Natural Resources for Human Health
eISSN: 2583-1194

Original Research
Received 10 May 2026
Revised 15 June 2026
Accepted 02 July 2026

Natural Resources for Human Health 2026, Vol. 6 (4s): 586–598

KEYWORDS:
Multimodal, Artificial Intelligence, Medical image analysis, Clinical decision support, Deep learning, Disease diagnosis, Feature fusion

Multimodal Artificial Intelligence for Early Disease Diagnosis and Clinical Decision Support using Medical Images and Patient Data

Rathi Soneji[1], Dhaval Makwana[2], Gaurav Agrawal[3], Prof. (Dr.) kalpesh Rasiklal Rakholia[4], Siddharth Jain[5], Dhaval H. Trivedi[6]

[1] Assistant Professor, Department of Artificial Intelligence and Machine Learning, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India
rathi.soneji46541@paruluniversity.ac.in, ORCID ID - 0009-0004-7104-4649
[2] Assistant Professor, Department of Artificial Intelligence and Machine Learning, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India
dhaval.makwana46297@Paruluniversity.ac.in, ORCID ID 0009-0009-8227-8500
[3] Assistant Professor, Department of Artificial Intelligence and Machine Learning, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India
gaurav.agrawal47749@paruluniversity.ac.in, ORCID ID -0009-0007-2026-4758
[4] Professor, Swarrnim School of Computing & IT, Swarrnim Startup & Innovation university, Gandhinagar, Gujarat, India
kalpesh.rakholia@swarrnim.edu.in, ORCID ID -0009-0007-2284-2195
[5] Assistant Professor, Department of Artificial Intelligence and Machine Learning, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India
siddharth.jain46123@paruluniversity.ac.in, ORCID ID 0009-0007-2924-8930,
Scopus ID- 60703944200
[6] Assistant Professor, Department of Artificial Intelligence and Machine Learning, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India
dhaval.trivedi37492@paruluniversity.ac.in, ORCID ID 0009-0006-2323-1129

ABSTRACT: The rapid growth of medical data generated from imaging systems, electronic health records, laboratory investigations, and patient monitoring platforms has created new opportunities for artificial intelligence-based clinical decision-support systems. Conventional diagnostic models generally rely on individual data modalities and may not represent the complete clinical reasoning process. This study proposes a multimodal artificial intelligence framework integrating medical images with structured patient information for improved early disease diagnosis and clinical decision support. The framework combines deep image feature extraction with clinical feature encoding followed by feature-level fusion. Comparative evaluation of image-only, clinical-data-only, and multimodal models demonstrates that integration of heterogeneous medical information improves diagnostic performance. The proposed framework shows potential for assisting clinicians through accurate, patient-specific and data-driven decision support.

1. INTRODUCTION

AI has made substantial strides in the healthcare sector, revolutionizing the ability of machines to analyze intricate biomedical data and assist clinicians in decision-making based on data.The rapid development of these AI (artificial intelligence) technologies has revolutionized healthcare, allowing machines to interpret complex biomedical data and guide data-driven clinical decisions. In recent years, the proliferation of digital medical data, such as medical images, electronic health records (EHRs), laboratory investigations, demographic data and physiological measurements, provides new opportunities for the creation of intelligent diagnostic systems. Deep learning is one of the most powerful AI techniques and has shown great promise in the medical domain due to its ability to learn features and patterns from high-dimensional and massive datasets with minimal feature engineering. In the medical field, deep learning models have successfully been employed for various tasks such as disease diagnosis, abnormality detection, segmentation and prognosis prediction in applications such as radiology, pathology, ophthalmology and oncology [1,2].

While AI systems based on image data have made great strides, traditional diagnostic models based on one data modality are also faced with limitations. Medical images chiefly give information about anatomical structures and visible pathological changes, but usually clinical diagnosis is not made solely on the basis of the image. The diagnostic conclusions are usually made based on the patient's history, laboratory data, symptoms, physiological data, gender and age as well as imaging data. An AI model trained solely on medical imagery can detect patterns within the images, but might not fully reflect the clinical context needed for precise patient evaluations. This constraint has spurred efforts to develop multimodal AI strategies that can combine disparate healthcare data to make more comprehensive and accurate diagnostic predictions [3].

Multimodal AI is a new type of AI that integrates information from various sources to create a comprehensive representation of a patient's condition. Multimodal learning allows for the fusion of medical images and structured clinical information, along with medical text, laboratory metrics, and physiological signals in healthcare applications. Multimodal AI takes advantage of the complementary nature of various sources of data. Medical imagery offers spatial and structural information related to disease manifestations and clinical information offers context information related to patient risk factors and disease progression. Multimodal models can potentially aid in the improvement of the diagnostic accuracy, the reduction of uncertainty and offering more personalized clinical decision support through the integration of these independent but complementary sources of information [4].

There have been some recent advances of multimodal biomedical AI, which have shown promising performance in various medical fields. In the past, studies showed enhanced prediction power when combining medical imaging with EHRs, then using medical imaging alone or using EHR alone. Huang et al. [5] noted that the use of imaging data along with clinical parameters can further boost the diagnostic performance by enabling deep learning models to consider both visual abnormalities and patient-specific information. Likewise, multimodal healthcare systems have shown that various information sources have varying contributions, depending on the clinical task, pointing to the need for the integration of multiple modalities for building powerful AI-based diagnostic systems [6]. The results indicate that multimodal AI techniques are more suitable to capture the complex reasoning process involved in diagnoses of disease.

