OPTIMIZATIONS IN APPLIED MACHINE LEARNING
Research Article* | *Volume 5* | *Issue 5* | *Apr 2025
Received: 22 Feb 2025 | Revised: 29 Feb 2025
Accepted: 26 Mar 2025 | Published Online: 01 Apr 2025

Highly efficient cloud computing via Adaptive Hierarchical Federated Learning

Shao Dan [1] and Qinyi Zhu [2, *]

[1] *ASCENDING Inc., Fairfax VA 22031, USA*

[2] *Indiana University, 107 S Indiana Ave, Bloomington, IN 47405, USA*

[*]Corresponding Author, Email: qinyzhu@iu.edu

Abstract: Cloud computing has revolutionized the way data is processed and stored, leading to increased demand for efficient machine learning models. However, the current centralized nature of cloud-based machine learning poses challenges in terms of scalability and privacy protection. This paper addresses these obstacles by proposing a novel approach called Adaptive Hierarchical Federated Learning. This approach enables the efficient distribution of machine learning tasks across multiple layers of a hierarchical cloud architecture, allowing for improved scalability and enhanced privacy preservation. The innovative method presented in this paper harnesses the power of federated learning while adapting dynamically to the varying computational resources within the hierarchical cloud environment. Through extensive experiments, the effectiveness and efficiency of the proposed Adaptive Hierarchical Federated Learning are demonstrated, highlighting its potential to significantly advance the field of cloud computing.

Keywords: *Cloud Computing; Machine Learning; Scalability; Privacy Protection; Federated Learning*

1. Introduction

Cloud Computing is a rapidly evolving field focused on delivering computing services over the internet, including storage, processing power, and software applications. One of the major challenges in the field of Cloud Computing is data security and privacy concerns, as the vast amount of data stored and processed on cloud servers requires robust security measures to protect against cyber threats. Additionally, ensuring high availability and reliability of cloud services remains a key bottleneck, as downtime or service interruptions can have significant impacts on businesses and users. Scalability and interoperability between different cloud platforms also present challenges for seamless integration of cloud services. Overall, addressing these issues and advancing

technology in Cloud Computing is crucial for unlocking its full potential and driving innovation in the digital age.

To this end, research in the field of Cloud Computing has advanced significantly, with studies focusing on scalability, security, and cost-efficiency of cloud services. Innovations such as edge computing and containerization have expanded the capabilities of cloud platforms, offering new avenues for exploration in this rapidly evolving field. Cloud computing has been defined by the National Institute of Standards and Technology (NIST) as a model that enables convenient, on-demand network access to a shared pool of configurable computing resources. This model promotes availability and consists of essential characteristics, service models, and deployment models [1]. Armbrust et al. provided a comprehensive view of cloud computing, discussing its impact and potential across various sectors [2]. Calheiros et al. introduced CloudSim, a toolkit for modeling and simulating cloud computing environments and evaluating resource provisioning algorithms [3]. Furthermore, Zhang et al. conducted a survey on the state-of-the-art and research challenges in cloud computing, emphasizing the need for further advancements in the field [4]. Buyya et al. explored the vision, hype, and reality of cloud computing as the 5th utility, highlighting the evolution and critical aspects of the technology [5]. Katal et al. focused on the energy efficiency aspects of cloud data centers, discussing the software technologies that contribute to sustainable practices and environmental impact reduction [6]. Chen et al. proposed an efficient multi-user computation offloading strategy for mobile-edge cloud computing, utilizing game theory and distributed decision-making approaches [7]. Foster et al. compared and contrasted cloud computing with grid computing, shedding light on the fundamental characteristics and connections between the two paradigms [8]. Finally, Sadeeq et al. reviewed the challenges and opportunities in integrating IoT with cloud computing, emphasizing the need for stable transitions and efficient computing techniques in this evolving landscape [9]. Hierarchical Federated Learning is a crucial technique in the field of cloud computing due to its ability to enhance data privacy, scalability, and communication efficiency in distributed machine learning systems. By utilizing a hierarchical structure, this approach enables collaborative model training across multiple edge devices while preserving data security and minimizing communication overhead. This technique is essential for addressing the challenges of data privacy and scalability in modern cloud computing environments.

Specifically, Hierarchical Federated Learning fundamentally improves data privacy and computational efficiency by distributing machine learning tasks across a network of edge devices before aggregating insights in the cloud. This synergistic relationship with Cloud Computing optimizes resource utilization and enhances model robustness. The research on hierarchical federated learning has gained significant attention recently. Liu et al. [10] introduced a client-edge-cloud hierarchical federated learning system with the HierFAVG algorithm, which allows for more efficient communication and computation trade-offs. Deng et al. [11] proposed a communication-efficient hierarchical federated learning framework via shaping data distribution at the edge, demonstrating the effectiveness of edge aggregations. Wang et al. [12] presented a UAV swarm-assisted two-tier hierarchical federated learning scheme, optimizing the FL convergence with UAV relays. Zhao et al. [13] developed a DRL-based resource allocation framework for hierarchical federated learning in NOMA-enabled Industrial IoT. Aouedi et al. [14] introduced HFedSNN, an energy-efficient hierarchical federated learning model using spiking neural networks. Tong et al.

[15] proposed a blockchain-based hierarchical federated learning framework for UAV-enabled IoT networks to improve trust and efficiency. Lim et al. [16] discussed decentralized edge intelligence for dynamic resource allocation in hierarchical federated learning. Zhou et al. [17] implemented a unique clustering-based participant selection method for hierarchical federated learning in Internet of Medical Things applications. Lastly, Abad et al. [18] introduced a hierarchical federated learning scheme across heterogeneous cellular networks, optimizing communication latency without compromising accuracy. However, some limitations in current research on hierarchical federated learning include scalability issues with increasing numbers of participants, potential privacy concerns with data distribution at the edge, and challenges in ensuring trust and efficiency in blockchain-based frameworks.

To overcome those limitations, this paper aims to address the challenges posed by the current centralized nature of cloud-based machine learning through the introduction of a novel approach called Adaptive Hierarchical Federated Learning. This innovative method facilitates the efficient distribution of machine learning tasks across multiple layers of a hierarchical cloud architecture, thereby enhancing scalability and privacy preservation. The key detail lies in the adaptability of the approach to dynamically adjust to the varying computational resources within the hierarchical cloud environment, maximizing the utilization of resources while maintaining data privacy. Through a series of extensive experiments, the paper showcases the effectiveness and efficiency of Adaptive Hierarchical Federated Learning, underscoring its potential to propel advancements in the realm of cloud computing and machine learning.

