iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** II **Month of publication:** February 2025

DOI: https://doi.org/10.22214/ijraset.2025.66970

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

Privacy - Preserving Anomaly Detection Using Federated Learning and Explainable AI

Divya R[1], Mrs. S. Ramya[2], Indhu Tharshini N[3]
Computer Science & Business Systems, Rajalakshmi Institute of Technology

Abstract: ***Anomaly detection is crucial for identifying security threats and system failures. Traditional methods often require centralized data collection and raising privacy concerns. This paper proposes an idea of privacy-preserving anomaly detection system using Federated Learning (FL), Explainable AI (XAI) and Generative Adversarial Networks (GAN). Federated Learning provides with decentralized training while preserving the data privacy and Explainable AI enhances model transparency, helping in decision making. By utilizing deep autoencoders for anomaly detection and SHAP/LIME for explainability, it ensures secure and interpretable anomaly detection across distributed environments. The proposed model is trained and validated using real-world datasets, demonstrating effectiveness in identifying anomalies while minimizing privacy risks.***
Keywords: Anomaly Detection, Privacy Preserving, Cybersecurity, Data Security, Explainable AI, Federated learning.

I. INTRODUCTION

With the increase in Digital interactions, anomaly detection becomes essential in cybersecurity, fraud detection, and system monitoring. Traditional anomaly detection systems rely on centralized data aggregation, raising privacy and security concerns. Federated Learning enables decentralized training across multiple nodes, ensuring that data remains localized. Additionally, integration of GANs ensures the model generalizes better to counter zero-day attacks. Federated Learning models often operate as black-box systems, making it difficult to interpret decisions. Hence, we integrated Explainable AI(XAI) techniques, providing transparency and trustworthiness to anomaly detection models. Additionally, the proposed system ensures real-time detection and integrates seamlessly into existing security infrastructures.

II. LITERATURE REVIEW

Anomaly detection has been a critical area of research due to its significance in various fields such as cybersecurity, fraud detection and system monitoring. Traditional approaches include statistical models, machine learning and deep learning techniques. But these methods rely heavily on centralized datasets, which pose challenges in terms of privacy, scalability and security of data along with regulatory compliance. The emergence of Federated Learning has revolutionized anomaly detection by enabling distributed learning without exposing raw data. Additionally, the explainability of AI models remains a concern in anomaly detection. Black-box Ai models, although powerful lack transparency, making it difficult for analysts and organizations to trust the detection outcomes. Explainable AI(XAI)bridges the gap by providing insights into model decisions, thereby increasing interpretability and trustworthiness.

A. Anomaly Detection Techniques

Traditional anomaly detection techniques include statistical models, clustering based methods, and deep learning architectures. Among these, Deep autoencoders have shown promising results in detecting deviations in network traffic patterns. But these methods often require large amounts of labeled data and can suffer from high false positives if not optimized effectively. To address these issues, we have proposed to use Generative Adversarial Networks (GANs) to generate synthetize attack data. This is effective in improving the model generalization.

B. Privacy concerns in Centralized Learning

Centralized learning poses significant security risks of data breaches and non-compliance with regulations like GDPR. Federated Learning mitigates these risks by keeping data distribute d while allowing model training in distributed nodes. It prevents direct data sharing, ensuring compliance with privacy regulations. To address these issues, Federated Learning (FL) has emerged as a promising privacy-preserving solution. McMahan introduced FL as a decentralized learning paradigm where data remains distributed across multiple clients, and only model updates (not raw data) are shared.

©IJRASET: All Rights are Reserved

This prevents unauthorized data exposure while maintaining strong machine learning performance. Recent work by Cheema demonstrated that FL-based anomaly detection significantly enhances security while reducing the attack surface, making it suitable for network intrusion detection and cybersecurity applications.

C. *Explainability in AI Models*

XAI techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) help understand AI decision making, fostering trust and compliance in critical applications. These tools enable security analysts to interpret and analyze anomaly detection results effectively.

