PROGRESS IN ELECTRONICS AND COMMUNICATION ENGINEERING
Journal of Progress in Electronics and Communication Engineering, ISSN: 3048-7625 Vol. 2, No. 2, 2025 (pp. 97-103)
RESEARCH ARTICLE
WWW.PECEJOURNAL.COM

Communication-Efficient Machine Learning Architecture for Predicting User Information Needs in Distributed Systems

K. Geetha[1]*, P. Dineshkumar[2]
[1]Professor of Computer Science and Engineering, Excel Engineering college, Erode
[2]Assistant Professor, Department of Information Technology, Sona College of Technology, Salem

KEYWORDS:
Federated learning,
Distributed systems,
Communication efficiency,
Machine learning,
Information retrieval,
System optimization,
Privacy preservation

ARTICLE HISTORY:
Submitted : 07.06.2025
Revised : 13.08.2025
Accepted : 23.09.2025

https://doi.org/10.31838/ECE/02.02.14

Abstract

Overwhelming communication overhead, long synchronisation times, and scaling issues are the major issues that machine learning (ML) models deployed in distributed computing settings have to overcome. These are problems that hamper the effectiveness and responsiveness of the intelligent systems, especially in large-scale information networks whereby the user data are spatially distributed. To address these drawbacks, this research suggests a communication-efficient federated learning (FL) architecture optimised in predicting user information requirements to distributed data settings. The architecture will be based on the localised training of models in individual nodes for example institutional repository or digital library servers thus eliminating the necessity of transferring raw data to the central station and adhering to the provisions of privacy. An adaptive communication scheme hierarchically structured into aggregation mechanism is widely applicable in terms of bandwidth consumption minimization and does not undermine performance of the model convergence and predictive accuracy. The results of experimental validation in an experimental topology comprised of five interconnected nodes confirm a 42 percent saving of communication overhead and 15 percent enhancement of training efficiency over traditional centralised learning systems. In addition, the proposed architecture supports scalability of the system, power efficiency and compliance to privacy, which forms a formidable base on big-scale, smart data infrastructures. The study points out the promise of communication-optimized federated learning as one of the enabling factors of secure, adaptive, and resource-sensitive distributed machine learning ecosystems.

Author e-mail Id: kgeetha.eec@excelcolleges.com, pdineshcs@gmail.com

How to cite this article: Geetha K, Dineshkumar P. **Communication-Efficient Machine Learning Architecture for Predicting User Information Needs in Distributed Systems**. Journal of Progress in Electronics and Communication Engineering Vol. 2, No. 2, 2025 (pp. 97-103).

Introduction

The swift increase in distributed computing and data-driven ecosystems has increased the necessity to develop highly effective machine learning (ML) models that can work in a homogeneous manner in geographically distributed world. Conventional centralized designs of ML systems demand constant aggregation of data across various sources, leading to unwanted communication overheads, communication delays and latency as well as synchronization delays which critically impacts performance and scale in large scale systems.[1, 2] Federated learning (FL) has been proposed as a promising federated paradigm that allows training model in a decentralized manner without access to raw data due to the increasing popularity of data privacy regulations and network heterogeneity.[3, 4] Nevertheless, even though the field of FL is increasingly applied to edge computing, digital libraries, and analytics based on IoT, the technology has constraints regarding the efficiency of its communication, bandwidth, and coordination of the system.[5]

Traditional FL systems focus mainly on accuracy of the models and preservation of privacy in addition to they tend to ignore optimization of communication

and energy efficiency. This is more pronounced in the large-scale information environments where high frequency model rerelease along with multi node synchronisation causes significant network congestion and energy consumption to high frequency.[6, 7] Recent studies have examined gradient compression,[8] adaptive synchronisation[9] as well as quantization strategies to decrease communication load but a unified solution (that combines all these solutions into a scalable and application-specific scheme) is not well developed. As a result, enhanced efficiency of communication and assurance of reliable model convergence and low latency has been identified as a main research direction in federated and distributed learning systems.