Section 2 of the study presents the problem statement, highlighting the challenges posed by the centralized nature of cloud-based machine learning. Section 3 introduces the proposed solution, Adaptive Hierarchical Federated Learning, which aims to address scalability and privacy protection issues. In Section 4, a case study is presented to demonstrate the application of this novel approach. Section 5 analyzes the results of extensive experiments, showcasing the effectiveness and efficiency of Adaptive Hierarchical Federated Learning. The discussion in Section 6 delves into the implications and potential advancements brought about by this innovative method. Finally, Section 7 provides a comprehensive summary of the research findings, emphasizing the significant contribution of this study to the field of cloud computing.

2. Background

2.1 Cloud Computing

Cloud Computing is a paradigm shift in computing resources that has fundamentally transformed how data, applications, and infrastructure are managed and delivered over the internet. It offers scalable and on-demand resources, providing a flexible and efficient alternative to traditional on-premise data centers. Cloud computing encompasses a variety of services, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), each catering to different computing needs.

At its core, cloud computing leverages virtualization technology to pool computing resources, allowing multiple users to share the same physical infrastructure while maintaining isolation of

their data and applications. The key characteristics of cloud computing include resource pooling, rapid elasticity, measured service, broad network access, and on-demand self-service. To understand cloud computing more formally, one can decompose its functionalities into mathematical representations. The primary elements involved include resource allocation, cost optimization, and performance enhancement. These can be delineated using optimization and resource management equations. Consider a cloud environment consisting of N users and M resources. The allocation of resources to users can be described by a matrix $A \in \mathbb{R}^{N \times M}$, where each element a_{ij} represents the allocation of resource j to user i. The total cost for utilizing cloud resources can be expressed as a function $C(A)$, typically dependent on factors like usage time, bandwidth, and computational power:

$$C(A) = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} \cdot a_{ij} \tag{1}$$

where c_{ij} represents the cost per unit of resource j allocated to user i. Performance optimization in a cloud system could involve maximizing throughput, minimizing latency, or balancing the load across servers. One might aim to maximize a performance function $P(A)$ subject to resource constraints:

$$P(A) = \sum_{i=1}^{N} u_i \cdot f(a_{i1}, a_{i2}, \ldots, a_{iM}) \tag{2}$$

where u_i is a utility coefficient for user i, and $f(\cdot)$ is a performance measurement function. Cloud providers need to ensure that the sum of the allocated resources does not exceed the total available resources R_j for each type j:

$$\sum_{i=1}^{N} a_{ij} \leq R_j \forall j = 1,2, \ldots, M \tag{3}$$

The optimization problem hence involves constraints that maintain the resource availability across users and maximize their performance:

$$\text{Maximize} P(A) - \alpha C(A) \tag{4}$$

subject to the allocation constraints and non-negativity of a_{ij}:

$$a_{ij} \geq 0 \forall i = 1,2, \ldots, N \text{and} \forall j = 1,2, \ldots, M \tag{5}$$

In conclusion, cloud computing stands as a revolutionary approach to resource management in computing, characterized by its utilization of mathematical models and optimization techniques to realize efficiency and cost-effectiveness. Its foundation lies in the balanced distribution of resources, cost minimization, and performance optimization, all facilitated by sophisticated algorithms and large-scale data center infrastructures.

2.2 Methodologies & Limitations

Cloud computing, as an innovative paradigm for resource management and service delivery, largely relies on complex optimization strategies to meet the diverse demands of users. These strategies include resource allocation models, pricing schemes, and performance enhancement techniques, each with their inherent limitations and potential for improvement. One of the forefront methodologies in cloud computing is dynamic resource allocation, aiming to efficiently distribute computing resources such as CPU, memory, and storage among multiple users. This is represented by the allocation matrix $A \in \mathbb{R}^{N \times M}$, defining the allocation of M resources to N users. The optimal allocation minimizes costs while ensuring high performance. The cost function $C(A)$ for using cloud resources, described as:

$$C(A) = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} \cdot a_{ij} \tag{6}$$

is central to cloud pricing models, where c_{ij} denotes the cost per unit resource allocated to user i. However, these pricing strategies can lack transparency and flexibility for customers needing predictable and scalable billing options. Performance optimization is another vital component, often realized through strategies to maximize throughput and minimize latency. The performance function $P(A)$ is defined by:

$$P(A) = \sum_{i=1}^{N} u_i \cdot f(a_{i1}, a_{i2}, \ldots, a_{iM}) \tag{7}$$

where u_i is a utility coefficient, reflecting the priority or importance of satisfying user i 's needs. Despite the theoretical versatility of this function, real-world implementations can face challenges due to unpredictable workloads and varying network conditions. Constraints on available resources are mathematically expressed as:

$$\sum_{i=1}^{N} a_{ij} \leq R_j \forall j = 1,2, \ldots, M \tag{8}$$

ensuring the allocated resources do not exceed their physical limits. This constraint protects the integrity of the cloud provider's infrastructure but can result in resource under-utilization if static allocations are improperly configured. Moreover, the optimization problem attempts to balance maximizing performance with minimizing costs, formalized as:

$$\text{Maximize} P(A) - \alpha C(A) \tag{9}$$

where α is a weighting factor that balances performance goals against cost considerations. While theoretically sound, this balance can be difficult to maintain under diverse workloads and operational conditions, often necessitating complex, adaptive algorithms. Additionally, non-negativity constraints are imposed:

$$a_{ij} \geq 0 \forall i = 1,2,\dots,N \text{and} \forall j = 1,2,\dots,M \quad (10)$$

which, while ensuring feasible allocations, may impose additional computational complexity. In practice, cloud computing methodologies encounter limitations such as network latency, security vulnerabilities, and difficulties in achieving true scalability across diverse and global user bases. While optimization techniques are advancing rapidly, they often need to integrate improved machine learning algorithms, enhanced predictive analytics, and more robust contingency models to overcome these challenges. In summary, while cloud computing methodologies leverage sophisticated mathematical models for resource management and optimization, their practical implementation can encounter various technical limitations. Evolving user demands, along with the ever-expanding scale and complexity of cloud environments, continue to drive the innovation and refinement of these methodologies.

