

<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *23rd IEEE Conference on Pervasive and Intelligent Computing, PICom 2025, Hakodate City, Nov 21-24.*

Citation for the original published paper:

Jamil, E., Garcia-Blas, J., Alawadi, S., Carretero, J. (2025)

A Deep Experimental Study of Ensemble-Based Phishing Detection in Centralised and Federated Settings

In: *Proceedings - 2025 IEEE Conference on Pervasive and Intelligent Computing, PICom 2025* (pp. 145-152). Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/PICom68402.2025.00025>

N.B. When citing this work, cite the original published paper.

©2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:bth-29303>

A Deep Experimental Study of Ensemble-Based Phishing Detection in Centralised and Federated Settings

Esraa Jamil*, Javier Garcia-Blas[†], Sadi Alawadi[‡], Jesus Carretero[†]

*Department of Programming, Al al-Bayt University, Jordan

Email: esraa_jamil@aabu.edu.jo

[†]Computer Science and Engineering Dept., Universidad Carlos III de Madrid, Madrid, Spain

Email: {fjblas, jcarrete}@inf.uc3m.es

[‡]Computer Science Department Blekinge Tekniska Högskola (BTH), Sweden

Email: sadi.alawadi@bth.se

Abstract—Phishing attacks remain a major cybersecurity threat in today’s increasingly interconnected world, exploiting human vulnerabilities to bypass traditional defenses. This study evaluates a range of Machine Learning and Deep Learning models for phishing detection using three datasets: phishing-only, legitimate-only, and a balanced merged dataset. The experiments were carried out using centralized and FL approaches, with further robustness evaluation under adversarial scenarios such as phishing injection and label-flipping attacks. The results demonstrate that DL models, particularly LSTM and MLP, delivered strong performance in both stand-alone and ensemble setups. Notably, FL models preserved comparable performance and even slightly better accuracy than centralized models, while enhancing privacy, scalability, and robustness. Under adversarial scenarios, FL models maintained high robustness, validating the practical effectiveness of privacy-preserving phishing detection in real-world environments. In centralized settings, LSTM achieved up to 96.00% accuracy on phishing data, while heterogeneous DL ensembles (LSTM, MLP, CNN) further enhanced performance to 97.13%. This advantage was especially evident during cross-distribution evaluations, where LSTM bagging showed strong generalization, achieving up to 85.73% accuracy. In federated settings, LSTM models reached up to 86.75% accuracy and an F1-score around 86%. Even under phishing injection and data poisoning attacks, the 10-client LSTM model showed the best balance with 83.33% accuracy, 85.67% F1-score, and the highest AUC of 91.40%. The 10-client CNN and 15-client MLP models also performed well, with accuracies around 83.7% and F1-scores above 85.5%, with MLP reaching an AUC of 91.13%.

Index Terms—phishing, federated learning, cybersecurity, machine learning, social engineering, privacy-preserving

I. INTRODUCTION

Nowadays, the rapid advancement of technology and widespread internet usage over the last decade have significantly expanded network scale and associated applications, with users exceeding two billion [1]. According to [2], internet users reached approximately 4.9 billion in 2020, highlighting the significant growth of digital connectivity in modern life. In particular, the COVID-19 pandemic has significantly increased

the reliance of people on online services, resulting in a 20% increase in the use of the internet from 50% to 70% during this period [3]. This wave of digitalisation not only transformed societal behaviour but also led to exponential growth in data generation. Based on the Dataprot report, the global deployment of Internet of Things (IoT) devices will exceed 29 billion by 2030, contributing to a data explosion [4].

However, this interconnected digital landscape has also introduced significant vulnerabilities. As data volumes and system complexity increase, cyberattacks have become more frequent, targeted, and sophisticated, revealing critical weaknesses in traditional security infrastructures [1] and posing substantial challenges for network security in accurately detecting intrusions. As an intrusion can be considered any attempt to compromise the confidentiality, integrity, or availability (CIA) of data, computer, network, or more to bypass the implemented security mechanisms [5]. The growing threats highlight the urgent need for advanced and adaptive intrusion detection systems capable of identifying and mitigating such attacks promptly.

Cyberattacks may target vulnerable computer networks, or they may exploit human vulnerabilities, as seen in social engineering (SE) attacks. SE attacks involve psychologically manipulating victims into compromising security measures. This manipulation can result in gaining unauthorized access to the target’s information or injecting malware into their system rather than directly attacking the system. Therefore, SE attacks exploit human emotions such as fear, trust, or curiosity to compromise the CIA of data [6].

SE attacks, especially phishing, have emerged as one of the most prevalent and damaging forms of cyber intrusion over the past decade [7]. These attacks can impose severe consequences, from financial loss, as in cases reported by the U.S. Federal Bureau of Investigation (FBI) where impersonators caused over \$2.3 billion in damages, to life-threatening disruptions in sectors like remote healthcare, where the Internet of Medical Things (IoMT) is increasingly used [8],

This work has been partially funded by the I+D+i project PID2022-138050NB-I00 (New scalable I/O techniques for hybrid HPC and data intensive workloads - SCIOT).

[9]. The “Ryuk” ransomware attacks on hospitals in France and California are stark examples of the real-world impact of such cyber threats. In response, regulatory frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) have been introduced to enforce stricter privacy protections [9].

