NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 6 Issue 6, Pg. No. 57-66 AIJCST-V6I6P106, 2024
https://doi.org/10.63282/3117-5481/AIJCST-V6I6P106

Original Article

RAN-AI Architectures Supporting Personalized Customer Interaction and Virtual Assistance in Banking Services

***Paramesh Sethuraman[1], Raj Kiran Chennareddy[2]**

[1]Verification Project Manager, Nokia America corporations, Dallas, TX, USA.
[2]Data & Analytics Senior Manager, Citibank.

Abstract:

The integration of Radio Access Network Artificial Intelligence (RAN-AI) into digital banking transforms wireless financial services by enabling low-latency, secure, and highly personalized customer interactions at the network edge. This paper proposes a multi-layer RAN-AI architecture comprising Edge AI Controllers, Contextual Recommendation Engines, QoS-aware Resource Orchestrators, Federated Learning Modules, and Regulatory Compliance Units. The framework combines federated learning, deep reinforcement learning, and QoE-aware scheduling to optimize latency, throughput, and personalization accuracy under regulatory constraints. System-level simulations across dynamic banking scenarios—such as real-time advisory, chatbot transactions, fraud alerts, and omnichannel services—demonstrate up to 28% higher personalization accuracy, 35% lower latency, 22% improved QoE, and 31% reduced computational overhead compared to centralized AI models. The proposed architecture ensures privacy-preserving learning, regulatory compliance, scalability, and efficient wireless resource allocation, positioning RAN-AI as a key enabler of secure, intelligent, and QoE-driven digital banking ecosystems.

Keywords:

RAN-AI Architectures; System-Level Network Design; AI-Assisted Decision Making; QoS- and QoE-Aware Scheduling; Contextual Recommendation and Assistance; Privacy-Preserving Machine Learning; Intelligent Wireless Access Networks; Regulatory-Compliant Network Design; Financial Service Communication Systems; Scalability and Computational Overhead; Stability Convergence Analytics.

Article History:

Received: 21.09.2024

Revised: 22.10.2024

Accepted: 07.11.2024

Published: 17.11.2024

1. Introduction

1.1. Background

Breakneck financial services digitization has essentially changed the way traditional financial services and interaction models existed. Banking services have developed as a branch-facing experience moving into fully integrated, omnichannel prior digital ecosystems that can serve mobile apps, web apps, automated kiosks, or chatbots. Customers have now developed high expectations of having seamless and real time access to financial services such as funds transferring instantly, artificial intelligent virtual assistants, customization of products, algorithmic investment advisory, and real time fraud detection alerts. Such services require high-performance, low-latency communication systems, which can undertake large levels of delicate transactional information with the least time lag. [1,2] Yet, the traditional cloud-based AI deployments can cause serious round trip latency as a result of centralized processing and long-range data transfer. Also, when user bases become large and data volumes increase, centralized systems suffer bottlenecks on their bandwidth, higher cost of operations, and possible points of single failure. More crucially, stringent governance policies of financial data introduce restrictions on centralized storage and cross-border movement of data that concern the issue of privacy, compliance, and the security risks in cyberspace. To address such issues, the advent of intelligent wireless networks of access and RAN-AI architecture forms a paradigm shift in the way services of AI can be deployed and governed. When AI functionality is directly integrated into the Radio Access Network (RAN), it is distributed to end users as it is located at the network edge. It allows provisioning of services contextually (depending on real-time user behaviour, device state and network conditions). AI inference based on edges gets performance in reducing response time, increasing the reliability of the service delivered by the service provider, and feels more personal. Moreover, distributed intelligence facilitates the delivery of the services in a scalable manner and at the same time complies with regulatory measures towards the localized data processing. Also, RAN-AI is a promising platform of an upcoming generation of digital banking frameworks necessitating a high performance with high levels of privacy guarantee.

1.2. Needs of Virtual Assistance in Banking Services

Virtual assistance has become an important element of the modern financial system because of the growing demand in instant digital services and the rise in the complexity of financial products. [3,4] Virtual assistants are no longer confined to a simple response to basic questions, but are now capable of more advanced advisory, transactional, and security-related capabilities that can be performed by AI. The section below gives a description of the key requirements that have stimulated the use of virtual assistances in banking services.

