RESEARCH ARTICLE

A Novel System for Detecting Model Poisoning Attacks in Federated Learning

Jagdeep Singh[1], **Saru Kumari**[2] **and Seema Agrawal**[3,*]

[1] Department of Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, India
[2] Department of Mathematics, Chaudhary Charan Singh University, Meerut, Uttar Pradesh, India
[3] Department of Mathematics, S.S.V. College, Uttar Pradesh, India

Abstract

Federated learning (FL) enables decentralized model training and enhances user privacy by keeping data on local devices. Despite these advantages, FL remains vulnerable to sophisticated adversarial attacks. Federated recommender systems (FRS), an important application of FL, are particularly susceptible to threats such as model poisoning. In this paper, we propose DyMUSA, a novel model poisoning attack tailored for FRS. DyMUSA exploits systemic vulnerabilities through dynamic user selection and adaptive poisoning strategies. Specifically, it leverages the Isolation Forest algorithm to identify anomalous users and generate poisoned gradients that compromise the integrity of the recommender system. Experiments conducted on real-world datasets demonstrate that DyMUSA significantly increases the exposure of targeted items while maintaining minimal impact on overall system performance.

Keywords: federated learning, attacks, privacy, real world data, datasets.

1 Introduction

Federated learning (FL) enables collaborative model training while preserving user privacy by keeping data on local devices and only exchanging model updates. This decentralized approach is particularly valuable in recommender systems (RS), which leverage FL to deliver personalized content on platforms like e-commerce, social media, and streaming services by analyzing user behaviors and preferences.

Federated recommender systems (FRS) enhance privacy and security by decentralizing the training process. However, this decentralization also introduces new challenges, especially in defending against sophisticated attacks such as model poisoning. These attacks manipulate the recommendation process by injecting malicious data or altering model parameters, which compromises the system's integrity and reliability. While traditional centralized RS are vulnerable due to their reliance on centralized data aggregation, FRS are not immune to adversarial actions, particularly those targeting the collaborative learning process.

Submitted: 20 November 2025
Accepted: 21 January 2026
Published: 07 February 2026

Vol. 2, **No.** 1, 2026.
10.62762/JRSC.2025.385825

***Corresponding author:**
✉Seema Agrawal
seemagrwl7@gmail.com

Citation
Singh, J., Kumari, S., & Agrawal, S. (2026). A Novel System for Detecting Model Poisoning Attacks in Federated Learning. *Journal of Reliable and Secure Computing*, 2(1), 27–38.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

In this paper, we present the Dynamic Malicious User Selection Attack (DyMUSA), an adaptive model poisoning attack specifically designed for FRS. DyMUSA employs an enhanced version of the Isolation Forest algorithm, combining Context-Aware and Weighted Isolation Forests to dynamically identify and exploit users exhibiting anomalous behaviors. By integrating contextual information and assigning different weights to features based on their importance, DyMUSA improves detection accuracy and effectiveness. Unlike static attacks, DyMUSA continuously adapts to changes in the training environment, increasing its impact while maintaining a low profile to avoid detection.

Our research makes several key contributions. First, we introduce DyMUSA, a dynamic attack framework that adapts to the evolving federated learning environment, thereby enhancing its stealth and effectiveness. Second, DyMUSA incorporates an improved anomaly detection mechanism, which leverages contextual data and weighted features to more accurately target users who exhibit abnormal behaviors. Third, the attack employs an adaptive poisoning strategy, where fake users and interactions are iteratively injected into the system to manipulate the recommendation model, ensuring the attack remains effective as the federated learning process evolves.

To validate the efficacy of DyMUSA, we conduct extensive experiments on real-world datasets. The results demonstrate that DyMUSA outperforms traditional poisoning attacks in both effectiveness and stealth, significantly impacting metrics such as exposure ratio and recommendation accuracy. These findings underscore the need for enhanced threat awareness, the development of dynamic detection methods, and the implementation of adaptive defense strategies in federated recommender systems. The research also highlights the importance of comprehensive security evaluations and paves the way for future studies focused on fortifying the security and stability of federated learning frameworks against evolving adversarial tactics.

1.1 Main Contributions

The main contributions are as follows:

1. We propose a new dynamic attack framework named DyMUSA that adapts to changes in the federated learning environment, increasing its stealthiness and effectiveness.
2. DyMUSA implicitly embeds contextual information and utilizes a novel detection paradigm to enhance its ability to target abnormal behavior users by mimicking them over typical users.
3. DyMUSA achieves higher attack success rates while remaining harder to detect than existing traditional methods - FedRecAttack and PoisonRec.

The rest of the paper is organized as follows: we review related work in Section 2, introduce our system model in Section 3, present experimental results and analysis in Section 4, and conclude the proposed work in Section 5.

2 Related Work

Modern recommender systems (RS) have significantly enhanced user experiences across digital platforms by aligning user preferences with available content [1, 3, 4]. Despite these advancements, RS are susceptible to various security threats, particularly poisoning attacks, where adversaries inject deceptive or harmful data into the training datasets or intervene during the model training process. Such attacks can promote low-quality products or spread misinformation, undermining the system's integrity [5–7].

Early research on poisoning attacks in RS introduced heuristic-based methods, including random, average, and love/hate attacks, which required minimal knowledge of the underlying data [8, 9]. Over time, more sophisticated strategies, such as bandwagon and relation attacks, were developed, leveraging principles like Zipf's law to strengthen connections between target and popular items [5, 10]. Although effective, these static methods became predictable and easier to detect [11].

In recent years, the focus has shifted towards attacks tailored to both centralized and decentralized RS architectures, particularly in federated learning frameworks. Federated recommender systems (FRS) are valued for their privacy-preserving characteristics, as user data remains local, with only model updates being shared [28]. However, this decentralized approach introduces new vulnerabilities that adversaries can exploit through model poisoning attacks. For example, FedRecAttack exemplifies how attackers can degrade a global model's performance by strategically manipulating local updates in a federated setting [33].

Attackers often operate with limited resources,

necessitating cost-effective and stealthy strategies to avoid detection [15, 16]. This requires blending malicious data with legitimate inputs to maintain a low profile [17, 18]. In response, two primary defensive strategies have emerged: poisoning data filtering and robust training. Poisoning data filtering involves identifying and removing malicious elements before they enter the system, employing techniques like supervised anomaly filtering, unsupervised clustering, and confidence-based filtering [19–21]. Robust training, on the other hand, enhances model resilience by incorporating methods such as mutual enhancement among multiple models and adaptive learning of confidence levels for each sample [24–26].