III. EXISTING SYSTEM

Current anomaly detection methods predominantly rely on centralized data processing, which comes with multiple challenges.

1) *Data Privacy Risks:* Storing sensitive data in a central location increases vulnerability to cyber-attacks and data breaches.
2) *Lack of interpretability:* Most machine learning based anomaly detection models function as black boxes, making it difficult to understand their predictions.
3) *High Computational Costs:* Centralized models require substantial computational resources, making scalability a challenge.
4) *Regulatory Compliance Issues:* Many regulations, such as GDPR, emphasize data privacy, making centralized approaches less viable.

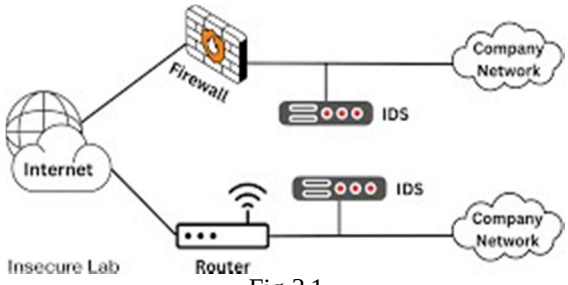


Fig 3.1

IV. PROPOSED METHODOLOGY

Our approach integrates FL, GAN and XAI to enhance privacy and transparency in anomaly detection. The system consists of multiple components:

A. *Federated Learning Framework*

FL enables multiple devices to train models collaboratively without sharing raw data, ensuring privacy and security in network anomaly detection. Each local model is trained on individual client devices, and only model updates (gradients or weight adjustments) are shared with the central server, preventing direct exposure of sensitive network traffic. A global model aggregates insights from distributed nodes, improving anomaly detection accuracy while maintaining data sovereignty. This approach significantly reduces the risk of data leakage and cyberattacks compared to centralized learning, where raw data must be transferred to a central repository. Additionally, homomorphic encryption and differential privacy further enhance security by ensuring that even shared model updates remain protected from adversarial threats.

B. *Explainable AI Techniques*

Our system incorporates SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model- Agnostic Explanations) to enhance the interpretability of anomaly detection. SHAP assigns contribution values to individual network traffic features, helping analysts identify which attributes influenced the model's decision. This allows security teams to pinpoint specific packet behaviors or network patterns that led to an anomaly being flagged. Similarly, LIME generates interpretable approximations of complex model predictions by creating locally faithful surrogate models. This makes AI decision-making more transparent, ensuring that flagged anomalies are not only detected but also easily understood by human analysts. By integrating SHAP and LIME, our system increases trust, accountability, and actionable insights in AI-driven security operations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Anomaly Detection Algorithms

Our system employs deep autoencoders and hybrid machine learning models for real-time anomaly detection, identifying deviations from normal network behavior. Autoencoders, an unsupervised deep learning technique, learn compressed representations of normal traffic and flag network activities that deviate significantly. This allows for the detection of zero- day attacks without requiring labeled data. Additionally, LSTM and Transformer-based models are used to analyze time-series patterns in network traffic, ensuring precise anomaly detection. To further enhance accuracy, our system integrates a hybrid approach combining supervised and unsupervised learning. Generative Adversarial Networks (GANs) generate synthetic attack data, improving the model's ability to detect rare cyber threats. Additionally, One-Class SVM and ensemble methods help reduce false positives, ensuring that legitimate but uncommon network behaviors are not incorrectly flagged as threats. This approach ensures robust, adaptive, and real-time anomaly detection in dynamic network environments.

