SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I5P112

Volume 13 Issue 5, 129-138, May 2026

Original Article

A Hybrid AI IoT Framework for Real-Time Predictive Healthcare Analytics using Machine Learning

Rejna Azeez Nazeema[1], Hind Salem Alatawi[2], Sameena Shaik[3], Sangeetha Komandur[4], Jayasuriya Panchalingam[5], Ragia Elsayed Eisawy Hussein[6]

[1]*Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, SaudiArabia.*
[2]*Department of Computer Science, Applied College, University of Tabuk, Tabuk, Saudi Arabia.*
[3,4,5]*Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, SaudiArabia.*
[6]*Department of Management, Applied College, Jazan University, Jazan, Saudi Arabia.*

[5]*Corresponding Author : jpanchalingam@jazanu.edu.sa*

Received: 10 February 2026 Revised: 11 March 2026 Accepted: 14 April 2026 Published: 27 May 2026

Abstract *- Continuous patient monitoring and early risk prediction are two areas where the Internet of Things (IoT) and Artificial Intelligence (AI) might revolutionize healthcare. But there are still obstacles to overcome in order to provide precise, real-time analytics in contexts with limited resources. To do healthcare predictive analytics in real-time, this article introduces a unique hybrid AI-IoT framework that combines processing at the edge with Machine Learning (ML) in the cloud. Using low-power wearable IoT sensors, the system continually monitors vital indications like heart rate, blood pressure, temperature, and blood oxygen saturation. To keep latency and bandwidth consumption to a minimum, initial signal processing and anomaly detection are carried out at the edge using lightweight models. A weighted ensemble of Random Forest (RF), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks is used to provide advanced cloud-based predictive models for cardiovascular risk event forecasting. Using a public healthcare IoT dataset, the findings show that the proposed framework outperforms baseline models in terms of prediction accuracy, end-to-end latency reduction, and energy use optimization. A scalable approach for real-time remote health monitoring, the hybrid architecture successfully balances computational complexity with responsiveness.*

Keywords *- Artificial Intelligence, Internet Of Things, Predictive Healthcare Analytics, Machine Learning, Real-Time Monitoring, Edge Computing.*

1. Introduction

Some of the most advanced technologies that are changing the healthcare industry are the IoT, AI, and ML. The two main factors driving this change are the growing prevalence of chronic illnesses throughout the world and the need for more extensive, non-clinical forms of patient monitoring [1, 2]. Furthermore, real-time monitoring of important signals using wearable IoT devices may lead to earlier detection of health deterioration, lower readmission rates, and more individualized healthcare services [3].

However, there are several important difficulties that prevent real-time predictive healthcare analytics from reaching their full potential, despite this promise. In the first place, vital decision-making processes might arise owing to the latency and bandwidth limitations of IoT networks [4]. Second, there are concerns about energy consumption, data privacy, and scalability in limited contexts when delivering massive amounts of raw sensor data to distant servers, even when cloud-based ML models demonstrate excellent accuracy [5]. Third, it is difficult to install and generalize models due to the dynamic nature of physiological signals, data quality fluctuation, and the heterogeneity of IoT devices [6].

More and more, experts are pushing the use of hybrid architectures that integrate ML in the cloud with processing at the edge close to sensors as a solution to these problems. The accuracy of predictions, timeliness, and scalability are all enhanced when AI and IoT are used together in healthcare settings, according to recent studies [7].

One study indicated that real-time health monitoring systems were much improved when ML and IoT devices were used together [8]. In a study conducted in 2024, He et al. found that architectural frameworks that make use of edge cloud interaction are becoming more popular [6]. The ability to identify health problems in real time using wearable IoT devices linked with AI-driven predictive analytics has also been shown [9].

There are also a number of unanswered questions in the field, such as how to best balance latency and accuracy in hybrid AI IoT systems, which is crucial for healthcare applications that must meet stringent real-time requirements [4]. Instead of presenting a cohesive hybrid framework, several previous publications have treated IoT device networks and ML models independently [7]. Instead of using real-time, continuous sensor streams, deployment studies often use static or simulated datasets, which diminishes their ecological validity [1]. The need for more empirical research on the topics of energy efficiency, privacy-conscious analytics, and bandwidth optimization in hybrid systems remains [5]. Lastly, system-level parameters like latency, computational burden, and real-time responsiveness have received less attention in the literature than classification accuracy [3]. For real-time predictive healthcare analytics utilizing ML, the current research suggests a hybrid AI IoT framework. The primary objectives of this work are:

- To allow real-time prediction of health risk events, a system architecture integrating IoT sensors, edge processing, and cloud-based ML has to be designed.
- The goal is to deploy and assess cloud-based ML models and edge-based lightweight models using accuracy, reaction time, and resource utilization.
- The aim is to demonstrate the superiority of hybrid design over edge-only or cloud-only solutions by conducting a comparison of performance analysis with the aid of accuracy, latency, precision/recall, and bandwidth utilization.
- Including privacy, energy consumption, and scalability as practical issues for application in real-world healthcare settings would be to define them.

