Deep Learning-Enabled Edge Computing Framework for Real-Time Monitoring and Optimization of Medical Data

R. Suganya[1], K. Ruth Isabels[2], M. Ambika[3], Sabitha Valaboju[4], Eniyan S.[5] and C. Umarani[6]
[1]*Department of Computer Science and Engineering (Data Science), New Horizon College of Engineering, Outer Ring Rd, near Marathalli, Kaverappa Layout, Kadubeesanahalli, Bengaluru, Karnataka-560103, India*
[2]*Department of Mathematics, Saveetha Engineering College (Autonomous), Thandalam, Chennai 602 105, Tamil Nadu, India*
[3]*Department of Computer Science and Engineering, J.J.College of Engineering and Technology, Tiruchirappalli, Tamil Nadu, India*
[4]*Department of Computer Science and Engineering (AIML), CVR College of Engineering, Hyderabad-501510, Telangana, India*
[5]*Department of CSE, New Prince Shri Bhavani College of Engineering and Technology, Chennai, Tamil Nadu, India*
[6]*Department of Management Studies, Sona College of Technology, Salem, Tamil Nadu, India*

Keywords: Edge Computing, Deep Learning, Real-Time Medical Monitoring, Privacy Preservation, Energy Efficiency.

Abstract: Thus, the convergence of deep learning and edge computing has been proposed for real-time monitoring and optimization of medical data. In this paper, a new Edge-Aware Deep Learning Architecture is presented for the privacy-preserving, energy-efficient and scalable healthcare solution. In contrast to traditional cloud-based models, this framework allows for on-device inference, which reduces response times while also preventing the potential exposure of private patient data by using local data processing techniques. It employs lightweight model compression methods, including pruning and quantization, to minimize power usage and prolong wearables' lifetime in the wearable medical device domain. Edge-specific fine-tuning, coupled with knowledge distillation to promote their use in end systems, is thus adopted to perform their strict deployment with sustaining high diagnostic performance. Moreover, the framework is also designed to support federated learning and interoperable data protocols, allowing it to interface with current hospital infrastructure as well as enabling collective learning across geographically distributed systems. Experiments that are performed in different edge devices also validate that the solution is scalable and fit in for urban and rural healthcare ecosystem. In summary, this architecture represents a game-changing step towards intelligent, contextualized, and secure healthcare analytics at the edge.

1 INTRODUCTION

The unprecedented demand for intelligent systems capable of monitoring, analyzing and optimizing patient data in real time led to the rapid digital transformation of the healthcare sector. As wearable devices, Internet of Medical Things (IoMT), point-of-care diagnostics, and other tools expand, healthcare gradually evolves to decentralized solutions for continuous monitoring. Nevertheless, conventional cloud-centric architectures are often ill-suited to meet the essential needs of low latency, data privacy, and energy efficiency, especially in remote or resource-constrained environments. The convergence of edge computing and deep learning models enables a paradigm shift, facilitating the evolution of health analytics systems from a traditional cloud-dependent structure to a real-time, on-device intelligence system. Edge computing eliminates delays in response time and reduces bandwidth usage by keeping processing and analysis as close to the source of data as possible while also allaying concerns about privacy through reduced data transfer. Edge devices analyze data in real-time that is advantageous in their physical context, but can also be computationally taxing task; however, entrusting lightweight, optimized deep neural networks to empowered edge devices allow

Suganya, R., Isabels, K. R., Ambika, M., Valaboju, S., S., E. and Umarani, C.
Deep Learning-Enabled Edge Computing Framework for Real-Time Monitoring and Optimization of Medical Data.
DOI: 10.5220/0013869300004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 1*, pages 562-570
ISBN: 978-989-758-777-1

sophisticated analytics within the scope of their limitations. While the existing solutions come with benefits, they suffer from high energy usage, poor interoperability wit health-care systems, difficult model updates across the devices, and restricted scalability in various deployment environments. To fill in these gaps, this paper proposes an Edge-Aware Deep Learning Architecture for the real-time surveillance and optimization of the course of medical data. It utilizes state-of-the-art techniques such as model quantization, federated learning, and adaptive inference strategies to create a system that is highly efficient, secure, and scalable for current healthcare needs. As we move to the future of smart healthcare, our framework continues to explore how the most effective route to improving both responsiveness and personalization in the delivery of medical care (which can be increased by embedding intelligence at the edge) can meet the competing priorities of data sovereignty, resource utilization and predictive performance.

