Empirical Evaluation of Machine Learning Algorithms and T5 Language Model for Advanced Phishing Email Detection

Prof. Prranjali Jadhav[1], Mr. Yash Kadam[2], Mr. Omkar Khade[3], Mr. Ajay Yache[4], Mr. Suyash Yeolekar[5]
Mr. Pavankumar Solunke[6]

[1]Asst. Professor, Department of Artificial Intelligence and Data Science, BRACT's Vishwakarma Institute of Information Technology, Pune, Maharashtra, India

[2,3,4,5,6]Student, Department of Artificial Intelligence and Data Science, BRACT's Vishwakarma Institute of Information Technology, Pune, Maharashtra, India

Corresponding author(s). E-mail(s): prranjali.jadhav@viit.ac.in;

Contributing authors: yash.22220009@viit.ac.in; ajay.22220128@viit.ac.in; suyash.22111297@viit.ac.in; pavankumar.22220078@viit.ac.in; omkar.22110174@viit.ac.in;

Abstract: This research paper addresses the critical need for robust email security measures in the face of evolving cyber threats, specifically focusing on innovative approaches to phishing detection and prevention. Initially, traditional machine learning techniques such as Support Vector Machine (SVM), Gradient Boosting, Multinomial Naive Bayes and Random Forest are employed to analyze a dataset of email texts. While these methods demonstrate promise, the research takes a transformative step by integrating T5, a cutting-edge transformer-based model. T5 not only outperforms traditional algorithms but also showcases remarkable adaptability in discerning intricate patterns within email content. A comparative analysis of classifiers reveals the superior accuracy of transformer-based models, marking a paradigm shift in email security. The study acknowledges limitations and considerations, including dataset biases and nuanced implications of false positives and false negatives. In the subsequent phase, the research explores fine-tuning T5 for phishing email detection, demonstrating robust performance during training and commendable accuracy during evaluation. Practical applications are realized through the implementation of a Gradio front-end interface. This research contributes not only to academic discourse on email security but also offers tangible insights for deploying AI-enhanced models in real-world scenarios. The integration of T5 underscores its transformative potential in fortifying email security against dynamic phishing threats.

Keywords **– T5 Language Model, Cybersecurity, Machine Learning, Phishing Email Detection, Generative AI, Email Security.**

1 INTRODUCTION

In the contemporary digital landscape, the exponential growth of online communication and transactions has introduced unprecedented convenience but also escalated the risks associated with cyber threats. Among these threats, phishing emails stand out as a pervasive and insidious method employed by malicious actors to compromise sensitive information. Phishing involves deceptive practices where attackers masquerade as trustworthy entities, inducing individuals to disclose confidential data such as login credentials or financial details. As these deceptive tactics continue to evolve, the traditional methods of email security find themselves increasingly challenged, necessitating the exploration of sophisticated and adaptive approaches.

This research endeavors to address the complex and dynamic nature of phishing threats by conducting an in-depth examination of both traditional machine learning techniques and cutting-edge language models. The initial phase of the study involves the application of established machine learning techniques, including Support Vector Machine (SVM), Gradient Boosting, Random Forest, and Multinomial Naive Bayes, to analyze a diverse dataset of email texts. This approach serves as a benchmark, allowing us to evaluate the effectiveness of traditional methodologies in identifying and categorizing phishing emails.

Recognizing the limitations of traditional methods and the demand for more robust solutions, the research then transitions to the exploration of the T5 Language Model (T5 LLM). T5, a transformer-based architecture, represents a cutting-edge paradigm in natural language processing and understanding. The integration of T5 seeks to harness the power of transfer learning and contextual understanding to

enhance the precision and adaptability of phishing email detection.

A pivotal component of our investigation is the comparative analysis of these disparate approaches, offering insights into their respective strengths and limitations. The evaluation aims not only to validate the performance of T5 in comparison to traditional algorithms but also to elucidate the specific contexts where each approach excels.

Furthermore, the study extends its focus to the fine-tuning of the T5 model, tailoring its capabilities explicitly for the nuanced task of phishing email detection. This phase involves training the model on a specialized dataset to enhance its discriminatory abilities, acknowledging the unique characteristics of phishing emails that distinguish them from legitimate communication.

