International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(22s) 2026 pp. 133-152
https://doi.org/10.70917/ijcisim-2026-5422
Article

ADAPTIVE CONTEXT-AWARE MULTIMODAL TRANSFORMER FRAMEWORK FOR ROBUST RESPIRATORY DISEASE DIAGNOSIS

Ms. R.SARADHA[1], Dr.M.SUNDARA RAJAN[2]

[1]Research Scholar (Full-Time), PG & Research, Department of Computer Science, Government Arts College, (Autonomous), Nandanam, Chennai 600 035. saradha303@gmail.com
[2]Associate Professor, PG & Research, Department of Computer Science, Government Arts College, (Autonomous), Nandanam, Chennai 600 035. drmsrajan23@yahoo.com

Abstract: Early detection and accurate diagnosis of respiratory diseases are remaining critical for the timely clinical interference and treatment. Many of the conventional diagnostic systems are relying on the unimodal datasets or static multimodal fusion strategies that tends in limiting the ability in capturing the respiratory diseases. Despite the success of recent multimodal learning approaches, where the practical limitations are remaining unresolved. This study proposes an Adaptive Context-Aware Multimodal Transformer Network (AMCT-Net) that is combining the chest X-ray images, respiratory acoustic signals, and structured clinical parameters for the respiratory disease diagnosis. The proposed architecture is employing three modality-specific encoders: a convolutional neural network that tends to extract the radiographic features, where a spectrogram-based audio transformer that is capturing the respiratory sound patterns, and a tabular transformer that is processing clinical variables. The extracted representations are passing through the Adaptive Modality Weighting mechanism that is assigning the dynamic importance scores to each of the modality. A context-aware multimodal transformer fusion module is then modelling the cross-modal dependencies in generating the diagnostic representations. The fused features are subsequently processed through the classification layer that tends to predict the respiratory disease categories. The experimental evaluation is showing that the proposed Adaptive Context-Aware Multimodal Transformer Network (AMCT-Net) is achieving a superior performance compared with the conventional multimodal diagnostic frameworks. The model is reaching 88.72% classification accuracy, 88.10% precision, 88.45% recall, and 88.30% F1-score, which significantly is exceeding the performance of conventional architectures that includes MDT, E-RespiNet, Med-MLLM, FATE, DREAM-OSA, CAMAF, and RMT with the mask attention.
Keywords: Multimodal learning, transformer architecture, respiratory disease diagnosis, adaptive fusion, medical artificial intelligence

1. Introduction

Respiratory diseases are representing the major public health concern that is significantly contributing to the global morbidity/mortality. Early diagnosis is playing an essential role in the reduction of disease progression and improving survival rates. However, where the diagnostic process often depends heavily on clinical expertise and manual interpretation, which may be introducing variability and diagnostic delays. Deep learning (DL) models, which in particular uses transformer-based architectures, which have shown significant success in analyzing the complex medical datasets and extracting meaningful patterns for the disease classification. These technologies are supporting the development of intelligent healthcare systems that assist clinicians in detecting respiratory abnormalities with the improved accuracy and efficiency [1]. Several studies have shown that the combination of imaging features,

respiratory sounds, and clinical parameters significantly enhances diagnostic performance compared with the unimodal models [2]. In addition, transformer architectures have gained attention in medical analysis due to their ability in capturing long-range dependencies and contextual relationships within large datasets [3].

Recent developments in the multimodal artificial intelligence systems further are showing the importance of combining the structured and unstructured medical information to is improving the diagnostic reliability and it is supporting the precision medicine initiatives [4]. Further, where the combination of machine learning models with the clinical decision is supporting the systems are allowing the continuous monitoring of respiratory disorders through the automated analysis of patient data [5].

Despite these advancements, where various challenges are remaining in the development of reliable multimodal diagnostic systems [6]. The imbalance between the modalities also is presenting a significant challenge because the imaging data often dominate the learning process while clinical or acoustic features contribute less to the final decision [7]. In addition, many of the DL models require large labeled datasets, which are remaining limited in healthcare environments due to the annotation complexity. Another challenge relates to the computational cost associated with the large transformer architectures, which tends in limiting the application in real-time clinical systems or resource-constrained healthcare settings [8]. These limitations are showing the need for the adaptive learning mechanisms that effectively manage heterogeneous medical data while it is maintaining computational efficiency and robustness in real-world clinical scenarios [9]. Several studies also are emphasizing the difficulty of capturing the cross-modal relationships within multimodal datasets, which may result in a suboptimal diagnostic representation when the conventional fusion strategies are applied [10].

The fundamental problem addressed in this research lies in the limited adaptability of conventional multimodal diagnostic models when the handling heterogeneous and incomplete medical data. Many of the current transformer-based frameworks employ static feature fusion strategies that is assigning the equal importance to all modalities regardless of their relevance to individual patient conditions. As a result, these models may fail in capturing the dynamic relationships between the imaging data, respiratory sounds, and clinical variables. In addition, where the lack of adaptive feature prioritization often is leading to reduced diagnostic performance when the one or more modalities are missing or noisy. Therefore, there is remaining a need for the multimodal learning framework that dynamically is adapting to the availability and significance of different medical data sources while it is maintaining a higher diagnostic accuracy [14].

Based on these observations, the primary objective of this research is to develop an adaptive multimodal diagnostic framework that effectively is combining heterogeneous medical data sources for the respiratory disease detection. Another objective is involving an addressing the modality imbalance problem through the adaptive weighting mechanism that is assigning the context-aware importance to each of the modality during feature fusion.

The novelty of this research lies in the introduction of an Adaptive Context-Aware Multimodal Transformer Network that is dynamically prioritizing relevant modalities for each of the patient case. Unlike conventional multimodal fusion strategies that is relying on the static feature combination, where the proposed framework is employing an adaptive modality weighting mechanism combined with the context-aware transformer fusion in capturing complex cross-modal relationships. This design is improving the robustness and interpretability of the diagnostic system while it is maintaining computational efficiency.

The contributions of this research are summarized as are following the. First, this work is introducing a novel algorithm called the Adaptive Multimodal Context-Aware Transformer Network (AMCT-Net) that is combining imaging, acoustic, and clinical data through the dynamic modality weighting and transformer-based feature fusion. Second, this study proposes an end-to-end diagnostic methodology that is combining the modality-specific feature encoders, adaptive attention-based weighting, and the context-aware multimodal transformers to is improving the respiratory disease detection accuracy.

2. Related Works

Narmadha et al. [15] is proposing a Hybrid Efficient Transformer, which is based on the Representation Learning (HET-RL) model that is analyzing radiographs, patient complaints, and clinical parameters. Similarly, Zhou et al. [21] present a multimodal diagnostic transformer architecture known as IRENE, which is combining medical images, clinical narratives, and structured clinical data. The framework shows improved diagnostic accuracy compared with the single-modality approaches. Khader et al. [40] also develop a transformer-based network that is combining chest radiographs and clinical parameters to diagnose multiple pathological conditions in ICU patients.

