International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 650-671
https://doi.org/10.70917/ijcisim-2026-5630
Article

An Explainable Hybrid Multimodal Learning Framework for Heart Disease Prediction Using Clinical and Imaging Data

Sameer Tembhurney[1], Shital Shende[2], Asakti Rautkar[3], Sneha Patle[4], Ravi Mangalsingh Thakur[5], Poonam Vishwas Meghare[6], Mikhal John[7] , Sonam Meshram[8]

[12345678]Department of Computer Science and Engineering, Ramdeobaba University (RBU), Nagpur, Maharashtra, India.
Email id : tembhurneysb@rbunagpur.in , spshende111@gmail.com , asakti.sg@gmail.com , patlesk_1@rbunagpur.in
thakursinghravi786@gmail.com, megharepv@rknec.edu, mikhaljohn8@gmail.com , meshramsu@rbunagpur.in

Abstract: The increasingly sophisticated nature of modern-day healthcare requires intelligent solutions that will help merge several types of datasets and produce more accurate disease prediction results. Conventional approaches that employ only either structured medical data or images ignore numerous essential factors that contribute to better understanding of the patient's condition. Therefore, this paper offers a new solution that allows integrating multiple technologies into the process of disease prediction. Machine learning, deep learning, as well as explainable AI, are incorporated into the system, which includes XGBoost model that analyzes structured data from the UCI Heart Disease dataset and an EfficientNetB0 neural network designed to classify skin lesions with the help of the HAM10000 dataset. An additional layer of explainability is provided by SHAP analysis of the clinical model and the utilization of Grad-CAM in order to visualize imaging results. Moreover, a simple technique that allows integrating results obtained with both approaches in one decision-making process is discussed. Clinical model achieves 88.52% accuracy and AUC equals 0.94, whereas the accuracy of the imaging model makes 76.88%.It ensures that decisions are made based on sound judgment through the utilization of the data from both approaches. It is important to note that although the data from both models is not linked to the same patients, the model demonstrates how AI multi-models can easily scale and be explained.
Keywords: Multimodal Learning, Explainable AI (XAI), Heart Disease Prediction, Skin Lesion Classification, Efficient-NetB0, XGBoost, Healthcare AI, Deep Learning

1. Introduction

[1]–[6]. However, these methods often work alone, which means they don't make full use of all the available information and may not provide the best possible diagnostic results. [1]–[6]. However, these methods often work alone, which means they don't make full use of all the available information and may not provide the best possible diagnostic results. [7]–[9]. In the last few years, multimodal learning has been identified as a potential solution to integrate diverse data sources that would help in making informed decisions [7], [10]–[12]. In the context of disease prediction, integrating clinical data sources such as medical history, physical measurements, and laboratory reports with imaging data sources like der-matoscopy or radiography provides valuable inputs [10], [13]–[15]. Developing a model that supports multiple data sources is challenging owing to issues such as varying data sources formats, unequal data availability across sources, feature ag-gregation complexities. [7], [8], [11], [16], [17].

Another critical challenge in modern healthcare AI systems is the lack of transparency. Black-box models, particularly deep neural networks, often fail to provide explanations for their predictions, limiting their adoption in clinical settings where trust and interpretability are essential [18]–[21]. Explainable artificial intelligence (XAI) techniques, such as SHAP and Grad-CAM, have been introduced to address this issue; however, their integration within multimodal frameworks is still limited and requires further investigation [18], [19], [21].

To tackle these challenges, this paper introduces a hybrid ensemble deep learning framework that combines structured clinical data with medical imaging data, along with explainability features. The framework uses an XGBoost-based ensemble model to predict heart disease using the UCI Heart Disease dataset [1], [2], and an EfficientNetB0-based convolutional neural network for classifying different types of skin lesions using the HAM10000 dataset [3], [4], [22]. Additionally, a fusion-based risk scoring method is introduced to merge predictions from both data types into a single decision-making system [7], [10]. The goal of this approach is to boost the accuracy of predictions while keeping the model interpretable, making it more practical for real-world clinical use. [18], [21]. The efficiency of the suggested framework has been demonstrated with the help of numerous tests. The clinical model operates with the accuracy level of 88.52% and the AUC value of 0.94, which means that its performance when working with structured data is impressive. As for the imaging model, it demonstrates 76.88% accuracy while working with the multiclass HAM10000 database. Thus, it can be regarded as highly efficient in recognizing key features from the images.

The main contributions of this work are summarized as follows:

Proposes a new multimodal disease prediction method that involves the use of structured clinical and medical imaging data [7], [10], [11].

Ensemble learning based on XGBoost to predict the occurrence of heart disease using tabular data [1], [2].

Deep learning-based image classification model for multiclass skin lesion detection using the EfficientNetB0 architecture [3], [4], [22].

Inclusion of the application of explainable AI, specifically SHAP for tabular data and Grad-CAM for images, to enhance interpretability of the models used [18], [19], [21].

Risk scoring using fusion techniques that integrate the results of various modalities for enhanced decision-making [7], [10], [13].

Comprehensive experimental evaluation demonstrating the effectiveness of the proposed approach across multiple datasets [1], [3], [4].

However, several essential caveats should be noted regarding the promising findings above. In particular, the clinical and imaging data utilized in the present work have been drawn from independent populations of subjects and not collected for the same patients, making it challenging to conduct an integrated patient-based analysis leveraging multi-modal data [7], [8]. In addition, the HAM10000 dataset employed for the skin lesion classification task suffers from an imbalanced distribution of classes, which may negatively impact the predictive performance of the model when dealing with rare ones [4], [22]. Furthermore, our approach to fusing various data types can be regarded as rather simplistic and may not be able to capture all complex dependencies inherent in the datasets under consideration. [7], [11]. The above limitations suggest potential directions for further investigation in the area of multi-modal data analysis. [10], [13].

In general, the presented approach provides a scalable and comprehensible way of disease prediction based on variousdata, which would aid in improving AI-based decision support systems in clinical settings. [18], [19], [21].

2. Literature Review

Advances in AI have had a significant effect on the health industry, particularly in disease prediction and medical imaging analysis [7], [23]–[25]. In this part, we consider the current literature in significant fields, highlighting their strengths and weaknesses and identifying the

Machine Learning for Clinical Data-Based Disease Prediction

Logistic Regression, SVM, Random Forest, and Gradient Boosting have been the traditional algorithms for disease prediction with clinical data [1]. Studies have found that ensemble models such as Random Forest and Gradient

Boosting are effective tools to detect complex patterns within clinical attributes, thus enhancing the prediction accuracy. [1], [2], [24].

XGBoost has been a trending model recently due to its capability to handle missing data, apply regularization techniques, and be highly effective on tabular data [1], [2]. Numerous researchers have proved that the XGBoost model is highly reliable and precise in predicting the risk of heart disease [2], [25], [26]. But these methods are solely based on structured data without incorporating any unstructured data like images. [7], [10], [11].

Limitation: Current clinical predictive models are confined only to tabular data and do not allow any type of multi-modal integration, resulting in incomplete diagnostic information [7], [10], [11].

