Privacy-Preserving Malware Detection Using Federated Learning and AI

Harshith G M[1], Gopimohan Mukherjee[2], K. C. Yashwanth[3], Vijayalakshmi M. M[4], B. Vijaya Nirmala[5]

[1]Harshith G M, Dept. of Information Science and Engineering, AMC Engineering College, Karnataka, India
[2] Gopimohan Mukherjee, Dept. of Information Science and Engineering, AMC Engineering College, Karnataka, India
[3] K C Yashwanth, Dept. of Information Science and Engineering, AMC Engineering College, Karnataka, India
[4] Vijayalakshmi M. M, *Dept. of Information Science and Engineering, AMC Engineering College, Karnataka, India*
[5] Vijaya Nirmala, Dept. of Information Science and Engineering, AMC Engineering College, Karnataka, India

---***---

Abstract – Malware threats have increased significantly with the rapid expansion of mobile devices, cloud platforms, and Internet-of-Things (IoT) ecosystems. Conventional malware detection systems typically rely on centralized machine learning models that require collecting large volumes of sensitive behavioral data from end-user devices. Such centralized data aggregation raises serious concerns related to privacy leakage, regulatory compliance, and user trust. To address these challenges, this paper presents a privacy-preserving malware detection framework based on Federated Learning (FL).

In the proposed approach, malware detection models are trained collaboratively across multiple client devices without transferring raw data to a central server. Each client locally trains a lightweight hybrid CNN–LSTM model using its private dataset and shares only encrypted or masked model updates. Secure aggregation and differential privacy mechanisms are incorporated to prevent reconstruction of individual client data and to provide formal privacy guarantees. The framework is evaluated using non-IID malware datasets distributed across multiple clients to simulate real-world deployment conditions. Experimental results demonstrate that the federated model achieves detection performance close to centralized training while significantly reducing data exposure. The findings indicate that federated learning offers a practical and scalable solution for privacy-aware malware detection in modern distributed environments.

Keywords: Federated Learning, Malware Detection, Privacy Preservation, Secure Aggregation, Differential Privacy, Edge AI.

1.INTRODUCTION

The increasing dependence on digital systems, mobile devices, and interconnected networks has resulted in a rapid rise in malware threats. Modern malware exhibits sophisticated behaviors such as polymorphism, stealth execution, and dynamic payload delivery, making detection increasingly complex. Traditional malware detection techniques, including signature-based and rule-based systems, struggle to cope with zero-day attacks and evolving threat patterns. Machine learning-based approaches have therefore gained popularity due to their ability to learn complex features and generalize to unseen samples.

Despite their effectiveness, most machine learning-based malware detection systems rely on centralized training, where raw behavioral data from end-user devices is collected and processed on a central server. Such centralized data aggregation introduces serious privacy risks, especially when sensitive system logs, application traces, or user-related information are involved. Additionally, regulatory frameworks increasingly restrict centralized storage of personal or device-level data.

Federated Learning (FL) offers a promising alternative by enabling collaborative model training without transferring raw data outside client devices. In an FL setting, each client trains a local model using its private dataset and shares only model updates with a central aggregator. This approach significantly reduces data exposure while still benefiting from collective intelligence. However, applying federated learning to malware detection presents unique challenges, including non-identical data distributions across clients, communication overhead, and vulnerability to adversarial updates.

This research addresses these challenges by proposing a privacy-preserving malware detection framework that integrates federated learning with secure aggregation, differential privacy, and communication-efficient training strategies.

1.1 Problem Statement

Existing malware detection systems largely depend on centralized architectures that require continuous data collection from user devices. While effective in controlled environments, these systems pose major privacy concerns and are vulnerable to data leakage, unauthorized access, and regulatory violations. Furthermore, centralized approaches struggle to scale efficiently across distributed environments such as enterprise endpoints and IoT ecosystems

Another limitation lies in the diversity of malware data across devices. Different users encounter different malware families, resulting in non-IID data distributions that degrade the performance of traditional centralized models. Additionally, bandwidth constraints and computational limitations on edge devices restrict the feasibility of frequent data transmission.

The absence of privacy-aware, scalable, and efficient malware detection mechanisms motivates the need for a decentralized solution. This research aims to address these limitations by designing a federated learning-based framework that ensures strong privacy protection while maintaining high detection accuracy.

