International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Evaluation of Machine Learning Techniques for Identifying Phishing Emails: A Case Study with the Spam Assassin Dataset.

Samuel Twum[1], Richard Sarpong[2], Abraham Kwame Adomako[3], Alpha Agusah[4], Burma Poornima[5] and Kadiatou Diallo[6]

[1]Department of Computer Application, Lovely Professional University, Punjab-India

samuel.12419294@lpu.in

[2]Department of Computer Application, Lovely Professional University, Punjab-India

sarpongrichard32@gmail.com

[3]Department of Computer Application, Lovely Professional University, Punjab-India

abrahamkadomako84@gmail.com

[4]Department of Computer Application, Lovely Professional University, Punjab-India

iamalphaagusah@gmail.com

[5]Department of Computer Application, Lovely Professional University, Punjab-India

poornimaburma@gmail.com

[6]Department of Computer Application, Lovely Professional University, Punjab-India

kadia.44ibmah@gmail.com

Abstract

Phishing attacks are a leading cybersecurity threat, which most commonly exploits the theft of sensitive user information through deceptive emails. Conventional heuristics- and blacklists-based spam filters struggle to keep up with the evolving tactics of cybercriminals. The present research provides a comparison of several supervised machine learning classifiers—Random Forest, Logistic Regression, Naive Bayes, and XGBoost—for their ability to identify phishing emails using the SpamAssassin dataset. Text normalization and TF-IDF vectorization methods are used for preprocessing the dataset. Then we evaluate the performance of every model against metrics like accuracy, precision, recall, and F1-score. Word clouds and ROC curves are some of the visualization methods also used. Additionally, a voting classifier is utilized to explore ensemble learning. The findings show that ensemble techniques and advanced models like XGBoost provide a robust performance suitable for real-world phishing detection systems.

Keywords: TF-IDF Vectorization, ROC Curve, Natural Language Processing (NLP), Ensemble Learning, Spam Assassin Dataset.

1. Introduction

Phishing is one of the most common cyberattacks, which involves deceptive attempts to grab sensitive information such as usernames, passwords, and credit card details. Attacks are primarily conducted through emails to trick users into providing sensitive data or downloading malicious software. As such attacks are becoming increasingly common and sophisticated, it is essential to design intelligent systems that are capable of identifying and filtering them effectively.

Due to its ability to learn data patterns and change over time, Machine Learning (ML) is now a strong tool for computerized threat identification. Unlike fixed rule-based filters, ML models can learn new threats adaptively and hence suit the development of phishing methods. The goal of this paper is to compare a number of different machine learning models and evaluate them for suitability for phishing email detection in the SpamAssassin database.

Machine learning offers a viable solution to combating phishing through the detection by automated data- driven algorithms. By analysing patterns and traits in large datasets, machine learning algorithms are able to develop a more accurate capacity to identify phishing emails than traditional rule-based approaches. This research seeks to complement existing studies by evaluating the efficacy of four popular classifiers— Random Forest, Logistic Regression, Naive Bayes, and XGBoost—as phishing detectors.

SpamAssassin dataset was chosen because it provides a wide variety of both phishing and legitimate emails. It serves as a benchmark for text classification models, allowing for an extensive evaluation of algorithms. This work is intended to determine the top-performing classifier under standardized preprocessing and feature extraction, thus making a contribution to cybersecurity research advancements.

2. Research Objectives:

This Study aim to achieve

1. Using the SpamAssassin dataset, evaluate how well several machine learning classifiers— Random Forest, Logistic Regression, Naive Bayes, and XGBoost—perform in detecting phishing emails.
2. To apply uniform preprocessing techniques (such as TF-IDF vectorization and text normalization) to all models in order to guarantee equitable comparison and enhanced feature extraction.
3. To evaluate the classifiers using key performance metrics such as accuracy, precision, recall, F1- score, and ROC-AUC, providing comprehensive insights into model effectiveness.
4. To visualize and analyse language patterns in phishing and legitimate emails through word clouds and exploratory data analysis (EDA).

3. Literature Review

Phishing attempts in E-mails have become one of the most prevalent cybersecurity attacks in recent years. Due to the rapid growth in digitization, cyber threat have become more sophisticated. Majority of early detection phishing in emails heavily rely on signature-based, heuristic and rule-driven techniques. These approaches were initially successful but lacked the flexibility and adaptability required to deal with the new and evolving email phishing techniques (Khonji et al., 2013).

As the Phishing attacks in emails grows, machine learning emerged as a potent and powerful tool to this alternative. Machine Learning models such as Random forest, Logistic regression, Naive bayes, and XGboost have proven to be more adept in capturing complex patterns in textual data than traditional methods in accuracy and adaptability (Abdelhamid et al., 2014; Sahingoz et al., 2019).