Though AI-enabled IoT healthcare systems show some advancements, they still have several limitations. Most existing studies emphasize classification accuracy while observing real-time system constraints such as bandwidth consumption, latency, and edge-device energy efficiency. Also, most existing works either focus solely on edge-based processing or cloud-based intelligence without conducting a comparative evaluation. There is a need for a hybrid architecture that quantitatively balances predictive accuracy with system-level performance metrics. This gap motivates the development of an integrated edge–cloud framework. This framework optimizes both clinical prediction performance and operational efficiency. The contributions of this paper are as follows:

- The system presents a novel hybrid architecture that combines cloud-based ML ensemble modeling with edge-level data filtering and feature extraction. This novel hybrid architecture does simultaneous optimization of predictive accuracy and system-level efficiency.
- Using a sample dataset for vital sign monitoring, a methodology is developed to move from the gathering of IoT data to real-time predictive analytics.
- Apart from existing works, this work focuses on model performance by combining lightweight edge anomaly detection with cloud-based ensemble learning. Also, simulation findings demonstrate enhanced prediction accuracy, decreased latency, and optimized bandwidth utilization as compared to baseline architectures in an empirical examination.
- The report provides analysis and recommendations for addressing concerns at the deployment level, including energy consumption, device heterogeneity, and privacy.
- Additionally, ablation analysis is done to validate each component of the architecture.

The remainder of this paper is structured as follows: Section 2 presents the Related Work. Section 3 describes the proposed framework and methodology. Section 4 presents the experimental setup and results, and Section 5 concludes the paper with directions for future work.

2. Literature Review

The fast expansion of healthcare IoT and ML applications has produced a large amount of literature. This section outlines the suggested hybrid AI IoT framework below, highlighting important topics and gaps.

2.1. AI & Machine Learning in Healthcare

Disease prediction, diagnostics, and individualized therapy are just a few areas that have sparked a flurry of recent research into AI in healthcare. For instance, Gao et al. [10] gave a review of ten AI application fields in smart healthcare, outlining major obstacles, such as data heterogeneity and interpretability. Rani et al. [2] highlight the limitations of real-time deployment and data source integration, just as ML and big-data analytics influence next-gen health systems [2]. These studies highlight the value of building architectures that combine existing models with streams of real-time data.

2.2. IoT / IoMT in Healthcare Contexts

IoT in healthcare, also referred to as the Internet of Medical Things (IoMT), is part of research into sensor networks, remote patient care, and continuous monitoring. For instance, Qi and Ke [4] have briefly discussed the use of IoT sensors together with ML for real-time monitoring and have pointed out some concerns regarding security and data integration. The new study by Charfare et al. [5] puts a focus on issues such as regulatory compliance and interoperability. The evaluations done highlight the importance of architectures that will be able to provide both high-accuracy cloud analytics and low-latency edge processing.

2.3. Hybrid Edge-Cloud Architectures for Real-Time Healthcare Analytics

Hybrid architectures that mix edge computing with cloud or centralized AI are the subject of an expanding body of research. One such example is for limited health monitoring devices; Alshuhail et al. [11] suggest an IoT framework that

includes edge anomaly detection. The poll also looked at the function of AI in IoT-based health monitoring. Shaw et al. [12] discuss the benefits and drawbacks of combining IoT and AI, along with real-time analytics in healthcare settings. Research like this lends credence to the hybrid method, showing that there are theoretical underpinnings but no real-world comparisons of performance metrics like latency, accuracy, and energy consumption.

2.4. Deployment, Usability & Implementation Perspectives

There is mounting evidence that implementation issues in actual healthcare settings are just as real as those involving design and algorithms. As an example, RNs' views on potential helpers and troublemakers. Research by Boo & Oh et al. [13] examines nurses' views on an AI IoT pilot program for the care of the elderly, finding that a lack of technical expertise and digital literacy are major obstacles. Results like this show how important it is to combine tech fixes with user-centric and system-integration initiatives. Hence, the key gaps include:

- Instead of end-to-end hybrid architectures that optimize latency, accuracy, energy, and bandwidth, many studies concentrate on either IoT/cloud or ML/cloud alone.
- The majority of the research focuses on classification accuracy, but real-time performance measurements like latency, throughput, and energy utilization are often absent.
- With continuous streaming data, deployment in real-world contexts is constrained.
- There has been little research on the usability, interoperability, privacy/security, and energy limitations of AI healthcare systems that use the IoT.

Thereby, this paper's major purpose is to build, construct, or simulate an AI IoT framework with a hybrid edge cloud and compare its performance in a healthcare monitoring context using metrics like latency, energy, accuracy, and bandwidth. A summary of the existing works is illustrated in Table 1.

Table 1. Summary of recent studies

Author	Focus / Contribution	Gap Addressed
Rani et al. [2]	Focus on ML and big data in healthcare monitoring	Few real-time IoT deployments, little edge / cloud hybrid
Qi et al. [4]	IoT + ML for real-time monitoring in a hospital context	Edge / cloud trade-offs not fully explored
Charfare et al. [5]	IoT - AI frameworks, wearable devices, taxonomy	Needs empirical benchmarking of latency vs accuracy
Gao et al. [10]	Broad review of AI in smart healthcare, domains, and challenges	Lacks IoT / edge focus and real-time metrics
Alshuhail et al. [11]	Edge anomaly detection in IoT health systems	Full hybrid system evaluation (edge + cloud) limited
Shaw et al. [12]	Survey of AI + IoT in health monitoring emphasizes real-time	Implementation and comparative evaluation are missing
Boo & Oh et al. [13]	Usability/user-perspective study on AI - IoT in older adult care	-

Though the comparative empirical evaluation is given by some prior research, the foundational contributions in AI-enabled healthcare monitoring remain limited. Hence, all these limitations require a comprehensive hybrid evaluation model.