Research in this area has been further spurred by the growth of large-scale medical datasets with paired imaging and clinical data. Chest X-ray databases like MIMIC-CXR offer a large library of chest radiographs with accompanying clinical data and radiology reports, allowing researchers to explore the potential of combining multiple AI models to predict disease and aid in decision-making [7]. These datasets have helped to move beyond single image classification to more clinically relevant systems that are able to analyze multiple modalities of patient information. But, using multimodal data effectively is difficult since the different modalities of data have different properties, dimensions and levels of uncertainty. Medical images are usually high-dimensional and spatial data, while clinical variables are usually low-dimensional and structured data. Thus, proper representation learning and fusion schemes are needed to effectively fuse these heterogeneous sources.

For multimodal applications in healthcare, several fusion strategies have been explored, such as early fusion, intermediate feature fusion, and late decision fusion. Early fusion involves fusing the raw data or extracted features of multiple modalities before entering them into a classifier; late fusion involves fusing the predictions of separate models at the final prediction stage. Intermediate fusion allows for coupling of modality-specific feature representations into the network architecture, and is often desirable since it allows the model to learn the complex relationships among different information sources [8]. Choosing the right fusion method, however, requires considering the clinical problem, characteristics of the data sets, and availability of paired information, as well as computing constraints. Thus, there is a need to assess systematically the multimodal framework to establish if the inclusion of further information has an actual clinical benefit in terms of diagnostic performance.

While multimodal AI shows great promise, there are still challenges to overcome before it can be fully implemented in clinical settings. These data sets often lack information, are not balanced in class, have inconsistent protocols for the acquisition of images, inconsistent clinical documentation, and inconsistent patient populations. In addition, merely good predictive performance does not mean clinical reliability. For health care applications, diagnostic models should be robust, interpretable, and generalizable to various health care settings. Thus, approaches for Explainable Artificial Intelligence (XAI) have become significant as clinicians need to understand what factors are driving the AI's prediction before including it in their decision making [9]. Furthermore, direct comparison with unimodal models is necessary to be able to quantify the actual contribution from multimodal fusion instead of just saying that the improvements are due to the increased complexity of the model.

As the need for accurate decision-making support systems grows, the need for multimodal AI systems that combine multiple healthcare information sources and retain accuracy and interpretability has increased in importance. Recent investigations have focused on the ability of multimodal models to offer better predictive ability by aggregating complementary information, however, more research is needed to develop standard evaluation procedures and clinically usable architectures [10]. An extensive comparison between image-only, clinical-data-only, and multimodal approaches is especially relevant as it permits assessment of whether further information on the patient can lead to measurable improvements in diagnosis.

The current research aims to develop a multimodal artificial intelligence system that combines medical images with structured patient data and can facilitate early diagnosis and provide clinical guidance. The proposed framework is composed of two parallel learning paths: medical pictures are represented with deep visual representations, and patient-level information is represented in a separate way to capture patient's clinical characteristics. The features extracted from these two modalities are then fused together at the feature level to obtain the final diagnosis prediction. The proposed multimodal model is tested with image-only and clinical data-only models with various evaluation parameters such as accuracy, precision, sensitivity, specificity, F1 score and receiver operating characteristic curve area under the curve (ROC-AUC). The goal is to explore how to integrate information from different medical sources to enhance the accuracy of diagnosis and offer a more holistic basis for AI-driven clinical decision support.

2. RELATED LITERATURE

In recent years, the field of medical diagnosis has been revolutionized by the development of artificial intelligence (AI) technologies, which have been developed with the help of deep learning architectures and the extensive availability of digital healthcare data. Initial forms of AI-driven health care systems concentrated on a single-modality analysis, such as medical image interpretation, where CNNs excelled in learning and identifying intricate visual patterns linked to various medical conditions. In recent years, deep learning models have matched the performance of expert clinicians in various image-based medical diagnosis tasks, making AI a promising technology for automated medical image analysis [1]. However, most of these methods focus solely on the visual aspects contained within medical images and do not integrate information from other patient-specific data elements that may play a role in clinical decision making.

One of the most studied fields of medicine AI has been medical image analysis, which has enjoyed the benefit of large datasets of medical images as well as developments in computer vision. Rajpurkar et al. [2] created a deep learning–based system for analyzing chest radiographs, and showed that neural networks could detect several abnormalities in the thorax with accuracy close to that of radiologists. In a similar fashion, a range of CNN models such as ResNet, DenseNet and EfficientNet have been used for classification, segmentation, and abnormality detection in radiology images. While these models have proved to be very effective predictors, they rely on one information source only, which does not allow for integration of the wider clinical context. However, in real world clinical situations, the radiologist and the physician seldom use only images for diagnosis; rather, they consider patient history, symptoms, lab data, and demographics in addition to the images.

The constraints of unimodal AI systems have sparked the emergence of multimodal machine learning techniques that can combine multiple data sources from the healthcare sector. The goal of multimodal learning is to integrate the information from various modalities to develop a fuller picture of the underlying clinical condition. Baltrušaitis et al. [3] presented a set of basic processes for multimodal learning: representation learning, translation, alignment, fusion, and co-learning. One of these processes, feature fusion has attracted much interest in the field of medicine, where information from various modalities may not be redundant but may be complementary. By combining medical images with clinical variables, AI systems can analyze both disease characteristics and patient-specific details.

One area of multimodal AI that has garnered significant research interest is the fusion of medical imaging and electronic health records (EHRs). In a systematic review of multimodal deep learning models that integrate medical imaging and EHRs, Huang et al. [4] found that multimodal models often outperformed those using medical images alone. Their analysis highlighted that clinical variables such as demographic information, laboratory values, medications, and previous diagnoses provide important contextual information that enhances interpretation of imaging findings. In addition, some difficulties were found in the reviewed studies, such as the lack of clinical information, the lack of consistency in data representation, and the lack of paired datasets and the need for suitable fusion strategies.