Furthermore, research has been conducted to improve the privacy and security of federated learning systems. For instance, the FedFast framework accelerates the training of FRS while maintaining high levels of security and privacy [37]. Additionally, lightweight anomaly detection mechanisms have been explored to promptly identify and mitigate potential threats in federated environments [38].

Beyond poisoning attacks, federated learning systems face other security threats, such as attribute inference and model inversion attacks, which aim to extract sensitive information from shared models. To counter these threats, researchers have proposed privacy-preserving techniques, including differential privacy, secure multi-party computation, and homomorphic encryption, which aim to protect user data while ensuring the effectiveness of recommendation models [12, 28].

In [27], Du et al. suggested a distributed foundation models for multi-modal learning in 6G wireless networks. This work explores cutting-edge AI techniques - pipeline parallelism, data parallelism, and multi-modal learning, to promote the sustainable growth of distributed multi-modal FMs in the 6G era. Pipeline parallelism can help alleviate communication bottlenecks by compressing activations and gradients while strategically allocating communication resources. For data parallelism, federated learning integrated with over-the-air computation (AirComp) speeds up gradient aggregation by merging communication and computation.

In [29], Hasan provides a comprehensive systematic review of federated learning (FL) as a privacy-preserving paradigm for enterprise decision systems, synthesizing findings from a large set of peer-reviewed studies. The paper highlights foundational FL algorithms (e.g., FedAvg, FedProx, SCAFFOLD) and the layered use of privacy mechanisms such as secure aggregation, differential privacy, homomorphic encryption, and multiparty computation to balance privacy guarantees with empirical robustness. It also identifies major vulnerabilities including model poisoning, backdoor attacks, and gradient leakage, and outlines practical defensive strategies like robust aggregation and anomaly detection. Sector-specific implementations in healthcare, finance, and other domains demonstrate how FL enables collaborative modeling while respecting data residency and governance requirements.

In [30], Feng et al. survey the security threats facing federated learning systems, categorizing key attack types—backdoor attacks, Byzantine attacks, and adversarial attacks—and reviewing their associated defense mechanisms. The survey emphasizes that the distributed nature and data inaccessibility in FL both protect privacy and introduce new vulnerabilities, making FL systems susceptible to attacks during training and inference phases. It also discusses the strengths and limitations of existing defenses, provides a threat taxonomy, and outlines future research directions aimed at strengthening FL's security posture.

In [31], Alansary et al. review emerging artificial intelligence (AI)-related threats in cybercrime, with a specific focus on zero-day attacks leveraging machine, deep, and federated learning techniques. The paper explores how advanced AI methods are exploited by cybercriminals to discover and exploit previously unknown system vulnerabilities that evade traditional signature-based defenses. Key challenges such as handling imbalanced data, generalization across diverse attack types, and computational trade-offs are discussed, alongside future research directions to enhance proactive detection and mitigation strategies.

In [32], Guo et al. investigate Gradient Inversion Attacks (GIA) in federated learning, analyzing how sensitive, private data can be reconstructed from shared gradient information despite FL's privacy goals. They categorize existing GIA methods into optimization-based, generation-based, and analytics-based approaches and evaluate their effectiveness and limitations in practical FL settings. Their analysis shows that while some attack variants are more practical than others, all pose serious privacy risks, and they propose a multi-stage defense pipeline

to guide the design of more robust FL frameworks.

This research contributes to the growing body of knowledge on RS security by introducing the Dynamic Malicious User Selection Attack (DyMUSA). DyMUSA employs advanced anomaly detection through an enhanced Isolation Forest algorithm to dynamically identify and exploit vulnerabilities in federated recommender systems. By introducing adaptive poisoning strategies that evolve with the federated learning process, this study emphasizes the need for robust defense mechanisms and aims to inspire future research to strengthen the security and resilience of federated learning frameworks.

In light of the inadequacies of traditional defenses, namely, limited adaptability, lack of dynamic response, high detection risk, vulnerability to evolving attacks, and inefficiency in sparse data, we present the Dynamic Malicious User Selection Attack (DyMUSA), an adaptive model poisoning attack specifically designed for FRS. Unlike existing attack strategies that do not evolve stealthily over time namely FedRecAttack [33] and PoisonRec [34], DyMUSA leverages an Enhanced Isolation Forest (EIF) algorithm to dynamically detect and exploit anomalous user behaviors in real-time. By integrating context-aware and weighted features into the isolation forest, DyMUSA adapts to changes in the federated learning process, generating poisoned gradients that are significantly harder to detect than those created by previous methods. The novelty of DyMUSA lies in its ability to continuously evolve its attack strategy by leveraging EIF as the global model adapts over time. This dynamic approach ensures that the poisoning attack remains effective while minimizing the risk of detection. DyMUSA achieves this by analyzing the current state of the system, and using contextual signals to identify and target the most vulnerable users. This contrasts with prior approaches that relied on fixed, predefined attack strategies, which often became less effective as the system evolved [35, 36].

Our extensive experimental evaluation shows that DyMUSA consistently outperforms existing attack frameworks, FedRecAttack and PoisonRec, achieving higher attack success rates while maintaining the robustness of the global model against detection.

3 System Model

Federated learning (FL) enables collaborative model training without sharing raw data, preserving user privacy by ensuring data remains on local devices while exchanging only model updates [6]. Recommender systems (RS) within federated learning environments use decentralized data to personalize content across various platforms, enhancing user satisfaction [1, 3, 11].

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ denote the set of users and $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$ denote the set of items. The interaction matrix $\mathcal{R} \in \mathbb{R}^{N \times M}$ represents user-item interactions, where $\mathcal{R}_{ij}$ is the interaction value between user u_i and item v_j.

We utilize LightGCN as the base model for our federated recommender system [2], though DyMUSA is applicable to various recommender models. LightGCN simplifies graph convolution networks by focusing on essential graph convolutional operations. The propagation rule in matrix form is given by:

$$\mathcal{E}^{(k+1)} = \tilde{\mathcal{A}}\mathcal{E}^{(k)} \tag{1}$$

where $\mathcal{E}^{(k)} \in \mathbb{R}^{(N+M)\times d}$ are the user and item embeddings at layer k, and $\tilde{\mathcal{A}}$ is the symmetrically normalized adjacency matrix of the user-item graph [2].

