QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

FEDERATED COGNITIVE DIGITAL TWINS FOR AUTONOMOUS AI AGENT INSIDER THREAT PREDICTION AND CONTAINMENT

Muhammad Danish
Email: contactdanish4@gmail.com
Department: Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, 54000, Pakistan
Mobeen Mazhar
Email: mobeenmazharmk421@gmail.com
Department: Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, 54000, Pakistan
Arslan Ali
Email: contactarslan4@gmail.com
Department: Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, 54000, Pakistan
Maryam Siddique
Email: maryamsiddique014@gmail.com
Department: Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, 54000, Pakistan
Mah Noor
Email: mahnoorsandhu32@gmail.com
Department: Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, 54000, Pakistan

Abstract

As users have legitimate access and business environments become more autonomous with the use of AI agents, insider threats continue to be a significant problem in the cybersecurity landscape. The current detection methods typically are reactive, require centralized data collection, and have privacy protection and scalability problems. In order to solve these problems, this paper introduces a Federated Cognitive Digital Twin (FCDT) framework for the autonomous prediction and containment of insider threats using AI agents. The proposed framework leverages Federated Learning, Cognitive Digital Twins, Long Short-Term Memory (LSTM) for behavioral prediction, and a Threat Risk Scoring mechanism to facilitate privacy-preserving collaborative learning and proactive threat detection in distributed environments. On one hand Federated Learning enables multiple organizations to train a global model without exposing sensitive data, and on the other hand, Cognitive Digital Twins continuously model behaviors and detect possible insider threats. To test the proposed framework, the UNSW-NB15 dataset was used, and the experiments were carried out in Google Colab. The model showed promising classification performance, with an accuracy of 84.21%, precision of 90.49%, recall of 79.70%, F1 score of 84.75% and ROC-AUC of 93.82%, indicating a good discriminating ability. It is observed from the experimental results that the proposed framework is effective in improving the insider threat prediction with data privacy and scalable cyber security intelligence. Overall, the proposed FCDT framework offers a promising approach to ensure the security of modern distributed enterprise systems in the face of dynamic insider threats and autonomous AI-agent risks.

Keywords: Federated Learning, Cognitive Digital Twin, Insider Threat Detection, Autonomous AI Agents, Cybersecurity, Long Short-Term Memory (LSTM), Threat Prediction, Privacy-Preserving Learning, Threat Risk Scoring, Distributed Artificial Intelligence.

1. Introduction

1.1 Insider Threat Landscape

The world of cloud computing, artificial intelligence (AI), distributed systems and autonomous digital services has rapidly changed how organizations operate. More and more enterprises are

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

putting technologies into practice to make them more efficient, automate decisions and help with their huge digital transformation project. These developments are very positive, but they also present complex cyber security issues, which are not easily resolved with conventional cyber security measures. Insider threats are one of the most serious and challenging cyber threats, as most of the malicious activity is likely to come from organizations or entities with authorized access to the resources of the organization itself [1]. Insider threats are those threats executed by trusted users, contractors, third party partners, compromised accounts or intelligent software entities within the organization's environment. Unlike external attackers, whose activities must go under security screens, insiders' actions go under security screens, and they have access to sensitive data and services that are privileged.

These can lead to data leaks, intellectual property theft, financial damage, business disruption, and damage to the brand. The complexity of enterprise systems has also been increasing, making identification of an insider threat even more difficult before the damage is done [2]. Most traditional insider threat detection methods are rule-based systems, signature matching and statistical anomaly detection. While these methods are successful in recognizing known attack patterns, they have a tendency to fail if attacks are more sophisticated and evolving. Modern insider threats often display very subtle deviations over time, resulting in a gradual, unnoticed change in behavior. Therefore, traditional security systems tend to trigger a lot of false alerts and offer less flexibility when adapting to changing enterprise scenarios [12].

There are recent advances in machine learning where more intelligent methods have been developed for behavioral analysis and for anomaly detection. Recurrent neural networks (RNNs), long short-term memory (LSTM) networks, autoencoders, and transformer-based models have proven effective in capturing intricate behavioral patterns from extensive sets of activity data. These methods enhance the ability of cybersecurity solutions to detect abnormal behavior and new attack patterns [13]. But there are many solutions out there that are reactive, sending you an alert once you've detected something suspicious, not predicting one before it happens.

1.2 Privacy Preserving Cybersecurity through Federated Learning

One of the biggest drawbacks of many machine learning cybersecurity solutions is that they require centralized data collection and processing. In traditional methods, information from various sources is gathered and fed into a central repository for modeling and analysis to be done from there.

These centralized architectures come with privacy issues, a tougher compliance problem and easy targets for cyberattacks. These restrictions are even more critical in distributed enterprise environments where data is created in various branches, departments, and cloud infrastructures spread out across the globe. To address these challenges, a novel approach called Federated Learning (FL) has been proposed, allowing collaborative machine learning while avoiding the sharing of the raw data [6]. In federated learning, each participant trains a model on their local data and only sends model parameters or updates to a server. This way, data can be kept private, while organizations can benefit from collective intelligence and better learning performance [5].

In cybersecurity applications, federated learning supports distributed anomaly detection, collaborative threat intelligence, and privacy-preserving security analytics [4]. While federated learning has these benefits and is an exciting approach, it is still not enough on its own to solve complex insider threat scenarios because it does not provide the contextual reasoning and the modelling of behavior.

1.3 Cognitive Digital Twins for Predictive Security

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

Digital Twin technology has attracted a lot of attention as a powerful tool to represent physical or digital entities within a virtual environment [7]. A digital twin dynamically keeps track of actual operational data, allowing for tracking, simulating and analysing the behaviour of the system. In the context of cybersecurity applications, digital twins can represent users, devices, applications, and network activities to create a better situational awareness and enable anomaly detection. Digital twins provide an up-to-date virtual representation of operational systems, which can help to inform security decisions [8]. The idea of Cognitive Digital Twins is an extension of the concept of the Digital Twins, with the addition of Artificial Intelligence, Adaptive Learning, Contextual Awareness, and Predictive Analytics [9].

The cognitive digital twins can learn behavioral patterns, understand the operational context, and predict future states from previous observations. These attributes make them especially well suited to a proactive cybersecurity landscape, where they need to be able to detect malicious behaviors early. Most current cognitive digital twin applications, however, are confined to individual contexts, and seldom take advantage of distributed intelligence-sharing frameworks to boost prediction precision across diverse organizations [10].

1.4 Security Challenges of Autonomous AI Agents

With the rise of autonomous AI agents, cybersecurity risk management now takes on a new dimension. An autonomous agent is able to execute tasks, utilize organizational resources, interact with enterprise systems and make operational decisions without significant human intervention [18] [19]. These features improve efficiency and automation, but also introduce new insider threat vectors that are not covered by traditional security solutions.

AI agents can be compromised, manipulated, or misaligned, which can lead to the misuse of their privileges, breaches of security policies, leaks of sensitive information, or the execution of harmful actions that can have a negative impact on organizational operations. Autonomous agents can act continuously, and change their behavior based on environmental conditions, which makes it more difficult to monitor and predict what they will do than for a human user [18]. With the growing use of AI-powered systems in critical business processes, there is a rising demand for security measures that can continuously monitor the AI-agent's activity and take automated countermeasures in the event of suspicious activity.

