*Journal of Discrete Mathematical Sciences & Cryptography*
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
DOI : 10.47974/JDMSC-2606

# Phishing email detection using MBFT-Lite : A multi-branch federated transformer framework

Anita Shrotriya [‡]
Sanvi Agarwal [†]
Tvisha Khanna [§]
Sunita Singhal *
*Department of Computer Science and Engineering*
*Manipal University Jaipur*
*Jaipur 303007*
*Rajasthan*
*India*

**Abstract**

Phishing attacks continue to pose a significant cybersecurity threat by exploiting users through deceptive emails. While traditional centralized phishing detection methods achieve high accuracy, they face critical limitations, including privacy concerns, scalability challenges, and susceptibility to data distribution shifts. To address these issues, MBFT-Lite (Multi-Branch Federated Transformer-Lite) is proposed—a novel, lightweight, privacy-preserving phishing detection framework based on federated learning. MBFT-Lite employs a dual-branch architecture that combines semantic analysis via a TinyBERT model with structural analysis through XGBoost models on metadata features. The TinyBERT models are aggregated using the FedAvg a lgorithm, while the XGBoost models remain decentralized to enhance privacy and reduce communication overhead. Predictions from both branches are fused using strategies such as weighted averaging, max-confidence voting, and a meta-classifier. Extensive evaluations across four diverse email datasets—CEAS 08, SpamAssassin, Nigerian Fraud Emails, and Nazario—demonstrate that MBFT-Lite achieves over 99.4% accuracy across clients and 97.9% accuracy on unseen data, while preserving data privacy and ensuring scalability.

***Subject Classification:*** *Primary 93A30, Secondary 49K15.*

***Keywords:*** *Phishing detection, Federated learning, TinyBERT, Fusion layer.*

[‡] *E-mail:* **anita.shrotriya@jaipur.manipal.edu**
[†] *E-mail:* **agarwalsanvi004@gmail.com**
[§] *E-mail:* **tvishakhanna1557@gmail.com**
* *E-mail:* **sunita.singhal@jaipur.manipal.edu (Corresponding Author)**

## 1. Introduction

The use of phishing attacks continues to be a major consternation of cybersecurity as they exploit social engineering through well-crafted deceptive emails to manipulate people. Traditional centralized phishing detection solutions, while effective, raise privacy concerns and create vulnerability through single points of failure [1-3]. Federated Learning (FL) addresses these concerns by enabling model training without sharing sensitive data, allowing decentralized learning across private datasets [4-5]. However, minimal research has been done on federated phishing detection, especially when it comes to integrating multimodal representation of data that combines semantic, wherein the content of the e-mail serves as a basis, and structural, where the metadata of the e-mail serves as the main focus features [6-8].

We present MBFT-lite, a privacy-preserving federated phishing system using TinyBERT for semantic analysis [9-10], while XGBoost is used to capture insights into structural metadata [11]. By integrating TinyBERT training via the Flower framework and merging XGBoost models through logistic regression, MBFT-Lite maintains high accuracy despite client data heterogeneity. This hybrid approach ensures minimal privacy leakage while improving detection performance.

MBFT-Lite introduces a federated, cross-device phishing detection architecture eliminating centralized data storage [7], achieving state-of-the-art results through a hybrid transformer-based model that integrates semantic understanding with metadata analysis. Comprehensive testing demonstrates strong generalization ability to novel unseen phishing datasets. MBFT-Lite differs from prior works is its unique fusion of transformer-based semantic modeling and interpretable metadata-based learning in a federated setup—addressing both privacy and generalization simultaneously. It supports multi-branch local training and centralized fusion logic and adaptive ensemble strategies (weighted averaging, max confidence voting, and a logistic regression-based meta-classifier), making it a regulation-friendly, deployable, and high-performing phishing detection solution relevant to real-world use cases across industries such as finance, education, and healthcare.

The rest of the paper is structured as follows: Section 2 reviews the related literature on phishing detection, federated learning, and hybrid deep learning approaches. Section 3 outlines the proposed methodology, including the MBFT-Lite architecture, local model training procedures, and fusion techniques. Section 4 presents the key algorithms employed in

the framework. Section 5 reports the experimental results, including federated learning performance, fusion strategy comparisons, and generalization evaluation. Section 6 provides a discussion of the findings, highlighting key observations and limitations. Section 7 concludes the paper by summarizing the main contributions and outlining directions for future research.