Deep Learning for Medical Image Analysis

The use of deep learning algorithms, particularly Convolutional Neural Networks (CNNs), has shown high efficacy in medical imaging segmentation and classification applications [3], [4], [17], [27], [28]. Some of these deep learning algorithms include ResNet. [3], [4], [20], [27], [29].

The HAM10000 database has been widely used as a benchmark for evaluating skin lesion classification models. [22]. Using transfer learning with pre-trained networks such as EfficientNetB0 has greatly boosted classification accuracy and cut down on training time [4]. However, even with these improvements, deep learning models are often seen as black boxes, which makes it hard to understand how they arrive at their predictions. [18], [19], [21].

Limitation: Image-based models lack interpretability and often struggle with class imbalance, particularly in multiclass medical datasets [4], [18], [19], [22].

Explainable Artificial Intelligence in Healthcare

Explainable Artificial Intelligence, or XAI, has become an important part of making AI models in healthcare more transparent and trustworthy [18], [19], [21]. Methods like SHAP, which stands for SHapley Additive exPlanations, are commonly used to show how different features affect predictions

TABLE I

Representative literature on multimodal disease prediction, medical imaging, and explainable AI

Ref.	**Paper**	**Domain / Data**	**Method**	**Main Contribution**	**Limitation / Relevance to Proposed Work**
[1]	Yang and Guan (2022)	Heart disease (clinical data)	SMOTE-XGBoost with feature optimization	Improves prediction on imbalanced clinical datasets	Limited to tabular data; no multimodal integration
[3]	Tajerian et al. (2023)	HAM10000 skin lesions	EfficientNet-B1 transfer learning	Achieves strong multi-class skin lesion classification	Image-only approach; lacks clinical integration
[4]	Gessert et al. (2020)	Skin lesion + metadata	Multi-resolution EfficientNet ensemble	Handles class imbalance and integrates metadata	Domain-specific; no general multimodal framework

[18]	van der Velden et al. (2022)	Medical imaging	XAI survey	Comprehensive review of XAI techniques in medical imaging		No implementation of predictive multimodal system
[8]	Pahud et al. (2024)	Multimodal imaging	Clinical XAI orchestration	Highlights need for explainable multimodal systems		Conceptual framework only
[7]	Demrozi et al. (2025)	Multimodal healthcare AI	Review of MMAI frameworks	Discusses fusion strategies and challenges in healthcare AI		No experimental validation
[22]	Tschandl et al. (2018)	HAM10000 dataset	Dataset benchmark	Provides large dermoscopic dataset		No prediction or explainability model
[2]	Ahmad et al. (2023)	Heart disease (UCI dataset)	ML models with feature selection	Achieves high accuracy (up to 94.6%)		No multimodal fusion or explainability

in models that work with tables of data. Meanwhile, Grad-CAM, which stands for Gradient-weighted Class Activation Mapping, helps provide visual insights into how CNN-based image models make their decisions. [15], [16], [18], [29].

These methods have been successfully applied in isolated domains; however, their integration within multimodal frameworks remains limited [18], [19], [21]. Furthermore, explanations generated by different modalities are often not unified, leading to fragmented interpretability [19].

Limitation: Existing XAI approaches are typically applied to single-modality systems and lack unified interpretability across multimodal architectures [7], [8], [13].

Multimodal Learning for Healthcare Applications

Multimodal learning combines data from various sources to enhance the accuracy of predictions [7], [10], [11], [13]. In healthcare, this means bringing together clinical records, medical images, genetic information, and data from sensors. Many research studies have looked into using multimodal methods for diagnosing diseases, and they have shown better results than relying on just one type of data. [10], [13], [30]. However, most existing multimodal systems need synchronized datasets, where all the different types of data come from the same patient [10], [11]. This requirement makes them less useful, since such datasets are usually hard to find. Also, the methods used to combine different data types in previous studies are often complicated and use a lot of computing power. [7], [9], [12].

Limitation: Current multimodal approaches depend on patient-level aligned datasets and often employ complex fusion mechanisms that reduce scalability [7], [11], [12].

Research Gap Analysis

- Based on the reviewed literature, the following research gaps are identified:
- Lack of frameworks that integrate clinical and imaging data without requiring patient-level alignment.
- Limited incorporation of explainable AI techniques within multimodal systems.
- Insufficient exploration of lightweight and scalable fusion strategies.
- Challenges in handling class imbalance and interpretability simultaneously.

Gap Summary Table

The summary in Table II clearly shows the shortcomings of current methods in various research areas. Clinical machine learning models work well with structured data but don't use information from imaging sources. Deep learning models for images perform well but have issues with understanding and handling imbalanced classes. Explainable AI helps with transparency but is mostly used in single-type data settings, making it less useful in complex healthcare situations. Even though multimodal learning improves predictions, these methods often depend on patient-level data and use costly ways to combine information, which makes them hard to scale and use in real-world settings.

by these findings, this study seeks to tackle the existing gaps in research by creating a combined system that handles both clinical and imaging data separately. It uses AI methods that can be understood and explained, across both types of data, and applies a simple way to combine the results. This method improves the accuracy of predictions while also making the system easier to use, more adaptable, and suitable for real-world healthcare settings with different types of data and environments.

Positioning of Proposed Work

To tackle the identified gaps, this paper suggests a hybrid multimodal framework that brings together clinical and imaging data without needing to match patients directly

TABLE II

Summary of Literature and Research Gaps

Category	Strengths	Limitations / Gaps
Clinical ML Models	High accuracy on structured data; interpretable models	Lack of multimodal integration and limited data diversity
Deep Learning Models	Strong performance in image-based tasks	Black-box nature and sensitivity to class imbalance
XAI Techniques	Enhances model transparency and interpretability	Mostly limited to single-modality systems
Multimodal Learning	Improved prediction using multiple data sources	Requires aligned datasets and complex fusion mechanisms

The framework uses a mix of ensemble learning and deep learning models along with explainable AI methods. It also presents a simple but effective approach for merging data when assessing risks. The technique ensures a good blend between effectiveness, interpretability, and scalability, which makes it suitable for healthcare applications.

3. System Model and Background Environment

The current segment describes the design of the hybrid multi-modal disease prediction system we have proposed. The system involves data from various sources and the assumptions on which its working is based.

Overview of the Proposed Framework

This framework is developed in order to take into account and integrate two types of information: the structured data collected from the patient's visit, and medical imaging data. The key advantage of this system over other multimodal frameworks is that no matching by the patient level is required for these two datasets.

The system is composed of three primary modules:

Structured Data Clinical Processing Module: This component processes structured clinical data and makes predictions about the probability of a disease using a multi-modal prediction model which is based on the XGBoost machine learning algorithm.It captures complex inter-feature relationships within table-formatted healthcare data.

Unstructured Data Imaging Processing Module: This component processes unstructured imaging data in the form of dermoscopy images and makes predictions using a CNN-based machine learning model that leverages the EfficientNetB0 neural network architecture.It learns valuable visual features and classifies skin lesions into multiple classes.

