Towards Trustworthy Email Phishing Detection: Integrating Multi-Modal Deep Learning, Federated Learning, and Explainable AI

Adetoye A. ADEYEMO Adetoye.adeyemo@tech-u.edu.ng
Department of Computer Engineering
Abiola Ajimobi Technical University
Ibadan, Oyo State, Nigeria

Ozichi N. EMUOYIBOFARHE Ozichi.emuoyibofarhe@bowen.edu.ng
Computer Science Programme
Bowen University
Iwo, Osun State, Nigeria

Adeyinka O. ABIODUN Aabiodun@noun.edu.ng
Africa Centre of Excellence on Technology Enhanced Learning
National Open University of Nigeria
Abuja, Nigeria

James O. Adegboye Olujoba.adegboye@federalpolyilaro.edu.ng
Department of Computer Science
Federal University of Technology
Ilaro, Nigeria

Sunday A. AJAGBE Ssaajagbe@pgschool.lautech.edu.ng
Department of Computer Engineering
Ladoke Akintola University of Technology
Ogbomosho, Oyo State, Nigeria

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Email phishing remains one of the major cybersecurity threats because it uses misleading strategies which keep evolving. This research presents a robust and standard phishing detection system that integrates multi-modal deep learning (DL) with federated learning (FL) and explainable artificial intelligence (XAI) to achieve better detection results while keeping user information secure. The proposed model uses a DL architecture which combines email text analysis with URL structural feature based on embedding, LSTM and feature fusion layers. Experiments were conducted using centralized and FL settings. In the centralized setup, the model recorded high performance with an accuracy of 99.6%, precision of 99.8%, recall of 99.5%, F1 score of 99.72% and AUC of 99.85% values surpassing 98%, benefiting from direct access to the complete training dataset. In contrast, the federated model obtained results that were slightly lower than the centralized system results but maintained competitive performance across all testing metrics while proving its ability to generalize from training data that was distributed among private locations. The

performance difference between the two systems results I attributed to two factors which include restricted global system access and the use of periodic data aggregation techniques in FL. The researcher made use of Explainable AI techniques which are LIME and perturbation analysis to show how the model used specific words to determine whether emails were phishing attempts or legitimate messages. It was established that centralized learning produced slightly higher metrics. However, FL offers a better solution because it protects user privacy while enabling effective detection of email phishing attacks.

Keywords: Email Phishing Detection, Multi-modal Deep Learning, Federated Learning, Explainable Artificial Intelligence, Privacy Preserving Machine Learning, LSTM Based Text Classification, URL Structural Feature Analysis, Cybersecurity Threat Detection.

1. Introduction

Email phishing has consistently been one of the main pervasive and dynamic cyber threat, posing serious risks to individuals and organizations all around the world. Phishing attacks are known for their deceiving efforts to obtain delicate and sensitive information such as usernames, passwords and credit card details by employing social engineering methods to manipulate their subtle targets [13]. Phishing attacks create both financial losses and reputation damage which businesses must guide against through effective and flexible detection systems [7]. The traditional phishing detection methods that rely on signature-based and basic rule encounters challenges in detecting contemporary phishing attacks which have become more sophisticated and unpredictable [3]. This concern has challenged the cybersecurity community to look into a more advanced artificial intelligence (AI) technologies, particularly deep learning as a solution to create an enhanced cybersecurity defences.

The introduction of DL model has brought about major changes to different areas of study which include natural language processing (NLP) and computer vision because it allows the models to extract complex patterns from broad datasets [9]. The field of cybersecurity has benefited from DL models and this has provided an effective method to detect hidden phishing pointers that traditional detection methods fail to pick up. Precisely, multi-modal DL that combine information from different data sources, holds a tremendous promise for improving phishing detection accuracy. Email phishing attacks often use various modalities, which includes textual content, hidden links and visual elements. Multi-modal DL models can create a more general understanding of an email's credibility by analyzing simultaneously these various features, thereby improving detection competencies [19].

Sensitive areas such as email security are faced with new difficulties because organizations use strong DL models for their operations, which create new problems about protecting user data (privacy) and understanding how the system works (interpretability). Training of precise DL models normally requires having access to large and centralized datasets which can presents difficulties due to privacy laws like GDPR and organizational unwillingness to share sensitive email data [6]. FL offers more effective answer to the existing problem. FL functions as a distributed ML system that allows several clients to jointly develop models without sharing their local data with each other. The system protects user information by sending only model parameters and gradients to the central server and this enables users to access shared intelligence resources [8]. The privacy protection characteristics of FL makes it a suitable technology for processing sensitive use cases that requires strict data protection as seen in phishing detection applications.

Also, the growing complexity and hidden operation of DL models has created difficulties for their deployment and usage in critical decision-making processes. The ability to explain model reasons for classifying an email as phishing, allow organizations to build user trust and then improving incident response capabilities and discovering new attack methods. This is where Explainable AI (XAI) plays a critical role. XAI techniques create models that describes their decision-making processes in a way that humans can understand their working process better [5]. By integrating XAI into multi-modal DL and FL frameworks for phishing detection researchers will be able to develop systems that are accurate and privacy-preserving but also interpretable and trustworthy. The detailed method presents a solution for different challenges which modern phishing detection systems face, leading to the development of email security systems that highlight both security and user experience.

The manuscript presents a novel email phishing detection method that make use of combined multi-modal DL with FL and explainable AI. Our system development uses the advantages of each system element to create a security solution that maintains user privacy and also providing accessible information system. The multi-modal DL system analyzes different email characteristics which include text content and URL parameters to build a comprehensive model that detects possible security threats. FL allows the model to undergo joint training by different organizations at the same time keeping all user information secure. Finally, the model utilizes explainable AI (XAI) methods which allows for human-readable predictions to build trust between users and the system and also improving threat assessment capabilities. The subsequent sections of this write-up presents a complete overview of current research which will show the technological growth made so far while our framework research aims to address specific scientific research lapses.