## 2. Literature Review

**Evolution of Phishing Detection:** Phishing detection has evolved from rule-based and classical machine learning (ML) models such as support vector machines and decision trees to advanced deep learning (DL) techniques [1]. Early ML approaches relied on textual and structural features but struggled against evolving phishing tactics. Transformer-based models such as BERT improved semantic understanding and detection accuracy, yet their dependence on centralized data raises privacy and security concerns [2].

**Federated Learning in Phishing Detection:** Federated learning (FL) has emerged as a promising paradigm to address privacy and data-sharing concerns in phishing detection. FL enables multiple organizations or devices to collaboratively train a global model without sharing raw data, thereby preserving user privacy and supporting compliance with data protection regulations such as GDPR and the EU AI Act. In FL, each participant trains a local model and only shares model parameters with a central server for aggregation.

Recent research demonstrates that FL can achieve detection performance comparable to centralized methods, particularly when data is balanced. However, with increasing participants or imbalanced data, challenges such as reduced model stability and slower convergence arise, with performance largely influenced by data distribution and model choice.

**Advances in Federated Frameworks:** Recent work has introduced hybrid frameworks that combine FL with continual learning strategies and attention-based classifiers [7]. These advanced approaches enable distributed systems to adapt to new and evolving phishing tactics in real time, while also addressing issues like catastrophic forgetting and model drift. Incorporating attention mechanisms and residual connections further improves the model's ability to focus on relevant phishing features and maintain performance over time.

## 3. Methodology

This section highlights the complete end-to-end methodology employed to design, train, and evaluate MBFT-Lite, a multi-branch phishing detection system based on federated learning. The architecture incorporates local training of XGBoost for metadata-based analysis and federated training of TinyBERT [10], [11] for semantic understanding of email content across distributed clients as displayed in Fig. 1. TinyBERT models are aggregated centrally using the FedAvg algorithm, ensuring privacy preservation, while XGBoost remains locally optimized. Following training, three fusion strategies—weighted average, max confidence voting, and a meta-classifier—are applied to combine predictions from both models and enhance classification reliability. Finally, the framework is evaluated using standard classification metrics and tested on an unseen dataset to assess its generalization capability.

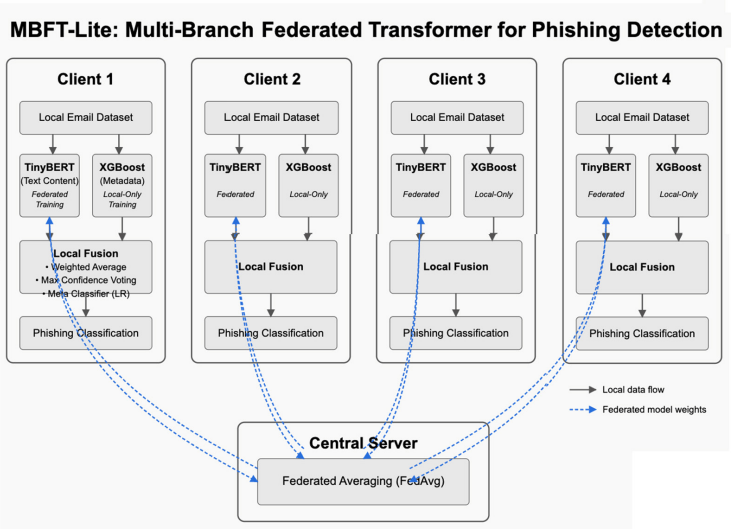


**Figure 1**

**System Architecture of MBFT-Lite (Multi-Branch Federated Transformer) Framework**

## 4. Algorithm

The two algorithmic components upon which MBFT-Lite is based are specified below to enable sound and privacy-sensitive phishing detection. These are: (1) the Federated Learning procedure based on the Federated