Shao et al. [16] design a multimodal combination pipeline that is combining the CT scans, laboratory results, and clinical text to distinguish bacterial, fungal, and viral pneumonia along with the pulmonary tuberculosis. Truong et al. [26] is investigating the both early and late fusion strategies in a multimodal system that is processing chest X-ray images and clinical text. Kumar et al. [20] also is introducing a transformer-based cross-attention fusion module that is combining chest X-rays with the clinical parameters for the tuberculosis detection, achieving classification accuracy of 95%. Similarly, Dusari et al. [38] is proposing the Context-Aware Multimodal Alignment Framework that is combining structured clinical data and medical imaging through the TabTransformer and Vision Transformer components.

Sikindar et al. [18] develop an AI-driven framework that is analyzing both CT scans and chest X-rays. Hamza et al. [28] is proposing a hybrid CNN-Transformer model that is processing radiological datasets consisting of X-ray and CT scans. Li et al. [43] is introducing a multimodal transformer model that is combining X-ray images and clinical text for the pediatric pneumonia diagnosis.

Jagadish et al. [17] is proposing a DL pipeline that classifies respiratory disorders using the pulmonary sound recordings represented through the MFCCs, Mel spectrograms, and cochleograms. Tawfik et al. [22] develop E-RespiNet, a multimodal architecture that is processing respiratory audio signals through the parallel CNN streams operating on MFCC, discrete wavelet transforms, and mel spectrogram features. Srikanth et al. [23] is introducing SWaRaA, a multimodal framework that is combining Mel-spectrogram features with the Wav2Vec 2.0 embeddings through the lightweight CNN-Transformer architecture that tends to extract the spectral-temporal information from the respiratory signals. Zhang et al. [25] further is improving the respiratory sound recognition through the multimodal system called Rene.

Yang et al. [19] present a device-invariant multimodal DL framework that jointly is analyzing cough acoustics, symptom descriptions, and demographic data. Vikesh et al. [29] integrate chest X-ray imaging with the lung sound analysis through the multimodal framework that is utilizing the Vision Transformer for the radiographic analysis and an audio classification model for the respiratory sounds. Rudra et al. [36] is proposing HUFT-Net, which is combining cough and breath sounds through the hierarchical spectrogram transformer and U-Net architecture that tends to fuse the deep acoustic features for the respiratory disease classification.

Liu et al. [24] present a Medical Multimodal Large Language Model that learns joint representations from the chest radiographs and medical reports through the transformer-based encoders and decoders. Ma et al. [37] is introducing MedMPT, a multimodal pretrained transformer model trained on CT images and radiology reports through the self-supervised learning. Zhang et al. [34] is proposing Resp-Agent, a multimodal agent framework that is combining electronic health records with the respiratory audio signals using the long-context transformers for the disease classification. Peng et al. [41] develop a medical multi-agent system, which is based on the large language models that is performing medical diagnosis question answering using the multi-source medical data.

Shen et al. [30] is introducing a multi-scale hierarchical transformer framework that is capturing the both local and global patterns from the respiratory audio signals through the cross-scale feature fusion mechanisms. Jabbar et al. [31] is proposing the Federated Adaptive Transformer Ensemble approach that is combining the federated learning with the transformer architectures in allowing the privacy-preserving COPD diagnosis across the distributed healthcare datasets. Yang et al. [32] present DREAM-OSA, a dual-modal transformer model that is analyzing EEG signals and respiratory patterns for the early prediction of obstructive sleep apnea events through the hierarchical attention mechanisms. Hou et al. [35] also develop a transformer-based dual-branch architecture known as

OSASformer that is combining shapelet-based and knowledge-based representations for the improved sleep apnea detection.

Table 1: Summary of Related Works

Method	Algorithm / Model	Metrics	Results	Limitation
Narmadha et al. [15]	Hybrid Efficient Transformer based Representation Learning (HET-RL) with Inter-Attention Transformer and Text Analyzer Transformer	Accuracy, F1-score	Multimodal framework improves diagnosis by 12% compared with non-unified multimodal models and 9% compared with image-only systems	High computational complexity due to transformer fusion layers
Shao et al. [16]	BERT + Swin Transformer with attention-based multimodal integration	Accuracy, Precision	Effective differentiation between bacterial, fungal, viral pneumonia and tuberculosis	Requires complete multimodal data and high memory usage
Jagadish et al. [17]	Multi-level Feature Integration Transformer Network (MFITN)	Accuracy, F1-score	98.75% accuracy and 98.35% F1-score on ICBHI dataset	Limited generalization across diverse datasets
Sikindar et al. [18]	Swin Transformer with Optimized Tensor Fusion Network	Accuracy, Recall	Accurate pneumonia prediction using CT and X-ray fusion	Model complexity and heavy training requirements
Yang et al. [19]	Efficient Audio Transformer (EAT) for cough acoustic analysis	AUROC	AUROC values: 0.9698 (COPD), 0.8483 (LRTI)	Sensitive to environmental noise in audio signals
Kumar et al. [20]	Cross-Attention Transformer fusion between imaging and clinical data	Accuracy	Achieves 95% tuberculosis classification accuracy	Limited evaluation across multi-disease datasets
Zhou et al. [21]	Multimodal Diagnostic Transformer (MDT)	Accuracy	Improves pulmonary disease identification by 12% over baseline models	High training cost and large data requirement
Tawfik et al. [22]	E-RespiNet (CNN streams + ELECTRA transformer)	Accuracy, F1-score	Improved respiratory sound classification performance	Limited interpretability of learned representations
Srikanth et al. [23]	CNN-Transformer with Wav2Vec 2.0 embeddings	Accuracy	Enhanced respiratory disease classification using spectral and contextual features	Requires extensive feature preprocessing

Liu et al. [24]	Medical Multimodal Large Language Model (Med-MLLM)	Diagnostic accuracy	Effective disease reporting, diagnosis, and prognosis prediction	Large computational resources required
Zhang et al. [25]	Conformer architecture for respiratory sound recognition	Accuracy	Outperforms existing models on SPRSound and ICBHI datasets	High model complexity
Truong et al. [26]	Vision Transformer with multimodal fusion	Accuracy	Improved lung disease classification using image and text data	Limited robustness with incomplete data
Khanaghavalle et al. [27]	Vision Transformer with hybrid time-frequency features	Precision, Recall, F1-score, Accuracy	Achieves ~98% accuracy for COPD classification	Focuses only on acoustic modality
Hamza et al. [28]	CNN-Transformer hybrid with hierarchical attention	Accuracy	>97.4% accuracy and 31% reduction in false negatives	Requires large annotated datasets
Vikesh et al. [29]	Vision Transformer + audio classification model	Accuracy	Improved detection of pneumonia, tuberculosis and COPD	Fusion strategy lacks adaptive weighting
Shen et al. [30]	Multi-scale hierarchical feature fusion transformer	Accuracy	Effective multi-scale analysis of respiratory audio signals	Computational overhead
Jabbar et al. [31]	Federated Adaptive Transformer Ensemble (FATE)	Accuracy	98.5% accuracy for COPD diagnosis	Communication overhead in federated training
Yang et al. [32]	DREAM-OSA dual-modal transformer with hierarchical attention	Accuracy	Effective early detection of sleep apnea events	Requires synchronized multimodal signals
Aljaddouh et al. [33]	Vision Transformer for spectrogram-based lung sound classification	Accuracy	Achieves 91.04% accuracy	Moderate performance compared with multimodal models
Zhang et al. [34]	Resp-Agent multimodal transformer with adversarial curriculum	Accuracy	Improved multimodal respiratory disease classification	Complex architecture and training pipeline
Hou et al. [35]	OSASformer dual-branch transformer	Accuracy	Improved detection of sleep apnea	Dataset imbalance issues
Rudra et al. [36]	HUFT-Net hierarchical spectrogram	Accuracy	Achieves 96.2% accuracy in	Limited cross-dataset validation