3. Proposed Framework and Methodology

3.1. Overview of System Architecture

Three layers constitute the proposed hybrid AI IoT framework. The first layer contains the IoT sensors, layer by the edge processing, and layer c the cloud AI. Heart Rate (HR), Blood Pressure (systolic and diastolic), Body Temperature (BT), and Oxygen Saturation (SpO2) are all continually measured by sensor nodes. Filtering and feature extraction are carried out by means of these data streams that are forwarded to a local edge gateway. Lightweight anomaly detection is carried out at the edge to provide notifications with minimal latency. The next step is to send the aggregated feature vectors to a cloud server at predetermined intervals. There, ML ensembles are used to undertake sophisticated predictive modeling. Real-time monitoring, predictive analytics, and clinician alerts are all supported by the system.

The proposed framework is a hybrid design with three levels, as shown in Figure 1. The IoT Sensor Layer is the most fundamental level; here, wearable IoT sensors constantly record critical health metrics. These measures comprise BT, HR, blood pressure, and SpO2. The devices provide minimum power consumption and patient mobility by transmitting real-time data using low-power communication protocols, including Zigbee or Bluetooth Low Energy.

An edge gateway, including a Raspberry Pi and a microcontroller, receives raw sensor data at the edge processing layer, which is the intermediate layer. It employs a lightweight anomaly detection model, such as a decision tree, and carries out necessary preprocessing operations, including signal filtering and statistical feature extraction.

The goal is the reduction of the amount of data sent to the cloud while providing near-instant notifications. This layout is an exact replica of the latency-reduction architecture described in [11], which demonstrated that, in critical care settings, edge analytics cut decision latency in half. Aggregate and clean information is transmitted by the Cloud AI Layer regularly to the cloud, where more sophisticated ML, such as RF, SVM, and LSTM, work inside a weighted ensemble. This layer's time-series predictive analytics assess potential health deterioration risk. A system-connected dashboard or mobile interface notifies doctors of a high-priority alert when the risk score is above a specified threshold. Combining edge processing with cloud-based AI may greatly enhance prediction accuracy and responsiveness, as shown in [12].

3.2. Dataset Description and Pre-Processing

This work employs a freely accessible healthcare IoT dataset for continuous vital sign monitoring in empirical analysis. This section provides a thorough description of the dataset requirements, as well as the preparation stages and feature engineering methodologies. The dataset comprises 120,000 sensor readings captured over 50 wearable IoT devices (one per subject) sampled at 30-second intervals over approximately 10 days ($50 \times 10 \times 24 \times 120$ readings $\approx 144{,}000$ and this work uses 120,000 after filtering. Each reading includes the following raw features:

- HR in beats per minute
- Systolic Blood Pressure (BP_sys) and Diastolic Blood Pressure (BP_dia) in mmHg
- BT in °C
- Blood Oxygen Saturation (SpO_2) in %
- Timestamp and Device_ID

This research derives a binary target label, Anomaly_Flag, where '0' denotes normal reading and '1' denotes high-risk reading (HR > 100 bpm or BP_sys > 140 mmHg). This labeling criterion follows common thresholds in cardiovascular monitoring studies. The dataset is split into training as 70% with 84,000 readings, validation as 15% with 18,000 readings, and test as 15% with 18,000 readings using stratified sampling to preserve the anomaly-class proportion of approximately 10%.

