Machine Learning Framework for Phishing Detection through Email using Imbalanced Data

Mher Mkrtumyan
mmkrtumyan29@gmail.com
Horizon Academic Research Program
Los Angeles, California, USA

ABSTRACT

Phishing emails remain a persistent cybersecurity threat, as attackers continue to employ increasingly sophisticated techniques to deceive users into disclosing sensitive information. Detection is particularly challenging in real-world environments, where legitimate emails vastly outnumber malicious ones. This study presents a dual-layer machine learning framework for phishing detection that independently analyzes sender metadata and email body content. The sender layer evaluates structural characteristics of email addresses, while the content layer extracts linguistic and statistical features from email text. Each layer produces a probability score representing the likelihood of phishing; these are subsequently integrated using a meta-classification model to generate a final decision. The framework is evaluated on a large, real-world dataset containing over 500,000 emails with a highly imbalanced class distribution. Experimental results demonstrate that the proposed approach provides robust and reliable performance under realistic conditions, highlighting the effectiveness of integrating multiple analytical perspectives for practical phishing detection.

KEYWORDS

Phishing Detection, Machine Learning, Dual-Layer Framework, Email Security, Imbalanced Data, Sender Metadata, Email Content Analysis, Random Forest, XGBoost, LightGBM, ROC Curve, Logistic Regression, Cybersecurity, Dataset Preprocessing

ACM Reference Format:
Mher Mkrtumyan. 2026. Machine Learning Framework for Phishing Detection through Email using Imbalanced Data. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 1)*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.5281/zenodo.18435064

1 INTRODUCTION

Phishing is a widespread form of cybercrime in which attackers attempt to deceive individuals into revealing sensitive information, such as passwords, financial details, or personal data. These attacks are most commonly carried out through fraudulent emails that impersonate trusted organizations or individuals [22]. As digital communication becomes increasingly central to daily life, phishing attacks have grown both in frequency and sophistication, posing a serious threat to individuals, institutions, and businesses.

Traditional phishing detection methods often rely on predefined rules or simple heuristics, such as blacklisted domains or keyword matching. While effective against basic attacks, these approaches struggle to detect modern phishing emails that closely resemble legitimate messages. Machine learning–based detection systems have therefore gained popularity due to their ability to identify complex patterns within large datasets. However, many existing systems focus on a single aspect of an email—either sender information or content—limiting their effectiveness in real-world scenarios.



Figure 1: Number of phishing attacks detected worldwide from 3rd quarter 2013 to 4th quarter 2024 [14].

Another major challenge in phishing detection is data imbalance. In real-world email traffic, phishing messages represent only a small fraction of total emails, yet they have disproportionately severe consequences. Models trained on artificially balanced datasets may report high accuracy but often fail to generalize to realistic environments. This study addresses these challenges by proposing a dual-layer machine learning framework designed specifically for large, imbalanced datasets.

1.1 Research Question

This research investigates the following question:

Can a dual-layer machine learning framework that independently analyzes sender metadata and email content improve phishing detection performance on large, real-world, imbalanced email datasets?

By separating sender-based and content-based analysis and integrating their outputs through a meta-model, this study aims to

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
IJSCAR VOL. 3, ISSUE 1, February 2026,
© 2026 The Author(s). Published by IJSCAR.
https://doi.org/10.5281/zenodo.18435064

evaluate whether a multi-perspective approach provides greater robustness and practical effectiveness than traditional single-layer systems.

1.2 Challenges and Recent Developments

Phishing detection presents several technical challenges due to the evolving nature of attack strategies. Attackers frequently manipulate both the visible content of emails and the underlying sender information to evade detection systems. Content-based detection models may fail when phishing emails contain minimal text, ambiguous language, or carefully crafted wording. Conversely, sender-based approaches may struggle when attackers use compromised accounts or realistic-looking domains.