Decision-Level Fusion-Based Decision Module: This component integrates the probabilities calculated by both components through decision-level fusion into a risk score.The fusion process is based on weighted integration.

The whole system uses a parallel setup, where each type of data is handled separately and then merged together. This approach cuts down on the need for perfectly timed data sets, while still keeping the predictions strong and reliable

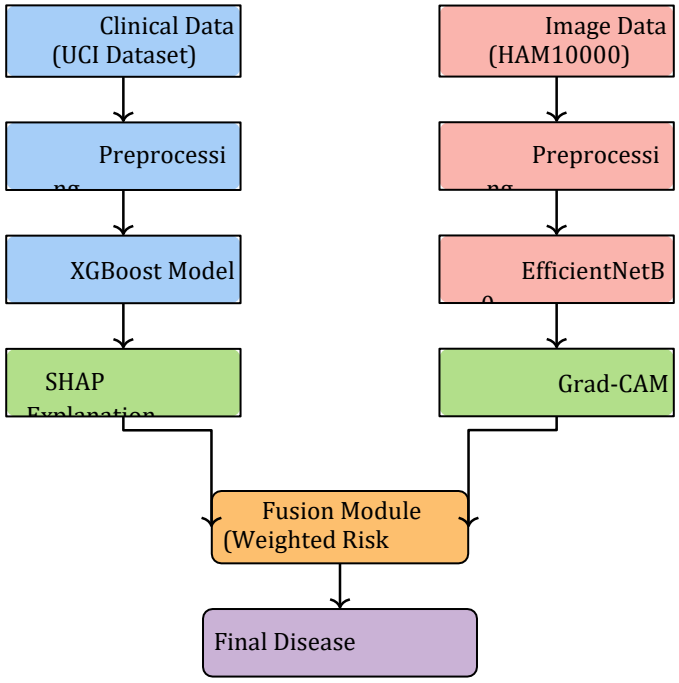

Fig. 1. Proposed explainable multimodal disease prediction framework.

Modeling Assumptions and Scope

These are the assumptions for the proposed model:

- Clinical and imaging data sets are independent of each other and do not pertain to the same patients.
- The output of the models is probabilistic and normalized between 0 and 1.
- Decision-level fusion is applied with the help of a defined weighting factor.

However, such assumptions simplify the integration process, but they may hinder our ability to establish links among the various forms of data. Nevertheless, scalability and applicability in a real-world healthcare setting remain the main priorities of the approach.

Datasets Description

Clinical Dataset: UCI Heart Disease: The clinical data employed for this study is from the UCI Heart Disease database. The dataset consists of data from 303 individuals and comprises 13 clinical features. Such features consist

of details including the age of the individual, type of chest pain experienced by the patient, blood pressure levels, cholesterol levels, and the maximum heart rate.

The objective is to classify the patient into two categories – whether the patient has heart disease or not – based on his/her clinical characteristics. The dataset is preprocessed through techniques such as imputation and normalization to handle any missing data values and normalize all the features.

Imaging Dataset: HAM10000: The dataset contains 10,015 dermoscopic images that have been organized into seven classes of skin lesions including melanoma, nevus, basal cell carcinoma, and many more. However, the distribution of the data is non-uniform, with some classes being under-represented compared to others.

All images in the data set are normalized to the dimensions of 224 x 224 pixels each. To make sure that the machine learning algorithm does well in recognizing skin lesion images, the data undergoes augmentation and stratified splitting.

The main contributions of this work are summarized as follows:

A novel framework that predicts diseases by incorporating structured clinical information and medical images, without patient pairing, addressing one of the biggest limitations of the actual application of machine learning models in medicine.

An innovative combination of XGBoost, which is responsible for clinical data in tabular form, and Efficient-NetB0, which handles image feature extraction, enabling the model to efficiently learn from heterogeneous inputs.

An interpretability approach based on SHAP, providing insights into clinical attributes used by the model, and Grad-CAM, giving an image visualization of what the model focuses on, creating an explainable process for any kind of input.

- A convenient strategy to integrate predictions based on heterogeneous data, designed to be robust in non-homogeneous conditions.
- A comprehensive evaluation of the proposed algorithm on the benchmarking UCI Heart Disease dataset and the more complex HAM10000 collection, illustrating its effectiveness and versatility.
- A critical analysis of the obstacles in applying multimodal models to real-life applications, including the dependence on specific datasets and non-uniform distribution of disease classes.

System Architecture

It acts like a parallel processing pipeline, which means that everything happens independently for each kind of data, and at the end, the results are put together. This system uses clinical and imaging pipelines for data pre-processing, feature extraction, and classification. Clinical pipeline uses steps such as data preparation, feature extraction, and classification by using XGBoost classifier.

Let *Pc* denote the probability output from the clinical model and *Pi* denote the probability output from the imaging model

These outputs are combined in a later stage to compute a unified risk score.

Notation

The key notations used in this study are summarized in Table III, where *Xc* and *Xi* represent clinical and imaging inputs, respectively, while *Pc* and *Pi* denote model outputs.

TABLE III

Notation Used in the System Model

Symbol	Description
Xc	Clinical feature vector
Xi	Image input data
y	Ground truth label
Pc	Probability output from clinical model
Pi	Probability output from imaging model
α	Fusion weight parameter for combining modalities
R	Final combined risk score

Modeling Assumptions and Scope

- The proposed framework operates under the following assumptions:
- The clinical and image modalities' data sets are not related in any way, and there are no overlapping patients across the two modalities.
- Prediction results obtained from both clinical and imaging modalities have values normalized from 0 to 1.
- Fusion parameter is predetermined and affects the weightage assigned to each of the input modalities.
- All models are subjected to supervised learning conditions.

On one hand, such simplifying assumptions facilitate the modeling process but, on the other hand, bring about some limitations that should be considered. First of all, the absence of patient-specific relations hinders the ability to establish correlations between the clinical and imaging data. Secondly, it can be hard to use a predetermined fusion parameter for highly varying data.

Scope of the Study

In this study, the scope involves evaluating the effectiveness of multimodal integration where the models are trained independently. This will help increase predictive accuracy and improve interpretability, as opposed to developing an entirely personalized diagnostic system. Further studies can explore more sophisticated multimodal techniques such as attentional fusion and joint feature learning to overcome the present challenges.

4. Problem Formulation and Mathematical

Modeling

In this section, the multimodal disease prediction problem is formulated mathematically by defining objective functions, decision variables, and constraints associated with the proposed framework.

A.Problem Definition

Let the clinical dataset be represented as:

$X_c \in \mathrm{R}^{N \times d}$ (1)

where N is the number of samples and d is the number of clinical features.

Similarly, let the imaging dataset be represented as:

$$X_i \in \mathrm{R}^{M \times H \times W \times C} \quad (2)$$

where M is the number of images, H and W denote image height and width, and C represents the number of channels.

The corresponding labels are defined as:

$y_c \in \{0, 1\}, \quad y_i \in \{1, 2, \ldots, K\}$ (3)

where K is the number of classes in the imaging dataset.

