Defending Against Text-Based Social Engineering Attacks Using Federated Adversarial Learning

Emdadul Haque Iram EMDADUL.HAQUE.IRAM@G.BRACU.AC.BD
Navid Nahiyan NAVID.NAHIYAN@G.BRACU.AC.BD
Musfika Jahan MUSFIKA.JAHAN@G.BRACU.AC.BD
Md. Nazmul Hasan MD.NAZMUL.HASAN4@G.BRACU.AC.BD
Department of Computer Science and Engineering
BRAC University, Dhaka, Bangladesh

Editor: Sakinat Folorunso, Roseline Ogundokun, and Fransisca Oladipo

Abstract

Text-based social engineering attacks such as phishing emails and scam messages have remained quite successful because language patterns and obfuscation patterns are constantly adapted by the adversaries, and centralized detection approaches have serious privacy concerns as well as insufficient real-world deployment. This thesis proposes a privacy-preserving and robust detection system by combining Joint Embedding Predictive Architecture (JEPA) representation learning process and federated learning, which can enable computationally costly analysis by multiple clients to collaboratively train the global model without exchanging the raw user data. To overcome modeling capacity in the heterogeneous (non-IID) clients distribution, a Mixture-of-Experts (MoE) design is proposed to avoid overload decision paths in various input characteristics and a Kolmogorov-Arnold Network (KAN)-based design is adopted to improve the expressive and interpretable feature transformations of the prediction head. The global model is trained by iterative local optimization and server-side aggregation while its robustness is induced by a federated adversarial learning stage which exposes the model to a set of adversarially perturbed samples/embeddings during model training. Experimental results on social engineering datasets show that the proposed JEPA-Federated-MoE/KAN pipeline consistently exhibits excellent detection performance with privacy preservation and offers enhanced flexibility against adversarial changes in comparison to federated training, which is conducive to a very pragmatic trade-off between detection accuracy, privacy guaranteeing and robustness to deploy effective FR methods in real messaging environments.

Keywords: Federated Learning; Adversarial Robustness; Social Engineering Detection; JEPA; Mixture-of-Experts; KAN; Phishing; Privacy Preservation; Federated Adversarial Learning.

1. Introduction

Text-based social engineering attacks, including phishing and smishing, are still considered one of the most widespread cyber threats, due to their ability to exploit the cognitive processes behind human decision-making, rather than relying on exploiting technical vulnerabilities (Aun et al., 2023). According to industry statistics, over one million phishing attacks occurred during Q1, 2025, alone (Anti-Phishing Working Group, 2025). The detection of these attacks is limited by two structural barriers: (i) communication data is protected by privacy laws, making it impossible to collect the data centrally; (ii) centrally

trained models are unable to generalize over heterogeneous user populations with non-IID data distributions.

Federated Learning (FL), is an approach to address the privacy gap by allowing models to be learned collaboratively without having access to the underlying data (McMahan et al., 2017). However, FedAvg suffers from performance degradation with non-IID data distributions, while still being prone to adversarial evasion attacks. At the same time, the dynamic, multi-modal nature of social engineering attacks, including text, HTML structure, and URL, requires advanced representation learning, which is not possible with traditional bag-of-words or token-based classification models.

This paper proposes a novel pipeline for detecting social engineering attacks, where (1) JEPA is used to encode the semantic information from the structure of the multimodal artifacts; (2) the proposed MoE-KAN is used to add classification capacity without requiring retraining the entire model; and (3) FAL is proposed to add adversarial perturbations during local training, providing improved robustness against evasion attacks.

1.1. Contributions

- A unified JEPA-based multimodal encoder for phishing and smishing attacks based on text, HTML, and URL modalities.
- A Mixture-of-Experts Knowledge-Aware Network (MoE-KAN) for building a prediction head to handle heterogeneous attack patterns for non-IID data in federated environments.
- Federated Adversarial Learning (FAL) within the embedding space for both efficiency and privacy preservation.
- Empirical evaluation for centralized, federated, and adversarial scenarios using large-scale datasets.

