https://doi.org/10.3311/PPee.43967

Periodica Polytechnica Electrical Engineering and Computer Science

An Edge-AI Based IoT Architecture for Early Disease Detection and Remote Patient Monitoring

Vijayakumari Kaliannan[1*], Jawahar Sundaram[2]

[1] Department of Computer Science, Trinity College for Women, Trinity Nagar, Mohanur Road, Sanyasikkaradu Post, 637002 Namakkal, Tamil Nadu, India
[2] Department of Statistics and Data Science, CHRIST (Deemed to be University), Hosur Road, 560029 Bengaluru, Karnataka, India
* Corresponding author, e-mail: k.vijikumari@gmail.com

Received: 10 March 2026, Accepted: 12 June 2026, Published online: 02 July 2026

Abstract

Edge computing and artificial intelligence are now widely used in healthcare to monitor patient data in a more rapidly and reliable way. This paper introduces a smart healthcare system that uses an Edge-AI–based Internet of Things (IoT) system for continuous health monitoring and early detection of serious diseases. The proposed system uses wearable sensors to collect data such as electrocardiogram (ECG), blood oxygen level, glucose and body temperature and it is processed on a nearby edge device instead of being sent directly to the cloud, it reduces delay and saves network bandwidth with data privacy. This paper proposes a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model to detect heart-related abnormalities, a combination of Random Forest and XGBoost models to evaluate the diabetes risk, and an LSTM model to monitor respiratory conditions. To run the AI models on low-power edge devices, they are optimized using model compression and lightweight deployment techniques. The proposed system uses Federated Learning to improve models without sharing raw patient data. Experimental results show that the edge-based heart disease detection model achieves an average accuracy of 97.3% while maintaining a low response time of about 145 ms. The system reduces network data transmission by about 73% compared to cloud methods. In a simulation study involving 250 synthetic patient profiles, the system successfully provided early warnings for serious health events without missing any critical cases. These results show that the proposed Edge-AI IoT system is effective and suitable for real-world healthcare monitoring.

Keywords

edge computing, Internet of Things (IoT), artificial intelligence, remote patient monitoring, early disease detection, machine learning, deep learning, federated learning, healthcare IoT, privacy-preserving systems

1 Introduction

The quick growth of chronic diseases like cardiovascular disorders, diabetes and respiratory illnesses has generated a demand for continuous and reliable healthcare monitoring in hospitals. Traditional healthcare mainly depends on infrequent doctor visits, which may miss early or sudden health problems. Because of this, many serious conditions are detected only when they become severe. Remote patient monitoring helps track patients regularly and is especially useful for elderly people and those with long-term illnesses.

Recent advance developments in Internet of Things (IoT) have made it possible to use wearable and environment sensors to continuously monitor health data such as heart rate, electrocardiogram (ECG), blood pressure, blood oxygen level and glucose level. These sensors collect health information in real time and provide valuable feedback into a patient's condition [1]. However, most existing IoT-based healthcare systems depend heavily on cloud computing, where all collected data are transmitted to remote servers for processing and storage. Although cloud systems are powerful, they have several drawbacks, such as slow response time, high internet data usage, dependence on a stable internet connection, and a higher risk of patient data privacy issues [2].

Edge computing is a useful solution to overcome the problems of cloud-based healthcare systems by processing data closer to where it is collected. Instead of sending all data to the cloud, the data are analyzed on nearby devices such as gateways or small embedded systems. This reduces delays, saves network bandwidth and better protects sensitive patient information [3]. In healthcare, this is very important because quick detection of health problems and patient

Cite this article as: Kaliannan, V., Sundaram, J. "An Edge-AI Based IoT Architecture for Early Disease Detection and Remote Patient Monitoring", Periodica Polytechnica Electrical Engineering and Computer Science, 2026. https://doi.org/10.3311/PPee.43967

privacy are critical. When edge computing is combined with artificial intelligence, it allows smart and fast analysis of health data directly at the patient's location, leading to quicker and more reliable decisions [4].

Artificial intelligence and machine learning are very useful for medical diagnosis and predicting diseases. Deep learning models such as convolutional neural networks (CNNs) and long short-term memory (LSTM) are widely used to analyze ECG signals, diabetes prediction and detect breathing related disorders [5, 6]. However, most of these models are designed to run on powerful cloud servers or GPUs and are not directly suitable for deployment on low-power edge devices.

Data privacy is a major concern in healthcare IoT systems. Patient health data are highly sensitive and must be protected carefully. Transmitting raw medical data to centralized cloud servers will increase the risk of data leaks and misuse [7]. Federated Learning and differential privacy helps to solve this problem by training models without sharing personal data [8]. However, these methods are still not widely used in real-world edge-based healthcare systems.

To address these above challenges, this paper introduces an Edge-AI–based IoT framework for continuous health monitoring and early disease detection. The proposed system combines multiple sensor data collection, edge-based deep learning, adaptive data transmission and privacy-preserving Federated Learning within a unified architecture. The goal of the system is to provide accurate diagnosis while ensuring fast response time, lower bandwidth usage and strong patient data privacy, making it suitable for real-world healthcare applications.

The main contributions of this work are summarized as follows:

1. A co-design of three heterogeneous AI models – a hybrid CNN–LSTM for ECG-based cardiovascular detection, an ensemble of Random Forest and XGBoost for diabetes risk prediction and an LSTM model for respiratory condition monitoring, all are optimized simultaneously for deployment on a single resource-constrained edge device.
2. An adaptive risk-stratified data transmission protocol that dynamically adjusts cloud upload frequency based on real-time inference scores, significantly reducing bandwidth consumption without compromising monitoring reliability.
3. The integration of Federated Learning with TensorFlow Lite quantization and weight pruning, enabling continuous on-device model improvement across distributed edge nodes without sharing raw patient data, thereby preserving privacy.