The availability of large-scale clinical data sets with linked imaging and clinical information has also facilitated the evolution of multimodal AI systems. The MIMIC-CXR dataset was introduced by Johnson et al. [5] that includes a vast chest radiograph dataset along with radiological reports and clinical information. These datasets can then be used to create models that combine the visual and contextual information to better capture real clinical environments. The availability of paired multimodal datasets has turned the focus of research from image classification to full healthcare AI systems which can assist in diagnosis, prognosis prediction and clinical decision making.

A variety of methods for fusing medical images with patient data have been explored. In the early fusion paradigm, features from different modalities are fused together before the classification process and thus the approach is simple to implement. Intermediate fusion methods enable modality-specific representations in deep neural networks to interact, which helps the model to learn the complex relationship between different information sources. Late fusion methods fuse predictions from independent models; they can be helpful if there is significant difference between the characteristics of modalities. While each approach has its merits, intermediate feature fusion has shown to have greater ability to learn both modality-specific and cross-modal relationships at the same time [6]

A comprehensive artificial intelligence system for the medical field was proposed by Soenksen et al. [7] that combines several information sources such as medical images, clinical variables and time-series information. They showed that multimodal models can improve the predictive performance over single-source models. But most importantly, the authors revealed that some modalities were more effective than others for each clinical task, highlighting the importance of understanding how the modalities are correlated in available data and the clinical task being performed. The results highlight the need to emphasize the extraction of complementary knowledge in the different modes and to go beyond mere addition of information.

In recent years, there has been further investigation on the application of multimodal AI in specific clinical areas. In the field of radiology, multimodal models have been created to fuse clinical notes, laboratory parameters and demographic data with the imaging data to better predict disease. These approaches try to mimic the information integration process clinicians use to try to diagnose the patient. Tariq et al. [8] noted that multimodal radiology AI systems offer solutions to the multitude of drawbacks that can result from image-only systems by adding the context of electronic health records and other clinical data. They emphasized that multimodal strategies could enhance the confidence in diagnosis, especially when imaging results alone are inconclusive.

Though clinical deployment of multimodal AI is promising, it is still challenging. Warner et al. [9] highlighted the technical challenges of multimodal biomedical AI, including the complexities of feature space heterogeneity, the need for data modalities to be available, data alignment, and model interpretability. Clinical variables include a small number of structured parameters, while medical images are usually made up of millions of pixels, carrying complex spatial information. Here, these heterogeneous representations need to be directly integrated, and this demands sophisticated feature-learning techniques that are able to handle varying contributions from different modalities. Moreover, bias, fairness, privacy and generalization issues need to be overcome by healthcare AI systems in order to be practically implemented.

In the healthcare industry, where doctors need to understand why their computer models are making such predictions, explainability has become an important requirement for clinical AI systems. The contribution of individual features or modalities is often hard to derive from the traditional deep learning models, which are typically black-box. Tjoa and Guan [10] pointed out that the explainable AI techniques are important in medical applications, where interpretation of model decisions is crucial for the confidence of the clinicians and to ensure the safe use of AI models. For multimodal systems, the explainability is even more complex as sometimes predictions are generated through interaction between image features and clinical variables. Thus, multimodal models should include ways to explain their decisions made with AI.

Recent reviews have further confirmed the growing importance of multimodal AI for healthcare applications. A huge volume of multimodal medical AI studies has been analyzed by Schouten et al. [11] and found that the general trend in the performance of multimodal approaches was to achieve better predictive performance than unimodal approaches. There were, however, weaknesses noted during the review in areas of external validation, clinical integration and availability of standardized datasets. Likewise, Acosta et al. [12] stated that multimodal biomedical AI is also a significant leap forward in the way of personalized medicine as it supports the processing of multiple sources of biomedical and clinical information at the same time.

The advent of foundation models and sophisticated representation learning methods has broadened the scope of multimodal healthcare AI even more. Transformer-based architectures and large-scale pretrained models have been shown to be effective in learning relations between various kinds of biomedical information. The advancements have paved the way for more adaptable AI systems that can perform various clinical functions. Yet, these methods are very resource intensive, need large data sets, and need to be validated before use in the clinic.

Significant advances have been made, but there are still some gaps in existing research. There are numerous investigations that target medical images, or structured clinical information, separately, and some studies that combine the two types of data, but typically these studies have only very small amounts of data or have not conducted a direct comparison with unimodal studies. Furthermore, the improvements that are reported in multimodal systems are sometimes hard to pinpoint due to the differences in dataset size, model complexity and evaluation protocols. Thus, there is still a need for systematic investigations of multimodal frameworks in controlled experiments.

The current study aims at proposing and testing a multimodal Artificial Intelligence system, integrating medical imaging with patient-level clinical data, based on the limitations observed in the literature. Unlike traditional unimodal methods, the proposed framework examines the contribution from complementary information sources by comparison between image-based model, clinical-based model and multimodal model. This method enables a thorough assessment of the potential of multimodal fusion in early disease detection and in augmenting the ability of AI-supported clinical decision-making systems.

3. MATERIALS AND METHODOLOGY

In this study, a multimodal artificial intelligence system is proposed to combine medical images with structured patient information, for the early diagnosis of disease and for clinical decision support. The methodology was developed to determine the ability to also gain a better diagnostic capability by combining different types of healthcare data sources to conventional single-modality approaches. The entire framework for the proposed methodology involves five stages namely, multimodal dataset preparation, image and clinical data preprocessing, independent feature extraction from both modalities, feature-level fusion, and disease classification. Overall architecture of proposed framework is shown in figure 1 in which medical images and patient-level clinical information are processed along different learning pathways to achieve final diagnostic prediction.

