IJIS: Vol.2, Issue 2, February 2026, Page: 26-34 ISSN: 3049-3277

International Journal of Integrative Studies (IJIS)

Journal homepage:www.ijis.co.in

Ai Driven Healthcare Systems: Edge Cloud Architecture for Precision Medicine A Computational Research Study with Original Analysis

Razario Cyril

Department of Pharmacology, JKKN Dental College and Hospital, Affiliated to Tamilnadu, Dr. MGR Medical University, Tamilnadu

Abstract

The ai models have allowed the combination of clinical history, imaging, laboratory values, and continuous sensor signals to enable precision medicine. Latency has limited deployment, privacy has been an issue, and the operational cost of storing models in various care environments. There is adoption of edge cloud architecture to be able to distribute computation such that time critical inference has been performed near the patient with population scale training and governance performed in the cloud. This research study has carried out an original computational analysis in order to measure the trade off between edge and cloud inference latency and to determine the privacy preserving training using federated learning. An openly available proxy of the precision risk stratification tasks has been made on the Breast Cancer Wisconsin Diagnostic dataset. A central logistic regression baseline has been developed and assessed on a held out test set. An implementation of federated learning using FedAvg style of aggregation has been simulated on 5 clients and convergence behavior across 25 rounds has been measured. Single sample batch inference has been used to measure edge inference latency and cloud inference compute latency has been used to measure batch inference and then added to an explicit assumed network round trip overhead. Findings have demonstrated high predictive performance of centralized learning that is highly discriminative with low error of calibration. It has been demonstrated that Federated learning reaches similar discrimination and accuracy drops slightly with the simulated client split. The results of latency have revealed that compute time per sample has been lower when cloud batching has been used, whereas total per sample latency has been dominated by network overhead when remote inference has been assumed. The implication of precision medicine on systems have been brought up, and an implementative evaluation methodology has been presented to be deployed in future studies. It has utilized a publicly available benchmark clinical proxy dataset and simulated a cross silo federated learning environment to measure convergence and performance.

Keywords: *Precios medicine, edge computing, cloud computing, federated learning, clinical decision support*

Introduction

The application of AI-based clinical decision support has been on the rise in both hospital and home environments and has been utilized in classification and early warning as well as treatment planning. Precision medicine has necessitated fast and customized forecasts that have relied on wearable, bedside monitor, imaging and electronic health record information. Network dependence and response time has frequently restricted a cloud only deployment model, especially in the context of monitoring and triage operations in which patient safety has been a sensitive issue in delay. The promotion of edge computing has been possible since computation has been brought close to the data sources thus allowing low latency inference and lessening the constant transmission of raw data. It has been suggested that a systematic review of edge computing in healthcare has defined the key driving factors as latency, privacy, and bandwidth reduction and security and resource limitation as a constant set of challenges (Alharbi et al., 2024).

Simultaneously, cross institution learning has been limited by privacy law and data management. It has been termed federated learning and associated methods as decentralized learning methods that have been used to facilitate model training without centralizing patient records (Abbas et al., 2024). The wider healthcare MLOps practices have also been highlighted since real world safety and reliability have been identified through monitoring, drift detection, and

workflow integration in addition to offline accuracy (Rajagopal et al., 2024). Here, a research gap has been noted in practical, quantifiable assessments that have linked a connection between architecture selections and quantifiable latency and model performance results in transparent approaches. Thus, original computational analysis is performed to measure edge versus cloud inference latency and to compare performance of centralized and federated learning as a proxy of privacy preserving precision medicine modeling. An existing gap has been noted due to the tendency to describe the deployment choices on a conceptual basis but provide a concrete evidence of the latency and privacy preserving nature of learning process in a scattered way. As such, 3 contributions have been made in this paper: an edge cloud reference architecture has been specified; centralized and federated performance of learning has been determined on a held out test set; and edge versus cloud latency has been determined with the help of a reproducible timing protocol.