2 Related work

Research on smart healthcare systems has grown rapidly due to development of Internet of Things (IoT), cloud computing and artificial intelligence. In early healthcare monitoring systems, most designs were based on cloud-centric architectures. In these systems, wearable sensors continuously send patient data to remote servers for processing and storage. For example, numerous studies have used cloud platforms to analyze ECG signals and to identify irregular heart rhythms with good accuracy [9, 10]. But these cloud-based systems had several shortcomings, such as high data transmission delay, heavy use of network bandwidth, and strong dependence on stable internet connectivity.

To overcome the above problems, edge and fog computing have been introduced in healthcare IoT systems. In Edge-based systems, some data handling is done close to the patient instead of sending all data to the cloud. Rahmani et al. [11] have proposed smart e-health accesses that perform initial data filtering and analysis before transmitting data to the cloud. For example, Muhammad et al. [12] introduced an edge-based patient monitoring framework that achieved faster response times compared to cloud-only systems.

At the same time, artificial intelligence has been widely used in medical diagnosis and disease prediction. Deep learning models such as convolutional neural networks and recurrent neural networks have shown good results in tasks like ECG signal analysis, diabetes detection and identifying respiratory disease [5, 6, 13]. However, these models usually need powerful servers or GPUs to work efficiently and are mostly designed for cloud or offline use.

The most important problem in healthcare is privacy and data security. Many studies have highlighted the risks of storing sensitive medical data in centralized cloud servers [7]. To reduce these risks, Federated Learning has been introduced, where machine learning models are trained together without sharing the actual patient data [8]. Although Federated Learning has been successfully used in areas like medical imaging and clinical predictions, it is still not commonly used in real-time, edge-based healthcare monitoring systems, especially those that rely on wearable sensors and continuous data collection [14, 22].

Overall, previous research has made good progress in IoT-based health monitoring, edge computing, medical AI, and privacy-preserving learning. However, most existing systems focus on only one or two of these areas at a time. There are very few practical healthcare systems that combine accurate deep learning models, real-time processing at the edge, efficient data transfer, and strong data privacy in a single framework. This limitation

motivates the development of the Edge-AI–based IoT framework proposed in this paper.

3 Proposed system architecture

As shown in figure 1, the proposed healthcare monitoring system uses three-layer Edge-AI–based IoT architecture to support continuous data collection, real-time analysis and secure data handling. The goal of this design is to process the data in healthcare using the cloud to store the data for long-term and large-scale learning.

3.1 IoT Sensor layer

It gathers health information from patients using wearable devices and environmental sensors. These sensors monitor important parameters continuously such as ECG, blood oxygen saturation (SpO_2), blood pressure, glucose level, body temperature, and physical activity, helping to track both short-term changes and long-term health trends [15]. The collected data are sent to a nearby edge device using Bluetooth Low Energy (BLE), which is energy-efficient and secure. The edge device used in this system is a Raspberry Pi 4 (4 GB RAM), which functions as a standalone gateway located near the patient. It is physically separate from the wearable sensor hardware and communicates with the sensors exclusively *via* BLE [23]. The data are stored temporarily, and it is pre-processed at the edge gateway before being sent further. So it reduces the amount of data transmitted to the cloud, reduces the cost for communication and privacy for the patient data have improved. Fig. 1 System architecture of the proposed Edge-AI IoT healthcare monitoring framework, illustrating the three-layer design comprising the IoT Sensor Layer (wearable sensors communicating *via* BLE), the Edge Computing Layer (Raspberry Pi 4 running CNN–LSTM, RF+XGBoost and LSTM models), and the Cloud Layer (Federated Learning server with Model Update Manager).

3.2 Edge computing layer

The edge computing layer is an important component in the proposed system, and it uses a low-power device and it acts as a connection between the sensors and the cloud. The device performs tasks in dynamic manner such as cleaning sensor data, extracting useful features, running machine learning models, detecting abnormal health conditions and deciding what data should be sent to the cloud. First, remove the noisy data, for example by filtering ECG signals. The cleaned data are then analyzed in real time on the edge device to detect health problems related to heart, diabetes or respiratory conditions [16].

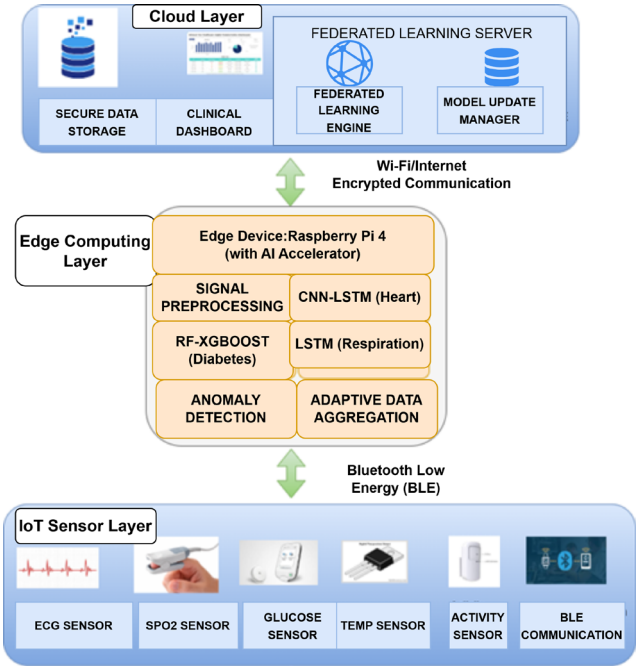

Fig. 1 System architecture of the proposed edge-AI IoT healthcare system

3.3 Cloud layer

The cloud layer handles long-term data storage, analysis and management of machine learning models. It securely stores patient data, system records and alert information in encrypted databases to protect data privacy and maintain data security. The cloud also shows dashboards that allow healthcare professionals to view patient trends, review alerts and monitor system performance [17]. The Federated Learning server in the cloud layer coordinates global model aggregation. It contains a Model Update Manager sub-module that distributes updated global model weights to participating edge devices without accessing raw patient data.