Given the dynamic and evolving nature of cyber threats, particularly phishing, there is a critical need for intelligent and adaptive security solutions. Intrusion Detection Systems (IDSs) have been developed as frontline defenses, capable of detecting anomalous behaviour and cyberattacks such as phishing, Denial of Service (DoS), and malware. By continuously analysing network traffic, IDSs play a vital role in preserving the CIA triad and ensuring the security of digital environments [10]. However, current IDS technologies must evolve to cope with the growing sophistication of attacks and protect increasingly complex infrastructures.

In this study, we devised an experimental methodology and a comprehensive set of experiments to evaluate the performance of various traditional ML and DL techniques, including Random Forest, Bagging, LSTM, and CNN, under a centralized and federated learning (FL) paradigm using a phishing attack dataset. Moreover, we extended the study scope by conducting more experiments under FL environments to ensure data privacy preservation, knowledge sharing across distributed parties, and efficient use of edge node resources without exposing raw data. The FL-trained models achieved performance metrics comparable to their centralised counterparts, with only a slight drop in accuracy. This is an acceptable trade-off considering the benefits of privacy, scalability, and computational cost. Notably, the federated model demonstrated greater robustness and the ability to generalize to previously unseen samples, making it well-suited for real-world scenarios where data distributions are dynamic and privacy constraints are critical.

The remainder of this paper is organized as follows. Section II provides a general overview of phishing attacks and FL. Section III summarizes the current research in phishing attack detection techniques. Section IV details the experimental methodology, including descriptions of the datasets, experimental phases, and evaluation metrics. Section V presents and discusses the results obtained from the experiments that were carried out. Finally, Section VI concludes the paper, summarizes key findings and outlines potential future directions.

II. BACKGROUND

This section provides the necessary background to facilitate a clear understanding of the concepts explored in this paper. It introduces and explains key topics relevant to the study, including phishing attacks and federated learning.

A. Phishing Attacks

Since their emergence in the mid-1990s, phishing attacks have become one of the most sophisticated and widespread

threats, targeting many internet users, governmental institutions, and service providers [11]. Beyond causing substantial financial losses, phishing also leads to reputational damage and a loss of trust [12]. Increasingly, phishing serves as an entry point for launching more advanced cyberattacks, such as malware infections, credential theft, and network infiltration [13]. Due to its dynamic nature, phishing lacks a universally accepted definition; the definitions vary depending on context and application. Generally, phishing can be defined as the process of deceiving targets into performing actions that benefit the attacker [14].

Phishing attacks are particularly dangerous due to their simplicity and effectiveness, often exploiting human errors rather than technical vulnerabilities. Techniques such as typosquatting (e.g., replacing “w” with “vv”) in URLs are used to mimic legitimate websites and deceive unsuspecting users [15]. Those attacks manifest in various forms, including fake websites, spoofed emails, QR phishing, SMS phishing, and DNS spoofing [16]. Among them, fraudulent emails remain the most common form of phishing, imitating legitimate sources to prompt victims into clicking malicious links or opening harmful attachments [17]. The COVID-19 pandemic significantly exacerbated this threat landscape, with phishing email volume reportedly increasing by over 600% during this period [18]. A typical phishing message might be: *“Dear valued customer, we have detected unusual activity in your bank account. To ensure uninterrupted service, please verify your details by clicking the link below”* [3]. Such messages prey on emotional triggers like urgency and fear, leading users to act without properly verifying the sender’s identity. Therefore, many phishing victims often fall prey to phishing attacks due to an insufficient assessment of the sender’s identity and a lack of proper education to recognize such threats.

B. Federated Learning

Federated Learning is a decentralized Machine Learning (ML) paradigm introduced by Google’s research team, designed to enable collaborative model training across distributed devices while preserving data privacy [19]. Unlike traditional centralized approaches that require raw data to be transferred to a central server, FL allows data to remain on local devices, such as smartphones, edge servers, or PCs, while only model updates are shared, meaning that the ML model moves to the data site. In a typical FL setup, a central server initializes an ML model with random weights and orchestrates iterative training rounds. In each round, the server involves all participants or selects a subset of participating clients and transmits the current model to them [20]. Clients then perform local training using their private data, commonly via Stochastic Gradient Descent (SGD), and return only the model updated parameters, weights, or gradients. These updates are aggregated at the server using techniques such as Federated Averaging (FedAvg), and the process is repeated iteratively until global model convergence [5]. This framework enables resource-constrained clients to benefit from global knowledge without compromising data privacy. By transmitting only

model parameters instead of raw data, FL significantly reduces communication overhead and mitigates privacy risks [21], [22]. Moreover, aggregating updates from a diverse range of local datasets improves the generalization capability of the global model. In security-sensitive domains, such as intrusion detection, this enables continual model refinement with real-time attack patterns gathered from heterogeneous sources [23].

III. RELATED WORK

Traditional ML and DL techniques have been extensively used for phishing detection. For instance, Butnaru et al. [24] trained five ML models using Kaggle and PhishTank datasets, with RF outperforming other classifiers. Yang et al. [25] presented a CNN-LSTM method with XGBoost, achieving high detection rates on PhishTank and DMOZ. He et al. [26] integrated Bi-LSTM with XGBoost to detect phishing and insider threats, achieving 98.38% accuracy. Maini et al. [27] applied a voting ensemble method of eight models, including RF, XGBoost and AdaBoost, achieving 93.6% accuracy.