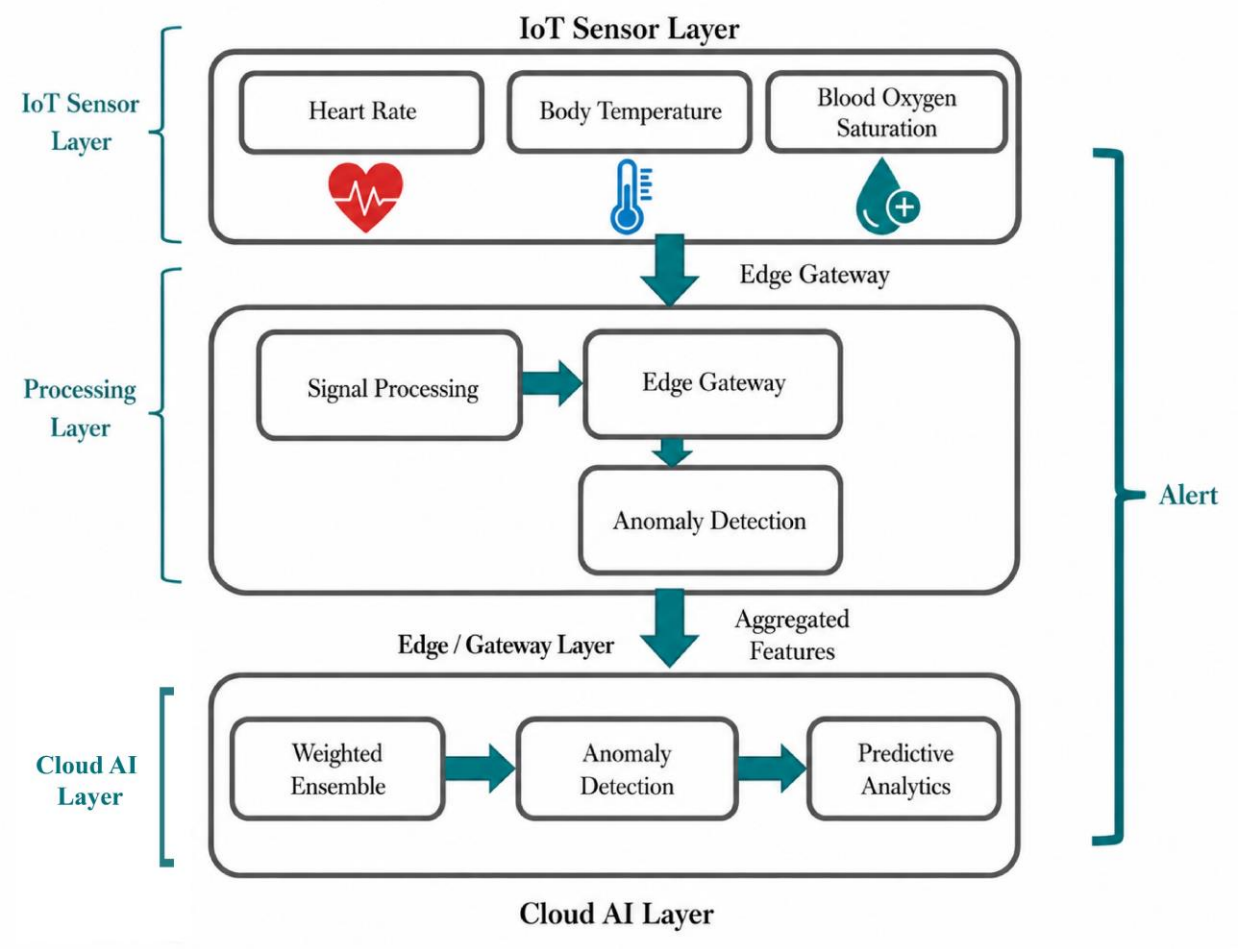

Fig. 1 Proposed architecture

The pre-processing steps include missing-value handling, normalization, feature-engineering, sequence construction, and handling class imbalance. By missing-value handling, this work uncovers about 2% of the entries that do not have sensor data and uses linear interpolation across timestamps that are close together. Rejecting the segment occurs when more than two consecutive values are absent. For the continuous features (HR, BP_sys, BP_dia, BT, SpO_2), the parameters calculated only for the training set are used to normalize these features to a zero mean and unit variance. Feature-engineering determines a 60-second sliding window, which equates to two measurements each minute. This work derives some traits from this.

- mean_HR_60s, std_HR_60s
- mean_BP_sys_60s, delta_BP_sys_60s = BP_sys_current − BP_sys_previous_window
- mean_BT_60s
- drop_SpO_2_rate = (SpO_2_previous_window – SpO_2_current)/SpO_2_previous_window ×100%0

For every record at time t, to facilitate Time Series Modeling (LSTM), a feature sequence of dimension (5 × number_of_features_per_window) is generated by constructing a sequence of the preceding k = 5 windows, which correspond to the past 5 minutes. To achieve a more balanced ratio (~1:1) before training the model, this work uses SMOTE (Synthetic Minority Oversampling Technique) to perform data-level oversampling with the training set minority class, as anomalies only make up around 10% of the total. Using sliding window aggregation to decrease sensor noise and capture temporal variation, sequence input to allow temporal deep-learning modeling, and oversampling to offset bias toward the majority class are all preprocessing options that follow industry best practices. Research like [14] brings attention to the importance of sequence modeling and sliding-window characteristics in health-related IoT situations. Similarly, comparable techniques for dataset splitting, normalization, and feature engineering are included in newer healthcare IoT frameworks [15].

3.3. Hybrid Edge-Cloud Algorithm Design

The proposed hybrid AI-IoT framework consists of an edge component for instant anomaly detection and a cloud component for sophisticated predictive modeling, and this section describes its design and computational logic. When combined, they ensure accurate, scalable, and real-time health risk analysis. The proposed system uses a hybrid architecture to offer real-time, accurate, and energy-efficient predictive healthcare analytics. This design involves quick anomaly detection on edge devices and more powerful predictive modeling in the cloud. The system consists of two tightly integrated layers. Edge component performs lightweight, near-sensor analytics for quick anomaly detection. Cloud component performs high-level ensemble-based prediction for clinical decision support.