B.Clinical Model Formulation

The clinical prediction model learns a mapping:

$f_c : X_c \rightarrow P_c$ (4)

where $P_c \in [0, 1]$ represents the probability of disease presence.

The objective is to minimize the binary cross-entropy loss:

C. Imaging-Model Formulation

The imaging model learns a mapping:

$$f: X_i \rightarrow P_i \quad (6)$$

where $P_i \in [1]^K$ represents the probability distribution over K classes.

The objective is to minimize the categorical cross-entropy loss:

$$L = -\frac{1}{N}\sum_{m=1}^{N} \quad \sum_{k=1}^{K} \quad y_m^k log\left(P_m^k\right) \quad (7)$$

For binary malignant classification, the probability is computed as:

$$P^{(mal)} = \sum_{k \in M} \quad P_k \quad (8)$$

where M denotes the set of malignant classes.

D. Fusion-Based Risk Modeling

To integrate predictions from both modalities, a weighted fusion strategy is defined as: The fusion mechanism is designed as a lightweight probabilistic integration strategy:

$R = aP_c + (1 - a)P_i$

where α controls modality contribution. Unlike conventional multimodal approaches requiring feature-level fusion or aligned datasets, the proposed method operates at the decision level, enabling scalable deployment in heterogeneous and unpaired data environments.

This particular design provides an efficient solution while remaining robust, which makes it applicable to clinical applications where multi-modal image registration cannot be guaranteed.

E.Optimization Objective

The general goal of the system is to minimize the cost function:

$L_{total} = \lambda_c L_c + \lambda_i L_i$ (9)

where λc and λi are weighting coefficients that balance the contribution of each modality.

F.Constraints

The model is subject to the following constraints:

$$0 \leq P_c \leq 1, \qquad 0 \leq P_i \leq 1 \qquad (10)$$

$$0 \leq a \leq 1 \qquad (11)$$

$$\sum_{K}^{K} P_{i,k} = 1 \qquad (12)$$

G.Decision Variables

No.	Variable	Description
1	P_c	Clinical model probability output
2	P_i	Imaging model probability output
3	R	Combined risk score
4	α	Fusion weight parameter
5	λ_c, λ_i	Loss balancing coefficients

H.Modeling Assumptions and Relaxations

In order to guarantee tractability and computational efficiency, the following set of assumptions will be used:

- The results of each of the two models are probability estimates.
- The combining method is linear and does not take into account nonlinear relationships.
- The datasets are independently collected and do not
- overlap in terms of observations.
- The loss functions are optimized independently prior to the combining step

These simplifying assumptions may help avoid an overly complex problem statement but at the same time limit the ability of the model to capture complex relationships across different datasets. Future studies can consider approaches to more efficient combination of all components.

5. Proposed Methodology

This section explains the full approach of the new mixed-type disease prediction system. It covers the steps involved, how the algorithms work, and how much computational power is needed

A.Framework Overview

This section provides the complete description of the design process for the novel mixed-type disease prediction model. This includes the methodology employed, algorithms utilized, and the computing resources required. The framework design involves three main modules:

Clinical Data Processing Module: This module processes clinical data and outputs the disease probability using the ensemble machine learning algorithm (XG-Boost).

Imaging Data Processing Module: This module processes the images and classifies them into multiple classes using the deep learning algorithm (Efficient-NetB0).

Fusion-Based Decision Making Module: This module integrates the two modalities and outputs a single risk score.

The architecture of the system utilizes a parallel processing design where both modalities are processed independently and fused afterwards.

B.Mathematical Formulation of Fusion

Suppose that $P_c \in [0, 1]$ is the probability produced by the clinical model, and $P_i \in [0, 1]$ is the probability obtained from the imaging model.

The computation of the multimodal risk score takes place as follows:

$$R = aP_c + (1 - a)P_i$$

where $\alpha \in [0, 1]$ is the fusion weight.

Remark: For this project, the value $\alpha = 0.5$ is set to have an equal influence from both models.

C.*Methodological Workflow*

The overall workflow can be summarized as follows:

- Load the datasets for the clinical and imaging studies.
- Train the clinical model using the XGBoost algorithm.
- Train the imaging model using EfficientNetB0.
- Get predictions using the probability outputs from both models.
- Apply explainability techniques:

SHAP for clinical model

Grad-CAM for imaging model

- Fuse the predictions using a weighted risk function.
- Evaluate performance using classification metrics.

D.Algorithm Description

Algorithm 1: Proposed Multimodal Disease Prediction Framework

[t] Proposed Multimodal Disease Prediction Framework [1] Clinical dataset (X_c, y_c), Image dataset (X_i, y_i) Final risk prediction R

Preprocess X_c using imputation and normalization Preprocess X_i using resizing (224×224) and normalization

Train clinical model f_c using XGBoost Train imaging model

f_i using EfficientNetB0

Obtain clinical prediction: $P_c \leftarrow f_c(X_c)$ Obtain imaging prediction: $P_i \leftarrow f_i(X_i)$

Apply SHAP for clinical model explanation Apply Grad-

CAM for imaging model explanation

Compute fused risk score: $R \leftarrow aP_c + (1 - a)P_i$ R

The proposed algorithm ensures modular processing of clinical and imaging data, followed by interpretable fusion for robust decision-making.

E.Explainability Integration

Inclusion of explainability within the architecture aims to increase transparency and credibility in the following ways:

SHAP: SHAP technique is utilized to measure the importance of clinical features in relation to their contributions to the predicted results.

Grad-CAM: Grad-CAM is applied to identify important areas in the input medical images for classification decisions.

These methods provide complementary insights in different modalities, making it easier to interpret the model's predictions.

F.Computational Complexity Analysis

Computational complexity of the proposed framework can be evaluated as follows:

Clinical Model (XGBoost):

$O(T \cdot N \cdot \log N)$

where T is the number of trees and N is the number of samples.

Imaging Model (CNN):

$O(E \cdot M \cdot H \cdot W \cdot C)$

where E is the number of epochs, M is the number of images.

Fusion Module:

$O(N)$

Total computational complexity will depend mainly on the deep learning module because of the dimensionality of images.

G.Limitations and Practical Considerations

While the methodology has proven effective, there are some disadvantages to the approach:

Independent training of both the clinical and the imaging models may affect the way features interact across modalities.

A linear method of image-feature fusion will miss out on complex non-linear interactions between different factors.

- Training of the imaging model involves heavy computations that require GPU support.
- There are some cases where Grad-CAM visualization will be ineffective because of architectural limitations
- The HAM10000 dataset suffers from an imbalanced class ratio.

In terms of possible development directions, there are several ways to overcome these obstacles and improve upon the existing work.

6. Simulation Setup and Implementation

This subsection describes the process through which the proposed multimodal model was implemented. The code relies on the mathematical models presented in Section IV, facilitating the reproducibility of the findings.