Table 1. Key Technologies and Their Roles in Insider Threat Prediction

Technology	Purpose	Strength	Limitation
Insider Threat Detection	Identify malicious internal activities	Detects suspicious behavior	Mostly reactive
Federated Learning	Collaborative model training without sharing raw data	Preserves privacy	Limited contextual reasoning
Digital Twins	Virtual representation of entities and systems	Continuous monitoring	Often deployed in isolation
Cognitive Digital Twins	Intelligent behavioral modeling and prediction	Supports predictive analysis	Limited distributed collaboration
Autonomous AI Agents	Automated decision-making and task execution	Enhances efficiency	Introduces new insider risks

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

1.5 Research Gap and Proposed Contributions

Although these areas of research have come a long way in insider threat detection, federated learning, digital twins, and AI-driven cybersecurity, the vast majority of the research in these areas has been done in isolation [18]. Most existing solutions are concerned with solving a particular component of security, including behavioral anomaly detection, learning without compromising privacy, system simulation or automated decision-making. Lack of any approach that can support privacy-preserving collaboration, intelligent behavioral modeling, predictive threat forecasting and autonomous response mechanisms at the same time [9], [17]. This means that current cybersecurity solutions are unable to effectively counter the challenges posed by the growing complexity of insider threats in today's enterprise environment. In order to overcome these challenges, a Federated Cognitive Digital Twin (FCDT) framework is proposed to tackle the Autonomous AI-Agent Insider Threat Prediction and Containment (AIRPIC) problem.

The proposed framework integrates the cognitive digital twin modeling with federated learning to facilitate the shared intelligence while respecting privacy across distributed environments. The framework simulates the actions of both human users and autonomous AI agents, forecasts probable malicious operations, and triggers effective containment measures prior to the occurrence of security incidents. The proposed approach combines distributed learning, behavioral reasoning, predictive analytics and autonomous response to offer a holistic solution for next-generation insider threat management.

The main findings of this research are summarized below:

1. Federated Cognitive Digital Twin for Distributed Insider Threat Prediction and Containment.
2. One consistent behavioral modeling for both human and autonomous AI agents.
3. Privacy-preserving Federated Learning (FL) mechanism for collaborative cybersecurity intelligence sharing.
4. Predictive threat analysis engine that can detect new insider threats prior to incident.
5. An autonomous containment strategy for dynamically responding to suspicious activities.

An integrated framework that integrates federated learning with cognitive digital twins and autonomous AI-agent security mechanisms. The rest of this paper is structured as follows: In Section 2, the existing literature on insider threat detection, federated learning, digital twins and autonomous cybersecurity systems are reviewed. The proposed methodology and the proposed system architecture are presented in section 3. Results and implications of the proposed approach are discussed in Section 4 with Future research directions outlined in Section 5.

2. Literature Review

As the world of cloud computing, distributed architectures, AI and intelligent automation has grown, the cybersecurity landscape has changed significantly. With increased interdependence of digital infrastructures, organizations are facing greater difficulty in protecting critical resources and sensitive information. Insider threats are among the most challenging cybersecurity threats to handle due to the fact that they have the ability to take place within trusted environments, leveraging legitimate identities and access privileges. An insider incident is often not detected by perimeter security defenses and is sometimes not discovered for quite a long time. Insider threats have been solved by many ideas behavioural analytics, Machine learning-based detection system, Federated learning framework, Digital twins, and AI-based monitoring system, etc.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

The methods presented have shown some success within certain application fields but most of the previously studied methods are limited to a single specific aspect of cybersecurity. This section conducts a review of the previous studies conducted in the field of insider threat detection, machine learning security analytics, federated learning, digital twin technology, cognitive digital twin, and autonomous AI-agent security. The proposed Federated Cognitive Digital Twin (FCDT) framework is derived from the strengths, limitations and research gaps revealed in the existing studies.

2.1 Insider Threat Detection

Insider threat detection has become a hot research topic in the field of cybersecurity, as any malicious action by an insider can have grave repercussions. Employees, contractors and business partners, compromised accounts, or other legitimate users of organization systems can be a source of insider threats. Insiders have a clear boundary of trust and work inside the fence, for this reason difficult to distinguish malicious activities from normal operations. Efforts for initial research were mainly on the rule-based detection systems and signature-based approaches. These tended to be based on a known attack signature, expert knowledge, and a pre-defined security policy to detect suspicious behaviour.

This worked well to identify the attack patterns they had already seen, but not for identifying new patterns or unusual behaviours that are not covered by the rules. [1] and [12] were unable to detect new attacks or deviations in behaviour from an attack pattern that do not conform to existing rules. This made it challenging for organizations to adjust to new techniques and fluctuating operational situations. To address these constraints, behavioral analytics methods that analyze how users engage with systems, their access to information, their communication histories, and other interactions have been introduced. Rather than using pre-defined rules, behavioral approaches create user activity profiles based on the normal user behavior and look for anomalies that could be a sign of malicious activity.

Deep learning models can be effectively trained in an unsupervised manner to detect unusual or abnormal behavioral patterns without the need for a large amount of labeled data [2]. They concluded that neural networks could be used to identify advanced insider operations that are hard to detect using traditional methods. Contextual Information: The importance of when analysing Insider behaviour [3]. They showed that the nature of the role, access rights, context, and temporal shifts in behavior affect insider threat prediction and automated mitigation.

2.2 Machine Learning and Deep Learning in Cybersecurity

It is no surprise that machine learning and deep learning algorithms are now essential parts of today's cybersecurity solutions. These approaches allow for the automatic extraction of features, pattern recognition and anomaly detection of large and complex datasets that may be hard to analyze with traditional methods. Data in a cyber environment is increasing and intelligent learning algorithms offer scalable solutions for detecting suspicious activities and identifying potential threats. There are several studies that have proven the effectiveness of machine learning techniques for cybersecurity applications. In large-scale environments, the deep learning architectures perform much better than the traditional statistical models in dealing with the anomaly detection problem [1].

Deep learning models can detect fine-grained behavioral patterns linked to malicious activities because they are able to learn complex data representations automatically. The deep learning based insider threat detection models based on enterprise system logs and user activity records [13]. The outcome of their research showed that they achieved better classification accuracy and better ability to recognize abnormal user behavior. Sequential and temporal cyber security data is particularly well suited for deep neural networks like recurrent neural networks (RNNs),

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

long short-term memory (LSTM) networks, convolutional neural networks (CNNs) and autoencoders. While machine learning tools offer significant benefits to cybersecurity, there are several significant obstacles to their use.

To begin with, several models demand vast amounts of centralized training information, raising privacy concerns and regulatory compliance problems. Second, training and maintenance of these models can require significant computational power, especially in the case of continually changing data sets. Third, many deep learning models are used as black-box systems which makes them less interpretable, and therefore less trustworthy to analysts for automated decision making. Yet another constraint is the lack of most machine learning frameworks that are heavily concentrated on classification and anomaly detection tasks instead of proactive behavioural forecasting. For this reason, they can typically only notice when something is amiss once it happens. This restriction inspires the creation of future-oriented and context-aware security frameworks that can predict threats before security incidents can occur.

2.3 Federated Learning for Privacy-Preserving Cybersecurity

With worries about data privacy and regulatory requirements, decentralized machine learning methods are gaining traction. Collaborative model training without sharing raw data has given rise to a new paradigm called Federated Learning (FL). Federated learning enables training of models on the local node and only sharing of model parameters, thereby offering a good balance between knowledge sharing and privacy protection. The Federated Averaging (FedAvg) algorithm has laid the groundwork for the modern federated learning systems [5].

They showed that this approach allows them to collaboratively build high-quality machine learning models without exposing sensitive information to the cloud. It was an innovation that had a tremendous impact on later studies in privacy-preserving machine learning. On the basis of this, Federated learning methodologies, challenges, and future research directions [6] were developed. They pointed out several key issues: communication efficiency, model convergence, system heterogeneity, and privacy protection. As federated learning continues to grow and branch out into other application scenarios, these are still active research areas. Some of the applications of federated learning in cybersecurity include distributed intrusion detection, malware analysis, anomaly detection, and collaborative threat intelligence systems. Federated learning can add an extra layer of security analytics for many organizations while maintaining the privacy of their sensitive operational information [4].