2. Literature Review

The field of email phishing detection has experienced crucial changes because cybercriminals have developed more advanced attack approaches and AI technology has reached new heights of development. This section presents a comprehensive review of the literature concerning multi-modal DL, FL and explainable AI in the context of phishing detection and highlighting key methodologies, findings and identified research gaps.

2.1. Multi-modal Deep Learning for Phishing Detection

Early phishing detection methods used single detection techniques that inspects email headers, email text and URL characteristics [1]. The methods was successful some cases but they failed to handle phishing attacks which used different communication methods together with various misleading techniques. The implementation of DL technology allows for advanced analysis through Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) which include Long Short-Term Memory (LSTM) networks. The networks establish superior ability to capture complex patterns which exist in both textual data and sequential data [12].

Recent research has emphasized the benefits of multi-modal DL for phishing detection. The multi-modal models can produce a complete email profile through their ability to combine different sources of data which include email body text, embedded URLs, sender

information and visual elements. For example, research have shown that linking LSTM-processed textual features with URL-based features can significantly increase detection accuracy than the single-modality methods [16].

Vulfin et al. in their research, came up with a multi-modal phishing website detection system which uses multi-modal analysis to improve its detection capabilities [17]. Similarly, Murhej and Nallasivan developed a multi-modal framework which incorporate EM-BERT for text analysis and SPCA for image and CSS feature extraction to achieve high accuracy results in multiple datasets [11].

All the above studies confirmed that different data types need to be investigated so that researchers can exceptionally fight the complex nature of phishing attacks. The process of bringing together different data types encounters difficulties and the system needs to handle several types of data in real-time for advanced multi-modal models.

2.2. Federated Learning for Privacy-Preserving Phishing Detection

The need for privacy-preserving ML has become important, especially in areas where there is a need to deal with sensitive user data like email communications. The traditional approach to centralized training needs the collection of various data, which must be stored on one main server, which causes major privacy issues and creating regulatory problems [6]. The FL system provides a decentralized solution that enables various clients to work together in developing a common global model while keeping their individual data secure [8]. This method proves to be especially effective for detecting phishing attacks where organizations are reluctant to provide their private email data which they consider sensitive and competitive information.

Multiple research studies have investigated the use of FL technology for cybersecurity purposes which include detecting spam and malware threats. Thapa et al. conducted one of the pioneering investigations into the use of FL for phishing email detection by comparing its performance to centralized learning which used RCNN and BERT models [13]. Their research results showed that FL can reach the same results as centralized methods when researchers use balanced datasets for testing. The researchers identified two main problems which emerged when more clients connected to their system as well as when their system encountered data that had extremely uneven distribution patterns. Li et al. presented FedPhishLLM which functions as a privacy-protecting and explainable phishing detection solution that merges FL with large language models to achieve enhanced privacy protections and better system performance. [10]. While FL provides a secure method for data protection, however, researchers need to solve three existing research problems which include solving non-IID data issues and creating effective aggregation methods and developing techniques to protect against adversarial attacks in federated systems. [18].

2.3. Explainable AI for Transparent Phishing Detection

The surge in the usage of complex DL models in essential systems requires researchers to develop methods that explain their functionality. The black-box behavior of various deep learning models creates three major problems because it prevents users from establishing trust in the system and makes debugging difficult and limits security analysts' ability to comprehend and respond to identified threats. Explainable AI (XAI) techniques aim to

bridge this gap by providing human-understandable insights into the reasoning behind AI model predictions [5]. The XAI system for phishing detection allows users and security personnel to see the reasons behind an email's classification as harmful, which builds trust in the system and supports better decision-making.

Researchers have examined different methods of XAI implementation for cybersecurity applications which include local interpretable model-agnostic explanations (LIME) and Shapley Additive Explanations (SHAP). Uddin et al. developed an explainable machine learning system for detecting phishing websites which utilizes ensemble learning methods (XGBoost LightGBM GBM) together with SHAP to show how their model makes predictions [14]. Alotaibi et al. created a web phishing classification system which uses XAI technology to improve both its interpretability and its ability to classify better [2]. The studies demonstrate that XAI technology improves trustworthiness and helps users discover essential elements which drive phishing identification, which will assist in creating effective defense strategies. The development of XAI methods which can perform computationally efficient real-time tasks while handling adversarial attacks and delivering useful explanations for multi-modal and FL systems remains an ongoing problem [15]. Alsuqayh et al. carried out a systematic literature review on feature engineering and XAI. The researchers identified lapses in transparency and user trust in current AI-based phishing detection systems [4].

The current literature shows clearly a trend toward cutting-edge and all-inclusive methodology for detecting phishing attacks. Multi-modal DL provides improved accuracy through its ability to process multiple data types, while FL is used for necessary privacy solutions that protect confidential email information. Explainable AI methods have become essential tools that establish trust and enable users to understand complicated model outcomes. Despite these advancements, multiple research gaps still exist because researchers need to create effective FL methods which work with non-IID data and they must develop systems which can handle multiple types of data while maintaining real-time capabilities and researchers need to create XAI methods which can interpret the combined multi-modal and federated model decisions. Our proposed research aims to contribute to filling these gaps by presenting a unified framework that leverages the synergistic benefits of these three critical areas.

3. Methodology

The research develops a privacy protective email phishing detection system for human users through its implementation of multi modal deep learning together with federated learning and explainable artificial intelligence. The methodology covers data preparation, feature representation, model design, centralized training, federated training and explainability analysis.

3.1. Dataset Preparation and Preprocessing

A publicly available phishing email dataset was used for this study. The dataset includes records that contain email text elements which show sender and receiver information and the subject and body content and structural elements that count the total embedded URLs. The target label indicates whether an email is benign or phishing.

