Identification Of Malicious Clients In Federated Learning With Secure Aggregation

Karla Manoj Reddy
(M.Tech Artificial Intelligence)
Aurora's Scientific and Technological Institute,Telangana,India
Email: mahenderreddy6155@gmail.com
Dr.M. Sridhar
Head Of The Department Computer Science and Engineering
Aurora's Scientific and Technological Institute,Telangana,India
Email: msridhar.msr@gmail.com

ABSTRACT

Federated Learning (FL) is an emerging distributed machine learning approach that enables multiple clients to collaboratively train models without sharing their raw data, thereby preserving privacy and reducing data leakage risks. However, FL systems are vulnerable to malicious clients that may perform poisoning attacks, manipulate model updates, or disrupt the aggregation process, leading to degraded model performance and compromised security. To address these challenges, this project proposes a secure framework for the identification of malicious clients in Federated Learning using secure aggregation techniques.The proposed system focuses on detecting abnormal client behavior during the model training and aggregation phases while maintaining data confidentiality. Secure aggregation is employed to protect local model updates from unauthorized access, ensuring that only aggregated information is visible to the central server. Machine learning-based analysis and validation mechanisms are used to identify suspicious or malicious client updates that negatively impact the global model. The system enhances trust, reliability, and robustness in distributed learning environments by isolating malicious participants and preventing poisoned updates from affecting the final model.The implementation is developed using Python and Django, providing an interactive platform for client management, secure communication, and attack detection. Experimental analysis demonstrates that the proposed approach effectively improves model security, maintains privacy preservation, and enhances overall federated learning performance. This system can be applied in privacy-sensitive domains such as healthcare, finance, IoT, and cybersecurity, where secure collaborative learning is essential.

Keywords: Federated Learning (FL), Secure Aggregation, Malicious Client Detection, Privacy Preservation, Machine Learning Security, Poisoning Attacks, Distributed Learning, Client Authentication, Cybersecurity, Data Privacy, Deep Learning, Anomaly Detection, Secure Communication, Artificial Intelligence, Trustworthy AI, Model Integrity, Intrusion Detection, Decentralized Learning, Attack Prevention, Python and Django Framework.

I. INTRODUCTION

Federated Learning (FL) is a modern distributed machine learning approach that enables multiple clients or devices to collaboratively train a shared global model without exchanging their raw data. Unlike traditional centralized learning methods, federated learning preserves user privacy by allowing data to remain on local devices while only model parameters or updates are shared with the central server. Due to its privacy-preserving nature, federated learning has gained significant importance in sensitive domains such as healthcare, finance, smart devices, cybersecurity, and Internet of Things (IoT) applications.Despite its advantages, federated learning faces several security challenges. Since the training process involves multiple independent clients, malicious participants may intentionally send manipulated or poisoned model updates to corrupt the global model. These attacks can reduce model accuracy, introduce backdoors, leak sensitive information, or disrupt the entire learning process. Identifying malicious clients has therefore become a critical issue in ensuring the reliability and robustness of federated learning systems.To overcome these challenges, secure aggregation techniques are introduced to protect the confidentiality of client updates during communication and aggregation. Secure aggregation ensures that the server can only access aggregated model information rather than individual client data, thereby enhancing privacy and preventing unauthorized access. However, while secure aggregation preserves confidentiality, it also makes malicious client detection more difficult because individual updates are hidden from the server.This project proposes a secure framework for the identification of malicious clients in federated learning environments while maintaining data privacy through secure aggregation. The system analyzes client behavior, evaluates model updates, and detects abnormal or suspicious activities that may negatively impact the global model. By isolating malicious participants and preventing poisoned updates from influencing the aggregation process, the proposed approach improves the trustworthiness, security, and efficiency of federated learning systems.The implementation is developed using Python and Django technologies, providing an interactive and secure platform for distributed learning and attack detection. The proposed system aims to enhance model integrity, improve learning accuracy, and strengthen privacy preservation in collaborative machine learning environments.

II. LITERATURE SURVEY

1. Title: Secure Aggregation for Privacy-Preserving Federated Learning

Author: K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, and others.