The final user and item embeddings are obtained by combining the embeddings from all layers:

$$\mathcal{E}_u^* = \sum_{k=0}^{K} \alpha_k \mathcal{E}_u^{(k)}, \quad \mathcal{E}_i^* = \sum_{k=0}^{K} \alpha_k \mathcal{E}_i^{(k)} \tag{2}$$

where α_k are hyperparameters determining the importance of each layer's embedding.

The predicted interaction score $\hat{\mathcal{R}}_{ij}$ is computed as the inner product of the final user and item embeddings:

$$\hat{\mathcal{R}}_{ij} = (\mathcal{E}_u^*)^T \mathcal{E}_i^* \tag{3}$$

In the federated learning framework, the central server maintains the global model parameters Θ, while each user u_i computes local updates. The server aggregates these updates to refine the global model [28].

3.1 Proposed Technique

The proposed Dynamic Malicious User Selection Attack (DyMUSA) is illustrated in Figure 1. DyMUSA leverages an enhanced Isolation Forest (EIF) algorithm that combines Context-Aware Isolation Forest (CAIF) and Weighted Isolation Forest (WIF) to dynamically identify users exhibiting anomalous behaviors [13, 14]. By incorporating contextual information and assigning

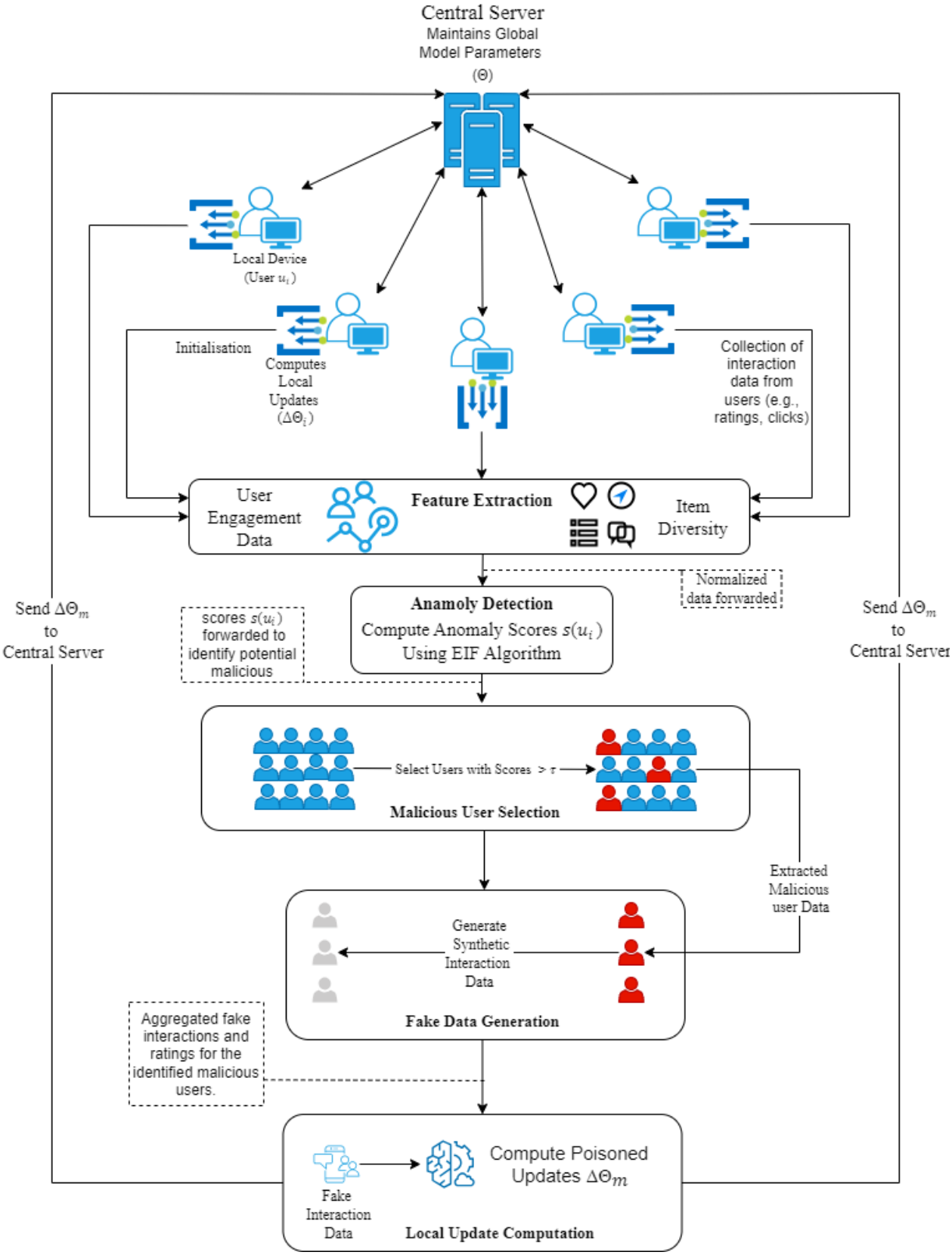

Figure 1. Architecture of proposed model DyMUSA.

weights to features based on their significance, this method enhances detection accuracy for abnormal user behavior [7].

- **Initialization**: Each local device (representing a user u_i) initializes and computes local updates ($\Delta\Theta$) based on user engagement data.

- **Feature Extraction**: Interaction data from users (such as ratings and clicks) is used to extract relevant features, including user engagement and item diversity metrics.
- **Anomaly Detection**: The extracted features are processed by the EIF algorithm, which computes anomaly scores $s(u_i)$ for each user to identify those with anomalous behavior.
- **Malicious User Selection**: Users with anomaly scores above a threshold τ are selected as potential malicious users, who significantly deviate from normal behavior.
- **Fake Data Generation**: For the selected malicious users, synthetic interaction data is generated to simulate interactions they did not perform, effectively poisoning the dataset.
- **Local Update Computation**: Using the fake interaction data, poisoned updates ($\Delta\Theta_m$) are computed locally to manipulate the global model.
- **Aggregation and Model Update**: The poisoned updates are aggregated and sent to the central server, updating the global model parameters (Θ).