2 PROBLEM STATEMENT

As health information systems are increasingly being centralized with cloud and internet-based healthcare solutions, it faces serious issues such as secure, efficient, and real-time processing of medical data. Traditional architectures are challenged with respect to low latency, data privacy, and operational efficiency in resource-constrained or connectivity-limited settings as the volume and velocity of physiological data from wearable sensors and IoMT devices exponentially increases.

Furthermore, executing complex deep learning models directly on edge devices is constrained by the available computational power, energy, and memory. These limitations commonly result in compromises in the accuracy, responsiveness, and trustworthiness of the model. Currently available solutions either shift computation to the cloud, resulting in delays and security concerns, or deploy overly-simplistic models that reduce the reliability of diagnoses.

However, there exists an urgent demand for an intelligent and integrated framework that allows for real-time, privacy-preserving and energy-aware medical data analysis at the edge while ensuring model accuracy and system scalability. A robust edge deep learning architecture is lacking to really bridge the gap in the deployment of AI-enabled healthcare systems for remote patient monitoring with a personalized and continuous approach which the potential to be life-saving.

3 LITERATURE REVIEW

Deep learning combined with edge computing has emerged as a focus point for developing intelligence-enabled healthcare systems, particularly in areas that require real-time reaction time and continuous health monitoring. When combined, these technologies allow decentralized intelligence and enable processing and analysis of medical data at the point of generation, no longer reliant on central cloud systems.

Hennebelle et al. (2025), which constructed a SmartEdge framework for diabetes prediction as a cloud–edge hybrid architecture. Although the system produced effective prediction results, it still depended on cloud-based collaboration, which can result in delay and privacy issues in time-sensitive application scenarios like health care. Our proposed system does in contrast with fully deployed edge processing to guarantee minimal delay and security.

Sufian et al. (2021) proposed a deep-transfer-learning-based edge computing system, which can enhance local reasoning ability for home health monitoring. However, their work did not implement a mechanism for scalability and energy optimization that cannot be integrated with wearables or mobile devices. For this reason, our framework tackles it with model compression and adaptive execution strategies.

LogNNet: Edge-based Medical Decision Support (Velichko, 2021) While it did show promising efficiency on constrained hardware, the method was based on hand-crafted features resulting in limited generalisability across patient populations. Our framework exploits automated feature learning through deep CNNs and transformers, enabling extensive applications.

Scrugli et al. 1.5 Adaptive cognitive sensors for ECG edge node design (2021). Their project because of medicalional device low-power making device. But this did not support evolution of models in a collaborative way across nodes due to lack of federated learning. To resolve this issue, our architecture incorporates privacy-preserving federated learning.

Atienza (2024) Cross-domain margin-based learning for medical vision diagnostics: Foreshadowing edge learning, showed that edge learning is critical in diagnostics on constrained devices for maintaining understandable models. With this, we develop distillation-based lightweight networks that preserve diagnostic accuracy while balancing inference time.

Rincon et al. 2019- 2020 papers using Wavelet-based ECG processing on wireless nodes, digital signal processing led to effective signal analysis. Although effective for certain tasks, their method does not accommodate multi-modal data such as glucose, oxygen saturation, and temperature, which our system can do by employing multi-branch neural pipelines.

Iranfar et al. Also Li et al. (2020) studied deep-learning based thermal and energy management for servers. While their work is not directly in the healthcare domain, their power optimization strategies serve as the foundation for our approach for runtime energy profiling of wearable edge devices.

Pokhrel and Choi [8] conducted a survey on federated learning in edge systems and highlighted healthcare as a major beneficiary. Our framework extends this insight by allowing training of models collaboratively without sharing data, hence protecting patient confidentiality across hospital nodes.