Table 1: Table of Past Research

Ref	**Methodology**	**Results**	**Research Gap Addressed/Identified**
[1]	FL with RCNN (THEMIS) and BERT models; FedAvg aggregation.	Comparable performance to centralized learning; RNN accuracy decreased by 1.8	
[13]	Multi-modal (EM-BERT for text/URL, SPCA for images/CSS); EAI-SC-LSTM for classification.	Accuracies of 99.6	
[16]	Ensemble learning (XGBoost, LightGBM, GBM) with SHAP for explainability.	High efficiency and transparency in predictions.	Traditional methods lack transparency and fail on novel phishing types.
[17]	Aquila Optimization Algorithm with XAI (XAIAOA-WPC).	Improved classification performance and interpretability.	Need for better optimization and interpretability in high-dimensional phishing data.
[14]	FL with Large Language Models (LLMs) and explainability features (Fed-PhishLLM).	Superior privacy and performance compared to baselines.	Privacy concerns in centralized LLM-based phishing detection.
[12]	Multi-modal analysis for web resource detection.	Improved efficiency through multi-source feature fusion.	Efficiency in real-time multi-modal analysis.
[19]	Systematic literature review on feature engineering and XAI.	Identified gaps in transparency and user trust in current AI-based phishing detection systems.	Transparency and user trust in AI-based phishing detection systems.

The researchers conducted a process which involved cleaning all textual fields before merging them together to form one unified text stream. Tokenization together with sequence padding were then applied to convert the text into standardized fixed length numerical formats. The team then performed normalization on numerical features that needed this process. The dataset was divided into three parts which included training, validation and testing data through a stratified method that maintained equal distribution of classes.

3.2. Multi Modal Feature Representation

Two complementary modalities were employed to improve detection performance:

Text modality: The combined email text was processed through an embedding layer. The text was then sent to a Long Short Term Memory (LSTM) network. The model uses this method to detect phishing email patterns which include urgent messages and fake offers and persuasive writing.

URL modality: The numerical URL feature was processed through a fully connected layer to create structural indicators which identify malicious emails.

The outputs of both modalities were joined together using feature concatenation and then processed in dense layers for final binary classification.

3.3. Centralized Learning Framework

The centralized learning system provided complete access to its training dataset through a single server. The multi modal deep learning model was trained end to end, using binary cross entropy loss and the Adam optimizer. The model make use of the standard evaluation metrics which are accuracy, precision, recall, F1 score and AUC. The researchers then observed the validation accuracy throughout the epochs to assess both the convergence and learning process stability.

3.4. Federated Learning Framework

The solution to data privacy concerns requires the use of FL technology. The training data was distributed across diverse simulated clients. Each client developed a comparable multi modal model through local training which used their local data. The trained model parameters were transmitted to a central server after local training finished and the server used FedAvg strategy to aggregate the data. The entire process operated without any exchange of actual email information.

3.5. Explainability Analysis

The implementation of XAI methods established better transparency which helped build user's trust. The system used Local Interpretable Model-agnostic Explanations (LIME) to identify the specific words that affected each prediction result. The system employed a perturbation analysis method to reduce tokens from email text and measure the resulting impact on prediction accuracy. The methods used in this study produce findings which humans can understand as they reveal how defenders identify legitimate threats and phishing attacks.

3.6. System Architecture

Figure 1 presents the Centralized model architecture. The architecture illustrates a three dimensional view of the centralized multi modal phishing detection model showing the text embedding and LSTM branch fused with the URL feature branch followed by fully connected layers for classification.