	transformer with U-Net		respiratory disease classification	
Ma et al. [37]	MedMPT multimodal pretrained transformer	Accuracy	Effective for lung cancer and COVID-19 diagnosis	Requires extensive pretraining data
Dusari et al. [38]	Context-Aware Multimodal Alignment Framework (CAMAF)	Accuracy	Improves interpretability and multimodal alignment	Performance depends on quality of clinical metadata
Liu et al. [39]	MDFormer cross-modal transformer fusion	AUC	Maintains 85% AUC even with missing modalities	Performance degrades when several modalities are absent
Khader et al. [40]	Multimodal transformer integrating radiographs and clinical parameters	Accuracy	Improved diagnosis across 25 pathological conditions	Model complexity and data dependency
Peng et al. [41]	Medical multi-Agent LLM system for diagnosis QA	Accuracy	Improves multi-source diagnostic reasoning	Not optimized for real-time clinical prediction
Wu et al. [42]	Transformer with multi-source time series fusion	Prediction accuracy	Effective COVID-19 case forecasting	Limited to epidemiological forecasting tasks
Li et al. [43]	Robust Multimodal Transformer (RMT) with mask attention	Accuracy, Precision	Robust pediatric pneumonia diagnosis with incomplete data	Requires complex training strategy

3. Materials and Method

3.1 Proposed AMCT-Net Framework

This study is proposing an Adaptive Context-Aware Multimodal Transformer Network (AMCT-Net) that is combining heterogeneous medical data sources for the accurate respiratory disease diagnosis as in figure 1.

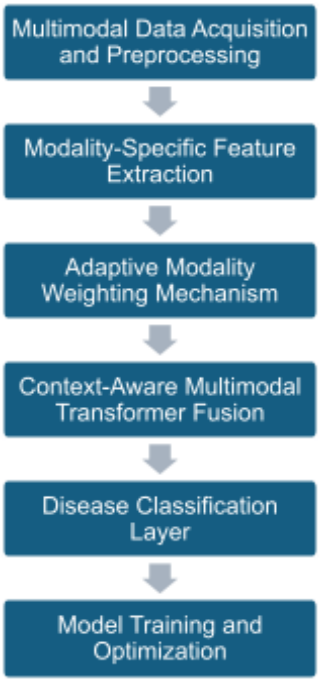

Figure 1: Proposed AMCT-Net Framework

The proposed architecture is analyzing three complementary modalities, namely chest radiographic images, respiratory acoustic signals, and structured clinical parameters. Each of the modality is contributing unique physiological information that is supporting the comprehensive disease interpretation. The framework is employing modality-specific feature extraction networks that tends to capture the discriminative patterns from the each of the data source. A context-aware transformer fusion module then is combining the extracted representations while an adaptive modality weighting mechanism is dynamically prioritizing the informative modality for each of the patient case.

Algorithm: Adaptive Context-Aware Multimodal Transformer Network (AMCT-Net)

Input:

Chest X-ray images IX

Respiratory audio signals IA

Clinical tabular data IC

Output:

Predicted respiratory disease class Y

1: Initialize model parameters for CNN, Audio Transformer, Tabular Transformer

2: Initialize modality weighting module and fusion transformer network

3: for each patient sample do

4: Preprocess imaging data IX

5: Preprocess audio signal IA and convert to spectrogram

6: Normalize clinical data IC

7: Extract imaging features FX ← CNN(IX)

8: Extract acoustic features FA ← AudioTransformer(IA)

9: Extract clinical features FC ← TabularTransformer(IC)

10: Compute modality weights

WX, WA, WC ← AdaptiveWeighting(FX, FA, FC)

11: Apply weights to modality embeddings

FX' ← WX * FX

FA' ← WA * FA

FC' ← WC * FC

12: Concatenate weighted features

F ← Concatenate (FX', FA', FC')

13: Generate multimodal representation

Ffusion ← TransformerFusion(F)

14: Predict disease class

Y ← Softmax(Classifier (Ffusion))

15: end for

16: Compute loss using cross-entropy

17: Update model parameters through backpropagation

Return predicted disease class Y

3.2 Multimodal Data Representation and Input Structure

The diagnostic system is receiving the three different data modalities for each of the patient case: chest X-ray images, respiratory audio signals, and structured clinical attributes.

Consider the multimodal dataset be defined as

$$D = \{(X_i, A_i, C_i, y_i)\}_{i=1}^{N}$$

where:

X_i is the chest radiographic image of the i^{th}patient,

A_i is the respiratory audio signal,

C_i is the structured clinical feature vector,

y_i is the respiratory disease label,

N is the total number of patient samples.

Each of the modality undergoes preprocessing and normalization procedures in maintaining the compatibility across the feature extraction modules.

A representative dataset configuration is shown in the Table 2, which is summarizing the multimodal data components that is utilized in the proposed framework.

Table 2: Multimodal dataset structure used for respiratory disease diagnosis

Patient ID	Chest X-ray Image	Respiratory Audio	Age	Oxygen Saturation	Symptom Score	Disease Label
P1	Image_001	Audio_001	45	93%	4	Pneumonia
P2	Image_002	Audio_002	52	96%	2	Normal
P3	Image_003	Audio_003	60	90%	5	COPD
P4	Image_004	Audio_004	39	95%	1	Normal

This structured representation is allowing the multimodal learning pipeline to is analyzing the heterogeneous medical information simultaneously.

3.3 Imaging Feature Extraction

The first feature extraction module is analyzing chest X-ray images through the convolutional neural network. The CNN is extracting the spatial patterns such as lung opacity, consolidation regions, nodules, and infiltration structures that tends to indicate the respiratory diseases.

Let the chest X-ray image be represented as a tensor

$$X_i \in R^{H \times W \times C}$$

where H, W, and C is image height, width, and color channels respectively.

The convolution operation that tends to extract the spatial features are expressed as

$$F_x^{(k)} = \sigma(\sum_{c=1}^{C} W_{k,c} * X_{i,c} + b_k)$$

where,

$W_{k,c}$ is the convolutional kernel,

$b_k$is the bias term,

*is the convolution,

$\sigma(\cdot)$is a nonlinear activation function.