DyMUSA's adaptive approach ensures it remains effective throughout the federated learning process by continuously identifying and exploiting malicious users.

To evaluate the performance of DyMUSA, we use metrics such as the exposure ratio at K ($ER@K$), normalized discounted cumulative gain (NDCG), hit rate at K ($HR@K$), and mean reciprocal rank (MRR).

4 Evaluation

The Dynamic Malicious User Selection Attack (DyMUSA) is meticulously designed to execute an advanced model poisoning attack on federated recommender systems. This section elaborates on the critical components and methodologies employed in the DyMUSA implementation.

The DyMUSA attack was simulated in a federated learning environment using PySyft, a library that enables privacy-preserving machine learning.

The DyMUSA attack operates by initializing specific parameters and data to effectively execute the attack. The key components include:

- *Interaction Matrix*: The interaction matrix $\mathcal{R} \in \mathbb{R}^{N \times M}$ contains user-item interaction data, serving as the basis for identifying user behavior and preferences.
- *User and Item Numbers*: Determines the number of users N and items M from the interaction matrix dimensions, providing an overview of the dataset's scale.
- *Attack Parameters*: Configures parameters such as the number of epochs, inner and outer epochs, and sizes for malicious users and feedback to ensure the attack's effectiveness.

The target items for the attack are selected using one of the following strategies:

- *Popular Items*: Items with the highest interaction counts.
- *Random Items*: Items selected randomly.
- *Category-Specific Items*: Items chosen from a particular category.
- *Low Interaction Items*: Items with minimal interactions.

New fake users are introduced into the dataset by:

- *User Embeddings Initialization*: Initializing new embeddings $\mathcal{E}_u^*$ for the fake users.
- *Updating Interaction Matrix*: Incorporating interactions for the fake users into the interaction matrix $\mathcal{R}$.
- *Reinitializing Recommender System*: Reinitializing the recommender system with the updated dataset.

DyMUSA employs an adaptive poisoning strategy to iteratively compromise the federated learning model. The attack evolves with the federated learning process by dynamically selecting malicious users and generating poisoned model updates based on their anomalous behavior.

- **Anomaly Score Computation:** For each user u_i in the interaction matrix $\mathcal{R}$, compute the anomaly score $s(u_i)$ using the EIF.
- **Dynamic Threshold Selection:** Update the threshold τ dynamically based on the percentile p of the anomaly scores:

$$\tau = \text{Percentile}(\text{AnomalyScores}, p) \quad (4)$$

Further adjust the threshold based on the detection rate DetectionRate and a target

detection rate β:

$$\tau = \tau + \alpha \times (\text{DetectionRate} - \beta) \tag{5}$$

- **Dynamic User Selection:** Select users with anomaly scores exceeding the updated threshold.

1. **Fake Data Generation:** DyMUSA generates fake interaction data for the identified malicious users. This fake data is designed to bias the model toward the attacker's target items.

   (a) **Fake User Initialization:**

   $$\text{FakeUser}_i = \{u_i \mid u_i \notin \mathcal{U}\} \tag{6}$$

   Here, $\mathcal{U}$ represents the set of real users, and FakeUser_i denotes a newly initialized fake user.

   (b) **Interaction Probability Computation:** For each fake user u_i and item v_j, compute the probability of interaction $P(u_i, v_j)$:

   $$P(u_i, v_j) = \sigma\left((\mathcal{E}_u^*)^T \mathcal{E}_j^*\right) \tag{7}$$

   where $\mathcal{E}_u^*$ and $\mathcal{E}_j^*$ are the embeddings for user u_i and item v_j, and σ denotes the sigmoid function.

   (c) **Bias Calculation:** Compute the bias for each item v_j to target specific items:

   $$\text{Bias}(u_i, v_j) = \begin{cases} \lambda, & \text{if } v_j \in \mathcal{T} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

   where $\mathcal{T}$ is the set of target items, and λ is a positive bias factor.

   (d) **Fake Rating Generation:** Generate fake ratings for each item v_j interacted with by fake user u_i:

   $$\text{FakeRating}(u_i, v_j) = \text{TargetValue} + \text{Bias}(u_i, v_j) \tag{9}$$

   Here, TargetValue is a predefined rating value intended to favor the target items.

   (e) **Fake Interaction Matrix Update:** Update the interaction matrix $\mathcal{R}'$ with the fake interactions:

   $$\mathcal{R}'_{ij} = \begin{cases} \text{FakeRating}(u_i, v_j), & \text{if } (u_i, v_j) \\ \in \text{FakeUserInteractions} & \\ \mathcal{R}_{ij}, & \text{otherwise} \end{cases} \tag{10}$$

2. **Model Update Computation:** Compute the model update $\Delta\mathcal{W}_u$ for each malicious user using the fake data and the global model from the previous round $\mathcal{W}_{t-1}$:

   $$\Delta\mathcal{W}_u = F(\mathcal{W}_{t-1}, \text{FakeData}_{u_i}) \tag{11}$$

3. **Accumulation of Poisoned Updates:** Accumulate the model updates from all malicious users:

   $$\Delta\mathcal{W}_t = \sum_{u_i \in \mathcal{U}_m} \Delta\mathcal{W}_u \tag{12}$$

4. **Global Model Update:** Update the global model $\mathcal{W}_t$ using the accumulated poisoned updates with a learning rate η:

   $$\mathcal{W}_t = \mathcal{W}_{t-1} + \eta\Delta\mathcal{W}_t \tag{13}$$

The adaptive poisoning strategy in DyMUSA leverages a dynamic approach to model poisoning within federated learning environments.