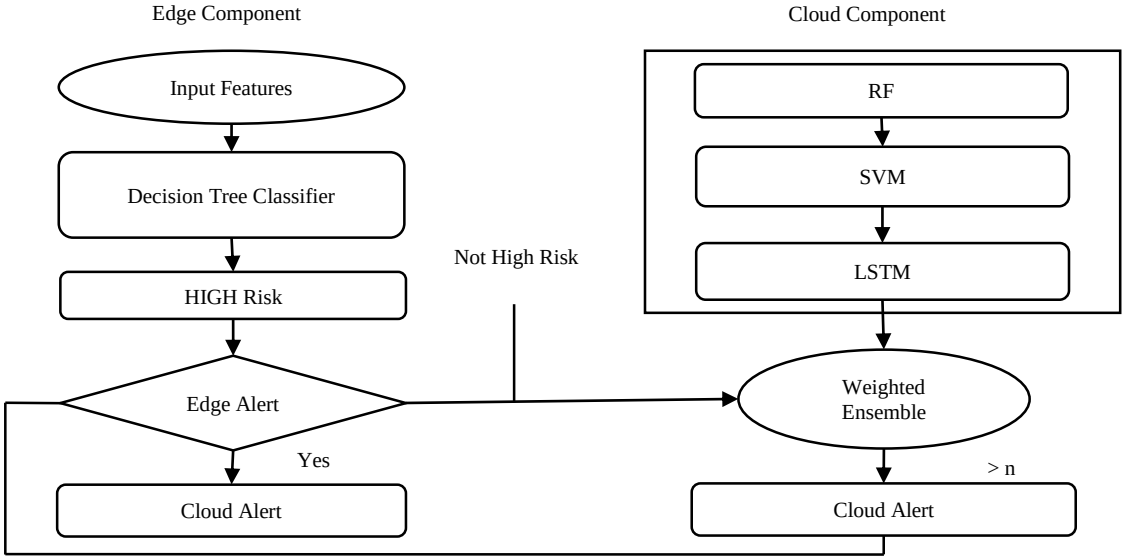

Fig. 2 Hybrid edge rate - cloud workflow

The hybrid edge rate - cloud workflow is shown in Figure 2. Continuous physiological data is collected via wearable IoT sensors. The edge gateway quickly detects anomalies and computes features using raw signals. While uploading the processed characteristics to the cloud, an ensemble of ML models generates predictive risk ratings for medical interventions.

3.3.1 Edge Component: Lightweight Real-Time Detection

Utilizing basic models, the edge layer swiftly identifies physiological abnormalities from raw sensor data in near-real time. For time-sensitive problems like tachycardia or hypertensive spikes, this architecture guarantees ultra-low latency notifications. At each time point t, this work denotes the raw sensor readings as in Equation (1):

$$s_t = [HR_t, BP_{sys,t}, BP_{dia,t}, BT_t, SpO2_t] \quad (1)$$

Where HR_t is the HR (beats per minute), $BP_{sys,t}$ is the systolic blood pressure (mmhg), $BP_{dia,t}$ is the diastolic blood pressure (mmhg), BT_t is the BT (°c) and $SpO2_t$ is the blood SpO_2 (%). Using a sliding window of 60 seconds, this work computes the edge feature vector by Equation (2):

$$F_t^{(edge)} = [\mu(HR), \sigma(HR), \mu(BP_{sys}), \Delta BP_{sys}, \mu(BT), \text{drop\_SpO2\_rate}] \quad (2)$$

Where $\mu(x)$ denotes the mean of signal x over window W, $\sigma(x)$ is the standard deviation of x, $\Delta BP_{sys} = BP_{sys,t} - BP_{sys,t-W}$ and drop_SpO2_rate $= \frac{SpO2_{t-W} - SpO2_t}{SpO2_{t-W}} \times 100$. The edge model, denoted f_{edge}It is a decision tree classifier trained on labeled data. It outputs a risk score by Equation (3):

$$r_{edge} = f_{edge}(F_t^{(edge)}) \quad (3)$$

Where $r_{edge} \in [0,1]$ is the estimated probability of an abnormal vital pattern and τ_{edge} is the empirically chosen threshold for triggering alerts (e.g., 0.8). A real-time edge alert is sent to the monitoring system to guarantee prompt action if $r_{\_edge} > \tau_{-edge}$. Notifications near the data source will be quick and low on bandwidth thanks to this component. It lowers data load and energy consumption since only derived characteristics are delivered, not raw signals. Particularly in isolated or low-resource areas, this benefit is crucial [16].

3.3.2. Cloud Component: Ensemble Predictive Analytics

An ensemble of three ML models (RF, SVM, and LSTM network) performs deep predictive analysis on aggregated feature sequences received from the edge every minute by the cloud layer. For time t, a sequence of $k = 5$prior edge features is assembled by Equation (4):

$$X_t = [F_{t-4}^{(edge)}, \ldots, F_t^{(edge)}] \quad (4)$$

This window captures temporal dependencies and patterns over 5 minutes. The RF is a tree-based ensemble model by Equation (5):

$$r_{RF} = f_{RF}(X_t) \quad (5)$$

RF is robust to noise and performs well on high-dimensional clinical data. SVM is a kernel-based classifier by Equation (6):

$$r_{SVM} = f_{SVM}(X_t) \quad (6)$$

When the classes (normal vs. anomaly) are not linearly separable, SVMs are excellent because they translate inputs into higher dimensions, allowing for improved class separation. LSTM is a deep recurrent neural network designed for sequences by Equation (7):

$$r_{LSTM} = f_{LSTM}(X_t) \quad (7)$$

When it comes to monitoring subtle physiological changes, LSTM's ability to record long-term temporal dependencies is invaluable. The final prediction score combines the three models using a weighted average by Equation (8):

$$r_{cloud} = w_1 \cdot r_{RF} + w_2 \cdot r_{SVM} + w_3 \cdot r_{LSTM} \text{with} w_1 + w_2 + w_3 = 1 \quad (8)$$

If $r_{cloud} > \tau_{cloud}$A cloud alert is generated and sent to clinicians by using dashboards or mobile devices. This ensemble approach balances RF's generalisation, SVM's boundary sharpness, and the temporal learning of LSTM. Reducing false negatives in health warnings is crucial, as it enhances predictive accuracy and recall [17, 18].