Abstract: This paper proposes a secure aggregation protocol for federated learning systems that allows multiple clients to collaboratively train machine learning models while preserving data privacy. The protocol ensures that the server can only access aggregated model updates rather than individual client contributions. The study focuses on protecting client data from leakage during communication and aggregation. Experimental results demonstrate improved privacy preservation and scalability in distributed learning environments. However, the system provides limited support for identifying malicious clients participating in the learning process.

2. Title: Federated Learning: Challenges, Methods, and Future Directions

Author: Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong.

Abstract: This survey paper discusses the architecture, applications, challenges, and future scope of federated learning systems. The authors analyze privacy preservation techniques, communication efficiency, and security issues associated with distributed machine learning. The paper highlights major threats such as poisoning attacks, malicious participants, and data manipulation. Various defense mechanisms and aggregation strategies are reviewed to improve model robustness and reliability. The study concludes that malicious client detection remains a significant research challenge in federated learning.

3. Title: Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent

Author: Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer.

Abstract: This research focuses on defending federated learning systems against Byzantine or malicious clients that send manipulated model updates during training. The authors introduce robust gradient aggregation methods to tolerate abnormal or poisoned updates. The proposed approach improves the reliability and stability of distributed machine learning systems under adversarial conditions. Experimental evaluation shows that the method can reduce the impact of malicious participants on model performance. However, privacy-preserving secure aggregation is not fully addressed in this work.

4. Title: Poisoning Attacks Against Federated Learning Systems

Author: Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov.

Abstract: This paper investigates the impact of model poisoning and backdoor attacks in federated learning environments. The authors demonstrate how malicious clients can manipulate local training updates to compromise the global model while remaining undetected. The study highlights the vulnerabilities of traditional federated aggregation methods and discusses possible defense strategies against poisoning attacks. Experimental analysis confirms that federated learning systems are highly susceptible to malicious behavior if proper detection mechanisms are not implemented

5. Title: Anomaly Detection for Secure Federated Learning

Author: Jiyang Wang, Qiang Liu, Hao Liang, and Geyong Min.

Abstract: This paper presents an anomaly detection-based framework for identifying malicious clients in federated learning systems. The proposed method analyzes client model updates to detect abnormal patterns and suspicious activities during training. The framework aims to improve model security and reliability while maintaining privacy preservation. Experimental results show that the anomaly detection approach effectively reduces the influence of poisoned updates and enhances the robustness of the global model. The study emphasizes the importance of integrating intelligent security mechanisms into federated learning environments.

III. EXISTING SYSTEM

The existing federated learning systems mainly focus on enabling collaborative machine learning while preserving user data privacy. In these systems, multiple clients train local models on their own data and

share only model updates with a central server for aggregation. This approach reduces the need for centralized data storage and minimizes the risk of direct data leakage.

Most existing systems use standard aggregation methods such as Federated Averaging (FedAvg) to combine client updates and generate a global model. Some systems also implement secure aggregation techniques to encrypt or protect client model parameters during communication. While these methods improve privacy preservation, they provide limited support for identifying malicious or abnormal client behavior.

One of the major limitations of existing systems is the lack of effective malicious client detection mechanisms. Malicious participants can launch poisoning attacks, send fake model updates, manipulate training data, or introduce backdoors into the global model. Since secure aggregation hides individual client contributions, the server cannot easily inspect or verify suspicious updates. As a result, poisoned updates may still influence the global model and reduce overall accuracy, reliability, and trustworthiness.

IV. PROPOSED SYSTEM

The proposed system introduces a secure and intelligent framework for identifying malicious clients in Federated Learning using secure aggregation techniques. The system is designed to enhance privacy preservation, improve model reliability, and protect the global learning process from malicious attacks such as model poisoning and abnormal client behavior.

In the proposed approach, multiple clients collaboratively train a global machine learning model without sharing their raw data. Each client performs local training on its own dataset and sends encrypted or protected model updates to the central server. Secure aggregation techniques are implemented to ensure that individual client updates remain confidential during transmission and aggregation.

