JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Multimodal Deep Learning for Pneumonia Detection Using Wearable Sensors: Toward an Edge-Cloud Framework

Emmanuel Onwuka Ibam [1] and Johnson Bisi Oluwagbemi [2,*]

[1] Department of Information Technology, Federal University of Technology, Akure 110001, Ondo State, Nigeria; e-mail : eoibam@futa.edu.ng
[2] Department of Computer Science, McPherson University, Seriki-Sotayo 340001, Ogun State, Nigeria; e-mail : oluwagbemijb@mcu.edu.ng
* Corresponding Author : Johnson Bisi Oluwagbemi

Abstract: Pneumonia remains a leading cause of morbidity and mortality worldwide, particularly in resource-limited settings and among elderly populations, where timely diagnosis and continuous monitoring are often constrained by limited clinical infrastructure. This study presents an edge–cloud–integrated framework for early pneumonia risk monitoring, leveraging multimodal wearable sensors and deep learning to support continuous short-duration monitoring. The proposed system is designed to operate in near real time under simulated deployment conditions, continuously acquiring and analyzing physiological signals (respiratory rate, heart rate, SpO_2, and body temperature) alongside event-driven acoustic biomarkers (cough sounds) within a distributed architecture. A lightweight edge module performs local signal preprocessing and anomaly triage, selectively transmitting salient information to a cloud-based multimodal deep learning model for refined risk estimation and interpretability analysis. The framework was evaluated using a multi-source dataset comprising public repositories (MIMIC-III and Coswara) and a clinically supervised wearable study conducted in two Nigerian hospitals, resulting in 718 hours of quality-controlled multimodal monitoring data. In a pooled multi-source evaluation, the system achieved an AUC of 0.95, while in a clinically realistic local-only evaluation, the AUC was 0.86, reflecting a consistent but preliminary diagnostic signal. These results highlight the importance of local data adaptation for real-world applicability and suggest that multimodal AI can provide meaningful early risk indicators under resource constraints. Beyond predictive performance, this work demonstrates the feasibility of integrating multimodal learning, edge–cloud computation, and explainable analytics into a deployment-aware, privacy-preserving monitoring framework for low-resource healthcare environments.

Keywords: Deep Learning; Edge–Cloud Computing; Explainable AI; Multimodal Learning; Pneumonia Detection; Remote Health Monitoring; Wearable Sensors; Wireless Health Systems.

Received: October, 17th 2025
Revised: January, 18th 2026
Accepted: : January, 19th 2026
Published: January, 20th 2026

1. Introduction

Pneumonia remains one of the leading causes of mortality worldwide, particularly in low- and middle-income countries where early diagnosis and continuous clinical monitoring are constrained by limited healthcare infrastructure [1]–[3]. Conventional diagnosis relies heavily on chest radiography and laboratory testing, which are resource-intensive, time-consuming, and unsuitable for remote or continuous monitoring scenarios [4], [5]. Recent advances in wearable sensing technologies and artificial intelligence have opened new opportunities for continuous, non-invasive respiratory health monitoring, particularly outside hospital settings [6]–[8].

However, most existing AI-based pneumonia detection systems remain unimodal, relying on either physiological signals or acoustic features alone. Such approaches are inherently fragile, as real-world sensor data are often incomplete, noisy, or context-dependent, leading to poor generalization across populations and environments. To address these limitations, multimodal learning—which integrates heterogeneous data streams such as physiological time series, cough acoustics, and demographic information—has emerged as a promising

DOI : 10.62411/jcta.14944

paradigm for more robust clinical inference [9]. By leveraging complementary modalities, multimodal fusion enables models to capture latent interactions that are inaccessible to single-modality systems, resulting in more stable and clinically meaningful predictions [10]–[13].

Despite its promise, multimodal health analytics faces three persistent challenges: (i) Data heterogeneity, where devices operate at different sampling rates and produce incompatible feature spaces; (ii) Computational constraints, as multimodal deep learning models are often too heavy for deployment on wearable or edge devices; and (iii) Domain shift and interpretability, where models trained on public datasets fail to generalize to local populations, and their decision processes remain opaque to clinicians [14]–[16].

Edge–cloud computing offers a principled solution to these challenges by partitioning intelligence across layers. The edge enables low-latency preprocessing and privacy-aware triage close to the user, while the cloud supports computationally intensive multimodal fusion, domain adaptation, and interpretability analysis [17]–[19]. This hierarchical design balances responsiveness, scalability, and clinical transparency, making it particularly suitable for resource-constrained healthcare environments.

Motivated by this paradigm, a multimodal learning strategy is adopted in which modality-specific feature representations are learned independently and subsequently integrated through a domain-adaptive fusion mechanism. Acoustic, physiological, and static data are processed using dedicated encoders to preserve the structural characteristics of each modality, while feature alignment across heterogeneous sources is enforced through adversarial domain adaptation. This design enables the model to learn representations that are both discriminative for pneumonia detection and robust to domain shifts arising from differences in acquisition settings, devices, and populations [20], [21]. To further support clinical transparency, explainable AI techniques based on SHAP are incorporated to provide post-hoc interpretability of the model's predictions, enabling examination of the contributions of each modality and feature group.

Within this framework, the edge–cloud architecture is treated not merely as a deployment option but as an integral component of the learning and inference process. Computationally lightweight operations are assigned to the edge layer to support timely preprocessing and triage, while resource-intensive fusion, adaptation, and interpretability analyses are performed in the cloud. Based on this design, a domain-adaptive multimodal deep learning framework for pneumonia detection using wearable sensors is presented, bridging local data acquisition with design-oriented scalable cloud-based intelligence while maintaining interpretability and deployment feasibility.

2. Related Works

2.1. Single-Modality Pneumonia Detection

Early studies on pneumonia detection predominantly relied on single data modalities, focusing on either acoustic or physiological signals. Audio-based approaches, such as those introduced in [2], [4], [22], [23], utilized cough and breath-sound features processed by convolutional neural networks (CNNs). While these methods achieved promising accuracy under controlled conditions, their performance was often sensitive to ambient noise, microphone variability, and demographic bias, which limited their generalizability in real-world settings.

Similarly, physiological-signal-based methods trained on large clinical datasets, such as MIMIC, demonstrated strong predictive performance for respiratory conditions; however, their applicability to wearable and real-time monitoring remained limited. Models trained on such data frequently exhibited poor cross-device consistency and reduced robustness when transferred to non-clinical environments [24]. As a result, unimodal frameworks remain vulnerable to signal degradation, missing data, and contextual variability, underscoring the need for multimodal approaches that integrate complementary physiological and acoustic information for more resilient diagnostics.

2.2. Multimodal Health Diagnostics

To address the limitations of unimodal systems, recent research has increasingly explored multimodal learning strategies that combine respiratory, cardiovascular, and acoustic data streams. Multimodal CNN–LSTM architectures have been reported to improve respiratory disease classification performance by 6–10% compared to unimodal baselines [25]–[28].

These improvements highlight the ability of multimodal models to capture complementary and correlated patterns across heterogeneous signals.

Despite these advances, most existing multimodal frameworks are implemented exclusively in cloud-based environments, requiring high bandwidth and continuous connectivity. This dependency limits their suitability for continuous monitoring in resource-constrained or remote settings. Furthermore, interpretability remains underexplored in multimodal healthcare analytics, with only a limited number of studies incorporating explainable AI techniques such as SHAP to clarify how individual modalities and features contribute to diagnostic outcomes. This gap limits clinical trust and hinders the adoption of multimodal AI systems in practice.

2.3. Edge–Cloud Computing in Healthcare

Edge–cloud architectures have emerged as a promising paradigm for design-oriented scalable and responsive healthcare systems. In such architectures, the edge layer performs low-latency preprocessing and privacy-preserving data handling close to the data source, while the cloud layer supports large-scale model training, global updates, and long-term analytics [17], [29], [30]. This hierarchical structure enables a balance between computational efficiency and real-time responsiveness, which is critical for continuous health monitoring in remote and resource-limited environments.

However, most existing edge–cloud implementations in healthcare remain unimodal or focus primarily on computational offloading without explicitly addressing the heterogeneity of multimodal health data. Signals originating from diverse sensors often differ in sampling rates, formats, and noise characteristics, complicating integration and learning. Moreover, many studies emphasize either system performance or computational efficiency, while overlooking the integration of explainability within the edge–cloud pipeline. Consequently, achieving domain robustness, interpretability, and resource efficiency within a unified framework remains an open research challenge.

2.4. Research Gap

Although significant progress has been made in multimodal AI for healthcare, several critical gaps persist. First, few frameworks explicitly address cross-domain generalization, particularly the adaptation of models trained on large public datasets to locally collected wearable data. Second, edge–cloud integration for scalable and near real-time deployment remains underdeveloped, as many systems continue to rely on centralized cloud-based inference. Third, model interpretability is often treated as an afterthought, leaving clinicians with limited insight into the physiological rationale behind AI-driven predictions.