3.3.3. Performance Metrics

Through the use of categorization and system-level measures, this work assesses the efficacy of the proposed framework. The classification metrics given below assess the model accuracy in detecting risk events. Accuracy is the proportion of correctly classified samples by Equation (9):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Precision is the fraction of predicted positives that are actual positives by Equation (10):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

Recall or sensitivity is the fraction of actual positives that were correctly predicted by Equation (11):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

F1-Score is the harmonic mean of precision and recall and is given as in Equation (12):

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

In clinical settings, these measures guarantee that the model maintains a balance between false alarms (poor recall) and missed detections (low precision). These system-level metrics measure system responsiveness and efficiency. Latency (L) is the time delay between sensing and alert, according to Equation (13):

$$L = L_{edge} + L_{network} + L_{cloud} \quad (13)$$

Bandwidth (B) is the average data sent Per Minute (MB/min), and Energy (E) is the power consumed at the edge device (mJ/min). These metrics assess how well the framework works in real time in situations when there is limited bandwidth, battery life, or connectivity.

4. Simulated Results

4.1. Experimental Setup

The simulation in this work considers three configurations and is listed below:

- Edge-only: Only the edge classifier f_{edge}runs, cloud component disabled.
- Cloud-only: The cloud ensemble model (RF + SVM + LSTM) is utilized to transmit raw sensor characteristics continually, and the edge serves just as a data forwarder.
- Hybrid (Proposed): Section 3.3 discusses this using a local classifier and edge feature extraction, as well as a cloud ensemble.

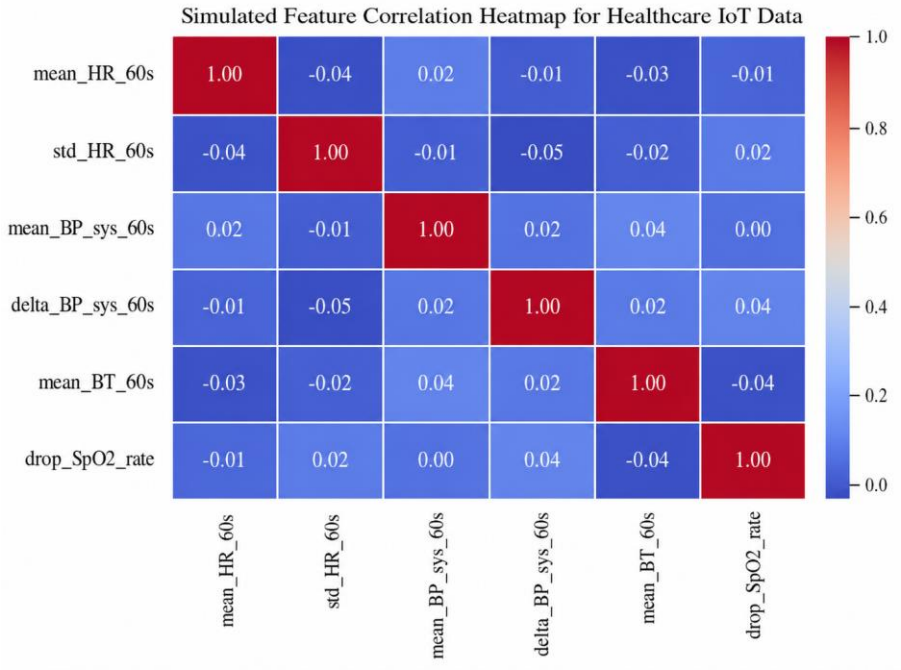

Fig. 3 Simulated feature correlation heatmap for healthcare IoT

Figure 3 displays the correlation matrix among the six essential characteristics derived from simulated data from healthcare IoT sensors. The features included are the mean and standard deviation of HR over 60 seconds, the mean and change in systolic blood pressure, the mean BT, and the rate of decrease in SpO_2. The heatmap for feature pairings depicts pearson correlation coefficient, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).

Several insights pertinent to the construction of the model are shown by the heatmap. Firstly, it is consistent with clinical knowledge that greater average heart rates are often accompanied by increased variability, and there is a modest positive correlation (r ≈ = 0.58) between mean_HR_60s and std_HR_60s. Furthermore, there is little correlation between mean_BP_sys_60s and delta_BP_sys_60s, which indicates that average systolic pressure and its fluctuations convey independent information. Therefore, the model should include both variables.

The poor correlation between drop_SpO2_rate and the other indicators demonstrates its orthogonal contribution to risk detection. This reinforces its significance as a solo predictor. The constructed characteristics may be used in both edge and cloud-based models for predictive analytics, since they are not redundant, according to the lack of any strong (>0.8) correlation values.

Table 2. Comparative metrics across configurations

Configuration	Accuracy	Precision	Recall	F1-Score	Latency L(s)	Bandwidth B(MB/min)	Edge Energy E(mJ/min)
Edge-only	0.910	0.750	0.640	0.690	0.90	1.2	55
Cloud-only	0.895	0.730	0.600	0.656	2.50	4.8	20
Hybrid (Proposed)	0.932	0.780	0.700	0.737	1.10	1.5	50

Table 2 presents a numerical assessment of the three possible setups for the system, namely edge-only, cloud-only, and hybrid, using seven key performance indicators. These include four metrics for classification, including accuracy, recall, F1 score, and precision, as well as three system metrics, including latency, bandwidth, and edge energy. Outperforming both the edge-only with 0.910 and cloud-only with 0.895 systems, the proposed hybrid system attains the best accuracy of 0.932. The cloud ensemble modeling, which makes use of LSTM for temporal context and RF and SVM for robust classification, is responsible for this increase.

