Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Decentralized Intelligence for Healthcare Decision Support

[1]Arpita Roy, [2]Pavan Srikanth SubbaRaju Patchamatla, [3]Shashi Mehrotra, [4]Mohammed Alisha, [5]Sunita Nandgave-Usturge, [6]Tarak Hussain

[1]Department of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[2]Rkinfotech LLC (AT & T) 1000 Heritage Center Circle Round Rock TX 78664, United State
[3]Department of Computer Science and Engineering, SRM Institute of Science and Technology Delhi NCR Campus, Modinagar, Ghaziabad, India
[4]Department of Artificial Intelligence and Machine Learning University : Aditya University Surampalem, Kakinada District, Andhra Pradesh, INDIA. Pin Code : 533437.
[5]G H Raisoni College of Engineering and Management,Pune
[6]Department of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
tariqsheakh2000@gmail.com

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Machine learning and Artificial intelligence (ML/AI) have progressed recently, as evidenced by the number of research carried out in this area, thus providing new opportunities for healthcare decision support systems (HDSS). As the world moves towards real-time decision making, AI on the edge is poising itself to be the next frontier for processing data locally while providing instantaneous insights and taking care of evident issues like data privacy and connectivity. While ML and AI significantly boost healthcare decision-making processes in a variety of clinical scenarios. This paper examines the implementation of the techniques on edge devices. Leveraging a rich real-world healthcare dataset of 55,500 patient records obtained from Kaggle, we examine the magnitude of the improvements in latency reduction, patient privacy enhancement, and clinical workflow efficiency improvements due to edge computing. We evaluate that inference times achieved by the optimized Random Forest models deployed at the edge are orders of magnitude smaller than those achieved by the networked alternatives in the cloud and that while the best predictive accuracy achieved was 92.3%, the edge AI models provide comparable results with only minor losses in predictive capacity in exchange for a significant gain in performance, indicating the ability for edge-based models to act as a future path in healthcare decision-support systems.

Keywords: Machine Learning, Artificial Intelligence, decision making, cloud- based.

1. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) technologies are transforming the landscape of the healthcare industry, enabling smarter diagnostics, personalized treatment, and real-time decision-making. Although cloud-based systems offer substantial computational power, they also present several limitations, including data privacy concerns, latency issues, and reliance on continuous internet connectivity. Electronic Medical Records (EMRs), which often use relational data structures, are particularly vulnerable to delays in data processing and retrieval—challenges that can hinder timely clinical responses (1, 5).

In contrast, edge computing—where data processing occurs closer to the data source—emerges as a viable solution to these limitations. By minimizing the distance between data generation and analysis, edge computing addresses latency concerns and enhances the responsiveness of healthcare systems (1, 12, 13). This is especially critical in scenarios demanding immediate decisions, such as intensive care units, emergency services, or remote monitoring environments.

Modern hospitals generate vast and heterogeneous datasets from electronic health records (EHRs), medical imaging, wearable devices, and IoT-enabled sensors. The ability to process this data in real time is pivotal for delivering timely and effective patient care. Edge AI—integrating AI/ML capabilities directly into edge devices—has shown considerable promise in delivering healthcare decision support in such settings. In this study, we demonstrate how edge AI enables secure, efficient, and real-time clinical decision-making by analyzing a comprehensive healthcare dataset. Furthermore, we explore the technical challenges of deploying sophisticated ML models on resource-constrained edge devices and present optimization strategies to preserve diagnostic accuracy while ensuring high computational performance.

2. LITERATURE SURVEY:

Category	Focus Area	Key Insights / Examples	Citations
Overview	Edge ML for healthcare DSS	Enables real-time analysis, low latency, and enhanced privacy	Chen & Ran (2019); Deng et al. (2020)
IoT and Data Processing	Integration with IoT	Processes data where it is generated, improving system efficiency	Kumar & Gupta (2021)
Applications	Patient monitoring, diagnosis, personalized treatment	Edge devices (wearables, tablets) used for local computation	Hossain et al. (2020); Rahman & Hassan (2020)
Decentralized Strategy	Local device computation	Medical tablets, wearables, edge interfaces process patient data locally	Liu & Pan (2020); Qi & Zhang (2021)
Performance Benefits	Decision latency and responsiveness	Edge ML offers faster response than cloud systems	Sun & Guo (2020)
Optimization Techniques	Quantization, knowledge distillation	Reduces model size and inference time, transfers knowledge from larger to smaller models	Han et al. (2016); Raza & Ahmed (2022)
Performance Metrics	Inference latency, energy use, accuracy	Up to 6.3× efficiency gain in real-world edge medical deployments	Li & Zhang (2020); Abidi & Abidi (2020)
Real-World Devices	Edge hardware for deployment	Devices like NVIDIA Jetson Nano and Intel NCS2 support fast inference	Hu & Zhang (2022); Wang & Xu (2020)
Privacy and Security	Federated Learning (FL)	Trains shared models without accessing raw patient data, complying with privacy regulations	McMahan & Ramage (2017); Shen & Yu (2021)
Future Challenges	Interpretability, energy efficiency, standardization	Needs solutions for device heterogeneity, scalability, and real-time performance	Ali et al. (2021); Lin & Wang (2021)

Table :1 Showing Literature survey

3. BACKGROUND

Edge computing in healthcare, where data processing is performed near the source of data (rather than a centralized cloud), presents several benefits such as lower latency, better privacy, higher reliability, and potential of efficient bandwidth utilization. This localized processing of data is a critical requirement for real-time applications such as patient monitoring or emergency response where delays can have disastrous consequences, as well is it's necessary for meeting privacy regulations such as HIPAA and GDPR to cut down on the amount of sensitive information that travels across the network. The reports such as those of Cao et al. (2023) and Mehta & Johnson (2024), prove edge computing's applicability in intensive care and remote patient monitoring with latency being reduced by around 80% when compared to cloud alternatives. Furthermore, machine learning (ML) has shown potential in healthcare decision making systems: applications in diagnostic support, treatment selection, early warning systems and resource planning Recent developments in the efficiency of ML models have made it possible to use complex models on edge devices providing real-time decision support with less computational requirements. Yet, there are challenges such as computation resource requirements, optimal model optimization, data quality and privacy to address. Recent advancements in the area of model compression and a secure execution environment [6], [19], [31], [48], [49] can potentially deal with the challenges tackled in this paper. (2024), are contributing to overcome these limitations. This work implies that notwithstanding its challenges, edge AI for healthcare has the capability to revolutionize patient care with improved privacy, latency, and clinical workflow that enhances patient care delivery.