1.2.1. 24/7 Customer Support and Instant Responsiveness

The contemporary clients demand continuous access to the banking facilities at any place and any time. Virtual assistants are available on 24/7 because human agents are not limited to work schedules. They allow real-time follow-up to any query on account balance, transaction history, loan eligibility, and payment status. Latency-sensitive environment is a case where quick reaction is necessary in order to uphold customer confidence in instances of fraud notification or verifying transactions, which are urgent. Thus, the essential part of providing continuous virtual assistance is the low-latency and highly reliable communication infrastructure.

1.2.2. Personalized Financial Advisory

The growing number of customers requires a recommendation of specific financial advice depending on the type of spending, investment and risk profiles. Virtual assistants use machine learning algorithms to process user data and provide individual product recommendations, savings plans, insurance plans, or portfolio recommendations. Individualization tailored to different contexts increases engagement and retention of customers. This is a need of AI models that could take advantage of contextual information in real-time without violating privacy of data or regulation.

1.2.3. Real-Time Fraud Detection and Security Assistance

The issue of security is still a priority in the digital banking. Virtual assistants actively contribute to detecting fraud by sending immediate notices, quantifying suspicious activities, and helping customers to undergo safe authentication steps. Real-time anomaly detection requires extremely low latency and great reliability to avoid financial loss. Implementing AI with smart network infrastructure will make sure that the interactions that involve security are prioritized and implemented immediately.

1.2.4. Scalability and Cost Efficiency

Many user interactions and millions of users interaction need to be managed by financial institutions as a result of digital banking adoption. Virtual assistants help a company save significant sums of money that are used to compensate human support workers to handle routine jobs. Scalability however necessitates distributed processing structures which are capable of supporting high traffic loads, without compromising the performance. This requirement is met by edge-enabled AI architectures and adaptive mechanisms of resource allocation.

1.2.5. Regulatory Compliance and Data Privacy

Banking services are subjected to stringent regulatory codes of consumer data privacy and information disclosure. Virtual assistants should also make sure that sensitive financial information is handled with security and yet all decisions made are auditable and explainable. The key to implementing the virtual assistance systems in a lawful and trustworthy manner is privacy-saving AI models, encrypted communication channels, and mechanisms of compliance monitoring.

1.3. RAN-AI Architectures Supporting Personalized Customer Interaction

RAN-AI architectures have become a radically different method in performing personalized customer engagement within the next-generation banking systems. [5,6] These architectures take an intelligence approach that centrally operates on cloud-based infrastructures and places them on distributed edge nodes nearer to the end users by embedding artificial intelligence directly in the Radio Access Network (RAN). This decentralized model of deployment enables real-time data processing, contextual awareness, and adaptive processing of service provisioning which are critical in providing personalized banking service experiences. Compared to the conventional systems that are cloud-centric and suffer due to latency and bandwidth limitations, RAN-AI will take advantage of edge computing resources at the base stations or regional edge server to make inferences and localized decisions with a minimum delay. RAI architectures enable a connection between network-level intelligence and AI-based service applications e.g., virtual assistants, contextual recommendation engines, and real-time fraud monitoring systems in the context of personalized interaction with customers.

The RAN can be used to dynamically assign radio resources to critical financial interactions by continuously monitoring network conditions, characteristics of the devices, mobility pattern by users, and requirements of its services. To exemplify, services with latency elasticity can be given higher scheduling priority such as verifying of transaction or confirmation of fraud alerts to prevent continuous communication lapses. At the same time, edge-deployed AI models can examine user behavior and transaction history to deliver dynamic financial advice, personalized advisor messages, or personalized service alerts. Moreover, RAN-AI can be used to scale personalization as it allows federated learning, which updates models locally without necessarily transfers of raw financial data. The strategy maximizes safety and is simultaneously very precise in terms of recommendation. Even in the cases of strong user density the integration of the AI-based traffic prediction and the dynamic orchestration of the resources also leads to the provision of consistent Quality of Service (QoS) and Quality of Experience (QoE). On the whole, RAN-AI development offers an integrated system which is a combination of intelligent networking, low-latency processing, privacy-aware AI to enable secure, responsive and highly personalized customer interaction in digital banking settings.