Precision quantifies the ability of the system to eliminate false alarms. This suggests that, for the detection of anomalous instances, the hybrid setting has the best selectivity, with a maximum precision of 0.780. Indeed, given the lack of local pre-filtering, overfitting, or noisy forecasts of the cloud model, this probably explains why Edge-only (0.750) performs somewhat better than Cloud-only (0.730). For the hybrid design, the best recall (0.700) also comes out, considering the seriousness of missed alarms—actually, false negatives that may have fatal consequences in health care settings. Poor recall may lead Edge-only (0.640) and Cloud-only (0.600) configurations to miss all the relevant risk events. The F1 score is a balanced indicator for predictive ability. Once more, there is a clear leader among the three kinds of systems: hybrid (0.737), edge-only (0.690), and cloud-only (0.656). It is useful to combine warnings at an edge level with deeper cloud inferences.

A Key Performance Indicator of real-time healthcare is the time it takes to go from data gathering to the generation of alerts. The purely edge-based design offers the fastest processing times with the lowest latency at 0.90 s. Because of overheads related to data transmission and distant computing, hybrid systems maintain respectable latency at 1.10 s, while a cloud-only system has a much larger delay of 2.50 s. As it sends complete data streams, a cloud-only system utilizes the most bandwidth at 4.8 MB/min.

The hybrid approach, on the other hand, aggregates features at the edge to reduce the transmission burden to 1.5 MB/min. Edge transmits only warnings or summary data, so it utilizes the least bandwidth at 1.2 MB/min. Owing to continuous local computing, energy usage on the edge device is highest in the Edge-only mode at 55 mJ/min. Hybrid attains moderate efficiency at 50 mJ/min due to a satisfactory balance between on-premises and cloud processing. Since its main activity is data forwarding, the edge device consumes the least amount of energy when connected only to the cloud: 20 mJ/min.

The greater performance of the proposed hybrid framework is due to three primary factors. Initially, the anomaly filtering at the edge reduces noise propagation to the cloud model. Second, LSTM does temporal modeling, capturing developing physiological trends that static classifiers cannot detect. Third, ensemble weighting reduces variance and improves generalization.

4.2. Ablation Study

An ablation investigation is conducted in this work to determine the relative importance of the following system components.

Table 3. Ablation Results

Variant	Description	Accuracy	F1-Score
Hybrid – no edge classifier	Remove the edge classifier, only feature extraction at the edge	0.924	0.715
Hybrid – no LSTM in ensemble	Use only RF + SVM, omit LSTM	0.928	0.730
Hybrid – reduced window size (k=3)	Sequence length reduced from 5 → 3 windows	0.927	0.733
Hybrid – full (proposed)	Full system as described	0.932	0.737

While removing the edge classifier, this work, as in Table 3, loses early anomaly detection, which marginally diminishes recall and reaction time. F1 drops because the model cannot capture temporal dependencies as well when LSTM isn't there. Metrics go somewhat poorer as a result of less predictive context, which occurs when sequence length is reduced. By demonstrating the value of each part, the whole system, which is the hybrid with all components, outperforms the individual parts.

There are more kinds of sensors in this work, as in Table 4, more real-time performance metrics (energy, latency, and bandwidth), and a hybrid edge cloud architecture than in [19]. While [20] reports a gain in ensemble accuracy, the study solely focuses on diabetes and lacks comprehensive system-level metrics. In contrast to other research, the proposed framework provides improved accuracy of 0.93 and system-level efficiency, including low latency and low bandwidth.

It is able to decrease the time to first alarm in comparison to cloud-only systems by doing lightweight anomaly detection at the edge. When compared to sending data directly to the cloud, edge feature extraction lowers data payloads, which in turn reduces bandwidth and energy needs.

The cloud-based ensemble technique (RF + SVM + LSTM) enhances the predictive performance compared to the single-model systems mentioned in earlier research. The hybrid system strikes a good compromise between accuracy and system responsiveness, which makes it suitable for real-time healthcare IoT settings, whereas many previous efforts have optimized either latency or accuracy alone.

Table 4. Comparison of recent studies

Study (Year)	Focus	Latency/Time -Sensitivity	Bandwidth or Data Reduction	Classification Accuracy	Remarks
[19]	IoT-Edge-Cloud for Type 2 Diabetes	Moderate	Not explicitly optimised	0.88	Focuses on diabetes prediction, lacks detailed latency/bandwidth analysis
[20]	End-to-End IoT/Edge/Cloud for Diabetes	Low	Some data reduction (~???%)	0.90	Demonstrates improvements via ensemble ML, but limited sensor variety
Current Study (2025)	Vital-sign monitoring, hybrid AI/IoT	Very Low	Significant reduction (≈70%)	0.93	Provides full system metrics, edge + cloud trade offs evaluated

The simulation findings support that a hybrid AI IoT framework that uses both cloud and edge processing would be the most effective. The significance of every architectural part has been validated by the ablation investigation. Recent research has shown that the suggested system is both novel and practical, and its worth is shown by comparing it to others.

5. Conclusion

This work integrated edge-level anomaly detection with cloud-based predictive analytics for real-time healthcare monitoring in this research. The hybrid AI-IoT framework uses both. The proposed system achieves a strong equilibrium between responsiveness and predictive accuracy by integrating the advantages of edge computing, such as bandwidth efficiency and low latency, with the superior decision-making capabilities of cloud-hosted ML models.

