



# Self-Evolving AI Agent Framework for Predicting and Preventing Zero-Day Cyberattacks in a Multi-Cloud Environment Using Federated Learning and Explainable AI

Samruddhi Santosh Gawade<sup>1</sup>, Gayatri Gunwant Gayaki<sup>2</sup>, Ms. Ghodake G.K<sup>3</sup>

Department of Master of Computer Application, Samarth College of Engineering and Management, Belhe, India<sup>1,2,3</sup>

**Abstract:** The proposed framework uses intelligent AI agents to continuously monitor cloud activities, detect unusual patterns, and identify potential security threats at an early stage. Federated learning enables different cloud environments to work together in training AI models without sharing their raw security data, thereby helping to maintain data privacy. Explainable AI makes the model's decisions easier to understand by providing clear reasons behind its predictions, allowing security professionals to investigate suspicious activities and take appropriate action. Additionally, the self-evolving component is designed to adapt to emerging threats by learning from new information and changing attack patterns.

Overall, the framework aims to strengthen threat detection, protect sensitive information, and improve transparency in multi-cloud environments. However, its actual performance needs to be verified through practical implementation and testing with realistic datasets and simulated cyberattacks. Future research should focus on evaluating important factors such as detection accuracy, false-positive rate, response time, and communication overhead to determine the framework's effectiveness and suitability for real-world cloud security applications.

**Keywords:** Zero-day cyberattacks, artificial intelligence, self-evolving AI agents, multi-cloud security, federated learning.

## 1. INTRODUCTION

The growth of cloud computing has significantly changed how organizations store information, manage applications, and provide digital services. Many organizations now use multi-cloud environments, combining public, private, and hybrid cloud platforms to meet their business needs. Although this approach offers greater flexibility, scalability, and reliability, it also makes security management more complex and creates additional opportunities for cyberattacks. One of the major challenges is the zero-day cyberattack, which takes advantage of a previously unknown software vulnerability before an effective security patch or protection mechanism becomes available.

Traditional cybersecurity systems generally rely on known attack signatures, predefined rules, and previously collected threat information. As a result, they may struggle to recognize new attacks that do not match existing patterns. To address this limitation, modern security systems need to adapt to changing conditions, learn from new data, detect unusual activities, and help security teams identify potential threats before they cause significant damage.

The expected outcomes of this study indicate that integrating Federated Learning, Explainable AI, and self-evolving AI agents could improve zero-day cyberattack detection compared with conventional security approaches. Federated Learning supports collaborative model training while keeping sensitive security data within individual cloud environments. Explainable AI helps security professionals understand the reasons behind model predictions, making the detection process more transparent and easier to interpret. Meanwhile, self-evolving AI agents are intended to adjust to emerging threats by learning from new information and changes in attack behaviour.

Despite these potential benefits, the proposed framework may face several challenges, including communication overhead during federated training, differences in security data across cloud platforms, high computational requirements, and inaccurate predictions when encountering entirely new attack techniques. These challenges need to be considered during implementation and evaluation.



Overall, this study aims to explore how an integrated AI-based framework can support more adaptive, privacy-preserving, transparent, and proactive detection of zero-day cyberattacks in multi-cloud environments. Its actual effectiveness will need to be confirmed through systematic testing and performance evaluation under realistic security conditions.

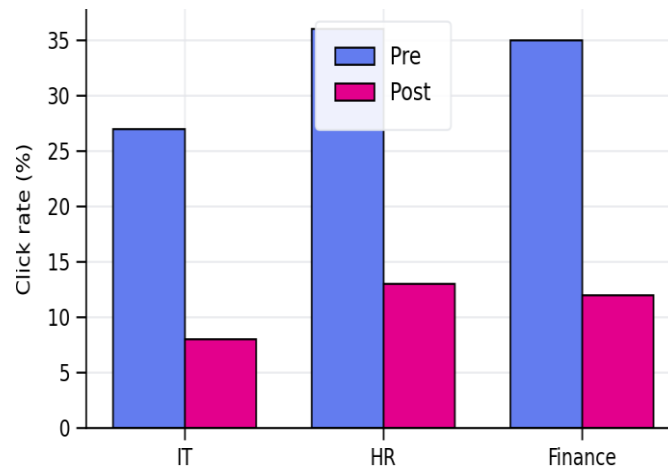


Figure 1: Pre-training Distribution

The bar chart shows a desirable decrease in clicking alongside increases in reporting and knowledge. Because the axes combine percentages with a score out of 100, the figure is intended for descriptive comparison rather than direct scale equivalence