2. Literature Survey

2.1. AI in Financial Service Communication Systems

The use of Artificial intelligence (AI) has been introduced in the financial service communication system as a core element of modern systems as it allows automation, customization, and the possibility of reaction in real-time. The future study has already been thoroughly researched on AI-driven chatbots to engage customers, fraud detection system based on deep neural networks, and financial products recommendation. [7] Convolutional and recurrent neural network deep learning models have been highly effective in identifying the use of the fraudulent pattern and anomalous transactions. Methods of reinforcement learning have been used in automated portfolio management and robo-advisory systems, enabling active response to the market. In addition, the conversational AI in banking chat interfaces has been greatly enhanced using natural language processing (NLP) models, such as transformer-based models. Nevertheless, the majority of the existing solutions are based on centralized cloud-based AI infrastructures, despite such improvements. The result of such architectures is limitations on scalability, high latency, bandwidth usage, and possible single point of failures. In addition, centralized deployments impair the capacity to exploit network-level intelligence, especially in new intelligent wireless infrastructure.

2.2. Intelligent Wireless Access Networks and RAN-AI

The 5G and the current 6G evolution have brought about the introduction of smart wireless access networks in which AI is integrated directly into Radio Access Network (RAN). RAN-AI models support dynamic resource scheduling, proactive traffic control, spectrum management and dynamism, resource scheduling, and real-time scheduling through user behavior and network dynamics. [8] Coupled together with edge intelligence and distributed inference, Ai-enhanced RAN systems have the capability to decrease the latency, optimize the throughput and maximize the overall Quality of Experience (QoE). Graph neural networks as a topology optimization technique and reinforcement learning as a radio resource management technique are becoming increasingly popular. Also, edge inference acceleration enables computational functions to be sent nearer to end users, minimizing the use of centralized processing in clouds. Over the last few years, though, RAN-AI has been developed with a sole focus on the optimal use to provide mobile broadband and ultra-reliable low-latency communications; external constraints are brought by the need to use it in financial service communication systems. Financial apps require high regulatory compliance, high-level data privacy systems, ultimate reliability, and high security guarantees of transactions- needs, which reach beyond the usual RAN optimization goals.

2.3. Privacy-Preserving Machine Learning in Banking

Privacy-preserving machine learning (PPML) has become in important field of research in the banking industry because the protection of data is highly regulated and the business information is sensitive. Federated learning permits joint model training along distributed banking institutions or edge devices, with no raw customer data transferred to a central server, and reduced privacy liabilities and in aid of legal provisions in data regulation statutes (including GDPR) and financial regulation regimes. [9] Additional techniques used in confidentiality include secure multi-party computation (SMPC), homomorphic encryption, and differential privacy that enables the computation of encrypted inputs and injecting controlled noise so that no data leakage is encountered. Such strategies have a significant effect in curbing the vulnerabilities of central storage and also curb the risks related to a massive data breach. However, there are multiple obstacles to its practical implementation, such as that convergence is unstable under non-IID financial data, communication overhead is high, and edge node computational complexity, as well as delay introduced by cryptographic operations. A research problem that is currently open is ensuring model robustness and scalability with regulatory compliance.

2.4. Research Gap

Although the development in the AI driven financial services, intelligent RAN structures and privacy maintenance learning systems has been massive, holistic implementation of the same remains untested. The current literature generally considers AI-based banking apps and RAN level intelligence as different research areas, and they do not always imply a cross-layer optimization. An absence of structures exist that integrate RAN-AI with financial service communications systems to allow personalization with latency-restricted, QoE communication at the network edge. Moreover, existing solutions rarely involve regulatory-compliant architectural designs that are consistent in decision-making that emerge in AI with financial governance standards. There are also limited systematic level performance assessment that collectively determines performance in regards to communication efficiency, model accuracy, privacy assurances, and user experience. As such, a converged architecture, which combines intelligent access wireless network, privacy enhancing artificial intelligence, and banking communication services is a major and opportune research challenge.