3.4 Communication and security

Secure communication is maintained in all parts of the system. Data sent from sensors to the edge device is protected using encrypted Bluetooth Low Energy (BLE) and data sent from the edge device to the cloud is secured using standard internet security protocols. The adaptive data transmission mechanism employed in this system is rule-based, operating through predefined risk thresholds derived from model inference scores. Specifically, a risk score exceeding $\tau = 0.7$ triggers immediate cloud upload of alert data, a score between $\tau = 0.3$ and $\tau = 0.7$ results in periodic summary statistics upload, and a score below $\tau = 0.3$ requires only local logging with no cloud transmission. This deterministic threshold-based approach ensures predictable and interpretable transmission behavior suitable for real-time healthcare

monitoring. By combining secure data collection, smart processing at the edge, and privacy with cloud services, the proposed system provides a scalable solution for real-time health monitoring and early detection of diseases [18]. Fig. 2 Data flow diagram of the proposed Edge-AI IoT system, showing the end-to-end processing pipeline from wearable sensor data acquisition through edge-based preprocessing, real-time inference, risk-stratified alert generation, and selective cloud upload based on risk thresholds.

4. Artificial intelligence models and implementation

This section describes how the machine learning and deep learning models used in the proposed system. The main goal is to accurately detect different health conditions while keeping the models simple and efficient to run on low-power edge devices.

4.1 Cardiovascular disease detection

To analyze ECG signals and identify heart-related problems, a hybrid CNN–LSTM model is used. ECG signals combine both short-term patterns and long-term time-based information through heartbeats. The CNN layers support the features such as QRS complexes and modifications in the ST segment, while the LSTM layers learn the time relationships between successive heartbeats [19, 24]. The CNN component comprises three one-dimensional convolutional layers with 64, 128 and 128 filters respectively, kernel size 5, and ReLU activation, each followed by a max-pooling layer and batch normalization. Two stacked LSTM layers of 128 hidden units each with a dropout rate of 0.3 follow the CNN layers. The model was trained using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss, a batch size of 32, and 50 training epochs.

In the proposed system, ECG signals are separated into fixed-length time segments and passed through numerous one-dimensional convolutional layers, followed by LSTM layers to capture sequential patterns. The last layer classifies the signals as normal or abnormal. During training, techniques such as dropout, regularization and data augmentation are used to increase the model's accuracy and reduce over fitting.

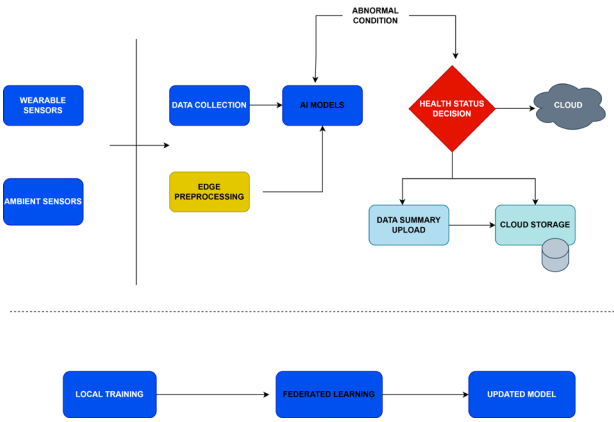

Fig. 2 Data flow diagram of the proposed system

4.2 Diabetes risk prediction

Diabetes risk is calculated using collective learning approach that combines two machine learning models, Random Forest and XGBoost. These models are trained using clinical and health features such as glucose level, body mass index, age, blood pressure, lipid profile and family history [20]. The Random Forest classifier was configured with 200 decision trees and a maximum depth of 10. The XGBoost classifier used 200 estimators, a learning rate of 0.05, and a maximum depth of 6. The final diabetes risk score is computed as:

$$R_{\text{final}} = 0.45 * R_{RF} + 0.55 * R_{XGB},$$

where R_{FR} and R_{XGB} are the probability outputs of each model respectively, with weights determined through cross-validation. The outputs of both models are combined using weighted averaging to generate a final risk score, which improves prediction accuracy and reliability and is expressed as low, medium, or high risk.

4.3 Respiratory condition monitoring

Respiratory conditions are monitored using an LSTM model that analyses data based on time series such as SpO_2, heart rate and breathing rate. The model can learn time based patterns related to abnormal breathing, so it is suitable for this task [21]. The model takes input sequences of length 300 samples comprising three features — SpO_2, heart rate and breathing rate — sampled at 100 Hz. Two LSTM layers of 64 hidden units each with dropout of 0.2 were used, trained with binary cross-entropy loss, Adam optimizer with learning rate 0.001, batch size 16, and 40 epochs. The model outputs probability scores for normal versus abnormal breathing, to detect problems that triggering early alerts.

