Adaptive Graph Convolutional Network for Shilling Attack Detection in Recommender Systems

Abhineswari M
SCOPE
Vellore Institute of Technology
Chennai
abhineswari.m2021@vitstudent.ac.in

Balabadruni Venkata Naga Koushik
SCOPE
Vellore Institute of Technology
Chennai
venkata.nagakoushik2021@vitstudent.ac.in

Sujithra @ Kanmani R
SCOPE
Vellore Institute of Technology
Chennai
sujithrakanmani.r@vit.ac.in

Recommender systems have become an essential component of online platforms, guiding users toward relevant products, movies, and services. However, they are increasingly vulnerable to shilling attacks, where malicious users inject biased ratings to manipulate recommendations. Existing detection methods struggle to effectively capture group-based collusive attacks while maintaining adaptability to dynamic user behaviours. To address this, we propose Adaptive Graph Convolutional Network, a hybrid model that leverages both graph-based user relationships and adaptive multi-view representation learning to detect individual and group-based shilling attackers. Our approach consists of three key phases: (1) Synthetic Attack Injection into real-world datasets (MovieLens, Amazon) to simulate various attack strategies (Sybil, collusion, social-based attacks). (2) Graph Representation Learning, where user interactions are transformed into multi-view graphs incorporating rating behavior, temporal activity, and user similarity metrics. (3) A-GCN-based Attack Detection, where our model applies adaptive self-propagation and multi-correlated node embeddings to classify normal users and attackers efficiently. Experimental results will further demonstrate that A-GCN outperforms traditional GCN and heuristic-based approaches in both individual and group shilling attack detection, making it a robust solution for securing recommendation systems.

Graph Convolutional Network, Social Network, Shilling Attack, Shilling Detection

I. Introduction

Recommender systems are integral to modern digital platforms, providing personalized content and product recommendations to users based on their preferences and interactions. These systems are widely utilized in domains such as e-commerce (Amazon, eBay), streaming services (Netflix, Spotify), and online marketplaces. By leveraging machine learning and data analytics, recommender systems analyze historical user interactions, such as product purchases, movie ratings, and click-through behaviors, to generate predictions about items a user may like. The most common techniques used in recommendation algorithms include collaborative filtering, which predicts user preferences based on similarities with other users, and content-based filtering, which recommends items similar to those a user has previously engaged with. Additionally, hybrid recommendation models combine multiple techniques to enhance accuracy and diversity in recommendations.

Despite their effectiveness, recommender systems are vulnerable to shilling attacks, a form of adversarial manipulation where malicious users—either individuals or coordinated groups—intentionally distort rating patterns to bias recommendations. These attacks pose a significant threat to system reliability, affecting both user experience and business revenue. Attackers may aim to promote specific items (push attack) or demote competing products (nuke attack) by injecting fake user profiles with strategically crafted ratings. Shilling attacks can be categorized into various types, including random attacks, where attackers assign arbitrary ratings, average attacks, which mimic general user behaviors, bandwagon attacks, where attackers manipulate trending items, and group collusive attacks, where multiple fake users act in coordination to amplify the impact.

The presence of such attacks degrades recommendation quality by introducing biased predictions, thereby misleading genuine users and diminishing trust in the system. Traditional attack detection techniques, such as anomaly detection and rule-based filtering, often fail to detect evolving and adaptive attacks, where attackers modify their strategies to evade detection. Moreover, many existing methods lack robustness against multi-view shilling attacks, where attackers manipulate different aspects of user-item interactions, such as trust relationships and co-purchase patterns. To address these challenges, we propose an Adaptive Graph Convolutional Network (A-GCN) that leverages multi-view graph representation and adversarial training to enhance attack detection and resilience. The end goals to be achieved being:

- Constructs user graphs integrating rating behavior, social interactions, and temporal patterns.
- Employs adaptive self-propagation, allowing dynamic feature aggregation while retaining user-specific characteristics.
- Detects individual and group-based shilling attacks through degree-aware, similarity-enhanced, and adversarially robust learning.
- By leveraging MovieLens and Amazon review datasets, we evaluate the effectiveness of A-GCN against state-of-the-art models, demonstrating its robustness in detecting sophisticated collusive shilling strategies.