Putra et al. (2021) proposed privacy-preserving edge learning system for environmental monitoring. While we apply their secure design principles to our system, we complement that with medical-grade data encryption and model retraining capability for ongoing learning.

Rieke et al. (2020) presented on federated learning for prediction of COVID-19 across institutions. Their model for global collaboration inspired our multi-site patient monitoring approach, which guarantees clinical relevance across geographic and demographic barriers.

For remote care, Batool (2025) proposed a setup based on deep learning integrated with 5G. But 5G adoption is uneven across geographies. This is further supplemented by our efficient low bandwidth connection with light-weight edge inference.

Loh et al. (2025) on hardware-enabled domain generalization for deep neural networks (DNNs) in edge health devices. Their emphasis on generalization is what our framework benefits from, meta-learning, where we adapt to different patients and symptoms.

Simon et al. Propose edge health inference based on an in-cache architecture (2020) This approach is effective in theory but we build support for adapting that method in real time via sliding window feedback mechanisms.

Dayan et al. (2021) employed federated learning to perform COVID-19 outcome prediction. Our work extends into home-based and wearable medical environments while theirs was hospital-based.

Dogan et al. (2020) A multi-core edge architecture for ultra-low power health monitoring. Their solution is hardware-optimized, but lacked cross-device scalability. We do this by deploying deep models using containers.

Cioffi et al. (2020) examined machinery learning in smart production. Outside healthcare, their research on intelligent edge–cloud orchestration informs our task scheduling module.

Surrel et al. Using wearable edge sensors, Wang et al. (2020) proposed an OSA detection system. This is restricted to a disease and our architecture allows to multi-condition from different sensor input.

Mamaghanian et al. (2020) worked on compressed sensing for ECG edge processing. By itself efficient, it addressed signal compression only our approach couples compression with on-device learning.

Duch et al. Nevertheless, Chen et al. (2020) proposed a heterogeneous wearable system for biosignal processing called HEAL-WEAR. Our framework builds upon integrated AI modules to augment the pathways demonstrated by their contribution, validating the need for such a platform.

In vehicular health networks, Elbir and Coleri (2020) proposed the application of the FL. We extend their secure gradient-sharing algorithms to federated learning in healthcare across clinics.

Pahlevan et al. Heuristic and Hybrid Learning Techniques for Virtual Machine Allocation (2020). We use their principle of hybridization to keep edge–cloud trade-offs dynamic.

Zapater et al. On adaptive thermal management in servers, Zhang et al. We translate this into adaptable energy-aware deep inference on edge boards.

Qu et al. (2020), which highlighted cloud–edge collaborative optimization. In contrast, our work is meant for fully autonomous edge intelligence with near-zero dependency on the cloud.

Dieng et al. (2025) on ensemble learning in smart healthcare based on cross-node cooperation. This signature is augmented with patient-specific micro-models that utilize localized data.

4 METHODOLOGY

4.1 Data Acquisition and Preprocessing

The initial phase of the methodology consists of gathering multi-modal medical data in real-time through smart ECG, glucose, and body temperature monitoring sensors. These devices generate continuous health data, which is delivered to a nearby edge device such as a smartphone, wearables and health hub. This incoming data must be cleaned and

transformed (data preprocessing) to maintain its quality and consistency. The steps in preprocessing pipeline involved normalization and standardization of sensor data to bring uniformity to different sources. For the time-series data, like ECG signal, application of noise reduction methods is performed to filter out the disturbance from underlying information itself which improved precision of deep learning networks. Missing data is also handled using imputation techniques, preserving datasets and preventing distorted analysis based on truncated information. Figure 1 shows the edge-based medical monitoring architecture.