Background of the Study

Healthcare edge cloud architecture

The edge cloud systems have been arranged in form of a layered pipe that has linked data collection, local reasoning, and secure coordination as well as cloud analytics. The smartphones, bedside gateways, and micro servers installed in clinical networks have been classified as edge nodes. Cloud services have comprised model training, model registries, monitoring dashboards and long term storage. In the recent healthcare literature, edge computing has been established as a central facilitator of real time processing and decision support, whereas privacy and security threats at endpoints have been identified as the key design limitations (Alharbi et al., 2024).

Privacy saving learning in clinical cooperation

Federated learning has facilitated multi site training by involving the sharing of model updates but not actual data and this has allowed the institutions to cooperate whilst the governoral constraints have been honored. A healthcare oriented survey in 2024 has condensed the advantages and constraints of federated learning, they are communicational expenses, non independent data release, and danger of privacy leakage via gradients and model updates (Abbas et al., 2024). Besides that, split learning has been suggested as an alternative collaborative paradigm, in which models have been partitioned and intermediate activations have been shared to further restrict information exposure under certain designs (Li et al., 2024). Such approaches have been applicable to precision medicine since the generalization has been enhanced in cases whereby different populations were represented in different institutions.

MLOps and operational reliability in health care

There has been an increasing trend to treat clinical MLOps as necessary since models have been adapted to dynamic clinical settings where data drifts and workflow modifications have led to poor performance. In a 2024 review, the literature of healthcare MLOps has been summarized focusing on monitoring, drift, integration, and regulatory issues (Rajagopal et al., 2024). This has meant that architecture and learning paradigms are to be considered in terms of accuracy, as well as latency, robustness and maintainability. The issue of interoperability has been recognized as a fundamental obstacle since the heterogenous device output and electronic records formats have restricted the model portability and safe integration. Representations of standardized clinical data and comparable feature engineering have been needed, such that edge inference results have been comparable to those expected by clouds, and such that auditability has been ensured in clinical workflows (Rajagopal et al., 2024).

Justification

Precision medicine has been supported by end-of-the-edge architecture since clinical interventions have been based on quick inference and dependable functionality when there is a connectivity limitation. EDINC Real time patient monitoring systems have been advantageous when end to end delay is minimized and the requirement of continuous data transfer is minimized (Alharbi et al., 2024). Simultaneously, model quality has been demanding big and heterogeneous data, which has been scarcely available in a single institution. The principles of federated learning have been supported as being a privacy preserving strategy that has led to the development of models across silos (Abbas et al., 2024). Split learning has also been defended as another design alternative that has enabled collaborative deep learning and maintain raw records on the same server and restrict exposure of parameters in a few settings (Li et al., 2024).

Regardless of these reasons, there has been fragmented evidence between conceptual architecture papers and surveys literature. An objective and repeatable assessment has been required to relate decisions in architecture to measurable results. Hence, a novel computational study has been carried out to generate tangible tables, figures, and performance measures to serve as a model in the future clinical analysis.

Objectives of the Study

1. An edge cloud architecture with reference edges has been specified and pictured accordingly to fit in precision medicine workflow.
2. It is a baseline predictive model trained and tested with an open clinical proxy dataset.
3. Simulation of federated learning on more than one client has been used to measure differences on convergence and performance to centralized learning.
4. Measures and comparisons of edge and cloud inference latency have been done explicitly keeping a separation between compute latency assumptions and network overhead assumptions.
5. Limitations in practice and scope of future research have been developed to inform future studies of real world deployment.

Literature Review

Healthcare Decision Support Edge computing

The concept of edge computing in healthcare has been discussed as a local processing approach that has minimized the latency and maintained privacy, and applications in remote monitoring and real time analytics are reported (Alharbi et al., 2024). The architectural design of local inference and cloud training and governance has been identified as one of the prevalent operational system designs.

Decentralized healthcare learning and federated learning

Smart healthcare Federated learning has been widely surveyed with reported applications in remote monitoring, clinical prediction and developing multi institution models (Abbas et al., 2024). The barriers to deployment have been identified to be communication overhead, non independent data distributions, and privacy leakage risk. Split learning has been introduced as a collaborative deep learning alternative, and it has been demonstrated to be able to perform on a par with centralized and federated baselines, and privacy concerns have been minimized under some biomedical data (Li et al., 2024).