2. Related Work

Traditional phishing detection using machine learning techniques relies on manually crafted features like TF-IDF and N-Grams, which are then used as input for support vector machines and decision trees, respectively (Wilk-Jakubowski et al., 2025). The emergence of deep learning techniques significantly boosted the overall accuracy of phishing detection, especially when using hybrid architectures like LSTM, BiGRU, CNN, and CNN-BiGRU, where a high accuracy of 97.2% is attained (Mahmud et al., 2024). The Transformer family of architectures, including BERT, RoBERTa, and DeBERTa, is close to the optimal solution for centralized architectures (Mahendru and Pandit, 2024). Recent studies also rely on LLM for phishing attack generation and detection (Heiding et al., 2024). All of the approaches, centralized access to the data is assumed. Federated learning for security-related tasks is a relatively new direction. In their work, McMahan et al. (McMahan et al., 2017) proposed the FedAvg algorithm. Later, FedPGT (Lai et al., 2024) and other approaches extended federated learning by including adversarial training for server-side attacks. In Chen et al. (Chen et al., 2021), the authors proposed certifiable robustness for federated learning. Although federated learning with a blockchain for smishing attacks is studied

(Ning et al., 2024), federated learning for text-based social engineering attacks remains a relatively unexplored direction. Although high centralized accuracy for social engineering attacks is attained by SOCIALBERT (Abobor and Josyula, 2023), no guarantees of privacy and robustness are provided. As far as the authors are aware, the integration of JEPA and MoE-KAN within a federated adversarial training framework represents a novel contribution in this field.

3. Proposed Methodology

Our pipeline process (Fig. 1), involves four steps: data unification and preprocessing, JEPA backbone pretraining, MoE KAN head training, and federated adversarial fine-tuning.

3.1. Multimodal JEPA Encoder

Each message instance is described by a canonical schema, which is composed of `body_plain`, `body_html` (empty for SMS), `url_list`, and `url_features`. Specialized transformers are used to generate modality-specific embeddings:

$$h^{\text{text}} = f_{\text{text}}(x^{\text{text}}), \quad h^{\text{html}} = f_{\text{html}}(x^{\text{html}}), \quad h^{\text{url}} = f_{\text{url}}(x^{\text{url}}). \tag{1}$$

The projection layer f combines the embeddings to form a joint representation:

$$z = \phi([h^{\text{text}}; h^{\text{html}}; h^{\text{url}}]) \in \mathbb{R}^d. \tag{2}$$

The JEPA objective learns a predictor to predict the target embeddings from context embeddings in the representation space, avoiding pixel/token reconstruction and promoting semantically rich, robust features with respect to distribution shifts (Assran et al., 2023; Saito et al., 2025). In the centralized pre-training phase, the entire model is learned, while in the federated phases, the backbone is frozen, significantly reducing communication costs only to the head parameters.

3.2. MoE–KAN Prediction Head

The frozen JEPA embedding z is forwarded through a Mixture-of-Experts (MoE) prediction head. A lightweight router g computes the gating weights, and N KAN expert networks operate individually on z:

$$y = \sum_i g_i(z) \cdot \text{KAN}_i(z), \qquad \text{top-}k \text{ sparsity enforced.} \tag{3}$$

KAN experts replace fixed MLP activations with learnable univariate B-spline functions for the edge. This allows for more flexible decision boundaries and requires significantly fewer parameters than MLPs (Liu et al., 2025). The load balancing auxiliary loss helps prevent expert collapse. The MoE–KAN prediction head is subjected to a warm start through centralized training before federated rounds begin.