1.2 Objective of The Project

The primary objective of this project is to design and implement a privacy-preserving malware detection system using federated learning that enables collaborative model training without exposing sensitive client data. The proposed system aims to overcome the limitations of centralized malware detection approaches by ensuring that raw data remains on user devices while only protected model updates are shared with a central server. By integrating secure aggregation and differential privacy techniques, the project seeks to minimize the risk of data leakage and inference attacks while maintaining high detection accuracy. Additionally, the system is intended to operate efficiently under non-IID data distributions and resource-constrained environments, making it suitable for real-world deployment across edge devices, enterprise endpoints, and IoT ecosystems.

2. RELATED WORK

Early malware detection techniques primarily relied on signature-based and heuristic methods, which are effective for identifying known threats but fail to detect new or evolving malware. To address these limitations, machine learning-based approaches were introduced, utilizing static and dynamic features such as opcode sequences, API calls, and system behavior. These methods improved detection accuracy but generally depend on centralized data collection, leading to privacy and scalability concerns.

Deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have further enhanced malware detection by automatically learning complex patterns from large datasets. Despite their effectiveness, such models require centralized training and extensive computational resources, making them less suitable for privacy-sensitive and distributed environments.

Federated Learning (FL) has emerged as a decentralized alternative that enables collaborative model training without sharing raw data. Studies applying FL to security applications demonstrate its potential to preserve privacy while maintaining competitive performance. However, challenges such as non-IID data distribution, communication overhead, and vulnerability to malicious updates remain.

To strengthen privacy and robustness, secure aggregation and differential privacy techniques have been integrated into federated learning frameworks. Secure aggregation prevents servers from accessing individual client updates, while differential privacy provides formal protection against inference attacks. Recent research also explores communication-efficient and robust aggregation methods to improve practical deployment. This work builds on these advances by applying an integrated federated learning framework to malware detection under realistic conditions.

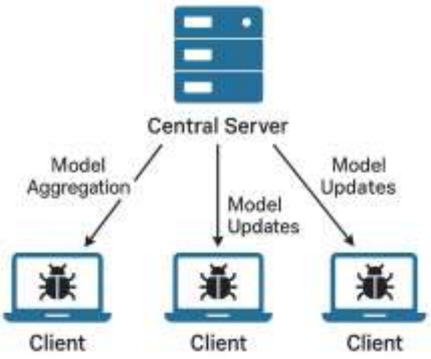

Fig. 1: Federated Learning Architecture for Malware Detection

3.METHODOLOGY

This section describes the design, working principle, and implementation of the proposed privacy-preserving malware detection system. The methodology follows a federated learning paradigm in which multiple client devices collaboratively train a shared model without transferring raw data to a central server. The overall approach is designed to ensure data privacy, scalability, and effective malware detection in distributed environments.

A. System Design and Architecture

The proposed system adopts a client–server architecture based on federated learning. Each client device locally stores malware and benign samples and performs model training using its private dataset. A central aggregation server coordinates the training process by distributing the global model and collecting protected updates from participating clients.

Unlike traditional centralized systems, raw malware data never leaves the client devices. Only encrypted or masked model updates are transmitted to the server, significantly reducing privacy risks. The server aggregates these updates to generate an improved global model, which is then shared with clients for subsequent training rounds.

B.Local Model Training

Each client trains a lightweight hybrid deep learning model consisting of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers. The CNN component extracts spatial features from malware representations such as opcode sequences or n-gram features, while the LSTM component captures sequential and temporal patterns in malware behavior.

Local training is performed for a limited number of epochs in each communication round to reduce computational overhead and communication cost. This approach allows the system to operate efficiently on edge devices with limited resources.

C.Federated Aggregation Process

After local training, each client prepares its model updates for transmission. To preserve privacy, secure aggregation techniques are applied so that individual updates cannot be inspected by the server. The server aggregates the received updates using a weighted averaging method based on the amount of local data available at each client.This iterative aggregation process continues across multiple rounds until the global model converges. Federated averaging enables the system to learn from diverse data distributions while maintaining collaborative intelligence across clients.