Federated Learning with FedAvg

**Require:** $K$ clients with local datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$, number of FL rounds $R$
**Ensure:** Global model $\mathcal{G}$
1: Initialize global model $\mathcal{G}$
2: **for** $r = 1$ to $R$ **do**
3: **for all** client $k = 1$ to $K$ **in parallel do**
4: Client $k$ initializes model $\mathcal{G}_k \leftarrow \mathcal{G}$
5: Client $k$ trains $\mathcal{G}_k$ locally on data $\mathcal{D}_k$
6: **end for**
7: Server aggregates models: $\mathcal{G} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \cdot \mathcal{G}_k$ ▷ $n_k$: samples in client $k$, $n$: total samples
8: **end for**
9: **return** Final global model $\mathcal{G}$

**Figure 2**

**Algorithm 1 for TinyBERT model training**

Averaging (FedAvg) algorithm, and (2) the Meta Classifier Fusion method based on the TinyBERT and XGBoost predictions combination.

This algorithm regulates the TinyBERT model training by various clients without exposing their local information. All clients train locally on their local email dataset, and after the completion of a set number of epochs, only the model updates are pushed to the central server. The server aggregates the updates using the FedAvg algorithm that takes an average of the weights based on the size of each client's dataset as shown in Figure 2.

The estimated class probabilities from XGBoost and TinyBERT, trained on each client, are subsequently meta-classified using a logistic regression-based meta-classifier. The fusion strategy leverages the power of both structural (metadata) and semantic (textual) analysis in order to produce more reliable final predictions as mentioned in Algorithm 2.

Meta Classifier Fusion

**Require:** TinyBERT probabilities $P_{tb}$, XGBoost probabilities $P_{xgb}$, Ground truth labels $Y$
**Ensure:** Trained Meta Classifier model $\mathcal{M}$
1: Create fusion dataset: $X_{meta} \Leftarrow \text{Concatenate}(P_{tb}, P_{xgb})$
2: Initialize logistic regression model $\mathcal{M}$
3: Train $\mathcal{M}$ on $(X_{meta}, Y)$
4: Predict final labels: $\hat{Y} = \mathcal{M}(X_{meta})$
5: Evaluate performance using accuracy, precision, recall, and F1-score
6: Save $\mathcal{M}$ for future inference on unseen data

**Figure 3**

**Algorithm 2 for XGBoost and TinyBERT, trained on each client**

## 5. Results

The model is evaluated in terms of standard classification metrics—accuracy, precision, recall, and F1-score—defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{I}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{II}$$

$$f1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precession} + \text{Recall}} \tag{III}$$

Here, *TP* refers to true positives (correctly identified phishing emails), *FP* to false positives (legitimate emails misclassified as phishing), and *FN* to false negatives (phishing emails misclassified as legitimate). These metrics guide performance assessment [13].

The fused probability score $P_{fused}$ from TinyBERT and XGBoost outputs is computed using weighted average fusion:

$$P_{fused} = \alpha * P_{tinyBERT} + \ (1 + \alpha) * P_{XGBoost} \tag{IV}$$

where $\alpha \in [0,1]$. In our experiments, $\alpha = 0.5$ was used. For meta-classifier fusion, the input feature vector is constructed as:

$$x = \left[ \frac{P_{TinyBERT}}{P_{XGBoost}} \right] \tag{V}$$

and passed to a logistic regression model to infer the final class label.

### 5.1 *Federated Learning Performance*

**Table 1**

**Federated Learning Evaluation Metrics per Client (TinyBERT)**

| Client | Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Client 1 | CEAS 08 | 0.975 | 1.000 | 0.970 | 0.976 |
| Client 2 | Nazario + CEAS Legit | 0.891 | 0.983 | 0.791 | 0.881 |
| Client 3 | Nigerian + CEAS Legit | 0.994 | 0.990 | 0.998 | 0.994 |
| Client 4 | SpamAssassin | 0.850 | 0.983 | 0.707 | 0.822 |

Four clients were simulated using heterogeneous datasets: CEAS 08, Nazario, Nigerian Fraud, and SpamAssassin. The TinyBERT model was fine-tuned locally [10] on each client using the FedAvg aggregation algorithm over five communication rounds. Table 1 summarizes the evaluation. Client 3 showed the best results due to balanced and diverse data, while Client 4 had a lower recall, indicating a higher false negative rate possibly due to noisy data.