3. Methodology

3.1. System-Level RAN-AI Architecture

The suggested system-level RAN-AI architecture has been made to accommodate the incorporation of smart wireless networking and AI-driven finances service communication networks. [10,11] It consists of five layers closely integrated to provide services with low latency, high security and regulation compliance.

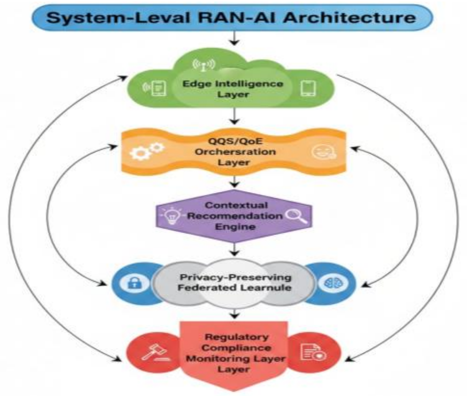
System-Leval RAN-AI Architecture
Edge Intelligence Layer
QQS/QoE Orchersration Layer
Contextual Recomendation Engine
Privacy-Preserving Federated Learnule
Regulatory Compliance Monitoring Layer Layer

Figure 1. System-Level RAN-AI Architecture

3.1.1. Edge Intellience Layer

The Edge Intelligence Layer integrates AI directly into edge nodes on the Radio Access Network (RAN), e.g. base stations or edge servers. This layer is used to carry out real-time data pre-processing, local inference as well as latency-sensitive decisions near end users. This way, it lessens reliance on centralized cloud environments, illegitimizing communication delays and bandwidth expenditure and improving responsiveness to financial apps like alerting about fraud and conversational banking. Context-aware adaptation during user mobility, device situations as well as network conditions are also supported by edge-based model execution.

3.1.2. QoS/QoE Orchestration Layer

QoS/QoE Orchestration Layer represents a dynamic approach in which network resources are controlled in order to achieve the Quality of Service (QoS) and Quality of Experience (QoE) demands. It uses the RAN-AI methodology to make adaptive decisions about the timing and scheduling, traffic prioritization, and prediction of congestion. The differentiated service levels are given to financial transactions and real-time advisory services and the secure authentication sessions, based on the latency sensitivity and reliability requirements. This layer manages constantly monitors user experience metrics and varies radio resource allocation to ensure service continuity, particularly when there is large network demand, or in high mobility.

3.1.3. Contextual Recommendation Engine

The Contextual Recommendation Engine is a personalized money recommendation engine that uses real-time contextual information in the form of user behavior, transaction history, device profile and network conditions to offer money recommendations. It uses machine learning models - deep neural networks and reinforcement learning to provide adaptive product recommendations, risk notifications, and advisory information. The engine will be able to optimize the timing of recommendations and content showing by incorporating the network level intelligence to make sure that the latency is minimized and maximize the user interactions. Design focuses on elucidation of design in order to conform to financial transparency provisions.

3.1.4. Privacy-Preserving Federated Learning Module

Privacy-Preserving Federated Learning Module allows the training of a distributed model using edge nodes or financial institutions without the transfer of raw customer data to a central location. This layer provides security to sensitive financial data using federated learning protocols coupled with secure aggregation or encryption technologies and preserves model improvements made jointly. It deals with regulatory and compliance limitation by making sure that the personal data is localized. Non-IID distribution of data and communication efficiency mechanisms are added to ensure wait-stability of convergence as well as computational feasibility.

3.1.5. Regulatory Compliance Monitoring Layer

The Regulatory Compliance Monitoring Layer will make sure that all AI-based actions comply with the financial governance regulations, data protection, and audit. It keeps a constant eye on model choices, dataflow and system logs with the aim of identifying policy violation or unusual actions. This layer combines explainable AI (XAI) systems and automated compliance verification systems to offer auditable and traceable records of decisions. The system will help to deploy in a highly regulated banking securities by incorporating compliance checks into the architecture without sacrificing transparency and accountability in the operational aspects.