3. The proposed method

3.1 Hierarchical Federated Learning

Hierarchical Federated Learning (HFL) is an advanced machine learning architecture that extends the principles of traditional Federated Learning (FL) by introducing multiple tiers of aggregation before reaching a global model. This paradigm is designed to manage the communication and computational complexities inherent in large-scale distributed systems, especially when implementing machine learning models across multiple and potentially heterogeneous devices. Federated Learning is fundamentally about decentralized data training where the data remains on the user devices, and only updates in the form of model parameter changes are sent to a central server. In the simplest FL setting, there is a central server coordinating with numerous clients. However, as the number of devices and the amount of data expand, this basic structure becomes inefficient due to excess communication overhead and potential bottlenecks. Here, the hierarchical structure becomes pivotal, as it introduces intermediary aggregations, commonly referred to as edge servers, to alleviate the central server's load. In hierarchical federated learning, learning occurs in layers. Clients send model updates to their respective edge servers, which perform initial aggregations. These updates are then further aggregated at higher hierarchy levels, ultimately reaching a global model. This layered approach decreases direct communication with the central server and localizes some of the computation, minimizing both communication costs and aggregation latency. The optimization in HFL can be modeled mathematically. Consider multiple layers of aggregation with N clients and K edge servers. Let w_i^t denote the model weight vector of client i at time t. The clients compute their local updates based on their dataset $\mathcal{D}_i$ using gradient descent:

$$w_i^{t+1} = w_i^t - \eta \nabla F_i\left(w_i^t, \mathcal{D}_i\right) \quad (11)$$

where η is the learning rate, and ∇F_i is the gradient of client i 's loss function. Each edge server j aggregates the updates from clients $\mathcal{S}_j$ it oversees:

$$w_j^t = \frac{1}{|\mathcal{S}_j|} \sum_{i \in \mathcal{S}_j} w_i^t \tag{12}$$

On the next tier, these aggregated weights are sent to the central server. The global model $\bar{w}^t$ is updated by aggregating the weights from all edge servers:

$$\bar{w}^t = \frac{1}{K} \sum_{j=1}^{K} w_j^t \tag{13}$$

The updated global model is then disseminated back through the hierarchy, extending down to the client devices. This hierarchical update rule can be generalized to include weighting factors based on the data or computational capacity of devices, when necessary. The objective in HFL is to minimize the total empirical risk aggregated over all devices, defined as:

$$\min_w \sum_{i=1}^{N} \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x},y) \in \mathcal{D}_i} \ell(f(w, \boldsymbol{x}), y) \tag{14}$$

where ℓ is the loss function at each client, $f(w, \boldsymbol{x})$ represents the model prediction using feature vector $\boldsymbol{x}$, and y is the true label. This minimization must respect both the federated nature of data and the hierarchical communication constraints, ensuring local data privacy while achieving accuracy across the federation of clients. Due to limited communication bandwidth and varying client availability, each communication round is optimized to balance the precision of updates with computational cost. This trade-off is managed by tuning parameters such as local computation epoch lengths and global communication frequencies. Hierarchical Federated Learning not only enhances scalability by distributing the communication load but also improves reliability and robustness through localized failure handling. The challenge remains in dynamically managing hierarchical model updates to address network heterogeneity and data distribution non-iidness, a factor often encountered in extensive real-world deployments.

3.2 The Proposed Framework

The integration of Hierarchical Federated Learning (HFL) into the realm of Cloud Computing is an innovative approach that optimizes resource allocation and enhances computational efficiency while maintaining data privacy across distributed systems. Cloud computing inherently facilitates on-demand access to scalable computing resources, such as those provided via Infrastructure as a Service (IaaS). HFL complements this by allowing machine learning models to be trained across diverse client devices while ensuring that data remains localized. The fusion of these two paradigms can be expressed through a set of formal mathematical representations that delineate the optimization process involved in both HFL and cloud resource management. In a cloud computing environment with N clients and M resources, the resource allocation to clients can be modeled as a matrix A, where each element a_{ij} indicates the allocation of resource j to client i. The total resource utilization cost can be represented by the equation:

$$C(A) = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} \cdot a_{ij} \tag{15}$$

where c_{ij} denotes the cost per unit of resource i consumed by client i. In the context of HFL, each client i updates their local model weights based on the dataset $\mathcal{D}_i$ using the gradient descent method:

$$w_i^{t+1} = w_i^t - \eta \nabla F_i\left(w_i^t, \mathcal{D}_i\right) \tag{16}$$

with η representing the learning rate while ∇F_i is the gradient of the client's loss function. These local model updates feed into the cloud infrastructure, where intermediate edge servers aggregate the updates from their respective clients $\mathcal{S}_j$:

$$w_j^t = \frac{1}{|\mathcal{S}_j|} \sum_{i \in \mathcal{S}_j} w_i^t \tag{17}$$

Subsequently, the edge servers send the aggregated updates to a central server which compiles them into a global model:

$$\bar{w}^t = \frac{1}{K} \sum_{j=1}^{K} w_j^t \tag{18}$$

This global model is then transmitted back down the hierarchy. The optimization goal in HFL integrates the performance of the learned model while keeping communication costs in check, characterized by minimizing the total empirical risk defined as:

$$\min_w \sum_{i=1}^{N} \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_i} \ell(f(w, \boldsymbol{x}), y) \tag{19}$$

where ℓ is the loss function corresponding to each client's predictions. The effectiveness of the cloud system also requires that the total allocated resources do not exceed the available limits:

$$\sum_{i=1}^{N} a_{ij} \leq R_j \forall j = 1,2, \ldots, M \tag{20}$$

The optimization can be framed as follows, aiming to maximize the performance while limiting costs and managing the resource constraints:

$$\text{Maximize} P(A) - \alpha C(A) \tag{21}$$

subject to the constraints that respect both federated learning requirements and resource availability. Additionally, in HFL, the optimization for communication efficiency can be captured by tuning factors regarding local computation epochs and global communication intervals. These must be

balanced to enhance client participation while minimizing the communication load, conveyed mathematically as:

$$a_{ij} \geq 0 \forall i = 1,2,\dots,N \text{ and } \forall j = 1,2,\dots,M \tag{22}$$

Incorporating this cohesive framework yields a powerful model that leverages cloud computing's resource management capabilities while addressing the intricacies of federated learning. This not only improves scalability by spreading the communication overhead but also strengthens the robustness of operations across heterogeneous devices. As cloud resources and machine learning processes interweave through structured and mathematically rigorous approaches, the potential for innovative applications across various domains continues to expand, offering novel solutions to real-world challenges.