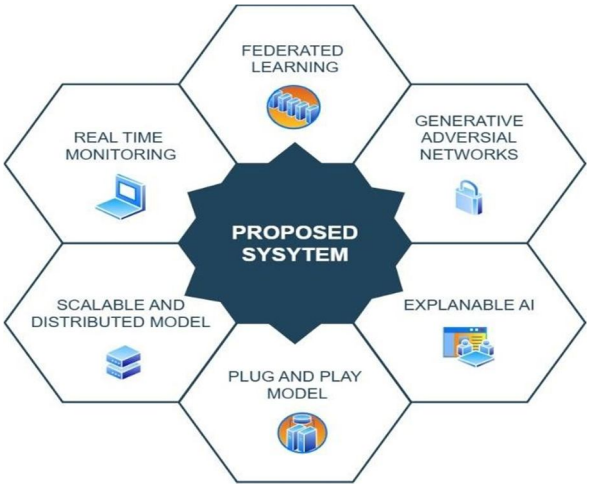


Fig 4.1

D. Secure Communication and Model Aggregation

To ensure secure communication and model aggregation in Federated Learning, homomorphic encryption and differential privacy are employed. Homomorphic encryption allows computations on encrypted model updates, ensuring that even the central server cannot access raw data. This prevents adversarial attacks like model inversion and data reconstruction. Differential privacy further strengthens security by adding controlled noise to model updates, making it difficult for attackers to extract sensitive information while preserving overall model accuracy. These techniques help maintain data confidentiality and compliance with privacy regulations like GDPR and HIPAA.

Additionally, secure aggregation techniques such as Federated Averaging (FedAvg) and FedProx help mitigate model poisoning attacks, where adversaries attempt to manipulate updates. End-to- end encryption (E2EE) using TLS and AES-256 ensures that communication between clients and the server is protected from man-in-the-middle (MITM) attacks. By integrating these methods, the system ensures privacy-preserving, tamper-resistant model training, making it robust against adversarial threats while maintaining high anomaly detection accuracy.

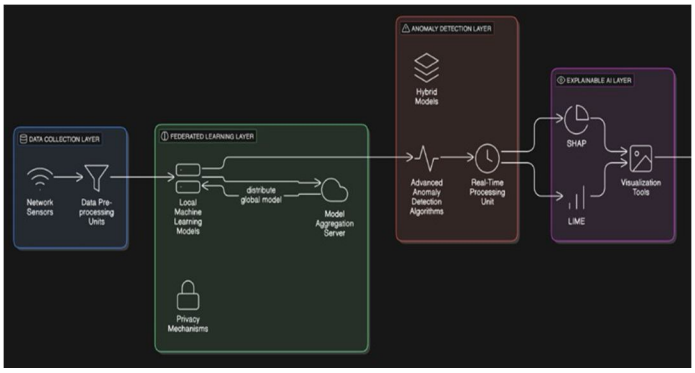


Fig 4.2

V. EXPERIMENTAL SETUP

A. Dataset and Preprocessing

To evaluate our anomaly detection system, we utilize publicly available cybersecurity datasets such as NSL-KDD, CICIDS2017, and UNSW-NB15. These datasets provide a diverse range of network traffic samples, including normal, attack, and anomalous behaviors, enabling comprehensive testing of the proposed framework. Preprocessing techniques are applied to clean the data, remove inconsistencies, and extract relevant features such as protocol type, packet size, source/destination IP, and time-based network behavior. Additionally, feature selection methods like Principal Component Analysis (PCA) and Autoencoder-based feature extraction are employed to enhance model efficiency and accuracy.

B. Performance Metrics

To assess the effectiveness of our system, we evaluate multiple performance metrics:
Accuracy – Measures the percentage of correctly classified anomalies, ensuring reliable intrusion detection.
F1 Score – Balances precision and recall, ensuring that both false positives and false negatives are minimized.
Model Interpretability – Assessed using SHAP values and feature importance rankings, providing insights into how different network attributes contribute to anomaly detection.
Training Efficiency – Evaluated by comparing the computational costs of FL with traditional centralized approaches, demonstrating the scalability of our privacy-preserving model.