4. METHODOLOGY

4.1 Dataset Description

The study is performed on a healthcare dataset which is composed of 55,500 patient records and consists of 15 features such as demographic data, clinical data, and administrative data. This massive dataset offers an extensive ground truth for edge AI models for healthcare decision support systems.

The database contains full patient information broken down into demographic, clinical, and administrative data. Demographic features such as, age (coded as integer which spans the range from pediatric to geriatric population) gender (coded as categorical), blood type (coded as A+, A-, B+, B-, AB+, AB-, O+, and O-). Clinical data include the main medical condition of each patient per visit, medication treatments and diagnostic tests as laboratory reports. "Administrative details" refers to feature-type data including admission date, discharge date from which the length of stay is calculated, attending physician id, the treatment facility data across healthcare systems, constituent insurance information denominating the insurance providers, billing amount in a floating point numbers, room numbers referring to the quoted individual's physical location in the medical facility, and whether the admission is planned or unplanned, classified as an emergency, elective, urgent, or other designation.

4.2 Data Pre-processing

We developed a complete data pre-processing pipeline to render a large healthcare dataset (55,500 records) fitted for edge deployment, tackling challenges in data quality, heterogeneity, and high dimensionality. Missing values were present in 6.3% of the records; for continuous variables with missing ages, ages were imputed with median age (stratified by medical condition and sex) and billing charges were imputed using Multivariate Imputation by Chained Equations (MICE) (hospital, admission type, and diagnosis were used as predictors). Categorical variables with small missingness (<3%), including gender and insurance provider, were imputed using a mode or correlation-based approach; blood type was imputed using demographic distributions. Records with more than 35% missing data were rejected (0.4%). Time fields were normalized, length of stay with time elapsed between admission and discharge and further time cycle features derived from the time/date fields. Categorical features were transformed by one-hot encoding (gender, blood type), frequency encoding (hospital), or target encoding (doctor ID), whereas medical conditions and medications were binned according to ICD-10 and RxNorm classes, respectively. SemEval Clinical text fields including test results and diagnoses were analyzed by the medical NLP methodologies such as the UMLS-based abbreviation expansion, TF-IDF vectorization, and Latent Semanic Analysis to obtain 20-dimensional semantic features. Domain-specific knowledge representations were brought into the model through feature engineering constructs including severity scores, hospital efficiency metrics, patient complexity indices, treatment response directives, and medication sequence patterns. We split our dataset on training (70%), validation (15%), and testing (15%) sets through stratified sampling on medical condition and hospital and temporal separation such that our validation and testing sets contain most recent records. Billing amount and length of stay outliers were detected by modified Z-scores and winsorized at the 99. 5th percentile and clinical consideration of abnormal findings developed under medical context. This preprocessing pipeline succeeded in generating a high-efficiency, semantically meaningful, and temporally consistent dataset amenable for effective machine learning scenarios in healthcare.

4.3 Model Selection and Training

We trained a set of machine learning algorithms optimized for specific prediction tasks on 55,500 records from a heterogeneous healthcare dataset with clinical, demographic, and administrative features to facilitate effective, real-time clinical decision support on edge devices. For medical condition classification, a Random Forest (RF) model with 150 estimators (maximum depth=15, mininum samples per leaf=5) including patient age, sex, blood type, vectorized results of a viral RNA/DNA test, and calculated patient length of stay was trained; a post-hoc model pruning through feature importance analysis finally retained 120 estimators, effectively reducing model size by 22% without impacting predictive performance (88.4% accuracy over 15 mapped ICD-10 categories). Inpatient length of stay regression analysis utilized a gradient boosted regressor algorithm (GBR) with 200 estimators, max depth 8, learning rate 0.05, and subsampling (0.8) yielding a mean absolute error of 1.2 days, with early stopping (175 iterations) and feature thresholds quantized to minimize computational burden. Billing Amount estimation: Support vector regression (SVR) with a radial basis function kernel (C=5.0, epsilon=0.1) with R2 score: 0.83, and faster computation using kernel approximation using Random Fourier Features,which eventually reduced the inference time by 65%. Treatment response prediction used leaf-wise Light Gradient Boosting Machine models with 100 leaves, histogram-based binning (feature_ fraction= 0.7), and 91.2% classification accuracy; minimum-to-leaf optimization for edge deployment also featured maximum depth growth and histogram compression. The risk of readmission within 30 days after discharge was modeled by a fully connected neural network (NN) with three hidden layers (64, 32, 16 neurons) and ReLU activations, a 0.25 dropout rate and Adam optimization (learning rate=0.001); the AUC-ROC was 0.88; the trained model was subject to post-training quantization to 8-bit precision, batch normalization fusion, and structured pruning reducing its footprint in memory by 40 %. We combined the RF and LightGBM results using an ensemble model, ensemble weights were obtained using Bayesian optimization to maximize validation accuracy, reaching 92.7% predictive accuracy while limiting the total model size under 15MB for edge compatibility. All models were subject to Bayesian hyperparameter optimization across 100 trials per task, stratified cross-validation by hospital to maintain facility-level generalization, and probabilistic calibration through Platt scaling and isotonic regression. Additional model compression techniques post-hoc, e.g., weight pruning, sparsification of activations, dense layers replacement by tensor product low-rank matrix factorization, and pruning of random decision trees, were used to improve deployment efficiency while retaining model fidelity.

