelligent systems into forensic workflows will become increasingly essential for maintaining both efficiency and thoroughness in uncovering digital evidence.

SCOPE OF THE STUDY

This study focuses on using machine learning models to classify emails as either spam or legitimate messages. It uses a labelled public dataset and compares the performance of different algorithms to determine which model is most accurate and reliable. The research is centred on analysing email text content and does not include real-time monitoring or image-based spam detection. Its primary aim is to build a foundation for better spam detection systems using traditional machine learning methods.

REFERENCES

1. Agarwal, K., & Kumar, T. (2018). Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization. *International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India*, 685-690.
2. Ameen, A. K., & Kaya, B. (2018). Spam Detection in Online Social Networks by Deep Learning. *International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey*.
3. E. Prasannakumar, D. M. (2016). Email Spam Detection Using Machine Learning Algorithms. *2016 IEEE Calcutta Conference (CALCON*, 388-393. doi:https://doi.org/10.1109/CALCON.2016.7890892
4. Geetha Gowri, S. D. (2022). machine learning. *International Journal of Computer Science and Engineering , 10(4)*, 45-51.
5. Harisinghaney, A., Dixit, A., Gupta, S., & Arora, A. (2014). Text and Image-Based Spam Email Classification Using KNN, Naïve Bayes and Reverse DBSCAN Algorithm. *International Conference on Optimization, Reliability, and Information Technology (ICROIT)*, 153-155.
6. Isra'a Abdul Nabi, Q. Y. (2021). Spam Email Detection Using Deep Learning Techniques. *Proceedings of the 2nd International Workshop on Data-Driven Security (DDSW 2021*, 88-94. doi:https://doi.org/10.1109/DDSW51458.2021.9443069
7. Kabir, T., Shemonti, A. S., & Rahman, A. H. (2018). Notice of Violation of IEEE Publication Principles: Species Identification Using Partial DNA Sequence: A Machine Learning Approach. *IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*.
8. Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. *IEEE Access*. Retrieved from https://doi.org/10.1109/ACCESS.2019.2954791
9. Md. Faisal, M. G. (2018). Email Based Spam Detection., (pp. 363-367). doi:https://doi.org/10.1109/CALCON.2018.8722606
10. Mohamad, M., & Selamat, A. (2015). An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification. *International Conference on Computer, Communications, and Control Technology (I4CT)*, 227-231.

11. Preeti Durgapal, A. N. (2021). 2021 International Conference on Computational Intelligence and Data Science (ICCIDS). doi:https://doi.org/10.1109/ICCIDS51481.2021.9442871

12. Shradhanjali, & Verma, T. (2017). E-Mail Spam Detection and Classification Using SVM and Feature Extraction. *Using SVM and Feature Extraction International Journal of Advance Research, Ideas and Innovation in Technology (IJARIIT)*. doi:ISSN: 2454-132X

13. Suryawanshi, S., Goswami, A., & Patil, P. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. *International Conference on Advances in Computing and Communication (IACC)*, 69-74. Retrieved from 10.1109/IACC48062.2019.8971582

14. U. Saranya, P. S. (2021). Email Spam Detection Using Deep Learning. 88-94. doi:https://doi.org/10.1109/DDSW51458.2021.9443069