These are especially useful when organizations are hesitant or unable to provide raw cybersecurity data. While federated learning offers many advantages, it is not a panacea for all cybersecurity needs. In most federated frameworks the emphasis is mainly on collaborative model optimization, and they lack contextual reasoning, behavioral simulation, and prediction decision-making capabilities. Moreover, federated learning settings are still at risk of model poisoning attacks, adversarial manipulation, and information leakage due to model updates. The constraints indicate that federated learning needs to be combined with other technologies to enable holistic insider threat management.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

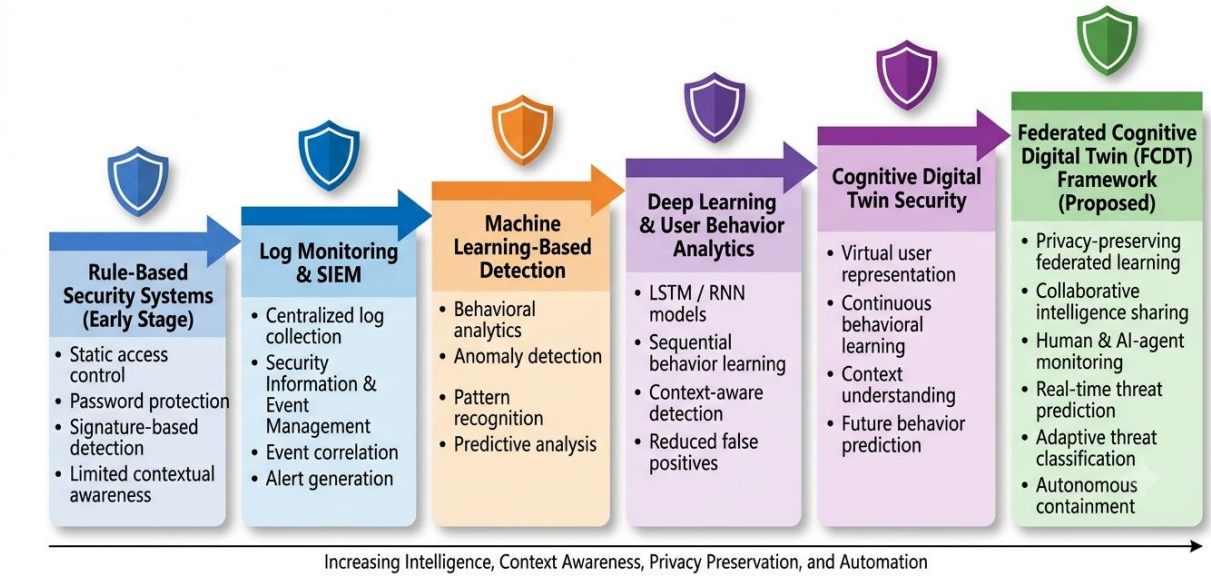

Figure 1: Evolution of Insider Threat Detection Technologies

Figure 1 shows the evolution of insider threat detection technologies, from the classic rule-based methods to the machine learning-based analytics and privacy-preserving federated learning frameworks.

2.4 Digital Twin Technology in Cybersecurity

Digital Twin technology has been a rising star as a way to make virtual representations of physical and digital systems. The digital twin is constantly fed data from its physical twin to provide real-time monitoring, simulation, diagnostics and predictive analysis. Digital twins can give organisations greater insight into system behaviour and performance through continuous synchronization. A comprehensive survey of the digital twin applications in cyber-physical systems, focussing on predictive analytics, system optimisation and operational management [7]. Their work also showcased the benefits of digital twins in enhancing situational awareness and enabling proactive decision-making in complex scenarios. Likewise, the enabling technologies and implementation issues for digital twin deployment such as scalability, data integration, interoperability, and data synchronization needs [8].

In cybersecurity, digital twins have been used to simulate network infrastructure, user activities, software systems, and network operation of an enterprise. Digital twins provide security analysts with simulated environments to test how a system will respond, attack it with various scenarios, and discover weaknesses without having to exploit them first. A Digital Twin based Insider Threat Detection framework using digital twin modelling and self-attention based deep learning architectures [9] was proposed. Their results showed that continuously modeling user activities using a digital twin has the potential to greatly enhance the performance of anomaly detection and behavioral analysis. While some cybersecurity solutions already exist for digital twins they have some significant limitations. Most focus on infrastructure level monitoring and not extensive behavioral intelligence. In addition, most of the implementations are standalone and cannot facilitate the sharing of intelligence or distributed learning between different environments. These weaknesses limit their usefulness in large and highly distributed cyber security environments.

2.5 Cognitive Digital Twins

Cognitive Digital Twins (CDTs) are the next generation of digital twin technology. Cognitive Digital Twins can incorporate AI, adaptive learning, contextual information, and reasoning to

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

go beyond simple monitoring and provide intelligent decision support. Unlike traditional digital twins that simply simulate a system's status, CDTs can understand behavior, learn from past observation and predict future results. The value of cognitive digital twins for using artificial intelligence to understand and reason about contextual information to enhance decision making processes [10]. Their research showed the possibility of intelligent virtual models to continuously improve knowledge about behavior and adjust to varying environmental circumstances. These features make cognitive digital twins especially promising for cybersecurity use cases where patterns of user behavior and threats change over time. For dynamic enterprises, cognitive digital twins can interpret behavioral history, context and interactions among systems to detect potential security anomalies before they become apparent. This type of prediction is an important advantage over traditional monitoring methods which only consider the state of the system at the present time. Nevertheless, cognitive digital twins are a relatively new field of research, and have not yet been widely used in an operational cybersecurity system. Current solutions tend to be platform and technology dependent and don't generally include distributed learning frameworks that can utilize intelligence from multiple organisations. Additionally, existing research on CDT tends to be limited to forecasting and tracking, and not threat response and AI-agent supervision. The identified gaps provide potential areas for additional innovation by combining cognitive digital twins with federated learning and autonomous security mechanisms.

2.6 Integration of Federated Learning and Digital Twins

The integration of Federated Learning and Digital Twins.The combination of Federated Learning and Digital Twins. Recently, the integration of Federated Learning (FL) with Digital Twin (DT) has become an exciting frontier in the field of cybersecurity.The fusion of Federated Learning (FL) and Digital Twin (DT) is a recent and intriguing research area for the next generation of cybersecurity systems. Federated learning facilitates collaboration between distributed entities without sacrificing privacy, and digital twins create virtual representations of users, devices, and operational environments that can be updated in real-time. These technologies can be integrated to enable intelligent cybersecurity solutions that can integrate distributed knowledge sharing and ongoing behavioural monitoring. A few researchers have investigated the benefits of digital twin on federated learning architectures. These technologies can be paired to allow enterprises to work together to enhance security models without risking their sensitive operational data. This is particularly significant in industries where privacy laws and data protection protocols are stringent, preventing direct data transfers between institutions. The FL-TWIN framework, a combination of federated learning and digital twin for intrusion detection use cases [17]. They showed that data can be preserved while collaborative model training can enhance the detection performance. Experimental findings indicated that increased resilience to adverse threats and increased adaptability to dynamic cyber environments. Although these promising results are encouraging, existing FL-DT integration mechanisms are still narrow in scope.

Current solutions are more about network intrusion detection and infrastructure monitoring and not insider threat prediction. Besides, they tend to be less cognitively sophisticated in their reasoning, have no mechanisms for predicting behavior and no containment strategies of their own. The combination of federated learning and digital twins is a great step forward, but further research is needed to create a complete framework and tackle the challenges associated with current insider threats in the context of human users and intelligent software.