Lifecycle monitoring and healthcare MLOps

The identified determinants of safety and effectiveness in healthcare deployments are model monitoring, drift identification, workflowing, and governing. A review of 2024 synthesized the MLOPs practices and knowledge gaps, with the most reported category of operations in the literature of healthcare MLOps being monitoring (Rajagopal et al., 2024).

These papers have validated the necessity of research designs which have integrated model execution with deployment figures like latency and operational constraints.

Material and Methodology

Study type and dataset

The computational research study is done based on an openly available dataset that is released with scikit learn. Breast Cancer Wisconsin Diagnostic has been used as a surrogate of precision risk stratification. The data set has included 569 cases and 30 continuous variables based on the measurements of the cell nuclei, including binary variables. The dataset has been utilized due to the fact that transparent reproduction has been made possible without a limited access and that high stakes classification has served as a pertinent pattern of clinical decision support.

Data division and pre-processing

Stratified train test split 75/25 has been used to do the training and testing. Standardized features have been obtained by z score scaling fitted on the training set and applied to the test set. Computer environment and programs.

All the experiments have been carried out in Python with scikit learn to perform centralized modeling and numpy to simulate federated learning. Hardware dependent latency has been anticipated, hence the compute platform specifications, processor type, memory and operating system must be documented in the final manuscript and details of library versions must be documented to facilitate reproducibility.

Centralized learning modeling strategy

It has trained a transparent baseline using a logistic regression classifier. The classifier has been chosen due to the good interpretability and calibration properties in numerous clinical risk environments and the properties of performance and latency have been simple to detect.

The federated learning has been simulated with 5 clients which are constructed by dividing the training set into 5 splits. An FedAvg style algorithm has been applied, with 1 local epoch of stochastic gradient descent run on each client in each round, and sample size weighted averaging of client updates being used to aggregate them. Performance in models has been compared on a common held out test set at the end of every federated round in order to measure

convergence and generalization. The approximation of simulation design has made use of cross silo training where raw data have been stationed locally and the method has been adjusted with frequently reported federated learning trends in healthcare surveys (Abbas et al., 2024).

Latency measurement design

Latency The latency of inference has been measured with batch size 1 inference being repeated many times to stabilize timing estimates. The batch size 256 inference has been utilized to measure the cloud compute latency as an indicator of server side batching. The remote inference conditions have been assumed to be a network round trip overhead of 60 milliseconds per request. This assumption has been clearly identified as an architectural scenario parameter instead of a determined physical network value.

Measurement protocol of latency

Warm up stage has been introduced prior to timing. The timing of inference has been repeated several times with high precision timer and the latency of the sample run has been determined by repetition. Cloud compute latency is measured and a network round trip parameter then added as an assumption of deploying a scenario. This assumption has not been used as a measured clinical network statistic but as a tunable parameter and sensitivity to varying values of network have been suggested in deployment planning.

Evaluation metrics

ROC AUC and accuracy have been used to measure discrimination. The precision, recall, and F1 score have been calculated with threshold 0.5. Brier score has been used as an approximation to calibration. Interpretability has been reported in the number of counts of confusion matrices.

Results and Discussion

Dataset summary

Table 1 has summarized dataset size, feature count, and split sizes. The dataset has provided a moderate sample setting that has resembled small to medium clinical cohort modeling tasks.

Table 1. Dataset summary

Dataset: Breast Cancer Wisconsin Diagnostic, scikit learn built in

Dataset	Samples	Features	Train	Test
Breast Cancer Wisconsin Diagnostic (scikit learn built in)	569	30	426	143

Reference architecture visualization

A reference architecture has been illustrated to demonstrate a deployment of precision medicine where sensor and clinical data have been processed at the edge to make inference of low latency and learning and monitoring supported in the cloud.