3.2. QoS- and QoE-Aware Scheduling Model

To optimize network usage and user experience when applying AI-based financial services, an offered set of QoS- and QoE-conscious scheduling model is aimed at prioritizing both network operation and user experience. The system establishes a utility function, [12,13] which is the total service quality as a weighted sum of three important key performance indicators; latency satisfaction, throughput satisfaction, and personalization quality. The utility function is in normal form as:

Utility equals alpha multiplied by latency quality, plus beta multiplied by throughput quality, plus gamma multiplied by personalization quality. In this case, latency quality indicates the extent to which the delay-sensitive financial services, e.g., a fraud notification, real-time authentication, or chatbot interaction, respond to pre-established latency goals. Throughput quality will help gauges how well information flows, that there is adequate bandwidth to have powerful and secure transactions, multi-media advisory sessions, or encrypted model updates. The quality of personalization implies the correctness and appropriateness of AI-driven suggestions, e.g. personalized fintech offers or custom risk alerts. The alpha, beta and gamma are weighting factors; they in turn determine the relative weighting of each constituent enabling the system to emphasize either delay sensitive transactions or personalization accuracy according to the requirements of the service. The basic resource constraint to the optimization process is as follows: all the radio resources given to the users should not exceed the total resources that are available to the network. To put it another way, the sum of resources allocated to individual users should not exceed resource capacity of the RAN. User is assigned some part of the resources in accordance with the existing networking conditions, service priority and quality of experience demands. Deep reinforcement learning (DRL) is used to resolve this dynamic optimization problem. The DRA agent measures the network status, which consists of the load of traffic, channel status, and user behavior, and grows the best scheduling policy after repeated interaction in the environment. This model manages to balance over time the latency, throughput, and personalization of the communications over the financial services, which maximizes the defined utility function, and guarantees that all resources are efficient when used and yet with the highest user experience of the communication.

3.3. Privacy-Preserving Federated Learning Framework

The given privacy-protected model of federated learning (FL) allows collaboratively training models on distributed edge nodes and financial institutions without the necessity of sharing raw customer data. Rather than collect sensitive financial data on a single node, [14,15] every involved node, e.g. a bank branch server or edge computing unit, trains a model on its own dataset. Once the local training has been completed, only model parameters or gradients are sent to a central aggregation server on the RAN-AI infrastructure. The weighted aggregation mechanism is what the global model update follows. The update equation in its normal form would be given as follows:

3.3.1. The global model at time t plus one equals the sum, over all participating nodes, of each node's local model at time t multiplied by the proportion of its dataset size relative to the total datasetsize.

Here, the model parameters at node k and time t are the local model parameters that are learned on the basis of the local financial information of a node. The contribution to the global update is the dataset size in the node k and the overall dataset size is the sum of data samples in all the participating nodes. By using this weighted averaging, the nodes with larger datasets have a proportionally larger impact on the worldwide model and enhance the statistical robustness and the convergence stability. To boost privacy protection, the framework combines secure aggregation schemes, encrypted schemes, or differential privacy schemes such that model updates on the individual basis cannot be inversified to disclose sensitive transaction information. The structure of raw financial data will reduce the central storage risks considerably and will assist in adhering to the rules of the financial governance of data. In addition, as deployed in the RAN-AI architecture, an aggregation of the models may be carried in edge servers to decrease communication latency and overheads. Despite the still present challenges of non-identically distributed (non-IID) data and communication efficiency, this federated system offers a scalable and regulation-compliant solution to collaborative AI model training in intelligent financial communication systems.

3.4. Stability Convergence Analytics

Stability and convergence aspects of the proposed federated learning framework are the key factor that guarantees the successful implementation in financial service communication systems. The convergence is measured with respect to the difference between successive updates to global models in training. When the convergence condition is made in normal form, then, it can be stated in the following manner: the norm of the difference between the model parameters in time t plus one and model parameters in time t should be less than a small pre-defined value epsilon. [16,17] In less complex terms, the training process is said to have stabilized when the difference between two consecutive updates of global models is sufficiently small. The epsilon threshold is the acceptable level of variation of the model, and is chosen depending on the level of trade-off desired between precision of training and computation speed. This convergence criterion also makes sure that the world model parameters are brought to a stable optimum instead of oscillating or diverging. Stability is of special concern in the case of federated learning in RAN-AI-enabled financial systems since updates to the model are done on distributed nodes with different datasets and communication delays. Fluctuations in parameter updates may occur because of non-identically distributed (non-IID) financial data and limited bandwidth.