To address these gaps, this study introduces a domain-adaptive, SHAP-explainable, edge–cloud-enabled multimodal deep learning framework for pneumonia detection. The proposed approach integrates heterogeneous physiological and acoustic data using adaptive feature alignment to enhance robustness across data sources. In addition, SHAP-based analysis is employed to quantify the relative contribution of each modality and feature group, providing interpretable insights aligned with clinical reasoning. By unifying distributed computation, multimodal learning, and explainable AI, the framework advances toward transparent, robust, and deployable pneumonia diagnostics suitable for low-resource environments.

3. Methodology

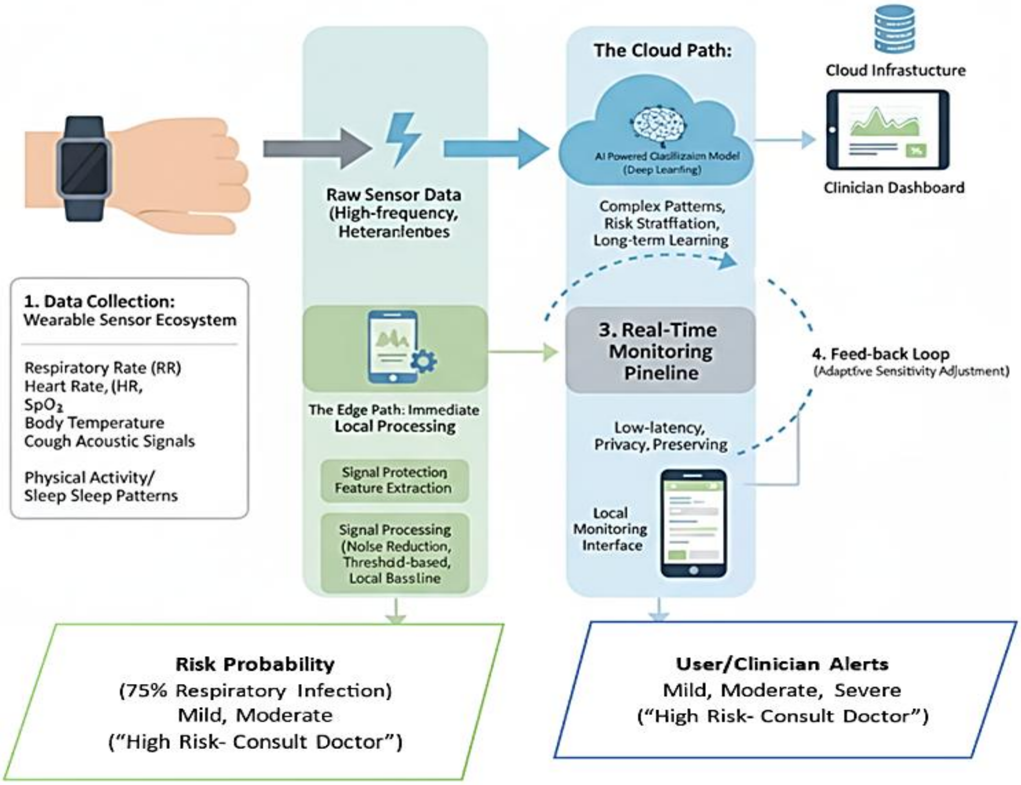
3.1. Framework Overview

This section presents a hybrid edge–cloud computing framework designed to support intelligent and continuous health monitoring using wearable sensors. As illustrated in Figure 1, raw sensor data are processed through two parallel and complementary pathways—edge and cloud—to balance low-latency responsiveness with computationally intensive analysis. In the proposed design, processing tasks are deliberately divided between the two layers to exploit the strengths of each. The edge layer prioritizes immediate, local processing close to the data source, while the cloud layer performs more complex, resource-intensive inference. This dual-processing strategy enables timely feedback without compromising analytical depth or model complexity.

The cloud processing path receives aggregated sensor data and applies a deep learning–based classification model trained to identify complex temporal and cross-modal patterns associated with health risk. Because cloud infrastructure provides design-oriented scalable computational resources, this module can execute multimodal fusion, long-term learning, and high-dimensional inference that would be infeasible on resource-constrained wearable devices. The cloud model's output consists of refined risk estimates and severity assessments, which are forwarded to the central monitoring pipeline for integration.

In parallel, the edge processing path performs immediate local analysis on or near the user's device. This module executes two primary functions. First, signal-processing operations are applied to clean raw sensor data, reducing noise and normalizing baseline variations, preparing the signals for rapid analysis. Second, a lightweight anomaly detection mechanism identifies deviations from the individual's typical physiological patterns, enabling early warning detection with minimal latency. The outputs of the edge module are delivered directly to the local monitoring interface and simultaneously transmitted to the cloud for contextual interpretation.

The outputs from both processing paths converge in the real-time monitoring pipeline, which serves as the system's central decision-making hub. By integrating cloud-level analytical insights with edge-level anomaly indicators, the pipeline generates a holistic assessment consisting of (i) a continuous risk probability score, (ii) a discrete severity category, and (iii) alert notifications for the user, caregiver, or clinician when predefined thresholds are exceeded. A key feature of the proposed framework is the adaptive feedback loop, also shown in Figure 1, which connects the monitoring pipeline back to the edge processing module. This mechanism allows cloud-derived risk assessments to recalibrate the sensitivity of local anomaly detection. When elevated risk is identified, the edge module can be dynamically adjusted to enhance responsiveness to early warning signs, enabling personalized, progressively adaptive monitoring over time.



Raw Sensor Data (High-frequency, Heteranlenbes
The Cloud Path:
AI Powered Clasifizaion Model (Deep Leanfing)
Cloud Infrastucture
Clinician Dashboard
Complex Pattems, Risk Straffiation, Long-term Learning
1. Data Collection: Wearable Sensor Ecosystem
Respiratory Rate (RR)
Heart Rate, (HR,
SpO_2
Body Temperature
Cough Acoustic Signals
Physical Activity/ Sleep Sleep Patterns
3. Real-Time Monitoring Pineline
4. Feed-back Loop (Adaptfve Sensitivity Adjustment)
The Edge Path: Immediate Local Processing
Low-latency, Privacy, Preserving
Signal Protection Feature Extraction
Signal Processing (Nolse Reduction, Threshdd-based, Local Bassline
Local Monitoring Interface
Risk Probability
(75% Respiratory Infection)
Mild, Moderate
("High Risk- Consult Doctor")
User/Clinician Alerts
Mild, Moderate, Severe
("High Risk- Consult Doctor")

Figure 1. Edge–cloud architecture for wearable-based pneumonia monitoring.

3.2. Data Sources and Study Design

3.2.1 Public Datasets

Two public datasets were used to pretrain and benchmark unimodal components of the framework:

- MIMIC-III Waveform Database [31], which provides clinically annotated physiological time-series data from ICU patients with and without pneumonia.
- Coswara Dataset [32], which contains crowd-sourced cough and breathing audio recordings collected under diverse environmental conditions.

These datasets were used exclusively for modality-specific representation learning and baseline validation, as they do not share common subjects and therefore cannot support synchronized multimodal fusion.

3.2.2. Local Clinical Dataset

To enable synchronized multimodal fusion under realistic acquisition conditions, a clinically supervised wearable monitoring study was conducted at the Federal University of Technology Akure Health Centre and McPherson University Health Centre in Nigeria between October 2024 and June 2025. Ethical approval was obtained from the institutional review board, and all participants provided written informed consent prior to enrollment. A total of 52 volunteers (29 male, 23 female; mean age 58.2 years) were monitored continuously for 24 hours using wearable sensors. Physiological signals were sampled at one-minute intervals, while cough events were captured using an event-triggered audio mechanism to reflect naturalistic recording conditions. Following quality control procedures, 1,912 high-quality cough recordings and over 1,000 hours of physiological data were retained for analysis.

Clinical ground-truth labels were adjudicated by attending clinicians based on physical examination and radiographic confirmation, resulting in 16 high-risk (pneumonia) and 36 low- to moderate-risk cases. While the cohort size reflects the practical constraints of supervised wearable data collection in real clinical settings, it provides a sufficiently controlled environment to examine multimodal signal coherence, cross-modal complementarity, and system-level behavior under synchronized conditions. Accordingly, this dataset is used to evaluate integrated fusion, domain adaptation, and edge–cloud interaction, rather than to estimate population-level diagnostic prevalence or statistically powered clinical accuracy.