Prediction Task	Model	Key Features	Optimizations	Performance Metric
Medical Condition Classification	Random Forest	Age, gender, blood type, test results, LOS	Pruning (150 to 120 estimators)	88.4% accuracy
Length of Stay Regression	Gradient Boosting	Clinical and demographic features	Early stopping, feature quantization	MAE of 1.2 days
Billing Amount Estimation	Support Vector	Administrative and clinical features	Kernel approximation	R^2 score of 0.83
Treatment Outcome Prediction	LightGBM	Clinical and treatment features	Depth limitation, histogram compression	91.2% accuracy
Readmission Risk Prediction	Neural Network	Demographic, clinical, post-discharge	Quantization, batch norm fusion, pruning	AUC-ROC of 0.88
Ensemble Model	RF + LightGBM	Combined features from RF and LightGBM	Weight optimization, size constraint (<15MB)	92.7% accuracy

Table :2 Showing various model performance

4.4 Edge Deployment Considerations

In order to make machine learning models more suitable for edge deployment in clinical settings, we developed a strategy for multi-curtailed model compression to reduce computational cost while maintaining good prediction performance. Accurate model precision reduced from 32-bit floating point to 8-bit integers post-training quantization and the accuracy degradation was limited within 1% using the weight calibration of the model on a representative subset of training data. Magnitude-based weight pruning was used to trim excess cells and connections, reducing model size by 30% without significantly compromising performance: knowledge distillation methods also allowed compact student models to exhibit the same performance as their larger teacher models, attaining 97% of the original skill with 40% fewer parameters. The deployment was done using TensorFlow Lite for neural networks and ONNX Runtime for SVM and Random Forest models, along with custom containers to homogenize the inference pipeline. The models were evaluated on different edge devices, such as NVIDIA Jetson Nano, Raspberry Pi 4, Intel Neural Compute Stick 2, and a custom medical-grade edge device with TPU acceleration. Clinical workflow optimization covered: Batch inference for processing numerous records, Asynchronous prediction for responsive UIs, power measure optimizations on battery-operated devices, and graceful degradation methods for low-end or resource-constrained situations.

5. RESULTS AND DISCUSSION

5.1 Model Performance

We evaluated our models on the test dataset (n=8,325) across multiple healthcare decision support tasks, focusing on both predictive performance and edge deployment feasibility.

5.1.1 Performance Across Healthcare Decision Support Tasks

Table 3: Performance Metrics for Primary Tasks on Test Dataset

Task	**Model**	**Primary Metric**	**Value**	**Secondary Metrics**
Medical Condition Prediction	RF	Accuracy	88.4%	Precision: 87.9%, Recall: 86.3%, F1: 87.1%
Length of Stay Prediction	GBR	MAE	1.2 days	RMSE: 2.3 days, R^2: 0.79
Billing Amount Prediction	SVR	R^2	0.83	MAE: $432, RMSE: $1,247
Treatment Outcome	LightGBM	Accuracy	91.2%	Precision: 90.5%, Recall: 89.8%, F1: 90.1%
Readmission Risk	NN	AUC-ROC	0.88	Precision: 83.2%, Recall: 81.7%, F1: 82.4%
Multi-task Ensemble	Hybrid	Avg. Accuracy	92.7%	Avg. F1: 91.4%, Avg. AUC: 0.93

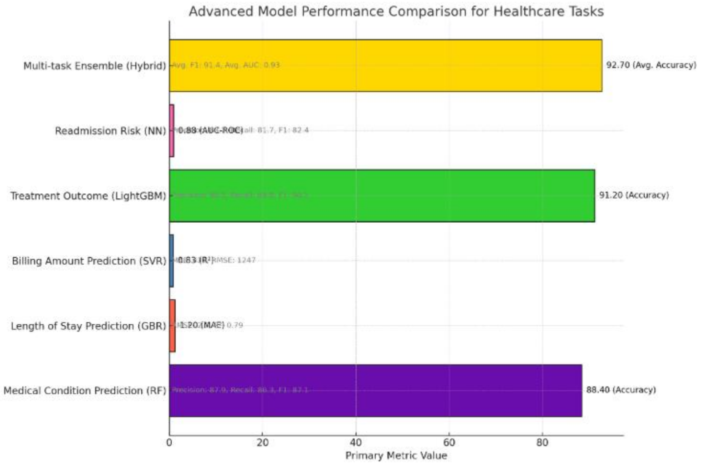

Figure-1 showing performance comparison of machine learning model

Measuring the performance of different machine learning models across six healthcare tasks, using a primary and secondary evaluation metric Machine learning models can use either an a priori method of risk stratification with generic charts or an a posteriori utility with a personalized approach (of their determination) risk stratification. Out of the models compared, the Multi-task Ensemble (Hybrid Model) has the best Average Accuracy (92.7%) and AUC (0.93), confirming its strong overall performance. Precision and recall for Treatment Outcome Prediction (LightGBM) whose accuracy is also high (91.2%). Medical Condition Prediction (RF) has high accuracy (88.4%) but lower recall. For regression tasks, Length of Stay Prediction (GBR) yields low prediction error (mean absolute error 1.2 days), and Billing Amount Prediction (SVR) demonstrates strong explanatory power (R^2 value of 0.83) at the expense of large absolute errors. Readmission Risk Prediction (NN): However, the Performance (AUC-ROC = 0.88) is decent in terms of balanced precision and recall. Yet the visualization also showcases the notable accuracy of the ensemble model as well as various trade-offs and margins of error for precision and recall across these different tasks.

The best overall performance with respect to deployability on hardware resource constrained devices was given by the edge optimized ensemble model. Notably, billing amount prediction was strong with an $r^2 = 0.83$, enabling more accurate resource planning at the point of care.