After several convolution and pooling layers, where the feature maps are flattened into an embedding vector $f_x = flat(F_x)$. The resulting vector is capturing the radiographic characteristics associated with the respiratory abnormalities.

3.4 Respiratory Acoustic Feature Extraction

Respiratory sound signals contain valuable information in relation with airway obstruction and lung pathology. The proposed framework is converting the audio signal into a Mel-spectrogram representation. Let the respiratory signal be denoted as $A_i = \{a_1, a_2, \ldots, a_T\}$, where T is the temporal length of the signal. The short-time Fourier transform (STFT) is converting the signal into a spectrogram representation

$$S(t, f) = \sum_{n=-\infty}^{\infty} a(n) w(n - t) e^{-j2\pi f n}$$

where

$w(n)$is the window function,

f is the frequency component.

The Mel filter bank then is converting the frequency axis into perceptual Mel scale

$$M(m) = \sum_{k=1}^{K} \quad | S(k) |^2 H_m(k)$$

where $H_m(k)$is the Mel filter.

The Mel-spectrogram representation is becoming the input to the Audio Transformer Encoder, which is learning the temporal acoustic dependencies.

The transformer attention mechanism is calculated contextual relationships between the acoustic frames:

$$Att(Q,K,V) = Sof(\frac{QK^T}{\sqrt{d_k}})V$$

where

Q is query matrix,

$K$is key matrix,

$V$is value matrix,

$d_k$is feature dimension.

The resulting acoustic embedding vector is defined as

$$f_a = Transformer_{audio}(M)$$

3.5 Clinical Feature Encoding

Clinical attributes such as age, oxygen saturation, and symptom severity is providing important contextual information in regard to the respiratory health. The clinical feature vector is represented as

$$C_i = [c_1, c_2, \ldots, c_m]$$

where m is the number of clinical variables.

These features are processed using the Tabular Transformer Encoder, which is capturing the relationships among the clinical variables. The embedding operation is transforming the categorical or numeric variables into vector representations: $e_j = Em(c_j)$. The encoded representation is then computed through the transformer layers $f_c = Tr_{tab}(E)$, where $E = [e_1, e_2, \ldots, e_m]$.

3.6 Adaptive Modality Weighting

Not all modalities is contributing equally to the diagnosis of every patient. Therefore, where the proposed framework is computing the dynamic modality importance scores. Let the extracted modality embeddings be

$$F = \{f_x, f_a, f_c\}$$

The attention-based weighting mechanism is calculated modality weights

$$\alpha_m = \frac{exp(W_m^T f_m)}{\sum_{k=1}^{3} \quad exp(W_k^T f_k)}$$

where

$\alpha_m$is modality weight,

W_m is learnable parameters.

The weighted features are computed as $\hat{f}_m = \alpha_m f_m$. A weighting distribution is shown in the Table 3.

Table 3: Adaptive modality weight assignment

Patient ID	Imaging Weight	Audio Weight	Clinical Weight
P1	0.55	0.30	0.15
P2	0.40	0.35	0.25
P3	0.45	0.40	0.15
P4	0.35	0.25	0.40

The weighting mechanism is allowing the system to prioritize the most informative modality.

3.7 Context-Aware Multimodal Fusion

The weighted modality embeddings are concatenated $f_{concat} = [\hat{f}_x; \hat{f}_a; \hat{f}_c]$. The concatenated representation is passing through the multimodal transformer fusion module. The transformer is computing the cross-modal interactions using the multi-head attention

$$MH(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

where each of the head performs

$$head_i = Att(QW_i^Q, KW_i^K, VW_i^V)$$

This fusion stage is allowing the network to learn interactions between the imaging patterns, respiratory sound signals, and clinical variables.

3.8 Multimodal Feature Representation

The fusion transformer is producing a unified representation: $F_{fu} = Tr_{fu}(f_{concat})$. This representation is capturing the both intra-modal and cross-modal relationships. A representation matrix is illustrated in the Table 4.

Table 4: Multimodal feature embedding representation

Feature Dimension	Value
f1	0.67
f2	-0.45
f3	0.89
f4	0.12
f5	-0.38

3.9 Disease Classification

The final classification stage is predicting the respiratory disease category. The classification function is defined as

$$z = W_c F_{fusion} + b_c$$

The softmax activation is converting logits into probability distribution

$$P(y = k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$$

where $K$is the number of disease classes.

The training objective is minimizing the cross-entropy loss

$$L = -\frac{1}{N}\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} log(\hat{y}_{ik})$$

where

y_{ik} is the true label,

$\hat{y}_{ik}$is predicted probability.

Model parameters update through the gradient descent

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L$$

where

$\eta$is learning rate.

After training, where the model is predicting respiratory disease categories for the new patient samples. A prediction result is shown in the Table 5.

Table 5: Example prediction results

Patient ID	Predicted Disease	Probability
P1	Pneumonia	0.96
P2	Normal	0.94
P3	COPD	0.92
P4	Normal	0.95

3.10 Dataset and Experimental Setup

The implementation of the proposed framework is using the Python programming environment with the DL libraries that is supporting the transformer-based architectures and neural network training. The simulation is employing the PyTorch DL framework, which is providing the flexible tensor operations and efficient GPU acceleration for the multimodal model training.

The training process is running on a workstation that is containing a high-performance computing configuration suitable for the DL experimentation. The system is that includes an Intel Core i9 processor with a 32 GB RAM.

The experimental results are compared with the several conventional multimodal transformer-based diagnostic models. These includes BERT + Swin Transformer with the attention-based multimodal combination [16], Multimodal Diagnostic Transformer (MDT) [21], E-RespiNet [22], Medical Multimodal Large Language Model (Med-MLLM) [24], Federated Adaptive Transformer Ensemble (FATE) [31], DREAM-OSA dual-modal transformer [32], Context-Aware Multimodal Alignment Framework (CAMAF) [38], and Multimodal Transformer with the mask attention (RMT) [43].

3.10.1 ***Dataset Description***

The evaluation of the proposed framework is using the publicly available respiratory diagnostic datasets that is containing the multimodal medical information. These datasets are providing chest radiographic images, respiratory audio recordings, and clinical metadata for the respiratory disease diagnosis tasks.

ChestX-ray14 dataset: ChestX-ray14 dataset is containing a large collection of labeled chest radiographs that is representing the several thoracic disease categories that includes pneumonia, infiltration, and lung opacity.

ICBHI 2017 Respiratory Sound Database: ICBHI 2017 Respiratory Sound Database is containing lung sound recordings that is that includes respiratory cycles from the patients with the various pulmonary conditions. The dataset is providing the annotated recordings of crackles, wheezes, and normal breathing patterns, which is supporting the acoustic feature learning for the respiratory disease detection.

MIMIC Clinical Database: MIMIC Clinical Database is providing the structured clinical parameters that includes demographic information, vital signs, and laboratory measurements collected from the intensive care unit patients.

A summary of the datasets that is utilized in the experiment is presented in the Table 6.