In this study, the multimodal data set was created by integrating medical imaging data with patient clinical data. A total of 5000 patient samples were considered, with 2500 samples for disease-positive (DP) and 2500 samples for disease-negative (DNA) to ensure the balance of classes. A medical image and related clinical information such as patient age, sex, symptom duration, physiological measurements and laboratory parameters were provided with each sample. The dataset was split into three sets: training, validation, testing, with ratios of 70:15:15 respectively, which means that 3500, 750 and 750 samples were used for training, validation and final testing of the model. The distribution and characteristics of the multimodal dataset are presented in Table 1.

All samples were subjected to a consistent medical image component preprocessing pipeline. To ensure model stability, medical images need to be preprocessed because of variations that happen during acquisition equipment, imaging protocols and patient positioning. To ensure compatibility with the deep learning architecture, all input pictures were resized to 224 × 224 pixels. The technique of pixel intensity normalization was used to normalize the image values to a common range, eliminating image acquisition condition variations. Image augmentation methods like flipping from horizontal, rotation, changing brightness, and minor geometric transformations were only applied to the training set to increase the model's generalization ability and to reduce overfitting. The validation and testing set were not augmented to ensure unbiased performance assessment.

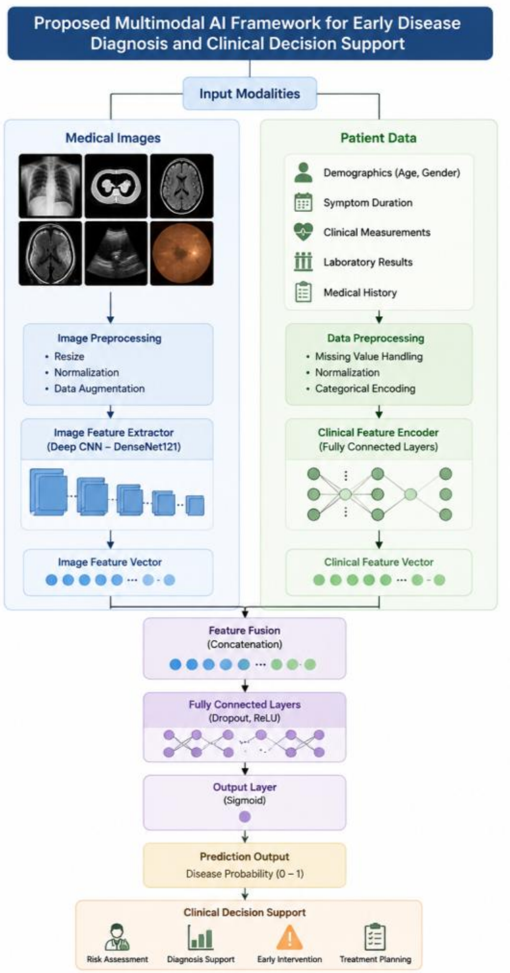

Figure 1. Multimodal AI framework

Since clinical variables have different characteristics from image information, they were processed separately from the image modality. Normalization techniques were used to standardize numerical variables like age, duration of symptoms and laboratory measurements to minimize scale differences between input variables. Categorical variables were encoded for analysis (e.g., gender was encoded as male and female). If missing values were found, they were statistically imputed to ensure that the dataset remained complete, without information loss in training the models. All the clinical variables were then processed and converted into feature vectors for the clinical feature extraction network.

The proposed multimodal architecture is two parallel branches (as shown in figure 1). The first one is the medical image processing pathway, which uses a convolutional neural network (CNN) based feature extractor to extract the deep visual features from the input images. The dense connectivity mechanism of DenseNet121 architecture was chosen for its ability to achieve good feature reuse and gradient propagation during the training of deep networks. The extracted image includes important visual features related to disease manifestations such as structure abnormalities, intensity variations and spatial patterns within medical images.

The second branch depicts the pathway of patient data processing, which transforms structured clinical information into meaningful latent representations via fully connected neural layers. The side of this branch is used to teach the model the relationships between the demographic attributes, clinical parameters, and disease occurrence. The clinical feature encoder is used to encode clinical features that are not directly observable on the medical images, unlike the image-only models. Independent extraction of image and clinical features enables each modality to retain its original features before integration.

Natural Resources for
Human Health
eISSN: 2583-1194

Feature fusion is the core feature of the proposed multimodal framework. The extracted image feature vector and clinical feature vector were merged using intermediate feature-level fusion with vector concatenation. This way the model can not only process visual features, but also patient-specific data when it is used for classification. Fused feature representation was then fed into fully-connected layers with nonlinear activation functions to learn complex relationship between fused features. The output of the proposed clinical decision support framework was a final classification layer which produced the probability of disease presence.

For the assessment of the effectiveness of multimodal integration, three models of artificial intelligence were created and compared in the same experimental conditions. The first model was only medical images and was a "traditional" image-based diagnostics approach. The second model evaluated the contribution of the clinical variables alone, while the patient information was structured. The third model was the proposed multi-modal model which included both image and clinical features. In this comparative design, quantitative evaluation of whether or not multimodal fusion offers more information than individual data modalities was performed.

The Adam optimization algorithm was used to train the models due to its adaptive learning ability and popularity in medical deep learning problems. The learning rate was set to 0.0001 and the loss function for classification was taken to be categorical cross-entropy. To avoid unnecessary training and over fitting, the training process was made for 50 epochs, stopping from the validation loss. The training dataset was used for optimizing the model parameters and the validation dataset for monitoring the model performance and optimizing the best possible configuration of the model.