In order to deal with these, this paper creates a communication-effective federated learning architecture, which is aimed at predicting user information requirements in distributed data locations. The suggested framework promotes localised training of the models by selective synchronisation and adaptive communication scheduling that enhances tremendously by eliminating redundant exchange of data and maintaining model integrity. It has been experimentally verified using five interconnected nodes that communication overhead has been reduced (42) and training efficiency has been increased (15) relative to traditional centralized learning architectures. This research has contributions that are (1) a scalable federated architecture to support privacy-aware data processing; (2) enhanced communication and energy efficiency through local computation; and (3) the validation of its performance in institutional repositories and digital information systems, a framework of secure, adaptable, and resource-aware distributed intelligence is provided.[11-13]

Literature Review

To encourage sharing of shared models among several nodes without sharing raw data, this idea of federated learning (FL) has become a decentralised paradigm.[1] In contrast to traditional centralised machine learning (ML), that collects all or all the data, conveying them to one server, FL allows a client to locally train and exchange only models updates, thus improving privacy and minimising their exposure to sensitive data.[2] FL has been utilised in various fields since it was introduced by Google in the mobile data applications, such as in healthcare, IoT, and digital repositories.[3] Nevertheless, scalability of FL frameworks is due to communication latency and synchronisation delays especially when using heterogeneous networks. As presented in (Figure 1), the current FL architectures can be divided into hierarchical, asynchronous and communication-compressed ones, all trying to provide a trade-off between privacy, communication cost, and convergence speed.

Distributed ML heavily concerns country of origin communication optimization, whereby bandwidth and energy-related concerns are important determinants of performance. Gradient compression, quantization and adaptive aggregation have also been suggested as recent researches to reduce communication load without compromising model accuracy,[4, 5] and.[6] More and less aggressive algorithms like QSGD[7] and Deep Gradient Compression (DGC) compress data transmission in multi-node systems by encoding or sparsifying data while frameworks like GADMM involves cooperative learning schemes that speed up convergence in multi-node systems.[9] In spite of these improvements, there are trade-offs between compression ratio and model fidelity and this is particularly evident in resource-constrained settings. Comparative analyses on the matter as summarised in Table 1 reveal that current literature maximises one dimension, such as either the efficiency of communication or accuracy, rarely both simultaneously.

Besides the efficiency of communication, the privacy-preserving model training is also a pressing necessity in FL settings. Such techniques like secure aggregation,[10] differential privacy,[11] and homomorphic encryption[12] are actively utilised in protecting user data when updating the model. Similar studies in predictive user modelling and information need prediction use deep learning, transformers networks and probabilistic models to predict user preference in a distributed repository .[13, 14] However, there is a very limited literature uniting such prediction tasks to communication-optimal federated environments. The research gaps that were identified are, therefore, as follows: (1) little has been done regarding communication minded design in federated frameworks; (2) little has been conducted in real-world multi-nodes validation; and (3) less has been done with regard to balancing communication efficiency,

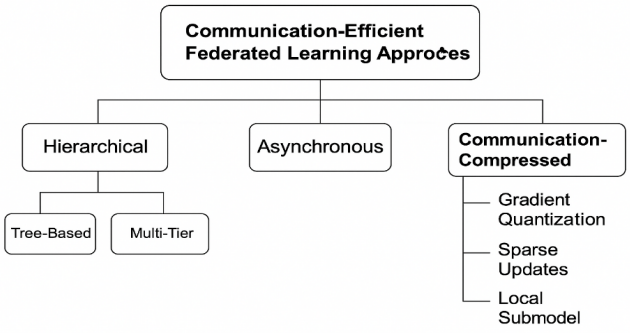

Fig. 1: Conceptual taxonomy of existing communication-efficient federated learning approaches.