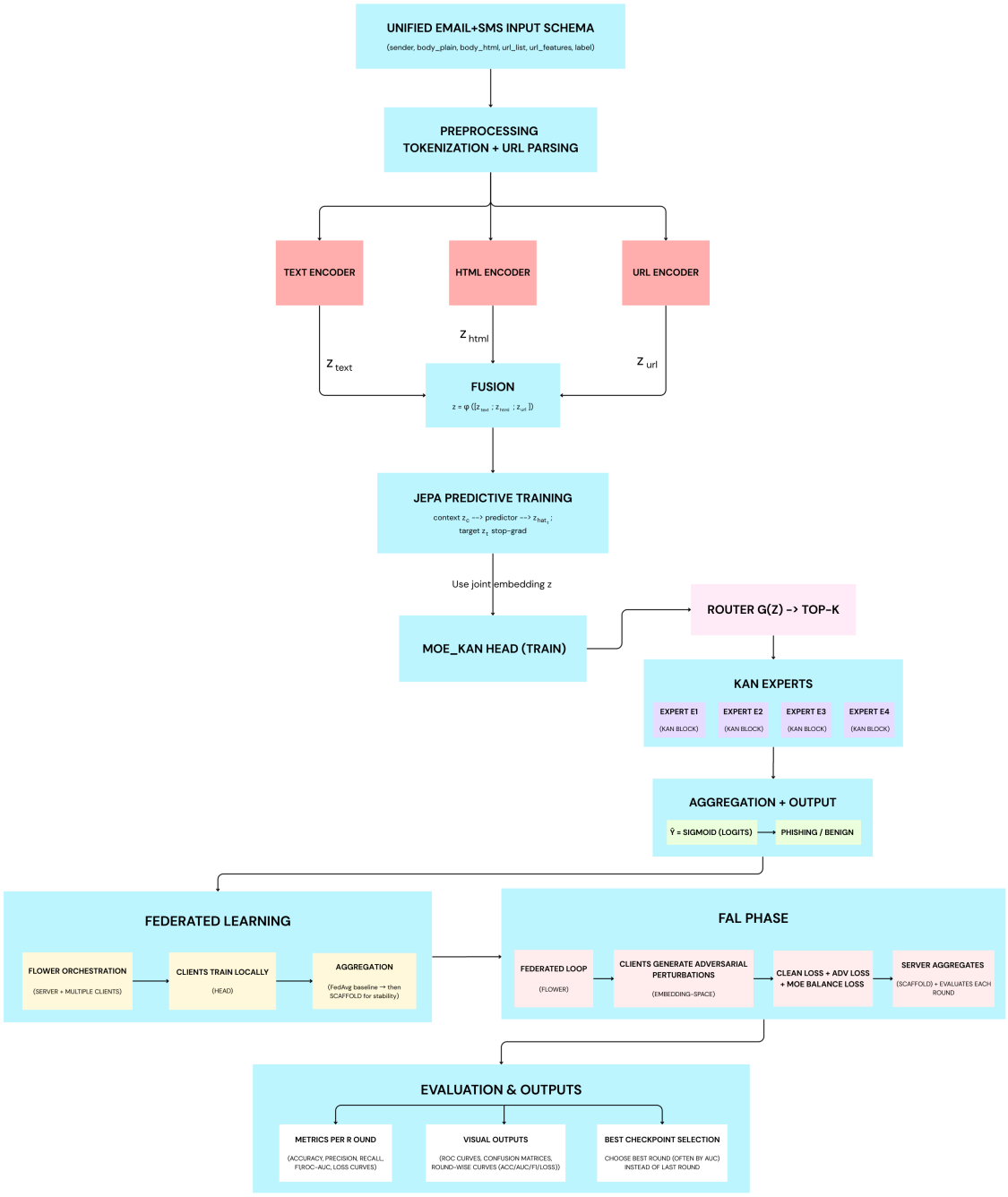

Figure 1: Overview of the proposed JEPA–MoE-KAN pipeline

3.3. Federated Learning with SCAFFOLD

Federated learning is carried out using the Flower framework together with the SCAFFOLD optimizer (McMahan et al., 2017). This optimizer utilizes control variates for correcting client drift when data is not IID. Ten client shards are used for federated client simulations. Only the MoE–KAN prediction head parameters are transmitted during federated rounds, and the JEPA backbone remains frozen. Server validation utilizes cached JEPA embeddings, separating validation from computation.