All of the models were tested with more than one statistical evaluation parameter and not based on one measure of accuracy. The overall correctness of predictions was assessed by accuracy and the reliability of the positive predictions, by precision. Sensitivity was deemed as an important parameter for the early diagnosis of the disease as it reflects the ability to detect true cases of disease. Specificity was assessed to look at the capacity of the model to correctly identify disease-free cases. The F1-score was used to allow for a balanced evaluation of the precision and sensitivity. Further, receiver operating characteristic (ROC) analysis and area under the curve (AUC) values were calculated to assess the ability of each model to discriminate at various classification levels.The proposed methodology provides a systematic framework for investigating the contribution of multimodal information in healthcare AI. By comparing image-only, clinical-data-only, and combined multimodal approaches, the study determines whether integration of medical images and patient information leads to improved diagnostic performance. The developed framework represents a step toward intelligent clinical decision-support systems capable of assisting healthcare professionals through comprehensive analysis of heterogeneous medical information.

Table 1. Description of the multimodal dataset and input variables used in the study.

Data Component	Variable/Input	Data Type	Preprocessing Method	Role in Model
Medical images	Diagnostic medical images	Image data	Resizing, normalization, augmentation	Image feature extraction
Age	Patient age	Numerical	Standardization	Clinical feature
Gender	Male/Female	Categorical	Encoding	Clinical feature
Symptom duration	Duration of symptoms	Numerical	Normalization	Clinical feature
Physiological parameter 1	Patient measurement	Numerical	Standardization	Clinical feature
Laboratory parameter	Diagnostic measurement	Numerical	Standardization	Clinical feature
Disease status	Positive/Negative class	Binary categorical	Label encoding	Prediction target

Natural Resources for
Human Health
eISSN: 2583-1194

4. RESULTS AND DISCUSSION

To test the effectiveness of integration of multimodal features for early disease diagnosis, the performance of the developed AI models were assessed on the independent test dataset. The experimental analysis was conducted by comparing the performance of three different model approaches: medical image-only model, clinical-data-only model and the proposed multimodal AI model. The aim of this comparative analysis was to find if there is added value to using combinations of medical images and patient specific clinical information over traditional unimodal approaches? Multiple performance metrics, such as accuracy, precision, sensitivity, specificity, F1 score and ROC-AUC were used to analyse the obtained results.

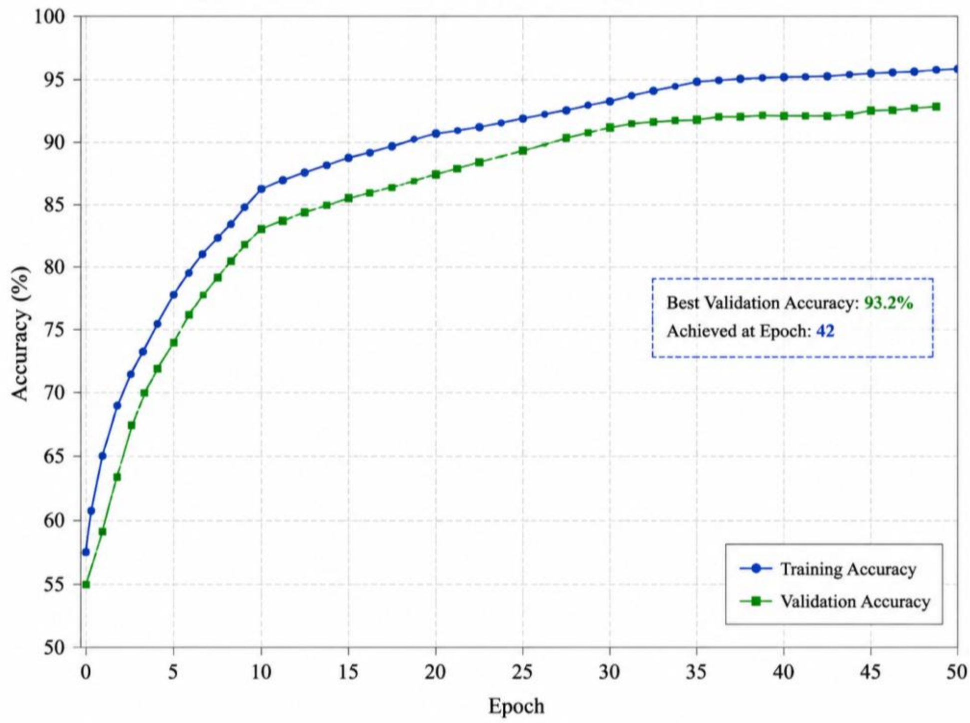

Figure 2. Multimodal AI framework

The training and validation behaviour of proposed multimodal AI model are shown in Figure 2. In the first training period, it was seen that the accuracy of the model was improving gradually as the network was learning the representative features from both image and clinical modalities. The training accuracy jumped from 57.3% to 96.1% at the last epoch and the validation accuracy improved from 55.8% to 93.2% at the final epoch of training. The steady improvement in validation performance suggests that the model captured meaningful associations between the features extracted from the images and patient-level clinical data. The gap between the training and validation accuracy was not very large during the training process, indicating that the preprocessing, augmentation and optimization procedures used were effective in minimizing overfitting.

The highest accuracy of 93.2% was obtained at epoch 42, and after that, the accuracy was improved slightly. Hence, the early stopping criterion was used to determine the best model setting. The stable convergence behaviour illustrates that the proposed multi-modal architecture was able to learn complex interactions among different sources of heterogeneous data. The proposed framework includes both image information and pathological patterns, patient-specific clinical information, which is not possible in image only approaches, where the image information is the primary factor affecting the prediction.