As a result, relying exclusively on either content or sender metadata is often insufficient. Modern phishing emails are designed to exploit the weaknesses of single-layer detection systems by appearing legitimate in one dimension while remaining malicious in another.

Recent research has increasingly favored hybrid detection frameworks that incorporate multiple feature types, including textual, structural, and behavioral characteristics. Ensemble learning and advanced machine learning models have demonstrated improved detection performance by capturing complementary signals across different feature domains. Despite these advancements, many studies still emphasize a primary feature modality, with limited integration of sender-specific information into content-focused models.

This study builds upon recent developments by explicitly separating sender and content analysis into two independent layers, allowing each to contribute uniquely to the final classification decision.

1.3 Overview of the Proposed Approach

To address the limitations of existing phishing detection systems, this study proposes a dual-layer machine learning framework. Rather than treating phishing detection as a single binary decision problem, the framework evaluates emails through two specialized analytical layers.

The first layer focuses on sender metadata, extracting structural features from email addresses to assess their legitimacy. The second layer analyzes the content of the email body using linguistic and statistical features. Each layer independently produces a probability score representing the likelihood that an email is phishing.

These scores are subsequently combined using a meta-classification model, which learns how to weigh sender and content information to produce a final prediction. This design enables the system to identify phishing attempts that might evade detection when only one type of information is considered.

2 RELATED WORK

Phishing detection has been extensively studied within the fields of cybersecurity and machine learning. Early detection systems primarily relied on rule-based methods, such as blacklists and keyword filtering. While straightforward to implement, these approaches proved vulnerable to evasion and required frequent manual updates [22].

More recent research has demonstrated the effectiveness of machine learning techniques for phishing detection [1] [2] [17]. Studies have reported high classification accuracy using algorithms such as Random Forest, Support Vector Machines, and gradient boosting methods. However, many of these studies evaluated their models on relatively small or artificially balanced datasets, limiting their applicability to real-world conditions.

Some researchers have focused on sender-centric detection methods, analyzing email headers and address structures to identify suspicious patterns [18]. Others have concentrated on content-based analysis, leveraging linguistic features to distinguish phishing emails from legitimate ones [21]. Hybrid approaches that combine multiple feature types have shown improved performance, particularly when ensemble learning techniques are applied [12] [16].

Recent advancements have also explored the use of feature selection and ensemble learning [6] [17] as well as transformer-based language models for phishing detection [7]. While these methods offer strong performance, they often require significant computational resources, which may limit their practicality in certain deployment scenarios.

The body of existing literature suggests that combining multiple analytical perspectives yields greater robustness. This study contributes to this area by proposing a structured dual-layer framework that integrates sender metadata and email content analysis while maintaining scalability and interpretability.

3 BACKGROUND

3.1 Phishing Attack Mechanism

Phishing is a social engineering attack in which attackers send deceptive emails that appear to originate from legitimate sources, such as financial institutions, technology companies, or employers. These emails often create a sense of urgency or authority to prompt immediate action.

Typically, a phishing email contains a link or attachment that directs the recipient to a fraudulent website designed to resemble a legitimate one. Once the user enters sensitive information, such as login credentials, the attacker captures and misuses this data. Because phishing attacks rely heavily on deception rather than malware, they can be difficult to detect using traditional security mechanisms.

3.2 Detection Challenges

Detecting phishing emails is increasingly difficult due to advances in attack sophistication. Many phishing emails now use realistic sender addresses, professional formatting, and carefully worded messages. Malicious links may be hidden behind legitimate-looking text, and attackers often adapt their strategies to bypass known detection rules.

Additionally, the imbalance between legitimate and phishing emails complicates machine learning training. Models may become biased toward predicting the majority class, resulting in high overall accuracy but poor phishing detection performance. These challenges highlight the need for detection systems that are both robust to evasion and effective under realistic data conditions.