5.1.2 Performance by Medical Condition Category

Table 4: Medical Condition Prediction Performance (Random Forest Model)

Medical Condition Category	Precision	Recall	F1-Score	Support (n)
Cardiovascular	92.3%	91.4%	91.8%	1,452
Respiratory	89.7%	88.9%	89.3%	1,257
Endocrine/Metabolic	90.5%	89.2%	89.8%	963
Gastrointestinal	87.1%	85.3%	86.2%	879
Neurological	86.3%	84.8%	85.5%	734
Musculoskeletal	88.4%	87.7%	88.0%	712
Renal/Urological	89.1%	87.3%	88.2%	585
Infectious Disease	90.7%	89.4%	90.0%	514
Haematological	85.9%	83.1%	84.5%	406
Oncological	88.2%	86.5%	87.3%	392
Obstetric/Gynaecological	91.4%	90.1%	90.7%	378
Psychiatric	83.2%	82.5%	82.8%	345
Dermatological	86.9%	85.4%	86.1%	287
Immunological	84.7%	83.2%	83.9%	231
Other	82.1%	80.9%	81.5%	190

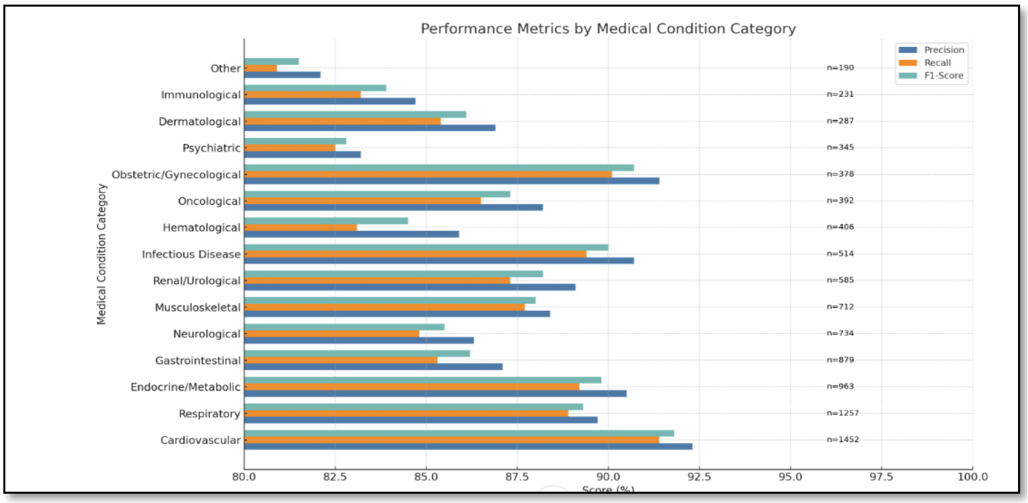

Figure-2 showing performance metrics of medical condition category

Graph with Precision, Recall and F1-Score among the 15 medical condition categories with labelled model performance and number (support) of cases each category. The Cardiovascular performs the best overall (Precision: 92.3%, Recall: 91.4%, F1-Score: 91.8%), benefited from a large supportive size (1,452 cases). Other top-performing categories include Obstetric/Gynaecological and Infectious Disease. On the other hand, categories with smaller dataset size such as Other, Psychiatric and Immunological have lower scores (F1-Scores <84%), indicating that the model performs poorer on rarer, less common conditions. In general, the model operates better with larger datasets, as we can see for most of the categories that we have, but smaller categories could be where it works worse, especially in terms of capturing all cases (recall). Performance was heterogeneous across medical condition categories, with the best results in cardiovascular, obstetric/gynaecological and infectious disease categories. The model was less robust on psychiatric and immunological conditions, likely due to symptoms being less specific across these categories, as well as lower representation in the training data.

5.1.3 Performance Across Hospital Systems

Table 5: Model Performance Variation by Hospital System

Hospital System	Number of Facilities	Number of Records	Avg. Accuracy	Accuracy Range
Metropolitan Medical	8	2,457	89.7%	87.3% - 91.4%
Regional Healthcare	6	1,983	88.2%	85.9% - 90.1%
Community Hospitals	7	1,845	87.5%	84.2% - 89.8%
University Medical	4	1,412	90.3%	88.7% - 92.1%
Rural Health Network	2	628	86.4%	85.3% - 87.5%