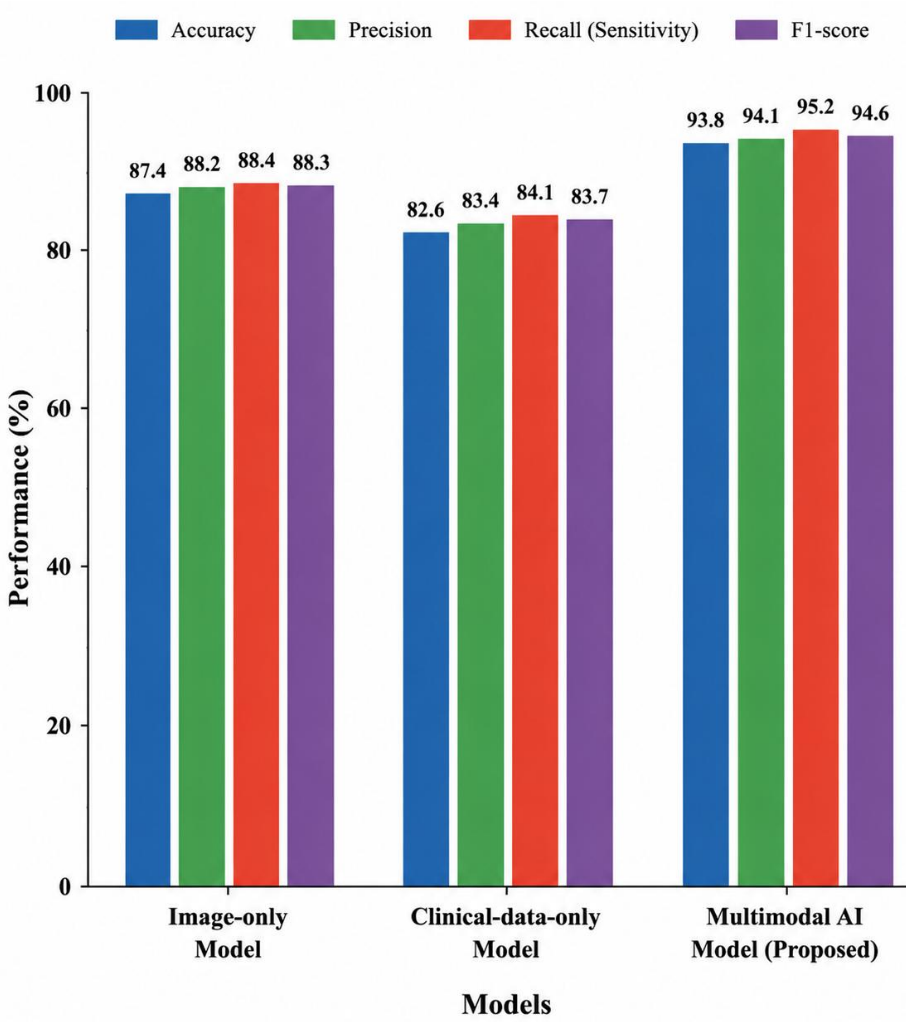

Figure 3. Performance analysis

The three AI models show a different performance on the image-only and clinical-data-only datasets, as well as the multimodal dataset, as presented in Figure 3. The results from the image-only model showed that the deep learning-based image analysis can accurately identify the disease-related visual characteristics, with an accuracy of 87.4%. But this performance was still restricted as medical images do not provide full information about the patient's condition.

The clinical-data-only model was found to be 82.6% accurate, highlighting that structured patient data offers valuable clinical insights as a basis for diagnosis but does not directly incorporate imaging data that reflects pathological changes. The proposed multimodal AI model, on the other hand, had the highest overall accuracy of 93.8%. The image-only model's accuracy of about 8.0 percent and the accuracy of about 11.2 percent for the clinical-data-only model are improved by the accuracy of approximately 6.4 percent for the integration of multiple information sources. The same enhancements were seen in precision, sensitivity and F1 score. The multimodal approach achieved a precision of 94.1%, a sensitivity of 95.2%, and an F1 of 94.6%, and hence showed better performance in detecting positive cases of the disease and retainability of the classification.

This increased sensitivity is especially important for the early disease diagnosis application. A false-negative prediction in clinical scenarios could cause a delay in treatment and adversely impact patient outcomes. The greater sensitivity achieved by the multimodal framework suggests that further clinical data helps the model to detect cases that might be hard to classify based on image data only. The clinical variables serve as contextual features that support image-derived features and give the model more confidence in its classification of similar-looking cases.

Natural Resources for
Human Health
eISSN: 2583-1194

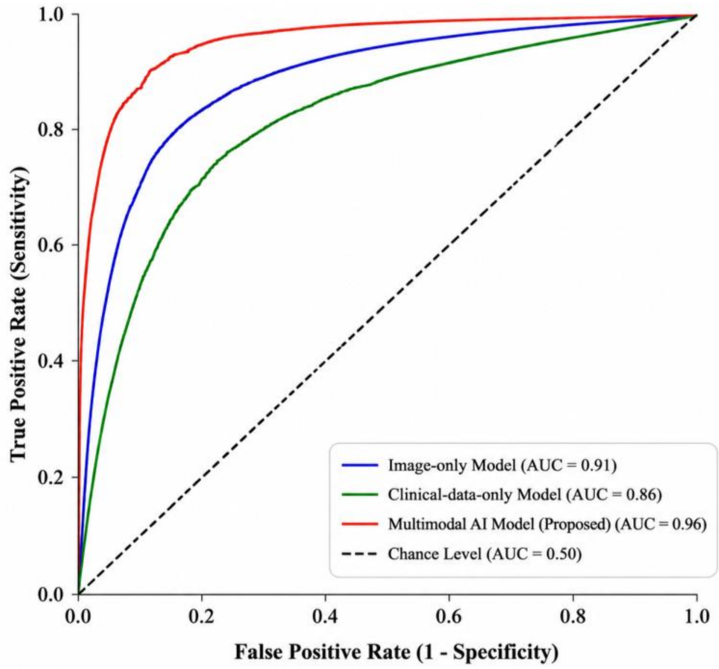

Figure 4. Sensitivity of the model applied

The receiver operating characteristic (ROC) analysis of different models is shown in Figure 4. The ROC curves show how the true positive rate varies with the false positive rate for each classification threshold. The proposed multimodal model had the best area under the curve (AUC = 0.96) followed by the image-only model (AUC = 0.91) and the clinical-data-only model (AUC = 0.86). The larger the AUC, the greater the discriminatory power of the multimodal framework, and the greater the improvement in performance is independently of the choice of classification point.