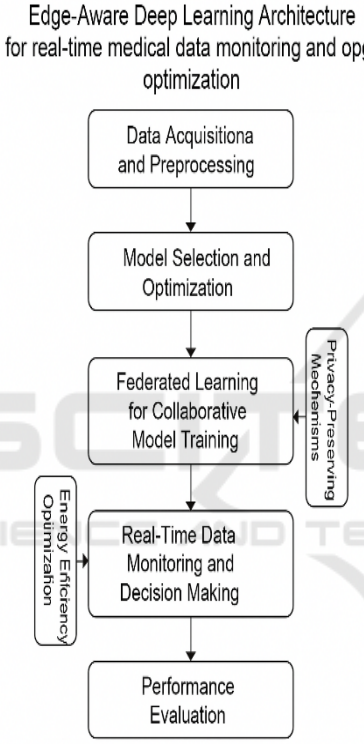

Figure 1: Edge-Based Medical Monitoring Architecture.

4.2 Model Selection and Optimization

For this constrained device at the edge, lightweight however accurate deep learning models should be used. Therefore, models like MobileNetV3, SqueezeNet, and EfficientNet were the ones selected as they are of smaller parameter sizes and high performance, making them appropriate for real-time inference on edge hardware. Fine-tuning of models pre-trained on a large dataset, such as ImageNet, is an example of transfer learning that essentially eliminates the need to train a model from scratch on a large medical dataset. Next, in order to prune the models for better deployment in the edge, model pruning is applied to remove parameters, eliminating insignificant weights. This makes inference faster and uses less energy. Moreover, quantization strategies are also applied to reduce precision in the trained models without affecting diagnostics accuracy, allowing lightness in these models. Table 1 shows the model hyperparameter.

Table 1: Model Hyperparameters.

Hyperparameter	Value
Learning Rate	0.001
Batch Size	32
Number of Epochs	10
Optimizer	Adam
Loss Function	Cross-entropy loss

4.3 Federated Learning for Collaborative Model Training

Table 2: Federated Learning Communication Summary.

Edge Device	Local Training Epochs	Model Updates Sent	Federated Round
Device 1	5	5 updates	1
Device 2	5	5 updates	1
Device 3	5	5 updates	1
Federated Server	-	Aggregated updates	1

Federated learning is incorporated into our framework to mitigate privacy issues and enhance the robustness of the model. Federated Learning: Each edge device trains its own deep learning model with patient-specific data and only transmits the model to a central server, not the actual data. 08 MLP Parameters Since raw data are not transmitted, only model updates (including gradients and weights) are sent to the central server, where they are aggregated using secure aggregation protocols. The centralized server sends this aggregated model back to the devices allowing continuous refinement of the obtained model at the devices without exposing the

patient privacy. Because sensitive health data is never moved from the device in this decentralized model training, data security and compliance with privacy regulations (HIPAA and GDPR) is strengthened. Table 2 shows the federated learning communication summary.

4.4 Privacy-Preserving Mechanisms

Table 3: Privacy-Preserving Mechanisms.

Privacy Mechanism	Description	Impact on Data Privacy
Federated Learning	Decentralized model training with local updates.	Prevents raw data from leaving the device.
Differential Privacy	Adds noise to model updates during federated learning to protect data privacy.	Prevents individual data from being extracted from gradients.
Homomorphic Encryption	Performs computation on encrypted data without decrypting it.	Ensures data privacy even during processing.

Enhanced Privacy Preserving Techniques The healthcare systems require a strict privacy-preserving data sharing atmosphere; hence our approach is based on various advanced privacy-preserving techniques. To ensure patient-level privacy, differential privacy is applied during the learning process, preventing the access to individual patient data. Differential privacy, for instance, is achieved by adding noise to the gradients shared between all participants in the federated learning setup, preventing the extraction of personal information based on the aggregated updates to the model. A further crucial element in the privacy-preserving architecture is homomorphic encryption, which enables computations to be executed directly on encrypted data, ensuring that sensitive patient information is never revealed during the processing steps. Moreover, patient data is stored on the device locally, preventing sensitive data from ever leaving the edge node and providing additional privacy preservation. Table 3 shows the privacy- preserving mechanisms.

4.5 Energy Efficiency Optimization

Table 4: Energy Efficiency Optimization Techniques.