3.2.3. Multi-source Dataset Composition and Rationale

Table 1 summarizes the composition of each dataset source, including sample counts, recording duration, and class distribution, providing a unified view of data balance and potential source bias. Table 2 describes the role of each dataset within the experimental design, clarifying how public and private data sources were used for pretraining, benchmarking, and multimodal evaluation.

Table 1. Dataset composition and label distribution

Dataset Source	**Data Type**	**Samples (n)**	**Duration (hours)***	**Pneumonia (%)**	**Healthy (%)**
MIMIC-III	Physiological waveforms	4,200	280	50.0	50.0
Coswara	Cough & breathing audio	3,100	190	41.9	58.1
Local Wearable	Multisensor time-series	3,500	248	58.6	41.4
Total	—	10,800	718	50.5	49.5

* Total duration refers to cumulative active recording time across all participants and devices, after filtering and cleaning.

Table 2. Dataset usage and experimental role.

Dataset	**Source**	**Modality**	**Access**	**Primary Use in Study**
MIMIC-III	PhysioNet	Physiological vitals	Public	Pretraining physiological encoder; unimodal benchmarking
Coswara	IISc Bangalore	Cough audio	Public	Pretraining acoustic encoder; audio feature learning
Local Clinical Dataset	FUTA & McPherson Health Centres	Multimodal (wearable + static)	Private	Multimodal fusion; fine-tuning; clinical evaluation

Importantly, record-level fusion was not performed across datasets because the public repositories and the local clinical dataset do not share common subjects. Instead, true multimodal fusion was evaluated exclusively on the local clinical dataset, where synchronized physiological, acoustic, and static data coexist for each participant. Public datasets were incorporated to enhance domain diversity and to pretrain modality-specific encoders, while cross-dataset experiments were used to assess robustness to domain shift rather than to construct synthetic patient profiles. This design ensures methodological validity while supporting scalable model development across heterogeneous data sources.

3.3. Preprocessing, Harmonization, and Domain Adaptation

Combining heterogeneous data sources introduces variability in sampling rates, feature distributions, and signal quality. To mitigate these effects, a unified preprocessing and harmonization pipeline was applied.

- Feature alignment and missing modality handling: Common physiological features were mapped to standardized units across datasets. Modalities absent in a given dataset were represented using binary masks concatenated with the feature vectors, allowing the model to learn modality-invariant representations without conflating missing values with low signal values.
- Per-source normalization: Each data source was normalized independently to prevent dominance by high-amplitude signals. Z-score normalization was applied to unbounded features:

$$\hat{x} = \frac{x - \mu_s}{\sigma_s} \tag{1}$$

where μ_s and σ_s denote the mean and standard deviation of source s. Bounded physiological signals were rescaled using min–max normalization, while audio features were normalized per utterance.

- Domain alignment: To reduce source-dependent variance, adversarial domain adaptation was implemented using a gradient reversal layer (GRL) prior to the fusion block [33]. Additionally, batch normalization statistics were re-estimated per source during fine-tuning, and optional Maximum Mean Discrepancy (MMD) regularization was applied in the latent space[34]. These strategies jointly encourage domain-invariant feature learning while preserving discriminative capacity.
- Temporal imputation and signal quality weighting: Short missing segments in wearable data were interpolated using cubic splines, while longer gaps were masked. Each modality was weighted by a signal-quality index derived from noise metrics, allowing the fusion layer to downweight unreliable channels.

3.4. Multimodal Model Architecture

The proposed model consists of modality-specific encoders followed by a balanced fusion mechanism, as summarized in Table 3, with the acoustic CNN sub-network illustrated in Figure 2.

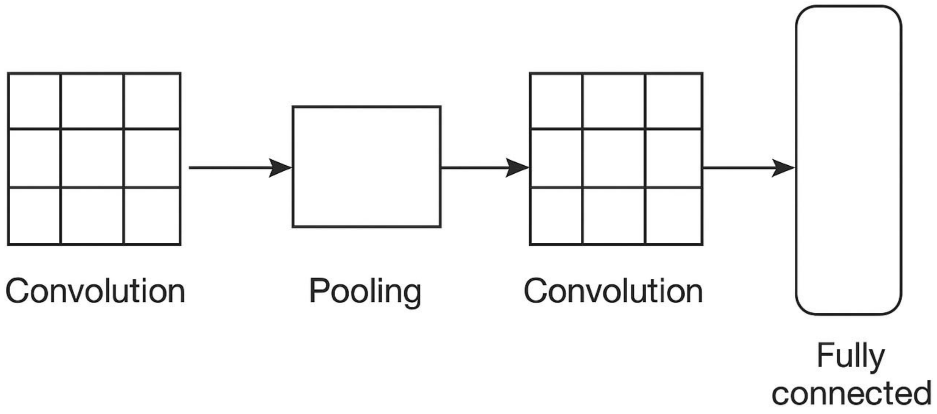


Figure 2. CNN architecture for cough spectrogram classification

Table 3. Multimodal model architecture and output dimensions.

Module	Layer / Component	Layer Type	Configuration	Output Dimension	Description
Acoustic Encoder (Audio Spectrogram – Coswara)	Input	—	128×128 Mel-spectrogram	(128, 128, 1)	Time–frequency representation of cough audio
	Conv Block 1	Conv2D + ReLU	32 filters, 3×3, stride 1	(126, 126, 32)	Local acoustic pattern extraction
	Max Pooling 1	MaxPooling2D	2×2	(63, 63, 32)	Temporal–frequency reduction
	Conv Block 2	Conv2D + ReLU	64 filters, 3×3	(61, 61, 64)	Mid-level acoustic feature learning
	Max Pooling 2	MaxPooling2D	2×2	(30, 30, 64)	Spatial abstraction
	Conv Block 3	Conv2D + ReLU	128 filters, 3×3	(28, 28, 128)	High-level acoustic representations
	Global Pooling	Global Average Pooling	—	(128,)	Fixed-length audio embedding
Physiological Encoder (Time-series – Wearable/MIMIC)	Input	—	10 features × 300 timesteps	(300, 10)	Multivariate physiological time-series
	Bi-LSTM Layer 1	Bi-LSTM	128 units	(300, 256)	Bidirectional temporal encoding
	Bi-LSTM Layer 2	Bi-LSTM	64 units	(300, 128)	Higher-order temporal abstraction
	Attention Layer	Attention	—	(128,)	Weighted temporal summarization
Static Feature Encoder (Demographic/Clinical)	Input	—	Age, sex, BMI, symptoms	(8,)	Encoded tabular static attributes
	Dense Layer	Dense + ReLU	32 units	(32,)	Nonlinear feature transformation
	Dropout	Dropout	0.2	(32,)	Regularization
Multimodal Fusion and Classification	Concatenation	—	[Audio(128) + Physio(128) + Static(32)]	(288,)	Multimodal feature fusion
	Dense Layer	Dense + ReLU	128 units	(128,)	Joint representation learning
	Dropout	Dropout	0.3	(128,)	Generalization control
	Output Layer	Dense + Softmax	2 units	(2,)	Pneumonia risk classification

3.4.1. Modality-specific Encoders

- Acoustic Encoder: A lightweight CNN based on MobileNetV2 [35] processes log-mel spectrograms of cough events.
- Physiological Encoder: A Bi-LSTM network [36] captures temporal dependencies in multivariate physiological time series.
- Static Encoder: Demographic and clinical variables are processed using a small dense network.

Each encoder outputs a fixed-dimensional embedding preserving the structural characteristics of its modality.

3.4.2. Fusion and Balancing Strategy

To prevent dominance of any modality, each embedding is projected into a shared latent space using learnable projection layers:

$$\tilde{h}_i = \mathrm{ReLU}(W_i h_i + b_i) \tag{2}$$

The balanced embeddings are then concatenated to form a unified representation:

$$H_{\text{fused}} = [\tilde{h}_{\text{CNN}} \| \tilde{h}_{\text{BiLSTM}} \| \tilde{h}_{\text{Static}}] \tag{3}$$

This fused vector is passed through fully connected layers to produce the final risk prediction. Alternative scalar modality-weighting was evaluated but omitted in deployment due to minimal performance gains and higher computational cost.

3.4.3. Learning Objective

Model training minimizes a composite loss comprising binary cross-entropy for pneumonia risk prediction and an auxiliary domain-adaptation loss. The output is interpreted as a continuous risk probability, which is subsequently mapped to categorical risk levels for monitoring purposes.