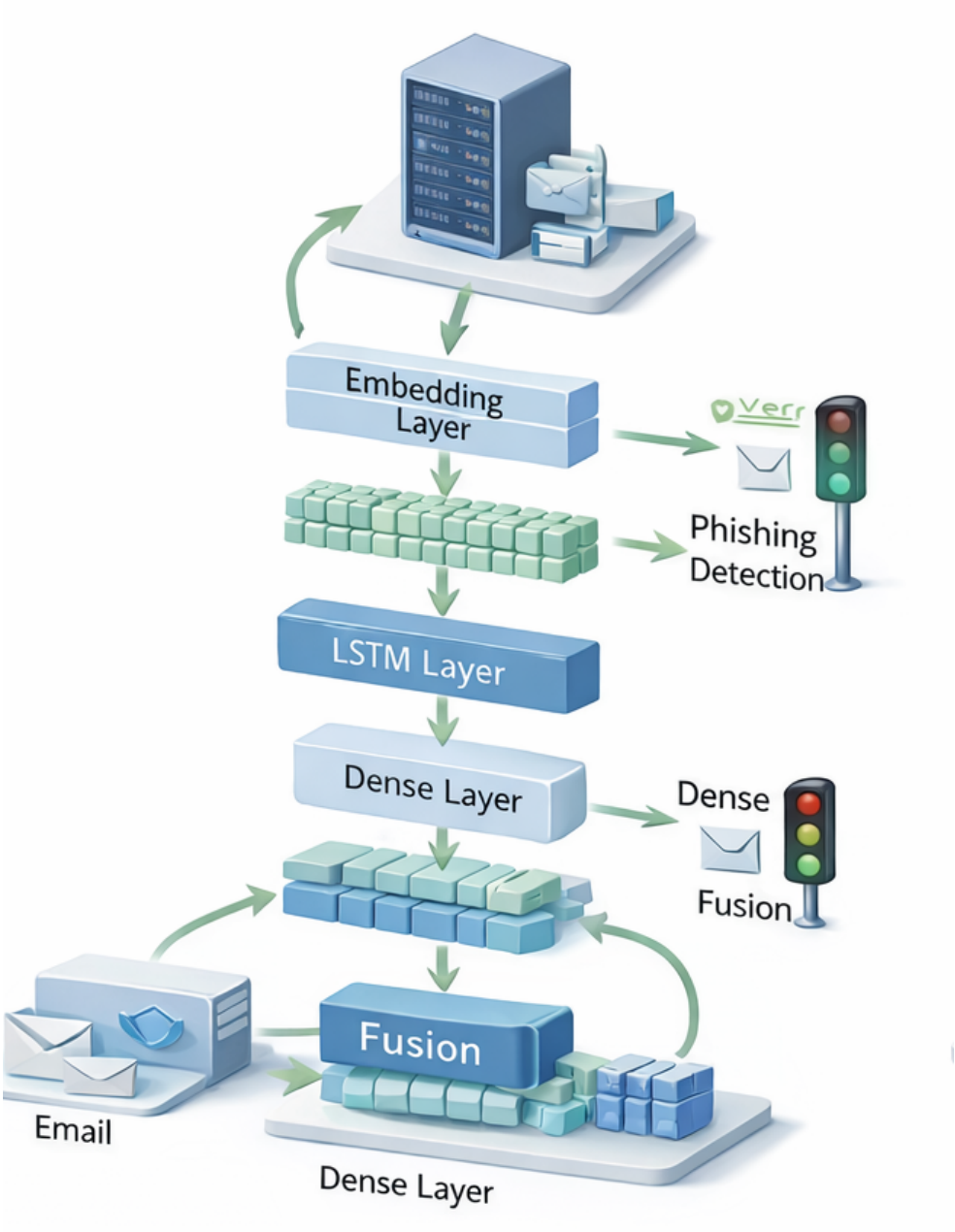

Figure 1: Centralized Model Architecture

Figure 2 presents the FL architecture. The architecture shows three dimensions of multiple client nodes which train the same multi modal model at their locations while they send updates to a central server through the Federated Averaging (FedAvg) algorithm.

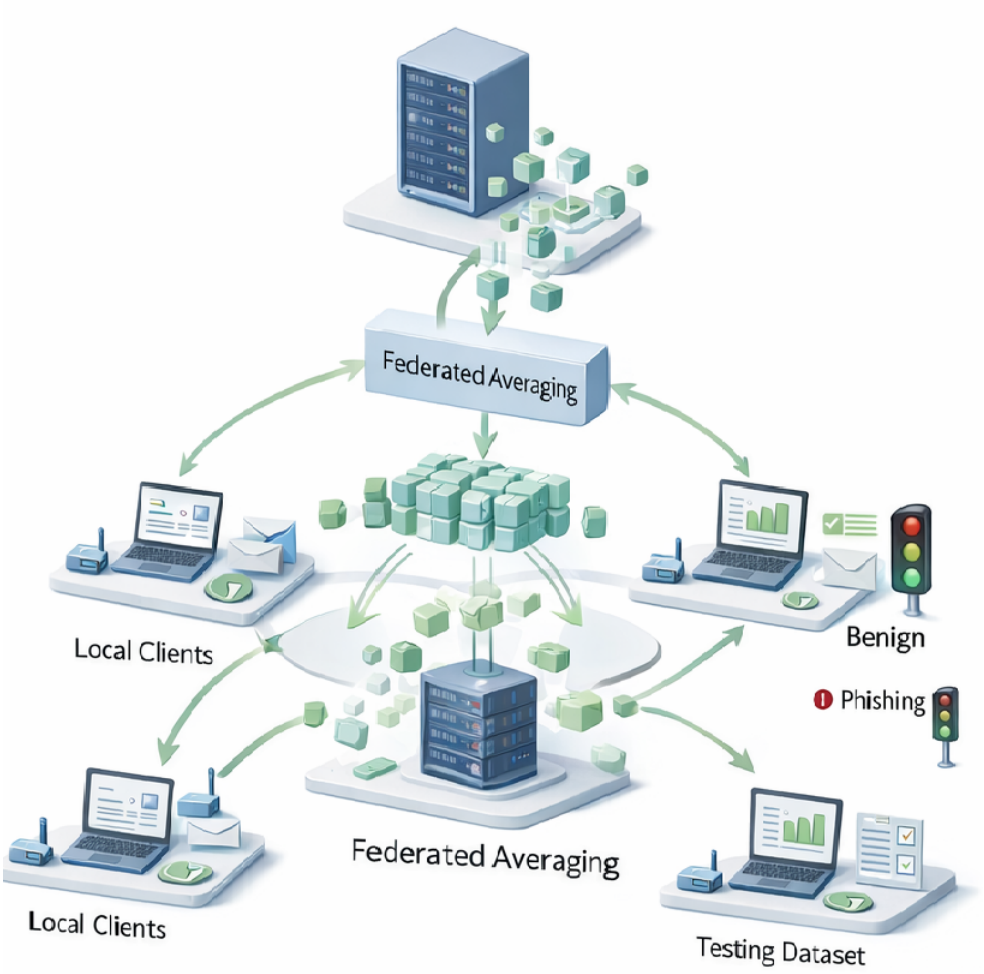

Figure 2: Federated Learning Architecture

3.7. Epoch Configuration

In the centralized training setup, the proposed multi modal deep learning model was trained on the entire training dataset for 10 epochs using a batch size of 64. This allowed the model to learn global patterns from the aggregated email data through repeated exposure to all training samples. Model performance was evaluated after each epoch using the validation set to ensure stable convergence and to prevent overfitting.

The FL experiment maintained identical model architecture throughout the testing process in order to maintain consistency. The training method was spread across five clients that possess different quotas of the training data. During each communication round, every client trained the model locally for one epoch using its private data. Once local training

was completed, the updated model parameters were sent to a central server, where they were combined using the FedAvg algorithm.

The federated training process was conducted through eight global communication rounds, which resulted in eight effective training epochs when considering the interaction between local updates and global aggregation. The system provides an equal trade-off between client learning depth and server synchronization rate which makes federated training comparable to centralized training methods.

To ensure a fair comparison between centralized and FL, the experiments needed to maintain identical hyperparameter settings which included batch size, optimizer, loss function for both centralized and FL testing. The Adam optimizer was used in both settings to provide stable and adaptive parameter updates during training.