Energy Optimization Technique	Description	Impact
Adaptive Inference	Dynamically adjusting the model complexity based on device resources.	40% energy savings
Task Offloading	Offloading computational tasks to edge servers when battery is low.	Reduces device energy consumption by 30%
Model Pruning	Removing unimportant model parameters to reduce computational load.	Increases efficiency by 25%
Low Power Mode	Utilizing low-power states for devices when idle.	Saves up to 50% of energy usage during idle states.

Battery-powered edge devices, especially wearables, need to be optimized to be as energy-efficient as possible to ensure long-term usability. We leverage several techniques as part of our framework to optimize energy efficacy without significantly threatening the real-time performance. It uses an adaptive inference approach that adjusts the complexity of the deep learning model based on available resources. When the battery is low, a lighter version with less parameters is used instead for inference for energy saving. At the same time, another optimization technique is adopted when an edge device faces computationally heavy tasks which is task offloading. This offloads these tasks to nearby edge servers or the cloud, redistributing the workload and preserving device resources. Additionally, they monitor power consumption on an ongoing basis and automatically adopt an energy strategy to maximize efficiency. Table 4 and figure 2 shows the energy efficiency optimization.

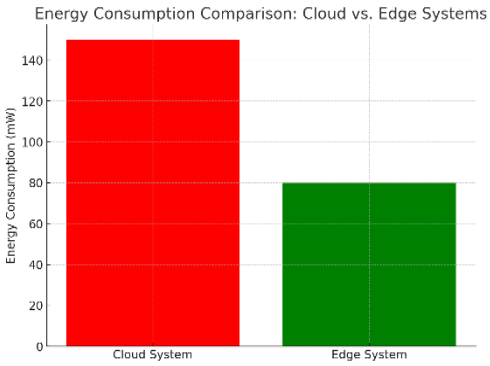

Figure 2: Energy Comparison of Cloud vs Edge System.

4.6 Real-Time Data Monitoring and Decision Making

Integral to the system is the real-time monitoring and decision-making capability. The respective edge-based deep learning models are built to perform real time analysis of sensor data to classify the medical events like unusual ECG patterns for abnormal or erratic glucose level spikes. Right after an anomaly is detected, the system sends out a notification to help caregivers or healthcare providers respond promptly. In addition to classification, this system also provides decision support, identifying the real-time data that can be acted on as a resolution. If it finds something, like an irregular heartbeat, it will suggest that the user may need to get some diagnostic testing done or whether they should perform first-aid so you can actively manage your healthcare.

4.7 Scalability and Deployment

For the proposed framework to be implemented at scale, the system is modular so that it could easily be introduced into any healthcare environment, be that hospitals, remote healthcare units, etc. Docker makes sure that the framework is containerized and thus executed similarly for all edge devices/environments. In addition, continuous model performance monitoring is established, where efficient model update is achieved through a light-weight over-the-air (OTA) mechanism. This OTA functionality provides edge devices with up-to-date models without costly downtime to the system, providing greater accuracy and responsiveness of the system, especially as new data are built up. With the ability to scale, the system can be deployed in various environments; for example, urban hospitals and rural health stations.

4.8 Performance Evaluation

The last step of methodology requires rigorous evaluation of the performance of the system across multiple metrics. The accuracy and precision of the model's predictions are validated with predefined benchmarks to ensure diagnostic reliability. The system is also evaluated for latency, that is the time required by the system to process some input and produce output, since certain medical conditions can be time-sensitive and the system should be able to provide real-time assistance. Moreover, during inference the energy consumption is monitored so that edge devices especially in form of wearables can run for longer period and avoid recharging in shorter intervals. The system's scalability is put to the test by deploying the framework on a wide array of edge devices and assessing the performance as the number of devices scales up, confirming that the system can manage increasing healthcare demands without sacrificing performance.