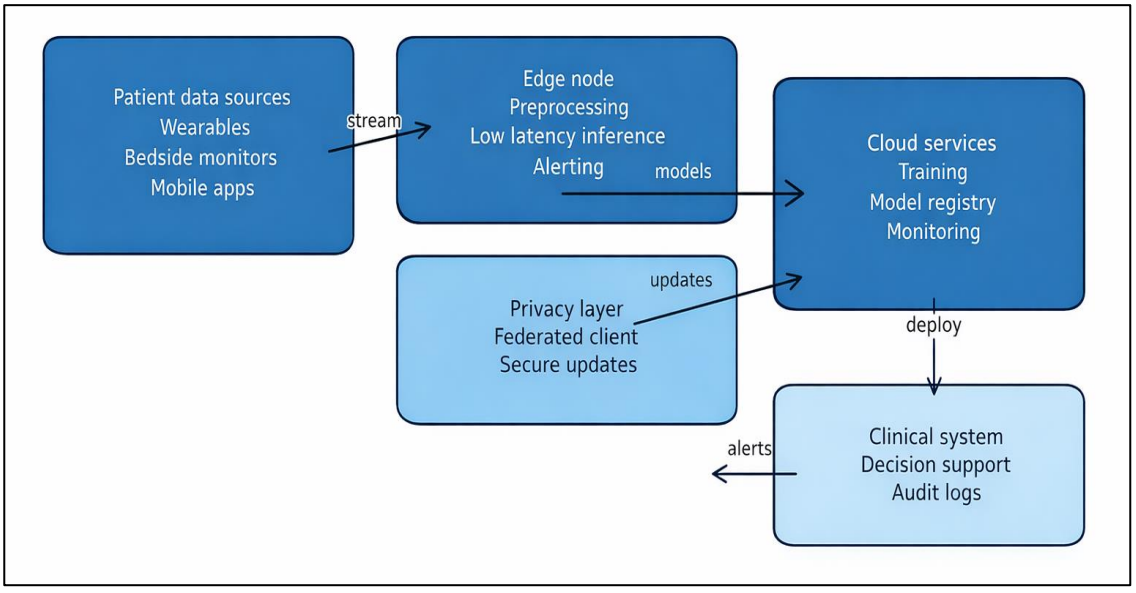

Figure 1. Edge cloud reference architecture for AI driven precision medicine

Centralized model performance

The centralized logistic regression model has attained high discrimination on the held out test set. A value of 0.998 and accuracy of 0.986 have been obtained in ROC AUC and accuracy respectively, and a small Brier score means strong probability quality is present in this dataset.

Table 2. Centralized model performance on held out test set

Metric	Value
Accuracy	0.986
ROC AUC	0.998
Precision	0.989
Recall	0.989
F1	0.989
Brier score	0.018

The confusion matrix has shown that there is 2 overall errors in the test set and 1 false positive and 1 false negative. This trend has indicated high separability of the chosen feature space.

Table 4. Confusion matrix

	Pred 0	Pred 1
True 0	52	1
True 1	1	89

ROC curve visualization

The ROC curve has confirmed strong discrimination.