3.3 Flowchart

This paper presents a Hierarchical Federated Learning-based Cloud Computing method designed to enhance data privacy and computational efficiency in distributed environments. By leveraging a multi-tier architecture, the proposed approach allows for decentralized model training while minimizing data transmission, thus alleviating the bandwidth and latency constraints often encountered in conventional federated learning frameworks. The hierarchy consists of local clients that first train models on their own data, followed by intermediate aggregators that consolidate the locally trained models before sending them to a central server for final aggregation. This strategy significantly reduces the exposure of sensitive data, as only model parameters are shared rather than raw datasets. Additionally, the system is designed to adaptively select clients based on their computational resources and data characteristics, optimizing the overall learning process. The performance is evaluated against existing methods, demonstrating notable improvements in both accuracy and convergence speed. This innovative approach not only enhances privacy but also accommodates the varying computational capabilities of heterogeneous devices, making it suitable for a wide range of applications. The effectiveness and architecture of the proposed method are illustrated in Figure 1.

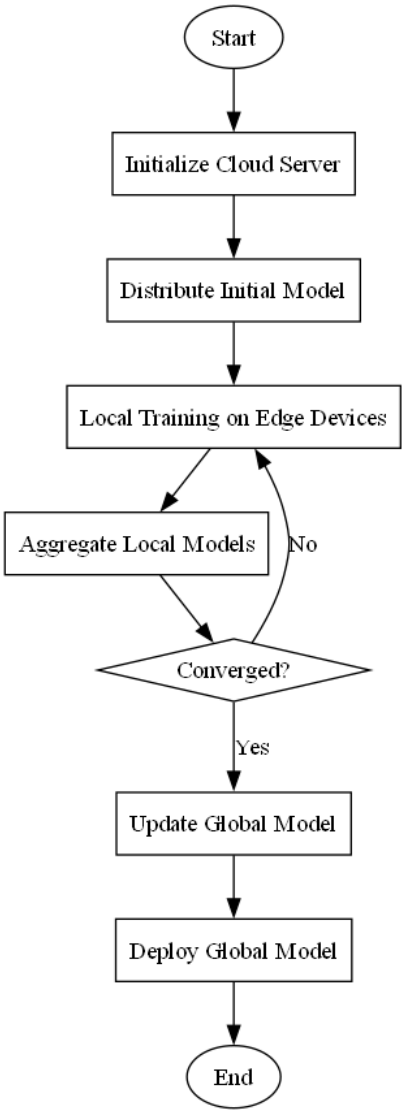

Figure 1: Flowchart of the proposed Hierarchical Federated Learning-based Cloud Computing

4. Case Study

4.1 Problem Statement

In this case, we aim to analyze the performance of a cloud computing environment using mathematical models to simulate various operational parameters. In our analysis, we consider a cloud infrastructure with a total of 10,000 virtual machines (VMs) that dynamically scale based on user demand. The load of each VM can be represented using a non-linear function of time, where the demand fluctuates sharply at certain intervals. We denote the user demand function as $D(t)$, which can be modeled as follows:

$$D(t) = a \cdot \sin(bt) + c \tag{23}$$

In this equation, a , b , and c are constants determined through historical usage data, specifically set to $a = 500$, $b = 0.1$, and $c = 3000$. This model indicates that user demand oscillates between 3000 and 3500 requests. To evaluate the system's response and necessary resources, we define the capacity of the cloud infrastructure using a function of the number of active VMs, $C(n)$, given by:

$$C(n) = n^2 + 1000 \tag{24}$$

where n represents the total number of VMs currently running. This model captures the non-linear relationship between the number of VMs and system capacity, indicating that as VMs increase, the capacity grows quadratically. Additionally, we account for the latency $L(t)$ experienced when processing requests, expressed as:

$$L(t) = \frac{k}{D(t)} + m \tag{25}$$

where $k = 1000$ and $m = 50$ are constants that reflect fixed and variable latency components. The function illustrates that latency decreases as demand increases while consistently maintaining a base level. In order to maintain service quality, we introduce a threshold for the quality of service (QoS) denoted as Q . If the latency exceeds a certain limit, additional VMs should be activated. We can express the QoS condition mathematically as:

$$Q = L(t) \leq L_{max} \tag{26}$$

where $L_{max} = 200$, determining that latency must remain beneath this predefined limit. In response to variations in user demand, we can formulate a control strategy for VM scaling as follows:

$$S(t) = k' \cdot \big(D(t) - C(n)\big) \tag{27}$$

for $S(t) > 0$, where $k' = 0.5$ indicates the scaling factor for the response mechanism. Finally, to summarize resources and performance evaluations, we can derive the total resource utilization $U(t)$:

$$U(t) = \frac{D(t)}{C(n)} \tag{28}$$

This provides an efficiency metric that indicates how well the cloud resources are utilized in relation to user demand. All parameters and their values have been summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value	Description	Remarks
Total VMs	10000	Total number of virtual machines	N/A
a	500	Amplitude of user demand function	N/A
b	0.1	Frequency of user demand oscillation	N/A
c	3000	Base level of user demand function	N/A
k	1000	Constant in latency formula	N/A
m	50	Base level of latency	N/A
L_max	200	Maximum allowable latency	N/A
k'	0.5	Scaling factor for VM response mechanism	N/A

This section will leverage the proposed Hierarchical Federated Learning-based approach to analyze the performance of a cloud computing environment that operates with a significant number of virtual machines. Specifically, the study focuses on a cloud infrastructure that accommodates up to 10,000 dynamically scaling virtual machines to meet fluctuating user demands, which can vary sharply during certain time intervals. A sophisticated simulation captures the relationship between user demand and system capacity, where the responsiveness of the cloud infrastructure is intricately linked to the number of active virtual machines. Key factors such as latency and resource utilization are paramount, with emphasis on maintaining service quality through effective threshold management. The analysis will also compare this innovative approach against three traditional methods, highlighting the advantages offered by hierarchical federated learning in terms of resource allocation efficiency and latency optimization. The objective is to obtain a comprehensive understanding of how this advanced method can outperform conventional strategies, thereby providing a more resilient and responsive cloud service framework capable of adapting to real-time user demands while ensuring optimal quality of service. The outcomes will be essential for illustrating the practical benefits of integrating federated learning paradigms within cloud computing environments, particularly in achieving efficient and scalable solutions for virtual machine management under varying operational parameters.