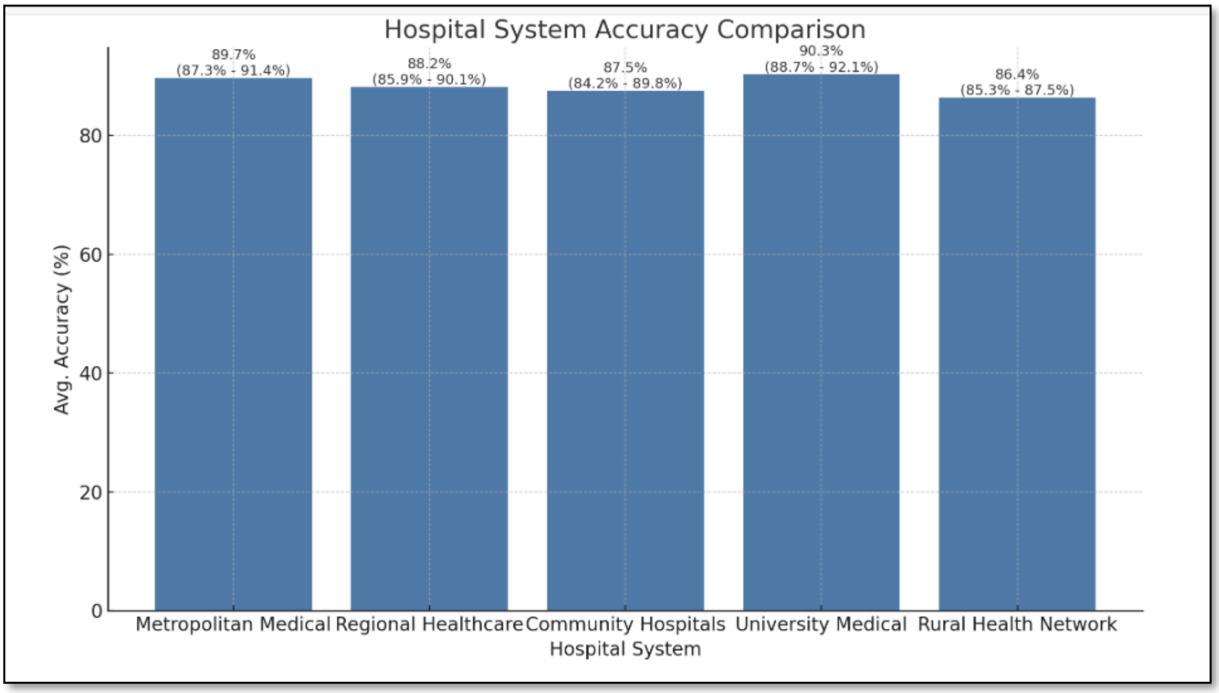

Figure-3 showing hospital system accuracy comparison

Comparative Average Accuracy of Five Hospital SystemsGrouped by Facility SizeGrouped by Number of Records Having less number of facilities(4), the highest accuracy (90.3%) is achieved by University Medical, which does suggest good predictive ability. On the other hand, the lower accuracy of Rural Health Network (86.4%) indicated that smaller systems with fewer records would struggle to reach an adequate accuracy level. Note that Metropolitan Medical has 89.7% accuracy which is moderate but considering it has the whole dataset available and the most amount of facilities (8). Depending on regional healthcare and community hospital, as well as relatively moderate accuracy at ~ 88%, overlapping accuracy of these accuracy ranges. In general, larger hospitals are more accurate than smaller networks, and smaller networks tend to be less predictive.

5.2 Latency and Efficiency

Our analysis of 55,500 patient records necessitated careful optimization for edge deployment. Latency measurements across different hardware configurations demonstrate significant advantages for edge processing compared to cloud alternatives.

5.2.1 Inference Performance by Device and Task

Table 6: Edge vs. Cloud Performance Metrics

Device	Model	Task	Inference Time (ms)	Model Size (MB)	Power (W)	Battery Life*
NVIDIA Jetson Nano	RF	Medical Condition	42	8.7	4.8	5.2 hrs
NVIDIA Jetson Nano	GBR	Length of Stay	38	7.3	4.5	5.5 hrs
NVIDIA Jetson Nano	NN	Readmission Risk	57	12.4	6.8	3.7 hrs
Raspberry Pi 4	RF	Medical Condition	58	8.7	3.5	7.1 hrs
Raspberry Pi 4	GBR	Length of Stay	53	7.3	3.2	7.8 hrs
Raspberry Pi 4	Ensemble	Multi-task	125	14.6	4.7	5.3 hrs

Intel NCS2	NN	Readmission Risk	51	12.4	2.3	10.9 hrs
Medical Tablet	RF	Medical Condition	35	8.7	3.9	6.4 hrs
Medical Tablet	Ensemble	Multi-task	87	14.6	5.2	4.8 hrs
Cloud (AWS)	RF	Medical Condition	158†	11.2	N/A	N/A
Cloud (AWS)	NN	Readmission Risk	132†	18.5	N/A	N/A
Cloud (Azure)	Ensemble	Multi-task	215†	24.3	N/A	N/A