2.7 Autonomous AI Agents and Emerging Security Risks

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

As more and more autonomous AI agents move into enterprise settings, new cybersecurity issues have emerged. Today's AI agents can perform tasks, communicate with organizational systems, analyze data and make operational decisions without the need for human involvement. These features offer significant improvements in terms of automation, efficiency, and scalability. But they also pose new security threats that conventional cybersecurity solutions aren't fully able to defend against.

In recent years, the development of large language models and systems incorporating autonomous agents has shown that AI systems can tackle complex reasoning, planning, and decision-making tasks.The recent progress in the development of large language models and agent-based autonomous systems has illustrated the capacity of AI systems to handle intricate reasoning, planning, and decision-making. The increasing capabilities of AI-driven agents and their ability to execute autonomously in a variety of application areas [19].

This is significant technological progress, but it also raises the spectre of inadvertent or harmful behaviour in trusted organisational spaces. AI agents that are unchecked, misaligned, misconfigured or breached can end up posing a threat to security. The possibilities of risk are unauthorized data access, privilege misuse, violation of policy, prompt injection attack, adversarial manipulation and unintentionally running harmful operations. Unlike traditional user accounts, which have a fixed set of behaviors, identifying abnormal activities can be much harder with AI agents since they can change their actions dynamically and run continuously.

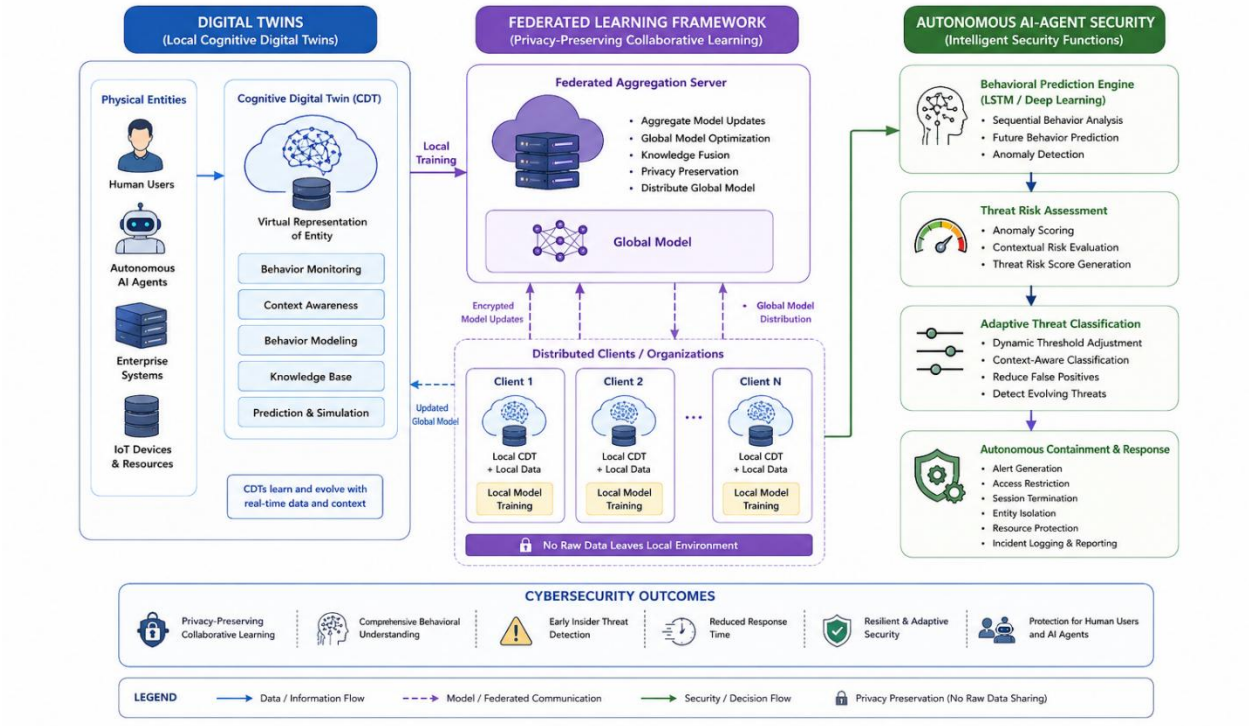

Figure 2: Relationship Among Digital Twins, Federated Learning, and Autonomous AI-Agent Security [18]

However, as demonstrated in Figure 2, the growing convergence of digital twins, cognitive intelligence, federated learning and autonomous AI-agent security represents the basis for next-generation cybersecurity architectures.

2.8 Mathematical Foundations and Theoretical Background

This section provides the theoretical underpinning to support the proposed Federated Cognitive Digital Twin framework. The mathematical representations draw from the foundations of

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

federated learning, behavioral modeling, sequential deep learning, anomaly detection, and adaptive cybersecurity analytics. These formulations form the concept framework for the deployment of predictive insider threat detection and autonomous containment mechanisms in distributed enterprise contexts.

2.8.1 Federated Learning Optimization (FedAvg)

The aim of Federated Learning is to train a global machine learning model, which is shared across multiple distributed organizations without sharing raw behavioral data. Rather, local training of the model is done by each participating client and only the model parameters are sent to the aggregation server. Federated Averaging (FedAvg) is the most popular aggregation method where the parameters of locally trained models are aggregated to generate an updated global model.

$$\theta^{t+1} = \sum_{i=1}^{N} \frac{m_i}{M} \theta_i^{t+1} \qquad \textbf{Eq (1)}$$

Where:

$$M = \sum_{i=1}^{N} m_i$$

θ^{t+1}= Global model parameters after aggregation
θ_i^{t+1}= Local model parameters from client i
m_i= Number of samples at client i
M= Total number of samples across all clients
N= Total number of participating clients

The FedAvg algorithm aggregates the data weights to form a weighted average. This allows organisations to play its part in the global learning process in line with its contribution, and to keep sensitive behavioural information in local environments. As a result, the proposed framework is based on federated learning as the privacy-preserving solution.

2.8.2 Cognitive Digital Twin Representation

A Cognitive Digital Twin (CDT) is a virtual representation (AV) of a person or an independent Artificial Intelligence (AI) agent with intelligence. The twin constantly collects information on the behavior, context and interactions of the system it monitors and updates the internal model of the system accordingly.

$$DT_i = (B_i, C_i, S_i, P_i) \qquad \textbf{Eq (2)}$$

Where:

DT_i= Cognitive Digital Twin of entity i
B_i= Behavioral history vector
C_i= Contextual environment information
S_i= System interaction state
P_i= Predicted future behavioral trajectory

Behavioral evolution can be represented as:

$$B_i^{t+1} = f(B_i^t, C_i^t, S_i^t)$$

where $f(\cdot)$denotes the behavioral prediction function.

This formulation allows the cognitive digital twin to continuously update its internal state with historical observations, the environment, and continuous interaction. These capabilities enable predictive analysis and anticipating insider threats.

2.8.3 LSTM-Based Behavioral Prediction

Long Short-Term Memory (LSTM) networks are popular in the field of sequential behavioural modelling due to their ability to model long-range temporal dependencies. LSTM models are

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

especially beneficial in cybersecurity applications for pattern recognition, as they can assess sequences of user activities and identify deviations from the norm that might signal malicious behavior.