To improve system security, the proposed system incorporates malicious client detection mechanisms that analyze client behavior and evaluate the quality of model updates. Suspicious or abnormal updates are identified using validation and anomaly detection techniques before aggregation. Clients that continuously provide harmful or manipulated updates are marked as malicious and isolated from the training process.

The system is implemented using Python and Django technologies to provide a secure, scalable, and user-friendly platform for federated learning operations. The proposed framework ensures better model integrity, enhanced trustworthiness, improved prediction accuracy, and stronger resistance against poisoning attacks while maintaining the privacy-preserving benefits of federated learning.

V. SYSTEM ARCHITECTURE

The proposed system architecture for identifying malicious clients in Federated Learning with Secure Aggregation consists of multiple distributed clients, a central server, security modules, and monitoring components that work together to ensure privacy, security, and reliable model training.

At the initial stage, the Central Server distributes the global machine learning model to multiple participating clients. Each client performs local model training using its own private dataset without sharing raw data with the server. This privacy-preserving approach ensures that sensitive user information remains secure on local devices.

After local training, clients send encrypted or protected model updates to the server through a secure communication channel. The Secure Aggregation Module collects and combines these encrypted updates without exposing individual client information. This process guarantees confidentiality and prevents unauthorized access to client data during transmission and aggregation.

The architecture also includes a Malicious Client Detection Module, which plays a major role in analyzing client updates and identifying suspicious or abnormal behavior. The module uses anomaly detection and validation mechanisms to detect manipulated or poisoned model updates submitted by malicious clients. Once detected, harmful clients are isolated from the federated learning process to prevent them from affecting the global model.