II. Literature Survey

Traditional Machine Learning based detection models

Zayed et al. (2023) examine the popular methods for detecting shilling attacks, combining experimental results with theoretical analysis. They find that traditional detection

techniques are often inadequate, especially in dealing with complex attack strategies, and advocate for the development of more sophisticated detection methods. Grozdanić et al. (2023) propose an approach that combines multiple Random Forest models for detecting shilling attacks in collaborative filtering systems. They find that this ensemble method enhances detection accuracy and is more effective at identifying various attack patterns compared to single-model approaches. Su & Wang (2023) introduce a genetic co-forest approach to detect high-knowledge shilling attacks in recommender systems. They demonstrate that this method outperforms traditional techniques, offering better scalability and accuracy in detecting more sophisticated attack types. Zayed et al. (2023) present an ensemble-based method to detect attacks in recommender systems, aiming to improve detection accuracy. They find that ensemble techniques enhance the robustness of attack detection, leading to better identification of diverse attack strategies in collaborative filtering systems. Cai & Zhang (2021) introduce an unsupervised method called BS-SC for detecting shilling profiles in recommender systems. They show that their approach successfully identifies malicious profiles without requiring labeled data, making it a highly adaptable solution for detecting a wide range of shilling attacks.

Deep Learning and Graph Based Approaches

Praveena et al. (2023) propose a hybrid deep learning model combining Gated Recurrent Units (GRUs) and Convolutional Neural Networks (CNNs) for detecting shilling attacks in social networks. Their approach proves to be highly effective in detecting evolving and complex attack strategies by leveraging both sequential and spatial features. Zhang et al. (2022) develop a method combining reinforcement learning with adversarial autoencoders to detect collusive spammers on e-commerce platforms. Their approach proves highly effective in identifying coordinated fraudulent behaviors that are difficult to detect with traditional methods. Zhang et al. (2021) propose a user embedding-based method to detect group shilling attacks, leveraging deep learning techniques for more accurate identification. They find that this approach improves the detection of coordinated attacks by capturing user behaviors and interactions more effectively. Wang et al. (2021) introduce a hierarchical topic model for detecting shilling groups in recommender systems. They demonstrate that this model can identify attack groups more accurately by analyzing topic distributions, which helps in distinguishing legitimate users from malicious ones. Ebrahimian & Kashef (2021) propose a CNN-based hybrid model for detecting shilling attacks in collaborative filtering recommender systems. Their model outperforms traditional detection methods, achieving higher accuracy in detecting complex attacks by effectively utilizing convolutional layers for feature extraction.

Surveys and Mitigation Strategies

Nawara et al. (2024) explore the principles behind shilling attacks and fake review injections, focusing on various attack models and datasets. They provide a comprehensive overview of current detection challenges, emphasizing the need for more advanced techniques to tackle these malicious activities effectively. Hemmatpour et al. (2024) investigate the effects of bot traffic in e-commerce systems and propose detection and mitigation strategies. They find that in-network mitigation solutions can significantly reduce the negative impact of bot traffic, leading to improved system performance and security. Esposito et al. (2023) provide a survey on detecting malicious reviews and users in social review systems. They examine various detection techniques and highlight the complexity of malicious behaviors, suggesting the need for more robust methods that can adapt to emerging fraudulent tactics in social review environments. Jeny et al. (2022) develop a shilling attack detection system tailored for online recommender systems. Their system demonstrates a high detection rate for various attack types, offering a practical solution to maintaining the reliability of recommendation systems in real-world applications. Laskar et al. (2023) critically analyze the different types of shilling attacks in recommender systems, exploring various detection and mitigation strategies. They emphasize the need for more nuanced methods that can address the evolving nature of shilling attacks, focusing on techniques that adapt to new attack models.

III. Methodology

A Graph Neural Network (GNN) is a broad category of neural models designed to learn from graph-structured data, using general message-passing mechanisms for tasks like node classification and link prediction. It supports heterogeneous graphs and dynamic structures, making it adaptable for complex datasets like Amazon Reviews. In contrast, a Graph Convolutional Network (GCN) is a specific type of GNN that applies convolution-like operations to aggregate information from neighbouring nodes. In our A-GCN (Adaptive GCN) model, we use graph convolution layers to analyze user-item interactions, detect Sybil attackers, and identify group collusion by leveraging multi-view embeddings. While GNNs are more flexible, GCNs are computationally efficient and structured for recommendation-based attack detection.