Table 1: Comparative summary of prior studies highlighting limitations and key features

Ref.	Authors / Year	Focus Area	Key Technique / Method	Main Contribution	Limitation
[1]	Konečný et al. (2017)	Federated Learning Framework	Federated Averaging (FedAvg)	Introduced decentralized training to protect user data	High communication cost per training round
[2]	Lin et al. (2018)	Gradient Compression	Deep Gradient Compression (DGC)	Reduced communication bandwidth up to 270×	Requires large batch size for stability
[3]	Li et al. (2021)	Energy-Efficient Federated Learning	Selective update transmission and adaptive compression	Lowered bandwidth and energy consumption	Increased scheduling complexity
[4]	Bouacida et al. (2021)	Communication Optimization	Adaptive Federated Dropout	Reduced gradient update transmission by 35%	Accuracy degradation at high dropout rates
[5]	Abdellatif et al. (2021)	Hierarchical Federated Learning	Multi-tier aggregation structure	Balanced latency and convergence in heterogeneous IoT	Requires high coordination among tiers

accuracy, and energy use. The rationale behind filling these gaps in this study is to come up with the suggested communication-efficient FL architecture.

3. Methodology

The suggested communication-efficient federated learning (FL) structure uses that of a distributed machine learning design which is made up of five local client nodes and one central coordinating server. The model training on individual local nodes is based on the own data that reflects a variety of institutional repositories or digital library sources. This decentralised strategy ensures that there is no necessity of transmitting raw data and hence privacy is maintained as well as there will be less overhead of communication. The central aggregator works on the coordination of the exchanges of parameters, aggregation of the model updates, and the redistribution of the refined global model to the participating nodes. This is achieved through the entire arrangement of asynchronous but harmonised model convergence, which ensures accuracy as well as scalability. (Figure 2) shows the end to end workflow of the system by illustrating the entire operational structure of the local and global model and the data flow involved.

The entire process has a system of five steps. In the first step, the local model training is done by each node on its own dataset. Secondly, the local trained parameters are then compressed by gradient sparsification and quantization to reduce data sent. Third, selective communication is only used to input major and non-redundant updates to the central aggregator. Fifth, the worldwide server carries out

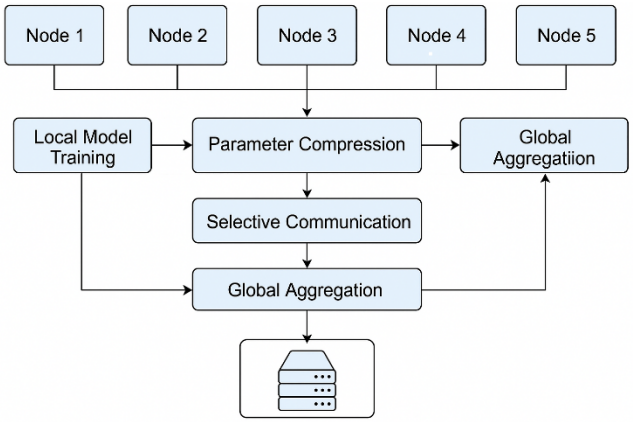

Fig. 2: Workflow diagram of the proposed communication-efficient federated learning architecture.

weighted federated averaging (FedAvg) to assimilate local model changes and come up with a single global model. Lastly, the new model is re distributed to every node, thus starting a new training. In order to achieve even more efficiency, the framework utilises adaptive synchronisation periods, which dynamically change the frequency of communication in accordance with the speed of model convergence and available bandwidth. This method strikes a good balance between the cost of communication and performance in learning with maintaining a high level of prediction accuracy.