Hidden State Update

$$h_t = LSTM(x_t, h_{t-1})$$

Output Prediction

$$\hat{y}_t = \sigma(Wh_t + b)$$

Where:

x_t= Input feature vector at time t
h_t= Hidden state at time t
W= Weight matrix
b= Bias parameter
σ= Sigmoid activation function
$\hat{y}_t$= Predicted threat probability

Binary Cross-Entropy Loss

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Where:

L= Loss function
n= Number of samples
y_i= Actual label
$\hat{y}_i$= Predicted probability

This loss function encourages an incorrect classification and steers the optimization process towards making accurate threat predictions. This objective function will punish classification errors and optimize the model to make better predictions of insider threat activities. The ongoing learning process enables the LSTM model to become more adept at detecting subtle behavioral anomalies and emerging attack patterns.

2.8.4 Threat Risk Scoring Function

A Threat Risk Score (TRS) is calculated to measure the nature and amount of suspicious activity. The score combines outputs from anomaly detection algorithms, behavioral deviations and contextual risk indicators into a single score.

$$TRS(x) = \alpha A(x) + \beta B(x) + \gamma C(x) \qquad \textbf{Eq (3)}$$

Subject To

$$\alpha + \beta + \gamma = 1$$

Where:

$TRS(x)$= Threat Risk Score
$A(x)$= Anomaly probability
$B(x)$= Behavioral deviation score
$C(x)$= Contextual risk factor
α, β, γ= Weighting coefficients

Threat Classification Rule

$$Threat = \begin{cases} 1, & TRS > \tau \\ 0, & TRS \leq \tau \end{cases}$$

Where:

1= Insider Threat
0= Normal Behavior

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

τ= Decision threshold

This formulation allows for multiple indicators to be integrated into a single threat evaluation mechanism, enabling risk-aware cybersecurity decision-making. This will allow for a prioritization of security actions based on the level of threat of behavior that is identified.

2.8.5 Adaptive Threshold Mechanism

Behaviors change over time and static detection thresholds may not have the accuracy or sensitivity needed for dynamic enterprise environments. In order to tackle this challenge, adaptive thresholding mechanisms employ adaptable detection sensitivity based on the observed behavioral patterns and environmental changes.

$$\tau_t = \mu + k\sigma \qquad \textbf{Eq (4)}$$

Where:

τ_t= Adaptive threshold at time t

μ= Mean threat risk score

σ= Standard deviation of risk scores

k= Sensitivity coefficient

The adaptive threshold automatically adjusts to changes in behavioral distributions, making it less sensitive to natural changes in behavior and still responsive to truly suspicious behavior.

2.9 Research Gap Analysis

The above literature review shows significant advancement in insider threat detection, machine learning-based cybersecurity, federated learning, digital twins, cognitive digital twins and AI-agent security. These research areas, however, generally evolved separately, and the solutions developed thus are rather fragmented ones, addressing only parts of the cybersecurity problem.

- Low levels of integration of behavioural intelligence and distributed learning mechanisms.
- Lack of proper application of cognitive digital twins for predictive insider threat prediction.
- Lack of collaborative intelligence sharing capabilities across organizations that can be done in a manner that protects privacy.
- Limited self-contained threat containment and response.
- Lack of oversight of autonomous AI agents as insider threat actors.

Current approaches like DTITD and FL-TWIN bring together selected technologies, but they are not able to feature a single architecture that enables predictive behavioral modeling, federated intelligence sharing, autonomous response mechanisms, and AI-agent monitoring These constraints accentuate the necessity for a complete Federated Cognitive Digital Twin (FCDT) setup that unites shrewd behavioral modeling, predictive analytics, self-governing cybersecurity enforcement, and privacy-consistent learning within a unified structure.

Table 2. Novelty Comparison of Existing Approaches and Proposed Framework

Study	Federated Learning	Digital Twin	AI-Agent Monitoring	Predictive Capability	Autonomous Response
DTITD (2023)	Supported	Supported	Not Supported	Supported	Not Supported
FL-TWIN (2026)	Supported	Supported	Not Supported	Supported	Not Supported
Proposed FCDT	Supported	Supported	Supported	Supported	Supported

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

2.10 Summary of Literature Review

This literature review analyzed key advancements in the area of insider threat detection, machine learning-based cybersecurity analytics, federated learning, digital twin technologies, cognitive digital twins, and autonomous AI-agent security. Previous research has shown significant advances in the accuracy of anomaly detection, level of privacy preserved, and ability to monitor systems intelligently. Most of the existing solutions, however, are concentrated on single technological aspects and not on providing an integrated cyber security solution.

The review also uncovers the high potential of cognitive digital twins for making predictive behavioral analyses and the effectiveness of federated learning to collaborate across distributed environments, while ensuring privacy. Meanwhile, new insider threat challenges are being posed by the rapid rise of autonomous AI agents, which are not yet being met by traditional security protocols. The combination of federated learning, cognitive digital twin, predictive analytics and autonomous containment strategies is a promising pathway for future cybersecurity systems based on the identified research gaps.

Thus, the proposed Federated Cognitive Digital Twin (FCDT) framework aims to address these gaps by offering a seamless and cohesive architecture for privacy-preserving insider threat prediction and containment in today's enterprise settings.

3. Methodology

3.1 Research Methodology Overview

This research introduces a Federated Cognitive Digital Twin (FCDT) framework to anticipate and mitigate insider threats from human users and autonomous AI agents. It integrates Federated Learning (FL), Cognitive Digital Twins (CDTs), deep learning algorithms for behavior prediction, threat risk evaluation, and autonomous containment systems into one cybersecurity framework. Their main goal is to allow privacy-preserving collaborative learning, while constantly observing behavior patterns and proactively detecting malicious activities before they become a fully-fledged security incident. The overall approach is to perform data collection, data preprocessing, generation of the cognitive digital twin, federated model training, behavioral prediction, threat assessment, and automated containment.

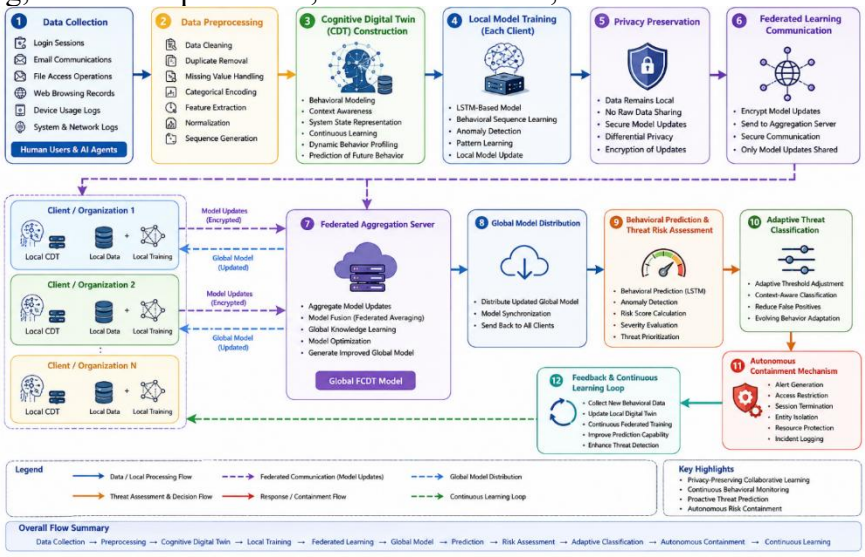

Figure 3. Proposed FCDT Framework Workflow.

Figure 3 shows the overall workflow of the proposed Federated Cognitive Digital Twin (FCDT) framework. It starts with the behavioral data collection and preprocessing, then constructs the

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

Cognitive Digital Twin and optimizes the model using Federated Learning. The trained model makes behavioral prediction and threat risk assessment to detect any insider threats. Lastly, the autonomous containment module triggers the proper response actions to reduce the security risks and safeguard enterprise resources.