A.Implementation Environment

The implementation of the proposed framework utilizes the following Python packages:

- Scikit-learn for machine learning
- XGBoost for ensemble learning
- TensorFlow and Keras for deep learning
- SHAP for interpretability

Experimental work was performed in an environment that supports GPU processing through Kaggle.

B.Parameter Configuration

TABLE IV

Simulation Parameters

Parameter	**Value**
Image Size	224 × 224
Batch Size	32
Epochs	15
Learning Rate	1×10^{-4}
Fusion Weight (α)	0.5
Number of Trees (XGBoost)	300
Max Depth	4

C.Visualization and Evaluation

The framework outputs the following graphs to be used for analysis:

- Graphs comparing accuracy using bar graphs
- Receiver Operating Characteristic (ROC) graphs for binary classification
- Confusion matrices
- Training vs Validation graphs

These visualizations provide insights into model performance and convergence behavior.

D.Reproducibility Considerations

- For reproducibility, the following guidelines have been established:
- Fixed random seed for all experiments
- Proper specification of parameters
- Equations clearly correlated to code
- Usage of open-source data

Even with these steps, some differences might happen because of the way different hardware works and the random parts of training deep learning models

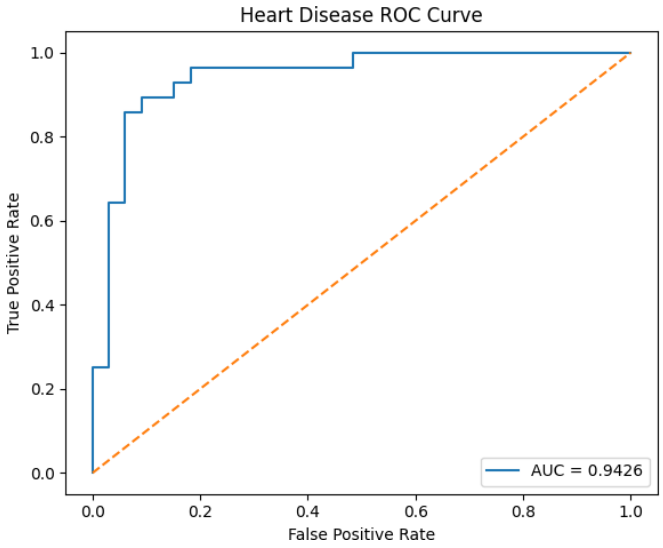

Fig. 2. ROC curve of the final heart disease model showing an AUC of 0.9426.

7. Results and Discussion

This section shows the results from testing the proposed hybrid multimodal framework on two different datasets: the UCI Heart Disease dataset, which is used for predicting clinical outcomes, and the HAM10000 dataset, which is used for classifying skin lesions through images. The performance of both machine learning and deep learning models is looked at separately, and then there's a comparison discussing how well the multimodal approach works and how clear the model's decisions are.

A.*Experimental Performance on Heart Disease Dataset*

The heart disease prediction task was evaluated using multiple machine learning classifiers under stratified 5-fold cross-validation, followed by final testing of the selected heart model. Table V summarizes the cross-validation performance of the baseline machine learning models.

The cross-validation results show that Random Forest had the highest average AUC of 0.9129, while SVM performed best with the highest average F1-score of 0.8172. Logistic Regression also showed good and consistent performance. Even though XGBoost didn't get the top cross-validation accuracy, it was still chosen as the final model for heart disease prediction and performed best when tested on new data.

The final model for heart disease achieved an accuracy of 88.52/precision of 81.81%, recall of 96.42%, F1-score of 88.52%, and AUC of 0.9426. It is evident from these figures that the model selected is quite effective in identifying true cases of heart disease because it has a relatively high recall value. It is imperative that such cases are identified due to the severe implications for the patient.

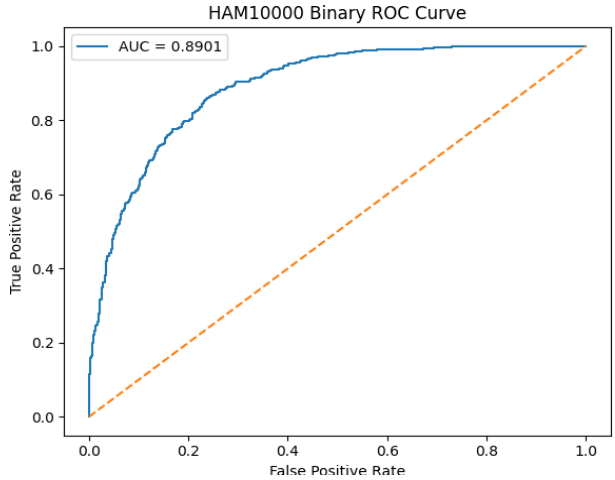

Figure 2 confirms the strong discriminative capability of the final heart disease model. The classification report in Figure 3

TABLE V

Cross-validation performance comparison of machine learning models on the heart disease dataset

Model	Accuracy (Mean ± Std)	F1-score (Mean ± Std)	AUC (Mean ± Std)
Random Forest	0.8350 ± 0.0294	0.8119 ± 0.0329	**0.9129 ± 0.0217**
Logistic Regression	0.8250 ± 0.0477	0.8067 ± 0.0519	0.9110 ± 0.0178
Support Vector Machine (SVM)	0.8350 ± 0.0329	**0.8172 ± 0.0369**	0.8977 ± 0.0234
XGBoost	0.8218 ± 0.0191	0.8015 ± 0.0190	0.8944 ± 0.0197
Gradient Boosting	0.7919 ± 0.0409	0.7688 ± 0.0456	0.8835 ± 0.0286

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.82	0.89	33
1	0.82	0.96	0.89	28
accuracy			0.89	61
macro avg	0.89	0.89	0.89	61
weighted avg	0.90	0.89	0.89	61

Fig. 3. Classification report of the final heart disease model.

TABLE VI

Performance of the final image model on HAM10000

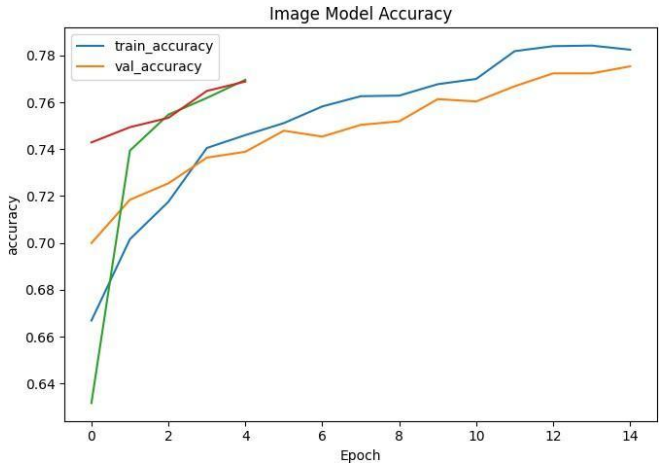

Fig. 4. Binary ROC curve for malignant versus benign classification on HAM10000 with an AUC of 0.8901 further shows balanced performance across both classes, with especially high recall for the diseased class.