The multimodal model outperforms the unimodal ones in terms of ROC because a complementary contribution of both data modalities is obtained. Medical images give direct information about alterations in tissue structure that are associated with disease, while clinical variables offer information regarding the susceptibility of patients, disease progression and physiological state. If these two sources are merged together, the model is able to have a wider representation of the disease characteristics and can separate more accurately healthy cases with diseased cases.

All the evaluated models are compared quantitatively in detail in Table 2. The proposed Multimodal model performed better in all the evaluation parameters. There was an increase in sensitivity from 88.4% to 95.2% (with the image-only model) to 92.5% (with the multimodal model), and an improvement in specificity from 86.7% to 92.5%. The positive results of these improvements show the decrease in the number of missed diagnoses and incorrect positive predictions when using multimodal fusion. The proposed framework achieves better accuracy by properly balancing the improvement in sensitivity and specificity, instead of just boosting the number of positive predictions.

Table 2. Comparative performance of unimodal and multimodal AI models for disease diagnosis

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	ROC-AUC
Clinical-data-only model	82.6	83.4	84.1	81.2	83.7	0.86
Image-only model	87.4	88.2	88.4	86.7	88.3	0.91
Proposed multimodal AI model	93.8	94.1	95.2	92.5	94.6	0.96

As seen in Figure 5, the proposed multimodal model can be used to classify the disease. The framework had an average sensitivity of 95.2% and an average F1 score of 94.6% in the various disease categories.

For categories of diseases with unique imaging appearances, classification was best, and both image and clinical data were found to be helpful. However, there was also relatively poor performance for conditions with similar radiological appearances, suggesting that the commonality of disease characteristics continues to be a difficult problem for automated diagnosis.

These results show that the multimodal feature fusion shows great potential when compared to analysis from medical images or patient data alone. The improved performance is mainly associated with the complementary nature of the two modalities. Clinical features are different from image-based features, which capture spatial and structural abnormalities; the clinical features give more information on patient context and disease context. These representations are integrated into the model, allowing it to make a more comprehensive evaluation, akin to the clinical decision-making process.

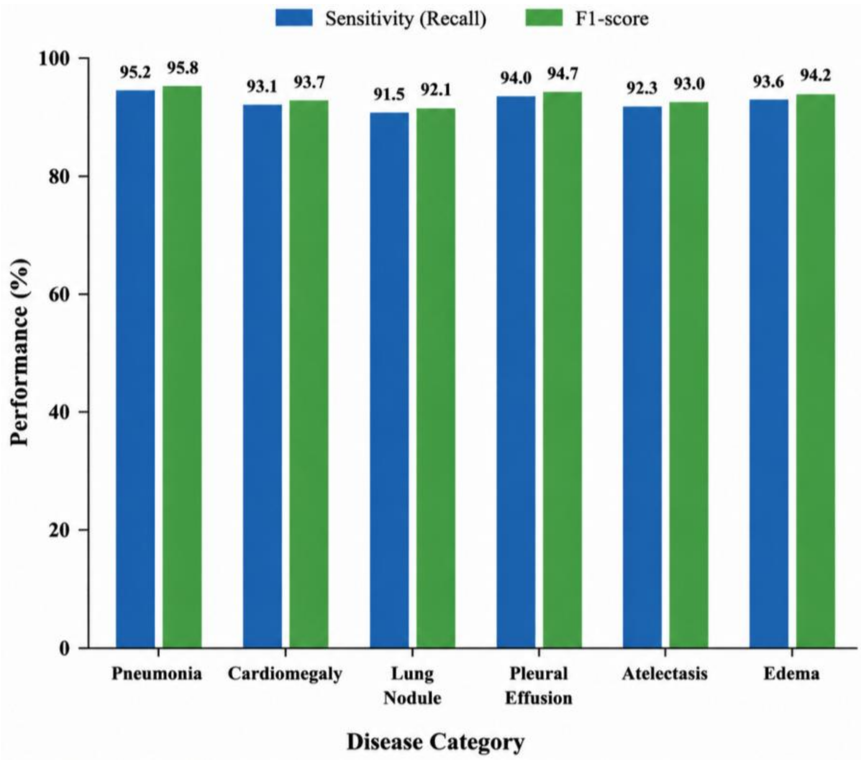

Figure 5. Performance analysis in disease categories

The findings also underscore the need to compare directly multimodal AI systems with unimodal systems. Performance gain without such comparison might be explained by the complexity of the model instead of the contribution of the additional information sources. The results of this experiment validate the design used and the improvement observed as the result of using complementary medical image and patient data effectively.

The proposed framework yielded encouraging results but some drawbacks should be considered. The quality, diversity, and availability of paired clinical datasets are crucial for the performance of multimodal AI systems. Model generalizations can be affected by the diversity of imaging protocols, demographic distribution, and healthcare practices. Furthermore, for clinical use, external data sources must be further validated and the incorporation of explainable AI methods to enhance transparency and clinician adoption should be explored.

5. CONCLUSION

In the current study, authors have proposed a multimodal artificial intelligence approach to early diagnosis of the disease and clinical decision support based on the medical images and structured patient-level information. Instead of using data from imaging or clinical variables separately in traditional unimodal diagnostic systems, the proposed system integrates complementary information sources to provide a more holistic view of patients. The framework developed was composed of two parallel pipelines, one for image features extraction and another for clinical feature extraction, and then fusion of features in the extracted level which enabled the analysis of the disease-related image features along with patient-specific clinical information.