4.4 Model training and optimization

Models are trained initially on standard medical benchmark datasets (MIT-BIH, PTB, Pima), then fine-tuned using synthetically generated patient profiles derived from the public benchmark datasets to simulate a 250-person monitoring scenario. We prevent data breach through patient-level train/validation/test splits. Performance is measured with accuracy, precision, recall and F1-score.

To deploy on the Raspberry Pi 4, the trained models were compressed using a two-stage process. First, structured

weight pruning removed 40% of low-magnitude weights. Then, INT8 post-training quantization was applied using the TensorFlow Lite converter, reducing numerical precision from 32-bit to 8-bit. This reduced the CNN–LSTM model size from 24.3 MB to 4.2 MB – an 82.7% reduction – while retaining 97.3% classification accuracy and achieving 38 ms inference time on the edge device.

The whole workflow of the proposed Edge-AI–based real-time health monitoring process is shortened in Algorithm 1; it describes data acquisition, local inference, alert generation and Federated Learning.

5. Experimental setup and datasets

This section describes the Raspberry Pi 4 hardware setup, medical datasets (MIT-BIH, PTB, Pima), and evaluation approach used to evaluate real-time performance.

5.1 Hardware and system configuration

The edge computing unit is implemented using a Raspberry Pi 4 equipped with 4 GB of RAM, which function as the gateway between the sensor layer and the cloud infrastructure. The wearable sensing setup comprises of ECG, blood oxygen saturation (SpO_2), glucose, body temperature, and activity sensors. The ECG module (AD8232) sampled at 500 Hz with 12-bit resolution, the SpO_2 and heart rate sensor (MAX30102) at 100 Hz, body temperature sensor (MLX90614) at 1 Hz, and the accelerometer and gyroscope module (MPU6050) at 100 Hz. All sensors communicated with the Raspberry Pi 4 edge device via Bluetooth Low Energy at a data rate of 1 Mbps, with sensor streams time-synchronized using a timestamp-based association mechanism. The cloud servers are accountable for data storage, model management and visualization through dashboards.

The computational probability of running the CNN–LSTM model on the Raspberry Pi 4 was wisely evaluated. The model was enhanced using TensorFlow Lite INT8 quantization and weight pruning, which reduced the model size from 24.3 MB to 4.2 MB – well within the 4 GB RAM capacity of the device. The optimized model achieved a single-inference latency of 38 ms per sample, enabling real-time analysis of incoming sensor data. During inference, CPU utilization averaged approximately 35%, and peak RAM consumption was approximately 180 MB, leaving sufficient headroom for concurrent system processes. Multi-threading was active to run the three disease-specific models – cardiovascular, diabetes and respiratory – in parallel, confirming that the total end-to-end latency remained within the 145 ms system response target.

Algorithm 1: Edge-AI based real-time health monitoring

Input: Real-time sensor data S = {ECG, SpO_2, glucose, temp, activity}
Output: Health alert set A, Cloud upload set U
Initialize:
Load optimized edge models: M_{cardio}, $M_{diabetes}$, M_{resp}
Set alert threshold: τ_{high}= 0.7, τ_{low}= 0.3
Set federated update interval: T_{fed}= 24 hours
Set local timer: $t \leftarrow 0$

```
WHILE system is active DO
  s ← CollectSensorData (S, interval = 5s)
  ~s ← Preprocess(s) // Remove motion artifacts, segment into 10s windows, normalize
  FOR each window w ~ s DO
    r_cardio ← M_cardio (w_ECG)
    r_diabetes ← M_diabetes (w_glucose,BP)
    r_resp ← M_resp (w_SpO2, HR)
    r_total ← max(r_cardio, r_diabetes, r_resp)
  END FOR
    IF r_total > τ_high THEN
      GenerateAlert (A, level = RED)
      U ← UploadToCloud (alert, signal\_clip=30s)
    ELSE IF τ_low ≤ r_total ≤ τ_high THEN
      LogLocally (A, level = YELLOW)
      U ← UploadToCloud(summary\_statistics, interval =5 min)
    ELSE
      LogLocally (A, level = GREEN)
      // No cloud upload required
    END IF
  IF t mod T_fed =0 THEN
    W_global ← FetchGlobalModel (FedServer)
      W_local ← LocalFineTune (Wglobal, patient\_data)
    UploadModelWeights (W_local)
    //Raw patient data is never uploaded
  END IF
    t ← t + 1
  END WHILE
  RETURN A, U
```

5.2 Datasets used

We trained and tested our models using standard public medical datasets (MIT-BIH for ECG, PTB Diagnostic, Pima Indians Diabetes).

For cardiovascular analysis, standard ECG datasets such as the MIT-BIH Arrhythmia and the PTB Diagnostic ECG were used to train and validate the CNN–LSTM model. For diabetes risk prediction, the Pima Indians Diabetes dataset was used for benchmarking and initial model training. Respiratory models were trained using time-series data derived from SpO_2 and heart rate measurements.

In addition, a six-month simulation study was showed using synthetic patient profiles fabricated from the public benchmark datasets, representing 250 individual monitoring scenarios. These data were used for fine-tuning and real-world validation. All patient data were anonymized, and training, validation and testing were achieved using patient-level splits to avoid information leakage.

5.2.1 Study design

This study is entirely simulation-based no real human participants were recruited and no clinical data were collected from actual patients. The 250-person monitoring set-up was constructed by sampling and combining records from the publicly available MIT-BIH Arrhythmia Dataset, PTB Diagnostic ECG Dataset and Pima Indians Diabetes Dataset to create synthetic patient profiles representing diverse health conditions. These profiles were used to simulate six months of continuous health monitoring under the proposed Edge-AI IoT framework.