Table 2: Epoch Configuration Parameters

Parameter	**Symbol**	**Value Used**	**Explanation**
Global Communication Rounds	R	8	Number of aggregation rounds perform
Local Epochs per Client	E	1	Each client trains locally for one epoch
Total Effective Epoch	R*E	8	Approximate training depth achieved i
Batch Size	-	64	Maintained across both training setups
Optimizer	-	Adam	Provides adaptive and stable learning u
Loss Function	-	Binary Cross Entropy	Used for binary phishing classification.

4. Results and Discussions

This section shows and discusses the experimental results which were obtained from the proposed multi modal DL framework in both centralized and FL settings. The study examines the classification performance learning patterns and explainability insights that the XAI methods provide. The results are presented in a series of figures which together demonstrate the effectiveness, robustness and interpretability of the proposed approach.

4.1. Confusion Matrix Analysis

Figure 3(a) and Figure 3(b) in this section are confusion matrices for the centralized and FL models respectively.

In the centralized setting shown in Figure 3(a), The model exhibits strong performance in identifying both benign and phishing emails through its testing results which show high true positive with 99.85% and true negative rates of 99.57%. The multi modal architecture shows effective performance in detecting phishing attempts because it achieves low misclassification rates of false positive of 0.15% and false negative of 0.43% through its ability to identify both semantic and structural elements of phishing attempts.

Figure 3(b) shows the confusion matrix which tests the performance of the FL model. The centralized model shows better results than the FL model because it produces more misclassifications (false positive of 0.46% and false negative of 1.04%) according to the data. This result established the fact that FL preserves most of the discriminative power of centralized training despite operating under strict data privacy constraints. Importantly, no

raw email data was distributed during the training, highlighting the practical applicability of the FL approach in real world email security environments.

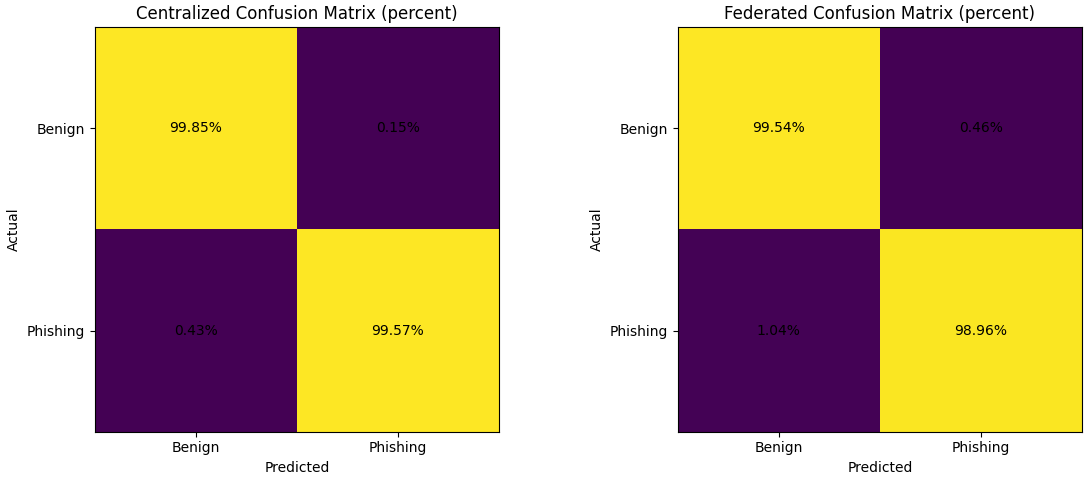

Figure 3: (a) Centralized model confusion matrix (b) FL confusion matrix

4.2. Federated Global Validation Accuracy

Figure 4 denotes the global validation accuracy of the federated model in all communication rounds.

The figure shows a gradual progress in validation accuracy as training progresses and this demonstrates that the global model has gotten a stable convergence. The observed slight instabilities between rounds occur because client systems experience different data distribution patterns which is typical in FL environments. The FedAvg aggregation strategy establishes its ability to successfully integrate local client knowledge into a unified global model because the upward trend continues.

This shows that FL attains reliable convergence through its privacy protection abilities, which makes it better option for organizations to use in their collaborative phishing detection efforts.

4.3. Centralized Versus Federated Learning Progress

Figure 5 did a comparison of the learning curve of centralized and federated models using validation accuracy.

The centralized model shows a faster convergence because of its direct access to the complete dataset at ones. On the other hand, the federated model displays a steady learning curve which shows the effect of distributed training and periodic aggregation. In spite of these differences, the two models reached similar validation accuracy results, this is because both demonstrate different performance levels which confirm that FL systems will attain their best results when they receive suitable system configuration.

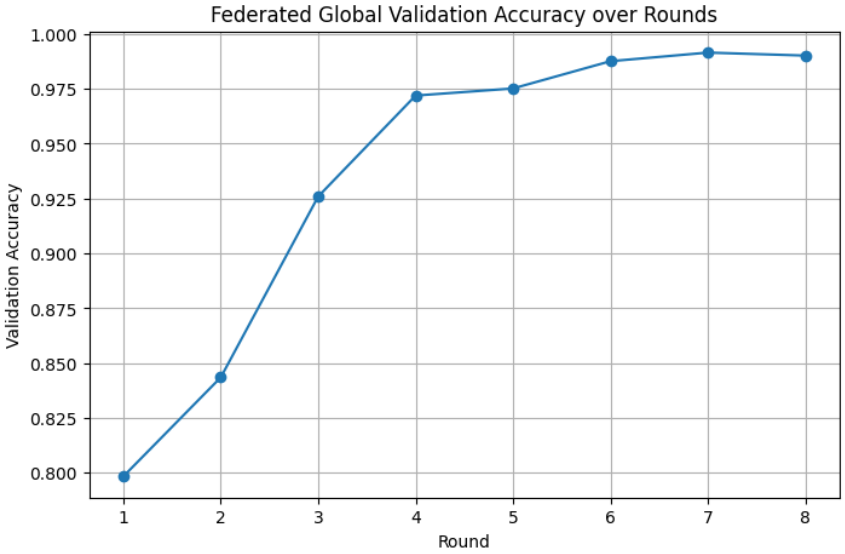

Figure 4: Federated Global Validation Accuracy over Rounds

This comparison discloses a crucial trade-off which are: centralized learning gives faster convergence rate and FL offers strong privacy assurances with only a slight performance compromise.