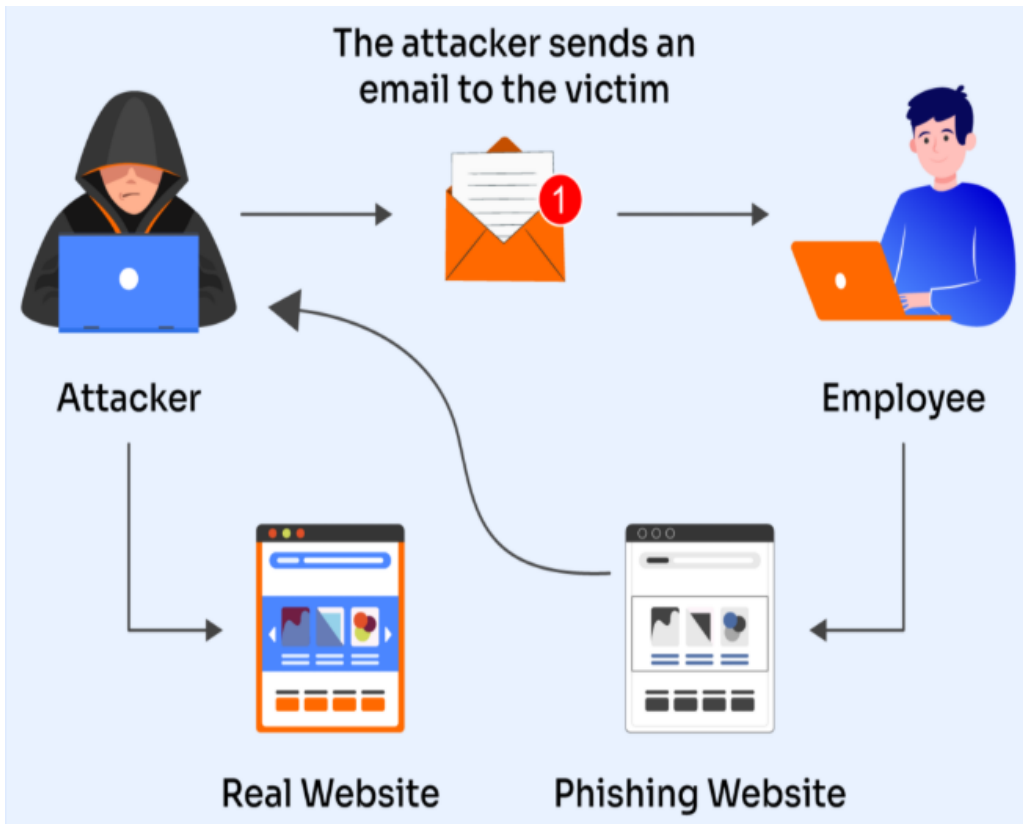


Figure 2: Phishing attack steps: from fake email creation to victim data theft [20].

4 DATASETS

To evaluate the proposed dual-layer framework, three publicly available email datasets were utilized. Rather than training separate models on each dataset, they were merged to create a comprehensive dataset representative of real-world email traffic.

The first dataset contains 39,154 phishing emails compiled from public sources [3]. The second dataset includes 38,810 labeled phishing and legitimate emails [4]. The third and largest dataset consists of 447,417 emails from the Enron Fraud Email Dataset [15]. These datasets were selected to provide diversity in email structure, content, and sender information.



Figure 3: Example of one of the datasets that our model trains on.

After preprocessing and integration, the final dataset comprised 524,735 emails. Of these, 479,710 (91.42%) were legitimate and 45.025 (8.6%) were phishing, closely mirroring real-world email distributions. This highly imbalanced dataset presents a challenging but realistic environment for evaluating phishing detection performance and ensures that results reflect practical deployment conditions.

5 SYSTEM IMPLEMENTATION

5.1 System Architecture

The proposed phishing detection system is implemented as a dual-layer machine learning framework designed to address the limitations of traditional single-layer approaches. Rather than relying on a single set of features or rules, the system independently analyzes two distinct components of an email: sender metadata and email body content.