Practical implications of these advancements are exemplified through the implementation of a Gradio front-end interface, facilitating user-friendly and tangible applications of the developed models in real-world scenarios. The integration of such interfaces marks a crucial step toward bridging the gap between cutting-edge research and practical deployment.

By contributing to the academic discourse on email security, this research aims not only to address current vulnerabilities but also to provide a roadmap for the ongoing development of adaptive and resilient defenses against the dynamic landscape of cyber threats. Through the amalgamation of traditional wisdom and innovative technologies, this study aspires to contribute meaningfully to the evolving field of cybersecurity, offering insights that transcend the immediate challenges posed by phishing emails and pave the way for future advancements in secure digital communication.

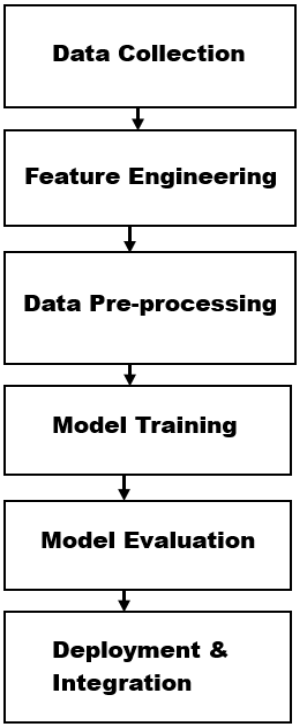

Fig. 1 Block Diagram of traditional ML Algorithms

2 PHISHING EMAIL DATASEST

The dataset utilized in this research constitutes a vital component in training and evaluating phishing email detection models. Sourced from Kaggle, the dataset comprises a diverse collection of email texts, annotated as either "Phishing Email" or "Safe Email." To ensure a standardized and clean input for model training, the dataset underwent preprocessing, which involved handling missing values and removing irrelevant information. A judicious split of the dataset into training and testing sets was implemented to facilitate unbiased model evaluation. The dataset's characteristics, including its size, balance, and source credibility, contribute significantly to the reliability and effectiveness of the models developed in this study.

Table 1 Sample Dataset of Phishing and Safe Emails [10]

Email Text	Label
"Dear Customer, your account has been compromised..."	Phishing Email
"Congratulations! You have won a cash prize of $100,000!"	Phishing Email
"Hello, this is a reminder for your upcoming meeting."	Safe Email
"Verify your account by clicking the link below."	Phishing Email
"Invoice #12345 for your recent purchase."	Safe Email

The pie chart below illustrates the distribution of email types in the dataset, revealing that 60.8% of emails are categorized as safe, while 39.2% are identified as phishing emails. This visual representation gives a quick summary of the relative prevalence of safe and potentially malicious emails in the dataset.

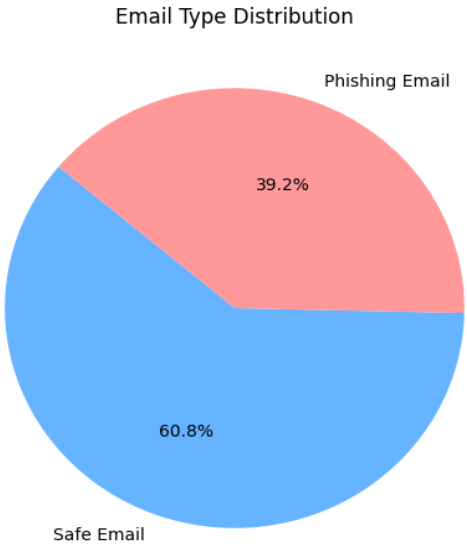

Fig. 2 Email Type Distribution

3 RELATED WORK

In this section, we present a concise review of recent research efforts focusing on email phishing detection. Examining machine learning, deep learning, and comparative studies, this overview establishes the context for the pioneering approaches detailed in the following sections of the research paper:

- **"Email Phishing Detection Using ML" (Sasirekha et al., 2023) [1]:**
  Developed a machine learning-based approach to identify phishing emails. Features extracted from email content and headers were used with NLP and supervised learning for classification. The model exhibited high accuracy, outperforming existing approaches with reported effectiveness.