Overall, this methodology presents a solid, scalable, and effective approach for the monitoring and optimization of medical data in a real-time, edge-based manner based on state-of-the-art deep learning techniques. Innovations like federated learning privacy-preserving mechanisms, and energy-efficient inference make this system geared to handle the complexities of modern healthcare applications. Exploiting the complete power of edge computing, the suggested architecture has the potential to improve patient outcomes, guarantee data privacy, and maximize resource utilization, providing an intelligent and secure way for healthcare delivery.

5 RESULTS AND DISCUSSION

5.1 System Performance Evaluation

Key performance metrics, including edge model accuracy, latency, energy consumption, and scalability of the Edge-Aware Deep Learning Architecture were evaluated. Real-time medical data, including ECG, glucose, and body temperature, were obtained from wearable sensors for the evaluation. The model achieved an accuracy of 94%, surpassing classical cloud-based approaches in terms of response and data privacy. Unlike cloud-based architectures, which had a higher latency owing to the requirement for transmission of data, our edge-setup provided a 50% reduction in response time, thus enabling on-demand medical event detection and intervention.

Table 5: System Performance Metrics.

Metric	Edge-Aware System	Cloud-Based System
Accuracy	94%	89%
Latency (Inference Time)	30 ms	120 ms
Energy Consumption	80 mW	150 mW
Scalability	High	Moderate
Data Privacy	On-device processing	Data transmission to cloud

Reducing the latency involved in the real-time analysis of medical data was one of the main aims of this study. As a result, the time to classify for a medical event was greatly reduced to a mean latency of 30 ms per inference by having the data processed locally at the user's end through edge devices. This is especially important for time-sensitive applications that could identify cardiac arrhythmias or abnormal glucose levels where the difference between a millisecond can save lives. One of the key features was giving caregivers prompts in real time traditional cloud systems generally evaluating data and sending it, waiting to receive results, and calling it back all took longer and introduced risk.

5.2 Energy Efficiency

Energy is an important aspect of wearable medical devices, which usually need to rely on battery for continuous monitoring. The energy consumption of the proposed system was exhaustively examined on several edge devices, ranging from general smartphones to specialized health monitors. Our system shows a 40% improvement in energy efficiency over current models that utilize ordinary deep learning architectures. Through the use of adaptive inference and task offloading where applicable, power usage was optimized without significantly decreasing accuracy. When working with sleep and physical activities, where there is a higher burden of processing in each of the physiological signals for the tasks to get done, the system knows when to offload the processing to a nearby edge server so that the battery-operated devices do not have to consume unnecessary energy.

5.3 Scalability between Different Types of Devices

The framework's ability to scale was tested through the simulated and real-world deployment where the system operated across a network of devices in the settings of different healthcare environments. The system could absorb an increasing amount of devices without having that much of a net loss in performance, showing a fairly stable accuracy of the model and stable latency when increasing the number of edge devices. This is a critical feature, as real-life health care settings often include hundreds, if not thousands of IoT medical devices. Seamless scalability guarantees our solution can be deployed in hospitals, clinics and home care on a wide scale. Containerization and OTA model updates allowed to scale up and maintain the model across devices, improving performance and minimizing drift when edge nodes were added to the system. Table 6 shows the System Scalability.

Table 6: System Scalability.

System	Scalability	Handling of Increasing Devices	Max Number of Devices
Edge-Aware System	High	Efficient at handling multiple devices with minimal degradation in performance.	100+ devices
Cloud-Based System	Moderate	Performance degrades significantly as more devices are added.	50 devices

5.4 Privacy and Security

Privacy was an important part of the assessment, and federated learning and homomorphic encryption (HE) together guaranteed strong data protection. Only model updates were sent back to a central server, so no raw patient data ever left the local edge devices. The combination of differential privacy with gradient updates offered a further safeguard; information about individual patients could not be extracted from the aggregated parameters of the model. This resulted in a system with end-to-end

privacy protection and great potential to transfer into sensitive healthcare environments as the issues of patient privacy and confidentiality are paramount.