4.1 Parameters Initialization

- Learning Rate η: This parameter controls how much the model weights are adjusted during training in response to the estimated error. A higher learning rate can lead to faster convergence but may also risk overshooting the optimal solution, while a lower learning rate provides more stable convergence but can be slower.
- Anomaly Threshold τ: This is a threshold value used to determine what constitutes an anomalous behavior in the model updates during the training process. If an update's magnitude exceeds this threshold, it may indicate a potential attack

The strategy begins with parameter initialization, setting up the initial global model $\mathcal{W}_0$, learning rate η, and anomaly score threshold τ. The system then computes anomaly scores for each user by analyzing the interaction matrix $\mathcal{R}$, contextual information c_i, and feature weights w_i. This involves calculating the average path length $\mathbb{E}(h(u_i, w_i, c_i))$ in the isolation trees. DyMUSA's adaptive poisoning strategy iteratively updates user and item embeddings to maximize the exposure of target items. The anomaly score $s(u_i)$ for each user u_i is calculated based on the EIF model. The set of malicious users $\mathcal{U}_m$ is defined. The objective function $\mathcal{L}$ to be minimized during the adaptive poisoning strategy is defined and the update rules for the embeddings are defined.

Users with scores $s(u_i)$ exceeding the threshold τ are identified as malicious. For these users, synthetic data is generated to simulate interactions, which are then used to compute poisoned model updates $\Delta\mathcal{W}_u$. These updates are calculated by applying a gradient function $F(\mathcal{W}_{t-1}, \text{FakeData}_u)$, ensuring that the malicious influence on the model is strategically incorporated. The global model $\mathcal{W}_t$ is iteratively updated with the aggregated poisoned updates $\Delta\mathcal{W}_t = \sum_{u_i \in \mathcal{U}_m} \Delta\mathcal{W}_u$, allowing the attack to adapt over multiple iterations T. This adaptability enables DyMUSA to continuously refine its strategy, ensuring the persistent influence of malicious data on the global model while evading detection and maintaining efficacy throughout the federated learning process.

The evaluation results indicate that DyMUSA consistently demonstrates superior performance across multiple datasets.

- **MovieLens-1M Dataset** [39]: DyMUSA not only achieves the highest HR@10 of 0.8058, MRR of 0.7560, and NDCG@10 of 0.7102, but also excels in maintaining a high exposure ratio of 0.6753. This underscores its capability to enhance the visibility of target items efficiently. The close margin between DyMUSA and other methods such as PoisonRec highlights its competitive edge, driven primarily by its ability to remain stealthy while effectively manipulating recommendation outcomes.

- **Amazon-Electronics Dataset** [40]: DyMUSA's showcases dominance, reflected in an HR@10 of 0.7735 and MRR of 0.7220. Its robust NDCG@10 score of 0.6768 and ER of 0.6355 signify its proficiency in influencing recommendation outcomes. This superiority not only demonstrates its effectiveness but also its adaptability in circumventing detection mechanisms.

- **Netflix-Prize Dataset** [41]: DyMUSA's performance is markedly superior with an HR@10 of 0.8215 and NDCG@10 of 0.7245. These high scores illustrate its exceptional ability to skew recommendation lists in favor of target items while minimizing detection risks, thus demonstrating a nuanced balance between efficacy and subtlety.

- **Yelp Dataset** [42]: DyMUSA illustrates effectiveness with an HR@10 of 0.7835 and MRR of 0.7330, outperforming other attacks. Its sustained high exposure ratio of 0.6455 suggests an adept manipulation of the recommendation system, enhancing its practical applicability in real-world scenarios that demand stealthy operations.

- **Steam-200K Dataset** [43]: DyMUSA's HR@10 of 0.8265 and NDCG@10 of 0.7300 emphasize its consistent efficacy across datasets. This adaptability across diverse environments underscores its strategic design, making it a formidable choice for executing potent attacks that remain under the radar.

The analytical assessment of DyMUSA reveals its strategic advantage in balancing effectiveness and stealth. By systematically analyzing metrics across various datasets, it is evident that DyMUSA's adaptive mechanisms allow it to exploit weaknesses in recommender systems, achieving substantial influence while minimizing the risk of detection. Its consistent outperformance of other methods highlights its sophisticated design, which is engineered to sustain long-term influence over recommendation models. This makes DyMUSA a leading model poisoning attack method, capable of exerting a strong, stealthy impact. The ability to adapt to evolving defenses further cements its role as a critical tool for adversaries seeking to manipulate federated recommender systems effectively.

This section evaluates the Dynamic Malicious User Selection Attack (DyMUSA) against existing defensive strategies, specifically focusing on poisoning data filtering and robust training methods. These evaluations use metrics such as Attack Success Rate (ASR), Exposure Ratio (ER), Detection Rate (DR), and Gradient Similarity Index (GSI) to measure DyMUSA's ability to bypass defenses. Additionally, defense metrics such as False Positive Rate (FPR), True Positive Rate (TPR), and Defense Success Rate (DSR) assess the effectiveness of these defensive strategies.

Poisoning data filtering focuses on identifying and removing malicious elements, such as counterfeit users and misleading data, before they influence the system [19–21]. DyMUSA effectively counters this defense by generating fake users whose behavior closely mirrors that of legitimate users, thereby evading detection mechanisms like supervised anomaly filters and unsupervised clustering methods [22, 23]. The adaptive nature of DyMUSA ensures that the generated interactions maintain high confidence scores, minimizing the likelihood of being flagged as anomalous.

Table 1. Comparison of attack performance against defense strategies.

Dataset	Attack	ASR	ER	DR	GSI	FPR	TPR	DSR
MovieLens-1M	AUSH	65%	0.68	45%	0.65	20%	55%	35%
	FedRecAttack	70%	0.72	40%	0.68	18%	60%	42%
	PoisonRec	73%	0.75	38%	0.71	16%	62%	45%
	DyMUSA	**85%**	**0.88**	**25%**	**0.90**	**10%**	**75%**	**65%**
Amazon-Electronics	AUSH	62%	0.64	50%	0.63	21%	50%	30%
	FedRecAttack	68%	0.70	42%	0.66	19%	57%	39%
	PoisonRec	71%	0.73	40%	0.70	17%	60%	42%
	DyMUSA	**83%**	**0.86**	**28%**	**0.88**	**12%**	**73%**	**63%**
Netflix-Prize	AUSH	60%	0.62	48%	0.61	22%	52%	33%
	FedRecAttack	67%	0.69	43%	0.65	20%	55%	38%
	PoisonRec	70%	0.72	39%	0.68	18%	58%	40%
	DyMUSA	**82%**	**0.85**	**27%**	**0.87**	**13%**	**70%**	**60%**
Yelp	AUSH	63%	0.66	49%	0.64	21%	53%	34%
	FedRecAttack	69%	0.71	44%	0.67	19%	56%	41%
	PoisonRec	72%	0.74	41%	0.69	17%	59%	43%
	DyMUSA	**84%**	**0.87**	**26%**	**0.89**	**11%**	**72%**	**62%**
Steam-200K	AUSH	61%	0.63	46%	0.62	20%	54%	32%
	FedRecAttack	66%	0.68	41%	0.64	18%	57%	37%
	PoisonRec	69%	0.71	37%	0.67	16%	60%	39%
	DyMUSA	**81%**	**0.84**	**29%**	**0.86**	**14%**	**69%**	**59%**