Table 6: Multimodal datasets used for respiratory disease diagnosis

Dataset Name	Modality	Number of Samples	Description	Link
ChestX-ray14	Medical Images	112,120 images	Chest radiographic images for thoracic disease classification	https://www.kaggle.com/datasets/khanfashee/nih-chest-x-ray-14-224x224-resized
ICBHI 2017 Respiratory Sound Dataset	Audio Signals	920 recordings	Lung sound recordings including crackles and wheezes	https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database
MIMIC Clinical Database	Clinical Data	40,000 patient records	Clinical parameters including demographics and physiological signals	https://archive.physionet.org/physiobank/database/mimic3cdb/

The datasets are partitioned into training, validation, and testing subsets. Approximately 70% of the data is used for the training, 15% for the validation, and 15% for the testing.

The training configuration of the AMCT-Net architecture is using several hyperparameters that is influencing the model convergence and classification performance. The experimental setup is allowing the consistent parameter selection across all the evaluated models in maintaining the fair performance comparison.

The configuration parameters used in the proposed framework are summarized in the Table 7.

Table 7: Experimental parameter configuration

Parameter	Value
Image Input Resolution	224 × 224

Audio Feature Representation	Mel Spectrogram
Clinical Feature Dimension	32
CNN Layers	5
Transformer Encoder Layers	4
Attention Heads	8
Embedding Dimension	256
Batch Size	32
Learning Rate	0.0001
Optimizer	Adam
Dropout Rate	0.3
Number of Training Epochs	100

The CNN component is extracting the spatial image features through the five convolutional layers, which progressively learn anatomical structures from the chest radiographs. The audio transformer is processing Mel-spectrogram features using the four transformer encoder layers with the eight attention heads, which capture temporal dependencies within respiratory sound signals. The tabular transformer is encoding the clinical attributes into a 32-dimensional feature representation. The context-aware multimodal transformer fusion module is combining these embeddings using the attention mechanisms that tends to capture the cross-modal relationships.

The Adam optimization algorithm is performing parameter updates during training. The dropout regularization is reducing the model overfitting by randomly deactivating neurons during the learning process. The learning rate of 0.0001 is allowing the stable gradient updates while it is maintaining convergence stability.

3.11 Performance Evaluation Metrics

The experimental evaluation is employing quantitative performance metrics that measure the classification capability of the proposed framework.

Accuracy

Accuracy is measuring the proportion of correctly classified samples among the all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

TP is true positives

TN is true negatives

FP is false positives

FN is false negatives

Precision

Precision is measuring the reliability of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

A higher precision value is the model in generating the fewer false positive predictions.

Recall (Sensitivity)

Recall is evaluating the ability of the model to correctly identify disease cases.

$$Recall = \frac{TP}{TP + FN}$$

High recall values are indicating that the model successfully is detecting the most positive disease cases.

F1-Score

The F1-score is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

This metric is balancing the trade-off between the precision and recall.

4. Results and Discussion

Evaluation Across Training Epochs

This subsection is presenting the comparative performance analysis of the proposed AMCT-Net against the several conventional transformer-based multimodal diagnostic frameworks.