Due to increasing concerns about data security and privacy, FL has attracted significant attention as a privacy-preserving approach. Thapa et al. [18] proposed an FL framework for phishing email detection using DL such as RNN, BERT and THEMIS. The framework was comprehensively evaluated across different settings, including data distribution, client numbers, and communication overhead, using diverse datasets like Nazario’s, Enron and Phishbowl. The study found that increasing the number of participants slightly reduced the accuracy of model convergence. Transfer learning was also shown to improve convergence precision. Overall, FL achieved high accuracy levels comparable to centralized learning, particularly with fewer organizations and balanced datasets. For example, THEMIS reached 97.9% accuracy, while BERT achieved 96.1% accuracy. Future work was suggested to enhance privacy by integrating additional techniques such as homomorphic encryption (HE) or differential privacy (DP). Lobner et al. [29] extended FL with local differential privacy (LDP) to enhance privacy in spam and ham email classification, effectively addressing FL’s vulnerability to gradient-based attacks. Using the 2020 Enron dataset (33,722 emails: 50.9% spam, 49.1% ham), the authors demonstrated that the performance and accuracy losses—common challenges associated with differential privacy—can be mitigated by applying a client-level F1-score threshold. In another decentralized phishing email detection study, Sun et al. [30] suggested the Federated Phish Bowl (FedPB) framework, which utilized global word embedding and long short-term memory (LSTM). FedPB achieved a competitive detection accuracy of 83% under varying numbers of participants and data heterogeneity. The dataset of 1,188 samples included 594 phishing emails collected from Microsoft 365 anti-phishing protection and 594 legitimate emails randomly selected from the Enron dataset. For future work, they aimed to investigate asynchronous learning to reduce communication delays between clients and the parameter server.

Yoon et al. [21] proposed an FL algorithm for voice phishing detection and conducted experiments using the CIFAR10 dataset, dividing it into two scenarios: one with balanced data consisting of 5,000 samples with similar features and another with unbalanced data ranging from 1,000 to 9,000 samples with varied features.

Khramtsova et al. [31] applied FL in Security Operation Centers (SOCs) to detect malicious URLs, using a dataset of over 700,000 URLs from sources such as PhishTank and URLHaus. The study examined various data partitioning scenarios and found that URL classification improved by up to 27% in all cases, with gains reaching up to 30% for smaller organizations, while a minor decrease of 0.5% was observed for larger agents. In future work, the authors plan to explore advanced aggregation methods, increase feature complexity, and investigate FL applications in other cybersecurity domains such as edge computing.

However, most prior studies rely on limited datasets or lack comprehensive adversarial robustness evaluations. In contrast, our study provides a systematic evaluation of phishing detection performance across three distinct dataset configurations: phishing-only, legitimate-only, and balanced merged datasets. We assess performance in both centralized and FL environments using diverse deep learning models, including ensemble variants. Furthermore, we evaluate model robustness against phishing injection and label-flipping attacks across varying FL client configurations (5, 10, and 15 clients). This comprehensive methodology provides a more realistic and privacy-preserving evaluation framework that addresses key limitations in existing literature. Table I presents a detailed comparison of related works, highlighting their methodological approaches, datasets utilized, and primary findings.

IV. EXPERIMENTAL METHODOLOGY

The following section presents the experimental methodology proposed in this work. It outlines the dataset employed for evaluation and provides a detailed description of the methodological approach used to conduct the experiments.

A. Datasets

We have used three datasets to evaluate the performance of phishing detection models: (1) a phishing URL dataset, (2) a legitimate URL dataset, and (3) a merged balanced dataset. The *phishing URL dataset* was collected from PhishTank, a platform that maintains a regularly updated repository of verified phishing URLs. The dataset comprises a total of 5,000 phishing samples that were randomly selected. The *legitimate URL dataset* was obtained from the CIC-URL-2016 dataset provided by the University of New Brunswick, which includes various types of URLs (e.g., benign, phishing, malware, spam). From the benign class, 5,000 legitimate URLs were randomly sampled from a total of 35,300 entries. To enable binary classification, the two datasets were combined into a *balanced dataset* of 10,000 entries, comprising equal parts of phishing and legitimate URLs (5,000 from each). The class labels were encoded as 0 for legitimate and 1 for phishing.

TABLE I: Summary of related works on Phishing Detection using ML, DL, and FL techniques.

Author(s)	Method(s)	Dataset(s)	Key Findings
Butnaru et al. [24]	RF, DT, SVM, etc.	Kaggle, PhishTank	RF outperformed other classifiers, achieving 99.29% accuracy.
Yang et al. [25]	CNN-LSTM + XGBoost	PhishTank, DMOZ	Achieved 98.99% accuracy using multidimensional URL features.
He et al. [26]	Bi-LSTM + XGBoost	Enron, Monkey.org	Detected phishing and insider threats with 98.38% accuracy.
Maini et al. [27]	Voting ensemble (RF, XGBoost, AdaBoost)	PhishTank, OpenPhish, moz.com, CIC-URL	The ensemble method achieved 93.6% accuracy.
Thaseen et al. [28]	ANN + Correlation-Based Feature Selection	NSL-KDD, UNSW-NB15	Achieved 98.45% and 96.44% accuracy; outperformed SVM, DT, and RF.
Thapa et al. [18]	FL with BERT, RNN, THEMIS	Enron, Nazario, Phishbowl	THEMIS reached 97.9% and BERT 96.1% accuracy; FL showed strong performance with few clients and balanced data.
Lobner et al. [29]	FL + Local Differential Privacy	Enron (2020)	Applied client-level F1-score threshold to reduce performance loss; 33,722 emails analyzed.
Sun et al. [30]	FedPB (FL + LSTM)	Microsoft 365, Enron	Achieved 83% accuracy; addressed data heterogeneity with global word embeddings.
Yoon et al. [21]	FL for Voice Phishing Detection	CIFAR10	Evaluated FL under balanced and unbalanced client data scenarios.
Khramtsova et al. [31]	FL in SOCs	PhishTank, URLHaus	FL improved URL detection by up to 30% for smaller organizations.