3.4. Federated Adversarial Learning (FAL)

Adversarial attacks are incorporated by perturbing the embedding space, not token space. This maintains privacy and circumvents costly backpropagation through the backbone:

$$z^{\text{adv}} = z + \delta^*, \qquad \delta^* = \arg\max_{\delta} \ell(\theta; z + \delta), \qquad \|\delta\| \leq \epsilon. \tag{4}$$

The client objective is a combination of clean and adversarial objectives:

$$\ell_{\text{total}} = \ell_{\text{clean}} + \alpha \cdot \ell_{\text{adv}} + \lambda \cdot \ell_{\text{balance}}. \tag{5}$$

A conservative ϵ and a single-step PGD attack are adopted for stability.

4. Empirical Evaluation

4.1. Dataset

Our methodology utilizes two consolidated datasets from publicly available sources. The Email dataset (182,709 samples) merges four phishing corpora including Enron, TREC, CEAS08, and curated social-engineering datasets (Champa et al., 2024; Engineering Ingegneria Informatica Spa, 2025; Miltchev et al., 2024; subhajournal, 2024), providing rich plain-text, HTML, and URL features. The SMS dataset (199,563 samples, 64.7% phishing) merges five smishing corpora including the UCI SMS Spam Collection and the Combined Labeled Smishing dataset (Mishra and Soni, 2022; Shaghayegh-HP, 2024; UCI Machine Learning Repository, 2012). Both datasets are normalized to a canonical six-field schema and partitioned into ten non-IID client shards for federated simulation.

4.2. Evaluation Metrics

The accuracy, precision, recall, F1-score, and ROC-AUC score are used as evaluation metrics. However, due to the nature of class imbalance in phishing datasets, ROC-AUC is used as the major evaluation metric. The binary cross-entropy (BCE) loss is also monitored to measure the model's stability. Confusion matrices and round-wise metric curves are used to detect non-IID drift and adversarial attack risks. Checkpointing using the best round on the basis of AUC is used instead of relying on the last round's result.

4.3. Results

The table 1 summarizes performance across all experimental tracks. The centralized training using the proposed method attains high AUC values for both email (0.982) and SMS (0.988) modalities, validating the efficacy of multimodal JEPA embeddings in terms of discriminative power. The proposed MoE-KAN warm-start approach attains an AUC score of 0.968, confirming its ability to learn diverse attack patterns using the lightweight expert network without requiring any further training.

Table 1: *Performance across centralized, federated, and adversarial settings.*

Track / Setting	**Acc**	**Prec**	**Rec**	**F1**	**AUC**
Email JEPA (Centralized)	0.930	0.931	0.965	0.948	0.982
SMS JEPA (Centralized)	0.946	0.947	0.971	0.959	0.988
JEPA + MoE–KAN Warm-start	0.887	NR	NR	NR	0.968
Email+SMS FL (Best Rd. 1)	0.851	0.853	0.943	0.895	0.908
Email+SMS FAL (Best Rd. 1)	0.841	0.851	0.928	0.887	0.896

In terms of federated learning (10 clients, using SCAFFOLD), we observe that the model's performance peaks in Round 1 with an F1 score of 0.895 and an AUC score of 0.908 and. However, this is followed by a decline in performance due to non-IID drift, a known issue in federated learning (Lai et al., 2024). This phenomenon suggests that we should focus on selecting the best model rather than relying on the final model's performance. The addition of FAL in this case results in an accuracy trade-off in exchange for robustness against adversarial embedding changes in Round 1 (AUC: 0.896, F1: 0.887). The SMS-only model attains an AUC score of 0.993 in Round 5 in the federated learning case.

4.4. Comparison with Baselines

Table 2: *Comparison with existing phishing / smishing detection methods.*

Model	**Acc**	**F1**	**AUC**
SVM (baseline)	0.677	0.773	0.619
Random Forest	0.739	0.745	0.761
CNN-BiGRU (Mahmud et al., 2024)	0.972	0.959	0.993
LSTM (Uddin and Sarker, 2024)	0.964	0.969	0.995
1D-CNNPD (Altwaijry et al., 2024)	0.989	0.984	0.999
Federated JEPA + MoE-KAN (Ours)	0.851	0.895	0.908
Federated JEPA + MoE-KAN + FAL (Ours)	0.841	0.887	0.896

Table 2 places our results in the context of the prior art. The centralized deep learning methods CNN-BiGRU and 1D-CNNPD outperform our federated approach in terms of raw accuracy; however, this is at the expense of data centralization. Meanwhile, we achieved competitive results in terms of AUC 0.908 under the federated constraint of no data sharing, and we achieved adversarial robustness, which is not possible in the baseline comparisons.