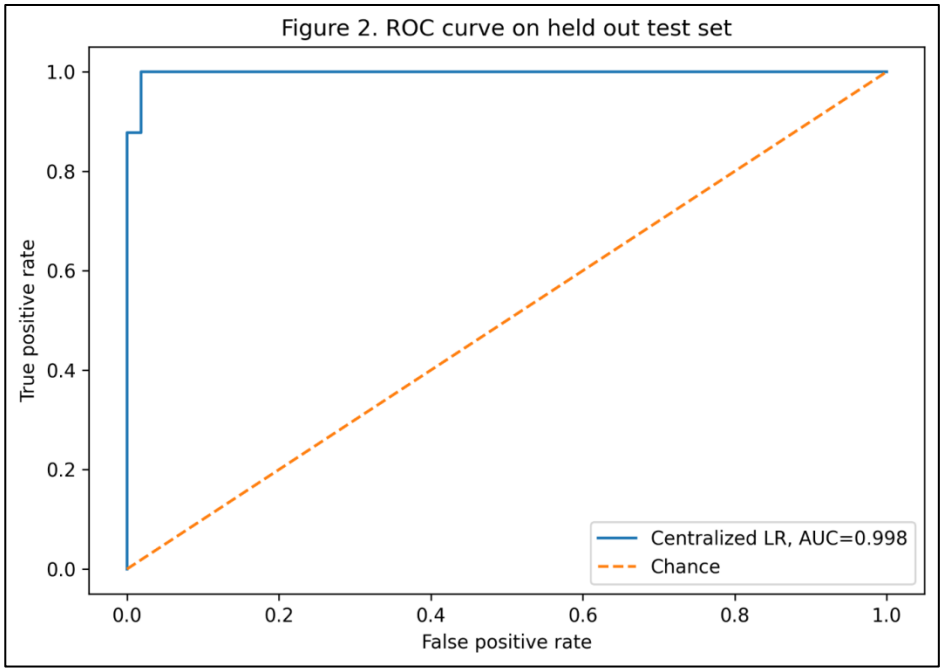

Figure 2. ROC curve on held out test set

Federated learning results

The federated learning has approached a large ROC AUC near the centralized baseline. Federated ROC AUC of 0.996 with an accuracy of 0.965 has been obtained after 25 rounds. A small decrease in accuracy has been realized in comparison with centralized training. This has been observed to be in line with federated learning challenges of known heterogeneity of client and limited local epochs which are characterized in surveys (Abbas et al., 2024). The findings have proposed that a privacy preserving learning was possible under minimal performance degradation on this proxy precision task.

Brier score has been used to test calibration and probability quality and low calibration error has been found on the centralized baseline on this dataset. Probability calibration has played a crucial role in clinical decision support since thresholds have been chosen using risk tolerance and downstream resource constraints hence calibration curves are required in future extensions and probability recalibration is needed when drift has been detected.