For each URL, 17 numerical features were extracted and grouped into three categories: address bar-based features (9), domain-based features (4), and HTML/JavaScript-based features (4). To maintain numerical consistency, all textual fields, such as domain names and raw URL strings, were excluded from the final dataset. All features were normalized using MinMax scaling to ensure compatibility with ML and deep learning models.

B. Methodology

The experimental methodology devised in this work was organized into four phases to assess phishing detection performance. All experiments have used an 80%-20% train-test split combined with 5-fold cross-validation to ensure robustness and generalizability. Model performance was evaluated using well-known evaluation metrics, including Accuracy, Precision, Recall, F1-Score, ROC-AUC and Cohen’s Kappa, where applicable.

1) **Phase 1:** In the first phase, both traditional ML models (RF, SVM) and DL models (LSTM, CNN, MLP) were trained and evaluated under centralized conditions. Four experimental configurations were considered:

- **Phishing-only:** models trained and tested exclusively on phishing samples.
- **Legitimate-only:** models were trained and tested on legitimate samples.
- **Legitimate-to-Phishing :** models were trained on legitimate samples and tested on phishing samples to assess cross-distribution generalization.
- **Merged Dataset:** A balanced dataset combining phishing and legitimate samples for binary classification tasks.

- 2) **Phase 2:** In this phase, ensemble learning was applied using the Bagging technique to enhance predictive performance and model stability. The experiments involved both homogeneous ensembles (e.g., LSTM bagging) and heterogeneous ensembles (e.g., LSTM, CNN and MLP combined). The same four datasets from Phase 1 were used to maintain consistency in comparative evaluation.
- 3) **Phase 3:** FL was implemented using the Flower framework, using the FedAvg aggregation algorithm. The balanced dataset was partitioned across 5, 10, and 15 simulated clients to emulate decentralized learning environments. DL models (LSTM, CNN and MLP) were trained over 10 communication rounds.
- 4) **Phase 4:** To assess model robustness and generalization under adversarial conditions, two experimental scenarios were introduced:

- **Unseen Phishing Samples:** A set of 500 previously unseen phishing URLs, held out from initial training, was injected during the test phase to evaluate the ability of centralized and federated models to generalize to novel threats.
- **Label-Flipping Attacks:** Simulated adversarial scenarios in which a subset of client datasets included deliberately mislabelled data to assess the resilience of FL models to poisoning attacks.

V. RESULTS AND DISCUSSION

This section presents the performance evaluation of various ML and DL models under both centralized and FL frameworks.

As shown in Figure 1, DL models demonstrated superior performance on the phishing dataset. Among the individual

models, LSTM performed best with 96.00% accuracy and an F1-score of 97.96%, followed closely by One-Class SVM (OCSVM) and Isolation Forest, both achieving over 95.70% accuracy. Ensemble bagging further enhanced results, particularly the heterogeneous ensemble of LSTM, MLP, and CNN, which reached 97.13% accuracy. However, adding classical models to the ensemble (LSTM-MLP-CNN-ISO-OCSVM) resulted in a slight accuracy drop to 97.00%.

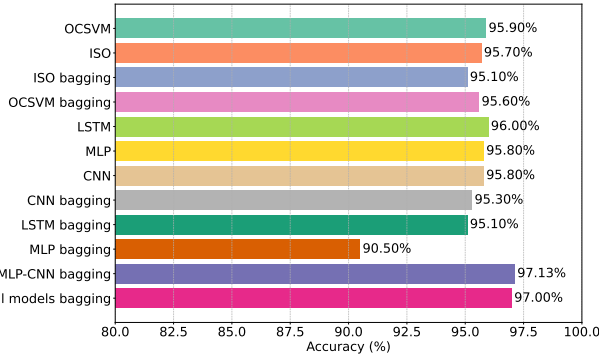


Fig. 1: Accuracy comparison of models and their bagging ensembles on the phishing dataset under centralized learning.

TABLE II: Performance comparison of individual models and their bagging ensembles trained on legitimate data and evaluated on phishing samples under centralized learning.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Kappa
Random Forest (RF)	63.67%	70.51%	46.95%	56.33%	63.67%	27.35%
SVM	72.72%	95.06%	70.95%	81.24%	76.25%	34.85%
RF Bagging	64.51%	72.24%	47.10%	56.98%	64.51%	29.03%
SVM Bagging	73.08%	94.73%	71.70%	81.60%	75.85%	34.82%
LSTM	85.38%	86.86%	97.84%	92.02%	17.28%	8.06%
MLP	85.59%	86.78%	98.25%	92.16%	29.36%	7.24%
CNN	85.71%	87.13%	97.88%	92.19%	29.84%	11.10%
LSTM Bagging	85.73%	87.14%	97.90%	92.21%	30.02%	11.20%
MLP Bagging	85.41%	86.72%	98.10%	92.06%	29.39%	6.44%
CNN Bagging	85.68%	87.14%	97.82%	92.17%	29.60%	11.22%
LSTM-MLP-CNN Bagging	85.69%	87.14%	97.84%	92.18%	29.53%	11.22%
LSTM-MLP-CNN-ISO-OCSVM Bagging	82.60%	83.21%	99.12%	90.47%	54.31%	1.42%