The effectiveness of DyMUSA is evident in its superior ASR and ER values across all datasets, as shown in Table 1. Robust training aims to enhance the resilience of recommendation models through strategies such as mutual enhancement among multiple models and adaptive learning of confidence levels for each data sample [24–26]. DyMUSA strategically exploits these methods by targeting the inconsistencies between models during mutual enhancement, gradually introducing poisoned data to bypass adaptive learning mechanisms designed to detect anomalies. This approach ensures that DyMUSA retains its effectiveness and stealth over extended periods.

Table 1 demonstrates DyMUSA's superior performance in reducing detection rates (DR) while maintaining high GSI scores across multiple datasets, confirming its ability to blend seamlessly into legitimate model updates.

Figure 2 represents the number of real attacks detected correctly divided by the total number of real attacks by AUSH, FedRecAttack, PoisonRec, and, proposed DyMUSA, using Movielens dataset.

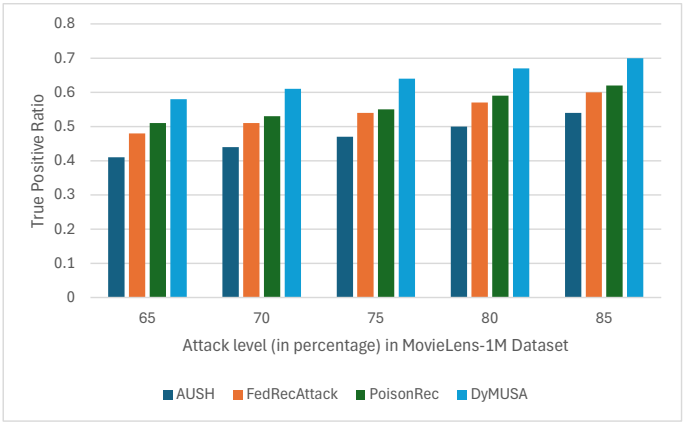

Figure 2. True Positive Ratio vs Attack level (in percentage) in MovieLens-1M Dataset.

Figures 3 and 4 represent the performance of the proposed DyMUSA Scheme. Using different sets of datasets, we evaluated the values of the performance metrics - Mean Anomaly Score, Temporal Change in Anomaly Scores, Model Updates, Attack Success Rate, Defense Success Rate, and, True Positive Rate.

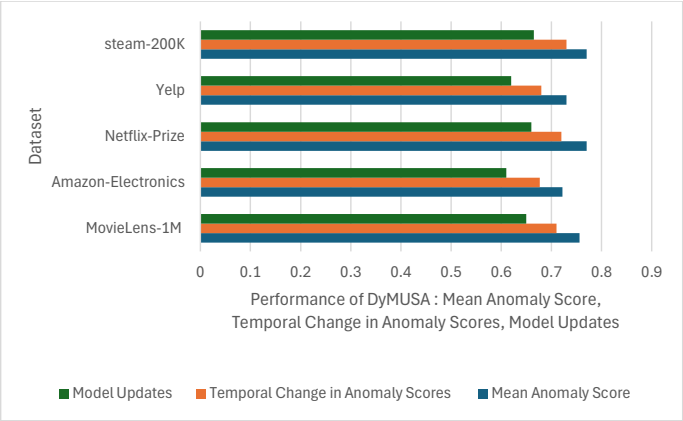

Figure 3. Performance of DyMUSA with different benchmark datasets.

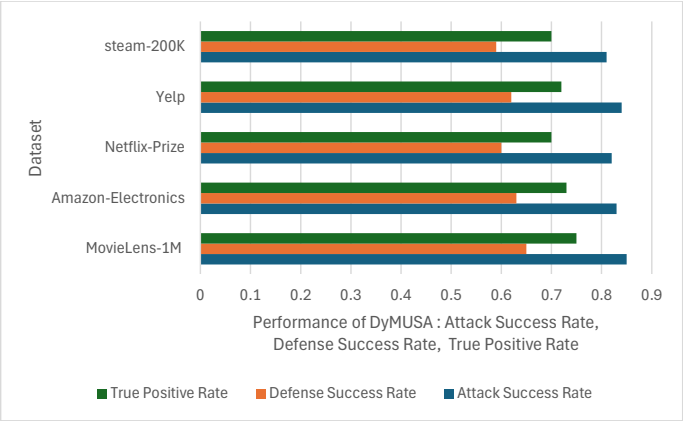

Figure 4. Performance of DyMUSA with different benchmark datasets.

5 Conclusion

In this paper, we presented DyMUSA, a sophisticated and adaptive model poisoning attack targeting federated recommender systems. Extensive evaluations on multiple real-world datasets demonstrate that DyMUSA achieves strong performance across key effectiveness and stealth metrics, including Hit Rate, Mean Reciprocal Rank, Normalized Discounted Cumulative Gain, Exposure Ratio, and stealthiness. The results show that DyMUSA can significantly manipulate recommendation outcomes while maintaining minimal impact on overall system performance, making the attack difficult to detect. The adaptability and stealth of DyMUSA expose critical vulnerabilities in federated recommender systems and underscore the urgent need for robust and adaptive defense mechanisms. These findings highlight the importance of continuous monitoring and the evolution of security protocols to safeguard federated systems against increasingly sophisticated poisoning attacks.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Fan, W., Zhao, X., Chen, X., Su, J., Gao, J., Wang, L., ... & Li, Q. A comprehensive survey on trustworthy recommender systems (2022). *CoRR, abs/2209.10117*.