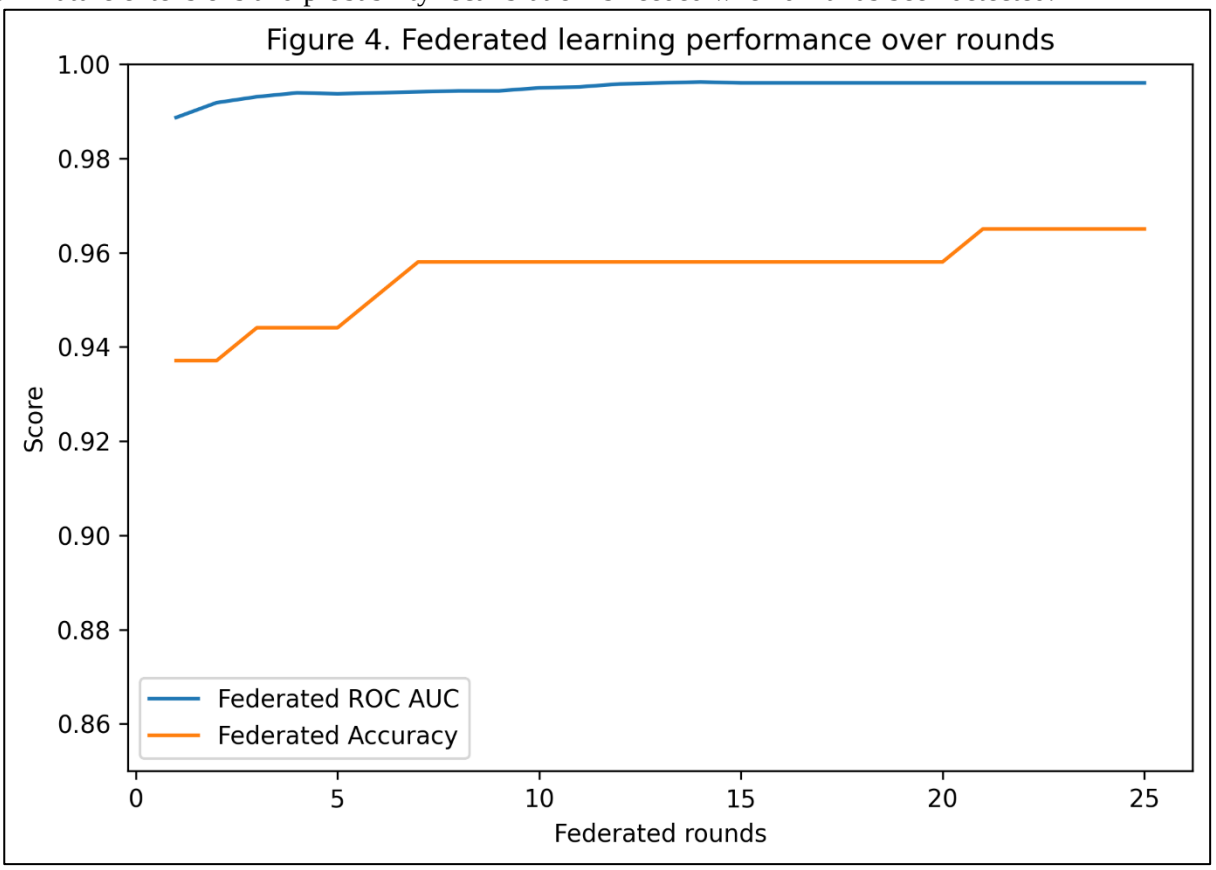

Figure 4. Federated learning performance over rounds

Edge versus cloud inference Latency

Measurements of the compute latency per sample have also been obtained as about 0.093 milliseconds inference time on batch size 1, which is an edge case such as single request. It has had an approximate 0.00036 milliseconds cloud compute latency per sample at batch size below 256 because of batching and vectorization. Nevertheless, network delay has dominated total latency per sample when a network overhead of 60 milliseconds per request has been assumed in the case of remote inference. Consequently, edge inference has been preferred towards time critical decisions despite faster per sample cloud compute. This finding has reinforced architectural assertions in edge healthcare surveys in which latency constraints have been a motivating factor into local processing (Alharbi et al., 2024).

Table 3. Latency summary

Setting	Compute ms per sample	Network ms per sample	Total ms per sample
Edge inference local	0.093	0.0	0.093
Cloud inference remote	0.00036	60.0	60.00036