D.Privacy Preservation Techniques

To further enhance privacy, differential privacy is incorporated at the client level. Before transmission, model updates are clipped to limit sensitivity and noise is added to prevent inference attacks. This ensures that information about individual data samples cannot be extracted from the shared updates. The combination of secure aggregation and differential privacy provides layered protection, making the system resilient against data leakage and malicious inference attempts.minimizing communication cost, the system becomes suitable for deployment .

such

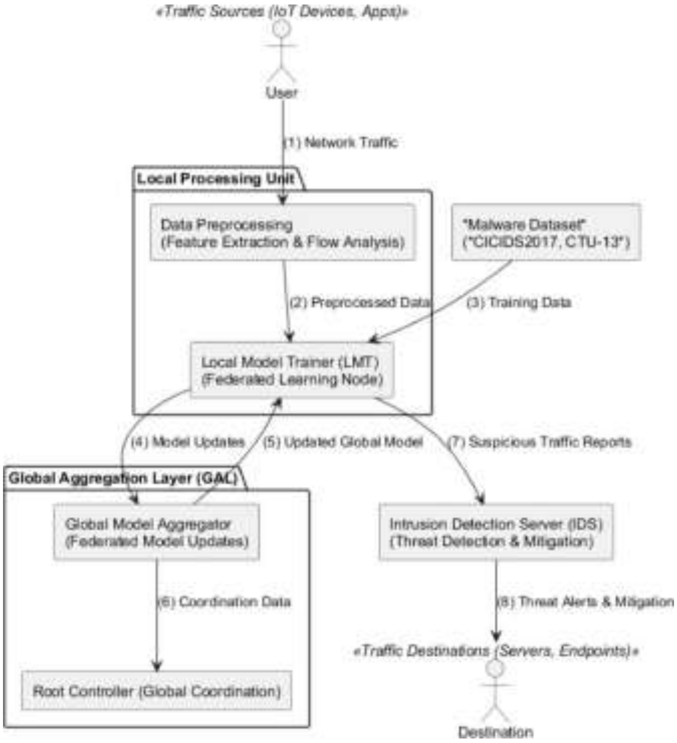
«Traffic Sources (IoT Devices, Apps)»
User
(1) Network Traffic
Local Processing Unit
Data Preprocessing
(Feature Extraction & Flow Analysis)
"Malware Dataset"
("CICIDS2017, CTU-13")
(2) Preprocessed Data
(3) Training Data
Local Model Trainer (LMT)
(Federated Learning Node)
(4) Model Updates
(5) Updated Global Model
(7) Suspicious Traffic Reports
Global Aggregation Layer (GAL)
Global Model Aggregator
(Federated Model Updates)
Intrusion Detection Server (IDS)
(Threat Detection & Mitigation)
(6) Coordination Data
(8) Threat Alerts & Mitigation
«Traffic Destinations (Servers, Endpoints)»
Root Controller (Global Coordination)
Destination

Fig 2.0: *Proposed Methodology Architecture*

4. RESULTS AND DISCUSSIONS

This section presents the performance evaluation and analysis of the proposed federated learning-based malware detection system. The system was tested under distributed and non-IID data conditions to simulate real-world deployment scenarios where different client devices encounter different malware patterns.

Detection Accuracy and Model Performance

The experimental results demonstrate that the federated learning model achieves high malware detection accuracy while preserving data privacy. The global model trained through federated aggregation performs closely to a centrally trained model, with only a marginal reduction in accuracy. This performance gap is primarily due to non-IID data distribution and privacy-preserving noise addition.

The hybrid CNN–LSTM architecture effectively captures both spatial and sequential characteristics of malware behavior. CNN layers extract discriminative features from malware representations, while LSTM layers model temporal dependencies, resulting in improved classification performance compared to traditional machine learning models.

Impact of Federated Learning

Impact of Federated Learning projection dimensionality improved accuracy. Bicubic interpolation produced the most consistent results for low-resolution face images captured at a distance, with significant improvements in recognition rates compared to nearest-neighbour and bilinear interpolation techniques. This validates PCA’s effectiveness in distributed or low-resolution surveillance setups.

Environmental Impact

Communication Efficiency Analysis

The integration of secure aggregation prevents the server from accessing individual client updates, ensuring confidentiality of local training information. Differential privacy introduces a controlled amount of noise to model updates, providing formal privacy guarantees. While this results in a slight decrease in detection accuracy, the trade-off is acceptable given the enhanced privacy protection.

The results indicate that combining secure aggregation with differential privacy offers strong resistance against data leakage and inference attacks without significantly degrading system performance.