[2] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020, July). Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (pp. 639-648). [CrossRef]

[3] Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016, September). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems* (pp. 7-10). 7-10.[CrossRef]

[4] Wang, X., He, X., Wang, M., Feng, F., & Chua, T. S. (2019, July). Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 165-174). [CrossRef]

[5] Gunes, I., Kaleli, C., Bilge, A., & Polat, H. (2014). Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review, 42*(4), 767-799. [CrossRef]

[6] Xia, G., Chen, J., Yu, C., & Ma, J. (2023). Poisoning attacks in federated learning: A survey. *IEEE Access, 11*, 10708-10722. [CrossRef]

[7] Zhang, H., Li, Y., Ding, B., & Gao, J. (2020, April). Practical data poisoning attack against next-item recommendation. In *Proceedings of the web conference 2020* (pp. 2458-2464). [CrossRef]

[8] Lam, S. K., & Riedl, J. (2004, May). Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web* (pp. 393-402). [CrossRef]

[9] Alonso, S., Bobadilla, J., Ortega, F., & Moya, R. (2019). Robust model-based reliability approach to tackle shilling attacks in collaborative filtering recommender systems. *IEEE access, 7*, 41782-41798. [CrossRef]

[10] Yu, J., Gao, M., Rong, W., Li, W., Xiong, Q., & Wen, J. (2017). Hybrid attacks on model-based social recommender systems. *Physica A: Statistical Mechanics and its Applications, 483*, 171-181. [CrossRef]

[11] Deldjoo, Y., Noia, T. D., & Merra, F. A. (2021). A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *Acm Computing Surveys (Csur), 54*(2), 1-38. [CrossRef]

[12] Rong, D., Ye, S., Zhao, R., Yuen, H. N., Chen, J., & He, Q. (2022, May). Fedrecattack: Model poisoning attack to federated recommendation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)* (pp. 2643-2655). IEEE. [CrossRef]

[13] Zhou, W., Wen, J., Qu, Q., Zeng, J., & Cheng, T. (2018). Shilling attack detection for recommender systems based on credibility of group users and rating time series. *PloS one, 13*(5), e0196533. [CrossRef]

[14] Gao, C., Wang, X., He, X., & Li, Y. (2022, February). Graph neural networks for recommender system. In *Proceedings of the fifteenth ACM international conference on web search and data mining* (pp. 1623-1625). [CrossRef]

[15] Cho, J. H., Sharma, D. P., Alavizadeh, H., Yoon, S., Ben-Asher, N., Moore, T. J., ... & Nelson, F. F. (2020). Toward proactive, adaptive defense: A survey on moving target defense. *IEEE Communications Surveys & Tutorials, 22*(1), 709-745. [CrossRef]

[16] Anelli, V. W., Deldjoo, Y., DiNoia, T., & Merra, F. A. (2021). Adversarial recommender systems: Attack, defense, and advances. In *Recommender systems handbook* (pp. 335-379). New York, NY: Springer US. [CrossRef]

[17] Fang, M., Yang, G., Gong, N. Z., & Liu, J. (2018, December). Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th annual computer security applications conference* (pp. 381-392). [CrossRef]

[18] Di Noia, T., Malitesta, D., & Merra, F. A. (2020, June). Taamr: Targeted adversarial attack against multimedia recommender systems. In *2020 50th Annual IEEE/IFIP international conference on dependable systems and networks workshops (DSN-W)* (pp. 1-8). IEEE. [CrossRef]

[19] Si, M., & Li, Q. (2020). Shilling attacks against collaborative recommender systems: a review. *Artificial Intelligence Review, 53*(1), 291-319. [CrossRef]

[20] Nawara, D., Aly, A., & Kashef, R. (2024). Shilling attacks and fake reviews injection: Principles, models, and datasets. *IEEE Transactions on Computational Social Systems*. [CrossRef]

[21] Nguyen, T. T., Quoc Viet Hung, N., Nguyen, T. T., Huynh, T. T., Nguyen, T. T., Weidlich, M., & Yin, H. (2024). Manipulating recommender systems: A survey of poisoning attacks and countermeasures. *ACM Computing Surveys, 57*(1), 1-39. [CrossRef]

[22] Lee, J. S., & Zhu, D. (2012). Shilling attack detection—a new approach for a trustworthy recommender system. *INFORMS Journal on Computing, 24*(1), 117-131. [CrossRef]

[23] Mazurowski, M. A. (2013). Estimating confidence of individual rating predictions in collaborative filtering recommender systems. *Expert Systems with Applications, 40*(10), 3847-3857. [CrossRef]

[24] Moskalenko, V., Kharchenko, V., Moskalenko, A., & Kuzikov, B. (2023). Resilience and resilient systems of artificial intelligence: taxonomy, models and methods. *Algorithms, 16*(3), 165. [CrossRef]

[25] O'Mahony, M., Hurley, N., Kushmerick, N., & Silvestre, G. (2004). Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology (TOIT), 4*(4), 344-377. [CrossRef]

[26] Joshi, P., Shaikh, M. Z., Varshney, N., & Dwivedy, B. (2024, March). Robustness Challenges in Deep Learning: Strategies for Enhancing Model Resilience. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 757-762). IEEE. [CrossRef]

[27] Du, J., Lin, T., Jiang, C., Yang, Q., Bader, C. F., & Han, Z. (2024). Distributed foundation models for multi-modal learning in 6G wireless networks. *IEEE Wireless Communications, 31*(3), 20-30. [CrossRef]

[28] Zhou, X., Xu, M., Wu, Y., & Zheng, N. (2021). Deep model poisoning attack on federated learning. *Future Internet, 13*(3), 73. 1-10.[CrossRef]

[29] Hasan, M. M. (2025). Federated Learning Models for Privacy-Preserving AI In Enterprise Decision Systems. *International Journal of Business and Economics Insights, 5*(3), 238-269. [CrossRef]

[30] Feng, Y., Guo, Y., Hou, Y., Wu, Y., Lao, M., Yu, T., & Liu, G. (2025). A survey of security threats in federated learning. *Complex & Intelligent Systems, 11*(2), 165. [CrossRef]

[31] Alansary, S. A., Ayyad, S. M., Talaat, F. M., & Saafan, M. M. (2025). Emerging AI threats in cybercrime: a review of zero-day attacks via machine, deep, and federated learning. *Knowledge and Information Systems, 67*(11), 10951-10987. [CrossRef]

[32] Guo, P., Wang, R., Zeng, S., Zhu, J., Jiang, H., Wang, Y., ... & Qu, L. (2025). Exploring the vulnerabilities of federated learning: A deep dive into gradient inversion attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [CrossRef]

[33] Rong, D., Ye, S., Zhao, R., Yuen, H. N., Chen, J., & He, Q. (2022, May). Fedrecattack: Model poisoning attack to federated recommendation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)* (pp. 2643-2655). IEEE. [CrossRef]

[34] Song, J., Li, Z., Hu, Z., Wu, Y., Li, Z., Li, J., & Gao, J. (2020, April). Poisonrec: an adaptive data poisoning framework for attacking black-box recommender systems. In *2020 IEEE 36th international conference on data engineering (ICDE)* (pp. 157-168). IEEE. [CrossRef]

[35] Fang, M., Cao, X., Jia, J., & Gong, N. (2020). Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)* (pp. 1605-1622).