Figure 2 plots the results of training and evaluating models on legitimate samples. OCSVM and Isolation Forest again showed strong performance, achieving 96.00% and 95.50% accuracy, respectively. In bagging mode, Isolation Forest slightly improved to 95.80%, while OCSVM bagging experienced a decrease to 93.60%. DL models recorded slightly lower individual scores, but the heterogeneous bagging ensemble (LSTM-MLP-CNN) improved the results to 96.50%. The highest performance was achieved by the hybrid LSTM, MLP, CNN, ISO, and OCSVM models, reaching 97.62% accuracy with an F1-score of 98.73%.

In the generalization experiment (see Table II), models were trained on legitimate data and tested on phishing samples. Traditional ML models such as RF and SVM showed limited capability, achieving 63.67% and 72.72% accuracy, respectively. In contrast, DL models demonstrated greater robustness. LSTM Bagging achieved the highest accuracy (85.73%) and the best F1-score (92.21%). Notably, extending the ensemble to include traditional ML models reduced accuracy to 82.60%

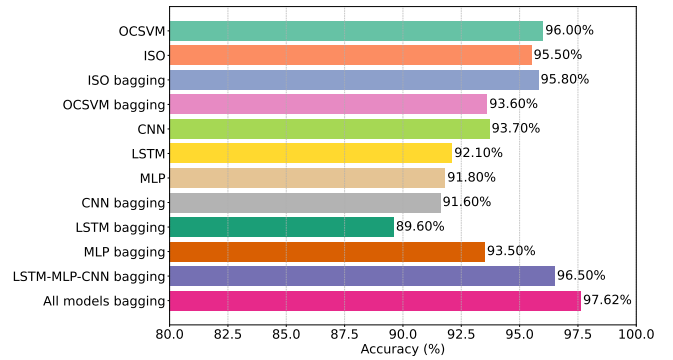


Fig. 2: Accuracy comparison of models and their bagging ensembles on the legitimate dataset under centralized learning.

but increased recall to 99.12%, reflecting higher sensitivity to phishing instances.

Table III depicts the results on the merged phishing-legitimate dataset. Classical classifiers demonstrated moderate performance; RF bagging achieved the highest accuracy (86.80%), with a strong Kappa score of 73.60%. Deep learning models continued to perform well, with LSTM bagging leading the group by achieving 84.95% accuracy and an F1-score of 83.02%, confirming that bagging improves model stability. While ensemble methods enhanced overall performance, combining traditional ML and DL models offered no significant additional gains.

TABLE III: Performance comparison of individual models and their bagging ensembles on the merged dataset under centralized learning.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Kappa
Random Forest (RF)	86.75%	92.49%	80.00%	85.79%	86.75%	73.50%
SVM	82.40%	96.42%	67.30%	79.27%	87.64%	64.80%
RF Bagging	86.80%	92.69%	79.90%	85.82%	92.50%	73.60%
SVM Bagging	82.65%	95.41%	68.60%	79.81%	87.43%	65.30%
LSTM	84.30%	96.73%	71.00%	81.89%	89.68%	68.60%
CNN	83.20%	97.98%	67.80%	80.14%	88.41%	66.40%
MLP	83.10%	96.75%	68.50%	80.21%	83.10%	66.20%
LSTM Bagging	84.95%	97.87%	72.09%	83.02%	89.96%	70.06%
MLP Bagging	84.00%	97.69%	70.32%	81.78%	89.00%	68.18%
CNN Bagging	83.05%	97.42%	67.90%	80.02%	87.91%	66.10%
LSTM-MLP-CNN Bagging	83.20%	96.89%	68.60%	80.33%	88.06%	66.40%
LSTM-MLP-CNN-RF-SVM Bagging	82.97%	97.14%	67.95%	79.95%	90.03%	65.95%

Across all centralized experiments, DL models consistently outperformed traditional ML classifiers, including both one-class and binary models, particularly in generalization scenarios. Bagging further improved model performance, with heterogeneous ensembles, especially those combining multiple DL models, showing notable improvements in both accuracy and F1-score. While incorporating classical models into ensembles occasionally improved recall, it could slightly reduce accuracy. Cross-domain testing (training on legitimate data and testing on phishing samples) revealed that classical models struggled to generalize, whereas DL models demonstrated strong robustness. In the balanced binary classification scenario using the merged dataset, all models performed well, validating the effectiveness of centralized learning for phishing detection.

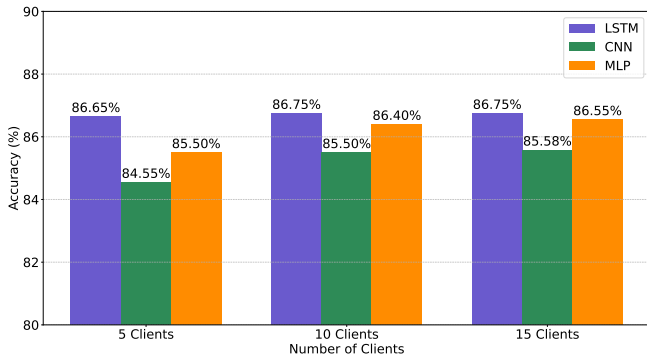


Fig. 3: Accuracy comparison of the LSTM, CNN, and MLP model under FL using 5, 10, and 15 clients. The experiment was conducted using 10 global aggregation rounds.