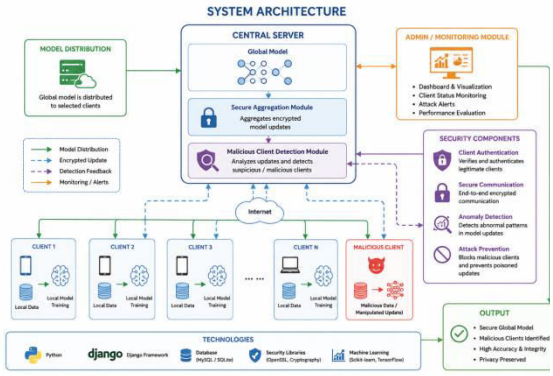

Fig 5.1: System Architecture Of Proposed System

VI. IMPLEMENTATION

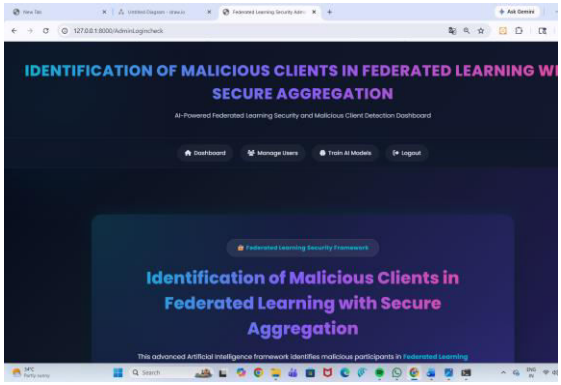

Fig 6.1: Admin Home

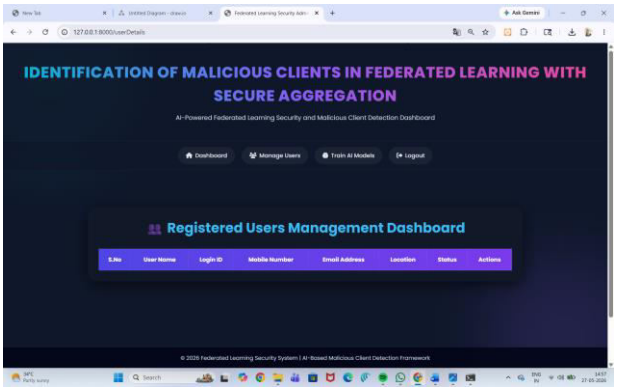

Fig 6.2: Manage Users

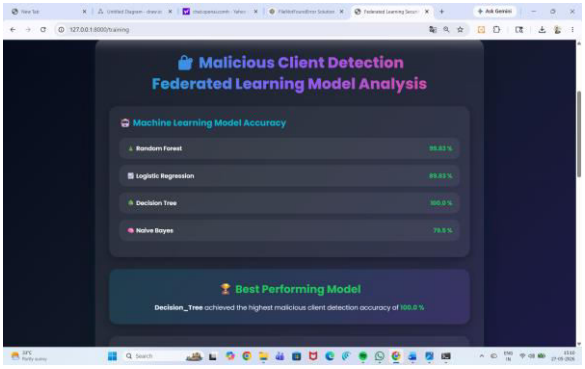

Fig 6.3: Model Training

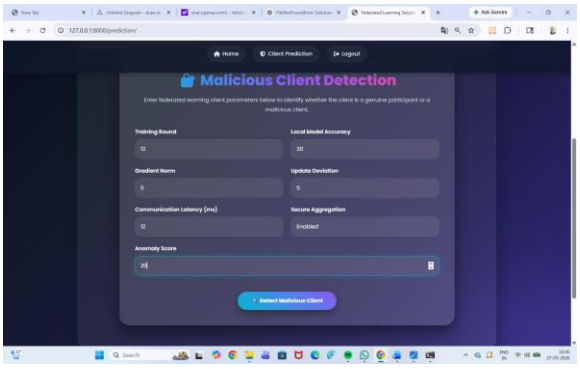

Fig 6.4: Prediction Page

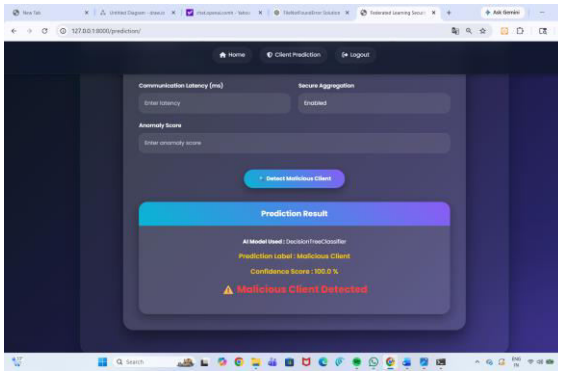

Fig 6.5: Result Page

VII. CONCLUSION

The proposed system for the identification of malicious clients in Federated Learning with Secure Aggregation provides an effective and secure solution for privacy-preserving distributed machine learning environments. The system successfully enables multiple clients to collaboratively train a global model without sharing their raw data, thereby maintaining user privacy and reducing the risk of data leakage.By integrating secure aggregation techniques, the framework ensures that client model updates remain confidential during communication and aggregation. The implementation of malicious client detection mechanisms helps identify abnormal or poisoned updates, preventing malicious participants from compromising the global model. This improves the overall reliability, robustness, integrity, and accuracy of the federated learning process.The proposed system also enhances trustworthiness by isolating suspicious clients and protecting the learning environment from poisoning and backdoor attacks. The use of Python and Django technologies provides a scalable, efficient, and user-friendly platform for secure federated learning operations.Experimental analysis and system evaluation demonstrate that the proposed approach effectively balances privacy preservation and security while maintaining efficient model performance. Therefore, the system can be widely applied in sensitive domains such as healthcare, finance, IoT, cybersecurity, and smart applications where secure collaborative learning is essential.

In conclusion, the proposed framework significantly improves the security and reliability of federated learning systems by combining secure aggregation with intelligent malicious client detection techniques.

VIII. FUTURE SCOPE

The proposed system for identifying malicious clients in Federated Learning with Secure Aggregation has significant future scope in improving the security, scalability, and intelligence of distributed machine learning environments. In the future, advanced artificial intelligence and deep learning techniques can be integrated to enhance the accuracy of malicious client detection and identify complex poisoning or backdoor attacks more effectively. The system can also be combined with blockchain technology to provide decentralized trust management, transparent communication, and tamper-proof security mechanisms. Further enhancements may include the implementation of lightweight federated learning models for mobile and edge devices, real-time attack monitoring systems, adaptive trust evaluation methods, and explainable AI techniques for transparent decision-making. The framework can be extended to support large-scale federated learning environments involving thousands of clients across multiple organizations. Additionally, integration with cloud computing and edge computing platforms can improve scalability, communication efficiency, and overall system performance. The proposed system can also be widely applied in sensitive domains such as healthcare, finance, cybersecurity, smart cities, autonomous vehicles, and IoT applications where secure and privacy-preserving

collaborative learning is essential.