5. Discussion

The warm-start AUC value (Fig. 2) of 0.968 implies that JEPA multimodal representations are more effective in encoding deception cues related to semantics and structure than surface features, and that MoE–KAN specialization effectively deals with diversity in attack types. The performance gap observed in federated learning compared to centralized learning results is as expected due to the IID sharding, which causes a classifier bias in the later rounds. The classifier's recall approaches 1.0 with a corresponding decrease in precision.

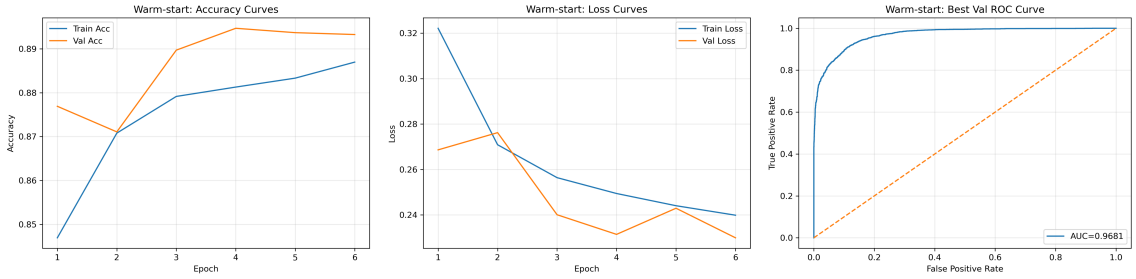

Figure 2: MoE_KAN training curves

The results of the FAL stage (Fig. 3) show that adversarial training in the embedding space is a computationally feasible approach for robustness in federated learning scenarios. It does not require any further data sharing, adds only one step of PGD per batch, and maintains AUC values above 0.89 for the best round. The compromise between clean and adversarial robustness is a reasonable price to pay for a security-critical application.

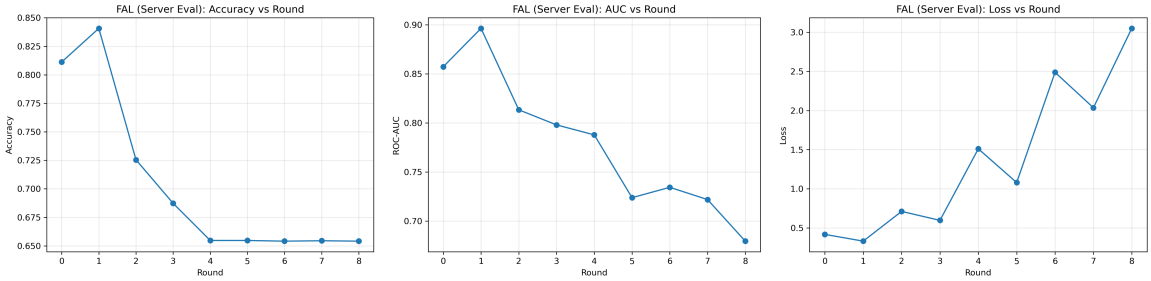

Figure 3: Federated Advarsarial Learning Curve

The following are the shortcomings of this research: The federated simulation was conducted on a single machine, but in a real-world scenario, there might be different convergence patterns for asynchronous clients. The figures from the original thesis provide further empirical results for training curves, confusion matrices, and ROC curves. Adversarial attacks on the token level have not been examined, and embedding-space FAL might underestimate the risk of text-level evasion.