[36] Zhang, K., Cao, Q., Sun, F., Wu, Y., Tao, S., Shen, H., & Cheng, X. (2025). Robust recommender system: a survey and future directions. *ACM Computing Surveys, 58*(1), 1-38. [CrossRef]

[37] Muhammad, K., Wang, Q., O'Reilly-Morgan, D., Tragos, E., Smyth, B., Hurley, N., ... & Lawlor, A. (2020, August). Fedfast: Going beyond average for faster training of federated recommender systems. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1234-1242). [CrossRef]

[38] Mothukuri, V., Khare, P., Parizi, R. M., Pouriyeh, S., Dehghantanha, A., & Srivastava, G. (2021). Federated-learning-based anomaly detection for IoT security attacks. *IEEE Internet of Things Journal, 9*(4), 2545-2554. [CrossRef]

[39] Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis), 5*(4), 1-19. 2015. [CrossRef]

[40] Suryam, A. (n.d.). Ecommerce electronics dataset [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/ecommerce-data/amazon-electronics-dataset

[41] Netflix-Inc. (n.d.). Netflix Prize data [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

[42] Yelp, Inc. (n.d.). Yelp Open Dataset [Data set]. Yelp. Retrieved from https://www.yelp.com/dataset

[43] Kohli, K. (n.d.). Steam dataset [Data set]. Kaggle. Retrieved from https://www.kaggle.com/code/kkohli/steam-dataset

Dr. Jagdeep Singh is an Assistant Professor in the Department of Computer Science and Engineering at Sant Longowal Institute of Engineering and Technology (SLIET), Longowal. He earned his Ph.D. in Computer Engineering from the University of Delhi. His research interests span Artificial Intelligence, Machine Learning, Delay-Tolerant Networks, and Cybersecurity. He is an AICTE-certified Master Trainer in High-Performance Computing and has published extensively in reputed journals such as IEEE Internet of Things, Journal of Ambient Intelligence and Humanized Computing, IET, and the International Journal of Communication Systems. He has also presented his work at several flagship international conferences, including IEEE ICC, IEEE Globecom, AINA, and IEEE CITS. Dr. Singh is a Senior Member of IEEE and INAE, and a member of the International Association of Engineers. He actively contributes to the research community through technical program committees and reviews for top-tier journals like IEEE IoT, IEEE TIFS, IEEE Access, and Wiley journals. He has organized multiple academic programs and has also received funding from AICTE for conducting advanced-level FDPs and short-term training programs. He is a recipient of the Shastri Conference and Lecture Series Grant (2021–22) and led Team EMMET to victory in the Prototype and Ideation Stages of the Indian Web Browser Development Challenge by MeitY, securing prize money of approximately INR 12,00,000. He was selected as a young faculty presenter in the AI theme at ESTIC 2025, organized by DST, Government of India, at Bharat Mandapam. He has delivered numerous expert lectures on cybersecurity and AI-based technologies in various national programs. (Email: jagdeep@ieee.org)

Dr. Saru Kumari is an Associate Professor with the Department of Mathematics, Chaudhary Charan Singh University, Meerut, Uttar Pradesh, India. She received her PhD in Mathematics in 2012 from Chaudhary Charan Singh University, Meerut, UP, India. She received India Research Excellence - Citation Awards - Women in Research-2023 by Clarivate Analytics. She has published more than 360 research papers in reputed international journals and conferences, including more than 300 research papers in various SCIE Indexed Journals such as IEEE TDSC, IEEE TII, IEEE JBHI, IEEE T-ITS, IEEE TCE, IEEE TGCN, IEEE IoTJ, Information Fusion, ACM TOIT, ACM TOMM, etc. She received the Best Paper award from the Journal of Network and Computer Applications, Elsevier in 2020, IEEE Consumer Electronics Magazine in 2022, and Vehicular Communication in 2022. She is on the editorial board of more than a dozen International Journals of high repute, under IEEE, Elsevier, Springer, Wiley, and others including SCI and SCIE journals such as IEEE Transactions on Intelligent Transport Systems, (SCIE); IEEE Systems Journal, (SCIE); Computer Standards & Interfaces, Elsevier (SCIE); AEÜ - International Journal of Electronics and Communications, Elsevier (SCIE); International Journal of Communication Systems, Wiley (SCIE); Concurrency and Computation: Practice and Experience, Wiley (SCIE); Telecommunication Systems, Springer (SCIE); Human-centric Computing and Information sciences, Springer (SCIE); Transactions on Emerging Telecommunications Technologies; Wiley (SCIE), etc. She has served as the Guest Editor of many special issues in SCIE Journals under IEEE, Elsevier, Springer, and Wiley. She has been involved in the research community as a Technical Program Committee (TPC) member or PC chair for more than a dozen international conferences of high repute. She is also a reviewer of dozens of reputed Journals, including SCI-Indexed Journals, under IEEE, Elsevier, Springer, Wiley, Taylor & Francis, etc. Her research interests include Cryptology, Information Security, Blockchain Technology, and Artificial Intelligence. (Email: saryusiirohi@gmail.com)

Dr. Seema Agrawal is an Associate Professor in the Department of Mathematics at S.S.V. College in Hapur, Uttar Pradesh, India. She has earned Ph.D. in Mathematics, and her areas of research is applied mathematics, optimization, and soft computing techniques. (Email: seemagrwl7@gmail.com)