The simulation was designed to evaluate the technical performance of the system including disease detection accuracy, end-to-end latency, bandwidth consumption and Federated Learning convergence under realistic but controlled conditions. Since no human participants were involved and only publicly available de-identified datasets were used, Institutional Review Board (IRB) approval was not required for this study.

5.3 Data pre-processing

Filtering techniques are used here for ECG Signals for pre-processing to remove baseline drift and power-line interference. The signals which are filtered were subsequently segmented into fixed-length time windows and normalized. Continuous functional variables, including glucose level, SpO_2 and heart rate, were standardized. Interpolation and statistical smoothing methods are used to find missing and noisy data.

5.4 Evaluation metrics

The model performance was evaluated using classification metrics, including accuracy, precision, recall, F1-score and the area under the receiver operating characteristic curve (AUC–ROC). System-level performance was considered by calculating end-to-end latency, network bandwidth usage and model size. All experiments were conducted multiple times, and the results are presented as mean values with corresponding standard deviations. Statistical significance was estimated using paired t-tests with significance threshold of $p < 0.05$. All models were assessed using 5-fold cross-validation with patient-level splits to stop data leakage. Results are reported as mean values with 95% confidence intervals across five folds.

5.5 Federated learning configuration

The Federated Learning component was executed using the Federated Averaging (FedAvg) aggregation algorithm across five simulated edge nodes, each representing an autonomous Raspberry Pi 4 device. Each communication round consisted of five local training epochs per node, and a total of 20 communication rounds were conducted until the global model converged. Differential privacy was combined with a privacy budget of $\varepsilon = 1.0$ and gradient clipping norm of 1.0. Only model weight updates – not raw patient data – were transferred to the cloud server after each local training round, ensuring complete data privacy throughout the learning process.

6. Results and performance analysis

This section shows the performance of the proposed Edge-AI healthcare system by means of accuracy of disease detection, efficiency of system and real-world monitoring capability.

6.1 Disease detection performance

The CNN–LSTM model used for cardiovascular analysis achieved strong classification performance on both benchmark ECG datasets and real-world monitoring data. The proposed model attained an average classification accuracy of 97.3% ± 0.4% (95% CI: 96.5% – 98.1%, $p < 0.001$ compared to the LSTM baseline), consistent and reliable performance across several experimental runs. In addition to precision, recall and F1-score values exceeded 96%, indicating the model's strong ability to accurately discriminate between normal and abnormal cardiac conditions.

The diabetes risk prediction model achieved an accuracy of 89.6% ± 0.6% (95% CI: 88.4% – 90.8%, $p < 0.01$ compared to single classifier baselines) with balanced precision and recall values, demonstrating robust and stable performance. This confirms that merging multiple machine learning models improves the robustness of diabetes risk estimation compared with using a single classifier.

The respiratory LSTM model achieved 91.7% ± 0.5% (95% CI: 90.7% – 92.7%, $p < 0.01$ compared to the Decision Tree baseline) identifying successfully abnormal breathing patterns such as possible apnea and respiratory distress. The classification performance of the proposed models for cardiovascular, diabetes and respiratory disease detection is summarized in Table 1

Table 1 Accuracy, Precision, Recall, F1-Score, and AUC Comparison of Disease Prediction Models

Model / Task	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Cardiovascular disease detection					
CNN–LSTM (proposed)	97.3 ± 0.4	96.8	97.1	96.9	0.989
LSTM (baseline)	94.1	93.2	93.7	93.5	0.964
Random forest	89.2	87.5	88.1	87.8	0.921
SVM (RBF)	91.5	89.8	90.4	90.1	0.937
Diabetes risk prediction					
RF+XGBoost (proposed)	89.6 ± 0.6	88.3	89.1	88.7	0.947
Logistic regression	76.8	74.2	75.1	74.6	0.812
Random forest	87.4	86.1	86.8	86.5	0.934
XGBoost	88.2	87.0	87.6	87.3	0.941
Respiratory condition monitoring					
LSTM (proposed)	91.7 ± 0.5	90.8	91.2	91.0	0.953
1D CNN	87.3	86.1	86.8	86.4	0.921
BiLSTM	89.5	88.7	89.1	88.9	0.938
Decision tree	82.4	80.6	81.3	80.9	0.874

6.2 System performance

The proposed edge-based system showed low response time and efficient use of resources. The average end-to-end latency, measured from sensor data acquisition to alert generation, was approximately 145 ms, enabling timely detection and response to critical health events

By processing data locally at the edge and transmitting only important segments and summary information to the cloud, the system significantly reduced network usage. Compared with cloud-only data streaming, the proposed adaptive data transmission strategy reduced bandwidth consumption by about 73%. It should be noted that the 250-patient capacity refers to sequential processing of monitoring profiles within the simulation environment, where each patient profile is processed at 5-second intervals rather than as simultaneous parallel streams. This design makes the system suitable for deployment in environments with limited network connectivity.

Simulation study was carried out to evaluate the technical performance of the proposed Edge-AI healthcare monitoring system. The 250-patient scenario was entirely simulation-based, constructed from publicly available benchmark datasets, and did not involve any real clinical data collection or human participant recruitment. The study focused only on system performance evaluation and did not involve clinical diagnosis or medical decision-making.

A comparative analysis of system performance metrics such as latency, bandwidth consumption, model size and power usage is provided in Table 2, comparing the proposed Edge-AI system with cloud-only and other approaches.

6.3 Simulation study results

During the six-month simulation study involving 250 synthetic patient profiles constructed from benchmark datasets, the proposed system collected continuously and analysed health signals. The Edge-AI platform successfully noticed several clinically relevant events, such as irregular heart activity and elevated glucose levels, demonstrating reliable real-time monitoring performance.