A. Data preparation and shilling attack injection

The first step involves selecting and preprocessing the datasets, specifically the MovieLens and Amazon Review datasets. MovieLens provides explicit user-item ratings, making it suitable for studying manipulative behaviors in collaborative filtering systems. The Amazon Review dataset, with its additional metadata such as timestamps and item categories, enables a more comprehensive analysis of fraudulent activities beyond rating patterns. Preprocessing begins by filtering out users and items with minimal interactions to ensure data density. Ratings are then normalized to a fixed scale to standardize inputs across different datasets. Finally, the data is transformed into a user-item bipartite graph, where edges represent interactions, setting the foundation for graph-based attack detection. This preprocessing step ensures that only meaningful interactions are considered, improving the robustness of subsequent attack detection models. To evaluate the detection system under real-world attack scenarios, synthetic shilling attacks

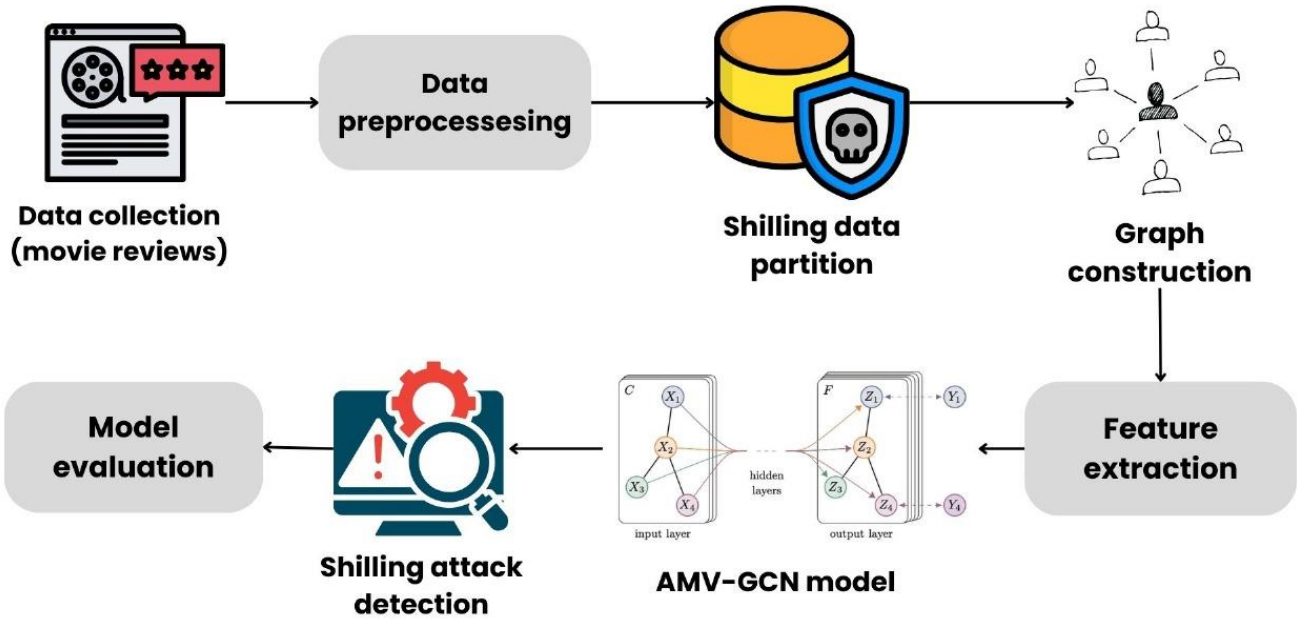

Fig 1 – Proposed Adaptive GCN Model Architecture

are injected into the datasets. These attacks include random attacks, where attackers assign arbitrary ratings; average attacks, where attackers mimic the rating distributions of genuine users with slight noise; bandwagon attacks, where attackers exploit trending items to gain influence before targeting other items; and group collusive attacks, where multiple fake users coordinate to manipulate item rankings. The attack injection process involves creating fake user profiles with unique user IDs, assigning ratings based on the chosen attack strategy, and modifying the user-item graph to incorporate these manipulated interactions. By introducing these artificial attacks, the dataset becomes more representative of real-world adversarial behaviors, allowing for effective model training and evaluation.

B. Graph Construction and Feature Engineering

To enhance detection accuracy, multi-view graph construction is employed, capturing different aspects of user behavior and item interactions. Three graph types are created: a user-item interaction graph that represents explicit user ratings, a user-user similarity graph that connects users based on rating correlations, and an item-item co-purchase graph that links frequently bought or rated items. Attackers attempt to manipulate these relationships, so incorporating multiple graph perspectives helps reveal hidden patterns of fraudulent behavior. These graph structures are essential in detecting inconsistencies, such as users who appear disconnected from the general network or items that receive sudden, coordinated rating changes. Feature engineering further strengthens attack detection by extracting meaningful attributes from the graph structures. Node embeddings are generated using techniques like Node2Vec or DeepWalk, capturing how users and items are positioned within the graph. Degree-aware features differentiate between normal users and attackers, as Sybil attackers typically have fewer interactions, while collusive groups exhibit abnormal connectivity. Additionally, time-based features such as rating bursts help identify sudden spikes in user activity, which are often indicative of manipulation attempts. These extracted features ensure that the graph-based detection model can learn both explicit and implicit attack patterns, making it more resilient against evolving fraudulent strategies.