3.5. Edge-side Adaptive Sensitivity Mechanism

To maintain consistent performance across varying sensor and environmental conditions, the edge module dynamically adjusts its alert threshold based on calibration feedback from the cloud. Let p_e denote the edge-predicted pneumonia probability for a given data segment. Periodically, the cloud computes a calibration offset based on the discrepancy between cloud- and edge-level risk estimates:

$$\Delta\tau = \eta\left(\mu_{p_c} - \mu_{p_e}\right) \tag{4}$$

where μ_{p_c} and μ_{p_e} denote the mean predicted probabilities produced by the cloud and edge models, respectively, over synchronized samples, and $\eta \in [0,1]$ is the adaptation rate controlling update smoothness. The edge decision threshold is then updated as:

$$\tau_e^{(t+1)} = \tau_e^{(t)} - \Delta\tau \tag{5}$$

If the edge model is under-sensitive relative to the cloud $\left(\mu_{p_c} < \mu_{p_e}\right)$ the threshold is decreased to increase sensitivity; conversely, if the edge is over-sensitive, the threshold is increased to reduce false alarms. This mechanism enables personalized sensitivity adjustment while preserving edge autonomy during intermittent connectivity, as threshold updates can be applied locally between cloud synchronizations. Algorithm 1 summarizes the adaptive sensitivity update procedure.

Algorithm 1. Adaptive Edge Sensitivity Update Using Cloud Feedback

```
INPUT: Edge prediction probabilities $p_e$ over last $N$ samples; cloud prediction probabilities $p_c$; adaptation rate $\eta$
OUTPUT: Updated edge decision threshold $\tau_e$
Collect local edge predictions $p_e$ from streaming sensor data
Compute summary statistics $\left(\mu_{p_e}, \sigma_{p_e}\right)$ at the edge
Transmit summary statistics to the cloud for calibration.
Compute cloud reference mean $\mu_{p_c}$ using cloud model inference
Compute calibration offset $\Delta\tau = \eta\left(\mu_{p_c} - \mu_{p_e}\right)$
Update edge threshold $\tau_e \leftarrow \tau_e - \Delta\tau$
Apply updated threshold for subsequent edge inferences until next synchronization
```

The methodology described above defines the learning, adaptation, and deployment mechanisms of the proposed framework, while the following section focuses on its empirical evaluation under controlled and clinically realistic conditions. Through the integration of multimodal learning, domain adaptation, and hierarchical edge–cloud computation, the proposed methodology provides a structured and deployment-aware approach to early pneumonia risk monitoring. The design emphasizes interpretability, robustness, and feasibility, forming the foundation for the experimental evaluation presented in Section 4.

4. Experimental Setup

This section describes the experimental protocol used to evaluate the proposed framework, including dataset splits, training configuration, evaluation metrics, and implementation details. The experimental design explicitly distinguishes between method validation and robustness analysis across heterogeneous data sources, in line with the study's multi-source nature.

4.1. Experimental Protocol and Evaluation Scenarios

To quantify both predictive performance and robustness to domain shift, experiments were conducted under four complementary evaluation scenarios:

- Public-only evaluation: In this setup, training and testing were performed exclusively on publicly available datasets. Physiological models were trained and evaluated on MIMIC-III waveform data, while acoustic models were trained and evaluated on Coswara. This scenario establishes unimodal baseline performance under controlled public data conditions.
- Private-only (local) evaluation : Training and testing were conducted solely on the locally collected clinical dataset from the two Nigerian hospitals. This scenario reflects the most clinically realistic deployment condition, as both training and inference data originate from the same acquisition environment, sensor hardware, and population.
- Cross-domain evaluation: To quantify domain shift, models were trained on public datasets and tested on local data, and vice versa. This analysis evaluates how differences in population, device characteristics, and acquisition protocols affect generalization, and highlights the need for domain adaptation.
- Combined (multi-source) evaluation: In this scenario, public and local datasets were pooled at the dataset level (not the record level) to evaluate model behavior under heterogeneous multi-source training. As discussed in Section 3.2, this setup is intended to assess robustness and representation learning under domain diversity rather than to estimate clinical deployment performance.

Across all scenarios, the same end-to-end multimodal evaluation pipeline was used, as illustrated in Figure 3, which shows how acoustic, physiological, and static inputs are integrated through the CNN, Bi-LSTM, and fusion network to produce pneumonia risk predictions. This unified pipeline ensures that performance differences across scenarios reflect data and domain effects rather than changes in model structure.

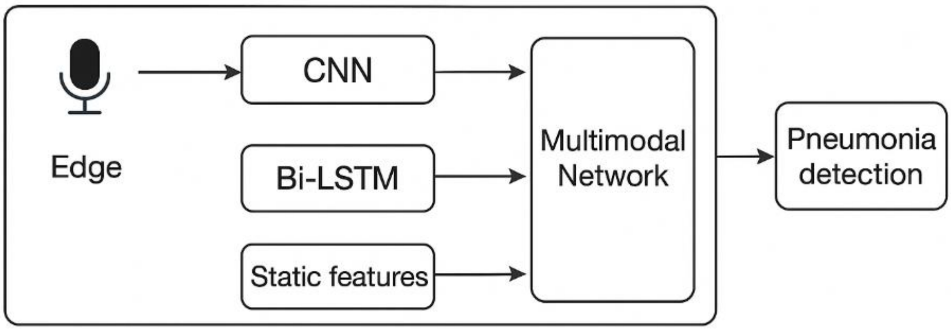


Figure 3. Multimodal deep learning framework for pneumonia detection

4.1.1. Baseline Models and Comparative Configurations

To contextualize the performance of the proposed multimodal framework, several baseline and comparative models were implemented using consistent preprocessing pipelines and training protocols. These baselines include classical machine learning models, unimodal deep learning models, and the proposed multimodal fusion network. Their configurations and key hyperparameters are summarized in Table 4.

Logistic regression was implemented using a standard convex optimization solver (liblinear) with L2 regularization. No learning rate was required, as optimization was performed using a second-order solver. SVM models serve as interpretable classical baselines for statistical physiological features, while the CNN and Bi-LSTM models represent strong unimodal deep learning baselines for acoustic and physiological modalities, respectively. The multimodal fusion network integrates all available modalities and serves as the proposed system, evaluated across the experimental scenarios described in Section 4.1, with modality availability determined by each scenario’s data composition.

Table 4. Baseline model configurations and hyperparameters

Model	Input Type	Key Layers / Parameters	Activation	Optimizer	Learning Rate	Dropout
Logistic Regression	Statistical features (d = 20)	Linear (L2 = 1×10^{-4}, liblinear solver)	Sigmoid	–	–	–
SVM (RBF)	Statistical features (d = 20)	Kernel = RBF, C = 1.0, γ = 0.1	–	–	–	–
CNN	128×128 log-mel spectrogram	Conv(32, 3×3) → Conv(64, 3×3) → GAP → Dense(128)	ReLU / Softmax	Adam	1×10^{-4}	0.5
Bi-LSTM	200×6 time-series	Bi-LSTM(64) → Dense(64)	Tanh / Softmax	Adam	1×10^{-4}	0.3
Multimodal Fusion Network	CNN + Bi-LSTM + Static features	Concat → Dense(128) → Dropout(0.3) → Dense(2)	ReLU / Softmax	Adam	1×10^{-4}	0.3

4.2. Training Configuration

All models were trained end-to-end using a unified optimization strategy to ensure fair comparison across unimodal, multimodal, and cross-domain evaluation scenarios. Unless stated otherwise, the same training configuration was applied to all experimental setups described in Section 4.1. The complete training configuration is summarized in Table 5.

Table 5. Training setup.

Parameter	Configuration	Justification
Epochs	120 (max)	Provides sufficient convergence time for multimodal feature learning without overfitting
Batch size	32 samples per update	Balances gradient stability and GPU memory constraints for multimodal data fusion
Optimizer	Adam (Kingma & Ba, 2015)	Adaptive learning suitable for mixed-modality data with sparse gradients
Initial learning rate	1×10^{-4}	Standard starting rate for CNN–LSTM-based multimodal learning
Learning rate schedule	Cosine annealing with warm restarts every 20 epochs	Encourages periodic exploration of new minima and stabilizes convergence
Beta$_1$, Beta$_2$	(0.9, 0.999)	Default Adam momentum parameters
Epsilon	1×10^{-7}	Numerical stability in variance updates
Weight decay	1×10^{-5}	Regularization to reduce overfitting in fusion layers
Dropout rates	0.2–0.3 (layer-dependent)	Enhances generalization during multimodal fusion
Gradient clipping	Global norm capped at 5.0	Prevents gradient explosion in recurrent layers
Early stopping	Patience = 10 epochs	Terminates training when validation loss plateaus
Model checkpointing	Best validation loss	Ensures optimal model state is used for evaluation
Cross-validation	5-fold stratified (per dataset source)	Robust performance estimation across heterogeneous data

Training was performed using the Adam optimizer with an initial learning rate of 1×10^{-4} , which provided stable convergence for both convolutional and recurrent components. To improve convergence stability and avoid poor local minima, a cosine annealing learning rate schedule with warm restarts every 20 epochs was employed. The learning rate at epoch t follows:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{\pi T_{\mathrm{cur}}}{T_{\max}}\right)\right) \tag{6}$$

where $\eta_{\max} = 1 \times 10^{-4}$, $\eta_{\min} = 1 \times 10^{-6}$, and T_{cur} resets every 20 epochs.