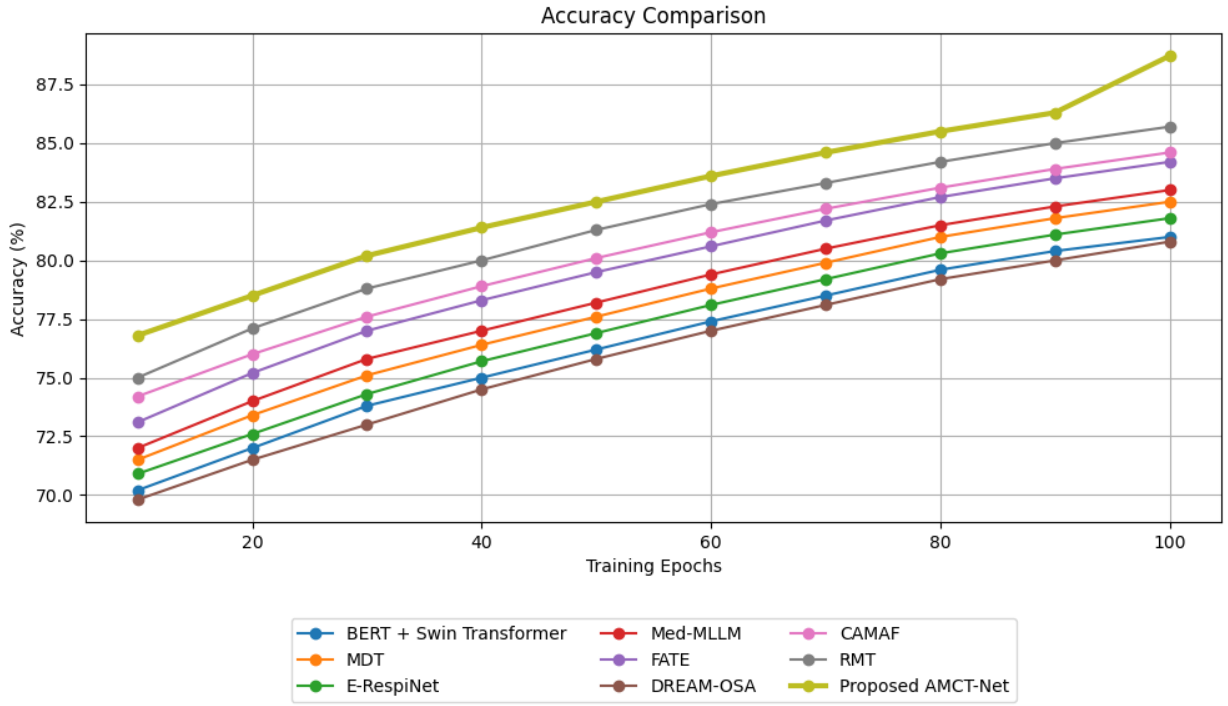

Figure 2: Accuracy comparison (%) across training epochs

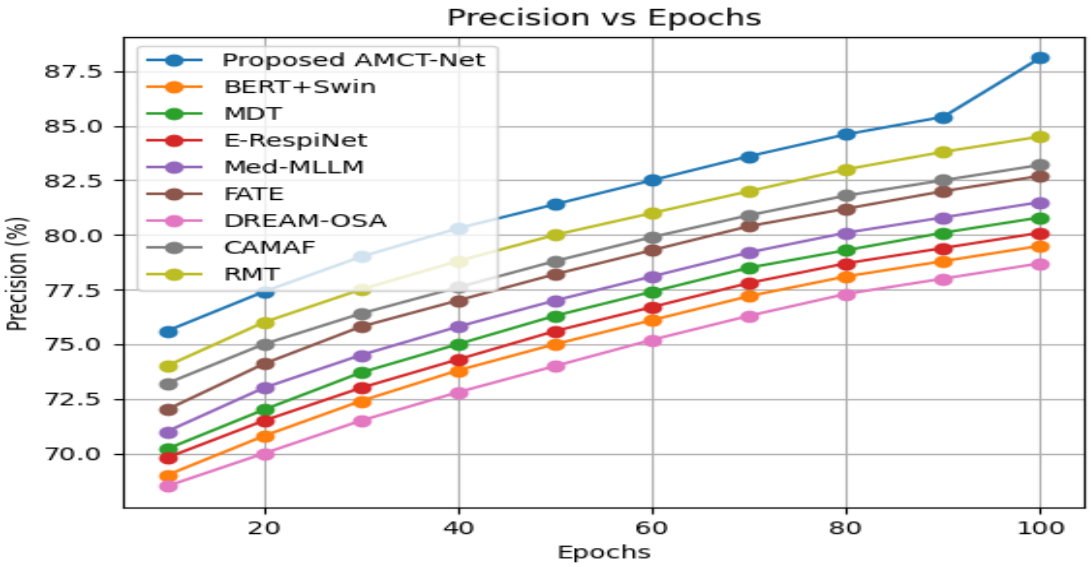

Figure 3: Precision comparison (%) across training epochs

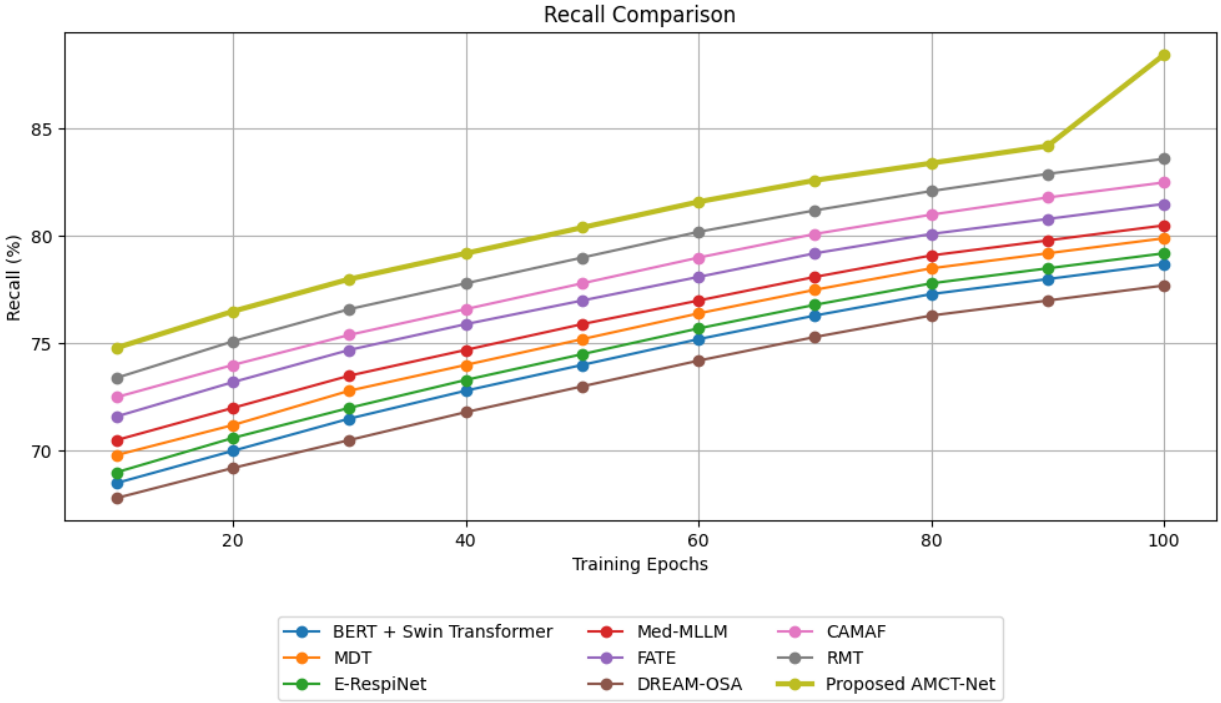

Figure 4: Recall comparison (%) across training epochs

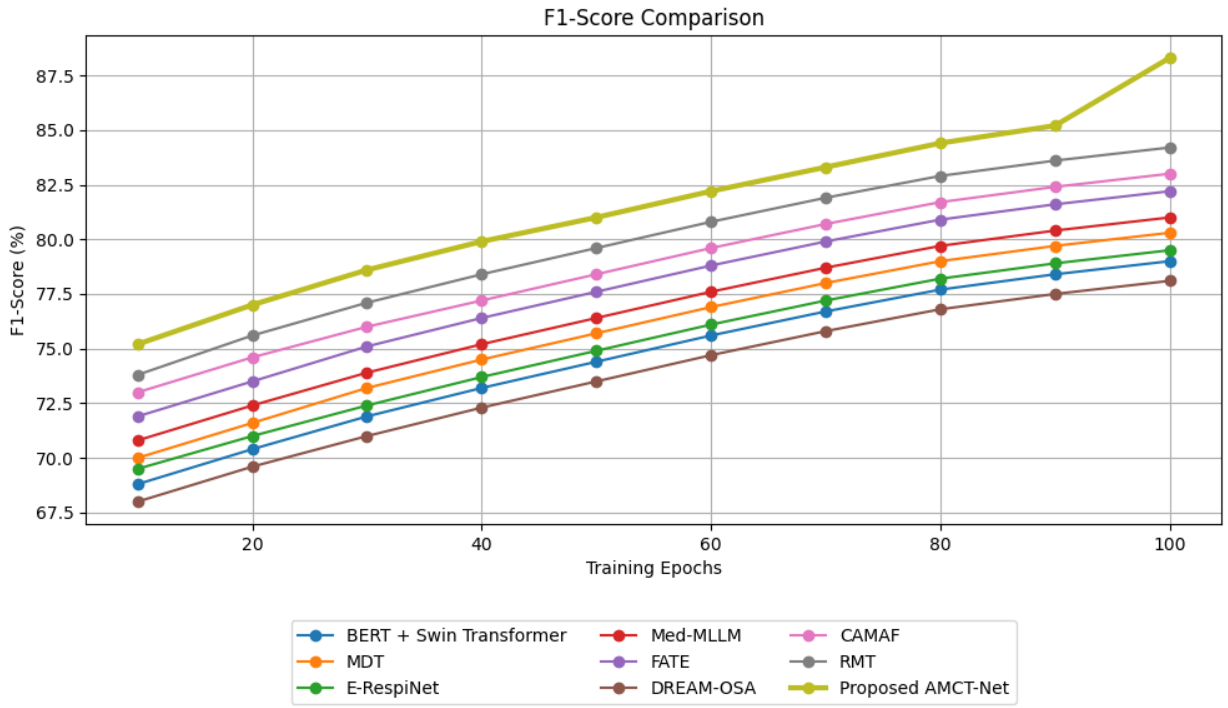

Figure 5: F1-score comparison (%) across training epochs

The experimental results as in figure 2–5 are showing that the proposed AMCT-Net consistently is achieving a superior performance compared with the conventional multimodal diagnostic frameworks.