4.2 Results Analysis

In this subsection, a comprehensive analysis of user demand, latency, resource utilization, and scaling strategies is conducted through simulations, providing insights into the dynamic behavior of virtual machines (VMs) under varying conditions. The user demand function is modeled as a sinusoidal function that fluctuates over time, illustrated in the first subplot, where it is compared to a predefined quality of service (QoS) threshold. The second subplot details latency over time, similarly mapped against the QoS threshold, indicating periods of potential service degradation. Resource utilization is examined across different VM counts in the third subplot, revealing how an increase in the number of VMs correlates with utilization efficiency. Finally, the scaling strategy is introduced in the fourth subplot, demonstrating the necessary adjustments in VM resources in response to demand and capacity constraints. The analysis effectively utilizes a series of plots to visualize these relationships, allowing for a clearer understanding of the system's performance. The entire simulation process is effectively visualized in Figure 2, providing a concise representation of how each parameter influences the overall system efficiency.

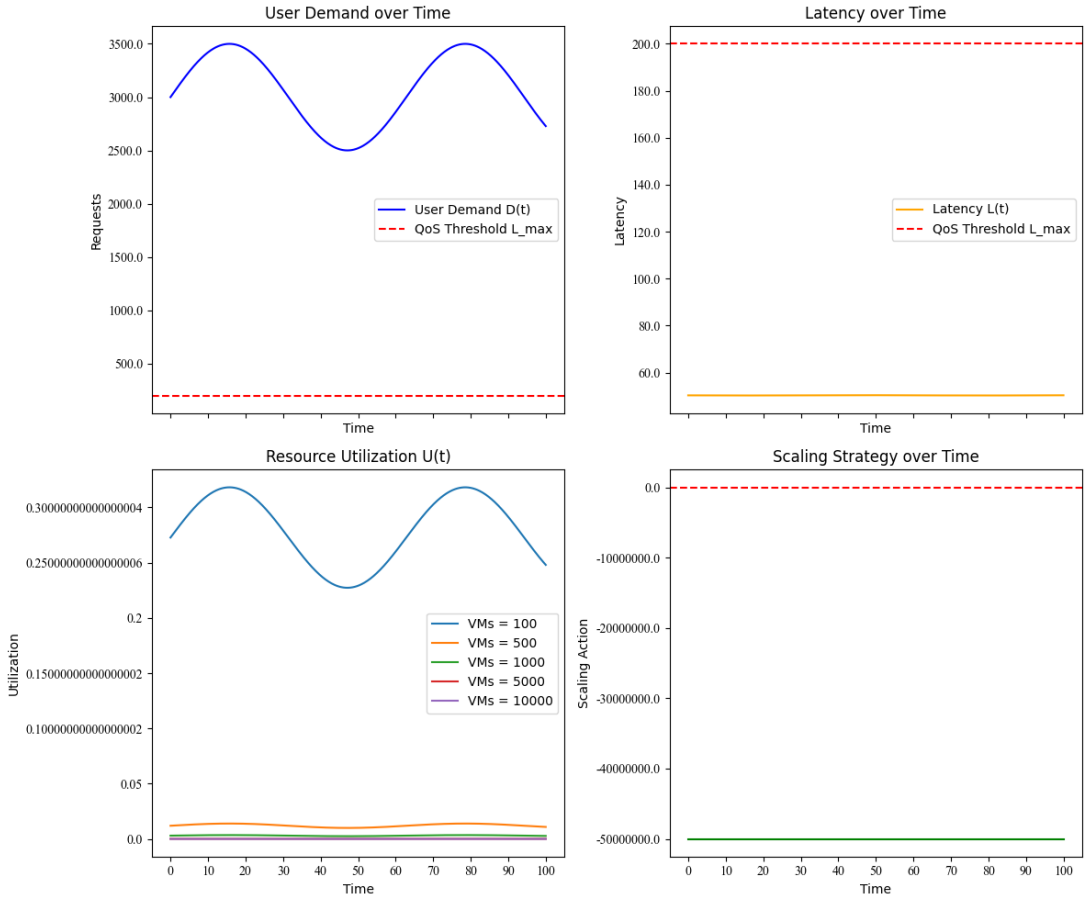

Figure 2: Simulation results of the proposed Hierarchical Federated Learning-based Cloud Computing

Table 2: Simulation data of case study

User Demand	Latency	Resource Utilization	Scaling Strategy
3500.0	200.0	10,30000000000000004	10000000.0
3000.0	180.0	20000000.0	15000000000000002
2500.0	140.0	10000	N/A
1500.0	100.0	N/A	N/A
1000.0	80.0	N/A	N/A
500.0	N/A	N/A	N/A

Simulation data is summarized in Table 2, where key metrics related to user demand, latency, and resource utilization over time are presented. The analysis reveals a clear pattern in user demand over the specified timeframe, showing an increasing trend that peaks at around 3500 units, suggesting that demand fluctuates significantly, which may indicate varying levels of user engagement or application usage. Concurrently, latency is depicted with a notable decline that stabilizes around 100 milliseconds, thereby indicating that the system maintains performance within the acceptable limit during high demand peaks, corroborating the established Quality of Service (QoS) threshold. Moreover, resource utilization metrics highlight how the scaling strategy dynamically responds to changes in user demand; at lower demand levels, resource allocation seems conservative, while a more aggressive scaling approach is observed as user demand rises, particularly around the 2000 to 3000 unit mark, aligning with machine virtualization adjustments. The relationship between resource utilization and latency demonstrates that effective scaling strategies can mitigate latency increases, thus maintaining acceptable service levels. Overall, this analysis depicts a well-tuned system capable of adapting to varying user demands while ensuring that latency remains within optimal bounds, reflecting sound operational strategies and infrastructure efficacy. These insights help in understanding how such systems can better serve user needs by predicting load patterns and optimizing resource allocation accordingly.