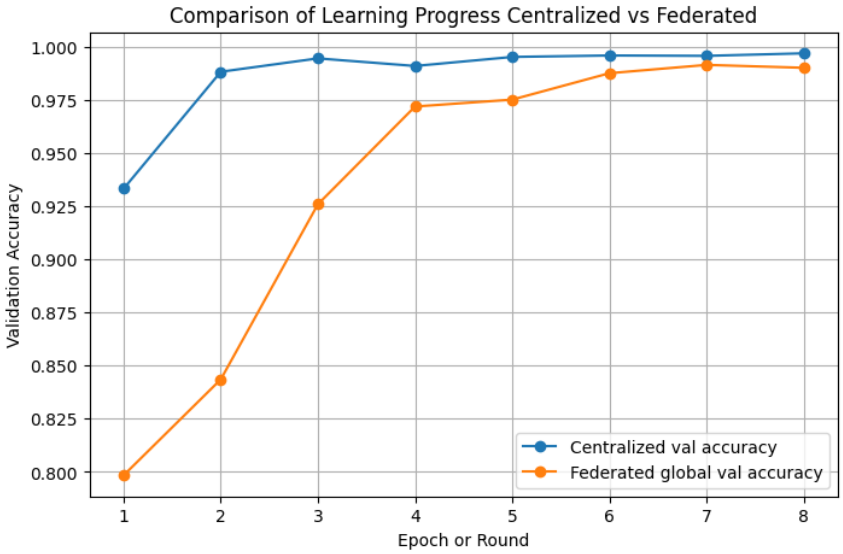

Figure 5: Centralized Vs Federated: Learning Progress

4.4. Explainability Analysis for Phishing Emails

The LIME feature importance together with textual explanations for a phishing email sample under the centralized model are presented in Figure 6 and Figure 7.

Figure 6 detailed the key words that contributed immensely to the phishing classification.

The model assigns greater importance to terms that describe urgent situations together with verification request terms and account suspension terms. The indicators match established phishing methods which verifies that the model detects essential meanings instead of random patterns. Figure 7, further visualizes these findings by showing which words have the greatest impact on the email content. The model decision process explanation which uses localized information helps security analysts and end users understand the classification.

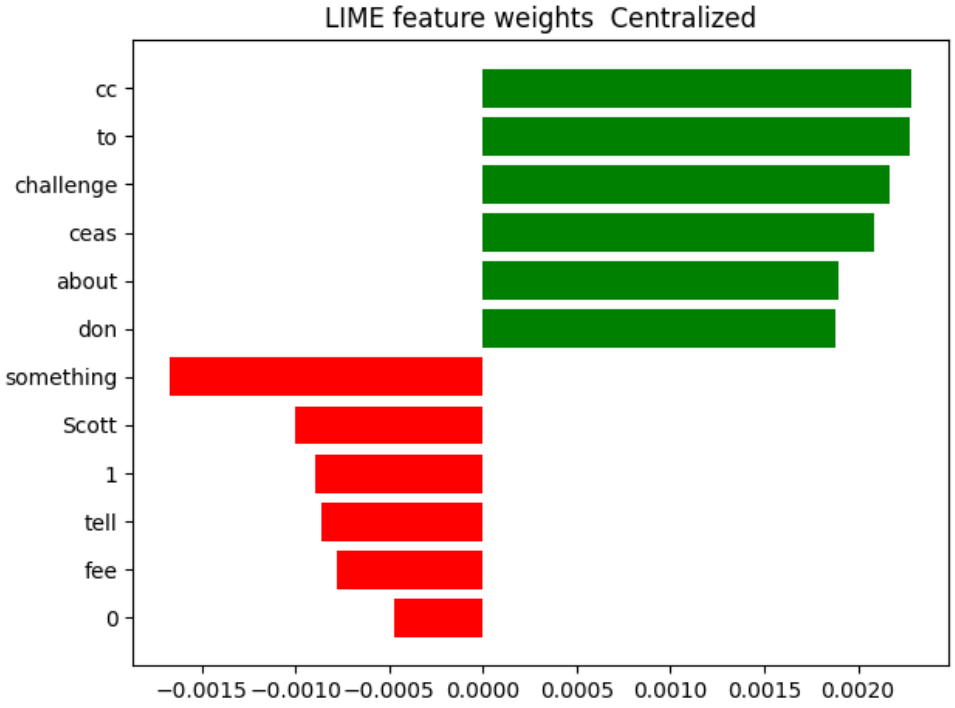

Figure 6: LIME feature weights Centralized phishing sample

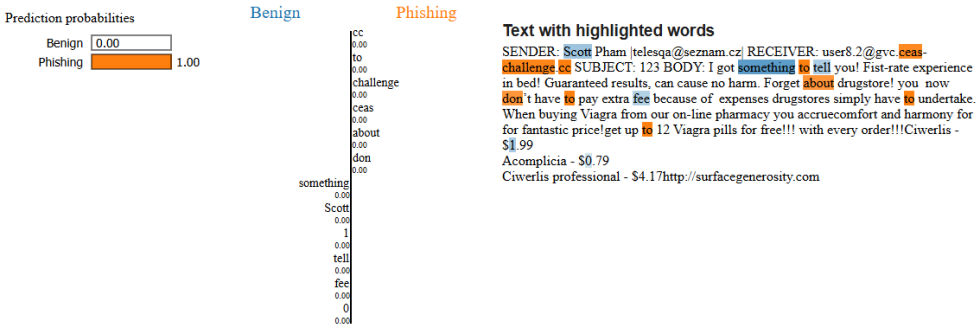

Figure 7: LIME explanations for phishing sample

4.5. Explainability Analysis for Benign Emails

Figure 8 and Figure 9 display the explainability results which were obtained from testing a benign email sample. The LIME model from Figure 8 allocates more weight to neutral and informational terms which appear in human day-to-day communication. The lack of urgency driven or misleading words contributes to the benign classification.