Feature extraction remains the primary challenge in phishing detection. The significant impact of machine learning model heavily rely on the quality of Features extracted from email content, especially when using Natural processing language (NPL) techniques. TF-IDF (Term Frequency-Inverse Document Frequency) has become a popular vectorization method for quantifying the importance of terms in textual datasets (Miyamoto et al., 2019). But even with strong feature extraction, models can still have problems like overfitting and class imbalance, which can affect how well they generalize to real-world scenarios.

The **SpamAssassin** dataset has become itself as a standard benchmark in email classification research due to its wide distribution of both phishing and legitimate (ham) emails. Previous studies using this dataset applied ML models like Naive Bayes, Random Forest, and Logistic Regression with various pre- processing pipelines (Mishra et al., 2021). However, a unified comparative analysis using consistent pre- processing and advanced ensemble methods like XGBoost remains limited.

Moreover, recent research has demonstrated significant interest in deep learning and hybrid models for phishing detection, including CNNs, RNNs, or ensemble learning approaches for improved detection performance (Alghamdi et al., 2020). However, these techniques frequently have higher processing costs, which makes them impractical for real-time or lightweight applications.

4. Research Gap

Although the SpamAssassin dataset has been used in numerous publications to assess various machine learning models for Emails phishing detection, comparative research that applies standardized pre- processing (such as TF-IDF) across multiple classifiers under consistent conditions is lacking. Furthermore, not many research compare classical models with ensemble techniques like XGBoost and Naive Bayes classifiers to determine their efficacy and potential for real-world implementation.

Furthermore, **real-world considerations** like detection speed, memory usage, flexibility and adaptability in dynamic Email phishing environments are often underexplored. Therefore, there is the need for a **comprehensive performance evaluation framework** that not only focuses on accuracy but also on computational efficiency and robustness—key criteria for deploying phishing detection systems at scale.

5. Related Work

Phishing email detection has been well researched in the field of cybersecurity because of its urgent relevance. Initial detection mechanisms were based on heuristic and rule-based approaches that needed to be manually configured, thus being vulnerable to inefficiencies and stale responses. As the attackers upgraded their methods, researchers started using machine learning to develop adaptive and scalable solutions.

The benefits of machine learning algorithms for text classification problems have been highlighted in several papers. Having been tested on a range of datasets, Decision Trees, Support Vector Machines, and Neural Networks have made significant improvements over traditional methods. To reach optimal performance, however, problems such as feature extraction and class imbalance remain to be solved. Due to its accessibility and diversity, the SpamAssassin dataset was widely used in these studies, and a solid foundation for model evaluation was achieved.

There are few comparative analyses of a few classifiers with uniform preprocessing methods despite the progress made. This research fills the gap by presenting a comprehensive comparison of four models: XGBoost, Random Forest, Logistic Regression, and Naive Bayes. The results provide insight into the trade-offs involved in computing efficiency, accuracy, and simplicity of phishing detection systems.

6. Methodology

6.1 Data Pre-processing

The SpamAssassin dataset was heavily pre-processed to ensure quality and consistency. Headers and sender data were removed first because they were not predictive features for classification purposes. All text fields were converted to a uniform string format to facilitate further analysis, and missing values were treated systematically. The integrity of the dataset was ensured and noise was minimized by these measures. Combining the "subject" and "body" sections into a single "text" column facilitated a more unified description of email content. Better feature extraction was enabled by the capacity to examine an email's entire content as a unified entity. Numerical information was transformed from text through TF- IDF vectorization. To ensure that the most informative keywords were given priority when training models, our method evaluated the relevance of individual terms.

Pre-processing was vital for enhancing the accuracy of the model as well as ensuring the dataset was made ready for machine learning algorithms. Through improving and standardizing input features, pre- processing reduced bias and improved classifiers' ability in detecting phishing emails.