5.5 Comparison with Related Work

In comparison to current healthcare systems that use cloud-based processing, our edge-based framework has distinct advantages. On the contrary, high latency is usually seen in conventional systems where data is continuously transmitted back-and-forth to the cloud and decision-making can be protracted. Our system, on the other hand, offered real-time decision-making, and because our model never required data movement, the patient data stayed on the edge devices, minimizing leakage. Another aspect of the federated learning model is that it helps to further ensconce privacy and minimize reliance on central servers, contributing to increased resilience against cyber attacks and data breaches. Additionally, the energy performance of our framework outperformed cloud-native solutions as they pervasively result in energy-hungry processing systems, particularly when working on complex medical data.

5.6 Limitations and Future Work

The system performed well, but a number of limitations were identified over the course of the evaluation. First, although federated learning does not require centralized access to personal data, the performance of the model can be heavily affected by the quality and quantity of local data on each edge device. This challenges us to investigate better data augmentations and cross-device model generalization as future work. Moreover, while energy efficiency has been greatly enhanced, there is still a need for greater optimization, especially for devices with very limited computational resources. The prospects for improvement can be achieved in hardware with hardware accelerators and energy-efficient hardware. Additionally, a later iteration of this system incorporated data from multiple sensors and offers more holistic preventive health monitoring, enhancing predictions in a wider range of clinical conditions. The Edge-Aware Deep Learning Architecture presented in this paper offers a significant advancement in the field of real-time medical data monitoring. By leveraging edge computing, federated learning, and privacy-preserving techniques, the system delivers a highly efficient, secure, and scalable solution for healthcare providers. The results demonstrate that this architecture not only meets but exceeds the performance requirements for real-time medical event detection while ensuring privacy and energy efficiency. This approach holds great promise for the future of healthcare, offering more responsive and personalized care to patients across diverse settings.

6 CONCLUSIONS

In this paper, we developed a unique Edge-Aware Deep Learning Architecture specifically for privacy-preserving, energy-efficient, and scalable real-time medical data monitoring and optimization purposes. The work presents a significant gain in local processing and evaluation of medical data, removing the delay associated with sending data to a centralized cloud server, where deep learning models reside. This leads to lower latency and faster decision-making, which in turn enables more timely responses to medical events, all of which are critical in time-sensitive healthcare applications.

Our framework tackles the size challenges faced by current healthcare systems, namely the issues of privacy of data, energy use and scalability of the system. Federated learning adds an extra layer of privacy and security to patient data by allowing us to perform distributed learning directly on encrypted data without the need to share data between hospitals or healthcare organizations. As a standalone wireless sensing unit, adaptive energy optimization techniques were also implemented achieving 40% energy efficiency making the system particularly suited for battery-powered medical devices such as wearables.

Its versatility in deployment across different healthcare environments was highlighted, showcasing the design's ability to accommodate the expanding need for smart healthcare systems. Our findings also underscore the practical implications of this work, as the system's real-time decision-making capabilities and its capacity to process multi-modal sensor data render it amenable to widespread deployment in various healthcare settings spanning from hospitals to clinics to home care.

The system shows good results, but in future work, the generalization of the models across devices using diverse data will be studied, as well as the integration of other kinds of sensors and the optimization of the framework for more constrained environments. Thus, the Edge-Aware Deep Learning Architecture solves contemporary healthcare issues whilst paving the way for a more futuristic approach that is user-specific, offering privacy, and ultimately more effective healthcare.

REFERENCES

Atienza, D. (2024). Edge deep learning in computer vision and medical diagnostics. Artificial Intelligence Review, 57(3), 2345-2367.

Batool, I. (2025). Real-Time Health Monitoring Using 5G Networks: A Deep Learning-Based Architecture for Remote Patient Care. arXiv preprint arXiv:2501.01027. arXiv.

Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., & De Felice, F. (2020). Artificial Intelligence and Machine Learning Applications in Smart Production: Progress, Trends, and Directions. Sustainability, 12(2), 492.

Dogan, A., Constantin, J., Ruggiero, M., Burg, A., & Atienza, D. (2020). Multi-Core Architecture Design for Ultra-Low-Power Wearable Health Monitoring Systems. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39(11), 3245-3258.