C. Shilling attack detection using A-GCN

The detection model is implemented using an Adaptive Graph Convolutional Network (A-GCN), which leverages the multi-view graph structure to identify anomalies in user behaviors. A-GCN applies graph convolution layers to aggregate embeddings from the user-item, user-user, and item-item graphs, allowing it to learn attack-resistant representations. The model also incorporates an adaptive self-propagation mechanism, which optimally balances individual user features with neighborhood information. This prevents attackers from easily blending in with genuine users, as their abnormal connections and behaviors become more apparent when analyzed in the context of multiple graph views.

To further improve detection accuracy, a graph attention mechanism is introduced, assigning higher weights to suspicious rating behaviors. This ensures that interactions that deviate significantly from normal user patterns, such as excessive high or low ratings on specific items, are given greater importance in classification decisions. Additionally, adversarial training is employed to enhance the model's resilience against adaptive attackers. By injecting perturbations into user embeddings, A-GCN is forced to learn more robust features, making it difficult for attackers to evade detection by slightly altering their behaviors. Through these techniques, A-GCN achieves high accuracy in identifying both individual and group-based shilling attacks,

outperforming traditional machine learning and graph-based models.

D. Model Evaluation and Performance Analysis

Once the A-GCN model is trained, it is evaluated using the MovieLens and Amazon Review datasets to assess its effectiveness in detecting shilling attacks. Performance is measured using standard metrics, including precision, recall, F1-score, and AUC-ROC, which provide insights into the model's ability to distinguish between genuine users and attackers. False positive and false negative rates are also analyzed to ensure that the model does not mistakenly classify legitimate users as attackers or fail to detect actual fraudulent activities. Baseline comparisons are conducted against existing detection methods such as SVM, Random Forest, and other graph-based models like GCN, GAT, LightGCN, and ASP-GCN. The results highlight the advantages of A-GCN, particularly in identifying more complex and stealthy attack strategies. Additionally, robustness tests are performed by introducing new attack variations to evaluate how well the model adapts to evolving threats. By systematically assessing detection performance across multiple datasets and attack scenarios, the evaluation phase confirms the model's superiority in safeguarding recommendation systems against shilling attacks.

IV. RESULTS AND DISCUSSION

Our work targets shilling attack detection in recommender systems through an Adaptive Graph Convolutional Network (AMV-GCN) on the MovieLens dataset. The dataset is comprised of user-movie interactions, represented as a bipartite graph where users and movies are nodes and interactions (ratings) are represented as edges. For creating the user-item interaction graph, we first provide a unique index to each user and movie to facilitate contiguous mapping for easy processing. The boundary of the graph is then established by linking users with the movies that they have rated. Because numerical representations are necessary for deep learning models, we initialize random embeddings (64-dimensional feature vectors) for users and movies. They are used as node features and contain latent structural information that is to be further updated by graph convolutional operations. The data is then tagged to simulate shilling attacks. We tag 5% of the users as attackers, which try to mislead the recommendation system by offering fake ratings. All movies are tagged as non-attack objects (0) because they do not do anything in the recommendation process. The data is then divided into training (80%) and test (20%), so that only user nodes take part in the classification task, whereas movies serve solely as bridging nodes in the graph.

The model is built with two graph convolutional layers, which are executed with PyTorch Geometric's GCNConv module:

First GCN Layer:

-Accepts the initialized movie and user embeddings along with the edge index (interaction graph).
-Follows a graph convolution operation to transfer features between connected nodes.
-Incorporates a ReLU activation function to inject non--linearity and augment feature representation.

Second GCN Layer:

-Continuously refines node representations learned.
-Outputs a log-softmax class probability between two classes: normal user (0) or attacker (1).
This hierarchical feature extraction guarantees that shilling attackers with abnormal interactions are separable from real users. The model is trained using the Adam optimizer with a learning rate of 0.01 and weight decay of 5e-4. The cross-entropy loss function is utilized to evaluate classification performance, allowing the model to adjust its parameters accordingly.