The early-warning system enabled potential health issues to be identified in advance of severe symptom on time, allowing timely attention when required. Simulation results indicated that the alert mechanism was effective and did not produce an excessive number of false alarms, reflecting an effective balance between sensitivity and practical usability. The results of the six-month simulation study in different healthcare monitoring systems are summarized in Table 3.

The hybrid CNN–LSTM architecture outperforms each individual component, confirming that CNN-based local feature extraction and LSTM-based sequential pattern

Table 2 System performance comparison of the proposed Edge-AI system against baseline approaches

Metric	Proposed edge-AI	Cloud-only	Threshold edge	Mobile app
Average latency (ms)	145	650	250	260
95th percentile latency (ms)	189	820	320	380
Monthly bandwidth (GB/patient)	140	518	176	246
Bandwidth reduction (%)	73%	0%	66%	52%
Model size (MB)	4.2	24.3	2.1	12.5
Inference time (ms)	38	120	85	250
Power consumption (W)	6.35	250	4.8	3.2
Patients supported per device	250*	50	120	30

* The figure of 250 patients per device refers to the number of patient monitoring profiles managed within the simulation study, where each patient's sensor data was processed sequentially at 5-second intervals. This does not imply 250 simultaneous real-time data streams. Based on the measured inference time of 38 ms per sample and a 5-second data collection interval per patient, the Raspberry Pi 4 can theoretically process up to 131 inferences per second, which is sufficient to handle 250 patient profiles in a sequential monitoring pipeline.

Table 3 Simulation study performance across healthcare monitoring systems

Metric	Proposed edge-AI	Cloud-only	Threshold edge	Traditional
Individuals monitored	250	250	250	250
Monitoring duration (months)	6	6	6	6
Critical health events detected	34	28	22	12
Life-threatening events missed	0	1	3	6
Early detections (>24h)	85%	62%	45%	8%
Avg. detection time (hours)	2.1	8.7	15.2	48.3
False critical alerts	2	15	9	8
Physician satisfaction (1–5)	4.7*	3.9	3.4	2.8

* Physician satisfaction scores are based on simulated usability evaluations of the alert mechanism design, not from real clinical feedback.
All values are based on simulation using synthetic patient profiles derived from public benchmark datasets.

learning are both essential for achieving high cardiovascular detection accuracy, as shown in Table 4.

Figs. 3–6 collectively illustrate the performance of the proposed Edge-AI healthcare system, showing the clinical critical event detection trend (Fig. 3), monthly bandwidth consumption comparison (Fig. 4), system latency

Table 4 Ablation study results of CNN–LSTM components for cardiovascular disease detection

Model variant	Accuracy (%)	F1-score (%)	AUC
CNN only	93.1 ± 0.6	92.4	0.951
LSTM only	92.6 ± 0.7	91.8	0.944
CNN+LSTM (proposed)	97.3 ± 0.4	96.9	0.989