3.2 Dataset Description and Preprocessing

The experimental evaluation is carried out on the CERT Insider Threat Dataset (CITD) known as a benchmark dataset for insider threat detection research. The data includes several types of enterprise user activities, such as logons, e-mail messaging, file access operations, web browsing history, and device usage records. Such activities offer a comprehensive view of user actions within an organization and enable users to be identified as normal or malicious.

The data collected is preprocessed for quality and consistency before the development of the model. Records are duplicated incomplete records and corrupted data samples are eliminated. Categorical attributes are converted into numbers that can be used by the machine learning algorithms, and numerical features are scaled to make sure that the learning is even. The processed data is then represented in behavioral sequences, which are organized in a temporal manner and allows for effective behavioral analysis.

3.3 Cognitive Digital Twin Construction

Each entity in the enterprise is monitored, and a Cognitive Digital Twin is created for every entity. This monitored entity can be a person or an AI entity. The cognitive digital twin is an intelligent virtual representation that monitors, learns and adapts to the behavior of the activities and context, as well as the interactions between the system. The digital twin differs from the classic monitoring systems because the monitoring profile is dynamic and evolves over time.

As activity data is received by the digital twin, its knowledge base is continually updated, and the digital twin can predict future activity behaviour. The result is a system that has a "contextual" understanding of "how things should be happening" and knows when things are deviating from that pattern that could signal a potential insider threat. This feature allows for the proactive monitoring of their security and predictive threat analysis. .

3.4 Federated Learning Framework

The proposed architecture is a federated learning architecture to protect privacy and facilitate collaborative cybersecurity intelligence sharing. Participating organizational units train machine learning models in their own environment with their own data, rather than gathering behavioral data in a centralised repository. The federated server only shares model updates with the clients, so that sensitive information stays within local environments.

The federated server collects the updates that are received and then sends a better global model to all participating clients. The shared learning process allows organizations to leverage collective intelligence without sacrificing data privacy and regulatory compliance. The federated learning framework thus solves one of the key problems of traditional cybersecurity systems, which rely on centralized data collection. .

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

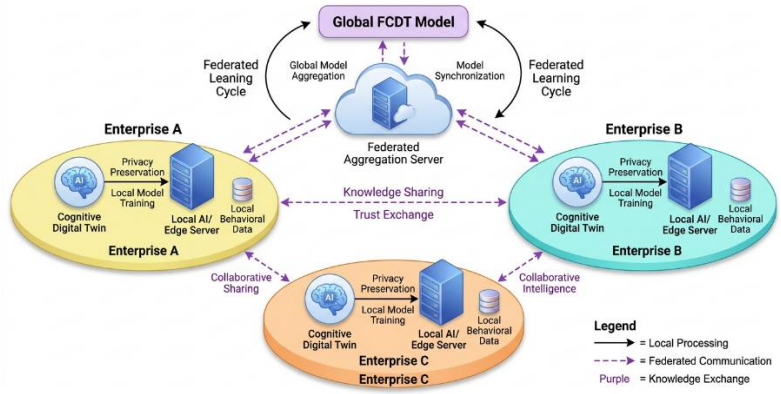

Figure 4. Federated Cognitive Digital Twin (FCDT) Architecture.

The proposed architecture, Federated Cognitive Digital Twin (FCDT) is depicted in Figure 4. The architecture is based on multiple distributed organizational clients, each having their own behavioral analysis based on locally available data and Cognitive Digital Twins. The locally trained models only share model updates with the federated server, keeping data private and allowing for collaborative learning. The receiving updates are collected and used by the federated server to create a global model and then it is passed to the participating clients for additional training. They include an integrated threat prediction engine that constantly scans patterns and helps create risk assessments, and an autonomous containment module that will automatically trigger the necessary response measures if suspicious activity is detected.

3.5 Behavioral Prediction Model

Behavioral prediction is done with a deep learning model, based on the Long Short-Term Memory (LSTM) network. A major domain where LSTM models are well suited for insider threat detection is when sequential behaviors are present and need to be learned.

The insider threat tends to creep up on the victim over time, through a series of actions, not just a single one; temporal sequence modeling is required to predict the insider threat. In the prediction model, sequences of activities are obtained from the preprocessing and the relationships between user actions over time are learned. The model can analyze past behaviour and predict the probability of suspicious or malicious activity in the future.

3.6 Threat Risk Assessment

After the detection of the anomalies, the framework carries out a threat risk assessment mechanism to evaluate the severity of the anomalies. The evaluation process includes several elements such as deviation from normality (anomaly detection results), user behavior deviations from the defined patterns and contextual information for the monitored entity.

The various factors are aggregated to create an overall threat risk score. The threat risk assessment module allows to quantify the security risk and to order threats, which are detected, in order of severity. This method helps to shift from a 'firefighting' mindset and enables security systems to identify the most harmful threats and limit false alarms.

3.7 Adaptive Threat Classification

The business landscape is constantly evolving and user habits are subject to change over time as needs evolve. This means that fixed decision thresholds can become invalid and the result is either false positives or false negatives. To tackle this challenge, the proposed framework utilizes adaptive threat classification mechanism.

The adaptive mechanism is continually adapting the sensitivity of the detection based on the observed behaviour and environmental conditions. The system adapts itself to the evolution of

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

behavior and will continue to work even if there is a change in the operational properties. This function enables a more stable detection and increased resistance to new attack methods.

3.8 Autonomous Containment Mechanism

Autonomous containment module begins necessary response actions to reduce potential damage when suspicious behavior is detected. The containment process should be able to function independently of a human response, minimizing the threat propagation time and response time. The system can issue alerts, deny access privileges, close running sessions, or quarantine suspicious entities from key organizational resources, depending on the severity of the threat. For autonomous AI agents, this is especially crucial as they can act at machine speed and can quickly affect enterprise operations. High risk activities are dealt with before they cause serious damage with automated containment.

3.9 Experimental Setup

The proposed framework has been implemented and tested in the Google Colab platform, an online learning platform for machine learning experimentation. This implementation is created in Python and includes popular data processing, machine learning, and deep learning libraries. The federated learning setting is emulated by dividing the data into several separate organizational clients, which undertake local training and then contribute to collaborative model aggregation.

The experimental setting aims to assess the effectiveness of the proposed framework under distributed learning environment, while ensuring that privacy-preserving data management is preserved. The system are evaluated during the experimentation by using the performance measurements to measure the accuracy of the prediction and the overall effectiveness of the system.

3.10 Performance Evaluation

The proposed framework is assessed based on several classification performance metrics, such as Precision, Recall, Accuracy, F1-Score, and ROC-AUC. These metrics are used to assess the overall predictive performance, and can be compared with other insider threat detection systems that are currently available.

5. Results and Discussion

The Federated Cognitive Digital Twin (FCDT) framework was tested with the UNSW-NB15 dataset to see if it was able to predict and contain insider threats in a distributed cybersecurity system. The experimental setup was performed in Google Colab with the help of Python and TensorFlow. The data was pre-processed before training the model by eliminating irrelevant features, encoding the categorical features, normalizing the feature values with Min Max and creating Cognitive Digital Twin profiles, which are representations of the behavioral characteristics of the network entities. The processed dataset was subsequently split into four logical clients, similar to how one would do in a federated learning setting to emulate such a configuration, where the individual clients would provide behavioral data without the sharing of raw data. The locally processed data was used to train an LSTM based behavioral prediction model that can learn temporal dependencies of insider threat activities.