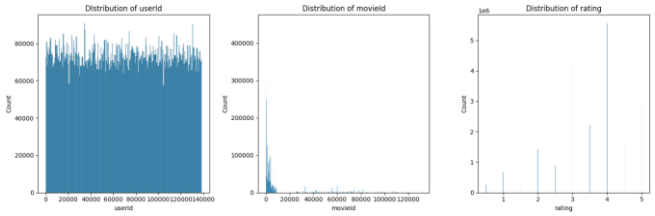

Fig 2 – Histogram distribution of userID, movieID and rating

The Distribution of userId (Left Plot)
The distribution looks quite even, indicating that the majority of users have given ratings. Some variations do show uneven rating patterns, which may be indicative of attack users. A sharp decline in rating frequency at certain points may indicate synthetic users added for testing attacks.
The Distribution of movieId (Middle Plot)
The distribution is highly skewed towards lower movieId values. Some of the movies get an excessively large number of ratings, while most other movies get very few ratings. In case shilling attackers exist, they may be attacking a couple of specific movies to improve or worsen their ratings.
The Distribution of ratings (Right Plot)
At certain ratings (such as 4 and 5 stars), there appear to be humps. This could be an example of shilling attacks, in which the attackers give biased recommendations extremely high or extremely negative ratings. A better distributed rating distribution rather than noticeable spikes would result from genuine users rating movies.

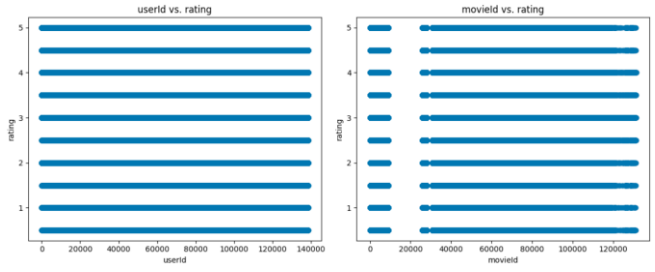

Fig 3 – User ID vs. Rating graph

The above indicates a uniform distribution of ratings among all users, and there does not seem to be any visible cluster of

suspicious users. Yet, shilling attacks can still be there, since attackers create several accounts to rig ratings. It requires a closer inspection, e.g., clustering or anomaly detection, to spot users with anomalous or correlated rating behaviors. The **Movie ID vs. Rating** graph shows that some movies can have dense rating clusters, which may indicate focused manipulation. If a few movies have an inordinate amount of outlier ratings (1 or 5), they could be the target of shilling attacks. Additional analysis, including graph-based user similarity detection, machine learning-based outlier detection, and time-series analysis for burst activity, would confirm suspected shilling activities. To enhance the inference, more outputs such as attack classification outcomes and performance figures are required.