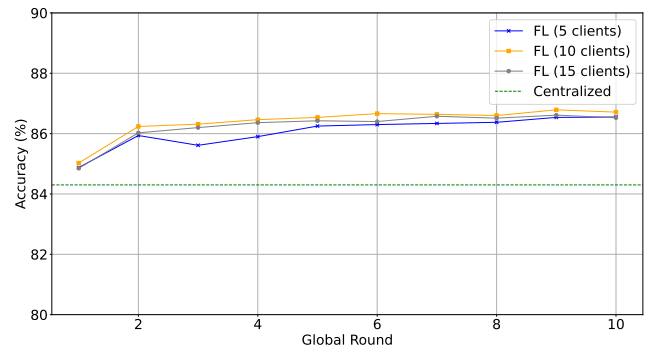
Figure 3 presents the performance of LSTM, CNN and MLP models under the FL framework with 5, 10, and 15 clients over 10 communication rounds on the merged dataset. Among the models, LSTM consistently achieved the highest accuracy, reaching up to 86.75%, with strong precision over 91% and recall around 81%, resulting in superior F1-scores of approximately 86% and Kappa values above 73%. closely followed by MLP with a peak accuracy of 86.55%, supported by similarly high precision and recall rates. CNN, while showing competitive performance, recorded slightly lower performance, with accuracy at 85.58% and the lowest Kappa values among the three models.

Overall, LSTM stands out as the top-performing DL model for phishing detection in both centralized and federated settings, with MLP as a strong alternative and CNN providing a competitive option. All three models showed stable accuracy across several client counts, proving their effectiveness in decentralized, privacy-preserving environments.

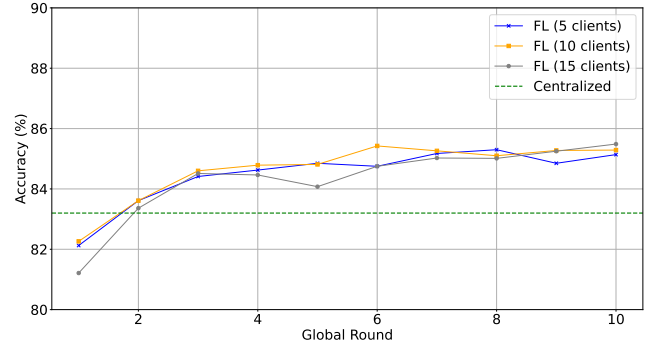
Additionally, these trends demonstrate that FL can offer not only comparable but often superior performance to centralized learning. They also highlight the robustness and scalability of FL for phishing URL detection in privacy-preserving environments.

To evaluate the robustness of both centralised and FL models under adversarial conditions, 500 previously unseen phishing samples, held out from the original dataset, were injected into the test set. Additionally, label flipping attacks were simulated by altering the labels within the local datasets of two clients. Figures 5 compare the accuracy of LSTM, CNN, and MLP models in centralized and FL setups (5, 10, and 15 clients) before and after phishing data injection. Detailed results are provided in Tables IV to V.

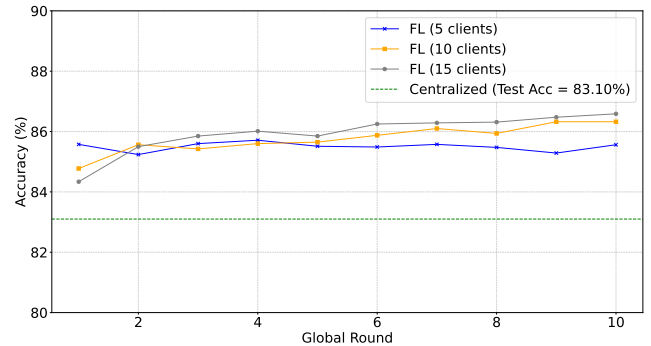
The LSTM model in FL with 10 clients initially recorded the highest accuracy 85.89% and precision 93.17% before any attack. Following the injection of held-out phishing samples and the simulation of label flipping, the model still maintained a high F1-score of 85.67% accuracy of 83.33%, and recall of 85.43%, indicating notable robustness. Similar trends were ob-



(a) LSTM model accuracy.



(b) CNN model accuracy.



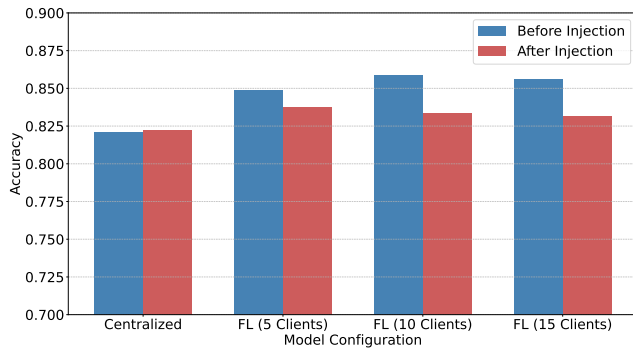
(c) MLP model accuracy.