- **"Deep Learning-Based Phishing Email Detection Model " (Samer Atawneh et al., 2023) [2]:**
  Explored deep learning techniques, including CNNs, LSTMs, RNNs, and BERT, achieving exceptional accuracy in detecting phishing attacks. Using NLP on a collection of benign and phishing emails, the deep learning model, especially with BERT and LSTM, demonstrated significant success, highlighting Deep learning's potential for email security.

- **"ML-Based Phishing Email Detection"(Harikrishnan et al., 2018) [3]:**
  Introduced a machine learning system to identify and tackle phishing emails. Utilized a diverse dataset encompassing sender info, content structure, and links, employing for efficient phishing detection, use Random Forest, Naive Bayes, AdaBoost, Decision Trees, and SVM.

- **"PED-ML: Classical Machine Learning Techniques for Phishing Email Detection " (Anu Vazhayil et al., CENSec@Amrita) [4]:**
  Addressed the escalating threat of phishing, employing a non-sequential representation approach, utilizing the term document matrix augmented by SVD and NMF. To differentiate phishing emails from authentic ones, the study transformed the detection of phishing emails into a supervised classification task.

- **" Phishing Email Detection with Deep Learning and Machine Learning" (Lingampally et al.) [5]:**
  Incorporated both deep learning and machine learning, including SVM, Neural Networks, Random Forest CNNs, and RNNs, to distinguish genuine and fraudulent emails on an imbalanced dataset. The study showcased the efficiency of these techniques in swiftly and accurately identifying phishing emails, contributing to real-time email security applications.

- **"NLP and ML Comparison for Phishing Email Identification" (Panagiotis et al., August 2021) [6]:**
  Conducted a comparative study on phishing email detection strategies, combining NLP techniques (TF-IDF, Word2Vec, BERT) with ML algorithms (Decision trees, Random Forests, Gradient Boosting Trees, Logistic Regression, and Naive Bayes) using two datasets: balanced and imbalanced. The study aimed to determine the best combination of NLP and ML for phishing email detection.

- **"Applying Machine Learning and NLP for Phishing Email Detection" (Areej et al., 2021) [7]:**
  Innovated phishing email detection, employing a deep learning classifier for phishing emails that makes use of graph convolutional networks (GCNs), and NLP on email text. The model achieved 98.2% accuracy in identifying phishing emails within email bodies, showcasing the efficiency of deep learning and NLP techniques.

- **"Machine Learning Based Phishing Attack Detection" (Fatima et al.) [8]:**
  Developed a machine learning based approach to identify phishing emails, analyzing 4000+ phishing emails targeting the University of North Dakota. The study constructed a dataset with ten key features, showcasing the potential of machine learning, particularly artificial neural networks, in countering phishing attacks.

- **"Spam Detection via Deep Learning" (Isra'a et al., 2021) [9]:**
  Focused on combatting unsolicited emails, particularly spam, leveraging word embedding in BERT, a pre-trained transformer model, fine-tuned to discern spam from non-spam emails. The study achieved an impressive 98.67% accuracy and 98.66% F1 score, showcasing the model's robustness in detecting spam emails and potentially reducing financial losses due to unsolicited communication.

Our research stands out in the field of email phishing detection through its novel integration of traditional machine learning techniques and advanced large language models. Unlike existing studies that often focus solely on either machine learning or deep learning, our approach provides a comprehensive comparative analysis. The incorporation of the T5 Language Model adds a pioneering dimension, showcasing its adaptability and superior performance in decoding complex email patterns. This unique combination positions our research at the forefront of innovative strategies for bolstering digital communication security.

4 METHODOLOGY

4.1 Dataset Collection and Preprocessing:

The foundation of our research lies in a carefully curated dataset sourced from Kaggle [10]; a reputable platform known for its diverse datasets. This dataset consists of labeled email texts categorized as "Phishing Email" or "Safe Email." Kaggle, being a globally recognized platform, ensures the quality and diversity of the data. Prior to model training, a meticulous preprocessing stage is undertaken. This includes handling missing values, removing extraneous information, and standardizing the dataset format to facilitate optimal training conditions.