### 5.3 *Local XGBoost Model Performance*

The XGBoost model was trained locally on each client and excluded from federated aggregation [11], leveraging its efficiency and interpretability on metadata while preserving data privacy. Table 2 shows the model's performance per client.

**Table 2**

**XGBoost Metadata Model Performance per Client**

| Client | Accuracy | Precision | Recall | F1-Score | Samples |
|---|---|---|---|---|---|
| Client 1 | 0.9560 | 0.9562 | 0.9560 | 0.9560 | 1000 |
| Client 2 | 0.9393 | 0.9408 | 0.9393 | 0.9392 | 626 |
| Client 3 | 0.9397 | 0.9398 | 0.9397 | 0.9396 | 1260 |
| Client 4 | 0.9375 | 0.9375 | 0.9375 | 0.9375 | 688 |

### 5.2 *Fusion Strategy Comparison*

Three post-federation fusion strategies were tested: Weighted Average Fusion, Max Confidence Voting, and a Meta Classifier (Logistic Regression) [12]. The average performance is illustrated in Fig. 4. The Meta Classifier outperformed the others with an F1-score of **0.9970**, narrowly surpassing the other two at **0.9969**, offering improved consistency across clients.

| Fusion Strategy | Average F1-Score |
|---|---|
| Meta-Classifier | 0.997049 |
| Maximum Voting | 0.996884 |
| Weighted Average | 0.996884 |

**Figure 4**

**Average F1-score of different fusion strategies. Meta Classifier slightly outperformed others**

### 5.4 *Evaluation of Unseen Data*

The trained meta-classifier was tested on a fully unseen dataset of 334 emails. It achieved a high accuracy of 97.9% and an F1-score of 97.91, as shown in Table 3. The classifier correctly identified 176 out of 181 phishing emails and made only 2 false positive errors. Fig. 5 and Fig. 6 demonstrate robust real-world performance.

**Table 3**

**Meta Classifier Evaluation on Unseen Dataset**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Class 0 | 0.9679 | 0.9869 | 0.9773 | 153 |
| Class 1 | 0.9888 | 0.9724 | 0.9805 | 181 |
| Macro Avg | 0.9784 | 0.9797 | 0.9789 | 334 |
| Weighted Avg | 0.9792 | 0.9790 | 0.9791 | 334 |

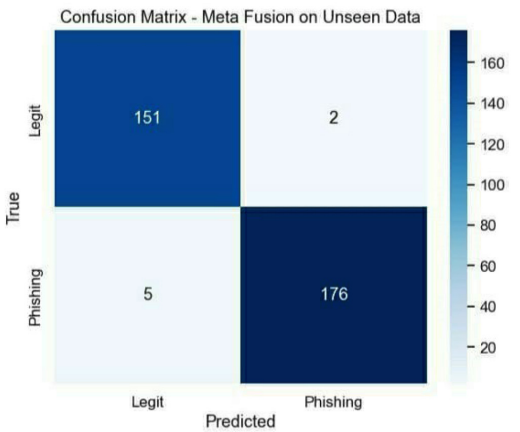


**Figure 5**

**Confusion matrix of the Meta Classifier on unseen email data. High TPR and low FPR achieved.**

## 6. Conclusion and Future Work

This paper introduces MBFT-Lite, a Multi-Branch Federated Transformer framework for privacy-preserving phishing detection that combines semantic learning via TinyBERT with structural analysis using XGBoost. Semantic analysis is performed by TinyBERT, trained in a