The first layer focuses on sender-related information extracted from the email address. Structural features such as domain characteristics, username patterns, and character composition are used to train a machine learning model. This model outputs a continuous probability score between 0 (legitimate) and 1 (malicious), representing the likelihood that an email is phishing based solely on sender metadata.

The second layer analyzes the email body by extracting linguistic and statistical features from the text. Multiple machine learning models are trained on these features, and their outputs are aggregated into a single content-based probability score.

The final stage of the system integrates the scores produced by both layers to generate a comprehensive phishing prediction. This modular architecture enables the system to detect phishing attempts that may evade detection when only one type of information is considered.

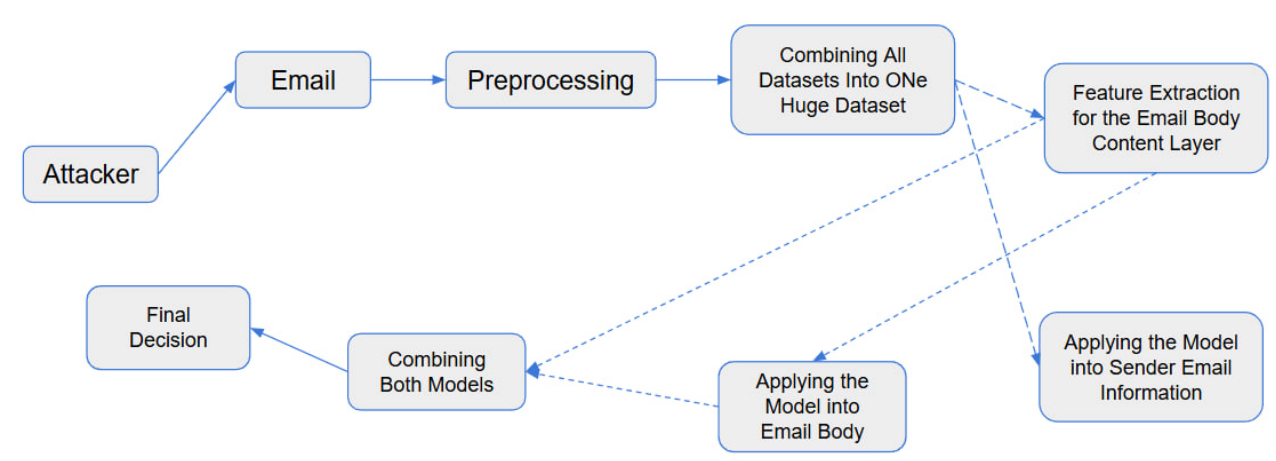


Figure 4: System Architecture of Our Dual-Layer Detection System

5.2 Data Preprocessing

Data preprocessing is a critical step in preparing raw email data for machine learning analysis. Email content often contains inconsistent formatting, hyperlinks, email addresses, numerical values, and special characters that can interfere with feature extraction.

To ensure consistency, all text is converted to lowercase. Hyperlinks and email addresses are removed, as they are highly variable and do not contribute meaningfully to textual pattern recognition. Non-alphabetic characters are eliminated, leaving only letters and spaces. Excess whitespace is removed to maintain uniform token separation.

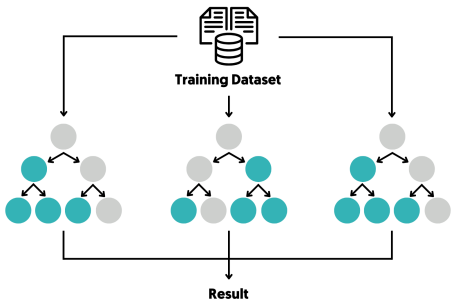


Figure 5: Structure of how Random Forest, XGBoost, and LightGBM work [10].