Models were trained for a maximum of 120 epochs with a batch size of 32, and early stopping was applied when the validation loss failed to improve for 10 consecutive epochs. Gradient clipping was applied with a global norm threshold of 5.0 to prevent instability during recurrent optimization. Dropout regularization (0.2–0.3, depending on layer) and L2 weight decay (1×10^{-5}) were used to mitigate overfitting in the fusion layers. Training typically converged after approximately 85 epochs. For reproducibility, all random seeds were fixed to 42, and the model state corresponding to the lowest validation loss was saved and used for subsequent evaluation.

4.3. Evaluation Metrics and Statistical Analysis

Model performance was evaluated using standard classification metrics commonly adopted in clinical decision support studies: Accuracy, Sensitivity (Recall), Specificity, Precision, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). Sensitivity was emphasized because minimizing missed pneumonia cases is clinically important, particularly in early risk-monitoring scenarios. To quantify uncertainty in model performance, 95% confidence intervals for AUC were estimated using bootstrap resampling with 1,000 iterations. Statistical comparisons between evaluation scenarios were conducted using the DeLong test for correlated ROC curves, enabling principled comparison of AUC values across models and data regimes. When multiple pairwise comparisons were performed, the Bonferroni correction was applied to control for Type I error.

In multi-source and cross-domain experiments, where class imbalance and domain heterogeneity were more pronounced, the Matthews Correlation Coefficient (MCC) was additionally reported as a balanced performance measure that remains informative under skewed class distributions. The specific combination of evaluation metrics reported for each experimental scenario is summarized in Table 6, ensuring transparent alignment between evaluation objectives and reported results.

Table 6. Evaluation protocol and metrics per experimental scenario

Experiment Type	Dataset Split	Cross-Validation	Evaluation Metrics	Purpose
Public-only	80/10/10	5-fold	Accuracy, Precision, Recall, F1, AUC	Baseline validation on public datasets
Private-only (Local)	70/15/15	5-fold	Accuracy, Precision, Recall, F1, AUC	Local generalization under realistic deployment conditions
Cross-domain (Public → Private)	Train public / Test private	–	AUC, F1	Quantification of domain shift and transfer performance
Combined (Domain-adapted)	Stratified merge (dataset-level)	–	Accuracy, F1, MCC	End-to-end robustness under heterogeneous data sources

4.4. Implementation and Deployment Simulation

Cloud-side training was performed on a workstation equipped with an NVIDIA RTX 3090 GPU. Edge-side inference was evaluated on an NVIDIA Jetson Xavier device to assess feasibility under constrained computational resources. All models were implemented in TensorFlow 2.12.

The edge–cloud interaction was simulated using a Wi-Fi (802.11ac) network with an average uplink bandwidth of 50 Mbps and latency ranging from 25–40 ms. Network latency was modeled as a Gaussian variable (μ=30 ms, σ=10 ms). To minimize bandwidth usage, only compressed multimodal feature vectors (approximately 150 KB per minute) were transmitted to the cloud. Edge inference latency was estimated using TensorFlow Lite profiling for models of comparable size (~2.7M parameters). Energy consumption was estimated using NVIDIA's tegrastats tool and the Jetson Energy Estimator, yielding an average draw of approximately 3.2 W during active inference. These estimates are consistent with low-power IoT deployment requirements. Cloud-side processing, including model retraining and SHAP-based interpretability analysis, was performed asynchronously and did not contribute to real-time latency.

4.5. Reproducibility and Transparency

To support reproducibility, all experiments were conducted with fixed random seeds, and training logs, learning rate schedules, and checkpointed weights were archived. Source code for model training and evaluation will be released upon publication, while a de-identified feature-level dataset derived from the local clinical study will be available upon reasonable request, in accordance with ethical approval and data use agreements. This experimental setup provides a rigorous and transparent evaluation of the proposed edge–cloud multimodal framework under controlled, clinically realistic, and cross-domain conditions. By explicitly separating performance assessment from robustness analysis, the protocol supports a nuanced interpretation of results presented in Section 5.

5. Results and Discussion

This section presents the experimental results of the proposed edge–cloud multimodal framework and discusses their implications in terms of diagnostic performance, robustness, deployment feasibility, and interpretability. Results are organized to progressively evaluate predictive accuracy, domain robustness, architectural contributions, system-level feasibility, and clinical transparency.

5.1. Overall System Performance Across Evaluation Scenarios

Table 7 summarizes the predictive performance of the proposed edge–cloud multimodal framework across four evaluation scenarios: public-only, private-only (local), combined multi-source, and cross-domain transfer. These scenarios were designed to separately assess unimodal baseline behavior, clinically realistic local performance, robustness under heterogeneous training conditions, and sensitivity to domain shift, rather than to provide a single estimate of population-level diagnostic accuracy. In the combined multi-source evaluation, the framework achieved its highest performance, with an AUC of 0.95 (95% CI: 0.93–0.96), sensitivity of 94.3%, and specificity of 90.1%. This result indicates a consistent, balanced multimodal diagnostic signal when heterogeneous data sources are jointly leveraged in a controlled, robust setting. Importantly, this scenario reflects representation learning and domain robustness rather than a realistic deployment condition, as public and local datasets are pooled at the dataset level.

Table 7. Performance under three dataset setups

Scenario	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score	AUC (95% CI)
Public-only (MIMIC & Coswara)	82.4 ± 3.1	80.2 ± 4.0	84.1 ± 2.7	0.81 ± 0.03	0.88 (0.85–0.90)
Private-only (Local)	79.7 ± 4.8	85.0 ± 5.0	77.3 ± 4.2	0.80 ± 0.04	0.86 (0.81–0.90)
Combined (Public + Local)	92.6 ± 2.0	94.3 ± 1.6	90.1 ± 2.3	0.92 ± 0.02	0.95 (0.93–0.96)
Public → Private	76.5 ± 3.9	79.1 ± 4.5	74.0 ± 4.1	0.77 ± 0.03	0.83 (0.79–0.87)
Private → Public	74.8 ± 4.5	78.6 ± 5.2	71.4 ± 3.8	0.76 ± 0.04	0.82 (0.78–0.86)

When evaluated exclusively on the local clinical cohort, the model achieved an AUC of 0.86 (95% CI: 0.81–0.90), with sensitivity of 85.0% and specificity of 77.3%. This private-only evaluation represents the most clinically realistic scenario, as both training and inference data originate from the same acquisition environment, sensor configuration, and population. Due to the limited number of pneumonia-positive cases in the local cohort, these results should be interpreted as evidence of consistent multimodal signal presence and system feasibility, rather than definitive clinical diagnostic accuracy. Cross-domain transfer experiments (public → private and private → public) exhibited a measurable performance drop, with AUC values of 0.83 and 0.82, respectively. This degradation reflects systematic differences in population characteristics, sensor properties, and acquisition conditions between public repositories and locally collected wearable data. Rather than indicating model failure, this behavior serves as a diagnostic indicator of domain shift and underscores the need for local adaptation for reliable deployment in real-world settings.

The results demonstrate that while public datasets provide useful pretraining signals for modality-specific encoders, stable performance in realistic clinical environments requires synchronized local data and domain-aware adaptation. Taken together, these findings support the central claim of this study: that multimodal fusion within an edge–cloud framework can produce consistent early risk signals under heterogeneous conditions, and that synchronized local data and domain-aware adaptation are essential for achieving stable behavior in realistic clinical settings.

5.2. Robustness to Domain Shift, Data Diversity, and Local Adaptation

When models trained on public datasets were evaluated on locally collected clinical data, performance decreased noticeably, with AUC dropping from 0.95 (combined multi-source evaluation) to 0.83 under cross-domain testing (Public → Private), as shown in Table 9. A similar degradation was observed in the reverse direction (Private → Public), confirming that domain shift affects both training-to-deployment and deployment-to-benchmark transfer scenarios. This degradation reflects differences in population demographics, sensor characteristics, and acquisition environments between public and local data sources. Public datasets such as MIMIC-III and Coswara are predominantly derived from Western populations and controlled recording environments, whereas the local dataset reflects real-world conditions in Nigerian clinical settings, including differences in baseline physiological ranges, ambient noise, and wearable sensor placement. These factors introduce systematic distributional shifts that cannot be resolved through naïve data pooling alone.