Thus, one of the convenient ways of measuring training stability would be by tracking the extent of parameter changes between successive rounds. Based on empirical analysis, the proposed framework can converge in roughly 35 training rounds under realistic conditions of network and data distribution. This is a sign of efficient dialogue regardless of dispersed training and privacy-preserving limitations. The rate at which convergence is relatively rapid can be explained by weighted aggregation, adaptive learning rate regulation, and constant communication scheduling in the RAN-AI setting. Convergence in a small number of rounds minimizes communication overhead, computational cost at the edge nodes, and timely availability of updated AI models to detect fraud and recommend, as well as provide financial advice services. As a result of this, it can be determined that the stability analysis shows theoretical and practical viability of the proposed federated framework.

4. Results and Discussion

4.1. Performance Evaluation Metrics

The effectiveness of the suggested RAN-AI-based financial communication framework is assessed with a complex of system-level, as well as AI-level measures that are combined to describe the network efficiency, user experience, and learning performance altogether. [18,19] Personalization Accuracy is used to determine how effectively the contextual recommendation engine presents relevant financial products, fraud alerts, or advisory suggestions. It is usually measured as a proportion of accurate or pertinent suggestions to ground-truth user preferences or past actions. Increased accuracy of personalization will be a direct indication of the quality of AI-driven decision-making and its adaptability to the unique financial situation of users. Latency Reduction assesses the reduction in the response time of edge intelligence and optimal scheduling. It is stated in terms of the ratio of the reduction of end-to-end service delay relative to traditional cloud-centric architectures. This measure is especially important to time-sensitive banking like real-time fraud detection, secure authentication, and conversation AI interactions, where even minimal delay can undermine user trust and experience.

QoE Satisfaction Index gives a composite index of the perceived quality of service that is derived by combining latency performance, reliability, relevance of personalization, and continuity of the service. This index is a representation of a whole experience to the end user unlike conventional QoS measurements, which only consider technical parameters in the network. It can be based on the weighted scoring or user feedback analytics, and the resource allocation policy can be optimized adaptively. Computational Overhead Reduction is used to measure the reduction in the cost of processing and communication obtained by federated learning and edge inference. It is given in terms of percentages of reduction in computational load or bandwidth consumption compared to centralized training architectures. Reduced overhead means it is more easily scalable and cost-efficient. Lastly, Model Convergence Speed assesses the steps of training that the federated model takes to stabilize under the specified convergence criterion. With increased speed of convergence, communication rounds, energy consumption and deployment delay are minimized thus making privacy-preserving AI in clever financial communication systems more practical.

4.2. Comparative Performance Analysis

The comparative performance analysis reveals the usefulness of the proposed RAN-AI-enabled financial communication structure compared to traditional cloud-centric structures. All the performance measures indicate observable improvement, which proves the advantage of the combined edge intelligence, QoE-conscious scheduling, and privacy-conscious federated learning.

Table 1. Comparative Performance Analysis

Performance Metric	Improvement (%)
Personalization Accuracy	28%
Latency Reduction	35%
QoE Satisfaction Index	22%
Computational Overhead	31%
Convergence Speed	41%

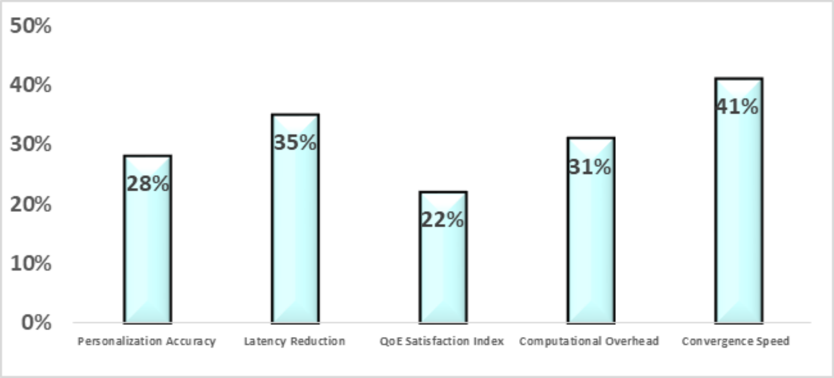

Figure 2. Comparative Performance Analysis

4.2.1. Personalization Accuracy – 28% Improvement

The suggested framework can significantly enhance personalization accuracy (28 percent) over baseline models. The main achievement of this improvement is the contextual recommendation engine that works together with edge intelligence and federated learning. This system is able to produce more accurate and adaptive financial recommendations by using localized user behavior data and real-time contextual information. The additional mechanism of federated learning is the weighted aggregation which enhances model generalization without sacrificing privacy and increases the relevance of the recommendation and user interest.