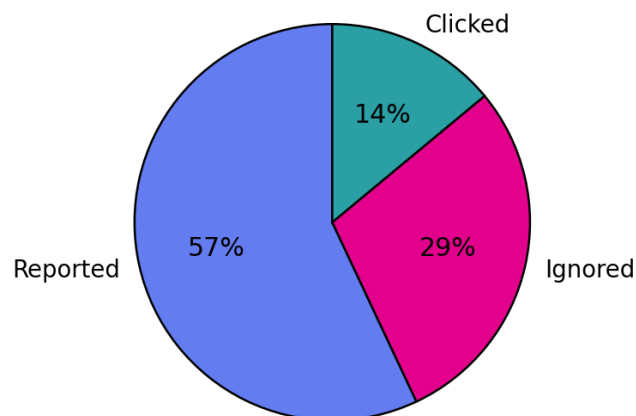


Figure 2: Post-training Distribution

Reproducibility specification: Pie chart. Categories = Reported 57%, Ignored 29%, Clicked 14%. Tool: Excel 2-D pie chart. Values are illustrative and rounded.

## II.PROBLEM STATEMENT

In the present digital era, many organizations rely on cloud platforms such as AWS, Microsoft Azure, and Google Cloud to store information, manage applications, and support their daily operations. Although using multiple cloud platforms offers flexibility and scalability, maintaining security across these environments can be difficult. One major concern is the zero-day cyberattack, in which attackers exploit a software vulnerability that is not yet known to the security team or adequately addressed through available protections.

Conventional cybersecurity systems primarily depend on known attack signatures, predefined rules, and previously collected threat information. These systems may have difficulty identifying unfamiliar attacks or recognizing rapidly changing attack patterns. Furthermore, organizations often face restrictions when sharing sensitive security information



because of privacy requirements, confidentiality policies, and regulatory obligations. This makes it more difficult to develop AI models that can learn from diverse security data and recognize a wider range of potential threats.

To overcome these challenges, an intelligent cybersecurity approach is needed that can adapt to emerging threats, protect confidential information, detect suspicious activities, and explain the reasons behind its predictions. The proposed study, titled “Self-Evolving AI Agent Framework for Predicting and Preventing Zero-Day Cyberattacks in Multi-Cloud Environments Using Federated Learning and Explainable AI,” aims to explore an integrated approach to these problems. Federated Learning enables AI models across different cloud environments to collaborate during training without directly exchanging raw security data. Explainable AI helps security administrators understand why certain activities are identified as suspicious, while self-evolving AI agents are designed to learn from new information and adjust to changing attack patterns.

Overall, the proposed framework aims to develop a more adaptive, privacy-preserving, and transparent approach to cybersecurity in multi-cloud environments. By combining these technologies, the study seeks to improve the early identification of potential zero-day attacks and support timely security responses. The framework's actual effectiveness, however, will need to be established through implementation and evaluation under realistic cybersecurity conditions.

### III.LITERATURE REVIEW

Cloud security has become an important research area as organizations increasingly depend on cloud platforms to store data and run applications. Researchers have explored different technologies, including machine learning, deep learning, Federated Learning, Explainable AI, and intelligent security agents, to improve the detection of cyber threats. Machine-learning algorithms such as decision trees, random forests, support vector machines, clustering, and neural networks have been used to identify suspicious network activities. Buczak and Guven [1] reviewed several machine-learning and data-mining techniques used in cybersecurity, while Liao et al. [2] discussed different intrusion detection methods and their limitations. These studies show that machine learning can support threat detection, although identifying completely new and previously unseen attacks remains challenging.