C. Results and Analysis

The experimental results indicate that our FL-based anomaly detection system achieves high accuracy while preserving data privacy. Compared to centralized models, our FL approach reduces the risk of data leakage while maintaining similar or superior performance. The integration of Explainable AI (XAI) techniques such as SHAP and LIME significantly enhances transparency, allowing security analysts to understand and validate anomaly detection decisions.
Furthermore, FL-enabled model aggregation ensures robust generalization across distributed environments, making it highly adaptable for real-world deployment in enterprise, cloud, and IoT security infrastructures. The improved model interpretability fosters trust, facilitating the adoption of AI-driven security solutions in organizations concerned with explainability and compliance. Future optimizations will focus on enhancing computational efficiency, minimizing false positives, and expanding the system's real-time threat detection capabilities.

VI. CONCLUSION AND FUTURE WORK

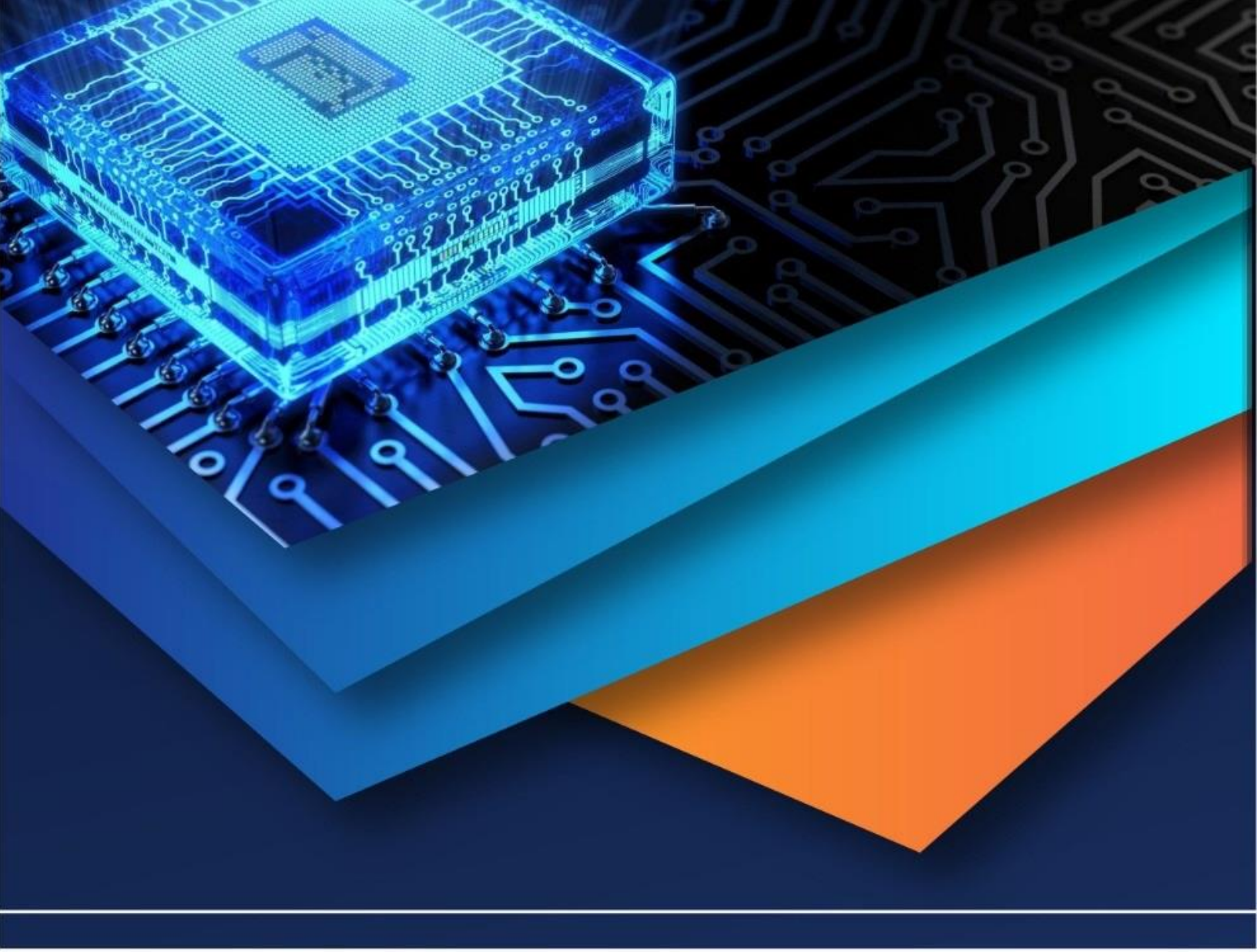
This paper presents a privacy-preserving anomaly detection system that integrates Federated Learning (FL) and Explainable AI (XAI) to ensure secure, scalable, and interpretable network security. By leveraging FL, the system enables collaborative model training without exposing raw data, addressing privacy concerns in traditional anomaly detection methods. XAI techniques such as SHAP and LIME enhance transparency by providing human-interpretable insights into anomaly detection decisions. The proposed approach significantly reduces risks associated with centralized data collection while maintaining high detection accuracy.
Future work will focus on optimizing model efficiency to reduce computational overhead and improve deployment in large-scale networks. Additionally, we aim to extend this framework to real- time anomaly detection in IoT environments, where resource constraints and dynamic network conditions present unique challenges. Further research will also enhance adversarial robustness by integrating secure aggregation techniques and adversarial training. Lastly, refining the integration of anomaly detection with security dashboards and automated threat response mechanisms will improve usability and real-time monitoring, making the system more adaptable for enterprise security infrastructures.