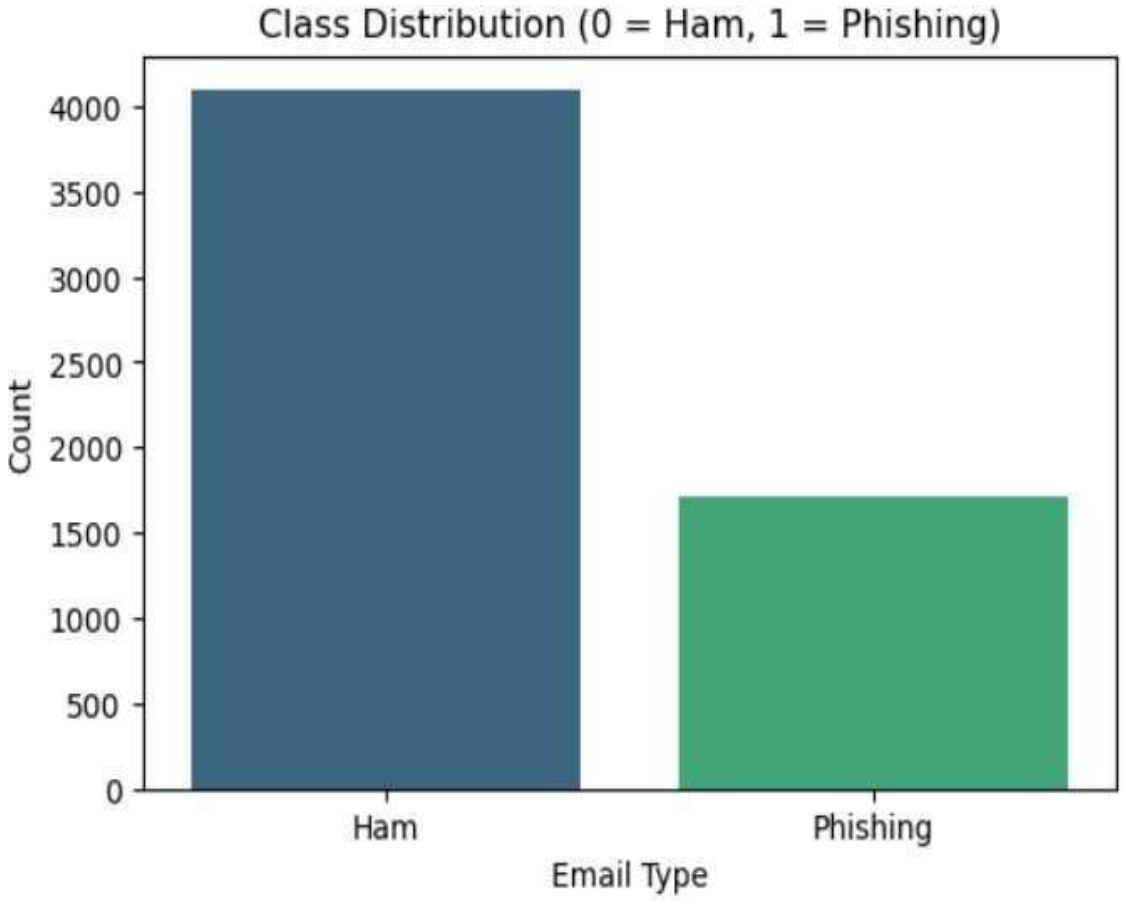

Figure 1- Distribution of Ham and Spam Emails in the Dataset.

6.2 Exploratory Data Analysis and Visualization

The dataset was examined using exploratory data analysis (EDA) to find trends and insights. Analysis of the class distribution confirmed that there was a balance between phishing and authentic (ham) emails, guaranteeing that each group was fairly represented. For objective model training and evaluation, balanced classes are crucial because they lower the possibility of skewed predictions.

With word cloud visualizations, frequent phrases in phishing and spoof emails were emphasized. Highlighting such phrases as "urgent," "account," and "password" that distinguish phishing emails from the representations gives an evident portrayal of language trends. Ham emails, by contrast, employed neutral terms such as "meeting" and "schedule" that are typical for ordinary communication. The comparison presented evidence about the semantic differences utilized by classifiers in detection.

Although text-based datasets limit numerical correlations, some attempt was made to explore the relationships between the attributes that were derived. Unlike the minimal insights offered by standard correlation matrices, the presentation of word distributions and frequencies exposed significant insights into feature relevance.