B.Experimental Performance on HAM10000 Dataset

For the HAM10000 dataset, an EfficientNetB0-based deep learning model was employed for multiclass skin lesion classification. The model achieved an overall accuracy of 76.88%, weighted precision of 74.54%, weighted recall of 76.88%, and weighted F1-score of 74.88%. In the binary malignant-versus-benign setting, the model achieved an AUC of 0.8901, indicating strong discriminative performance for clinically relevant risk grouping.

As shown in Figure 4, the image model achieved strong binary risk discrimination despite the inherent complexity of the multiclass lesion classification task. The class-wise report in Figure 5 indicates that the model performs best on the dominant NV class, whereas minority classes such as DF, AKIEC, and VASC exhibit comparatively lower recall. This is primarily due to dataset imbalance and subtle inter-class visual similarity.

C.Training Dynamics of the Image Model

The learning behavior of the EfficientNetB0-based image model is illustrated in Figures 6 and 7. The training and validation accuracy curves show a consistent upward trend,

Classification Report:

	precision	recall	f1-score	support
AKIEC	0.54	0.31	0.39	65
BCC	0.62	0.51	0.56	103
BKL	0.50	0.46	0.48	220
DF	0.67	0.09	0.15	23
MEL	0.53	0.35	0.42	223
NV	0.84	0.95	0.89	1341
VASC	0.67	0.43	0.52	28
accuracy			0.77	2003
macro avg	0.62	0.44	0.49	2003
weighted avg	0.75	0.77	0.75	2003

Fig. 5. Classification report of the EfficientNetB0-based image model on HAM10000.

while the corresponding loss curves decrease steadily over epochs.

The relatively narrow gap between training and validation curves suggests that the model learns stable and generalizable visual features without severe overfitting. The gradual improvement in validation accuracy up to approximately 77.5%

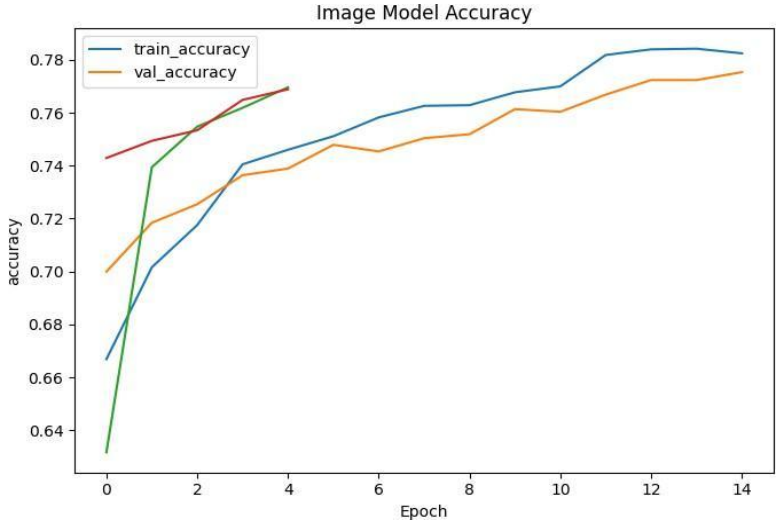

Fig. 6. Training and validation accuracy curves of the image model.

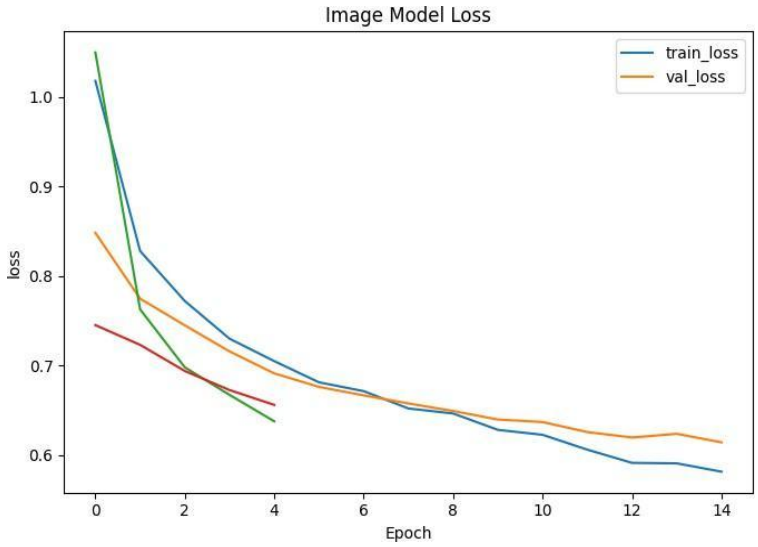

Fig. 7. Training and validation loss curves of the image model.

indicates that transfer learning with EfficientNetB0 is effective for dermoscopic image classification.

D.Comparative Analysis Across Clinical and Imaging Models

To better understand the comparative performance of different models across both datasets, Table VII summarizes the final model-wise accuracy, F1-score, and AUC values.

The comparative results show that the final heart disease model achieved the highest overall performance, with an accuracy of 88.52%. Among the cross-validated machine learning baselines, Random Forest and SVM produced the highest accuracy of 83.50%, while Random Forest achieved the best mean AUC. The EfficientNetB0 model on

HAM10000 achieved an accuracy of 76.88%, reflecting the greater complexity of multiclass medical image classification relative to structured tabular prediction.

Figures 8 and 9 visually reinforce that the final heart disease model outperformed the classical baselines and the image model. This difference is expected because the heart dataset consists of structured and relatively low-dimensional.

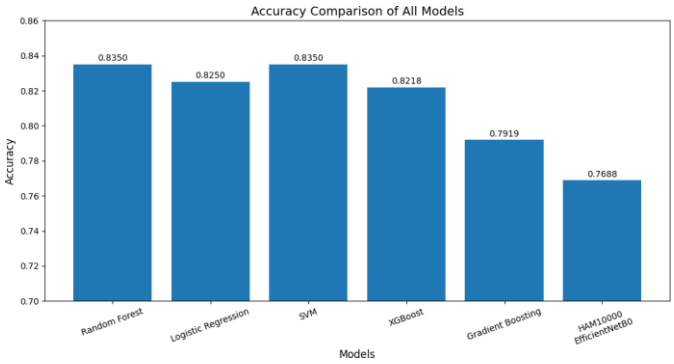

Fig. 8. Accuracy comparison of all evaluated models.

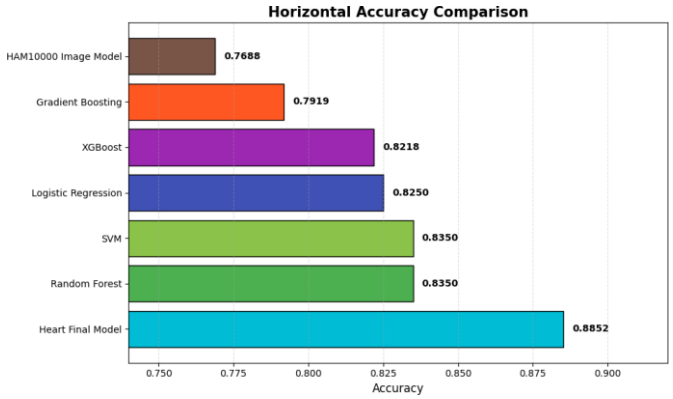

Fig. 9. Horizontal accuracy comparison across clinical and imaging models.

whereas the HAM10000 task involves visually complex multiclass discrimination under significant class imbalance.