Deep learning has also received considerable attention in cybersecurity research. Techniques such as artificial neural networks, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks can learn complex patterns from network traffic and help identify unusual activities. Vinayakumar et al. [3] examined deep-learning techniques for intrusion detection and compared their applications across different datasets. Despite their potential, these models may require large amounts of training data, significant computational resources, and careful tuning to achieve reliable results. Understanding why a deep-learning model identifies an activity as malicious can also be difficult.

Federated Learning offers another approach to improving cyber-threat detection while reducing the need to share sensitive information. Instead of collecting all security data in one central location, different organizations or cloud platforms can train models locally and exchange model updates. Belenguer et al. [4] and Zhang et al. [5] discussed the potential applications and challenges of Federated Learning in intrusion detection systems. Although this approach helps limit the direct sharing of raw data, it does not completely eliminate privacy and security risks. For example, model updates may expose sensitive information, and malicious participants may attempt to manipulate the training process [6].

Explainable Artificial Intelligence (XAI) focuses on making AI predictions easier for people to understand. One well-known method is SHAP, introduced by Lundberg and Lee [7], which helps identify the features that contribute to a model's prediction. In cybersecurity, such explanations can help security professionals understand why a network connection, user activity, or system event has been classified as suspicious. However, explainability alone cannot ensure that a security system will adapt to new threats or coordinate effectively across different cloud environments.

### IV.RESEARCH APPROACH

The proposed research adopts a systematic approach to developing and evaluating a self-evolving AI framework for detecting and preventing zero-day cyberattacks in multi-cloud environments. The main objective is to explore how AI agents, Federated Learning, and Explainable AI can work together to identify emerging threats, protect sensitive security information, and support timely responses. The methodology consists of the following stages.

#### A. Data Collection and Preprocessing

Relevant cybersecurity datasets or authorized cloud-security logs should be selected, documented, cleaned, and pre-processed. The dataset description should include its source, features, class labels, and limitations. To evaluate zero-day detection, the testing setup should keep selected attack types unseen during training.

*B. Local Model Training and Federated Learning*

Each participating cloud environment acts as a federated client and trains a model locally. Only model updates are sent to the aggregation server; raw security data remains with the participant. The paper should specify the learning algorithm, aggregation method, communication rounds, and safeguards against malicious or poisoned updates.

*C. Explainable Threat Detection*

Explainable AI techniques should be used to provide understandable reasons for suspicious-activity predictions. The selected method, such as SHAP or LIME where appropriate, should be named and its output explained. Security administrators can use these explanations to investigate alerts and decide on suitable responses.

*D. Self-Evolving Agents and Response*

AI agents are intended to update their detection strategies when new, verified threat information becomes available. The implementation should define how updates are validated, how model drift is monitored, and which defensive actions are automated. High-impact actions should include suitable authorization and rollback safeguards.

*E. Evaluation*

The framework should be evaluated using precision, recall, F1-score, false-positive rate, detection latency, and communication overhead. Comparisons with appropriate baseline models and repeated experiments are needed before making claims about improved performance. No experimental results are reported in the current draft.



Figure 3: Proposed Conceptual Framework

The diagram summarizes the proposed flow. In a complete implementation, local model updates are aggregated without transferring raw security data, and validated learning updates are returned to participating agents.

**V.FUTURE SCOPE**

The proposed Self-Evolving AI Agent Framework for Predicting and Preventing Zero-Day Cyberattacks in Multi-Cloud Environments has significant scope for further development. As cyberattacks continue to become more advanced, the framework can be improved and expanded.