Importantly, the observed performance drop should not be interpreted as a failure of the model, but rather as a diagnostic indicator of domain bias and limited generalization inherent in single-source training. By incorporating locally collected data into the combined training regime and applying domain adaptation strategies (Section 3.3), the model partially mitigated this shift, recovering performance to an AUC of 0.95. This demonstrates that local adaptation is essential for achieving stable performance in real-world deployment environments. Beyond methodological implications, these findings highlight a broader equity issue in clinical AI. Models trained exclusively on large public datasets may not generalize reliably to underrepresented populations, even when overall accuracy appears high in benchmark settings. The results of this study therefore underscore the necessity of local data integration—not as a refinement step, but as a core design principle for responsible, globally deployable healthcare AI systems.

5.3. Comparison with Prior Studies

To contextualize performance, Table 8 compares the proposed system with representative state-of-the-art unimodal methods, all evaluated on the same public datasets. While prior studies focused on single-modality learning, the proposed framework integrates multimodal fusion and domain-adaptive fine-tuning on a clinically paired dataset.

Table 8. Comparison with prior studies

Study	Methodology	Dataset Used	Accuracy (%)	AUC
Brown et al. [4]	CNN (Audio)	Coswara	84.2	0.88
Sharma et al. [28]	VGGish + LSTM (Audio)	Coswara	86.5	0.90
Liu et al. [17]	Bi-LSTM (Vitals)	MIMIC	88.1	0.91
Proposed (Ours)	CNN–BiLSTM–Static (Fusion)	Local Clinical*	92.6 ± 2.0	0.95

*Sub-networks were pretrained on public datasets; fusion and evaluation were performed on synchronized local clinical data.

The observed performance gain arises from three complementary factors: (i) multimodal fusion that captures both physiological and acoustic manifestations of respiratory disease, (ii) explicit domain adaptation that reduces distributional mismatch between public and local data, and (iii) interpretability-aware modeling that aligns predictions with known clinical markers. Importantly, this comparison highlights methodological advancement rather than direct dataset-level superiority, as the proposed model is evaluated under more realistic clinical conditions. These results suggest that performance improvements are not solely attributable to model architecture, but to the integration of learning, adaptation, and deployment design. In this sense, the proposed framework extends prior work by addressing not only detection accuracy, but also robustness and clinical viability.

5.4. Ablation Study: Contribution of Each Modality

Table 9 presents the ablation results evaluating the contribution of each modality under controlled removal settings.

Table 9. Ablation results

Configuration	Modalities Used	Dataset Source	Accuracy (%)	AUC (95% CI)
Model 1	Physiological	Public + Local	87.8 ± 3.1	0.90 (0.88–0.92)
Model 2	Audio	Public + Local	85.2 ± 2.7	0.88 (0.85–0.90)
Model 3	Static	Local	79.5 ± 3.9	0.83 (0.80–0.86)
Model 4	Physio + Audio	Local	91.4 ± 2.4	0.94 (0.91–0.95)
Model 5	Physio + Static	Local	89.7 ± 2.8	0.92 (0.89–0.94)
Model 6	Audio + Static	Local	86.9 ± 2.9	0.89 (0.87–0.91)
Model 7	Full (All)	Local	92.6 ± 2.0	0.95 (0.93–0.96)

The results confirm that multimodal synergy, rather than model capacity alone, drives performance gains. Physiological and acoustic features consistently form the dominant diagnostic pair, while static variables act as contextual enhancers that improve calibration but contribute less standalone discriminative power. Notably, the performance gap between unimodal and multimodal configurations (AUC 0.90 → 0.95) demonstrates that the system's strength lies in cross-modal complementarity, not in overfitting to a single signal source. This supports the design choice to prioritize synchronized multimodal fusion on the local clinical dataset.

5.5. Edge–Cloud Adaptation and Deployment Feasibility

The adaptive sensitivity mechanism described in Section 3.5 was evaluated through deployment simulation. Figure 4 illustrates the feedback loop, while Table 10 summarizes latency and energy consumption under different configurations.

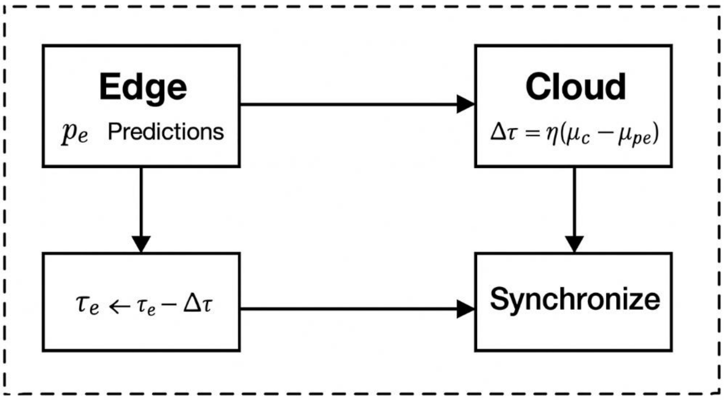

Edge
p_e Predictions
Cloud
$\Delta\tau = \eta(\mu_c - \mu_{pe})$
$\tau_e \leftarrow \tau_e - \Delta\tau$
Synchronize

Figure 4. Adaptive Sensitivity Update Loop

Table 10. Edge–cloud simulation results

Mode	Edge Load (%)	Latency (ms)	Power (W)	Comment
Edge-only	100	135	3.2	Real-time, higher energy
Cloud-only	0	250	<1	Network-dependent
Hybrid	60	160	2.4	Balanced, optimal

Hybrid processing achieves a favorable trade-off between responsiveness and efficiency, remaining well below the 1-second latency threshold for continuous monitoring. These results indicate that adaptive edge–cloud coordination is not only feasible but also practically beneficial, enabling real-time monitoring without excessive energy or bandwidth costs. While these findings are based on simulation rather than longitudinal deployment, they provide evidence that the proposed architecture can operate within the constraints of wearable and low-resource clinical environments.

5.6. Interpretability and Clinical Consistency Analysis

To assess whether the proposed multimodal framework produces clinically meaningful and transparent predictions, model interpretability was analyzed using SHapley Additive exPlanations (SHAP). SHAP enables both global and local explanations by quantifying each feature's contribution to model outputs [37]–[39], providing insight into how multimodal signals jointly influence pneumonia risk estimation.

Interpretability analysis was conducted using Gradient SHAP (SHAP v0.43) on the trained multimodal model. For each evaluation fold, SHAP values were computed for 1,000 randomly sampled instances, equally distributed across the MIMIC, Coswara, and local wearable datasets, to ensure balanced representation across sources. Feature values were normalized per source prior to explanation to avoid scale-induced bias.

5.6.1. Global Feature Importance

Table 11 summarizes global feature importance measured as mean absolute SHAP values aggregated across all samples, while Figure 5 presents the corresponding beeswarm plot..

Table 11. Global feature importance

Rank	Feature	Source Modality	Mean	Std.	Interpretation
1	Oxygen Saturation (SpO_2)	Wearable / MIMIC	0.162	0.041	Lower SpO_2 strongly increases pneumonia risk
2	Cough Energy (dB)	Audio (Coswara)	0.143	0.035	Strong cough intensity correlates with infection
3	Respiratory Rate	Wearable / MIMIC	0.129	0.037	Elevated rate distinguishes pneumonia
4	Temperature	Wearable	0.111	0.030	Fever contributes to higher risk prediction
5	$MFCC_{13}$ (Spectral Flatness)	Audio	0.094	0.028	Voice spectral distortion signals respiratory distress
6	Heart Rate Variability	Wearable	0.087	0.026	Reduced HRV indicates physiological stress
7	Age	Static	0.076	0.021	Older patients more likely to be high-risk
8	Cough Duration (ms)	Audio	0.062	0.019	Prolonged cough associated with pneumonia
9	Respiration-Temperature Interaction	Fused	0.057	0.018	Nonlinear interaction between features
10	Device Signal Quality Index	Static	0.049	0.017	Low SQI slightly lowers confidence in detection

Across all datasets, oxygen saturation (SpO_2), cough energy, and respiratory rate consistently emerged as the most influential predictors of pneumonia risk. These features are well-established clinical indicators of respiratory compromise, suggesting that the model’s decision-making process is grounded in physiologically meaningful biomarkers rather than spurious correlations.

Importantly, acoustic features derived from cough signals (e.g., cough energy, $MFCC_{13}$, and cough duration) were complementary to physiological vitals. This supports the design choice of multimodal fusion, as audio-based biomarkers captured respiratory distress patterns that were not fully represented in vital signs alone. Static demographic features (age, device signal quality index) contributed secondary but stabilizing effects, primarily modulating risk estimates in borderline cases. The presence of a fused interaction term (respiration–temperature interaction) among the top ten features further indicates that the model learns nonlinear relationships across modalities, reflecting the multifactorial nature of pneumonia symptoms.