Fig. 4: Accuracy over global rounds for DL models in federated learning with different numbers of clients (5, 10, and 15). (a) LSTM model. (b) CNN model. (c) MLP model.

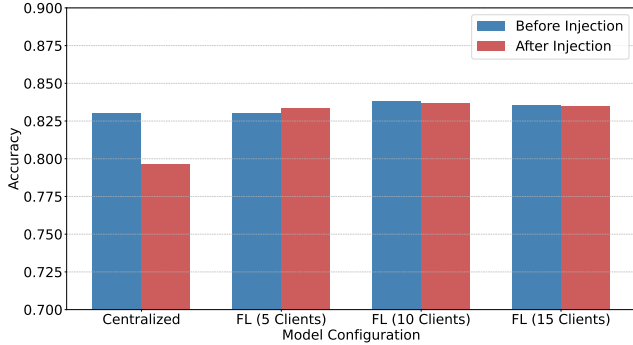
served with the CNN and MLP models. Following the attack, both models sustained F1-scores above 85% across across all clients except MLP with 5 clients. The AUC values remained stable, with the MLP model reaching a peak of 91.13%. These findings confirm FL's resilience against adversarial attacks, performing comparably or better than centralized methods under such conditions.

VI. CONCLUSION AND FUTURE WORK

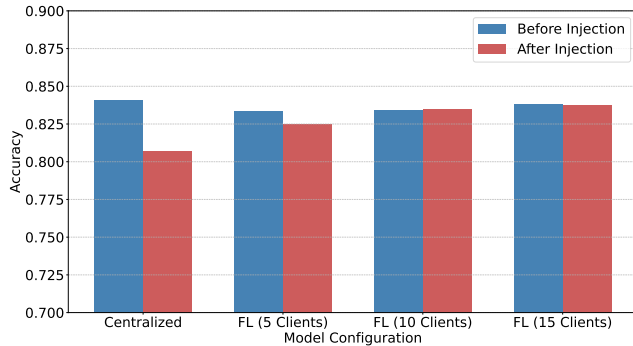
This paper presents a multi-phase evaluation methodology for phishing detection models under both centralised and FL environments. A variety of traditional ML and DL



(a) LSTM model accuracy.



(b) CNN model accuracy.



(c) MLP model accuracy.

Fig. 5: Accuracy comparison of centralized and federated DL models before and after adversarial scenarios. (a) LSTM model. (b) CNN model. (c) MLP model.

TABLE IV: Performance of LSTM, CNN, and MLP models in FL after holding out phishing samples.

Model	Clients	Accuracy	Precision	Recall	F1-Score	AUC	Kappa
LSTM	5	84.89%	87.98%	78.89%	83.19%	90.12%	69.50%
	10	85.89%	93.17%	75.78%	83.58%	91.18%	71.44%
	15	85.63%	92.65%	75.67%	83.30%	90.98%	70.92%
CNN	5	83.05%	82.99%	80.78%	81.87%	89.72%	66.00%
	10	83.79%	84.99%	79.89%	82.36%	90.34%	67.40%
	15	83.58%	84.35%	80.22%	82.23%	90.31%	67.00%
MLP	5	83.32%	84.66%	79.11%	81.79%	89.35%	66.43%
	10	83.42%	83.20%	81.44%	82.31%	90.36%	66.71%
	15	83.84%	84.36%	80.89%	82.59%	90.55%	67.53%

architectures, including ensemble configurations, were ex-

TABLE V: Performance of LSTM, CNN, and MLP models in FL after adversarial scenarios.

Model	Clients	Accuracy	Precision	Recall	F1-Score	AUC	Kappa
LSTM	5	83.75%	91.94%	79.07%	85.02%	90.23%	67.50%
	10	83.33%	85.92%	85.43%	85.67%	91.40%	65.80%
	15	83.13%	85.66%	85.36%	85.51%	91.32%	65.30%
CNN	5	83.33%	88.52%	82.07%	85.17%	90.38%	66.20%
	10	83.71%	88.48%	82.86%	85.58%	90.80%	66.90%
	15	83.46%	87.12%	84.07%	85.57%	90.83%	66.20%
MLP	5	82.50%	85.10%	84.86%	84.98%	90.19%	64.02%
	10	83.50%	88.62%	82.29%	85.33%	90.99%	66.54%
	15	83.75%	89.09%	82.21%	85.51%	91.13%	67.08%

amined across multiple scenarios, including phishing-only, legitimate-only, and merged datasets. LSTM consistently outperformed other models in both centralised and FL paradigms, while MLP and CNN also demonstrated competitive results. The FL models, implemented using the Flower framework, achieved performance comparable to their centralised counterparts while preserving data privacy and enhancing scalability. Robustness evaluations under adversarial conditions, including the injection of unseen phishing data and label-flipping attacks, demonstrated the resilience of the FL models, which exhibited only slight performance degradation.

For future work, we intend to extend this study by incorporating advanced privacy-enhancing methods such as differential privacy and secure aggregation. Furthermore, investigating asynchronous FL approaches could further improve scalability and robustness, particularly in settings with larger client pools such as smart cities.