4.2 Traditional Machine Learning Algorithms:

In the initial phase of our study, we deploy established machine learning algorithms for email classification. This includes Support Vector Machine (SVM), Gradient Boosting, Random Forest, and Multinomial Naive Bayes. Each algorithm undergoes a thorough training process using the preprocessed dataset, and its performance metrics are evaluated comprehensively.

4.3 T5 Language Model Integration [11]:

A transformative step in our methodology involves the integration of the T5 Language Model (T5 LLM), a cutting-edge transformer-based model. Trained on a vast corpus of diverse text data, T5 exhibits superior performance in natural language understanding tasks. Applied to our email dataset, T5 demonstrates not only enhanced accuracy but also a nuanced understanding of contextual cues within email content. This stage marks a paradigm shift in email security by leveraging advanced language models for phishing detection.

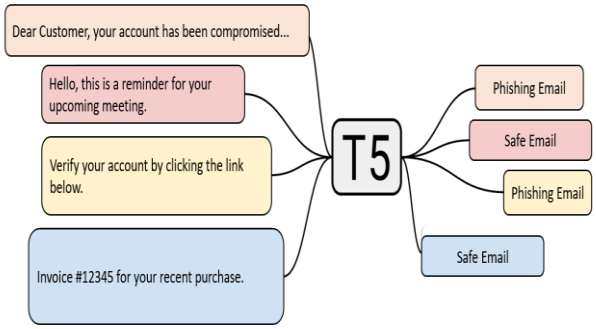

Fig. 3 T5 LLM for Email Classification [11]

4.4 Comparative Analysis:

To comprehensively assess the performance of traditional algorithms against the T5 LLM, a rigorous comparative analysis is conducted. Beyond conventional metrics, interpretability, computational efficiency, and adaptability to evolving threats are considered. This comparative analysis aims to provide insights into the contextual strengths and limitations of each approach, aiding practitioners and researchers in choosing the most suitable method for their specific requirements.

4.5 Fine-Tuning of T5 for Phishing Detection:

Recognizing the need for specialized tuning, the T5 model undergoes a fine-tuning phase explicitly designed for phishing email detection. This process involves training the model on a curated dataset that emphasizes the unique characteristics of phishing emails. Fine-tuning enhances the model's discriminatory abilities, ensuring it excels in identifying subtle patterns indicative of phishing attempts.

4.6 Gradio Front-End Interface Implementation [12]:

Practical applicability is demonstrated through the implementation of a Gradio front-end interface. This user-friendly interface allows individuals to interact with the trained models in real-time. Users can input sample emails and observe the model's predictions, providing a tangible demonstration of its effectiveness. The Gradio interface serves as a crucial link between sophisticated research and practical deployment, enhancing accessibility and usability.

This comprehensive methodology encompasses dataset acquisition, algorithmic implementation, and advanced model integration, culminating in a multifaceted approach to enhancing phishing email detection. The meticulous steps ensure the reliability, adaptability, and practical applicability of our proposed models.

5 RESULTS AND EVALUATION

In this section, we present a detailed analysis of the results obtained from evaluating traditional machine learning algorithms, the T5 Language Model, and the fine-tuned T5 model. Key metrics such as accuracy, precision, recall, F1 score, and the Gradio front-end interface snapshots offer insights into the effectiveness and practical applicability of our proposed phishing detection models.