6. Conclusion

We propose a privacy-preserving and adversarially robust framework for detecting text-based social engineering attacks using JEPA multimodal representation learning, MoE-KAN expert specialization, and Federated Adversarial Learning. The effectiveness of the proposed framework is validated using large-scale email and SMS phishing attack data, achieving centralized detection AUC ≈ 0.98 and competitive federated detection AUC ≈ 0.908 without relying on centralized raw user data. Federated Adversarial Learning provides robustness to adversarial attacks at a small cost in detection accuracy. The work offers a deployable blueprint to address the critical issues of representation learning, client heterogeneity, and adversarial robustness in the context of social engineering detection, which have been completely neglected in prior federated learning solutions in this area.

Possible future research directions include the design of personalized federated learning strategies to address instability in federated learning due to non-IID data, incorporating differential privacy, using multilingual data, and establishing adversarial robustness at the text level.

Acknowledgement

All thanks to Almighty Allah for giving us the courage and clarity to complete this research. We would like to express our sincere gratitude to our supervisor, Annajiat Alim Rasel, and our co-supervisor, Zayed Humayun, from the Department of Computer Science and Engineering, BRAC University, for their continuous guidance, valuable feedback, and motivation throughout the entire research period. Finally, we would also like to thank our families for their cooperation, belief in us, and continuous support.

References

M. Abobor and D. P. Josyula. Socialbert: A transformer based model used for detection of social engineering characteristics. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 174–178, 2023.

Najwa Altwaijry, Isra Al-Turaiki, Reem Alotaibi, and Fatimah Alakeel. Advancing phishing email detection: A comparative study of deep learning models. *Sensors*, 24(7):2077, 2024. doi: 10.3390/s24072077. URL https://www.mdpi.com/1424-8220/24/7/2077.

Anti-Phishing Working Group. Phishing activity trends report, 1st quarter 2025. Technical report, Anti-Phishing Working Group (APWG), July 2025. URL http://www.apwg.org. Analysis by Greg Aaron.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Yichiet Aun, Ming-Lee Gan, Nur Haliza Binti Abdul Wahab, and Goh Hock Guan. Social engineering attack classifications on social media using deep learning. *Computers,*

Materials & Continua, 74(3):4917–4931, 2023. doi: 10.32604/cmc.2023.032373. URL https://doi.org/10.32604/cmc.2023.032373.

Arifa I. Champa, Md. Fazle Rabbi, and Minhaz F. Zibran. Phishing email curated datasets (ceas_8, enron, ling, nazario, nigerian, spamassasin, trec), 2024. URL https://zenodo.org/records/8339691. Accessed: 2026-02-03.

C. Chen, B. Kailkhura, R. Goldhahn, and Y. Zhou. Certifiably-robust federated adversarial learning via randomized smoothing. In *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, 2021. doi: 10.1109/MASS52906.2021.00032.

Engineering Ingegneria Informatica Spa. Multiclass nlp dataset for phishing and social engineering threat detection, 2025. URL https://zenodo.org/records/15235123. Accessed: 2026-02-03.

F. Heiding, B. Schneier, and J. Bernstein. Devising and detecting phishing emails using large language models. *IEEE Security & Privacy*, 2024. URL https://ieeexplore.ieee.org/document/10466545.

W. Lai, Z. Xu, and Q. Yan. Fedpgt: Prototype-based federated global adversarial training against adversarial attack. In *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2024. doi: 10.1109/CSCWD61410.2024.10580613.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. In *International Conference on Learning Representations (ICLR)*, 2025.

Sakshi Mahendru and Tejul Pandit. Securenet: A comparative study of deberta and large language models for phishing detection. In *2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI)*, pages 160–169. IEEE, July 2024. doi: 10.1109/BDAI62182.2024.10692765.

Tanjim Mahmud, Md. Alif Hossen Prince, Md. Hasan Ali, Mohammad Shahadat Hossain, and Karl Andersson. Enhancing cybersecurity: Hybrid deep learning approaches to smishing attack detection. *Systems*, 12(11):490, 2024. doi: 10.3390/systems12110490. URL https://www.mdpi.com/2079-8954/12/11/490.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 1273–1282. PMLR, 2017.