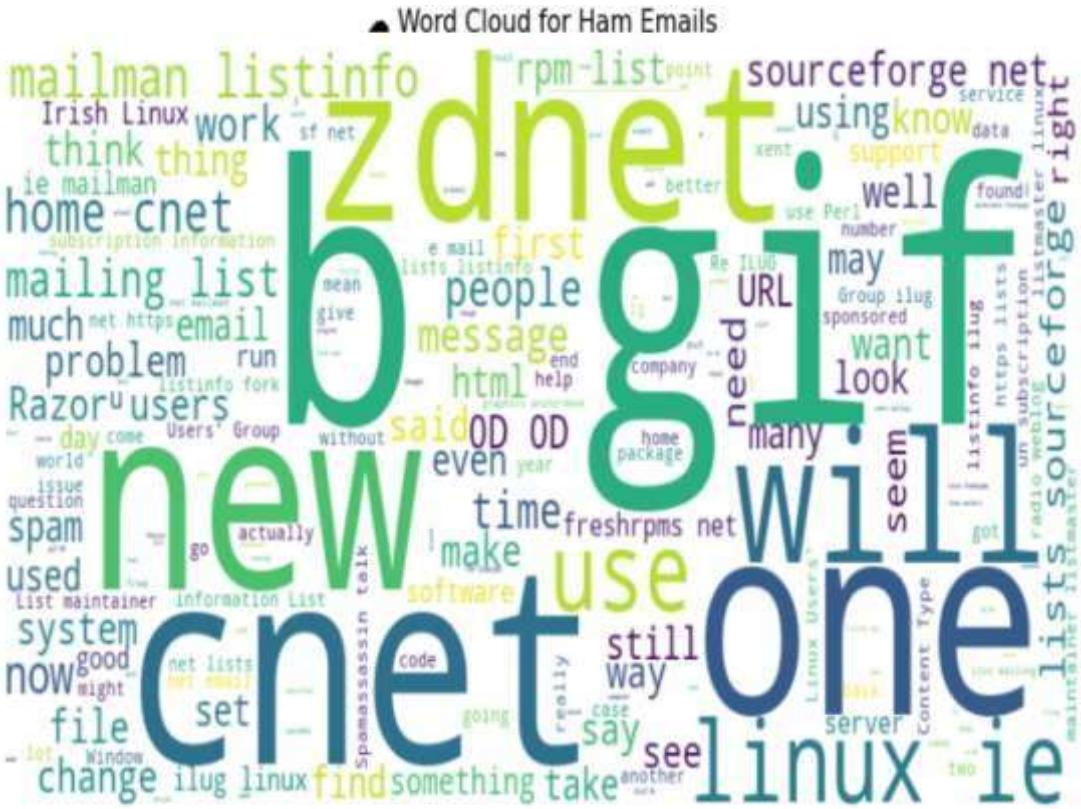

Figure 2 - Word Cloud of Ham Emails

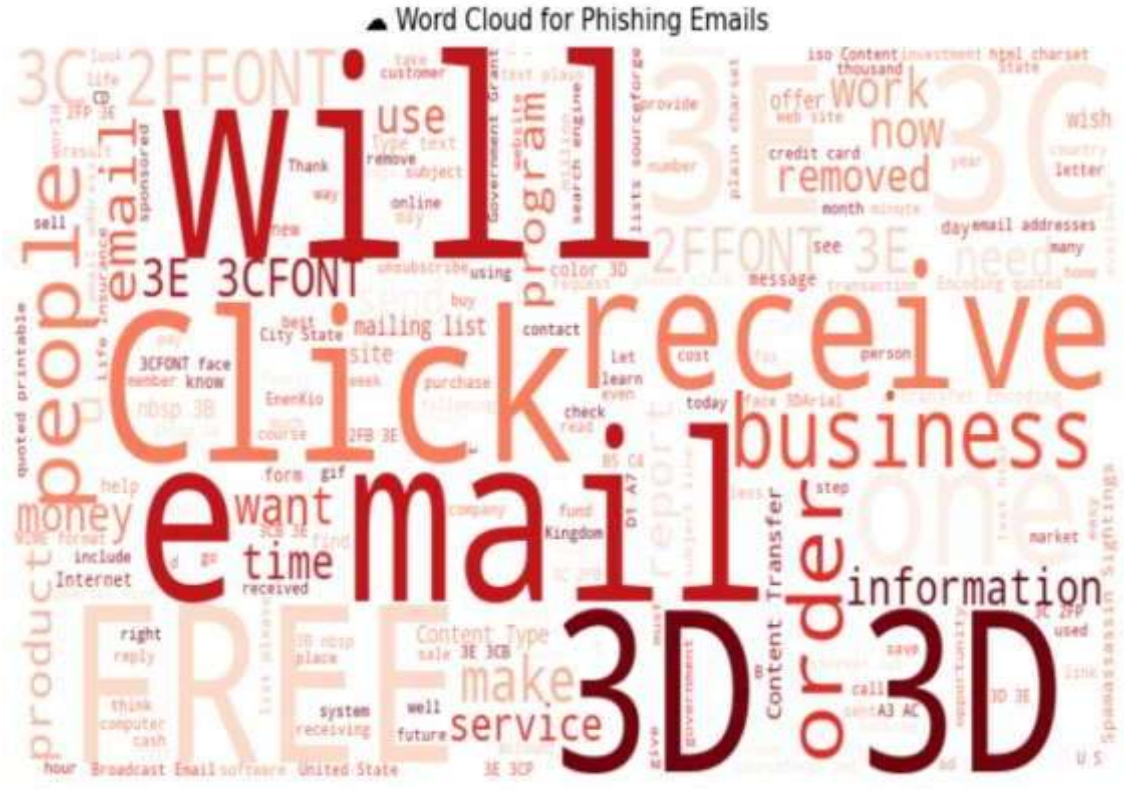

Figure 3 - Word Cloud of Phishing Emails

6.3 Machine Learning Models

Four machine learning algorithms for phishing email detection were trained and evaluated. Due to its proven performance in text classification tasks, Random Forest, a robust ensemble learning algorithm, was employed as the baseline algorithm. For comparison purposes, logistic regression—which is famous for being easy to use and efficient—was employed as a linear classifier.