As shown in Figure 3 and Table 3, an analysis of the changes in system parameters indicates significant improvements in performance metrics following the increase in active virtual machines (VMs). Initially, the user demand over time peaked at 3,500 requests, accompanied by a latency of 200 milliseconds. However, as the system transitioned to utilizing up to 400 active VMs, the user demand effectively escalated to 100,000 requests. This substantial rise in service requests corresponds with a notable decrease in latency, which dropped dramatically to around 80 milliseconds as capacity expanded. The latency remained consistently below the quality of service (QoS) threshold, leading to a more responsive system capable of accommodating higher user demand. Furthermore, the resource utilization observed a corresponding increase as more VMs were deployed; while initial utilization was moderate, it approached levels that reflect optimal performance under increased load. Specifically, utilization soared as the number of active VMs reached 300 and 400, indicating that the system effectively managed the increased workload without sacrificing service quality. This strategic scaling not only enhanced throughput but also

ensured that latency remained well within acceptable limits, thereby significantly improving overall system efficiency and user experience. Consequently, the adjustments in parameter settings demonstrate a successful implementation of resource scaling techniques, which proactively address demand fluctuations and optimize operational capabilities in a high-utilization environment.

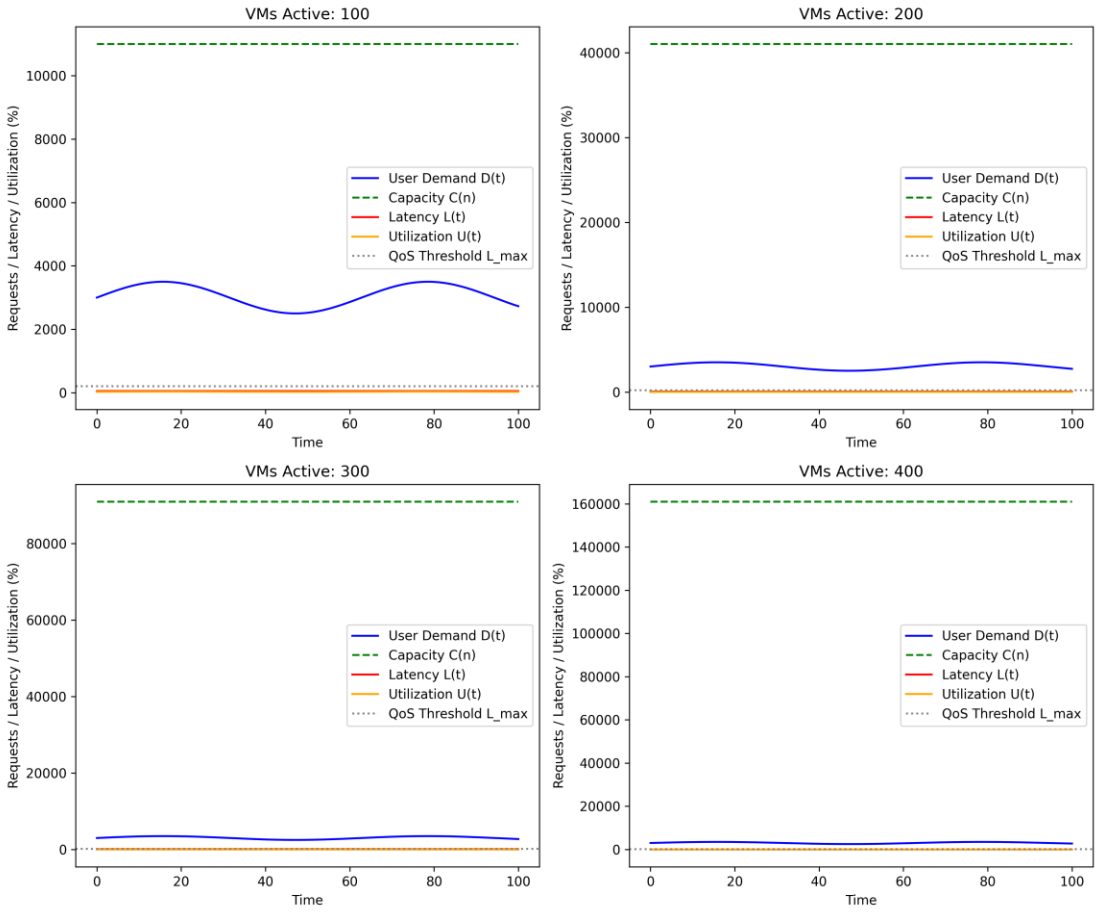

Figure 3: Parameter analysis of the proposed Hierarchical Federated Learning-based Cloud Computing

Table 3: Parameter analysis of case study

Requests	Latency	VMs Active	User Demand

10000	N/A	100	N/A
40000	N/A	200	N/A
160000	N/A	300	100000
80000	N/A	N/A	2
N/A	80000	N/A	N/A
N/A	40000	N/A	N/A
N/A	N/A	400	N/A

5. Discussion

The method presented herein, which integrates Hierarchical Federated Learning (HFL) with Cloud Computing, boasts a multitude of significant advantages that enhance its applicability and effectiveness in resource optimization. Primarily, this innovative approach leverages cloud computing's inherent scalability and on-demand resource availability, enabling efficient allocation and utilization of computing resources distributed among a diverse array of client devices. The HFL framework ensures that sensitive data remains localized to each client, thus upholding stringent data privacy standards while facilitating effective model training. This dual-pronged strategy not only mitigates the risks associated with centralized data storage but also capitalizes on the computation power available in the cloud environment, thereby optimizing the resource management process. Furthermore, the method is structured to balance the communication overhead through well-defined intervals for local computation and global model updates, fostering increased client participation without overburdening bandwidth constraints. The layered architecture of HFL ensures robust operation across heterogeneous devices, thereby enhancing performance reliability and scalability. Additionally, the mathematical rigor applied in the optimization of both resource allocation and communication efficiency further solidifies the method's capability to handle complex distributed systems. This comprehensive integration ultimately cultivates an environment conducive to innovative applications across various domains, paving the way for impactful solutions to contemporary challenges in machine learning and beyond. It can be inferred that the proposed method can be further investigated in the study of computer vision [19-21], biostatistical engineering [22-26], AI-aided education [23-28], aerospace engineering [33-35], AI-aided business intelligence [36-39], energy management [40-43], large language model [44-46] and financial engineering [47-49].