Figure 9 visually highlights this observation by highlighting benign terms within the email body. The contrast between phishing and benign explanations validates that the model learns specific linguistic patterns for each class which further strengths confidence in its decision making process.

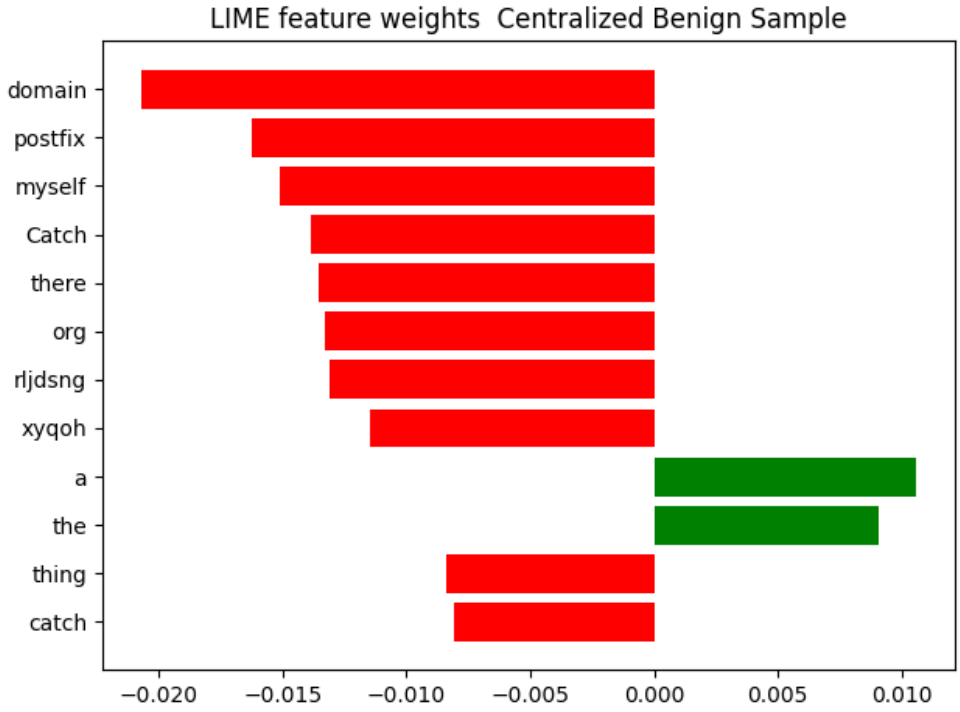

Figure 8: LIME feature weights Centralized Benign sample

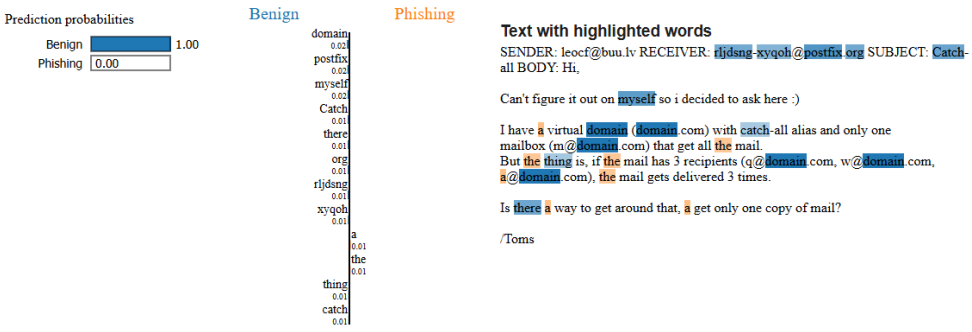

Figure 9: LIME explanations for Benign sample

4.6. Performance Comparison Between Centralized and Federated Models

Figure 10 shows the performance results of both centralized models and federated models with the use of five evaluation metrics which includes: accuracy, precision, recall, F1 score and AUC.

The centralized model recorded higher scores across all metrics compared to FL model because it have access to all the training data at ones. The federated model contend closely behind the centralized model because it maintains strong results in all the assessment tests. The small performance difference shows how FL technology helps keep model quality at the same time making sure that there is data privacy.

The result shows that the proposed multi modal federated model provides a practical balance between detection accuracy, privacy preservation and explainability.