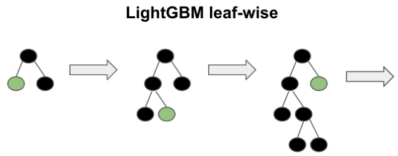


Figure 6: Structure of how LightGBM works [5].

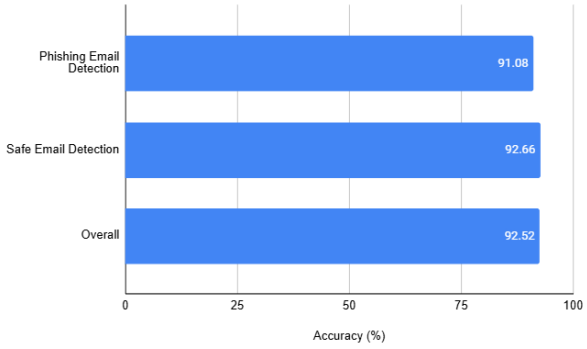


Figure 7: The final Detection Results.

In cases where words are improperly concatenated, a word segmentation tool is applied to restore meaningful word boundaries. After cleaning, the processed text is reconstructed into a single structured string and saved to a separate file. This preprocessing pipeline ensures that the data is standardized and suitable for subsequent feature extraction and model training.

5.3 Dataset Integration

Following preprocessing, the cleaned datasets are merged into a single unified dataset. Combining multiple datasets increases the diversity and volume of training data, allowing the models to learn more generalized patterns.

Each dataset is first standardized to ensure consistent formatting. Sender information is reconstructed where necessary by combining username and domain components, and email addresses embedded within text are extracted. Column names are normalized across datasets to maintain a consistent schema, including sender information, cleaned email content, and phishing labels.

Once standardized, the datasets are concatenated into a single comprehensive dataset. The resulting dataset is stored for use in training both the sender-based and content-based detection layers.

5.4 Email Body Feature Extraction

To enable machine learning models to analyze email content, textual data must be transformed into numerical features. A set of statistical and linguistic features is extracted from each email body.

These features include total word count, number of unique words, and average word length. Lexical diversity is calculated as the ratio of unique words to total words, providing insight into writing style variability. Word repetition metrics are also computed, as phishing emails often rely on repeated keywords.

Additionally, the system identifies and counts commonly used phishing-related terms, such as “verify,“ “account,“ and “urgent.“ The proportion of these keywords relative to the total word count is used as an indicator of suspicious content. All extracted features are combined into a numerical feature vector and saved for model training and evaluation.

5.5 Sender-Based Model Implementation

The sender-based detection layer is implemented using a Random Forest classifier, an ensemble learning method that aggregates predictions from multiple decision trees to improve accuracy and robustness [8].

Emails with missing sender information or labels are excluded from training. Each sender address is divided into username and domain components, which are normalized to lowercase. A variety of structural features are extracted, including length measurements, digit and special character counts, character repetition patterns, entropy values, and the presence of phishing-related keywords.

The dataset is split into training (70%) and testing (30%) subsets to evaluate model performance. The trained Random Forest model outputs both a binary classification and a probability score representing the likelihood that an email is phishing based on sender metadata. These outputs are stored for later integration with the content-based layer.

5.6 Content-Based Model Implementation

The content-based detection layer utilizes multiple machine learning algorithms, including Random Forest, XGBoost, and LightGBM. These models are selected due to their strong performance on structured feature sets and large-scale datasets [9] [13].

Using the extracted email body features, the dataset is divided into training and testing sets. Each model learns patterns associated with phishing behavior based on linguistic and statistical characteristics of the email content. After training, each model generates a probability score for every email.

To enhance prediction stability, the final content-based suspicion score is computed as the average of the scores produced by the three models. This ensemble approach reduces model-specific bias

and improves overall reliability. The resulting scores are saved for use in the final classification stage.