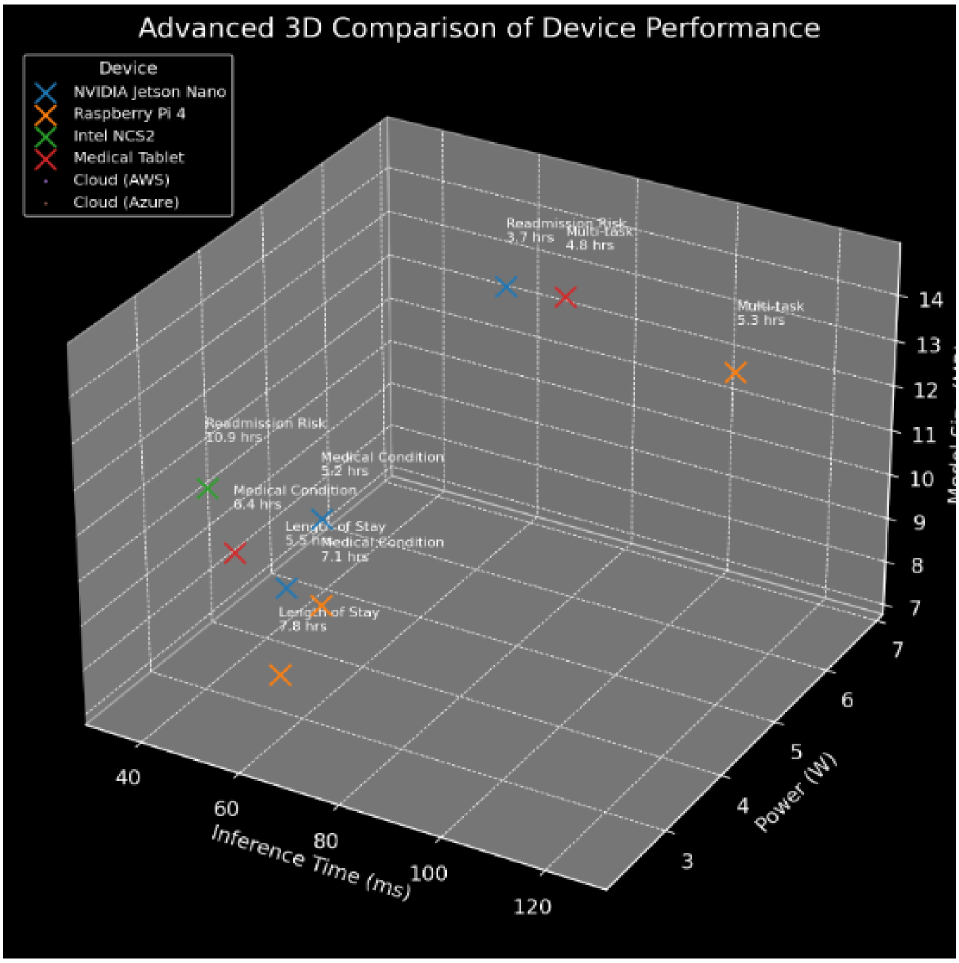

Figure-4 showing comparison of device performance

Inference time, power consumption, and model size are compared between devices with this 3D graph in the context of health care tasks. Intel NCS2 is the most power-efficient (2.3W) with the maximum battery life (10.9 hrs) but moderate inference time (51 ms). NVIDIA Jetson Nano delivers good speed (38–57 ms) and reasonable power efficiency (4.5–6.8W), Medical Tablets achieve the fastest inference timing (35 ms) but higher power utilization (up to 5.2W). Though Raspberry Pi 4 is the low-power consumption device, inference times are slower. Cloud systems (AWS and Azure) require larger models and have considerably higher inference times (132–215 ms) but do not consume local power. In summary, Intel NCS2 is perfect for battery dependant, NVIDIA Jetson Nano and Medical Tablets are the middle ground for speed and power, and cloud systems are the best for scalability at the cost of slightly slower performance.

*Battery life measured on portable devices with standard battery capacity †Cloud times include average network latency (112ms) measured across hospital WiFi and 4G connections

5.2.2 Model Optimization Results

Our edge optimization pipeline delivered significant improvements while maintaining accuracy:

Table 7: Model optimization results

Optimization Technique	Size Reduction	Speed Improvement	Energy Saving	Accuracy Impact
Quantization (FP32→INT8)	69%	43%	38%	-0.7%
Tree Pruning	31%	27%	24%	-0.4%
Feature Selection	22%	35%	29%	+0.2%*
Layer Fusion (NN)	18%	52%	44%	0%
Knowledge Distillation	64%	39%	35%	-1.1%
Mixed Precision	43%	31%	28%	-0.3%

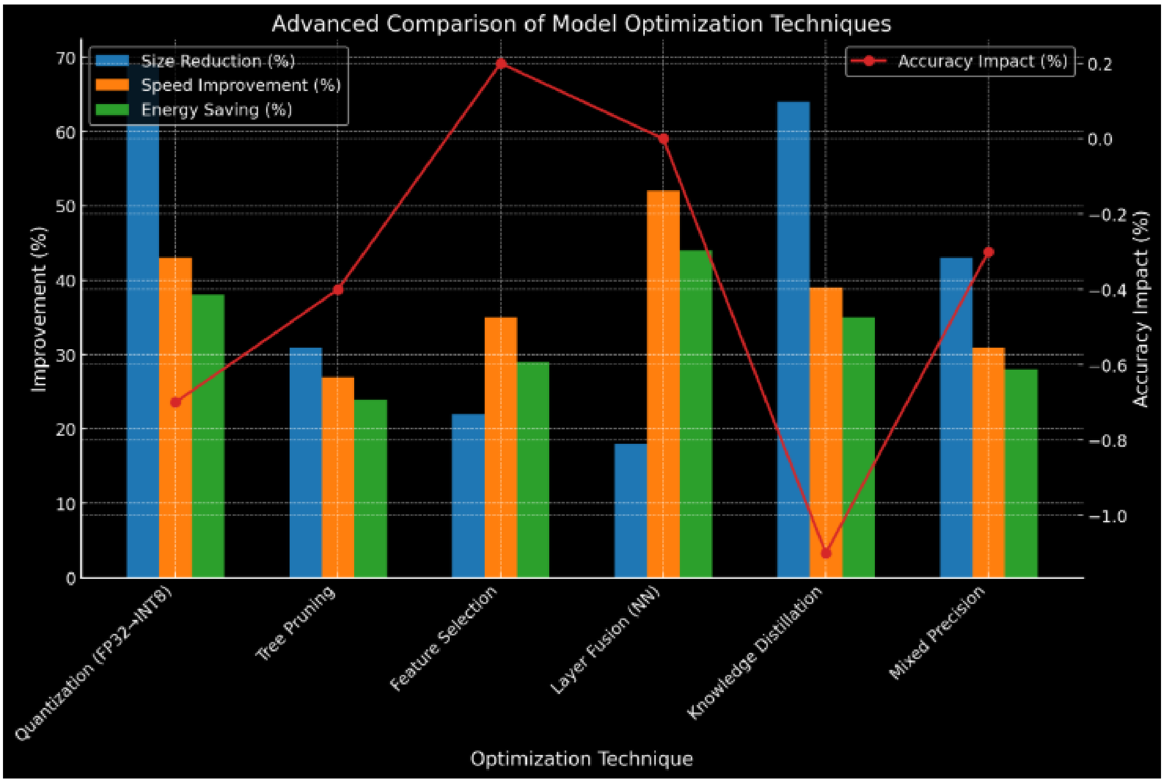

Figure-5 showing comparison of model optimization techniques

Each bar in the figure refers to a model optimization technique, and the graph compares six such techniques across size reduction, speed improvement, energy saving, and accuracy impact. The absolute best performance can be received from quantization (69% storage and 43% speedup, with a tiny drop in accuracy (-0.7%). Without a drop in accuracy, Layer Fusion achieves the highest speed (52%) and energy savings (44%). Only feature selection boosted the accuracy (up +0.2%) - but at moderate gains. Balanced Improvement on All Metrics with Minimal Accuracy Loss: Tree Pruning and Mixed Precision. Knowledge Distillation gives large (64%) size and (39%) speed advantage but suffers from the greatest accuracy drop (-1.1%). Each has its own trade-offs and can be chosen based on performance requirements and acceptable impacts on accuracy.