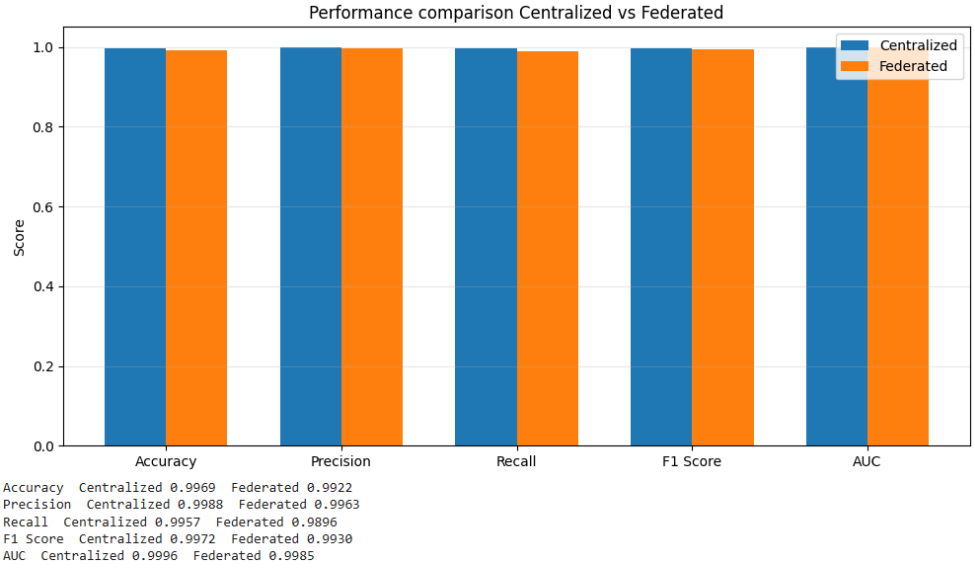

Figure 10: Performance comparison of centralized and FL models

5. Conclusion

From our findings, the experimental results show that the developed system successfully combines multi modal DL, FL and XAI to detect email phishing. The centralized model establishes a strong performance baseline and the federated model achieved comparable results without exposing sensitive email data. The system becomes suitable for real world security environments and this is because the introduction of explainability techniques improve both transparency and trustworthiness.

References

[1] J. Aljabri, N. Alzaben, N. Nemri, S. Alahmari, S. D. Alotaibi, S. Alazwari, et al. Hybrid stacked auto-encoder with dwarf mongoose optimization for phishing attack detection

in internet of things environment. *Alexandria Engineering Journal*, 106:164–171, 2024. doi: 10.1016/j.aej.2024.06.070.

[2] S. R. Alotaibi, H. K. Alkahtani, M. Aljebreen, A. Alshuhail, M. K. Saeed, S. A. Ebad, W. S. Almukadi, and M. Alotaibi. Explainable artificial intelligence in web phishing classification on secure IoT with cloud based cyber physical systems. *Alexandria Engineering Journal*, 110:490–505, 2025. doi: 10.1016/j.aej.2024.09.115.

[3] F. S. Alsubaei, A. A. Almazroi, and N. Ayub. Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. *IEEE Access*, 12:8373–8389, 2024. doi: 10.1109/ACCESS.2024.3351946.

[4] N. Alsuqayh, A. Mirza, and A. Alhogail. Exploring feature engineering and explainable AI for phishing website detection: A systematic literature review. *International Journal of Electrical and Computer Engineering (IJECE)*, 15(6):5863–5878, 2025. doi: 10.11591/ijece.v15i6.pp5863-5878.

[5] M. C. Calzarossa, P. Giudici, and R. Zieni. Explainable machine learning for phishing feature detection. *Quality and Reliability Engineering International*, 40:362–373, 2024. doi: 10.1002/qre.3411.

[6] HIPAA Compliance Assistance. *Summary of the HIPAA Privacy Rule*. Office for Civil Rights, Washington, DC, USA, 2003.

[7] IBM. Cost of a data breach report 2019. Technical Report 8, Computer Fraud & Security, 2019. Retrieved from: https://insights.integrity360.com/hubfs/2019-cost-of-a-data-breach-report-04_03025203USEN.pdf.

[8] P. Kairouz and H. B. McMahan. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

[9] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

[10] W. Li, S. Manickam, and Y. Chong. FedPhishLLM: A privacy-preserving and explainable phishing detection mechanism using federated learning and LLMs. *Journal of King Saud University – Computer and Information Sciences*, 37, 2025. doi: 10.1007/s44443-025-00267-0.

[11] M. Murhej and G. Nallasivan. Multi-modal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-based EAI-SC-LSTM. *Frontiers in Communications and Networks*, 6:1587654, 2025. doi: 10.3389/frcmn.2025.1587654.

[12] K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman. A systematic review on deep-learning-based phishing email detection. *Electronics*, 12(21):4545, 2023. doi: 10.3390/electronics12214545.

[13] C. Thapa, J. W. Tang, A. Abuadbba, Y. Gao, S. Camtepe, S. Nepal, M. Almashor, and Y. Zheng. Evaluation of federated learning in phishing email detection. *Sensors*, 23(9):4346, 2023. doi: 10.3390/s23094346.

[14] K. M. M. Uddin et al. Explainable machine learning for phishing site detection: A high-efficiency approach using boosting models and SHAP. *The Journal of Engineering*, 2025. doi: 10.1049/tje2.70110.

[15] V. V. G. and P. A. Thomas. Explainable AI for phishing detection: Techniques, challenges, and experimental validation. In *2025 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 64–68. IEEE, 2025. doi: 10.1109/RAICS66191.2025.11332593.

[16] B. Van Dooremaal, P. Burda, L. Allodi, and N. Zannone. Combining text and visual features to improve the identification of cloned webpages for early phishing detection. In *Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES '21)*, pages 1–10. Association for Computing Machinery, 2021. doi: 10.1145/3465481.3470112.

[17] A. Vulfin, A. Sulavko, V. Vasiliev, A. Minko, A. Kirillova, and A. Samotuga. A multi-modal phishing website detection system using explainable artificial intelligence technologies. *Machine Learning and Knowledge Extraction*, 8(1):11, 2026. doi: 10.3390/make8010011.

[18] A. Wainakh, E. Zimmer, S. Subedi, J. Keim, T. Grube, S. Karuppayah, A. Sanchez Guinea, and M. Mühlhäuser. Federated learning attacks revisited: A critical discussion of gaps, assumptions, and evaluation setups. *Sensors*, 23(1):31, 2022. doi: 10.3390/s23010031.

[19] L. Yang, J. Zhang, X. Wang, Z. Li, Z. Li, and Y. He. An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. *Expert Systems with Applications*, 165:113863, 2021. doi: 10.1016/j.eswa.2020.113863.