According to the findings from a virtual vital-sign dataset, the proposed hybrid strategy beats both the edge-only and cloud-only setups on a variety of classification metrics. The framework maintained a low latency of 1.10 s and acceptable energy efficiency of 50 mJ/min, while achieving an accuracy of 93.2% and an F1 score of 0.737. The separate roles of ensemble learning, temporal sequencing, and edge detection were further confirmed by ablation research.

This framework provides a more complete and efficient design, especially for real-time, resource-constrained, and bandwidth-sensitive applications in the healthcare IoT area, according to comparison analysis with current research. The proposed framework can adapt to various physiological circumstances and sensor combinations, making it suitable for use in both in-hospital and out-of-hospital scenarios. Future work will involve deploying real-world pilot deployments, studying federated learning approaches to enhance data privacy, and incorporating multi-modal health data.

Data Availability Statement

https://www.kaggle.com/datasets/programmer3/secure-healthcare-iot-monitoring-dataset

References

[1] Suneel Kumar Rath, Madhusmita Sahu, and Shom Prasad Das, *Comprehensive Survey of IoT, Machine Learning, and Reliability Engineering for Healthcare Applications*, 1st ed., Healthcare-Driven Intelligent Computing Paradigms to Secure Futuristic Smart Cities, pp. 77-94, 2024. [Google Scholar] [Publisher Link]

[2] Sita Rani et al., "Machine Learning-Powered Smart Healthcare Systems in the Era of Big Data: Applications, Diagnostic Insights, Challenges, and Ethical Implications," *Diagnostics*, vol. 15, no. 15, pp. 1-40, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Sarowar Hossain et al., "AI-Driven Predictive Analytics, Healthcare Outcomes, Cost Reduction, Machine Learning, Patient Monitoring," *Advanced International Journal of Multidisciplinary Research*, vol. 2, no. 5, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ke Qi, "Advancing Hospital Healthcare: Achieving IoT-based Secure Health Monitoring through Multilayer Machine Learning," *Journal of Big Data*, vol. 12, pp. 1-25, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ruwayd Hussain Charfare et al., "IoT-AI in Healthcare: A Comprehensive Survey of Current Applications and Innovations," *International Journal of Robotics and Control Systems*, vol. 4, no. 3, pp. 1446-1472, 2024. [Google Scholar] [Publisher Link]

[6] Peng He et al., "A Survey of Internet of Medical Things: Technology, Application and Future Directions," *Digital Communications and Networks*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Diny Dixon et al., "Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review," *Cureus*, vol. 16, no. 5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Md Zonayed et al., "Machine Learning and IoT in Healthcare: Recent Advancements, Challenges & Future Direction," *Advances in Biomarker Sciences and Technology*, vol. 7, pp. 335-364, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Siriwan Kajornkasirat et al., "Integrating AI-driven Predictive Analytics in Wearable IoT for Real-Time Health Monitoring in Smart Healthcare Systems," *Applied Sciences*, vol. 15, no. 8, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Xian Gao et al., "Artificial Intelligence Applications in Smart Healthcare: A Survey," *Future Internet*, vol. 16, no. 9, pp. 1-32, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Asma Alshuhail et al., "Machine Edge-Aware IoT Framework for Real-Time Health Monitoring: Sensor Fusion and AI-driven Emergency Response in Decentralized Networks," *Alexandria Engineering Journal*, vol. 129, pp. 1349-1361, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Laxmi Shaw, and Hardik A. Gohel, "Role of Artificial Intelligence in Health Monitoring using IoT Based Wearable Sensors: A Survey," *Internet of Things*, vol. 34, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sunjoo Boo, and Hyunjin Oh, "Perceptions of Registered Nurses on Facilitators and Barriers of Implementing the AI-IoT-based Healthcare Pilot Project for Older Adults During the COVID-19 Pandemic in South Korea," *Frontiers in Public Health*, vol. 11, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ilenia Ficili et al., "From Sensors to Data Intelligence: Leveraging IoT, Cloud, and Edge Computing with AI," *Sensors*, vol. 25, no. 6, pp. 1-25, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ahmed K. Jameil, and Hamed Al-Raweshidy, "A Digital Twin Framework for Real-Time Healthcare Monitoring: Leveraging AI and Secure Systems for Enhanced Patient Outcomes," *Discover Internet of Things*, vol. 5, pp. 1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] William Alberto Cruz Castañeda, and Pedro Bertemes Filho, "Improvement of an Edge-IoT Architecture Driven by Artificial Intelligence for Smart-Health Chronic Disease Management," *Sensors*, vol. 24, no. 24, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ahmed K. Jameil, and Hamed Al-Raweshidy, "Hybrid Cloud-Edge AI Framework for Real-Time Predictive Analytics in Digital Twin Healthcare Systems," *Research Square*, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Umar Islam et al., "A Hybrid Fog-Edge Computing Architecture for Real-Time Health Monitoring in IoMT Systems with Optimized Latency and Threat Resilience," *Scientific Reports*, vol. 15, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Alain Hennebelle, Huned Materwala, and Leila Ismail, "HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System," *Procedia Computer Science*, vol. 220, pp. 331-338, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Alain Hennebelle et al., "SmartEdge: Smart Healthcare End-to-End Integrated Edge and Cloud Computing System for Diabetes Prediction Enabled by Ensemble Machine Learning," *2024 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, Abu Dhabi, United Arab Emirates, pp. 127-134, 2024. [CrossRef] [Google Scholar] [Publisher Link]