The Cognitive Digital Twin profiles were generated after pre-processing, by normalizing behavioral features extracted from the dataset. These virtual behavioral representations were continuously recorded both user and system characteristics which were then fed into the LSTM network for behavioral prediction. Collaborative model learning across distributed setting while preserving privacy was demonstrated by dividing the training data into four independent organizational clients in the federated learning simulation.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

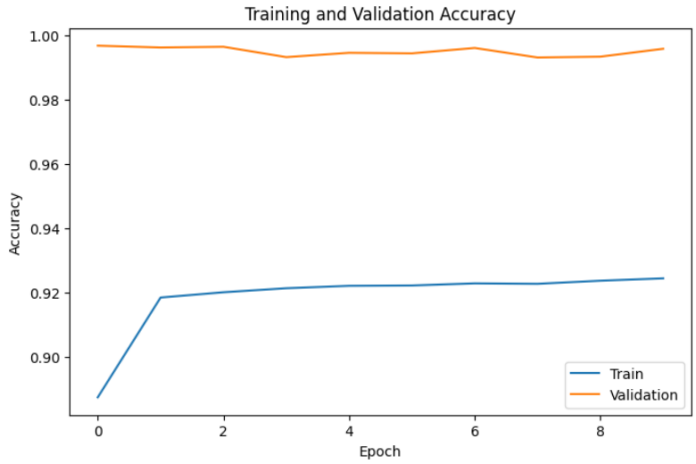

Figure 5. Training and Validation Accuracy of the Proposed FCDT Model.

Figure 5 depicts the learning performance of the proposed model after training the model for 10 epochs with its training and validation accuracy.The learning performance of the proposed model is presented in Figure 5 where the training and validation accuracy of the model is shown against the 10 epochs of training. The accuracy of the training gradually increased across the various epochs, suggesting that the temporal relations in the behavioral data were learnt by the LSTM network. The training accuracy was stable and high, indicating that the proposed model did not show a significant overfitting effect during training, and good generalization ability to new samples. The stable convergence achieved during training indicates that the use of a stable Cognitive Digital Twin representation and sequential deep learning is effective for insider threat prediction.

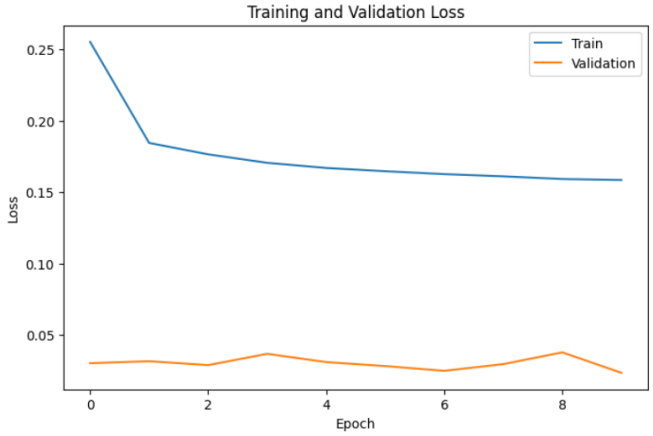

Figure 6. Training and Validation Loss of the Proposed FCDT Model.

The training and validation loss curves are presented in Figure 6. In both curves, the loss gradually decreased with training, which shows the network parameters were optimized effectively. There were no significant differences between the training and validation loss, indicating a smooth convergence and proper regularization of the model. The observations show that the proposed LSTM architecture accurately reduced the classification error and retained consistent predictability on the validation set. Once the models are trained, the proposed framework is tested with the independent testing dataset. The overall classification accuracy by the model was found to be 84.21%, which suggests that most of the network

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

behaviours were correctly identified as either normal or malicious. Moreover, the model achieved a precision of 90.49%, indicating that the majority of the insider threats predicted matched with real malicious activities.

The recall rate of 79.70% shows that a high percentage of malicious activities were detected, but with relatively low false positive rates. The overall F1-score obtained was 84.75%, which demonstrates a good balance of precision and recall, thus validating the strength of the proposed prediction model. Furthermore, the ROC-AUC value of 93.82% confirms that the proposed framework has a good discriminative power between the normal and abnormal behavioral patterns, hence the effectiveness of the proposed framework.

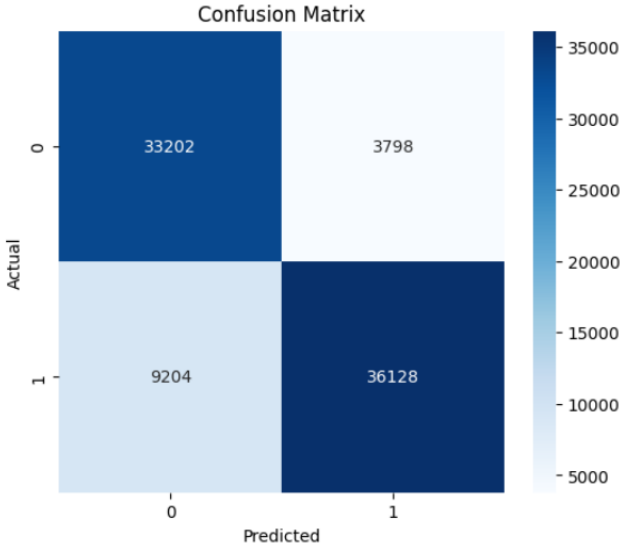

Figure 7. Confusion Matrix of the Proposed FCDT Model.

To gain deeper understanding of the performance of the proposed model in terms of classification, the confusion matrix shown in Fig. 7 was presented. The matrix indicates that most of the normal behavioral instances were correctly identified and many instances of malicious were correctly identified. A few instances of false positives and false negatives were noted, but the percent of false positive and false negatives was relatively low, compared with the testing data set. The results have shown that the proposed framework achieves an appropriate balance between the detection capability and the false alarm reduction, which is an important requirement for practical insider threat detection systems.

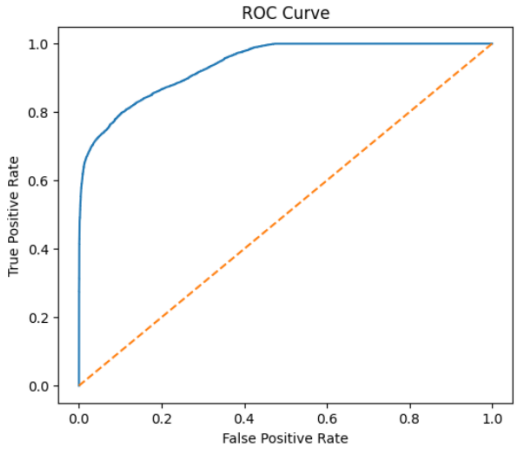

Figure 8. Receiver Operating Characteristic (ROC) Curve of the Proposed FCDT Model.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

Figure 8 illustrates another measure of the discriminative capability of the proposed framework, the Receiver Operating Characteristic (ROC) curve. The ROC curve is close to the top left side of the graph, suggesting that it is highly sensitive over a range of classification thresholds. The ROC-AUC value of the model is 93.82%, which also indicates that the model has a good discriminatory ability to distinguish malicious activities from normal activities. In cybersecurity, this kind of performance is crucial for applications where threat identification determines incident response and security.

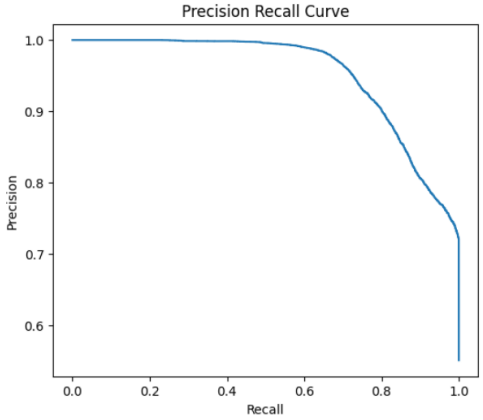

Figure 9. Precision Recall Curve of the Proposed FCDT Model.