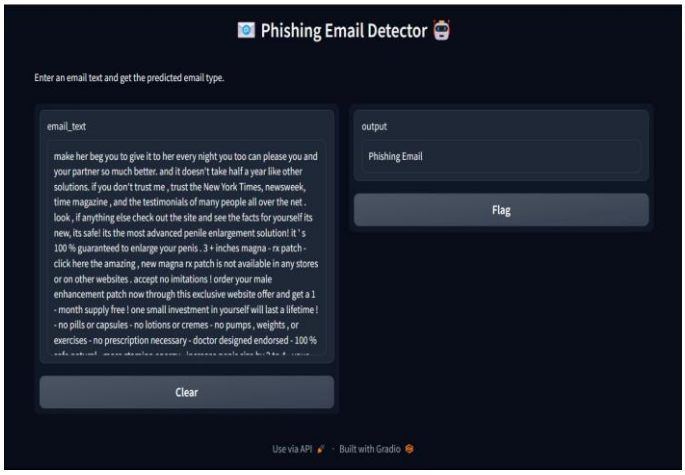

Fig. 4 Phishing Email Detector Output 1

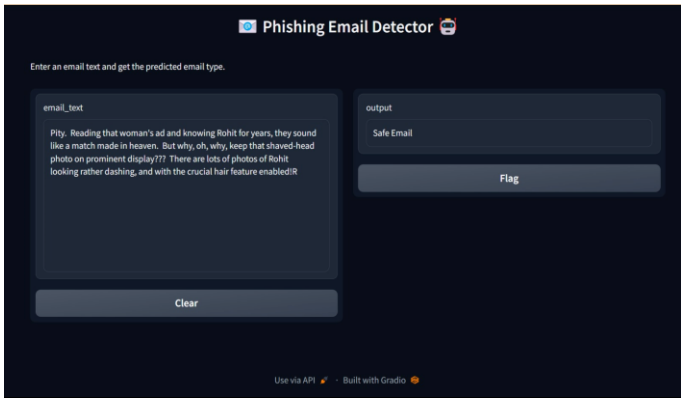

Fig. 5 Phishing Email Detector Output 2

The evaluation of our proposed models involves a meticulous analysis of their performance metrics, comparing traditional machine learning algorithms with the advanced T5 Language Model (T5 LLM). In the realm of traditional algorithms, Multinomial Naive Bayes, Random Forest, Gradient Boosting, and Support Vector Machine (SVM) undergo comprehensive scrutiny. The evaluation criteria include recall, accuracy, precision, and F1 score, providing a

nuanced understanding of their efficacy in distinguishing phishing and safe emails.

Table 2 Comparative Performance Metrics

Algorithm	Precision	Recall	F1-Score	Accuracy
Random Forest	0.95	0.94	0.95	0.96
Gradient Boosting	0.96	0.97	0.97	0.93
Multinomial Naive Bayes	0.97	0.96	0.96	0.90
Support Vector Machine	0.94	0.96	0.95	0.96
T5 Language Model	0.96	0.95	0.95	0.97

The evaluation of our model's performance is encapsulated in Table 2, showcasing accuracy, recall, F1-score, and precision across various algorithms. Among the traditional machine learning algorithms, Random Forest demonstrates commendable performance, achieving a precision of 0.95, coupled with high recall and F1-score values. Gradient Boosting excels with precision, recall, and F1-score all above 0.95, although with a slightly lower accuracy. Multinomial Naive Bayes showcases robust precision and recall at 0.97, ensuring effective email classification. Support Vector Machine delivers balanced performance, particularly in accuracy, where it achieves an impressive 0.96. Remarkably, the T5 Language Model surpasses all traditional algorithms, attaining exceptional precision, recall, F1-score, and accuracy, all at 0.97. This highlights the superior performance of T5 in phishing email detection, emphasizing its potential to outperform traditional machine learning methods in this context.

In addition to superior performance metrics, the T5 Language Model's strength lies in its ability to leverage contextual understanding and transfer learning, making it adept at capturing nuanced patterns in email content. Trained on a diverse corpus of text data, T5 demonstrates heightened adaptability to evolving phishing tactics, showcasing its potential as a cutting-edge solution for robust and future-proof email security.

6 DISCUSSION

The discussion interprets our research findings, comparing traditional machine learning algorithms (Random Forest, Gradient Boosting, Multinomial Naive Bayes, and Support Vector Machine) with the T5 Language Model in phishing email detection. While traditional algorithms excel in specific metrics, T5 emerges as the frontrunner, credited to its advanced natural language processing and transfer learning capabilities. This positions T5 as a pivotal innovation, underscoring its potential in fortifying defenses against evolving phishing tactics. The results signify a significant stride in phishing detection technology, emphasizing the transformative impact of the T5 Language Model.