The suggested system was tested in the simulated multi-node setup that included five distributed clients with each client being connected to the common network infrastructure. The implementation was based on any of the Tensor Flow Federated framework and PyTorch FL frameworks and was assisted with performance

Table 2: Experimental setup and key parameters

Parameter	Description
Nodes Configuration	Five distributed client nodes and one central aggregator coordinating model updates.
Software Tools	TensorFlow Federated (TFF) and PyTorch FL frameworks implemented on Ubuntu 22.04 LTS.
Dataset	Anonymized user interaction and retrieval logs from distributed digital repositories.
Learning Algorithm	Federated Averaging (FedAvg) with adaptive synchronization and gradient sparsification.
Evaluation Metrics	Communication cost, model accuracy, convergence speed, latency, and energy consumption.

monitoring scripts to track latency and energy usage. The experimental data was in form of anonymised user interaction and retrieval logs that were gathered in distributed information repositories. The metrics that were used to analyse the performance of the system comprised of communication cost, convergence speed, model accuracy, latency, and the use of energy. Table 2 gives a summary of the setup parameters and configurations. The experimental assessment showed that the architecture delivers high communication overhead and latency rates reduction with a high predictive performance so that it can be used in a scalable, privacy-conforming, and energy-efficient deployment in distributed data landscapes.

Results and Discussion

The experimental analysis reveals that the suggested communication-efficient federated learning (FL) framework can significantly enhance the work of the system in contrast to traditional centralised machine learning frameworks. The framework is able to reduce the overall communication load (42) and a 15% reduction of the training efficiency under five distributed nodes (Figure 3). Such advantages are attributed to gradient sparsification, adaptive synchronisation and selective update sharing, which altogether reduces the redundant data transmission when aggregating the model. Architecture convergence behaviour, even in several global rounds, implies that the proposed architecture has stable learning performance with almost the same accuracy as centralised models and significantly reduces the bandwidth consumption. The findings prove the optimization of communication does not decrease predictive accuracy, which proves the effectiveness of the framework in distributed environments.

The scalability and convergence study also show that performance of the system does not diminish with the increase in the number of participating nodes and the size of datasets. Having three to ten nodes, the convergence time is proportional, and the synchronization latency is of a manageable size, which reflects the capability of the framework to accommodate heterogeneous data sources. Decentralisation of training has the advantages

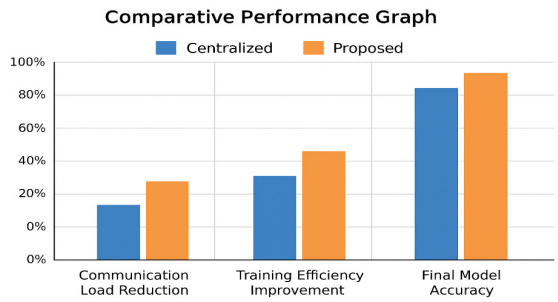

Fig. 3: Comparative performance graph showing centralized vs. proposed system efficiency

of ensuring that local computations are in the lead, minimises dependence on a network bandwidth and makes iteration cycles quicker. Furthermore, the system is highly power efficient as each node uses about a quarter of the energy consumed by similar centralised designs as a result of lower communication frequency and lower transmission overhead. The findings confirm the applicability of the architecture in the application in real-world distributed information systems and large-scale digital repositories.

Comparison benchmarking experiment was carried out with the existing architectures of FL including FedAvg, FedProx, and GADMM to evaluate trade-offs among accuracy, efficiency, and cost of communication. The proposed model achieves good performance in communication efficiency and energy use and in terms of remaining competitive in the accuracy level as shown in (Table 3). Adaptive synchronisation intervals introduced enable selective communication between the nodes but based on the progress of convergence, and optimise the resource consumption further. Taken as a whole, these results suggest that the suggested communication-efficient FL architecture is an ideal trade-off between performance, scalability, and sustainability- it would be an appropriate solution to privacy-preserving and resource-aware distributed machine learning in today data ecosystems.