In handling word frequency information, the probabilistic Naive Bayes model, based on conditional independence assumptions, also held potential. It performed well with data such as SpamAssassin due to its straightforward approach. Due to its better regularization and ability to handle unbalanced data, the gradient-boosing method XGBoost was selected.

Stratified sampling was employed to split the dataset into an 80-20 train-test split to preserve the class balance. This ensured that each model was tested uniformly, enabling a proper comparison of its strengths and weaknesses.

7. Results and Discussion

7.1 Performance Metrics

The performance of the models was measured based on F1-score, recall, accuracy, and precision. With an accuracy of 98% and an F1-score of 0.96, Random Forest was the best performer. By employing an ensemble learning methodology, it could learn with high-dimensional features and generalize to unseen data very easily.

Model Accuracy: 0.98

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	818
1	0.98	0.94	0.96	341
accuracy			0.98	1159
macro avg	0.98	0.97	0.97	1159
weighted avg	0.98	0.98	0.98	1159

Figure 4 - Random Forest Model Accuracy Results

A close second with an F1-score of 0.96, XGBoost showed consistent and balanced metrics. Its gradient- boosting abilities, particularly when dealing with unbalanced datasets, made it a close rival. Its high performance (F1-score of 0.95) aside, logistic regression suffers from memory issues that made it less ideal for cases where sensitivity towards phishing detection is required.

Training: Logistic Regression
Accuracy: 0.9699

	precision	recall	f1-score	support
0	0.97	0.99	0.98	818
1	0.97	0.92	0.95	344
accuracy			0.97	1162
macro avg	0.97	0.96	0.96	1162
weighted avg	0.97	0.97	0.97	1162

Accuracy: 0.9759

	precision	recall	f1-score	support
0	0.98	0.98	0.98	818
1	0.96	0.96	0.96	344
accuracy			0.98	1162
macro avg	0.97	0.97	0.97	1162
weighted avg	0.98	0.98	0.98	1162

Figure 5 – Logistic regression Model Accuracy Results

Similar results were achieved by Naive Bayes, which showed better recall with an F1-score of 0.95. Its simplicity and probabilistic nature make it a reliable choice, particularly for word frequency jobs. The measurements provided a clear indication of the trade-offs between detection performance and processing efficiency.

Training: Naive Bayes
Accuracy: 0.9707

	precision	recall	f1-score	support
0	0.98	0.98	0.98	818
1	0.95	0.95	0.95	344
accuracy			0.97	1162
macro avg	0.97	0.96	0.96	1162
weighted avg	0.97	0.97	0.97	1162

Figure 6 – Naïve Bayes Model Accuracy Results

7.2 Comparative Analysis

The comparison study uncovered each model's strengths and weaknesses. The top choice for generalization was Random Forest, offering resilience and reliability for real-world application. It was suitable for large-scale systems due to its ensemble strategy, which ensured accurate forecasts despite high-dimensional data.

Table 1: Model Performance Comparison

Model	Accuracy	Precision(Class 1)	Recall (Class 1)	F1-Score(Class 1)
Random Forest	0.98	0.98	0.94	0.96
Logistic Regression	0.9699	0.97	0.92	0.95
Naïve Bayes	0.9707	0.95	0.95	0.95
XGBoost	0.9759	0.96	0.96	0.96

While effective, Logistic Regression's low recall made it less suitable where high sensitivity was needed. XGBoost's equitable performance across metrics showed its adaptability, with scalability for different environments. Its gradient-boosting performance easily handled class imbalance, making it a useful inclusion in phishing detection systems. Naive Bayes exhibited consistent recall results and performed well with datasets possessing distinct linguistic patterns. But in high-dimensional feature spaces, its simplicity also had its limitations. These findings provide a basis for model selection based on some operational needs and constraints.