Communication Efficiency Analysis

Communication optimization techniques such as parameter quantization and selective update transmission significantly reduce bandwidth consumption. Experimental observations show that communication overhead is reduced with minimal impact on detection accuracy. This makes the proposed system suitable for deployment on mobile and edge devices where bandwidth and power resources are limited.

Overall Discussion

The results confirm that federated learning is a viable and effective approach for malware detection in distributed environments. The proposed system successfully balances accuracy, privacy, and efficiency. Compared to centralized approaches, the system offers improved data security and regulatory compliance while maintaining competitive performance.

To visualize the training progress and detection outcomes, a web-based dashboard was developed. The dashboard provides real-time insights into model performance, client participation, and malware classification results, enabling effective monitoring of the federated learning process.

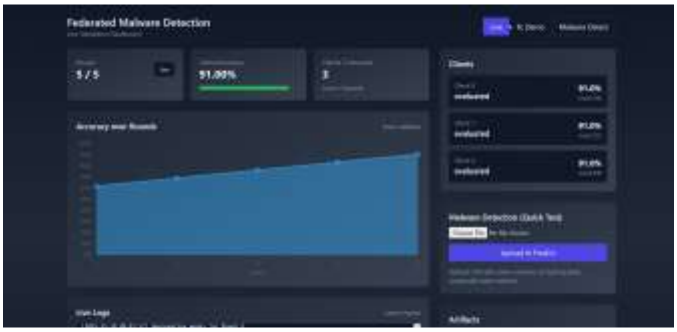

Fig 3.0: *dashboard*

As shown in Fig. 3, the dashboard displays key metrics such as detection accuracy, training rounds, and classification status. This visualization assists in understanding the convergence behavior of the global model and offers transparency into the system's operation. The dashboard also supports easier analysis and validation of results during experimentation.

5. CONCLUSION

This research presented a comprehensive privacy-preserving malware detection framework based on federated learning, aimed at overcoming the limitations of conventional centralized detection systems. By enabling decentralized model training across multiple client devices, the proposed system ensures that sensitive malware data remains local while still allowing collaborative intelligence to be achieved. This design significantly reduces privacy risks associated with data aggregation and aligns well with modern data protection and regulatory requirements.

The integration of a hybrid CNN–LSTM model proved effective in capturing both spatial and temporal characteristics of malware behavior. Experimental evaluation under non-IID data distributions demonstrated that the federated model achieves detection performance close to that of centralized training approaches, with only a marginal reduction in accuracy. The use of secure aggregation ensured that individual client updates could not be reconstructed by the server, while differential privacy provided formal guarantees against inference and membership attacks. Although these privacy mechanisms introduce slight computational and accuracy trade-offs, the overall impact remains acceptable for real-world security applications.

In addition to privacy preservation, the proposed framework addressed practical deployment challenges such as communication overhead and resource constraints. Communication-efficient techniques, including parameter quantization and selective update transmission, significantly reduced bandwidth usage, making the system suitable for edge devices, enterprise endpoints, and IoT environments. The inclusion of a monitoring dashboard further enhanced system transparency by providing real-time visualization of training progress, detection outcomes, and overall system performance.

Overall, the proposed federated malware detection system demonstrates a balanced trade-off between accuracy, privacy, and efficiency. It offers a scalable and cost-effective solution for modern distributed computing environments where centralized data collection is undesirable or impractical. Future research can focus on extending this work through personalized federated models, asynchronous training mechanisms, and semi-supervised learning techniques to further improve adaptability and robustness against evolving malware threats. The outcomes of this study indicate that federated learning is a promising direction for building next-generation, privacy-aware cybersecurity systems.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, AMC Engineering College, for providing the necessary infrastructure and academic support to carry out this research. We are thankful to our project guides and faculty mentors for their continuous guidance, encouragement, and valuable suggestions throughout the course of this work. We also acknowledge our peers and contributors for their assistance in experimentation and discussions, which helped improve the quality of this research.

REFERENCES

1. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

2. K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pp. 1175–1191, 2017.

3. C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

4. P. Kairouz et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

5. R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pp. 1310–1321, 2015.

6. A. Pektaş and T. Acarman, "Learning Malware Behaviors Using Machine Learning Techniques," *IEEE Transactions on Cybernetics*, vol. 49, no. 7, pp. 2582–2593, 2019.

7. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.