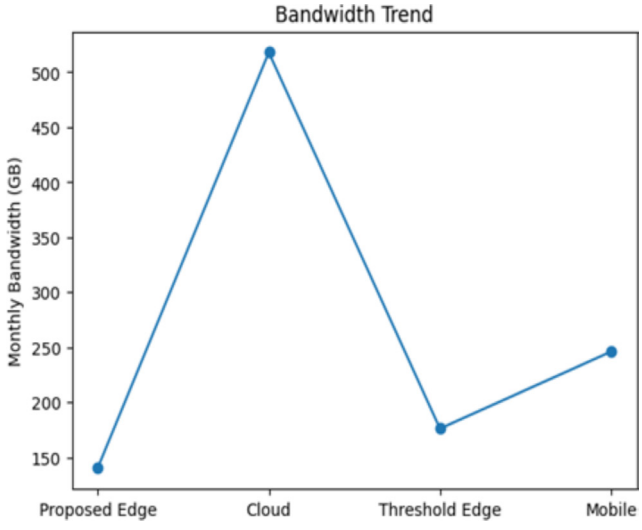

Fig. 3 Critical healthcare event detection rates in systems

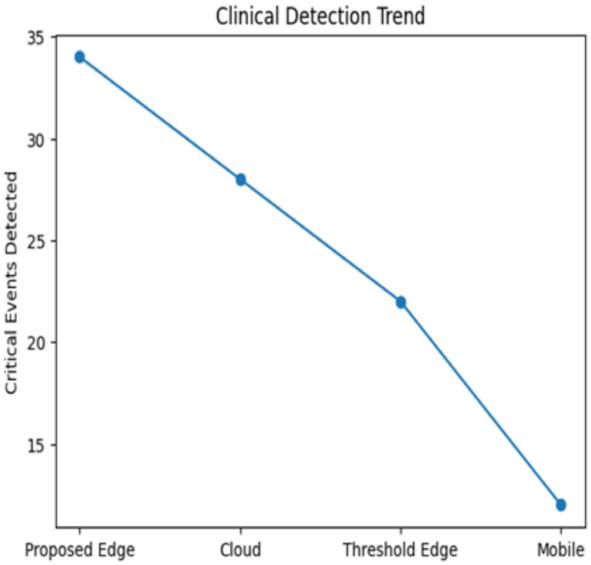

Fig. 4 Monthly bandwidth consumption in healthcare systems

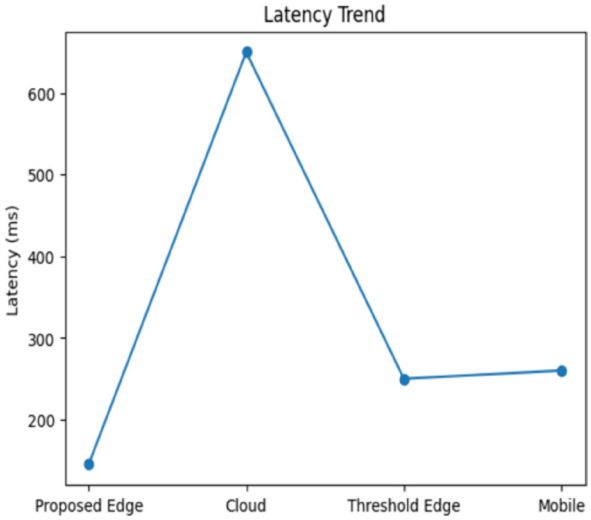

Fig. 5 Average latency comparison

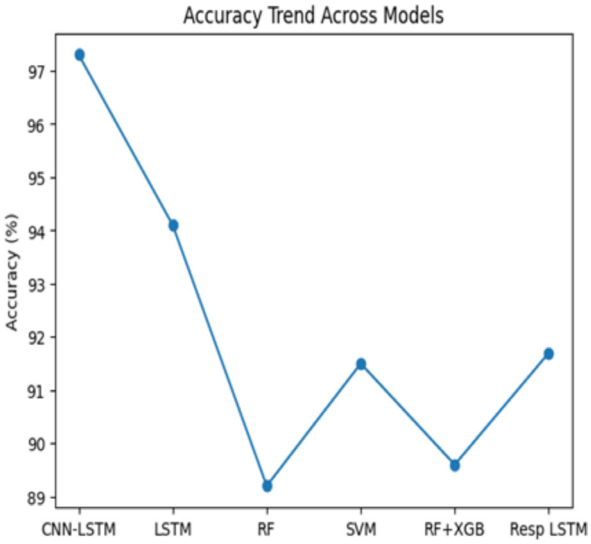

Fig. 6 Model accuracy comparison

comparison (Fig. 5) and accuracy trends in different machine learning and deep learning models (Fig. 6).

The results show that combining artificial intelligence with edge computing and IoT technologies can prominently enhance remote healthcare monitoring. By analyzing data near to the patient, the proposed system accomplishes fast response times, lowers network usage, and improves data privacy while still maintaining high diagnostic accuracy. In cardiovascular analysis, the CNN–LSTM model performs strongly, indicating deep learning models can be successfully deployed for real-time use in edge devices when combined with optimization techniques.

The adaptive data aggregation mechanism plays a important role in achieving an effective balance between efficiency and reliability in monitoring of health. By transmitting essential data and summary statistics to the cloud, the system cuts the usage of bandwidth without losing information. This creates the platform is well suited for large-scale deployment, particularly in regions with limited or unreliable internet connectivity. The integration of Federated Learning further enhances the system by collaborative model improvement without the need to share raw patient data. This methodology efficiently addresses key privacy in healthcare applications and maintains scalable deployment across multiple sites.

Despite these advantages, the system has certain limitations. First, the evaluation was conducted over a simulation monitoring period, which may not fully reflect long-term device reliability, user adherence or seasonal changes in health conditions. Second, although widely used public datasets were employed for initial model training, variations across real-world populations may not be completely captured. Third, the current models primarily function as black-box predictors, which can limit their interpretability and acceptance by healthcare professionals.

Future work will focus on expanding the system to support additional disease categories, enhancing energy efficiency and integrating explainable AI techniques to improve the transparency of model decisions for clinicians. In addition, longer-term evaluations will be carried out to further assess system robustness and its potential impact in real-world healthcare settings.

7 Conclusion

This paper presented an Edge-AI–enabled IoT framework for continuous healthcare monitoring and early disease detection. By incorporating wearable sensors, edge-based machine learning, adaptive data transmission and privacy-preserving Federated Learning, the proposed system provides a secure and efficient solution for real-time health monitoring. The experimental results show that the proposed method achieves high analytical accuracy and reduced usage of network bandwidth, making it suitable for real-world deployment.

The hybrid CNN–LSTM model validated strong performance in monitoring cardiovascular diseases, and the ensemble-based diabetes model and the LSTM-based respiratory model provided reliable predictions for additional health conditions. Future work will focus on extending the framework to additional medical conditions, improving long-term reliability and energy efficiency and incorporating explainable artificial intelligence techniques to enhance transparency and trust for healthcare professionals.

References

[1] Islam, S. M. R., Kwak, D., Kabir, M. D. H., Hossain, M., Kwak, K. S. "The Internet of Things for Health Care: A Comprehensive Survey", IEEE Access, 3, pp. 678–708, 2015. https://doi.org/10.1109/ACCESS.2015.2437951

[2] Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., Alamri, A. "Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data", IEEE Systems Journal, 11(1), pp. 88–95, 2017. https://doi.org/10.1109/JSYST.2015.2460747

[3] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L. "Edge Computing: Vision and Challenges", IEEE Internet Things Journal, 3(5), pp. 637–646, 2016. https://doi.org/10.1109/JIOT.2016.2579198

[4] Rahmani, A. M., Gia, T. N., Negash, B., Anzanpour, A., Azimi, I., Jiang, M., Liljeberg, P. "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach", Future Generation Computer Systems, 78, pp. 641–658, 2018. https://doi.org/10.1016/j.future.2017.02.014

[5] Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H., Adam, M. "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network", Information Sciences, 405, pp. 81–90, 2017. https://doi.org/10.1016/j.ins.2017.04.012