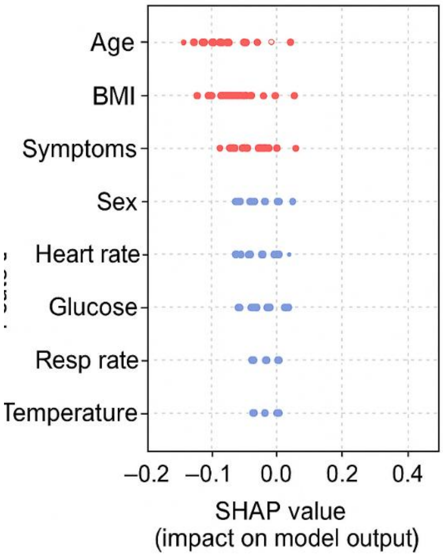


Figure 5. SHAP summary plot (beeswarm)

5.6.2. Local Case Explanations

To examine individual decision behavior, local SHAP explanations were analyzed for representative true-positive, false-negative, and false-positive cases. Table 12 summarizes the relative contribution of each modality, while Figure 6 illustrates a representative force plot for a pneumonia-positive case.

Table 12. Feature-level SHAP contribution by modality

Modality	Top SHAP Features	Contribution (%)
Physiological	SpO$_2$, respiratory rate	41.3
Audio	Cough energy, MFCC$_{13}$	36.7
Static	Age, temperature	22.0

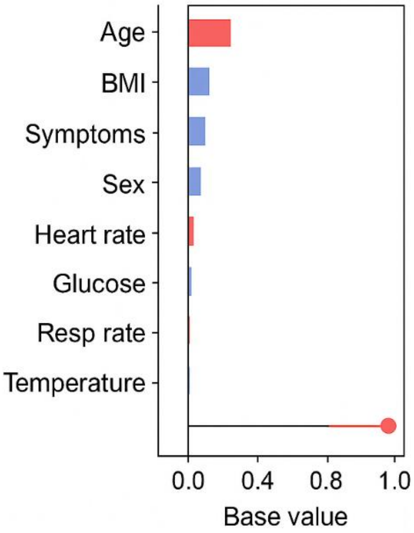


Figure 6. SHAP force plot (local case)

Local explanations reveal that high-risk predictions are typically driven by a convergence of abnormal vitals (low SpO_2, elevated respiratory rate) and strong acoustic markers (high cough energy), while static features amplify or attenuate the final risk score. In contrast, false negatives were often associated with weak or noisy acoustic signals combined with mild physiological abnormalities, indicating that sensor quality and transient conditions remain important sources of uncertainty. False positive cases commonly involved temporary fever or irregular breathing patterns; SHAP correctly highlighted these features as dominant contributors, demonstrating that the model's errors are traceable and clinically interpretable rather than opaque.

5.6.3. Cross-Dataset Consistency of Explanations

To evaluate whether interpretability remains stable across heterogeneous sources, SHAP importance values were averaged per dataset (Table 15).

Table 13. Cross-dataset SHAP comparison

Feature	MIMIC	Coswara	Local	Comment
SpO_2	0.175	—	0.151	Universal pneumonia indicator
Respiratory rate	0.142	—	0.133	Consistent across sensors
Cough energy	—	0.163	0.095	More discriminative in Coswara
Temperature	0.118	—	0.107	Slightly higher importance locally
HRV	—	—	0.092	Unique to wearable data

While absolute SHAP magnitudes varied across datasets, the relative ordering of clinically relevant features remained largely consistent. Physiological features dominated MIMIC-derived explanations, acoustic features were more prominent in Coswara, and wearable-specific signals (HRV, SQI) emerged only in local data. This pattern confirms that the model adapts its reasoning to available modalities without altering its core clinical logic.

5.6.4. Summary of Interpretability Findings

- Overall, the SHAP analysis provides converging evidence that the proposed framework is:
- Clinically grounded, as dominant features align with established pneumonia biomarkers.
- Multimodally coherent, with different modalities contributing complementary evidence rather than redundant signals.
- Robust across domains, as interpretability patterns remain stable despite population and device differences.
- Transparent and auditable, enabling clinicians to trace both correct and incorrect predictions to specific physiological or acoustic factors.

These findings indicate that interpretability is not merely an auxiliary visualization step, but an integral component of the edge–cloud framework, supporting trust, debugging, and eventual clinical adoption.

6. Conclusion

This study investigated the feasibility of an edge–cloud integrated multimodal framework for early pneumonia risk monitoring using wearable sensors. By combining physiological signals, cough acoustics, and static clinical attributes within a domain-adaptive learning architecture, the proposed system demonstrated consistent diagnostic signal across public datasets, locally collected clinical data, and cross-domain evaluation scenarios. Rather than optimizing for a single dataset, the framework was explicitly designed to address data heterogeneity, deployment constraints, and clinical transparency, which are critical for real-world use in resource-limited settings. A key contribution of this work is the demonstration that multimodal fusion, when coupled with domain adaptation and hierarchical computation, can yield robust, interpretable risk estimates without continuous cloud connectivity. The edge–cloud architecture enables low-latency local triage while preserving the ability to perform computationally intensive fusion and explanation in the cloud. Interpretability analysis using SHAP further showed that model predictions are driven by clinically meaningful biomarkers, such as oxygen saturation, respiratory rate, and cough energy, supporting alignment between the model's reasoning and established clinical knowledge.

From a practical perspective, the deployment in Nigerian primary healthcare centers illustrates how AI-based monitoring systems can augment limited clinical capacity. Continuous, non-invasive monitoring offers a complementary layer of early warning in environments where diagnostic resources and staffing are constrained. Importantly, the system does not aim to replace clinical judgment, but to provide decision support that is transparent, auditable, and adaptable to local conditions. Despite these encouraging findings, several limitations should be acknowledged. The clinical cohort size remains modest, and although the total monitoring duration is substantial, larger, more diverse populations are needed to validate generalizability. In addition, interpretability was evaluated post hoc and not yet integrated into real-time edge interfaces, which may limit immediate clinical usability. The adaptive edge mechanism was evaluated in controlled simulations rather than in continuous long-term deployment, and future studies should assess its stability under extended real-world operation.

Future work will focus on expanding multi-site clinical studies across diverse ecological and demographic settings, integrating self-supervised learning to reduce dependence on labeled data, and developing lightweight on-device interpretability mechanisms. These directions are essential steps toward translating multimodal AI monitoring systems into sustainable, clinically accepted tools for global respiratory health..

Author Contributions: Conceptualization: J.B.O. and E.O.I..; Methodology: J.B.O.; Software: J.B.O..; Validation: J.B.O. and E.O.I..; Formal analysis: J.B.O.; Investigation: J.B.O.; Resources: E.O.I.; Data curation: J.OB.O..; Writing—original draft preparation: J.B.O.; Writing—review and editing: E.O.I; Visualization: J.B.O.; Supervision: E.O.I.; Project administration: J.B.O.; funding acquisition: E.O.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The source code used to train the models and perform the analysis presented in this study is publicly available in a GitHub repository (detail will be release on request): https://github.com/oluwagbemijb/Edge-Cloud-Enabled-Multimodal-Framework-for-Real-Time-Pneumonia-Detection-Using-Wearable-Sensor. The raw physiological and acoustic data collected from the clinical study are not publicly available due to patient privacy restrictions outlined in our Institutional Review Board approval and the informed consent agreement. To facilitate reproducibility and verification of our findings, a de-identified minimal dataset is available from the corresponding author upon reasonable request. This dataset contains the final feature vectors used to train the models and the corresponding ground-truth labels. Access will be granted to qualified researchers upon completion of a data use agreement intended to ensure that the data is used for non-commercial research purposes only.

Acknowledgments: The authors acknowledge the use of AI-assisted tools for language refinement and proofreading of this manuscript. All scientific content, methodological decisions, data analysis, and interpretations remain the full responsibility of the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] World Health Organization (WHO), “Pneumonia,” *WHO Fact Sheets*, 2021. https://www.who.int/health-topics/pneumonia#tab=tab_1

[2] D. Bhattacharya *et al.*, “Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection,” *Sci. Data*, vol. 10, no. 1, p. 397, Jun. 2023, doi: 10.1038/s41597-023-02266-0.

[3] P. O. Adebayo, F. Basaky, and E. Osaghae, “Leveraging Variational Quantum-Classical Algorithms for Enhanced Lung Cancer Prediction,” *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 307–323, Dec. 2024, doi: 10.62411/jcta.10424.

[4] C. Brown *et al.*, “Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2020, pp. 3474–3484. doi: 10.1145/3394486.3412865.