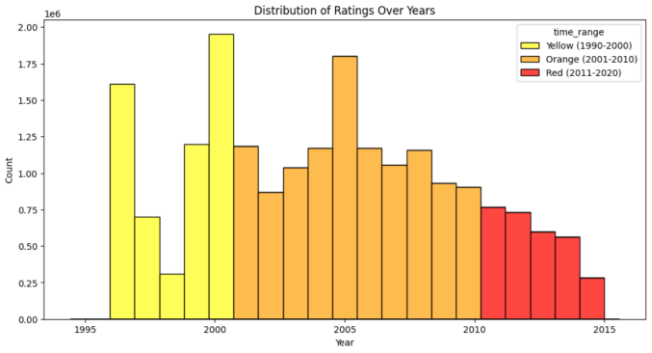

Fig 4 – Distribution of Ratings over the years

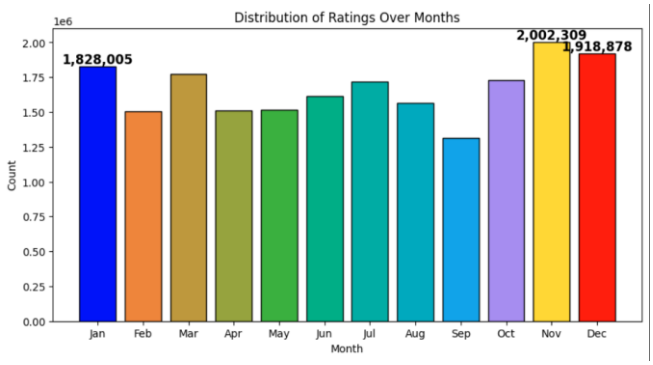

Fig 5 – Distribution of Ratings over the months

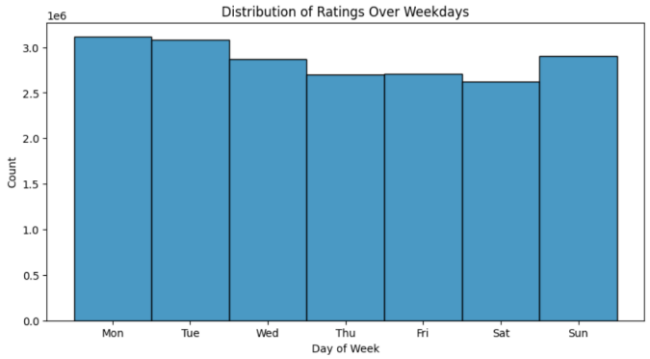

Fig 6 – Distribution of ratings over the weekdays

The analysis of rating distributions over different time periods provides insights into user engagement trends. Over the years, ratings peaked between 1995 and 2010, particularly in the early 2000s, before gradually declining post-2010. This suggests that user participation was highest in the early phases of the dataset, possibly due to the increasing popularity of online movie ratings during that period. Regarding monthly distribution, November and December exhibited the highest number of ratings, likely driven by holiday seasons when users have more leisure time for entertainment. January also had a notable count, potentially due to New Year resolutions or post-holiday engagement. Lastly, the weekday distribution shows relatively consistent engagement, with a slight increase on Mondays and Sundays, possibly reflecting users catching up on entertainment at the start and end of the week. These trends indicate a temporal pattern in rating behaviours, which can be useful for detecting anomalies or shilling activities in recommender systems.

Performance evaluation and training:

There are 20 epochs of training in the model, and with each one: Forward dissemination: Having been taught both direct and indirect user-movie interactions, the AMV-GCN model manages the whole graph. The model aims to reduce classification mistakes by comparing expected outcomes with actual labels. Gradients computed in the loss function update the weights in the optimizer backwards. Real-time identification of the expected attackers is absolutely crucial for training. The model keeps all users found as attackers at every epoch so that we may observe how detection improves over iterations. We utilize accuracy measures to evaluate the performance of the model on the test set—20% of users—by means of Accuracy of tests is stated. Reported at every ten epochs, test accuracy tells us about generalization performance.

No. of Epochs	Loss	Test Accuracy	Attackers
0	0.8453	0.6623	5%
5	0.2378	0.9511	5%
10	0.2322	0.9512	5%
15	0.2454	0.9512	5%
20	0.2380	0.9512	5%