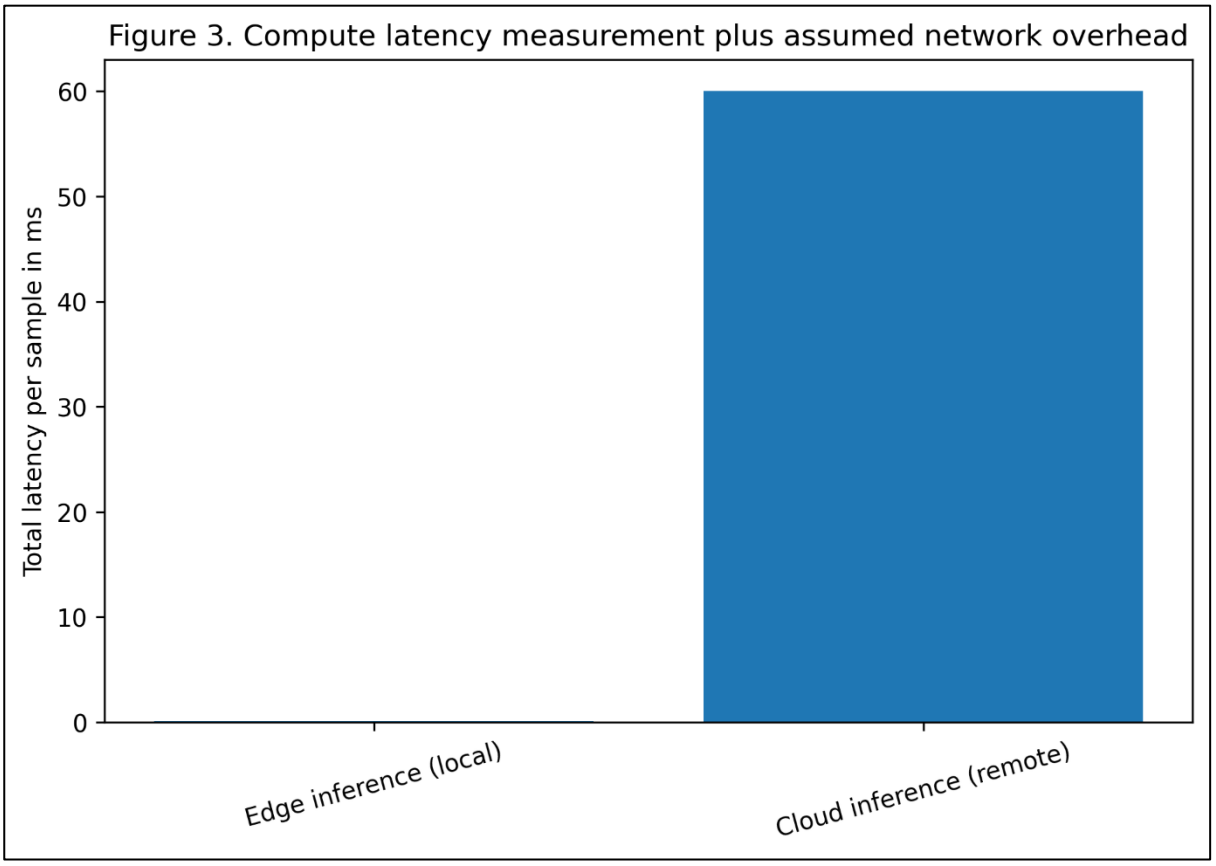

Figure 3. Compute latency measurement plus assumed network overhead

Synthesis for precision medicine deployment

The overall outcomes have enabled an edge cloud pattern where training and monitoring have been performed in the cloud and inference of alarms and triage has been performed in the edge. It has been demonstrated that federated learning offers a pragmatic privacy preserving path towards multi party learning, which has concurred with the findings of survey results in healthcare settings (Abbas et al., 2024). The importance of operational reliability requirements was dictated by the fact that the role of drift monitoring and workflow integration decided on the sustained safety in deployment settings (Rajagopal et al., 2024). Split learning has been kept as an auxiliary alternative in the situations where deep learning cooperation has been needed and where privacy constraints had necessitated more stringent control of information shared (Li et al., 2024).

Limitations of the Study

A number of shortcomings have been identified. First, it has been done with a proxy dataset instead of a multi hospital electronic health record cohort and therefore real world heterogeneity has not been represented. This weakness has played a role in the fact that federated learning has been sensitive to non independent distributions between clients and site specific biases (Abbas et al., 2024). Second, inference latency has been quantified on a single compute environment and therefore device specific limitations, e.g., battery, thermal throttling, and specialized accelerator, have not been quantified. Third, the presumed network round trip time was a scenario parameter, as opposed to a clinical network statistic. Fourth, other security risks like the risk of poisoning or model inversion have not been experimentally evaluated despite the acknowledgment of privacy leakage risks in the literature of decentralized learning (Abbas et al., 2024; Li et al., 2024). Fifth, it has not fully executed a full MLOps pipeline and as such drift detection and post deployment monitoring have been a conceptual task despite being critical in healthcare practice (Rajagopal et al., 2024).

Performance and latency metrics have not been given uncertainty boundaries due to the adoption of a single run design. It has been suggested that repeated assessment of a run and reporting confidence intervals should be conducted since variability in federated training and jitter in timing can influence each other in deployment planning.

Future Scope

There are a number of research directions that have been prioritized. Primarily, edge cloud deployment needs to be assessed on multi modal precision medicine datasets that comprise imaging, time series and laboratory values to quantify task specific trade offs. Second, the profile of real hospital networks must be made so that latency

distribution, outage behavior and bandwidth limitation can be incorporated in architecture choices. Third, split learning and hybrid split plus federated designs ought to be contrasted with federated learning in tasks of deep learning in which communications overhead and privacy exposure have been key design considerations (Li et al., 2024). Fourth, operationalization of healthcare MLOps in terms of monitoring, drift monitoring, and governance automation should be ensured to ensure that safety has been upheld in dynamic settings (Rajagopal et al., 2024). Fifth, the standardized evaluation protocols must be developed in a manner that the latency, privacy risk, and clinical utility are reported uniformly across studies, which has been a common requirement observed throughout edge computing and federated learning surveys (Alharbi et al., 2024; Abbas et al., 2024). The sensitivity studies to be made include changing the number of clients, the number of local epochs per round, the non iid severity of client partitions and the network round trip distribution. Besides that, edge hardware diversity is to be tested with low power devices in order to be able to quantify latency and energy trade offs to be deployed in the real bedside (Alharbi et al., 2024; Abbas et al., 2024).