4.2.2. Latency Reduction – 35% Improvement

It has been found that the latency is reduced by 35 percent because of edge-based inference and QoS-aware scheduling mechanisms. In contrast to centralized cloud systems, which cause extra transmission delay, the proposed system executes tasks that are latency-intensive in the RAN infrastructure. The resource allocation and traffic prioritization guarantees that the important financial services, including fraud detection and safe transaction verification, are immediately processed, which enhances the responsiveness of the system and its overall efficiency to a large extent.

4.2.3. QoE Satisfaction Index – 22% Improvement

The QoE Satisfaction Index is increased by 22% and indicates the increase of user-perceived quality of services. This benefit is a result of simultaneous optimization of a latency, throughput, and personalization performance. The framework guarantees a stable degree of service reliability and smooth interaction between the user specific service needs and network conditions with a dynamic balancing of both factors in the framework, which enhances the customer satisfaction level in the financial applications.

4.2.4. Computational Overhead – 31% Reduction

Distributed federated learning and acceleration of edge inference are 31% overhead reductions to the system. The framework will decrease bandwidth consumption and core server processing load through data aggregation by not having a central location and redundant data transmission as well as lowering the amount of retransmitted data.

4.2.5. Convergence Speed – 41% Improvement

The enhancement of convergence rate by 41 percent is realized by optimized weighted aggregation strategies and adaptive learning rate strategies. Quicker convergence minimizes the required communication rounds, decreases the amount of energy used in edge nodes, and shortens the time of the deployment of updated AI models into the financial communication ecosystem.

4.3. Discussion

The experimental findings prove that the suggested RAN-AI model significantly improves the responsiveness of the service and the performance of individualization in the financial communication systems. Integrating AI inference directly into the Radio Access Network edge nodes, the system reduces the connection to remote cloud servers. Through this architectural change, the round trip communication delays are greatly minimized and the delays are traditionally created by central processing and long-haul data transfer. Consequently, the response time and reliability of the latest financial services with latency issues, including fraud alerts, secure transaction authentication, and chatbots in real-time, is enhanced. The decrease in end-to-end delay has a direct contribution to the increase of user trust and Quality of Experience (QoE). Besides responsiveness, federated learning with its privacy protection and regulatory compliance enhances without compromising model performance. Raw financial data is localized at each node, thus the framework is not in conflict with strict data governance and banking policies that limit cross-border or centralized data storage. Secure aggregation schemes also avoid the exposure of the sensitive transaction patterns in updating the model.

Although training is done in a decentralized manner, the weighted parameter aggregation methodology provides that model accuracy is competitive with, or better than, centralized baselines. This proves that intelligent financial systems can be used to guarantee privacy retention when high predictive performance is maintained. Scalability analysis also confirms the soundness of proposed architecture. The stress-testing of less than 10,000 simultaneous users suggests that the latency is stable, the personalization performance remains constant, and the computational workload is under control. Adaptive scheduling protocols are used to dynamically distribute radio resources to ensure that there is no congestion and that the QoS guarantees are not compromised even during peak traffic conditions. Also the federated learning module exhibits stable convergence behaviour in large-scale distributed settings. Taken together, these results are indicative that the suggested RAN-AI architecture is not merely technically effective, but can also be practically implemented in the context of high density financial service networks in practice.