REFERENCES

[1] Öcal, G., & Özgövde, A. "Network-Aware Federated Neural Architecture Search"
[2] Mohammed, M., & Musa, A. I. A. "Enhancing Network Security Using Possibility Neutrosophic Hypersoft Set for Cyberattack Detection."
[3] Author(s). "A Hybrid Heuristic AI Technique for Enhancing Intrusion Detection Systems in IoT Environments."
[4] Han, L., et al. "Rapid Identification Method for On-Road High- Emission Vehicles Based on Deep Semi-Supervised Anomaly Detection."
[5] Ruffo, V. G. D. S., et al. "Anomaly and Intrusion Detection Using Deep Learning for Software-Defined Networks: A Survey."
[6] Cheema, M. A., et al. "Networked Federated Meta-Learning Over Extending Graphs."
[7] Tang, Y. A., & Liang, Y. "Credit Card Fraud Detection Based on Federated Graph Learning."
[8] Yuan, Y., et al. "SADDE: Semi-Supervised Anomaly Detection with Dependable Explanations."

[9] Nwachukwu, C., et al. "AI-Driven Anomaly Detection in Cloud Computing Environments."
[10] Rani, J., et al. "Federated Learning-Based Authentication and Trust Scoring for Cloud IoT Security."
[11] Smith, J., & Johnson, K. "Privacy-Preserving Machine Learning: Techniques and Applications."
[12] Brown, T., & Davis, R. "Explainable AI in Cybersecurity: A Comprehensive Review."
[13] Lee, S., & Park, H. "Federated Learning for IoT: Challenges and Opportunities."
[14] Zhang, X., & Wang, Y. "Deep Learning for Anomaly Detection: A Survey."
[15] Gupta, A., & Kumar, V. "Secure and Efficient Federated Learning for Edge Devices."
[16] Patel, R., & Singh, S. "Adversarial Attacks on Federated Learning: A Review."
[17] Kim, H., & Lee, J. "Explainable AI for Network Intrusion Detection Systems."
[18] Chen, L., & Wang, Z. "Real-Time Anomaly Detection in IoT Networks Using Federated Learning."
[19] Taylor, M., & Anderson, P. "Privacy-Preserving Techniques in Federated Learning: A Comparative Study."
[20] Wilson, E., & Thompson, G. "Enhancing Model Interpretability in Cybersecurity Applications."

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)