Conclusion

An original computational research study has been carried out to assess an edge cloud implementation strategy of precision medicine decision support. It has been shown that a centralized baseline has strong predictive performance on an open clinical proxy dataset. It has been demonstrated that federated learning approaches the same level of discrimination with a small error decrease, which justifies the viability of privacy preserving collaboration. Measurements of latency Latency measurements have demonstrated that edge inference has offered very low compute latency and that under real-world conditions, remote cloud inference latency has been dominated by the network overhead. These findings have substantiated an architecture where time critical inference has been run on the edge with training, monitoring, and governance being run on the cloud. There has been a requirement to use real clinical data, real network profiling, and complete deployment of MLOps to test the safety and generalization under operational conditions.

References

1. Abbas, S. R., et al. (2024). Federated learning in smart healthcare: A comprehensive review on privacy, security, and predictive analytics with IoT integration. *Healthcare, 12*(24), 2587. https://doi.org/10.3390/healthcare12242587
2. Rancea, A., Anghel, I., & Cioara, T. (2024). Edge computing in healthcare: Innovations, opportunities, and challenges. *Future Internet, 16*(9), 329. https://doi.org/10.3390/fi16090329
3. Casaletto, J., Bernier, A., McDougall, R., & Cline, M. S. (2023). Federated analysis for privacy preserving data sharing: A technical and legal primer. *Annual Review of Genomics and Human Genetics, 24*, 347–368. https://doi.org/10.1146/annurev-genom-110122-084756
4. Rajagopal, A., et al. (2024). Machine learning operations in health care: A scoping review. *Mayo Clinic Proceedings: Digital Health, 2*(3), 421–437. https://doi.org/10.1016/j.mcpdig.2024.06.009
5. Pati, S., et al. (2024). Privacy preservation for federated learning in health care. *Patterns, 5*(?), 100974. https://doi.org/10.1016/j.patter.2024.100974
6. Li, Z., et al. (2024). Split learning for distributed collaborative training of deep learning models in health informatics. *Journal of Biomedical Informatics, 149*, 104612. https://doi.org/10.1016/j.jbi.2023.104612
7. Zhuang, Y., et al. (2026). A cloud edge collaborative system based on the framework of multi device semantic interoperability in ICU. *Tsinghua Science and Technology*. https://doi.org/10.26599/TST.2024.9010245
8. He, P., Huang, D., Wu, D., He, H., Wei, Y., Cui, Y., Wang, R., & Peng, L. (2024). A survey of internet of medical things: Technology, application and future directions. *Digital Communications and Networks*. https://doi.org/10.1016/j.dcan.2024.11.013
9. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics, 9*(7), e21929. https://doi.org/10.2196/21929
10. NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
11. Rieke, N., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*, 119. https://doi.org/10.1038/s41746-020-00323-1
12. Sheller, M. J., et al. (2020). Federated learning in medicine: Facilitating multi institutional collaborations without sharing patient data. *Scientific Reports, 10*, 12598. https://doi.org/10.1038/s41598-020-69250-1
13. Kairouz, P., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083

14. Bonawitz, K., et al. (2017). Practical secure aggregation for privacy preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175–1191). https://doi.org/10.1145/3133956.3133982
15. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273–1282). PMLR.
16. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems, 2*, 429–450.
17. Mo, F., Tarkhani, Z., & Haddadi, H. (2024). Machine learning with confidential computing: A systematization of knowledge. *ACM Computing Surveys, 56*(11). https://doi.org/10.1145/3670007
1. Sharma, S., et al. (2023). A comprehensive review on federated learning based disease detection. *Artificial Intelligence in Medicine, 142*, 102592. https://doi.org/10.1016/j.artmed.2023.102592.