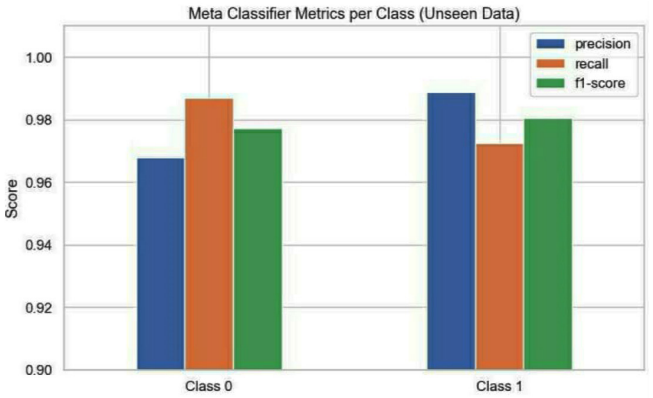


**Figure 6**

**Per-class precision, recall, and F1-score of the Meta Classifier on the unseen test dataset. The model performs consistently across both legitimate (Class 0) and phishing (Class 1) emails.**

federated manner with each client fine-tuning on local email content and sharing only model weights. In parallel, XGBoost models are trained solely on local metadata features and remain entirely on each client. The two branches are fused locally through weighted averaging, max-confidence voting, or a logistic-regression meta-classifier, delivering high accuracy and strong generalization on previously unseen phishing samples. Experimental results confirm both client-level and global performance gains, demonstrating the framework's practicality for real world deployment.

Future enhancements will focus on incorporating Differential Privacy and Secure Aggregation, introducing domain adaptation and client-specific fine-tuning to mitigate data heterogeneity, experimenting with advanced fusion approaches such as neural ensembles, and extending evaluation to multilingual and demographically diverse datasets to ensure equity, fairness, and broad applicability.

## References

[1] E. G. Dada, J. S. Bassi, H. Chiroma, S. I. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: Review, approaches and open research problems," *Heliyon*, vol. 5, no. 6 (Jun. 2019).

[2] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism," *IEEE ACCESS,* vol. 7, pp. 56, 56329 - 56340 (2019).

[3] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, "SoK: A comprehensive reexamination of phishing research from the security perspective," *IEEE Commun. Surveys Tuts.,* vol. 22, no. 1, pp. 671–708 (2019).

[4] B. Y. Lin, C. He, Z. Ze, H. Wang, Y. Hua, C. Dupuy, and S. Avestimehr., "FedNLP: Benchmarking federated learning methods for natural language processing tasks," in *Findings of the Assoc. Comput. Linguistics: NAA*CL 2022, pp. 157–175 (Jul. 2022).

[5] L. U. Khan, S. R. Pandey, N. H. Tran, W. Saad, Z. Han, M. N. Nguyen, and C. S. Hong, "Federated learning for edge networks: Resource optimization and incentive mechanism," *IEEE Commun. Mag.*, vol. 58, no. 10, pp. 88–93 (Oct. 2020).

[6] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 15908-15919 (2021).

[7] 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*, vol. 23, no. 9, Art. no. 4346 (2023).

[8] M. Alazab, S. P. R. Rm, P. K. R. Maddikunta, T. R. Gadekallu, and Q. V. Pham, "Federated learning for cybersecurity: Concepts, challenges, and future directions," *IEEE Trans. Ind. Informatics*, vol. 18, no. 5, pp. 3501–3509 (May 2021).

[9] M. Revathi, "Exploring the efficacy of federated-continual learning nodes with attention-based classifier for robust web phishing detection: An empirical investigation," arXiv preprint arXiv:2405.03537 (2024).

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding,"

in Proc. 2019 Conf. North American Chapter of the Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186 (Jun. 2019).

[11] P. Dadheech, A. Kumar, and V. Singh, "An optimization of bitmap index compression technique in bulk data movement infrastructure," *TARU J. Sustain. Technol. Comput.*, vol. 1, no. 2, pp. 81–91 (2019), doi: 10.47974/2019.TJSTC.003.

[12] M. Vasti and A. Dev, "Ensemble learning based predictive modelling on a highly imbalanced multiclass data," *J. Inf. Optim. Sci.*, vol. 45, no. 8, pp. 2141–2164 (2024), doi: 10.47974/JIOS-1778.

[13] S. V. Devika, B. Unhelkar, S. S. Shankar, P. Chakrabarti, and S. Arvind, "Managerial perspectives: Utilizing MLP-CNN for predicting and classifying DDoS attacks," *J. Inf. Optim. Sci.*, vol. 45, no. 8, pp. 2071–2079 (2024), doi: 10.47974/JIOS-1651.

*Received July, 2025*