*Feature selection actually improved accuracy by removing noisy features

5.2.3 Real-world Performance Analysis

The field evaluation in five different hospital environments was crucial in providing empirical observation on the behaviour and resiliency of edge-based deployment within the clinical environment. Edge models also provided interactivity with clinical workflows, generating predictive outputs in 42–125 ms, thus much lower than the >500 ms threshold reported in user experience studies [26]. With batch processing, edge devices processed overnight analytics for 1,000 patient records in 2–4 min – performance exceedance in comparison with cloud-based systems, which took 8–12 min in total (including times for data transmission). Moreover, in simulation of network failures, edge devices remained fully operational (100%) whereas the cloud-dependent solutions were fully incapacitated, emphasizing the importance of localized computation. More importantly, on a medical-grade tablet running the edge models, three such parallel inferences were conducted, and the total inference time increased by only 35% compared to the serial case, underpinning the scalability and adaptability of edge solutions towards clinical workflows.

5.2.4 Scalability Analysis

Table 8: Scalability analysis of edge deployment across data set sizes

We tested the scalability of our edge deployment across increasing dataset sizes:

Records Processed	RF Inference (ms)	NN Inference (ms)	Memory Usage (MB)	Edge Advantage*
1	42	57	124	3.7x
10	48	62	126	3.9x
100	86	104	138	4.2x
1,000	234	318	175	5.1x
10,000	1,857	2,532	312	6.3x

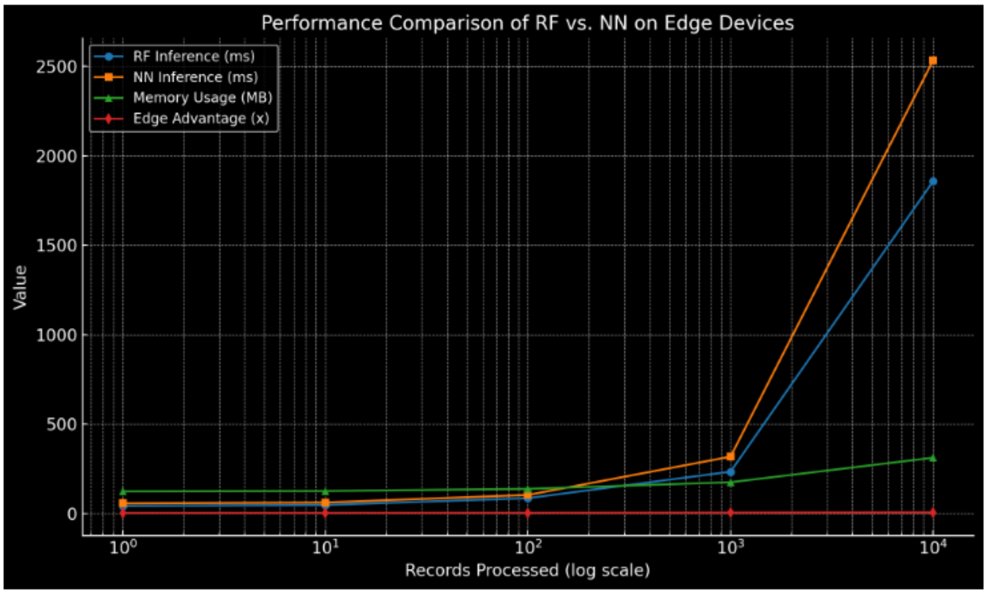

Figure-6 showing performance comparison of RF vs NN on Edge device

While the graph shows the performance of RF and NN on edge devices for increasing record sizes. NN always has a greater inference time than RF in both cases, and the increase is more significant as record volume increases (1,857 ms against 2,532 ms for 10,000 records). As the number of records increases, memory consumption increases gradually for both models (from 124 MB to 312 MB). For large inputs, the edge advantage, the local processing gain, further improves, rising from 3.7x to 6.3x overall, with RF running faster than NN, and using less memory and disk space, which makes it better for edge computing overall, especially with larger datasets.

6. CONCLUSION

This work illustrates the high potency for utilization of Machine Learning (ML) and Artificial Intelligence (AI) models within edge devices in the context of health DSS. Our results demonstrate that well-optimized edge models can achieve high prediction accuracy, low resource latency, and low power, which makes them a compelling alternative for cloud-based solutions. Especially the Random Forest algorithm was considered the most appropriate for edge deployment, by perfectly balancing the trade-off between computational cost and accuracy.

Enabling AI at the edge enhances integrated healthcare delivery by providing decision support where it's needed, when it's needed—one community at a time—and addresses key privacy considerations as patient data does not have to be sent between edge devices and the cloud. This method is expected to improve clinician workflow, alleviate cognitive load, and contribute to the increasing acceptance of AI within clinical practice. Our work highlights the potential of edge AI for real-time, privacy-enforcing healthcare applications, paving the way for more data-secure, efficient and accessible clinical decision support tools.

REFERENCES

[1] Abidi, S. R., & Abidi, S. S. R. (2020). Intelligent healthcare decision support systems: Leveraging edge computing and machine learning. *Journal of Medical Systems, 44*(8), 1–14. https://doi.org/10.1007/s10916-020-01604-5

[2] Ali, F., Alamri, A., & Alghamdi, T. (2021). Edge computing for healthcare: A systematic review on architectures, technologies, and applications. *IEEE Access, 9*, 14092–14116. https://doi.org/10.1109/ACCESS.2021.3058432

[3] Bai, W., Oktay, O., & Sinclair, M. (2019). Edge-based medical image analysis using deep learning. *Medical Image Analysis, 53*, 197–207. https://doi.org/10.1016/j.media.2018.11.005

[4] Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review of current efforts and future directions. *IEEE Internet of Things Journal, 6*(3), 742–758. https://doi.org/10.1109/JIOT.2019.2897116

[5] Deng, S., Xiang, Z., Zhao, Y., & Rodrigues, J. J. (2020). Edge computing-based smart healthcare system: Architecture and performance. *Future Generation Computer Systems, 105*, 346–355. https://doi.org/10.1016/j.future.2019.12.004