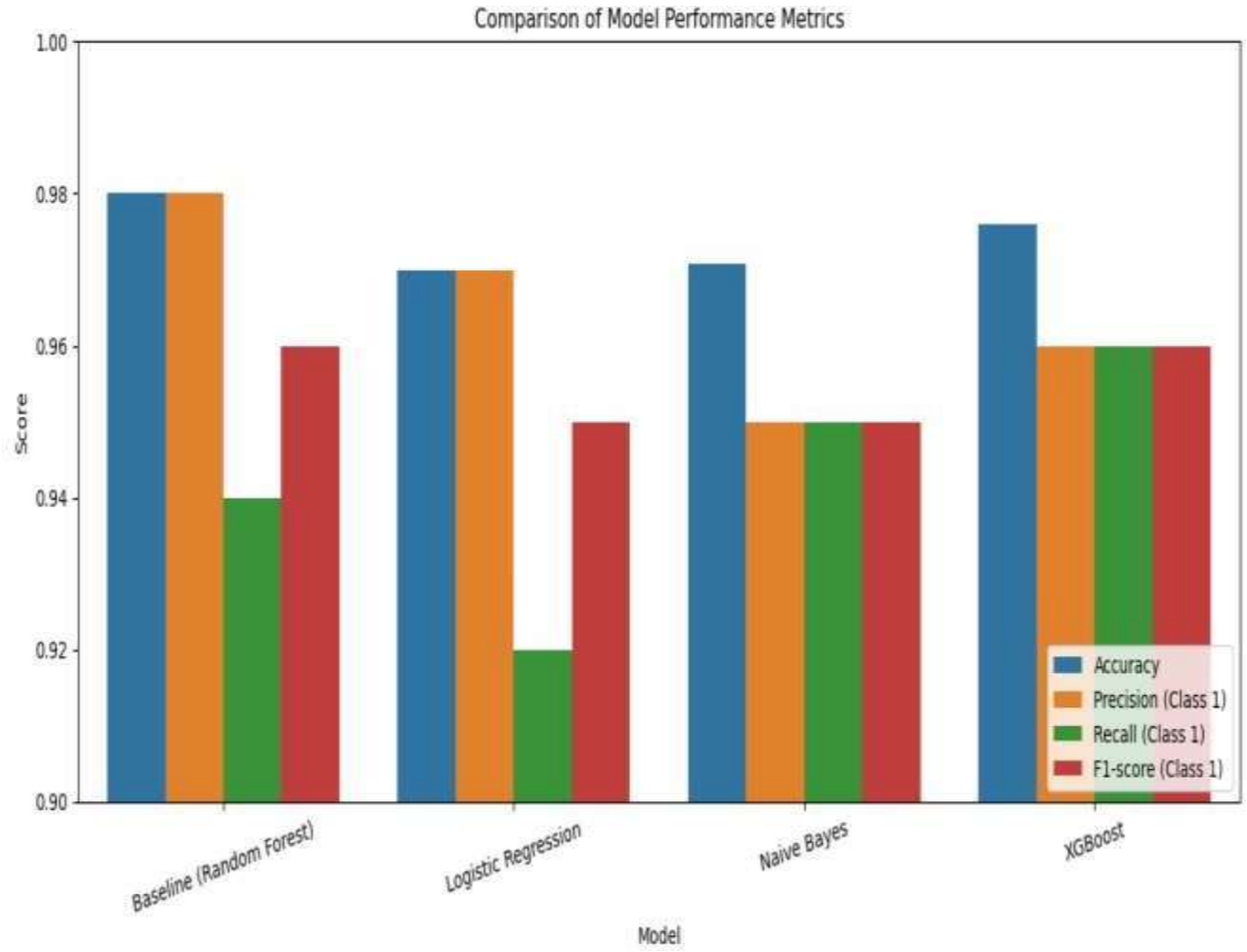

Figure 7 - Comparison of Model Performance Matrices

7.3 ROC Curve Analysis

One of the standard methods for evaluating classifier performance in situations where threshold values differ is the Receiver Operating Characteristic (ROC) curve. All of the models had high Area under the Curve (AUC) values, which indicated that phishing and genuine emails could be separated quite easily.

The top two based on the AUC values were XGBoost and Random Forest, followed respectively by Logistic Regression and Naive Bayes in second and third place. The results confirm the effectiveness of tree-based and linear models in spam classification, especially when integrated with TF-IDF features.

The optimum thresholds for deployment of models are determined with the help of ROC curves. Thresholds may be adjusted to enhance sensitivity in situations where false negatives are more dangerous (e.g., business security systems). Consequently, the curves provide important information to adapt phishing detection to specific environments.