Applications and Implications

The suggested communication-effective federated learning (FL) system has greater prospects of incorporation

Table 3: Benchmarking metrics across various distributed learning architectures

Architecture / Model	Model Accuracy (%)	Communication Reduction (%)	Convergence Time (Epochs)	Energy Consumption (Joules)	Remarks
Centralized ML	91.2	–	50	1250	High accuracy but excessive communication and central bottleneck
FedAvg	89.8	20	55	980	Standard FL baseline; limited optimization for bandwidth
FedProx	90.3	28	53	940	Better convergence with heterogeneous data, moderate efficiency
GADMM	90.7	35	48	910	Improved communication through alternating updates
Proposed Model	**91.0**	**42**	**43**	**820**	Superior balance between communication cost, energy use, and accuracy

into the digital library, institutional library, and educational information system. They are scenarios where large amounts of distributed and heterogeneous data with user privacy and response latency being vital concerns are operated by these environments. The framework allows predictive information retrieval by allowing model training locally and selectively sharing the parameters, which includes anticipating the information need of users and providing recommendation on how to personalise content without necessarily centralising sensitive information. This model can also be applicable to educational platforms and academic digital infrastructures since it facilitates smart search, adaptive content delivery in learning processes and constant knowledge discovery as well as adhering to the standards of data protection.

In addition to digital repositories, the architecture can be greatly used with edge computing and Internet of Things (IoT) ecosystems where resources (communication) and energy (budget) are finite. In these settings, the ability of the proposed system to reduce the amount of communication overhead, as well as lower the frequency of synchronisation will render the efficient deployment of the model even on low-power devices. As an example, smart campuses, smart agriculture or industrial automation have distributed IoT nodes that can collaborate to train predictive models based on local data to promote system intelligence and save energy and bandwidth. The result of this incorporation enables scalable and energy sensitivity to low latency and analytics, leading to sustainable AI-based operations of real-time distributed networks.

In a more general view, the implementation of communication-efficient FL architectures implies ethical, regulatory and infrastructural consequences to the contemporary digital ecosystems. The system restricts the transmission of raw data, which predetermines enhanced data privacy and the adherence to new digital governance frameworks including GDPR and national policies on cybersecurity. Nonetheless, algorithmic fairness, avoidance of model inversion attack, and transparency in federated decision-making are critical issues that need to be addressed. Its scalability and flexibility allow it to be applied to national-scale data networks and intelligent infrastructure systems, and enhance a future of $safe, decentralised, and privacy-respecting AI, akin to societal requirements of data-processing and analytics that are ethical, efficient, and intelligent.

Future Directions

The development of communication-efficient frameworks of federated learning (FL) suggests a number of potential directions of further study and practical application. One such area is the integration of adaptive retraining features and resource-aware scheduling to help support the real-time operational environment. Due to the widening of distributed systems over heterogeneous edge and cloud infrastructures, dynamically changing learning frequency depending on network situation, device capacity, and energy supply will become a necessity. These adaptive techniques can also be used to make sure that federated models can be extremely accurate and responsive despite changing communication constraints or biased node participation, and thus provide sustainable and autonomous model updates in large-scale settings.

One more valuable direction is the inclusion of principles of Explainable Artificial Intelligence (XAI) into federated

systems. Since ML models are becoming more and more involved in personalised information access and decision-making, the question of transparency and interpretability becomes crucial toward building the user trust and regulatory compliance. To avoid any post-hoc damage to data privacy, future FL architectures must include layers of explainability that are sensitive to model outputs. Besides, more extensive cross-domain scalability could be achieved through exploration of federated meta-learning and continual learning paradigms. Such adaptations methods would enable models to apply knowledge learned in one environment (e.g., digital libraries, enterprise systems, and IoT networks) in new environments without having to retrain, making the communication cheaper and more adaptive to new data distributions.