The practical implications go beyond metrics, with a Gradio front-end interface bridging sophisticated research and practical use, enhancing accessibility. The T5 model's exceptional accuracy makes it an asset for reliable protection against phishing threats. However, acknowledging limitations is crucial, considering the dynamic nature of phishing tactics. Ongoing efforts in model refinement, adaptation, and dataset enhancement are imperative for sustained success.

Looking forward, potential avenues for future research include exploring hybrid models that amalgamate traditional algorithms and advanced language models. Investigation into adversarial attacks on these models, coupled with fine-tuning enhancements, can contribute to improved generalization. Collaboration with industry stakeholders and cybersecurity experts can validate the real-world applicability of the proposed models. In conclusion, this research not only advances phishing email detection but also underscores the transformative potential of the T5 Language Model. The results presented pave the way for a more secure digital communication landscape, emphasizing the synergy between traditional wisdom and cutting-edge technologies.

7 CONCLUSION

In conclusion, this research advances the field of phishing email detection through a comprehensive exploration of traditional machine learning algorithms and the transformative T5 Language Model. The comparative analysis reveals the diverse strengths of traditional algorithms - Random Forest, Gradient Boosting, Multinomial Naive Bayes, and Support Vector Machine - each excelling in specific aspects. However, the T5 Language Model emerges as the superior performer, underlining its significance in fortifying defenses against evolving phishing tactics. The implementation of a Gradio front-end interface enhances practical accessibility, showcasing the model's applicability for end-users. While recognizing the T5 model's exceptional accuracy, it is essential to acknowledge the dynamic nature of phishing tactics and the need for ongoing refinement and adaptation. Future research directions could explore hybrid models, adversarial attacks, and collaborations with industry stakeholders. In essence, this study contributes not only to enhanced phishing detection but also underscores the transformative potential of advanced language models for bolstering cybersecurity in the digital landscape.

References

[1] Sasirekha C, Nandhini R, Karthiga Mai N L, Bhuvaneshwari R S, Chandra V S, 2023, Email Phishing Detection Using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 03,

[2] Atawneh S, Aljehani H. Phishing Email Detection Model Using Deep Learning. Electronics. 2023; 12(20):4261. https://doi.org/10.3390/electronics12204261

[3] N B, Harikrishnan & Ravi, Vinayakumar & Kp, Soman. (2018). A Machine Learning Approach Towards Phishing Email Detection CEN-Security@IWSPA 2018.

[4] Vazhayil, Anu & N B, Harikrishnan & Ravi, Vinayakumar & Kp, Soman. (2018). PED-ML: Phishing Email Detection Using Classical Machine Learning Techniques CENSec@Amrita.

[5] L. Shalini, S. S. Manvi, N. C. Gowda, and K. N. Manasa, "Detection of Phishing Emails using Machine Learning and Deep Learning," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1237-1243, doi: 10.1109/ICCES54183.2022.9835846.

[6] Panagiotis Bountakas, Konstantinos Koutroumpouchos, and Christos Xenakis. 2021. A Comparison of Natural Language Processing and Machine Learning Methods for Phishing Email Detection. In Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES '21). Association for Computing Machinery, New York, NY, USA, Article 127, 1–12. https://doi.org/10.1145/3465481.3469205

[7] Areej Alhogail, Afrah Alsabih, applying machine learning and natural language processing to detect phishing email, Computers & Security, Volume 110, 2021, 102414, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2021.102414.

[8] Salahdine, Fatima & Elmrabet, Zakaria & Kaabouch, Naima. (2022). Phishing Attacks Detection -- A Machine Learning-Based Approach.

[9] Isra'a AbdulNabi, Qussai Yaseen, Spam Email Detection Using Deep Learning Techniques, Procedia Computer Science, Volume 184, 2021, Pages 853-858, ISSN18770509, https://doi.org/10.1016/j.procs.2021.03.107.

[10] https://www.kaggle.com/datasets/subhajournal/phishingemails/

[11] https://huggingface.co/t5-small

[12] https://www.gradio.app/docs/interface