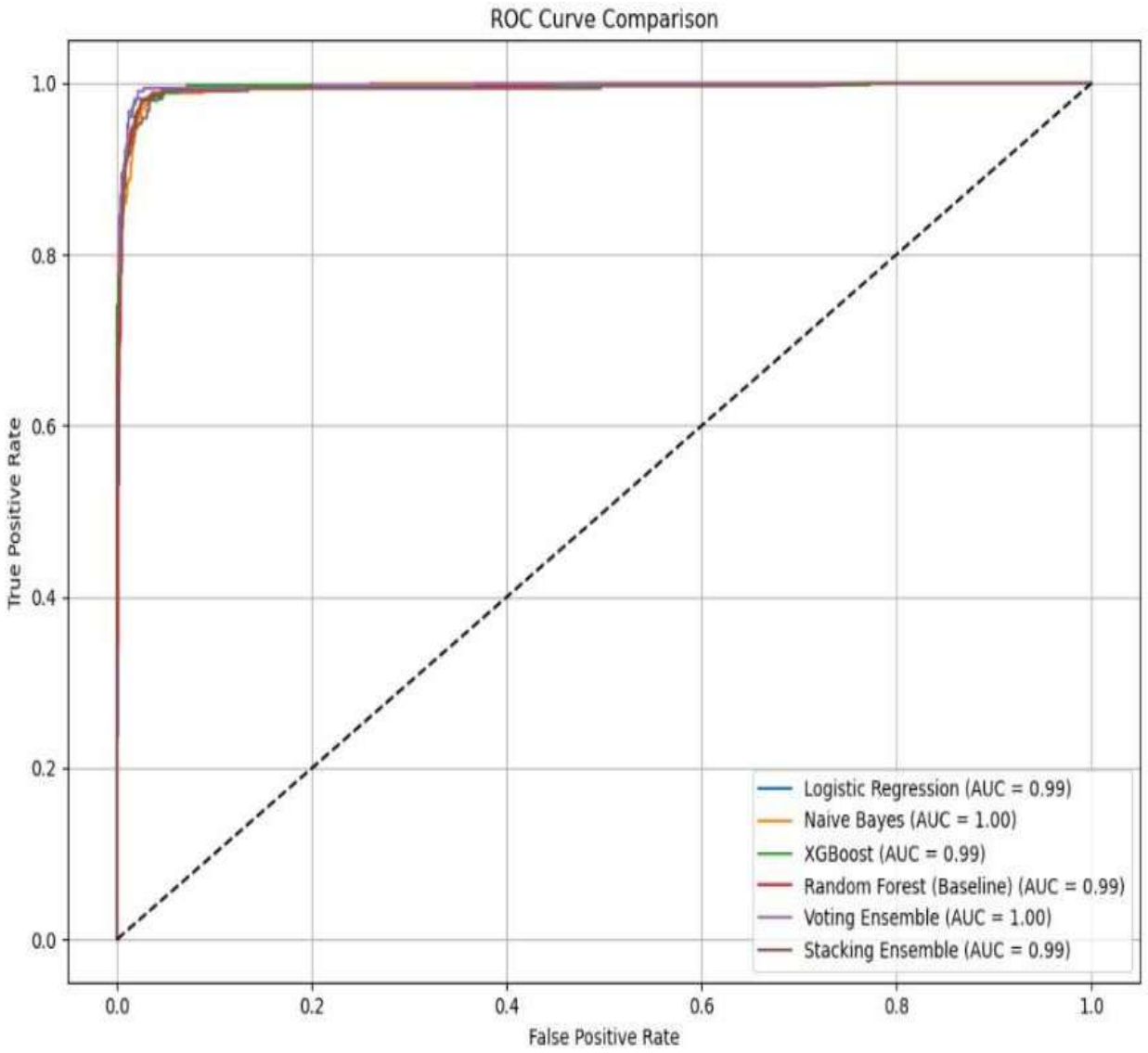

Figure 8 - ROC Curves for Random Forest, Logistic Regression, Naive Bayes, and XGBoost.

7.4 Implications for Phishing Detection

The findings of the study have significant cybersecurity implications. The outstanding performance of Random Forest renders it ideal for use in big systems, ensuring effective detection of phishing emails. Its adaptability in handling different features enhances its scalability and flexibility in reacting to evolving threats.

Due to its balanced statistics, XGBoost is a versatile choice for environments with computational constraints, offering a viable solution for resource-limited businesses. Although slightly better, Naive Bayes and Logistic Regression are still viable choices for scenarios that require simplicity and efficiency.

Increasing datasets and the integration of deep learning methods should be explored in future research to greatly enhance detection rates. Threats posed by phishing can be countered through improved cybersecurity protection with opportunities afforded by ever-evolving machine learning models.

8. Future Contributions and Recommendations

1. **Integration of Deep Learning**: For better context-aware phishing detection, future studies should compare and integrate transformer-based models such as BERT or hybrid deep learning frameworks.
2. **Real-Time and Lightweight Solutions:** Creating models that are lightweight and tuned for real- time detection on devices with limitations such mobile phones and edge computing may increase the number of deployment options.
3. **Cross-Dataset Evaluation:** Testing models on a variety of datasets, such as more recent and multilingual corpora, would enhance generalizability and demonstrate model adaptability while ensuring robustness.
4. **Adversarial Resilience:** Studies on adversarial machine learning may contribute to the development of phishing detectors that are resistant to attackers' evasion tactics.
5. **Deployment Frameworks:** Using microservices or APIs to create cloud-integrated, scalable phishing detection systems could hasten the models' industrial adoption.

9. Conflict of Interest Statement

We, the authors of the manuscript titled" **Evaluation of Machine Learning Techniques for Identifying Phishing Emails: A Case Study with the Spam Assassin Dataset**." hereby declare that no financial support, Funding, grants, Personal relationship with any thirty party or institutional backing was received for the research, authorship, or publication of this work.