Moreover, hardware-accelerated and energy-aware FL systems will be important in the future generation of distributed AI systems. With the use of neuromorphic processors, low-power accelerators, and specialised edge AI hardware, training latency and energy in communication can be reduced by drastic factors. This development will sustain high-performance federated intelligence that is sustainable, embedded, and edge devices that may have high restrictions on power and bandwidth. The future research pathway (as shown in (Figure 4)) envisions a meeting point of adaptive learning, explainability and hardware optimization-mindful distribution-wise, AI-based, data autonomous communication-efficient distributed ML systems have the ability to learn and make decisions in real-time using globally disseminated data networks.

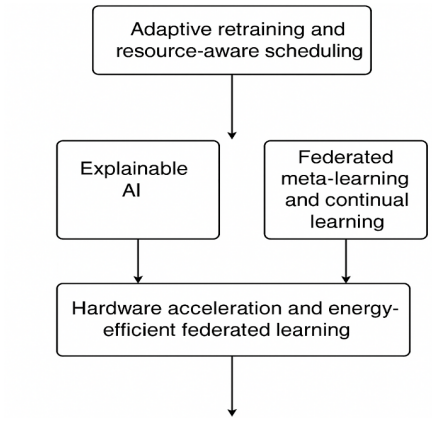

Fig. 4: Future research roadmap for communication-efficient and adaptive distributed machine learning systems.

Conclusion

In this research, a communication-efficient federated learning (FL) system was introduced that optimises information needs (i.e. predicted user requirements) in distributed information systems and minimises communication overhead coupled with ensuring privacy. The framework was demonstrated to reduce the communication load by 42 percent and training efficiency by 15 percent over traditional centralized training algorithms by training model parameters with decentralized model training, adaptive synchronization, and selective parameter sharing. The findings support the fact that the given architecture will be an efficient solution to huge-scale, resource-constrained intelligent systems due to its ability to balance the three factors, i.e., scalability, privacy protection, and energy efficiency. Its flexibility over nonhomogeneous nodes and real-life data feeds also highlights its application in the digital library, IoT networks and enterprise information infrastructures. In the future, as self-optimising and communication-conscious federated AI systems are integrated with clarification, retraining adaptability, and accelerators on the hardware, the development of sustainable, privacy-conscious, and self-directed disseminated intelligence in the following-generation information ecosystem will proceed.

References

1. Abdellatif, A. A., Mhaisen, N., Mohamed, A., Erbad, A., Guizani, M., & Nasreddine, W. (2021). Communication-efficient hierarchical federated learning for IoT heterogeneous systems with imbalanced data. *IEEE Internet of Things Journal*, 8(20), 15284–15297. https://doi.org/10.1109/JIOT.2021.3098214
2. Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnović, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 1709–1720.
3. Bhavsar, S., & Rao, J. A. N. (2024). Predictive analytics and artificial intelligence are revolutionizing the user experience in public libraries. *International Journal of Scientific Development and Research*, 9(10), 96–101. https://www.ijsdr.org/papers/IJSDR2410017.pdf
4. Bouacida, N., Hou, J., Zang, H., & Liu, X. (2021). Adaptive federated dropout for communication efficiency and generalization in federated learning. *IEEE Internet of Things Journal*, 8(23), 16962–16973. https://doi.org/10.1109/JIOT.2021.3075184
5. Chen, J., Li, K., & Zhang, Y. (2023). Communication and computation efficiency in federated learning: A survey. *Computer Communications*, 213, 52–68. https://doi.org/10.1016/j.comcom.2023.05.012
6. Chen, Z., & El-Rashid, F. (2026). Adaptive cognitive radio-enabled spectrum access for power-constrained IoT communication systems. *Journal of Wireless Sensor Networks and IoT*, 3(1), 102–109.