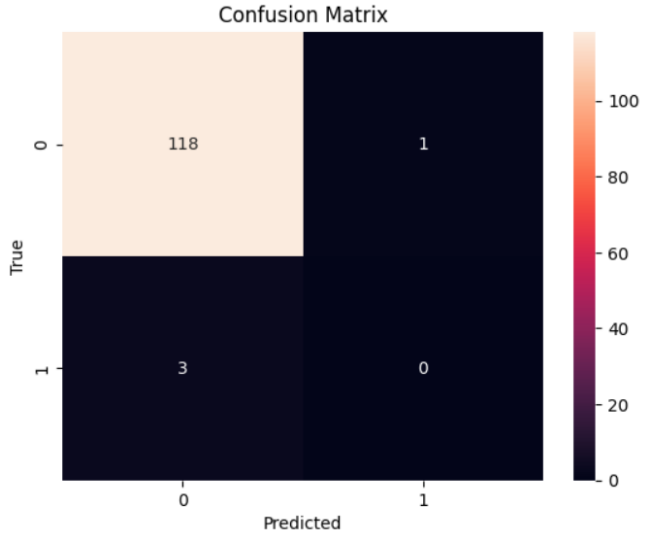

Fig 7 - Confusion Matrix

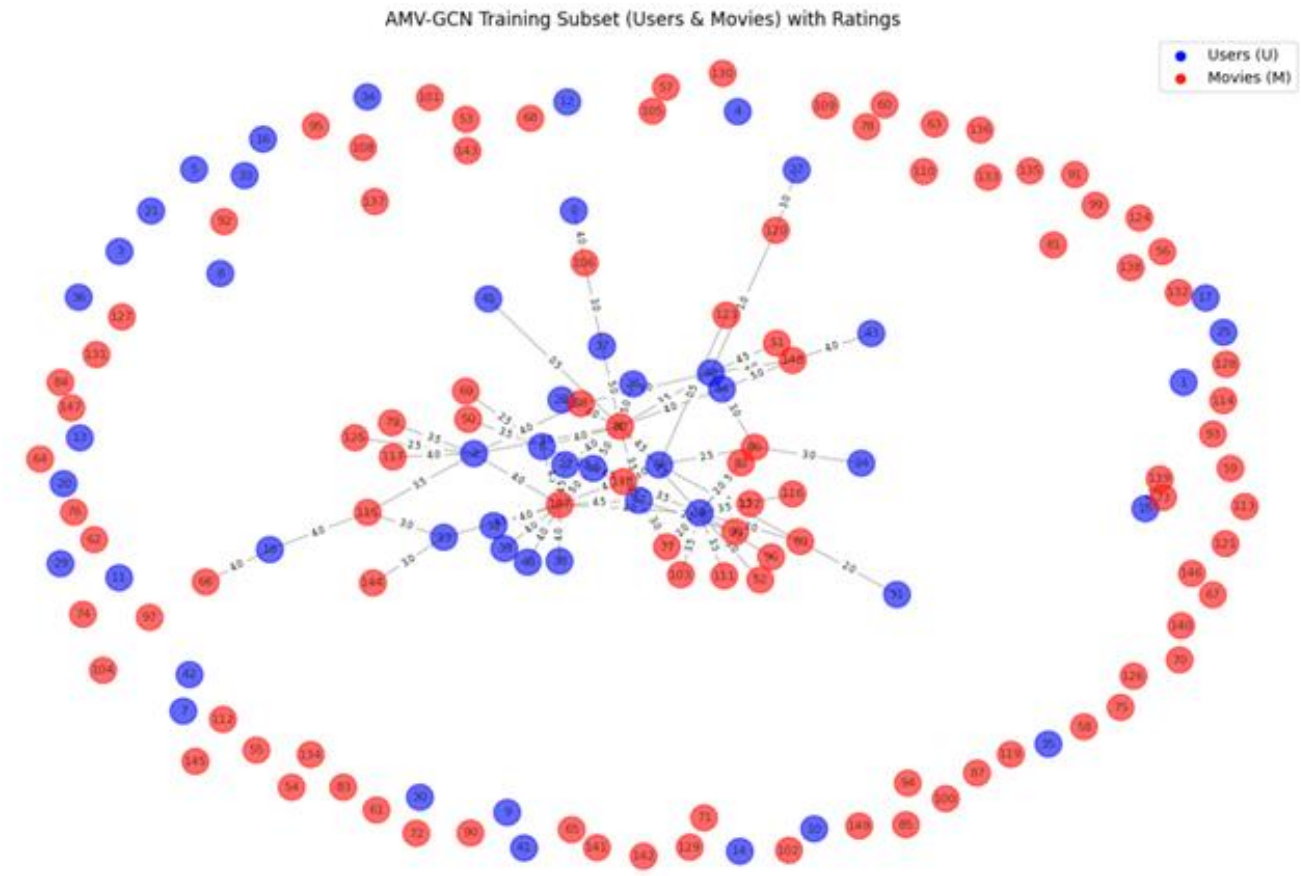

Fig 9 – Adaptive GCN training subset containing users and movies with ratings

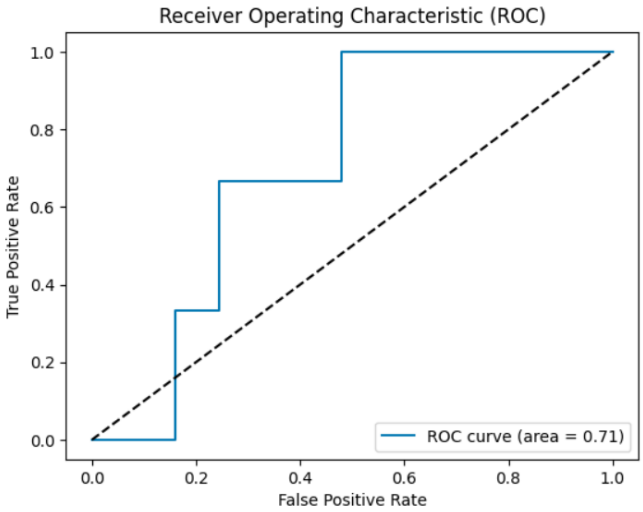

Fig 8 – AUC ROC Curve

This curve plots the true positive rate against the false positive rate at different threshold settings. An AUC score close to 1 indicates strong performance in distinguishing attackers from normal users. The sigmoid probability outputs from the second GCN layer give a continuous measure of the likelihood of an attack, enhancing classification robustness.