[5] A. B. Rashid, J. Asma, K. Barua, and D. Das, “An Enhanced Deep Learning Framework for Pneumonia Detection in Chest X-rays,” *SN Comput. Sci.*, vol. 6, no. 5, p. 472, May 2025, doi: 10.1007/s42979-025-04017-x.

[6] Z. Jia, H. Huth, W. Q. Teoh, S. Xu, B. Wood, and Z. T. H. Tse, "State of the Art Review of Wearable Devices for Respiratory Monitoring," *IEEE Access*, vol. 13, pp. 18178–18190, 2025, doi: 10.1109/ACCESS.2025.3529437.

[7] Y. Li *et al.*, "Integrated wearable smart sensor system for real-time multi-parameter respiration health monitoring," *Cell Reports Phys. Sci.*, vol. 4, no. 1, p. 101191, Jan. 2023, doi: 10.1016/j.xcrp.2022.101191.

[8] M. B. Teferi and L. A. Akinyemi, "Deep Learning-Based Cross-Cancer Morphological Analysis: Identifying Histopathological Patterns in Breast and Lung Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 235–248, Oct. 2024, doi: 10.62411/faith.3048-3719-36.

[9] D. Lin, Y. F. Ji, J. A. A. McArt, and J. Li, "SynLS: A novel diffusion-transformer framework for generating high-quality wearable sensor time series data to enhance health monitoring," *biorxiv*. May 15, 2025. doi: 10.1101/2025.05.11.653212.

[10] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Brief. Bioinform.*, vol. 23, no. 2, Mar. 2022, doi: 10.1093/bib/bbab569.

[11] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," *npj Digit. Med.*, vol. 5, no. 1, p. 171, Nov. 2022, doi: 10.1038/s41746-022-00712-8.

[12] N. Fadul, M. F. Alaskar, K. B. Jillahi, and D. B. El-Khaled, "Generative AI in Healthcare: An Analytical Review of Models, Clinical Applications, and Decision-Support Implications," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 587–615, Dec. 2025, doi: 10.62411/faith.3048-3719-298.

[13] J. B. Oluwagbemi, A. E. Mesioye, and R. S. Akinbo, "Depress-HybridNet: A Linguistic-Behavioral Hybrid Framework for Early and Accurate Depression Detection on Social Media," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 3, pp. 432–444, Sep. 2025, doi: 10.62411/faith.3048-3719-266.

[14] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *Sci. Rep.*, vol. 12, no. 1, p. 17981, Oct. 2022, doi: 10.1038/s41598-022-22514-4.

[15] M. Salvi *et al.*, "Multi-modality approaches for medical support systems: A systematic review of the last decade," *Inf. Fusion*, vol. 103, p. 102134, Mar. 2024, doi: 10.1016/j.inffus.2023.102134.

[16] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," *Nat. Med.*, vol. 28, no. 9, pp. 1773–1784, Sep. 2022, doi: 10.1038/s41591-022-01981-2.

[17] M. Liu, S. Suh, J. F. Vargas, B. Zhou, A. Grünerbl, and P. Lukowicz, "A Wearable Multi-modal Edge-Computing System for Real-Time Kitchen Activity Recognition," in *Communications in Computer and Information Science*, 2025, pp. 132–145. doi: 10.1007/978-981-97-9003-6_9.

[18] M. D. Toruner *et al.*, "Artificial Intelligence-Driven Wireless Sensing for Health Management," *Bioengineering*, vol. 12, no. 3, p. 244, Feb. 2025, doi: 10.3390/bioengineering12030244.

[19] M. C. Kelvin-Agwu, B. O. Tomoh, and A. Y. Forkuo, "Development of AI-Assisted Wearable Devices for Early Detection of Respiratory Diseases," *J. Front. Multidiscip. Res.*, vol. 6, no. 1, pp. 64–72, 2025, doi: 10.54660/.IJFMR.2025.6.1.64-72.

[20] H. Ren *et al.*, "CheXMed: A multimodal learning algorithm for pneumonia detection in the elderly," *Inf. Sci. (Ny).*, vol. 654, p. 119854, Jan. 2024, doi: 10.1016/j.ins.2023.119854.

[21] H. Wang *et al.*, "Developing and Validation of a Multimodal-Based Machine Learning Model for Diagnosis of Usual Interstitial Pneumonia," *Chest*, Sep. 2025, doi: 10.1016/j.chest.2025.08.034.

[22] M. S. Sunarjo, H. Gan, and D. R. I. M. Setiadi, "High-Performance Convolutional Neural Network Model to Identify COVID-19 in Medical Images," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 19–30, Aug. 2023, doi: 10.33633/jcta.v1i1.8936.

[23] K. Pyar, "Segmentation Performance Analysis of Transfer Learning Models on X-Ray Pneumonia Images," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 64–74, Jun. 2024, doi: 10.62411/faith.2024-10.

[24] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 396–406, Mar. 2024, doi: 10.62411/jcta.10106.

[25] J. Li *et al.*, "Diagnostic assistance method for RR-TB/MDR-TB patients under treatment based on CNN-LSTM," *Sci. Rep.*, vol. 15, no. 1, p. 38035, Oct. 2025, doi: 10.1038/s41598-025-21955-x.

[26] J. Colin and N. Surantha, "Interpretable Deep Learning for Pneumonia Detection Using Chest X-Ray Images," *Information*, vol. 16, no. 1, p. 53, Jan. 2025, doi: 10.3390/info16010053.

[27] J. Xu and Y. Wang, "FMT:A Multimodal Pneumonia Detection Model Based on Stacking MOE Framework," in *2025 8th International Conference on Information and Computer Technologies (ICICT)*, Mar. 2025, pp. 517–521. doi: 10.1109/ICICT64582.2025.00087.

[28] S. Sharma and K. Guleria, "A Deep Learning based model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks," *Procedia Comput. Sci.*, vol. 218, pp. 357–366, 2023, doi: 10.1016/j.procs.2023.01.018.

[29] A. Rancea, I. Anghel, and T. Cioara, "Edge Computing in Healthcare: Innovations, Opportunities, and Challenges," *Futur. Internet*, vol. 16, no. 9, p. 329, Sep. 2024, doi: 10.3390/fi16090329.

[30] A. Ali, "Exploring the Transformative Potential of Technology in Overcoming Educational Disparities," *Int. J. Multidiscip. Sci. Arts*, vol. 2, no. 1, Jul. 2023, doi: 10.47709/ijmdsa.v2i1.2559.

[31] I. Sanches *et al.*, "MIMIC-BP: A curated dataset for blood pressure estimation," *Sci. Data*, vol. 11, no. 1, p. 1233, Nov. 2024, doi: 10.1038/s41597-024-04041-1.

[32] N. Sharma *et al.*, "Coswara -- A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," *arXiv*. Aug. 11, 2020. doi: 10.21437/Interspeech.2020-2768.

[33] Y. Song, Z. Liu, J. Wang, R. Tang, G. Duan, and J. Tan, "Multiscale Adversarial and Weighted Gradient Domain Adaptive Network for Data Scarcity Surface Defect Detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021, doi: 10.1109/TIM.2021.3096284.

[34] H. Lim, B. Kim, J. Choo, and S. Choi, "TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation," *arXiv*. Feb. 18, 2023. [Online]. Available: http://arxiv.org/abs/2302.05155

[35] B. Isgor and M. Koklu, "Lightweight Hybrid Model for Bone Fracture Detection Using MobileNetV2 Feature Extraction and Ensemble Learning," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 3, pp. 521–533, Dec. 2025, doi: 10.62411/faith.3048-3719-284.

[36] A. O. Eboka *et al.*, "Resolving Data Imbalance Using a Bi-Directional Long-Short Term Memory for Enhanced Diabetes Mellitus Detection," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 1, pp. 95–109, May 2025, doi: 10.62411/faith.3048-3719-73.

[37] T. R. Noviandy, G. M. Idroes, and I. Hardi, "An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 84–95, Aug. 2024, doi: 10.62411/faith.2024-16.

[38] J. B. Oluwagbemi, O. V. Oyetayo, and E. O. Ibam, "SysFungiNet: A Multi-Omics Data Fusion Framework with Explainable AI for Bioactive Prioritization," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 661–679, Jan. 2026, doi: 10.62411/faith.3048-3719-304.

[39] A. Hamza, W. Hussain, H. Iftikhar, A. Ahmad, and A. M. Shamim, "Evaluating Open-Source Machine Learning Project Quality Using SMOTE-Enhanced and Explainable ML/DL Models," *J. Comput. Theor. Appl.*, vol. 3, no. 2, pp. 206–222, Nov. 2025, doi: 10.62411/jcta.14793.