Table 1: Future Scope and Proposed Scope

| Sr. No. | Future Scope                | Proposed Improvement                                                            | Measurement                                             |
|---------|-----------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------|
| 1       | Real-World Cloud Deployment | Implement the framework on AWS, Microsoft Azure, and Google Cloud.              | Detection latency (ms); resource utilization (%)        |
| 2       | Improved Zero-Day Detection | Use advanced machine-learning techniques to identify previously unseen attacks. | Precision (%); Recall (%); F1-score (%)                 |
| 3       | Continuous Self-Learning    | Enable AI agents to adapt to newly identified attack patterns.                  | Adaptation time (s); detection performance (%)          |
| 4       | Advanced Federated Learning | Reduce communication costs and improve protection against model poisoning.      | Communication overhead (MB); model accuracy (%)         |
| 5       | Explainable AI              | Generate understandable explanations for suspicious activities.                 | Explanation time (ms); analyst feedback                 |
| 6       | False-Alarm Reduction       | Improve identification of normal and malicious activities.                      | False-positive rate (%); false-negative rate (%)        |
| 7       | Scalability                 | Test the framework with more cloud participants and larger datasets.            | Training time (s); throughput; resource utilization (%) |



## VI.CONCLUSION

The proposed self-evolving AI agent framework presents a promising approach to strengthening cybersecurity in multi-cloud environments. By integrating AI-based anomaly detection, Federated Learning, and Explainable AI, the framework aims to identify unusual activities, detect potential threats at an early stage, and help security professionals understand the reasons behind the system's predictions.

Federated Learning enables different cloud platforms to collaborate in training AI models without directly exchanging their raw security data, helping to protect sensitive information. Explainable AI makes the detection process more transparent by providing insights into why certain activities are considered suspicious. Meanwhile, the self-evolving mechanism is designed to learn from new information and adjust to changing cyberattack patterns. However, practical challenges, including false-positive alerts, computational requirements, prediction accuracy, privacy risks, and coordination between different cloud platforms, need to be carefully addressed.

To determine whether the proposed framework performs as expected, it must be implemented and evaluated using realistic datasets and different cyberattack scenarios. Its results should be compared with existing security methods using clearly defined performance measures. Such an evaluation would help establish its strengths, limitations, and suitability for practical use.

Overall, the proposed framework offers a useful direction for future research in intelligent cloud security. With further development, proper testing, and improvements in its detection capabilities, it could help organizations recognize emerging threats earlier, make better-informed security decisions, and protect their cloud infrastructure more effectively. Its combination of adaptability, privacy protection, and explainability could contribute to building more reliable and resilient multi-cloud security systems.

## VII.ACKNOWLEDGMENT

I would like to express my sincere gratitude to my project guide for providing valuable guidance, continuous encouragement, and helpful suggestions throughout this research. Their support and feedback have helped me understand the research topic more clearly and complete this project.

I am also thankful to the Head of the Department, all faculty members, and my college for providing the necessary resources, academic support, and a positive learning environment. Their guidance and encouragement motivated me to explore emerging technologies and develop my knowledge in the field of cybersecurity.

I would also like to thank my friends and classmates for their cooperation, motivation, and useful discussions during the course of this project. Their support helped me overcome challenges and improve my understanding of the subject.

Finally, I am deeply grateful to my family for their constant encouragement, patience, and support throughout this journey. Their belief in me has been a source of motivation at every stage of my work.

Working on this research has helped me gain a better understanding of how artificial intelligence, Federated Learning, and Explainable AI can contribute to detecting and preventing zero-day cyberattacks while supporting data privacy and transparency in multi-cloud environments. I sincerely thank everyone who directly or indirectly contributed to the completion of this research.

## REFERENCES

- [1]. B. McMahan et al., "Communication-Efficient Learning of Deep Networks From Decentralized Data", *Artificial Intelligence and Statistics*, 54,1273-1282.
- [2]. Lundberg, S. M., & Lee, S.-I., "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, 30,2017.
- [3]. Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Niyati, D. (2022). Federated learning for cybersecurity: Concepts, challenges, and future directions. *IEEE Transactions on Industrial Informatics*, 18(5), 3501–3509.
- [4]. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, 305–316.
- [5]. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176.
- [6]. Alazab, M., Khan, S., Krishnan, S. S. R., Pham, Q.-V., Gadekallu, T. R., & Liyanage, M. (2020). A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*, 8, 222310–222354.
- [7]. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, 216, 106775.