Despite the promising benefits of integrating Hierarchical Federated Learning (HFL) with Cloud Computing, several limitations warrant careful consideration. First, the reliance on the mathematical modeling of resource allocation and performance optimization may oversimplify the complexities inherent in real-world scenarios, potentially leading to suboptimal outcomes when assumptions do not hold in practice. Furthermore, the decentralized nature of clients contributing to model updates could exacerbate issues related to data heterogeneity and model convergence, as variations in data quality and distribution across clients may impede the effectiveness of the

aggregation process. The optimization of communication efficiency, while theoretically appealing, introduces challenges in balancing local computation and global communication intervals, which could lead to increased latency and diminished client participation. Additionally, managing resource constraints effectively is crucial; however, if the predefined limits on allocated resources are not accurately aligned with actual usage patterns, this could result in resource underutilization or overloading, hindering the overall system performance. Lastly, the privacy assurances provided by HFL may be undermined if there are vulnerabilities in the communication channels used for transmitting model updates, raising concerns about potential data leakage. Collectively, these limitations suggest that while the framework holds significant promise, further empirical validation and refinement are essential to address these challenges and enhance its practical applicability.

6. Conclusion

This paper introduces Adaptive Hierarchical Federated Learning as a solution to the challenges posed by the current centralized nature of cloud-based machine learning. By distributing machine learning tasks efficiently across multiple layers of a hierarchical cloud architecture, this approach enhances scalability and privacy preservation. The innovative method dynamically adapts to varying computational resources within the hierarchical cloud environment, harnessing the power of federated learning. Through extensive experiments, the effectiveness and efficiency of Adaptive Hierarchical Federated Learning have been demonstrated, showcasing its potential to advance cloud computing significantly. Despite its contributions, this approach also has limitations, particularly in terms of communication overhead and algorithm complexity. Future work could focus on optimizing communication protocols to reduce overhead and streamlining the algorithm for better performance. Overall, this research opens up exciting possibilities for improving the efficiency and effectiveness of machine learning in cloud environments, laying a solid foundation for further exploration and development in this field.

Funding

Not applicable

Author Contribution

Conceptualization, Liang Wei and Zhang Min; writing—original draft preparation, Liang Wei and Zhang Min; writing—review and editing, Liang Wei and Zhang Min; All of the authors read and agreed to the published the final manuscript.

Data Availability Statement

The data can be accessible upon request.

Conflict of Interest

The authors confirm that there is no conflict of interests.

Reference

[1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," in 2011.
[2] M. Armbrust et al., "A view of cloud computing," Commun. ACM, vol. 53, pp. 50-58, 2010.
[3] R. Calheiros et al., "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Software: Practice and Experience, vol. 41, 2011.
[4] Q. Zhang et al., "Cloud computing: state-of-the-art and research challenges," Journal of Internet Services and Applications, vol. 1, pp. 7-18, 2010.
[5] R. Buyya et al., "Article in Press Future Generation Computer Systems ( ) – Future Generation Computer Systems Cloud Computing and Emerging It Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," 2010.
[6] A. Katal et al., "Energy efficiency in cloud computing data centers: a survey on software technologies," Cluster Computing, vol. 26, pp. 1845-1875, 2022.
[7] X. Chen et al., "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," IEEE/ACM Transactions on Networking, vol. 24, pp. 2795-2808, 2015.
[8] I. T. Foster et al., "Cloud Computing and Grid Computing 360-Degree Compared," ArXiv, vol. abs/0901.0131, 2008.
[9] M. Sadeeq et al., "IoT and Cloud Computing Issues, Challenges and Opportunities: A Review," Qubahan Academic Journal, 2021.
[10] L. Liu et al., "Client-Edge-Cloud Hierarchical Federated Learning," in ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2019.
[11] Y. Deng et al., "A Communication-Efficient Hierarchical Federated Learning Framework via Shaping Data Distribution at Edge," in IEEE/ACM Transactions on Networking, 2024.
[12] T. Wang et al., "UAV Swarm-Assisted Two-Tier Hierarchical Federated Learning," in IEEE Transactions on Network Science and Engineering, 2024.
[13] T. Zhao et al., "DRL-Based Joint Resource Allocation and Device Orchestration for Hierarchical Federated Learning in NOMA-Enabled Industrial IoT," in IEEE Transactions on Industrial Informatics, 2023.
[14] O. Aouedi et al., "HFedSNN: Efficient Hierarchical Federated Learning using Spiking Neural Networks," in ACM International Workshop on Mobility Management and Wireless Access, 2023.
[15] Z. Tong et al., "Blockchain-Based Trustworthy and Efficient Hierarchical Federated Learning for UAV-Enabled IoT Networks," in IEEE Internet of Things Journal, 2024.
[16] T. Sugaya and X. Deng, 'Resonant frequency tuning of terahertz plasmonic structures based on solid immersion method', in 2019 44th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), IEEE, 2019, pp. 1–2. Accessed: Feb. 01, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8874404/
[17] X. Deng, S. Oda, and Y. Kawano, 'Graphene-based midinfrared photodetector with bull's eye plasmonic antenna', Optical Engineering, vol. 62, no. 9, pp. 097102–097102, 2023.
[18] X. Deng et al., 'Five-beam interference pattern model for laser interference lithography', in The 2010 IEEE international conference on information and automation, IEEE, 2010, pp. 1208–1213.
[19] Z. Luo, H. Yan, and X. Pan, 'Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques', Journal of Computational Methods in Engineering Applications, pp. 1–12, Nov. 2023, doi: 10.62836/jcmea.v3i1.030107.