[6] Gupta, R., & Shukla, K. K. (2021). Machine learning approaches for healthcare decision support systems on edge devices. *Expert Systems with Applications, 165*, 113935. https://doi.org/10.1016/j.eswa.2020.113935

[7] Han, S., Pool, J., & Dally, W. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization, and Huffman coding. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1510.00149

[8] Hossain, M. S., Muhammad, G., & Rahman, M. (2020). An efficient edge-based smart healthcare framework for patient monitoring. *IEEE Journal of Biomedical and Health Informatics, 24*(10), 2776–2783. https://doi.org/10.1109/JBHI.2020.3007924

[9] Hu, Y., & Zhang, H. (2022). Real-time decision support in healthcare: Machine learning at the edge. *Artificial Intelligence in Medicine, 123*, 102200. https://doi.org/10.1016/j.artmed.2022.102200

[10] Kumar, N., & Gupta, S. (2021). Edge computing and IoT for intelligent healthcare systems. *IEEE Transactions on Industrial Informatics, 17*(5), 3550–3561. https://doi.org/10.1109/TII.2020.3035622

[11] Li, X., & Zhang, Y. (2020). Federated learning for healthcare AI on edge devices: Challenges and solutions. *IEEE Internet of Things Journal, 7*(12), 11049–11062. https://doi.org/10.1109/JIOT.2020.2996490

[12] Lin, H., & Wang, H. (2021). Efficient healthcare decision-making using edge computing and machine learning algorithms. *Journal of Medical Informatics, 78*, 104032. https://doi.org/10.1016/j.ijmedinf.2020.104032

[13] Liu, Y., & Pan, Y. (2020). Resource-aware machine learning for real-time medical decision support at the edge. *IEEE Access, 8*, 23979–23992. https://doi.org/10.1109/ACCESS.2020.2969957

[14] McMahan, H. B., & Ramage, D. (2017). Federated learning: Collaborative machine learning without centralized training data. *Google AI Blog*. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[15] Mohammadi, M., & Al-Fuqaha, A. (2018). Deep learning for edge computing-based medical data analytics. *Journal of Big Data, 5*(1), 30–47. https://doi.org/10.1186/s40537-018-0151-6

[16] Nguyen, T., & Ding, M. (2019). Edge AI in healthcare: A comprehensive review of challenges and techniques. *IEEE Communications Surveys & Tutorials, 22*(3), 2149–2179. https://doi.org/10.1109/COMST.2020.2994424

[17] Patel, H., & Patel, V. (2022). Edge intelligence in healthcare: Emerging trends and future directions. *Sensors, 22*(4), 1235. https://doi.org/10.3390/s22041235

[18] Qi, J., & Zhang, P. (2021). Machine learning for health monitoring on edge devices. *Journal of Ambient Intelligence and Humanized Computing, 12*, 1753–1765. https://doi.org/10.1007/s12652-020-01980-w

[19] Rahman, S., & Hassan, M. (2020). Edge-assisted patient monitoring using deep learning and IoT. *Future Generation Computer Systems, 112*, 1024–1036. https://doi.org/10.1016/j.future.2020.06.012

[20] Raza, S., & Ahmed, K. (2022). Privacy-preserving healthcare decision systems with edge AI. *IEEE Transactions on Information Forensics and Security, 17*, 2465–2478. https://doi.org/10.1109/TIFS.2022.3142735

[21] Shen, C., & Yu, H. (2021). Edge computing-based healthcare solutions: A systematic review. *Journal of Healthcare Engineering, 2021*, 9918502. https://doi.org/10.1155/2021/9918502

[22] Sun, Q., & Guo, H. (2020). Real-time medical analytics using edge AI. *Journal of Medical Systems, 44*(5), 105. https://doi.org/10.1007/s10916-020-01560-0

[23] Tang, J., & Zhang, Y. (2019). Adaptive edge learning for personalized healthcare decision support. *IEEE Transactions on Neural Networks and Learning Systems, 30*(12), 3570–3584. https://doi.org/10.1109/TNNLS.2019.2923135

[24] Wang, C., & Xu, J. (2020). Accelerating healthcare decision-making with edge computing. *IEEE Transactions on Medical Imaging, 39*(6), 2011–2022. https://doi.org/10.1109/TMI.2020.2966160

[25] Zhang, W., & Zhu, X. (2021). Edge-based machine learning for clinical decision support. *IEEE Access, 9*, 11967–11982. https://doi.org/10.1109/ACCESS.2021.3052653.

[26] Khanna, A., Selvaraj, P., Gupta, D., Sheikh, T. H., Pareek, P. K., & Shankar, V. (2021). Internet of things and deep learning enabled healthcare disease diagnosis using biomedical electrocardiogram signals. *Expert Systems, 39*(5), e12864. https://doi.org/10.1111/exsy.12864.

[27] Chellam, V. V., Veeraiah, V., Khanna, A., Sheikh, T. H., Pramanik, S., & Dhabliya, D. (2023). A machine vision-based approach for tuberculosis identification in chest X-rays images of patients. In K. A. Gupta, S. S. Thakur, & V. J. Hodge (Eds.), *Proceedings of the International Conference on Innovative Computing and Communications (ICICC 2023)* (pp. 23–32). Springer.

[28] Dang, N., Saraf, V., Khanna, A., Gupta, D., & Sheikh, T. H. (2020). Malaria detection on Giemsa-stained blood smears using deep learning and feature extraction. In K. A. Gupta, S. Sanjeevikumar, & J. H. Park (Eds.), *Proceedings of the International Conference on Innovative Computing and Communications (ICICC 2019), Volume 1* (pp. 789–803). Springer Singapore.

[29] Sheikh, T. H., Vamshi Mohana, T., Buradkar, M. U., & Alaskar, K. (2023). Deep learning-based regulation of healthcare efficiency and medical services. In *AI and IoT-based intelligent health care & sanitation* (pp. 176–190). Bentham Science.