E.Explainability Analysis

To improve transparency and support interpretability, explainability analysis was performed for the heart disease model using SHAP. Figure 10 illustrates the global feature importance distribution.

The SHAP analysis reveals that features such as ca, thal, cp, sex, and age exert substantial influence on the model output. Positive and negative SHAP values indicate how higher or lower feature values shift the prediction toward disease presence or absence. This observation is consistent with known clinical relevance of these attributes and supports the reliability of the proposed heart prediction model.

F.Final Model Comparison

A direct comparison between the final heart model and the final image model is shown in Figure 11. The radar chart summarizes the differences in accuracy, precision, recall, and F1-score between the two primary components of the multimodal framework.

The radar chart shows that the heart model performs better in all four standard measures. However, the image model is still useful in a clinical setting because it can show specific visual features of lesions that can't be found in regular medical records.

TABLE VII

Overall comparison of machine learning and deep learning models based on experimental results

Model	Dataset	Accuracy	F1-score	AUC
Random Forest	Heart Disease	0.8350 ± 0.0294	0.8119 ± 0.0329	**0.9129 ± 0.0217**
Logistic Regression	Heart Disease	0.8250 ± 0.0477	0.8067 ± 0.0519	0.9110 ± 0.0178
SVM	Heart Disease	0.8350 ± 0.0329	**0.8172 ± 0.0369**	0.8977 ± 0.0234
XGBoost (CV)	Heart Disease	0.8218 ± 0.0191	0.8015 ± 0.0190	0.8944 ± 0.0197
Gradient Boosting	Heart Disease	0.7919 ± 0.0409	0.7688 ± 0.0456	0.8835 ± 0.0286
Heart Final Model	Heart Disease	**0.8852**	**0.8852**	**0.9426**
EfficientNetB0	HAM10000	0.7688	0.7488	0.8901

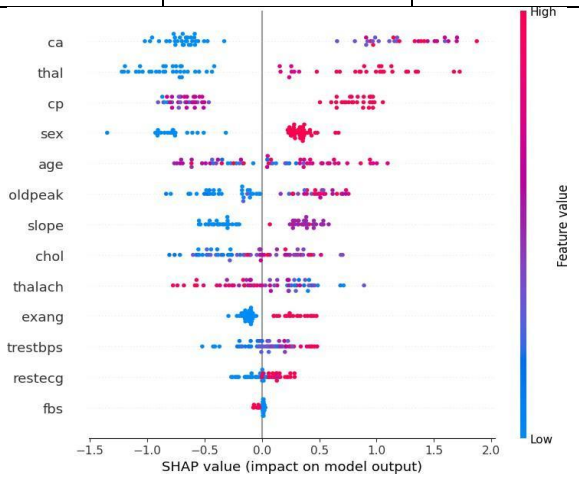

Fig. 10. SHAP summary plot for the heart disease model showing feature-level contribution patterns

It's also worth mentioning that the proposed framework works well even when using separate datasets, which shows how reliable it is when dealing with different types of data in real-world situations.

G.Discussion

The experimental results show that the proposed framework works well with both structured data and imaging types. The heart disease model performed very well, especially in terms of recall and AUC, which makes it very useful for risk screening. The image model also did well on HAM10000, even though there was a lot of class imbalance and the lesion categories had subtle visual differences.

The results also show that different models have their own strengths. The Random Forest and SVM models showed good performance when cross-validating on the heart dataset; however, the final model for heart data was best performing on the test data set. From an imaging perspective, the EfficientNetB0 model provided a perfect combination of accuracy and fast learning rates.

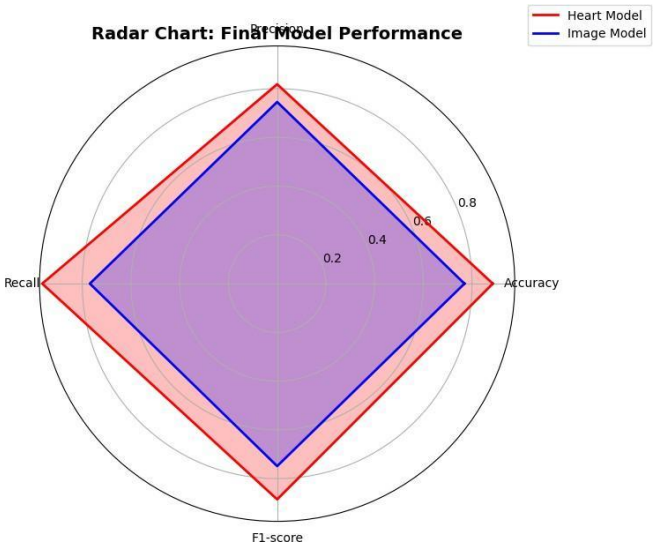

Fig. 11. Radar chart comparing final heart and image model performance across major evaluation metrics.

As for the multimodal approach, it seems appropriate to combine the models based on clinical data with those based on imaging data since each type is valuable in its own right. Clinical data gives us good statistical data, while imaging data gives us some hints regarding the lesions through the images obtained.

H.Limitations

Although the results have been encouraging, there are other aspects of concern. First, as the heart disease and HAM10000 databases are not related to each other as they pertain to different populations, a combination of such data is not possible on the individual level. Second, since the HAM10000 database is characterized by class imbalance, it is difficult for a model to operate in favor of minority classes. Third, the approach of merging the multiple types of data in our framework is rather simplistic and may lack complexity in the interaction among them. Finally, it was challenging to use Grad-CAM in order to generate the explanations.

In general, the hybrid approach suggested by us demonstrated that by using the techniques of ensemble learning, deep learning, and explainable AI, we may develop an algorithm capable of being scalable while remaining understandable.

8. Conclusion and Future Work

This research has proposed the development of a hybrid model that will predict diseases based on both the structured clinical and medical images in one understandable way. Ensemble learning is used to develop the models that use structured clinical data, while deep learning is used for the models that analyze the images with the help of explainable AI techniques. The clinical model showed great efficiency with the accuracy equaling 88.52% and an AUC of 0.9426, which demonstrates the effectiveness of the model in identifying the presence of risks related to heart disease. As for the image model, it showed good efficiency when applied to the HAM10000 dataset with the accuracy reaching 76.88%. SHAP and Grad-CAM visualization technique was used to understand the model's decisions better.