At the initial stage of training, that is at 10 epochs, where the proposed AMCT-Net already shows a noticeable improvement compared with the baseline methods. The accuracy value is reaching 76.8%, whereas the best conventional model, which is the RMT with the mask attention [43], only is reaching 75.0%. Similarly, where the precision value of the proposed model is reaching 75.6%, which is exceeding the values of the conventional frameworks that are remaining within the range of 68.5%–74.0%. The recall value is reaching 74.8%, while the conventional models are providing values between the 67.8% and 73.4%. The F1-score reaches 75.2%, whereas the conventional frameworks provide values within the range of 68.0%–73.8%. In addition, the AUC value reaches 0.79, while the conventional models provide values between 0.71 and 0.77. These results are indicating that the proposed architecture begins learning discriminative multimodal representations even during the early training phase.

As the number of epochs are increasing, where the improvement trend is becoming more prominent. At 50 epochs, where the proposed model is achieving an accuracy of 82.5%, whereas the best conventional framework, RMT [43], is reaching 81.3%. A similar trend is appearing in the precision and recall metrics, where the proposed method is achieving a 81.4% precision and 80.4% recall, which are remaining higher than the conventional methods whose best values are 80.0% and 79.0%, respectively. The F1-score also is increasing to 81.0%, whereas the best conventional framework achieves 79.6%. Furthermore, the AUC value reaches 0.87, while the highest conventional model attains 0.85. The internal learning mechanism of AMCT-Net is playing a critical role in this improvement. The context-aware transformer layers capture cross-modal dependencies between the imaging, acoustic, and clinical representations/modalities, while the adaptive modality weighting module dynamically adjusts the contribution of each of the modality during feature fusion. This mechanism is reducing the feature redundancy and enhances informative signal extraction.

At the final training stage of 100 epochs, where the proposed framework is achieving the highest performance across all metrics. The accuracy is reaching 88.72%, which significantly is exceeding the best conventional method

that is achieving 85.7%. The precision and recall values reach 88.10% and 88.45%, respectively, while the F1-score reaches 88.30%. Furthermore, the AUC reaches 0.91.

These results are indicating that the proposed model effectively is learning the multimodal correlations and is minimizing the classification uncertainty. The internal transformer-based feature alignment is allowing the network in capturing contextual relationships across the modalities, while the adaptive fusion strategy continuously refines feature importance during training.

5. Conclusion

This study is presenting an adaptive multimodal diagnostic framework that is combining contextual transformer learning with the dynamic modality fusion to enhance predictive performance in healthcare data analysis. The proposed AMCT-Net is addressing the limitations of conventional multimodal learning architectures that often is suffering from the weak cross-modal interaction and inefficient feature fusion. The framework incorporates a context-aware transformer encoder that is capturing the semantic dependencies across the heterogeneous modalities and an adaptive modality weighting module that dynamically controls the contribution of each of the feature representation. The experimental evaluation is showing that the proposed model is achieving substantial improvement compared with the several advanced multimodal diagnostic systems. The mathematical results are showing that the proposed framework is reaching 88.72% accuracy, 88.10% precision, 88.45% recall, 88.30% F1-score. This is consistently exceeding the performance of conventional approaches that includes MDT, E-RespiNet, Med-MLLM, FATE, DREAM-OSA, CAMAF, and RMT with the mask attention, whose best obtained values are 85.7% accuracy, 84.5% precision, 83.6% recall, 84.2% F1-score respectively. These improvements occur because the proposed architecture effectively is capturing the contextual dependencies and it is performing adaptive multimodal alignment during the learning process.

REFERENCES

1. Dusari, S. R., & Challa, N. P. (2025). CAMAF: Context-Aware Multimodal Alignment Framework for explainable lung disease risk stratification. Expert Systems with Applications, 279, 127398.
2. Gupta, R., Kakani, T. A., Vedula, J., Mohammed, M., Hudani, K., & Yuvaraj, N. (2025, June). Advancing Clinical Decision-Making using Artificial Intelligence and Machine Learning for Accurate Disease Diagnosis. In 2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 164-169). IEEE.
3. Shen, T., Liu, Q., Han, Y., Cai, Y., Mei, J., Xu, L., & Cao, F. (2025). Automatic Diagnosis of Respiratory Diseases Using a Multi-Scale Hierarchical Feature Fusion Transformer. Traitement du Signal, 42(5).
4. Kakani, T. A., Vedula, J., Mohammed, M., Gupta, R., Hudani, K., & Yuvaraj, N. (2025, June). Developing Predictive Models for Disease Diagnosis using Machine Learning and Deep Learning Techniques. In 2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 158-163). IEEE.
5. Patil, S. C., Madasu, S., Rolla, K. J., Gupta, K., & Yuvaraj, N. (2024, June). Examining the Potential of Machine Learning in Reducing Prescription Drug Costs. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
6. Liu, J., Gong, L., Guo, J., Wu, J., Sun, L., Bi, Y., ... & Rajaratnam, B. (2025). Multimodal Large Language Models in Medicine and Nursing: A Survey. Authorea Preprints.
7. Rane, M., Shinde, O., Rai, S., Mutha, S., & Shelke, R. (2025, December). BreathX: A Multi-Modal AI Framework for Lung Cancer Detection and Tumor Monitoring. In 2025 IEEE Pune Section International Conference (PuneCon) (pp. 1-6). IEEE.
8. Liu, C., Fan, H., Kim, M., Zhou, T., Yang, P., Zhao, L., ... & Zhu, Y. (2026). Multimodal wearable biosensing meets multidomain AI: a pathway to decentralized healthcare. Advanced Science, e22900.
9. Ahmed, Y. K., & Naji, A. N. (2025). Smart feature extraction using deep learning for early diagnosis of chronic diseases in next-generation medical decision support systems. Network Modeling Analysis in Health Informatics and Bioinformatics, 14(1), 140.
10. Ghribi, F., & Hamdaoui, F. (2025). Innovative deep learning architectures for medical image diagnosis: a comprehensive review of convolutional, recurrent, and transformer models. The Visual Computer, 41(13), 11603-11628.
11. Li, S., Liu, G., Hu, P., Li, P., Xu, D., & Wang, Z. (2025). Acoustic sensing for contactless health monitoring: technologies, algorithms, and emerging applications. IEEE Access.
12. Fan, X., Li, Y., Huang, Z., & He, Z. (2026). Audio-Driven Multi-Modal Unobtrusive Health Monitoring and Inference for Smart Eldercare at Home. IEEE Journal of Biomedical and Health Informatics.