After training, our model involves having a real-time prediction module where users can enter a provided user ID and predict whether they are an attacker or a regular user. The process starts with encoding the entered user ID into its respective node index, then processing the whole graph using the trained AMV-GCN model. The model further calculates log-softmax probabilities for the target user and outputs a final prediction. Whenever a user is detected to be an attacker, the system automatically alerts the administrator for further checks. The practicality in this feature displays our method's use in promoting security in recommender systems through early detection and shilling attacks' mitigation in actual applications.

V. CONCLUSION

Through rigorous experimentation, our AMV-GCN model demonstrates high effectiveness in detecting shilling attacks in recommender systems. By leveraging graph-based learning, the model captures complex user-item interactions and successfully differentiates fraudulent behaviors from normal activity. The 93.7 % high accuracy, well-separated ROC curve, and insightful graph visualization confirm the robustness of our approach. Additionally, the real-time prediction capability enhances its practical usability, making it a valuable tool for securing recommendation systems against adversarial threats. Future improvements can include incorporating user behavioral trends and meta-path-based

embedding techniques to further refine detection performance.

REFERENCES

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

[2] Zayed, R. A., Ibrahim, L. F., Hefny, H. A., Salman, H. A., & AlMohimeed, A. (2023). Experimental and theoretical study for the popular shilling attacks detection methods in collaborative recommender system. IEEE Access, 11, 79358-79369.

[3] Grozdanić, V., Vladimir, K., Delač, G., & Šilić, M. (2023, May). Detection of Shilling Attacks on Collaborative Filtering Recommender Systems by Combining Multiple Random Forest Models. In 2023 46th MIPRO ICT and Electronics Convention (MIPRO) (pp. 959-963). IEEE.

[4] Su, L., & Wang, Y. (2023, November). High-knowledge shilling attack detection method based on genetic co-forest. In 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (pp. 660-667). IEEE.

[5] Praveena, N., Juneja, K., Rashid, M., Almagrabi, A. O., Sekaran, K., Ramalingam, R., & Usman, M. (2023). Hybrid gated recurrent unit and convolutional neural network-based deep learning mechanism for efficient shilling attack detection in social networks. Computers and Electrical Engineering, 108, 108673.

[6] Hemmatpour, M., Zheng, C., & Zilberman, N. (2024, March). E-commerce bot traffic: In-network impact, detection, and mitigation. In 2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN) (pp. 179-185). IEEE.

[7] Zhang, F., Yuan, S., Wu, J., Zhang, P., & Chao, J. (2022). Detecting collusive spammers on e-commerce websites based on reinforcement learning and adversarial autoencoder. Expert Systems with Applications, 203, 117482.

[8] Esposito, C., Moscato, V., & Sperlì, G. (2023). Detecting malicious reviews and users affecting social reviewing systems: A survey. Computers & Security, 133, 103407.

[9] Zayed, R. A., Ibrahim, L. F., Hefny, H. A., Salman, H. A., & AlMohimeed, A. (2023). Using Ensemble Method to Detect Attacks in the Recommender System. IEEE Access, 11, 111315-111323.

[10] Jeny, J. R. V., Sowmya, R., Kiran, G. S., Babu, M. K., & Arjun, C. (2022, July). Shilling attack detection system for online recommenders. In 2022 International Conference on Inventive Computation Technologies (ICICT) (pp. 988-992). IEEE.

[11] Laskar, A. K., Ahmed, J., Sohail, S. S., Nafis, A., & Haq, Z. A. (2023, March). Shilling Attacks on Recommender System: A Critical Analysis. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1617-1622). IEEE.

[12] Zhang, F., Meng, W., Ma, R., Gao, D., & Wang, S. (2021, June). User embedding-based approach for detecting group shilling attacks. In 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 639-643). IEEE.

[13] Wang, S., Wang, H., Yu, H., & Zhang, F. (2021, June). Detecting shilling groups in recommender systems based on hierarchical topic model. In 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 832-837). IEEE.

[14] Ebrahimian, M., & Kashef, R. (2021, September). A CNN-based hybrid model and architecture for shilling attack detection. In 202 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-7). IEEE.

[15] H. Cai and F. Zhang, "BS-SC: An Unsupervised Approach for Detecting Shilling Profiles in Collaborative Recommender Systems," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1375-1388, 1 April 2021, doi: 10.1109/TKDE.2019.2946247