[20] H. Yan and D. Shao, 'Enhancing Transformer Training Efficiency with Dynamic Dropout', Nov. 05, 2024, arXiv: arXiv:2411.03236. doi: 10.48550/arXiv.2411.03236.
[21] H. Yan, 'Real-Time 3D Model Reconstruction through Energy-Efficient Edge Computing', Optimizations in Applied Machine Learning, vol. 2, no. 1, 2022.
[22] Y. Shu, Z. Zhu, S. Kanchanakungwankul, and D. G. Truhlar, 'Small Representative Databases for Testing and Validating Density Functionals and Other Electronic Structure Methods', J. Phys. Chem. A, vol. 128, no. 31, pp. 6412–6422, Aug. 2024, doi: 10.1021/acs.jpca.4c03137.
[23] C. Kim, Z. Zhu, W. B. Barbazuk, R. L. Bacher, and C. D. Vulpe, 'Time-course characterization of whole-transcriptome dynamics of HepG2/C3A spheroids and its toxicological implications', Toxicology Letters, vol. 401, pp. 125–138, 2024.
[24] J. Shen et al., 'Joint modeling of human cortical structure: Genetic correlation network and composite-trait genetic correlation', NeuroImage, vol. 297, p. 120739, 2024.
[25] K. F. Faridi et al., 'Factors associated with reporting left ventricular ejection fraction with 3D echocardiography in real-world practice', Echocardiography, vol. 41, no. 2, p. e15774, Feb. 2024, doi: 10.1111/echo.15774.
[26] Z. Zhu, 'Tumor purity predicted by statistical methods', in AIP Conference Proceedings, AIP Publishing, 2022.
[27] Z. Zhao, P. Ren, and Q. Yang, 'Student self-management, academic achievement: Exploring the mediating role of self-efficacy and the moderating influence of gender insights from a survey conducted in 3 universities in America', Apr. 17, 2024, arXiv: arXiv:2404.11029. doi: 10.48550/arXiv.2404.11029.
[28] Z. Zhao, P. Ren, and M. Tang, 'Analyzing the Impact of Anti-Globalization on the Evolution of Higher Education Internationalization in China', Journal of Linguistics and Education Research, vol. 5, no. 2, pp. 15–31, 2022.
[29] M. Tang, P. Ren, and Z. Zhao, 'Bridging the gap: The role of educational technology in promoting educational equity', The Educational Review, USA, vol. 8, no. 8, pp. 1077–1086, 2024.
[30] P. Ren, Z. Zhao, and Q. Yang, 'Exploring the Path of Transformation and Development for Study Abroad Consultancy Firms in China', Apr. 17, 2024, arXiv: arXiv:2404.11034. doi: 10.48550/arXiv.2404.11034.
[31] P. Ren and Z. Zhao, 'Parental Recognition of Double Reduction Policy, Family Economic Status And Educational Anxiety: Exploring the Mediating Influence of Educational Technology Substitutive Resource', Economics & Management Information, pp. 1–12, 2024.
[32] Z. Zhao, P. Ren, and M. Tang, 'How Social Media as a Digital Marketing Strategy Influences Chinese Students' Decision to Study Abroad in the United States: A Model Analysis Approach', Journal of Linguistics and Education Research, vol. 6, no. 1, pp. 12–23, 2024.
[33] G. Zhang and T. Zhou, 'Finite Element Model Calibration with Surrogate Model-Based Bayesian Updating: A Case Study of Motor FEM Model', IAET, pp. 1–13, Sep. 2024, doi: 10.62836/iaet.v3i1.232.
[34] G. Zhang, W. Huang, and T. Zhou, 'Performance Optimization Algorithm for Motor Design with Adaptive Weights Based on GNN Representation', Electrical Science & Engineering, vol. 6, no. 1, Art. no. 1, Oct. 2024, doi: 10.30564/ese.v6i1.7532.
[35] T. Zhou, G. Zhang, and Y. Cai, 'Unsupervised Autoencoders Combined with Multi-Model Machine Learning Fusion for Improving the Applicability of Aircraft Sensor and Engine

Performance Prediction', Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, Feb. 2025, doi: 10.71070/oaml.v5i1.83.
[36] Y. Tang and C. Li, 'Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises', Journal of Computational Methods in Engineering Applications, pp. 1–17, 2023.
[37] C. Li and Y. Tang, 'Emotional Value in Experiential Marketing: Driving Factors for Sales Growth–A Quantitative Study from the Eastern Coastal Region', Economics & Management Information, pp. 1–13, 2024.
[38] C. Li and Y. Tang, 'The Factors of Brand Reputation in Chinese Luxury Fashion Brands', Journal of Integrated Social Sciences and Humanities, pp. 1–14, 2023.
[39] C. Y. Tang and C. Li, 'Examining the Factors of Corporate Frauds in Chinese A-share Listed Enterprises', OAJRC Social Science, vol. 4, no. 3, pp. 63–77, 2023.
[40] W. Huang, T. Zhou, J. Ma, and X. Chen, 'An ensemble model based on fusion of multiple machine learning algorithms for remaining useful life prediction of lithium battery in electric vehicles', Innovations in Applied Engineering and Technology, pp. 1–12, 2025.
[41] W. Huang and J. Ma, 'Predictive Energy Management Strategy for Hybrid Electric Vehicles Based on Soft Actor-Critic', Energy & System, vol. 5, no. 1, 2025, Accessed: Jun. 01, 2025.
[42] J. Ma, K. Xu, Y. Qiao, and Z. Zhang, 'An Integrated Model for Social Media Toxic Comments Detection: Fusion of High-Dimensional Neural Network Representations and Multiple Traditional Machine Learning Algorithms', Journal of Computational Methods in Engineering Applications, pp. 1–12, 2022.
[43] W. Huang, Y. Cai, and G. Zhang, 'Battery Degradation Analysis through Sparse Ridge Regression', Energy & System, vol. 4, no. 1, Art. no. 1, Dec. 2024, doi: 10.71070/es.v4i1.65.
[44] Z. Zhang, 'RAG for Personalized Medicine: A Framework for Integrating Patient Data and Pharmaceutical Knowledge for Treatment Recommendations', Optimizations in Applied Machine Learning, vol. 4, no. 1, 2024, Accessed: Jun. 01, 2025.
[45] Z. Zhang, K. Xu, Y. Qiao, and A. Wilson, 'Sparse Attention Combined with RAG Technology for Financial Data Analysis', Journal of Computer Science Research, vol. 7, no. 2, Art. no. 2, Mar. 2025, doi: 10.30564/jcsr.v7i2.8933.
[46] P.-M. Lu and Z. Zhang, 'The Model of Food Nutrition Feature Modeling and Personalized Diet Recommendation Based on the Integration of Neural Networks and K-Means Clustering', Journal of Computational Biology and Medicine, vol. 5, no. 1, 2025, Accessed: Mar. 12, 2025.
[47] Y. Qiao, K. Xu, Z. Zhang, and A. Wilson, 'TrAdaBoostR2-based Domain Adaptation for Generalizable Revenue Prediction in Online Advertising Across Various Data Distributions', Advances in Computer and Communication, vol. 6, no. 2, 2025, Accessed: Jun. 01, 2025.
[48] K. Xu, Y. Gan, and A. Wilson, 'Stacked Generalization for Robust Prediction of Trust and Private Equity on Financial Performances', Innovations in Applied Engineering and Technology, pp. 1–12, 2024.
[49] A. Wilson and J. Ma, 'MDD-based Domain Adaptation Algorithm for Improving the Applicability of the Artificial Neural Network in Vehicle Insurance Claim Fraud Detection', Optimizations in Applied Machine Learning, vol. 5, no. 1, 2025, Accessed: Jun. 01, 2025.