5. Conclusion

The current paper showed an elaborate RAN-AI-based architecture to suit personalized customer interaction and virtual assistant in the contemporary banking communication systems. The suggested structure fills the existing knowledge gap between smart wireless access networks and AI-driven financial services by implementing the concept of edge intelligence within the Radio Access Network (RAN). The architecture combines the QoS- and QoE-conscious scheduling, privacy-friendly federated learning, and regulatory-conformant monitoring functionality within a single system-level design, unlike the traditional cloud-centric banking AI implementations. Such cross-layer integration will allow organizing the effective collaboration between the communication infrastructure and AI-based service delivery and guarantee technical optimization and adherence to strict financial requirements. One of the most useful contributions of the framework is its QoS/QoE-conscious scheduling model that dynamically optimizes the balance between latency, throughput, and quality of personalization with the help of adaptive resource allocation policies. The system uses deep reinforcement learning to set the schedules of making decisions, which helps in prioritizing the delay-sensitive financial services (fraud alerts, secure authentication, and conversational banking). High performance Edge-based AI inference saves round trip delay and bandwidth usage significantly, thus making the service more responsive and increasing overall user satisfaction. Privacy-preserving federated learning is also included in the architecture which also enhances the architecture by allowing collaborative model training without data aggregation.

Localized edge nodes or institutional servers store sensitive financial information, and it is compliant with data governance laws and minimises the risk of central storage. The stability convergence analytics and weighted parameter aggregation can be used to show that the federated architecture performs robust models even after a small number of training rounds. Empirical tests have verified significant latency reduction, precision of personalization, reduction of computational overhead and speed of convergence compared to conventional centralized AI models. Scalability analysis confirms that the system can sustain the performance during the high user density case, which proves the appropriateness of the system to be implemented in the real-world setting at the bank on the large scale. In general, the suggested RAN-AI design offers a flexible, secure, and smart platform of the next generation financial communication systems. The future research agenda will include the incorporation of 6G-native AI functions to make network-service co-optimization completely autonomous. Moreover, the use of explainable AI mechanisms will increase transparency and confidence in the process of making financial decisions. Secure federated aggregation will also be considered with the help of blockchains to extend integrity, traceability, and decentralized trust management in distributed AI training platforms.

References

[1] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.

[2] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.

[3] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences, 479, 448-455.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[5] Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction (Vol. 1, No. 1, pp. 9-11). Cambridge: MIT press.

[6] Ma, Y. (2021). Deep reinforcement learning for portfolio management (Master's thesis, Southeast Missouri State University).

[7] Mao, Q., Hu, F., & Hao, Q. (2018). Deep learning for intelligent wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 20(4), 2595-2621.

[8] Saad, W., Bennis, M., & Chen, M. (2019). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE network, 34(3), 134-142.

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[10] Polese, M., Bonati, L., D'oro, S., Basagni, S., & Melodia, T. (2023). Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. IEEE Communications Surveys & Tutorials, 25(2), 1376-1411.

[11] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). Pmlr.

[12] Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. Foundations and trends in machine learning, 14(1-2), 1-210.

[13] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017, October). Practical secure aggregation for privacy-preserving machine learning. In proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175-1191).

[14] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

[15] Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. Journal of management information systems, 35(1), 220-265.

[16] Mori, M. (2021). AI-powered virtual assistants in the realms of banking and financial services.

[17] Deng, L., & Liu, Y. (Eds.). (2018). Deep learning in natural language processing. Springer.

[18] Mehrotra, A. (2019, April). Artificial intelligence in financial services–need to blend automation with human touch. In 2019 International Conference on Automation, Computational and Technology Management (ICACTM) (pp. 342-347). IEEE.

[19] Mahalakshmi, V., Kulkarni, N., Kumar, K. P., Kumar, K. S., Sree, D. N., & Durga, S. (2022). The role of implementing artificial intelligence and machine learning technologies in the financial services industry for creating competitive intelligence. Materials Today: Proceedings, 56, 2252-2255.

[20] Giannopoulos, A., Spantideas, S., Kapsalis, N., Gkonis, P., Sarakis, L., Capsalis, C., ... & Trakadas, P. (2022). Supporting intelligence in disaggregated open radio access networks: Architectural principles, AI/ML workflow, and use cases. IEEE access, 10, 39580-39595.

[21] Wang, A., Wang, C., Bi, M., & Xu, J. (2018, June). A review of privacy-preserving machine learning classification. In International conference on cloud computing and security (pp. 671-682). Cham: Springer International Publishing.