Radoslav Miltchev, Dimitar Rangelov, and Evgeni Genchev. Phishing validation emails dataset, 2024. URL https://research.utwente.nl/en/datasets/phishing-validation-emails-dataset/. Accessed: 2026-02-03.

Sandhya Mishra and Devpriya Soni. Sms phishing dataset for machine learning and pattern recognition, 2022. URL https://data.mendeley.com/datasets/f45bkkt8pr/1. Accessed: 2026-02-03.

Weiguang Ning, Yingjuan Zhu, Caixia Song, and Jiwei Gao. Blockchain-based federated learning: A survey and new perspectives. *Security and Privacy*, 2024. doi: 10.1002/spy2.435.

A. Saito, P. Kudeshia, and J. Poovvancheri. Point-jepa: A joint embedding predictive architecture for self-supervised learning on point cloud. In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7348–7357, 2025. doi: 10.1109/WACV61041.2025.00714.

Shaghayegh-HP. Combined labeled smishing robust model training dataset, 2024. URL https://github.com/shaghayegh-hp/Smishing_Dataset. Accessed: 2026-02-03.

subhajournal. Phishing email dataset, 2024. URL https://www.kaggle.com/datasets/subhajournal/phishingemails. Accessed: 2026-02-03.

UCI Machine Learning Repository. Sms spam collection dataset, 2012. URL https://archive.ics.uci.edu/ml/datasets/sms+spam+collection. Accessed: 2026-02-03.

Mohammad Amaz Uddin and Iqbal H. Sarker. An explainable transformer-based model for phishing email detection: A large language model approach, 2024. URL https://arxiv.org/abs/2402.13871v2. arXiv preprint arXiv:2402.13871v2.

Jacek Lukasz Wilk-Jakubowski, Lukasz Pawlik, Grzegorz Wilk-Jakubowski, and Aleksandra Sikora. Machine learning and neural networks for phishing detection: A systematic review (2017–2024). *Electronics*, 14(18):3744, 2025. doi: 10.3390/electronics14183744.

Appendix A

JEPA-MoE/KAN-FAL Supplements

Dataset

Email Dataset: 182,709 samples; SMS Dataset: 199,563 samples. These data are extracted from freely available datasets referenced in the references section. Both datasets have been processed according to the canonical format of six fields and partitioned into ten non-IID client shards before training.

The datasets can be accessed from the following link: https://drive.google.com/drive/folders/1zMzBEZd6RWXoM73L4u7Z1E4yKWgGD5sq?usp=drive_link

Environment Setup

All experiments are carried out using Python version 3.10 and PyTorch version 2.1. The federated learning experiment runs on the Flower (flwr) framework. See instructions for installing PyTorch at pytorch.org and instructions for Flower at flower.ai.

To run the experiments, the following packages must be installed:

- `flwr` (1.7.0)

- `torch` (2.1.0)
- `transformers` (4.38.0)
- `scikit-learn` (1.4.0)
- `pandas` (2.1.0)
- `numpy` (1.26.0)

It is highly recommended to install the dependencies in a virtual environment.

Potential Errors

During federated training with Flower, you might encounter a CUDA out-of-memory error. This issue can be alleviated by decreasing the batch size or number of active clients per round.

Parameters

The following variables can be adjusted during training for customizations:

- `num_clients`: number of federated clients (default: 10)
- `num_rounds`: number of federated rounds
- `local_epochs`: number of local training epochs per round
- `num_experts`: number of KAN experts in the MoE head
- `top_k`: top-k sparsity in expert gating
- `epsilon`: perturbation bound ϵ for FAL
- `alpha`: adversarial loss weight in the objective function
- `lr`: learning rate for local optimization

Results

After pretraining, the AUC and F1 scores for Email and SMS tracks will be shown in the terminal. After each federated round, round-wise performance measures will be logged, and the best checkpoint will be automatically saved based on AUC. Training curves, confusion matrices, and ROC curves will be saved in the `outputs` directory after finishing the training process. The results and outputs can be accessed from here: https://drive.google.com/drive/folders/1j0O7iC_EDBFoX7-hMK9BObI8uBSRdBcR?usp=drive_link