5.7 Dual-Layer Score Integration

The final classification stage integrates the sender-based and content-based probability scores using a LightGBM meta-model. This model learns how to optimally combine the two scores to produce a final phishing prediction.

The combined dataset is divided into training and testing subsets for meta-model evaluation. LightGBM is particularly effective for this task due to its ability to model non-linear relationships between features. This allows the system to account for cases in which one layer indicates high risk while the other appears benign.

The meta-model outputs a final probability score for each email. A range of classification thresholds is evaluated to identify the value that best balances phishing detection accuracy and false positive rates. Emails with scores exceeding the selected threshold are classified as phishing, while those below are labeled as legitimate. This integrated approach enhances detection robustness under realistic, highly imbalanced conditions.

6 RESULTS

The proposed dual-layer phishing detection framework was evaluated on a large, real-world dataset consisting of 524,735 emails, of which approximately 8.6% were phishing messages. This highly imbalanced distribution closely reflects real-world email traffic and presents a challenging environment for machine learning models.

The system achieved an overall classification accuracy of 92.52%. The phishing email detection rate was 91.08%, indicating that the majority of malicious emails were correctly identified. The safe email detection rate reached 92.66%, demonstrating strong performance in accurately classifying legitimate messages.

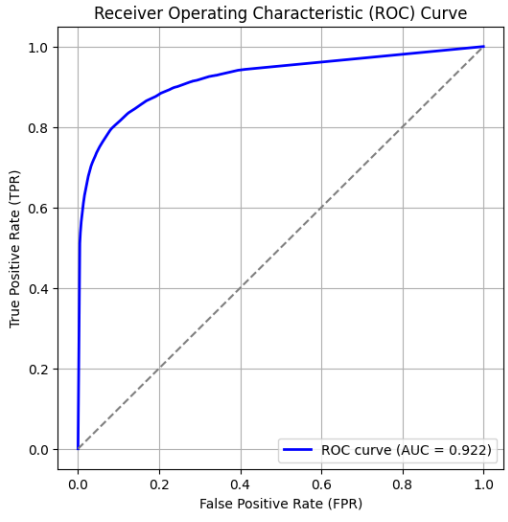


Figure 8: The ROC curve shows how well the model can distinguish between phishing and safe emails. With an AUC score closer to 1, the model demonstrates high effectiveness in separating the two classes.

Although several prior studies have reported higher accuracy values—often exceeding 95%—these results are typically obtained using much smaller datasets with artificially balanced class distributions [1] [2] [19]. In contrast, the dataset used in this study is significantly larger and more imbalanced; the results obtained in this work are more representative of real-world and operational performance [3] [4] [15].

6.1 Feature Group Impact and Ablation Discussion

To better understand the contribution of different feature groups, additional analysis was conducted on feature importance within the trained models. Results from the Random Forest classifier indicate that URL based features—such as encoded links— and sender-related attributes—such as domain length, structural irregularities, and suspicious keyword presence—exert a strong influence on phishing detection performance.

Content-based linguistic features, including keyword density, word repetition, and lexical diversity, also contributed meaningfully to classification accuracy. While these features were individually less dominant than some sender-based features, they provided complementary information that improved detection when combined.

This analysis supports the design choice of a dual-layer architecture. By integrating sender metadata and email content features, the framework captures distinct and largely independent signals, leading to improved robustness under highly imbalanced conditions.

6.2 Scalability and Latency Considerations

Scalability and computational efficiency are critical factors for real-world email security systems. The proposed framework was designed with modularity in mind, allowing each layer to operate independently and enabling future extensions.

During evaluation, batch processing experiments demonstrated that the system could analyze approximately 2,000 emails per second on standard hardware. The average classification time per email remained below 0.5 seconds. These results suggest that the framework can scale linearly with dataset size and is suitable for near real-time deployment in enterprise or institutional email environments.