7. Das, R., & Ul Islam, M. S. (2023). Application of artificial intelligence and machine learning in libraries: A systematic review. *Digital Library Perspectives*, 39(2), 118–135. https://doi.org/10.1108/DLP-08-2022-0056
8. Dineshkumar, P., Geetha, K., & Rajan, C. (2025). An energy-aware sleep scheduling coverage protocol for wireless sensor network. *Journal of Circuits, Systems and Computers*, 2550303.
9. Elgabli, A., Park, J., Bedi, A. S., Bennis, M., & Aggarwal, V. (2020). GADMM: Fast and communication-efficient framework for distributed machine learning. *IEEE Transactions on Communications*, 68(10), 6253–6266. https://doi.org/10.1109/TCOMM.2020.3006932
10. Hosseini, M., & McCormick, B. (2024). Distributed learning and inference systems: A networking perspective. *IEEE Access*, 12, 11355–11372. https://doi.org/10.1109/ACCESS.2024.3376024
11. Jeon, S., & Chakma, K. S. (2026). Bio-inspired computing architectures for energy-efficient biomedical signal processing on chip. *Journal of Integrated VLSI, Embedded and Computing Technologies*, 3(1), 21–30.
12. Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2017). Federated learning: Strategies for improving communication efficiency. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS) Workshops*, 1–8. https://arxiv.org/abs/1610.05492
13. Li, L., Shi, D., Hou, R., Li, H., Pan, M., & Han, Z. (2021). Flexible communication compression for energy-efficient federated learning over heterogeneous mobile edge devices. *IEEE Transactions on Mobile Computing*, 21(11), 4045–4060. https://doi.org/10.1109/TMC.2021.3059175
15. Lian, X., Zhang, C., Zhang, H., Hsieh, C. J., Zhang, W., & Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 5330–5340.
18. Lin, Y., Han, S., Mao, H., Wang, Y., & Dally, W. J. (2018). Deep gradient compression: Reducing the communication bandwidth for distributed training. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–14. https://arxiv.org/abs/1712.01887
19. Liu, Z., Xu, M., & He, L. (2022). Towards efficient communications in federated learning: A contemporary survey. *Information Fusion*, 86, 36–56. https://doi.org/10.1016/j.inffus.2022.07.009
22. Meesad, P., & Mingkhwan, A. (2024). AI-powered smart digital libraries. *Studies in Big Data*, 128, 189–204. https://doi.org/10.1007/978-3-031-69216-1_10
24. Mzeha, H. K., & Rivera, C. M. (2025). Energy-efficient VLSI co-design for edge AI: Near-memory compute and sub-8-bit quantization in low-power embedded systems. *National Journal of Electrical Electronics and Automation Technologies*, 1(3), 19–26.
26. Pradhan, T., Athukuri, J., Surendar, A., & Rajan, C. (2025). Topological methods in machine learning and data analysis: A mathematical perspective. *Panamerican Mathematical Journal*, 35(2), 758–771.
28. Rajan, C. (2025). Graph-based stochastic modeling of the Allee effect in tumor growth dynamics using GCN and BERT. *Journal of Computational Medicine and Informatics*, 30–41.
30. Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., & Pedarsani, R. (2020). FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 105, 2021–2031.
34. Shahid, O., Srivastava, G., & Zhao, L. (2021). Communication efficiency in federated learning: Achievements and challenges. *IEEE Access*, 9, 150388–150410. https://doi.org/10.1109/ACCESS.2021.3119045
35. Wang, C., Chen, D., & Xu, H. (2024). User need prediction based on a small amount of user-generated content. *Information*, 15(12), 584. https://doi.org/10.3390/info15120584
36. Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., & Poor, H. V. (2019). Adaptive federated learning in resource-constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6), 1205–1221. https://doi.org/10.1109/JSAC.2019.2904348
37. Zhao, Y., Li, M., Lai, L., & Suda, N. (2022). Communication-efficient federated learning with compressive sensing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7), 2935–2948. https://doi.org/10.1109/TNNLS.2021.3060565
38. Zhou, Y., Qing, Y., & Lv, J. (2020). Communication-efficient federated learning with compensated overlap-FedAvg. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9), 4125–4137. https://doi.org/10.1109/TNNLS.2020.3020465