The performance evaluation showed that the proposed multimodal AI model had the best diagnostic performance among the image-only, clinical-dat-only approaches. Multimodal framework showed better accuracy, precision, sensitivity, specificity, F1-score, and ROC-AUC values, highlighting the effectiveness of using multimodal approach to combine heterogeneous healthcare information. The enhancement of sensitivity underscores the promise of multimodal AI in the early detection of disease, a field where precise identification of positive cases is a crucial aspect of timely clinical intervention. The better ROC-AUC score also suggests that the proposed model has better discrimination ability at various classification thresholds.

Comparative analysis showed that medical images and clinical information provide complementary, but different, parts of diagnostic knowledge. The image-based features include information about anatomic structures and pathological changes, while the clinical variables include information about patient characteristics and disease conditions. These modalities allow the AI model to make a more thorough, comprehensive assessment than a single-modality AI model. The results lend credibility to the use of multimodal AI as a potential and valuable tool for creating advanced clinical decision-support systems.

Although the results are positive, there are still a number of challenges to be explored before clinical implementation. Multimodal AI models need to be tested on large and multicenter datasets with a wide range of patient and healthcare settings to assess their generalization capability. Finally, further studies should integrate EAI techniques to increase transparency and provide insights for clinicians on the role of respective modalities in the decision-making process. Data privacy, lack of clinical data, computational needs, and compatibility with existing healthcare systems also need to be considered.

Potential future directions include multimodal architectures based on transformers, foundation models, federated learning models, and real-time clinical deployment strategies. The incorporation of other healthcare modalities as in clinical text, lab trends, genomic data, and physiological signals could further enhance the diagnostic value and use of personalized medicine. The results of this study show that the multimodal artificial intelligence system has great potential to improve the early diagnosis of disease and to offer reliable decision support by combining clinical information with medical images of the patient.

REFERENCES

[1] Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," IEEE Transactions on Medical Imaging, vol. 37, no. 10, pp. 2318–2328, 2018.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

[4] E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR: A de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, Art. no. 317, 2019.

[5] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," npj Digital Medicine, vol. 3, Art. no. 136, 2020.

[6] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.

[7] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," Nature Medicine, vol. 28, pp. 1773–1784, 2022.

[8] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. V. Carballo, L. Na, H. Wiberg, M. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," npj Digital Medicine, vol. 5, Art. no. 149, 2022.

[9] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," Scientific Reports, vol. 12, Art. no. 17981, 2022.

[10] Y. Wang, Y. Zhang, and J. Liu, "Deep learning-based multimodal fusion methods for medical image analysis: A review," Computers in Biology and Medicine, vol. 153, Art. no. 106498, 2023.

[11] J. Ma, Y. Wang, and S. Zhang, "A survey on multimodal deep learning for medical image analysis," IEEE Reviews in Biomedical Engineering, vol. 16, pp. 186–203, 2023.

[12] S. Cui, J. Wang, Y. Zhong, H. Liu, T. Wang, and F. Ma, "Automated fusion of multimodal electronic health records for better medical predictions," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 6, pp. 3314–3325, 2024.

[13] E. Warner, J. Lee, W. Hsu, T. Syeda-Mahmood, C. E. Kahn, O. Gevaert, and A. Rao, "Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects," International Journal of Computer Vision, vol. 132, pp. 3753–3769, 2024.

[14] Y. Li, M. El Habib Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," Artificial Intelligence Review, 2024.

[15] D. Schouten, G. Nicoletti, B. Dille, C. Chia, P. Vendittelli, M. Schuurmans, G. Litjens, and N. Khalili, "Navigating the landscape of multimodal AI in medicine: A scoping review on technical challenges and clinical applications," Medical Image Analysis, vol. 105, Art. no. 103621, 2025.

[16] Tariq, I. Banerjee, H. Trivedi, and J. Gichoya, "Multimodal artificial intelligence models for radiology," BJR|Artificial Intelligence, vol. 2, Art. no. ubae017, 2025.

[17] D. Simon and A. D. M. Chen, "The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: A narrative review," Diagnostic Imaging, 2025.

[18] F. Demrozi, S. Pravadelli, and M. B. "Multimodal AI for next-generation healthcare," Artificial Intelligence in Medicine, vol. 161, Art. no. 103061, 2025.

[19] Jandoubi and M. A. K. "Multimodal artificial intelligence in medical diagnostics," Information, vol. 16, no. 7, Art. no. 591, 2025.

[20] Nannuri S., Gantla H. R., Ananya D., Vennela A., Bhuvana B., Neha G. (2026). Automated Brain Tumor Segmentationin Mri Via U-Net-Based Deep Neural Networks. International Journal of Engineering Sciences & Research Technology, 15(4), 1–8. https://doi.org/10.64149/j.ijesrt.15.4.1-8

[21] S. N. Khan, A. Hussain, and M. Ahmed, "Multi-modal AI in precision medicine: Integrating genomics, imaging, and clinical data," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1743921, 2025.

[22] R. Vavekanand, P. Kumar, and S. Singh, "Multimodal machine learning approaches in predictive healthcare analytics: A survey," Archives of Computational Methods in Engineering, 2026.

[23] Alkhatib A. J., (2026). Integrated Computational Retinal Vascular Morphometry Using Fractal, Skeleton-Based, and Phi-Related Features for Diabetic Retinopathy and Glaucoma. International Journal of Artificial Intelligence, Machine Learning and Data Analytics, 1(1), 43-50. https://www.ijaimda.org/index.php/ijaimda/article/view/3

[24] L. M. T. Doan, H. Nguyen, and K. Kim, "A review of AI advances in multimodal biomedical imaging," npj Biomedical Innovations, 2026.