7 COMPARISON TO OTHER METHODS

When evaluated on the same large, imbalanced dataset, several conventional machine learning models exhibited notable performance degradation, as illustrated in Figure 10. This decline highlights a common limitation of single-layer approaches when applied to realistic email distributions.

In contrast, the proposed dual-layer framework maintained stable performance across both phishing and legitimate email classes. By independently analyzing sender metadata and email content, the system reduces reliance on any single feature group and mitigates the risk of evasion strategies targeting only one layer.

It is important to note that differences in reported performance across studies are influenced by dataset composition, feature selection, and evaluation methodology. Therefore, the results of this study should be interpreted as competitive rather than as a claim

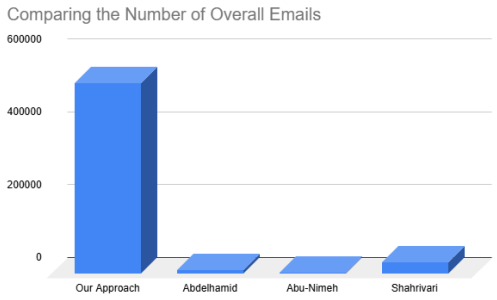


Figure 9: Comparing the Number of Overall Emails [1] [2] [19].

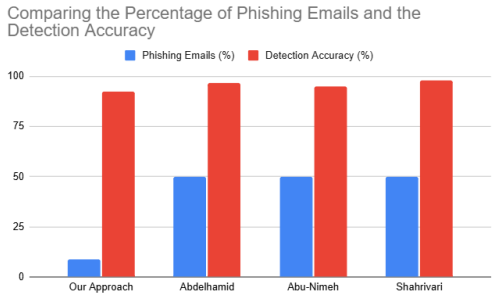


Figure 10: Comparing the Percentage of Phishing Emails and the Detection Accuracy [1] [2] [19].

of superiority. Nonetheless, the framework demonstrates clear advantages in robustness and practical applicability under realistic conditions.

8 LIMITATIONS AND CHALLENGES

Despite its strengths, the proposed framework has several limitations that warrant discussion. First, many publicly available email datasets lack complete metadata, restricting the depth of sender-based analysis [3] [4] [15]. As a result, the sender layer relies primarily on structural features of email addresses rather than richer contextual information [18].

Second, certain phishing attacks—often referred to as zero-day phishing—may originate from legitimate-looking or compromised accounts. In such cases, both sender metadata and content features may appear benign, increasing the likelihood of false negatives [11].

False positives also remain a challenge. Some legitimate emails may contain unusual formatting or structural characteristics that resemble phishing patterns, leading to incorrect classification [18]. Additionally, the content-based models rely on relatively lightweight linguistic features. Due to resource constraints, large language models were not employed, limiting the system's ability to interpret nuanced language, sarcasm, or multilingual phishing attempts [7].

Finally, the current meta-classification approach uses a single LightGBM model. While effective, more complex ensemble strategies and neural meta-learners may further improve performance [12] [16].

9 FUTURE WORK

Future research will focus on enhancing both detection accuracy and practical usability. Additional sender-related features—such as message timestamps, recipient counts, and historical communication patterns—may provide valuable contextual information for identifying sophisticated phishing attempts.

The integration of machine learning algorithms, such as Decision Trees and lightweight deep learning architectures may achieve higher accuracy and robustness. Advanced natural language processing models, including transformer-based architectures such as BERT or RoBERTa, may improve the system's ability to detect subtle and context-dependent phishing content. Expanding support for multilingual emails and image-based phishing is also a priority.

From an application perspective, future iterations of the system may be deployed as a web-based tool or browser extension, enabling users to assess email authenticity in real time. User feedback and real-world testing would further inform system improvements and usability enhancements.

10 CONCLUSION

This study presented a dual-layer machine learning framework for phishing email detection that integrates sender metadata analysis with email content analysis. By treating these components independently and combining their outputs through a meta-model, the system achieves strong performance under realistic, highly imbalanced conditions.