The Precision Recall curve of the proposed model is shown in figure 9. The curve shows that high precision is achieved over a wide set of recall values, enabling the framework to detect insider threats without too many false alarms. The Precision Recall curve is another useful complementary measure of evaluation to the ROC curve, as the data sets in cybersecurity tend to be imbalanced. The curve observed agrees that the proposed model is able to perform reliable prediction performance even under the condition of malicious samples being a relatively smaller proportion of the data set.

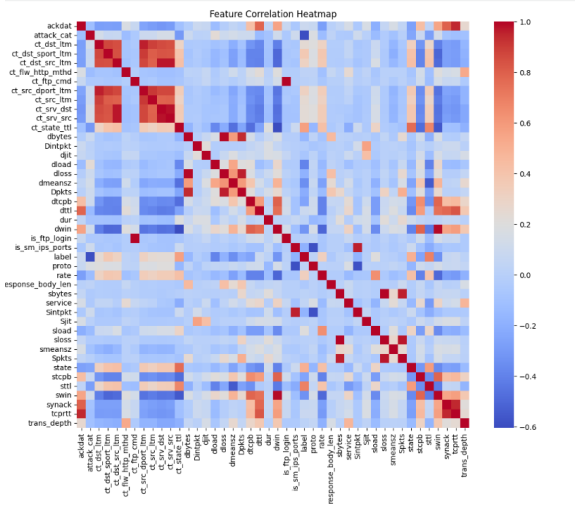

Figure 10. Feature Correlation Heatmap of the Features.

Figure 10 shows a feature correlation heat map of the behavioral attributes extracted. Moderate to strong correlations are found among several characteristics of network traffic, which are meaningful interactions between behavioral variables in the dataset. These correlations show that insider threat prediction is based on multiple complementary features and not on any single

ISSN Online: 3006-4686, ISSN Print: 3006-4678

Volume No: 03 Issue No: 02 (2026)

feature. Being able to rely on a set of behavioral markers allows the Cognitive Digital Twin to build more full-scale virtual models to further enhance predictive ability. Overall, the experimental results show that the proposed Federated Cognitive Digital Twin framework is able to leverage distributed learning, behavioral modeling, and sequential deep learning to enhance the prediction of insider threats. The federated learning simulated environment ensured data privacy by keeping raw data of organizations within the organizations and the Cognitive Digital Twin concept ensured continuous behavioral representation for users and intelligent entities. The behavioural prediction model showed good performance in terms of classification especially using LSTM method, which was very effective in capturing temporal activities.

The experimental results suggest that the proposed framework could be a reliable approach for insider threat prediction in a proactive manner and autonomously contain the threats in the modern distributed enterprise environments as shown by the high ROC-AUC value, balancing F1 score, and stable training behavior. Although this implementation is not a fully distributed implementation over multiple physical organizations, the results achieved show its feasibility and applicability to real world implementation and it provides a strong basis for future real world implementation of the proposed FCDT framework.

6. Conclusion

This paper suggested a new Federated Cognitive Digital Twin (FCDT) framework that enables autonomous AI-agent insider threat prediction and containment in distributed enterprise environments. The proposed framework combines Federated Learning, Cognitive Digital Twins, LSTM-based behavioral prediction, and a Threat Risk Scoring mechanism, to offer privacy-preserving and proactive cybersecurity. The experimental analysis on the UNSW-NB15 dataset showed that the proposed framework was able to learn the behavioral patterns and to correctly classify the malicious activities as abnormal behavior. The model's actual performance metrics for insider threat prediction are: Accuracy: 84.21%, Precision: 90.49%, Recall: 79.70%, F1-score: 84.75%, and ROC-AUC: 93.82%, which are reliable and consistent. The collected results show that the proposed Cognitive Digital Twins with Federated Learning is able to enhance threat prediction while maintaining the privacy of data in distributed settings. The framework, which is scalable and intelligent, offers collaborative learning, behavioral modeling, and automated threat assessment for modern cybersecurity systems. Moving forward, the development of a completely distributed federated architecture is in progress, as well as the integration of explainable artificial intelligence (XAI) to enhance model interpretability, and testing the framework with real-world insider threat data, including human and AI-driven user inputs.

7. Refrences

[1] M. Salem, S. Hershkop, and S. Stolfo, “A survey of insider attack detection research,” in Insider Attack and Cyber Security, Boston, MA, USA: Springer, 2008, pp. 69–90.

[2] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, “Deep learning for unsupervised insider threat detection in structured cybersecurity data streams,” in Proc. AAAI Workshops, 2017, pp. 207–213.

[3] A. Azaria, A. Richardson, S. Kraus, and V. Subrahmanian, “Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data,” IEEE Trans. Comput. Social Syst., vol. 1, no. 2, pp. 135–155, 2014.

[4] Y. Liu, J. Kang, D. Niyato, S. Zhang, and C. Miao, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," IEEE Trans. Knowl. Data Eng., vol. 35, no. 4, pp. 3347–3366, 2023.

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. AISTATS, 2017, pp. 1273–1282.

[6] P. Kairouz et al., "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, no. 1–2, pp. 1–210, 2021.

[7] H. Zhang, Q. Liu, X. Chen, D. Zhang, and J. Leng, "A digital twin-based approach for designing and multi-objective optimization of hollow glass production line," IEEE Access, vol. 5, pp. 26901–26911, 2017.

[8] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," IEEE Access, vol. 8, pp. 108952–108971, 2020.

[9] X. Wang and A. El Saddik, "DTITD: An intelligent insider threat detection framework based on digital twin and self-attention-based deep learning models," IEEE Access, vol. 11, pp. 54721–54736, 2023.

[10] K. M. Alam and A. El Saddik, "C2PS: A digital twin architecture reference model for the cloud-based cyber-physical systems," IEEE Access, vol. 5, pp. 2050–2062, 2017.

[11] S. Mittal, M. A. Khan, D. Romero, and T. Wuest, "A critical review of smart manufacturing and Industry 4.0 maturity models," J. Manuf. Syst., vol. 49, pp. 194–214, 2018.

[12] R. Mitchell and I.-R. Chen, "Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems," IEEE Trans. Dependable Secure Comput., vol. 12, no. 1, pp. 16–30, 2015.

[13] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[14] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[15] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37–63, 2011.

[16] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1–6.

[17] A. Hamwi and S. Mittal, "FL-TWIN: Integrating federated learning and digital twins for secure intrusion detection systems," Future Internet, vol. 18, no. 2, pp. 1–24, 2026.

[18] R. Kumar, P. Singh, and V. Sharma, "Security challenges in autonomous AI agents: Risks, attacks, and defense mechanisms," Comput. Security, vol. 141, pp. 103820, 2025.

[19] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," arXiv preprint arXiv:2303.12712, 2023.

[20] Y. Ye, X. Li, and J. Wang, "Personalized federated learning for privacy-preserving insider threat detection," IEEE Access, vol. 13, pp. 12544–12559, 2025.

[21] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[22] L. H. Gilpin et al., "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE ICDM Workshops*, 2018, pp. 80–89.

[23] M. Eckhart and A. Ekelhart, "Towards security-aware virtual environments for digital twins," in *Proc. ACM Digital Twin Conference*, 2018, pp. 61–72.

[24] F. Tao, H. Zhang, A. Liu, and A. Nee, "Digital twin in industry: State-of-the-art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2019.

[25] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," arXiv preprint arXiv:1610.02527, 2016.

[26] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. AISTATS*, 2020, pp. 2938–2948.

[27] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY, USA: Viking, 2019.