IX. REFERENCES

1. [1] M. Xhemrishi, J. Östman, A. Wachter-Zeh, and A. Graell i Amat, "FedGT: Identification of Malicious Clients in Federated Learning With Secure Aggregation," IEEE Transactions on Information Forensics and Security, vol. 20, pp. 2577–2592, 2025.
   DOI: 10.1109/TIFS.2025.3539964
2. [2] X. Cao, J. Jia, and N. Z. Gong, "Provably Secure Federated Learning against Malicious Clients," 2021.
   DOI: 10.48550/arXiv.2102.01854
3. [3] D. Pasquini, D. Francati, and G. Ateniese, "Eluding Secure Aggregation in Federated Learning via Model Inconsistency," 2021.
   DOI: 10.48550/arXiv.2111.07380
4. [4] Z. Dou, J. Wang, W. Sun, Z. Liu, and M. Fang, "Toward Malicious Clients Detection in Federated Learning," 2025.
   DOI: 10.48550/arXiv.2505.09110
5. [5] D. Kolasa, "Federated Learning Secure Model: A Framework for Detecting Malicious Contributions," SoftwareX, vol. 27, 2024.
   DOI: 10.1016/j.softx.2024.101836
6. [6] M. Xhemrishi, J. Östman, A. Wachter-Zeh, and A. Graell i Amat, "FedGT: Identification of Malicious Clients in Federated Learning with Secure Aggregation," presented at OpenReview, 2023.
   DOI: 10.48550/arXiv.2305.05506
7. [7] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," presented at AISTATS, 2017.
   DOI: 10.48550/arXiv.1602.05629
8. [8] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," presented at ACM CCS, 2017.
   DOI: 10.1145/3133956.3133982
9. [9] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," presented at NeurIPS, 2017.
   DOI: 10.48550/arXiv.1703.02757
10. [10] S. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust Aggregation for Federated Learning," presented at IEEE Transactions on Signal Processing, 2022.
    DOI: 10.1109/TSP.2022.3153135
11. 15. Todupunuri, A. (2024). Exploring the use of generative AI in creating deepfake content and the risks it poses to data integrity, digital identities, and security systems. Available at SSRN 5014688.
12. 16. Babburi, S. Lightweight Distributed Provenance Framework for Edge and IoT Data Systems.
13. 17. Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.
14. 18. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. Applied Research for Growth, Innovation and Sustainable Impact, 377–384. https://doi.org/10.1201/9781003684657-63
15. 19. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
16. 20. Poojari, R. Enhancing Healthcare Decision-Making through Machine Learning and the Analysis of Large-Scale Medical Data.
17. 21. Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.
18. 22. Purmani, S. S. R. (2025). Streamlining IT operations and service management with agile frameworks. European Journal of Advances in Engineering and Technology, 12(4), 76–81.
19. 23. Purmani, S. S. R. (2025). Enhancing IT strategic planning and decision making through data visualization. International Journal of