Evaluation on a large real-world dataset demonstrates that the dual-layer approach improves robustness and reduces reliance on any single feature source. The results indicate that combining multiple analytical perspectives is more effective than traditional single-layer detection methods.

Overall, this work provides a practical and scalable approach to phishing detection and contributes to ongoing efforts to improve email security in real-world environments. Future enhancements may further strengthen the framework's adaptability and detection capabilities.

REFERENCES

[1] Abdelhamid, N., Ayesh, A., Thabtah, F.: Phishing detection based associative classification data mining. Expert Systems with Applications **41**(13), 5948–5959 (2014)
[2] Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S.: A comparison of machine learning techniques for phishing detection. In: Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit. pp. 60–69 (October 2007)
[3] Alam, N.A.: Phishing email dataset. Kaggle (May 24 2024), https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset/data
[4] Amgain, P.: Phishing and legit emails .csv file. Kaggle (June 8 2025), https://www.kaggle.com/datasets/prabhatamgain/phishing-and-legit-emails
[5] Ashish Saini: Unleash the power of lightgbm in python 3: Your path to machine learning master. https://innovationyourself.com/lightgbm-in-machine-learning/ (2023), accessed: 2025-10-28
[6] Basnet, R.B., Sung, A.H., Liu, Q.: Feature selection for improved phishing detection. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. pp. 252–261. Springer Berlin Heidelberg, Berlin, Heidelberg (June 2012)

[7] Bhowmick, A., Hazarika, S.M.: Machine learning for e-mail spam filtering: review, techniques and trends. arXiv preprint arXiv:1606.01042 (2016)
[8] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
[9] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 785–794 (August 2016)
[10] Dida.do: What is random forest? https://dida.do/what-is-random-forest (2025), accessed: 2025-09-08
[11] Divakaran, D.M., Oest, A.: Phishing detection leveraging machine learning and deep learning: A review. arXiv (May 16 2022), https://arxiv.org/abs/2205.07411
[12] Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S.B., Joga, S.R.K.: Phishing detection system through hybrid machine learning based on url. IEEE Access **11**, 36805–36822 (2023)
[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems **30** (2017)
[14] Petrosyan, A.: Number of global phishing attacks 2024. Statista (April 23 2025), https://www.statista.com/statistics/266155/number-of-phishing-attacks-worldwide/
[15] Rao, A.S.: Enron fraud email dataset. Kaggle (December 28 2023), https://www.kaggle.com/datasets/advaithsrao/enron-fraud-email-dataset
[16] Rao, R.S., Kondaiah, C., Pais, A.R., Lee, B.: A hybrid super learner ensemble for phishing detection on mobile devices. Nature (May 15 2025), https://www.nature.com/articles/s41598-025-02009-8
[17] Sahoo, D., Liu, C., Hoi, S.C.: Malicious url detection using machine learning: A survey. arXiv preprint arXiv:1701.07179 (2017)
[18] Sanchez, F., Duan, Z.: A sender-centric approach to detecting phishing emails. In: 2012 International Conference on Cyber Security. pp. 32–39. IEEE (December 2012)
[19] Shahrivari, V., Darabi, M.M., Izadi, M.: Phishing detection using machine learning techniques. arXiv (2020), https://arxiv.org/abs/2009.11116
[20] Valimail: Phishing prevention best practices: A guide. https://www.valimail.com/resources/guides/guide-to-phishing/phishing-prevention-best-practices/ (2025), accessed: 2025-09-08
[21] Verma, R., Shashidhar, N., Hossain, N.: Detecting phishing emails the natural language way. In: European Symposium on Research in Computer Security. pp. 824–841. Springer, Berlin, Heidelberg (September 2012)
[22] Zuraiq, A.A., Alkasassbeh, M.: Phishing detection approaches. In: 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS). pp. 1–6. IEEE (October 2019)

Received 12 September 2025; Accepted 26 December 2025