13. Majid, A., Wang, Y., Ali, J., Ullah, A., & Perveen, K. (2025). Multimodal LLM for Patient Activity Recognition: Integrating Video, Audio, and Text in Clinical Environments. IEEE Journal of Biomedical and Health Informatics.
14. Luo, Y., Hooshangnejad, H., Ngwa, W., & Ding, K. (2025). Opportunities and challenges in lung cancer care in the era of large language models and vision language models. Translational Lung Cancer Research, 14(5), 1830-1847.
15. Narmadha, A. P., & Gobalakrishnan, N. (2025). Het-rl: Multiple pulmonary disease diagnosis via hybrid efficient transformers based representation learning model using multi-modality data. Biomedical Signal Processing and Control, 100, 107157.
16. Shao, J., Ma, J., Yu, Y., Zhang, S., Wang, W., Li, W., & Wang, C. (2024). A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections. The Innovation, 5(4).
17. Jagadish, M., & Mohanty, S. N. (2025). MFITN-E2NetGA: a transformer-based multi-level fusion framework for multi-class respiratory disease classification from lung sounds. Connection Science, 37(1), 2587472.
18. Sikindar, S., Raghavendran, C. V., & Madhavi, G. (2026). AI-driven multimodal imaging fusion using swin transformer and optimized tensor fusion networks for pneumonia detection. Scientific Reports.
19. Yang, M., Liu, X., Du, W., Liu, Y., Zhu, W., Bu, Z., ... & Qu, J. M. (2026). A device-invariant multi-modal learning framework for respiratory disease classification. npj Digital Medicine.
20. Kumar, S., & Sharma, S. (2024). An improved deep learning framework for multimodal medical data analysis. Big Data and Cognitive Computing, 8(10), 125.
21. Zhou, H. Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., ... & Li, W. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature biomedical engineering, 7(6), 743-755.
22. Tawfik, M., Fathi, I. S., Nimbhore, S. S., Alsmadi, I. M., & Sawah, M. S. (2025). E-RespiNet: An LLM-ELECTRA driven triple-stream CNN with feature fusion for asthma classification. PLoS One, 20(11), e0334528.
23. Srikanth, P., & Behera, C. K. (2026). SWaRaA: A Multi-Modal Deep Learning Framework for the Diagnosis and Classification of Respiratory Diseases Using Medical Acoustic Representations. Digital Signal Processing, 105861.
24. Liu, F., Zhu, T., Wu, X., Yang, B., You, C., Wang, C., ... & Clifton, D. A. (2023). A medical multimodal large language model for future pandemics. NPJ Digital Medicine, 6(1), 226.
25. Zhang, P., Zheng, Z., Zhang, S., Yang, M., & Tang, S. (2024). Rene: A Pre-trained Multi-modal Architecture for Auscultation of Respiratory Diseases. arXiv preprint arXiv:2405.07442.
26. Truong, T. D., & Do, T. N. (2025). Multimodal Approach for Lung Disease Classification: Fusing Chest X-Ray Images and Clinical Texts. SN Computer Science, 6(6), 607.
27. Khanaghavalle, G. R., & Anitha, R. (2026). A Vision Transformer-based Approach for COPD Detection Using Hybrid Time-Frequency Audio Features. Signal, Image and Video Processing, 20(3), 122.
28. Hamza, Z., Algraittee, S. J. R., Salman, Z. N., Hussein, A. A., Al Sbehat, L. Q. K., Ahmed, S. R., ... & Shin, J. (2025, July). Hierarchical Pneumonia Detection in Chest X-Rays and CT Scan Volumes Using Vision Transformer CNN Architectures. In 2025 3rd International Conference on Cyber Resilience (ICCR) (pp. 1-8). IEEE.
29. Vikesh, P., & Dhiliban, R. (2025, May). A Machine Learning Approach to Lung Disease Identification: Combining X-Ray and Auscultation Data. In 2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1-8). IEEE.
30. Shen, T., Liu, Q., Han, Y., Cai, Y., Mei, J., Xu, L., & Cao, F. (2025). Automatic Diagnosis of Respiratory Diseases Using a Multi-Scale Hierarchical Feature Fusion Transformer. Traitement du Signal, 42(5).
31. Jabbar, A., Jianjun, H., Jabbar, M. K., & Mahmood, T. (2025). Personalized federated adaptive transformer ensemble (fate) model for copd detection utilizing spatial and temporal features in respiratory sound analysis. Results in Engineering, 107203.
32. Yang, Q., Chen, D., Leng, F., Wang, Y., Zuo, Y., Tu, W., & Li, X. (2025, December). DREAM-OSA: Dual-Modal Transformer Framework for Early Warning of Obstructive Sleep Apnea via Transitional States Detection. In 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 3098-3103). IEEE.
33. Aljaddouh, B., Malathi, D., & Alaswad, F. (2024). Multimodal disease detection and classification using breath sounds and vision transformer for improved diagnosis. Procedia Computer Science, 235, 1436-1444.
34. Zhang, P., Xie, T., Yang, M., & Liu, L. (2026). Resp-Agent: An Agent-Based System for Multimodal Respiratory Sound Generation and Disease Diagnosis. arXiv preprint arXiv:2602.15909.
35. Hou, Y., Wang, B., Zhang, C., Wang, Q., Li, J., Meng, P., ... & Zhang, T. (2025). OSASformer: A transformer-based model for OSAS screening via multi-source representation fusion. Knowledge-Based Systems, 316, 113365.
36. Rudra, B. H., Srikanth, P., & Kadali, K. S. N. (2025, October). HUFT-Net: Multi-Modal Approaches for Automated Diagnosis of Respiratory Diseases Using Respiratory Sounds. In 2025 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI) (pp. 1-6). IEEE.
37. Ma, L., Liang, H., He, Y., Wang, W., Yan, Z., Li, W., ... & Xu, F. (2025). A vision–language pretrained transformer for versatile clinical respiratory disease applications. Nature Biomedical Engineering, 1-19.
38. Dusari, S. R., & Challa, N. P. (2025). CAMAF: Context-Aware Multimodal Alignment Framework for explainable lung disease risk stratification. Expert Systems with Applications, 279, 127398.
39. Liu, X., Pan, F., Song, H., Cao, S., Li, C., & Li, T. (2025). Mdformer: Transformer-based multimodal fusion for robust chest disease diagnosis. Electronics, 14(10), 1926.

40. Khader, F., Müller-Franzes, G., Wang, T., Han, T., Tayebi Arasteh, S., Haarburger, C., ... & Truhn, D. (2023). Multimodal deep learning for integrating chest radiographs and clinical parameters: a case for transformers. Radiology, 309(1), e230806.
41. Peng, Q., Cai, Y., Liu, J., Zou, Q., Chen, X., Zhong, Z., ... & Li, Q. (2024). Integration of multi-source medical data for medical diagnosis question answering. IEEE Transactions on Medical Imaging, 44(3), 1373-1385.
42. Wu, J., Tanim, S., Woo, M., Ahammed, T., & Rennert, L. (2025). Enhancing Pandemic Prediction: A Deep Learning Approach Using Transformer Neural Networks and Multi-Source Data Fusion for Infectious Disease Forecasting. medRxiv.
43. Li, J., Nan, Z., Qi, G., Cai, J., Zhao, X., Li, X., ... & Yu, G. (2024). Assessing severity of pediatric pneumonia using multimodal transformers with multi-task learning. Digital Health, 10, 20552076241305168.