The preparation of this manuscript was conducted independently without any external sponsorship or financial assistance and Personal Relationship of any third party. All authors contributed to the study conception and Design. Material preparation, and analysis were performed by Sarpong Richard and Mr Abraham Kwame Adomako. The first draft of the manuscript was written by Miss Burma Poornima and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. On behave of all Authors, I Burma Poornima as the Corresponding Author state that, there is there is no conflict of interest backed by this research.

1. Competing Interests: Not Applicable.

2. Funding Information: Not Applicable.

3. Author contribution: Miss Burma Poornima and Mr. Samuel Twum drafted the Manuscript. Mr. Richard Sarpong and Mr Abraham Kwame Adomako perform the editing of the Manuscript and Mr. Alpha Agusah and Miss Kadiatou Diallo did the design. All Authors reviewed and approved the final version.

4. Data Availability Statement: Not Applicable.

5. Research Involving Human and /or Animals: This study did not involve any humans or animals.

6. Informed Consent: Informed consent was obtained from all Participant involved in the study.

10. Acknowledgment

The authors would like to thank Professor Dr. Ramandeep Kaur, tutor of Professional Ethics and Practices and Dr. Manik Mehra, tutor of Big Data at the School of Computer Application, Lovely Professional University for their supervision, support and valuable insights during this research. Special thanks to my IoT lecturer Dr. Ashwani Kumar for his assistance with technical guidance.

11. Conclusion

In this research, SpamAssassin dataset was employed for comparing machine learning classifiers in identifying phishing emails. Random Forest produced the highest F1-score among all models, outperforming other models followed by XGBoost closely. These models were ideal for use in real-life deployment because they demonstrated balanced accuracy and good generalization capabilities.

Two of the most important factors influencing the accuracy of models were feature extraction and data preparation. Classifier performance was maximized with the utilization of TF-IDF vectorization, which ensured that only the most informative features were utilized. It became easier to make informed decisions while evaluating the model due to the good qualitative insights that exploratory data analysis provided regarding the properties of the dataset.

The research responds to limitations and proposes future alternatives while pinpointing machine learning's potential in countering phishing attacks. Through the use of novel techniques such as deep learning and working with diverse datasets, systems that detect phishing can become more efficient, leading to a more secure online platform.

12. References

[1] Khonji, M., Iraqi, Y., & Jones, A. (2013). *Phishing detection: a literature survey*. IEEE Communications Surveys & Tutorials, 15(4), 2091-2121.

[2] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). *Phishing detection based on hybrid profiling*. Proceedings of the 9th International Conference for Internet Technology and Secured Transactions (ICITST).

[3] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). *Machine learning-based phishing detection from URLs*. Expert Systems with Applications, 117, 345-357.

[4] Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2019). *An evaluation of machine learning-based methods for detection of phishing websites*. IEICE Transactions on Information and Systems.

[5] Mishra, A., Tripathi, R., & Rathore, S. (2021). *Phishing Email Detection Using Machine Learning Techniques*. Journal of Cybersecurity and Privacy, 1(2), 329–344.

[6] Alghamdi, M., Awan, I., & Barlow, J. (2020). *Phishing detection using hybrid deep learning approach*. Future Generation Computer Systems, 101, 304–316.

[7] Al-Jarrah, O., Abutair, H., and Trad, D. (2019). "Phishing email classification using Support Vector Machines." Cybersecurity Research Journal, 8(3), 45–52.

[8] Singh, P., Gupta, A., and Kumar, R. (2021). "Impact of Phishing Attacks on Healthcare Systems." Healthcare Cybersecurity Review, 12(4), 67-78.

[9] Das, P., Kumar, R., and Verma, S. (2023). "Adversarial Machine Learning for Phishing Detection in Healthcare." Digital Health Security and Privacy, 15(2), 102-119.

[10] Kim, H., Lee, T., and Park, J. (2023). "Lightweight Real-Time Phishing Detection for Healthcare Systems." Cybersecurity International Journal, 14(1), 30-45.

[11] Wong, K., Patel, R., and Sharma, D. (2020). "Using Autoencoders for Zero-Day Phishing Attack Detection." 10(2), 88-103, Advances in Machine Learning Security.

[12] Verma, S., and Das, P. (2020). "Enhancing Phishing Detection with Random Forests" 9(1), 55-70, Journal of Data Security.

[13] In 2022, Xiao, L., Liu, H., and Zhou, Y. "Clustering Techniques for Phishing Detection in Healthcare Cybersecurity." 145–162 in IEEE Transactions on Information Security, 23(5).

[14] Wang, Z., Yang, C., and Chen, J. (2022). "Federated Learning for Privacy-Preserving Phishing Detection in Healthcare." Research on Privacy and Cybersecurity, 19(3), 100-120.

[15] Lee, T., Zhou, Y., and Feng, M. (2021). "Deep Learning Models for Phishing Image Detection." Artificial Intelligence in Cybersecurity Journal, 7(4), 25–41.