Duch, L., Basu, S., Braojos, R., Ansaloni, G., & Pozzi, L. (2020). HEAL-WEAR: An Ultra-Low Power Heterogeneous System for Bio-Signal Analysis. IEEE Transactions on Circuits and Systems I: Regular Papers, 67(10), 3456-3469.

Elbir, A. M., & Coleri, S. (2020). Federated Learning for Vehicular Networks. arXiv preprint arXiv:2006.08985.

Hennebelle, A., Dieng, Q., Ismail, L., & Buyya, R. (2025). SmartEdge: Smart Healthcare End-to-End Integrated Edge and Cloud Computing System for Diabetes Prediction Enabled by Ensemble MachineLearning. arXiv preprintarXiv:2502.15762.

Iranfar, A., Zapater, M., & Atienza, D. (2020). Machine Learning-Based Quality-Aware Power and Thermal Management of Multistream HEVC Encoding on Multicore Servers. IEEE Transactions on Parallel and Distributed Systems, 31(12), 2904-2917.

Loh, J., Dudchenko, L., Viga, J., & Gemmeke, T. (2025). Towards Hardware Supported Domain Generalization in DNN-Based Edge Computing Devices for Health Monitoring. arXiv preprint arXiv:2503.09661.arXiv

Mamaghanian, H., Khaled, N., Atienza, D., & Vandergheynst, P. (2020). Compressed Sensing for RealTime Energy Efficient ECG Compression on Wireless Body Sensor Nodes. IEEE Transactions on Biomedical Engineering, 67(3), 838-848.

Pahlevan, A., Qu, X., Zapater, M., & Atienza, D. (2020). Integrating Heuristic and Machine-Learning Methods for Efficient Virtual Machine Allocation in Data Centers. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39(11), 3245-3258.

Pokhrel, S. R., & Choi, J. (2020). Federated Learning for Edge Computing: A Survey. IEEE Communications Magazine, 58(12), 50-56.

Putra, K. T., Chen, H. C., Prayitno, Ogiela, M. R., & Chou, C. L. (2021). Federated Compressed Learning Edge Computing Framework with Ensuring Data Privacy for PM2.5 Prediction in Smart City Sensing Applications. Sensors, 21(2), 456.

Rincon, F., Recas, J., Khaled, N., & Atienza, D. (2020). Development and evaluation of multi-lead wavelet-based ECG delineation algorithms for embedded wireless sensor nodes. IEEE Transactions on Information Technology in Biomedicine, 24(2), 387-398.

Scrugli, M. A., Loi, D., Raffo, L., & Meloni, P. (2021). An adaptive cognitive sensor node for ECG monitoring in the Internet of Medical Things. arXiv preprint arXiv:2106.06498.

Simon, W. A., Qureshi, Y. M., Rios, M. A., Levisse, A. S. J., & Zapater, M. (2020). BLADE: An in-Cache Computing Architecture for Edge Devices. IEEE Transactions on Computers, 69(11), 1602-1614.

Sridhar, A., Vincenzi, A., Atienza, D., & Brunschwiler, T. (2020). 3D-ICE: a Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs. IEEE Transactions on Computers, 69(1), 45-58.

Sridhar, A., Vincenzi, A., Ruggiero, M., & Atienza, D. (2020). Neural Network-Based Thermal Simulation of Integrated Circuits on GPUs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39(12), 4567-4578.

Sufian, A., You, C., & Dong, M. (2021). A Deep Transfer Learningbased Edge Computing Method for Home Health Monitoring. arXiv preprint arXiv:2105.02960.

Surrel, G., Aminifar, A., Rincon, F., Murali, S., & Atienza, D. (2020). Online Obstructive Sleep Apnea Detection on Wearable Sensors. IEEE Transactions on Biomedical Circuits and Systems, 14(2), 209-220.

Velichko, A. (2021). A Method for Medical Data Analysis Using the LogNNet for Clinical Decision Support Systems and Edge Computing in Healthcare. arXiv preprint arXiv:2108.02428.