REFERENCES

- [1] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, p. e4150, 2021.
- [2] M. T. Jafar, M. Al-Fawa'reh, M. Barhoush, and M. H. Alshira'h, "Enhanced analysis approach to detect phishing attacks during covid-19 crisis," *Cybernetics and Information Technologies*, vol. 22, no. 1, pp. 60–76, 2022.
- [3] R. Goenka, M. Chawla, and N. Tiwari, "A comprehensive survey of phishing: mediums, intended targets, attack and defence techniques and a novel taxonomy," *International Journal of Information Security*, pp. 1–30, 2023.
- [4] M. Al-Fawa'reh, J. Abu-Khalaf, P. Szweczyk, and J. J. Kang, "Malbot-drl: Malware botnet detection using deep reinforcement learning in iot networks," *IEEE Internet of Things Journal*, 2023.
- [5] S. Agrawal, S. Sarkar, O. Aouedi, G. Yenduri, K. Piamrat, M. Alazab, S. Bhattacharya, P. K. R. Maddikunta, and T. R. Gadekallu, "Federated learning for intrusion detection system: Concepts, challenges and future directions," *Computer Communications*, 2022.
- [6] P. Burda, L. Allodi, and N. Zannone, "Cognition in social engineering empirical research: a systematic literature review," *ACM Transactions on Computer-Human Interaction*, vol. 31, no. 2, pp. 1–55, 2024.
- [7] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers & Security*, vol. 68, pp. 160–196, 2017.
- [8] R. Abdilllah, Z. Shukur, M. Mohd, and T. M. Z. Murah, "Phishing classification techniques: A systematic literature review," *IEEE Access*, vol. 10, pp. 41 574–41 591, 2022.
- [9] A. López Martínez, M. Gil Pérez, and A. Ruiz-Martínez, "A comprehensive review of the state-of-the-art on security and privacy issues in healthcare," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [10] J. Hautsalo, "Using supervised learning and data fusion to detect network attacks," 2021.

- [11] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139–154, 2021.
- [12] A. I. Elkhawas, T. M. Chen, and I. Gashi, "Privacy-preserving federated learning for phishing detection," *IEEE Technology and Society Magazine*, 2025.
- [13] B. Naqvi, K. Perova, A. Farooq, I. Makhdoom, S. Oyediji, and J. Porras, "Mitigation strategies against the phishing attacks: A systematic literature review," *Computers & Security*, p. 103387, 2023.
- [14] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, p. 563060, 2021.
- [15] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or not phishing? A survey on the detection of phishing websites," *IEEE Access*, vol. 11, pp. 18 499–18 519, 2023.
- [16] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University-Computer and Information Sciences*, 2023.
- [17] S. Zhuo, R. Biddle, Y. S. Koh, D. Lottridge, and G. Russello, "Sok: Human-centered phishing susceptibility," *ACM Transactions on Privacy and Security*, vol. 26, no. 3, pp. 1–27, 2023.
- [18] 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, p. 4346, 2023.
- [19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [20] M. Tahir, T. Mawla, F. Awaysheh, S. Alawadi, M. Gupta, and M. I. Ali, "SecureFedPROM: A Zero-Trust Federated Learning Approach with Multi-Criteria Client Selection," *IEEE Journal on Selected Areas in Communications*, 2025.
- [21] J. Y. Yoon and B. J. Choi, "Privacy-friendly phishing attack detection using personalized federated learning," in *International Conference on Intelligent Human Computer Interaction*. Springer, 2022, pp. 460–465.
- [22] S. Alawadi, A. Ait-Mlouk, S. Toor, and A. Hellander, "Toward efficient resource utilization at edge nodes in federated learning," *Progress in Artificial Intelligence*, vol. 13, no. 2, pp. 101–117, 2024.
- [23] R. Doriguzzi-Corin and D. Siracusa, "Flad: adaptive federated learning for ddos attack detection," *Computers & Security*, vol. 137, p. 103597, 2024.
- [24] A. Butnaru, A. Mylonas, and N. Pitropakis, "Towards lightweight url-based phishing detection," *Future internet*, vol. 13, no. 6, p. 154, 2021.
- [25] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE access*, vol. 7, pp. 15 196–15 209, 2019.
- [26] D. He, X. Lv, X. Xu, S. Chan, and K.-K. R. Choo, "Double-layer Detection of Internal Threat in Enterprise Systems Based on Deep Learning," *IEEE Transactions on Information Forensics and Security*, 2024.
- [27] A. Maini, N. Kakwani, B. Ranjitha, M. Shreya, and R. Bharathi, "Improving the performance of semantic-based phishing detection system through ensemble learning method," in *2021 IEEE mysore sub section international conference (MysuruCon)*. IEEE, 2021, pp. 463–469.
- [28] I. Sumaiya Thaseen, J. Saira Banu, K. Lavanya, M. Rukunuddin Ghalib, and K. Abhishek, "An integrated intrusion detection system using correlation-based attribute selection and artificial neural network," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 2, p. e4014, 2021.
- [29] S. Löbner, B. Gogov, and W. B. Tesfay, "Enhancing Privacy in Federated Learning with Local Differential Privacy for Email Classification," in *International Workshop on Data Privacy Management*. Springer, 2022, pp. 3–18.
- [30] Y. Sun, N. Chong, and H. Ochiai, "Federated Phish Bowl: LSTM-Based Decentralized Phishing Email Detection," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2022, pp. 20–25.
- [31] E. Khrantsova, C. Hammerschmidt, S. Lagraa, and R. State, "Federated learning for cyber security: Soc collaboration for malicious url detection," in *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2020, pp. 1316–1321.