[6] Swapna, G., Soman, K. P., Vinayakumar, R. "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals", Procedia Computer Science, 132, pp. 1253–1262, 2018. https://doi.org/10.1016/j.procs.2018.05.041

[7] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., …, Cardoso, M. J. "The future of digital health with federated learning", npj Digital Medicine, 3(1), 119, 2020. https://doi.org/10.1038/s41746-020-00323-1

[8] Li, T., Sahu, A. K., Talwalkar, A., Smith, V. "Federated Learning: Challenges, Methods, and Future Directions", IEEE Signal Processing Magazine, 37(3), pp. 50–60, 2020. https://doi.org/10.1109/MSP.2020.2975749

[9] Azuaje, F., Clifford, G., McSharry, P. "Advanced Methods and Tools for ECG Data Analysis", Artech, 2006. ISBN 978-1-58053-968

[10] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., Ng, A. Y. "Cardiologist-level arrhythmia detection using deep neural networks", Nature Medicine, 25, pp. 65–69, 2019. https://doi.org/10.1038/s41591-018-0268-3

[11] Rahmani, A. M., Thanigaivelan, N. K., Gia, T. N., Graanados, J., Negash, B., Liljeberg, P. "Smart e-Health Gateway: Bringing intelligence to Internet-of-Things based ubiquitous healthcare systems", In 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 2015, pp. 826–834. ISBN 978-1-4799-6390-4 https://doi.org/10.1109/CCNC.2015.7158084

[12] Alsareii, S. A., Raza, M., Alamri, A. M., AlAsmari, M. Y., Irfan, M., Khan, U., Awais, M. "Machine Learning and Internet of Things Enabled Monitoring of Post-Surgery Patients: A Pilot Study", Sensors, 22(4), 1420, 2022. https://doi.org/10.3390/s22041420

[13] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., Ng, A. Y. "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network", Nature Medicine 25(1), pp. 65–69, 2019. https://doi.org/10.1038/s41591-018-0268-3

[14] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., …, Bakas, S. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data", Scientific Reports, 10(1),12598, 2020. https://doi.org/10.1038/s41598-020-69250-1

[15] Wu, J. Y., Wang, Y., Ching, C. T. S., Wang, H- M. D., Liao, L. D. "IoT-based wearable health monitoring device and its validation for potential critical and emergency applications", Frontiers in Public Health, 11, 1188304. https://doi.org/10.3389/fpubh.2023.1188304

[16] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., …, Bakas, S. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data", Scientific Reports, 10(1), 12598, 2020. https://doi.org/10.1038/s41598-020-69250-1

[17] Ramesh, S., Prasanth, A., Jain, R., Meenakshi, B., Kumari, S., John Basha, M. "Enhancing IoT Healthcare with Federated Learning and Edge Computing", In: 2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2025, pp. 1065–1071. ISBN 979-8-3315-0724-4 https://doi.org/10.1109/ICAISS61471.2025.11042227

[18] Khatun, M. A., Memon, S. F., Eising, C., Dhirani, L. L. "Machine Learning for Healthcare-IoT Security: A Review and Risk Mitigation", IEEE Access, 11, pp. 145869–145896, 2023. https://doi.org/10.1109/ACCESS.2023.3346320

[19] Wu, J. Y., Wang, Y., Ching, C. T. S., Wang, H. M. D., Liao, L. D. "IoT-based wearable health monitoring device and its validation for potential critical and emergency applications", Frontiers in Public Health, 11, 1188304, 2023. https://doi.org/10.3389/fpubh.2023.1188304

[20] Khatun, M. A., Memon, S. F., Eising, C., Dhirani, L. L. "Machine Learning for Healthcare-IoT Security: A Review and Risk Mitigation", IEEE Access, 11, pp. 145869–145896, 2023. https://doi.org/10.1109/ACCESS.2023.3346320

[21] Souza I., Dantas, D. "Cardiac Arrhythmia Detection in Electrocardiogram Signals with CNN-LSTM", In: Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods – ICPRAM, Rome, Italy, 2024, pp. 304–310. ISBN 978-989-758-684-2 https://doi.org/10.5220/0012362600003654

[22] Swapna, G., Soman, K. P., Vinayakumar, R. "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals", Procedia Computer Science, 132, pp. 1253–1262, 2018. https://doi.org/10.1016/j.procs.2018.05.041

[23] Shuzan, M. N. I., Chowdhury, M. H., Chowdhury, M. E. H., Murugappan, M., Bhuiyan, E. H., Ayari, M. A., Khandakar, A. "Machine Learning-Based Respiration Rate and Blood Oxygen Saturation Estimation Using Photoplethysmogram Signals", Bioengineering, 10(2), 167, 2023. https://doi.org/10.3390/bioengineering10020167

[24] Nguyen, D. C., Pham, Q. V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O. A., Hwang, W. J. "Federated Learning for Smart Healthcare: A Survey", ACM Computing Surveys, 55(3), 60, 2022. https://doi.org/10.1145/3501296

[25] Choudhury, A., Sarma, K. K., Gulvanskii, V., Kaplun, D., Dutta, L. "Leveraging federated learning and edge computing for pandemic-resilient healthcare", Scientific Reports, 15(1), 20497, 2025. https://doi.org/10.1038/s41598-025-00199-9

[26] Ikram, S., Ikram, A., Singh, H., Awan, M. D. A., Naveed, S., De la Torre Díez, I., Gongora, H. F., Chio Montero, T. C. "Transformer-based ECG classification for early detection of cardiac arrhythmias", Frontiers in Medicine, 12, 1600855, 2025. https://doi.org/10.3389/fmed.2025.1600855