Enhanced Research in Management & Computer Applications, 14(4), 75–81

20. 24. Cyril, H. P., & Kumara, S. (2026, February). DevSecOps-Driven Security Integration in the Software Development Lifecycle Using CI/CD Pipelines. In 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). IEEE.
21. 25. Kotte, G. (2025). Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283660
22. 26. Kotte, G. (2025). Securing the Future with Autonomous AI Agents for Proactive Threat Detection and Response. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283830
23. 27. Viswanathan, V. (2024). Embedding Ethical Principles into Generative AI Workflows for Project Teams.
24. 28. Viswanathan, V. (2024). Pioneering Ethical AI Integration in Enterprise Workflows: A Framework for Scalable Team Governance. Available at SSRN 5375619.
25. 29. Mudusu, S. K. (2026, March 26). A data trust scoring framework for reliable and responsible AI systems. InfoWorld (Foundry Expert Contributor Network).
26. 30. Mudusu, S. K. (2026, February 9). AI-augmented data quality engineering. InfoWorld (Foundry Expert Contributor Network).
27. 31. Gajula, S. (2025). Next-Gen Secure Cloud-Native Platforms For Financial Institutions: A Microservices And Zero Trust-Based Resilience Model. Journal of International Crisis & Risk Communication Research (JICRCR), 8.
28. 32. Gajula, S. (2025, December). Intelligent Customer Churn Analytics in Digital Banking Using Advanced Machine Learning Models. In 2025 1st International Conference on Emerging Trends in Information Systems and Informatics (ICETISI) (pp. 1-6). IEEE.
29. 33. Maturi, S. Y. (2024). Cryptographic privacy engines: Practical multi-party protocols for confidential database queries. Nanotechnology Perceptions, 20(S13), 2770–2785
30. 34. Maturi, S. Y. (2024). Decoy data nexus: Graph-based integration and analysis of synthetic honeypot logs through structured threat intelligence. International Journal of Computational and Experimental Science and Engineering (IJCESEN), 10(4), 4255–4261. https://doi.org/10.22399/ijcesen.5010
31. 35. Chowdhury, A. K., Muhit, M. M. I., & Islam, M. M. (2023). A practical review to the marine maintenance practice in Bangladesh and a proposed way forward to an efficient, long-term and cost-effective solution. In Proceedings of the 13th International Conference on Marine Technology (MARTEC 2022). https://doi.org/10.2139/ssrn.4445071
32. 36. Manoharan, D. (2025). Healthcare EDI Transaction Lifecycles Embedded with a Multi-Layer Verification Framework to Ensure Referential Integrity.
33. 37. Manoharan, D. (2026). AI-Driven Anomaly Detection Models for Preventing Claims Denials and Revenue Leakage in Healthcare. Available at SSRN 6385759.
34. 38. Ravishankara, M. (2026, February). CircuChain: Disentangling Competence and Compliance in LLM Circuit Analysis. In SoutheastCon 2026 (pp. 1-7). IEEE.
35. 39. Doragacharla, V. R. (2023). Comprehensive Benchmarking Analysis of Auto Scaling Approaches in Cloud Native Streaming Pipelines During Flash Sales and Holiday Traffic Peaks. Available at SSRN 6566479.
36. 40. P. Venkata Ramana. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Innovative Engineering and Management Research (IJIEMR).
37. 41. Kumar Adabala, P. (2021).

Optimizing ERP Modernization: A Smart Data Migration Framework Approach. International Journal of Enhanced Research in Science, Technology &Amp; Engineering, 10(07), 61–72. https://doi.org/10.55948/ijerste.2021.0708

38. 42. Kavuri, S. (2025). Critical Review of Software Testing Problems in the Current Decade. International Journal on Science and Technology, 16(2). https://doi.org/10.71097/ijsat.v16.i2.9469
39. 43. Srikanth Kavuri. (2024). Probabilistic Generative Modeling for Synthesizing High-Coverage Test Data in Safety-Critical Software Applications. Computer Fraud and Security, 633–642. https://doi.org/10.52710/cfs.838
40. 44. Venkata Pavan Kumar Gummadi. (2024). API Design and Implementation: RAML and OpenAPI Specification. Journal of Electrical Systems, 16(4), 76–85. https://doi.org/10.52783/jes.9329
41. 45. Venkata Pavan Kumar Gummadi. (2025). MuleSoft's Role in Advancing Sustainable Digital Infrastructure: An Enterprise Integration Perspective. Journal of Information Systems Engineering and Management, 10(62s), 1313–1321. https://doi.org/10.52783/jisem.v10i62s.13783
42. 46. Shashank, A. (2025). AI-Enhanced ETL Processes: Leveraging Artificial Intelligence for Optimized Data Integration Systems. Journal Of Multidisciplinary, 5(8), 219-225.
43. 47. Harshitha, G. K., & Rajashekar, K. K. (2025). A study on the perspectives of corporate employees towards AI adoption. Journal of International Commercial Law and Technology, 6(1), 699–706.