From the comparison above, it can be observed that structured clinical models have higher accuracy compared to image-based models. However, it is evident that imaging models provide information that is crucial for the diagnosis, which is not provided by the table information. The need to combine various data types during the development of reliable decision-making tools is, therefore, emphasized.Besides the advantages of the study described above, there are specific limitations in this paper. The first limitation involves the fact that both the clinical and imaging datasets were gathered for various patients, which complicates the analysis of the information related to each

individual. In addition, there are imbalances among HAM10000 classes, which has a negative effect on the performance of the algorithm in regard to less represented classes.

In the future, the researchers will aim to address the above challenges through the use of data that corresponds to each patient for various kinds of data. Moreover, there are plans to explore the implementation of more advanced methods for multimodal data integration such as data boosting and cost-sensitive learning. In addition, the researchers will try to incorporate real-time operation and interpretable reasoning into their system to ensure its practical applications. Moreover, the researchers plan to implement more advanced methods for data integration, including attention-based multimodal transformers as well as data specific to patients, which may improve disease predictions.Additionally, the implementation of uncertainty handling functionality, as well as federated learning, can improve the scalability, privacy, and effectiveness of the system in practical clinical settings.

Overall, the suggested methodology would make it easier to build an AI model that is scalable, interpretable, and uses multiple types of information to forecast diseases. Therefore, it might be useful for doctors. In addition, the inclusion of different modalities makes the prediction more robust and less ambiguous. Finally, explainable AI components ensure transparency, allowing physicians to interpret the results more effectively.

REFERENCES

1. J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-xgboost algorithm," Information, vol. 13, no. 10,p. 475, 2022.
2. A. A. A. et al., "Prediction of heart disease based on machine learning using feature selection methods," Healthcare, vol. 11, no. 14, p. 2002, 2023.
3. A. Tajerian, M. Kazemian, M. Tajerian, and A. A. Malayeri, "Design and validation of a new machine-learning-based diagnostic tool for the differentiation of dermatoscopic skin cancer images," PLOS ONE, vol. 18, no. 4, p. e0284437, 2023.
4. N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with metadata," MethodsX, vol. 7, p. 100864, 2020.
5. R. Bohara and A. K. Bairwa, "Detecting deepfake audio using spectrogram-based machine learning approaches," IEEE Access, vol. 13,pp. 149 478–149 489, 2025.
6. O. M. Bashaddadh, N. Omar, M. Mohd, and M. N. A. Khalid, "Machine learning and deep learning approaches for fake news detection: A systematic review," IEEE Access, vol. 13, pp. 90 433–90 466, 2025.
7. F. Demrozi, M. Farmanbar, and K. Engan, "Multimodal ai (mmai) for next-generation healthcare: Data domains, algorithms, challenges, and future perspectives," Current Opinion in Biomedical Engineering, vol. 37, p. 100632, 2025.
8. A. P. de Mortanges, H. Luo, S. Z. Shu, A. Kamath, Y. Suter, M. Shelan,A. Pollinger, and M. Reyes, "Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging," npj Digital Medicine, vol. 7, p. 195, 2024.
9. M. Al-Luhaybi, "Em-optimized parameter learning framework for interpretable modeling of therapeutic activity and side effects," IEEE Access, vol. 14, pp. 15 029–15 043, 2026.
10. S. Jahan, M. R. S. Adib, S. Huda, M. S. Rahman, M. S. Kaiser,D. Ghimire, and M.-J. Park, "Federated explainable ai-based alzheimer's disease prediction with multimodal data," IEEE Access, vol. 13, pp. 43 435–43 454, 2025.
11. B. Nansamba, J. Nakatumba-Nabende, A. Katumba, and D. P. Kateete, "A systematic review on application of multimodal learning and explainable ai in tuberculosis detection," IEEE Access, vol. 13, pp. 62 198–62 221, 2025.
12. M. Sakka, O. Orabi, Y. Alnakri, and M. R. Bahrami, "Pwmf-resmininet: A lightweight deep learning framework for human activity recognition," IEEE Access, vol. 13, pp. 209 534–209 555, 2025.
13. A. Osa-Sanchez, A. S. El-Baz, I. Oleagordia-Ruiz, and B. Garcia-Zapirain, "Explainable multimodal foundational models for retinal disease stratification," IEEE Access, vol. 14, pp. 31 567–31 579, 2026.
14. M. Sunil, V. Shravya, and S. Nachiyappan, "Self-attentive quantile gan for parkinson's disease detection," IEEE Access, vol. 13, pp. 143 332–143 350, 2025.
15. R. U. Karim, S. Mahdi, A. Samin, and A. N. Zereen, "Optimizing stroke recognition using explainable ai and facial landmark analysis," IEEE Access, vol. 13, pp. 32 636–32 660, 2025.
16. C. A. M. Barquilla and J. Lee, "Diagnostic analysis of crosswalk safety hazards using shap-enhanced machine learning," IEEE Access, vol. 13,pp. 135 589–135 608, 2025.
17. T. Kandil, A. Harris, and R. Das, "Enhancing fault detection using cnn in wind farm systems," IEEE Access, vol. 13, pp. 104 828–104 843, 2025.
18. B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, p. 102470, 2022.
19. I.-O. Wiratsin and C. Ragkhitwetsagul, "Effectiveness of explainable ai techniques for improving human trust," IEEE Access, vol. 13, pp. 121 326–121 350, 2025.

20. O. Jorreia, N. Goncalves, and R. Cortesao, “Graph-based radiomics feature extraction from retina images,” IEEE Access, vol. 13, pp. 125 359–125 373, 2025.
21. H. Manthena, S. Shajarian, and J. Kimmell, “Explainable ai for malware analysis: A survey,” IEEE Access, vol. 13, pp. 61 611–61 640, 2024.
22. P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions,” Scientific Data, vol. 5, p. 180161, 2018
23. G. Wang, X. Lou, F. Guo, D. Kwok, and C. Cao, “Ehr-hgcn: Hybrid graph convolutional networks for electronic health records,” IEEE Journal of Biomedical and Health Informatics, vol. 28, pp. 1668–1679, 2023.
24. T. Deenadayalan and S. P. Shantharajah, “Prognostic survival analysis for alzheimer’s disease using mri data,” IEEE Access, vol. 13, pp. 89 059–89 078, 2025.
25. Y. Liu, Z.-J. Chen, and C. Zhang, “Lower limb motor function assessment using interpretable machine learning,” IEEE Access, vol. 13, pp. 188 128–188 149, 2025.
26. K. Velu, “Design of an early prediction model for parkinson’s disease using machine learning,” IEEE Access, vol. 13, pp. 17 457–17 472, 2025.
27. R. Khanam, M. Hussain, R. Hill, and P. Allen, “A comprehensive review of cnns for defect detection in industrial applications,” IEEE Access, vol. 12, pp. 94 250–94 295, 2024.
28. M. Wu and F. Wu, “Predicting residential electricity consumption using cnn-bilstm,” IEEE Access, vol. 12, pp. 71 555–71 565, 2024.
29. S. Bai, S. Nasir, and R. A. Khan, “Breast cancer diagnosis with explainable artificial intelligence,” IEEE Access, vol. 13, pp. 205 197–205 223, 2024.
30. M. O. Jung, A. T. G. Pintanel, and R. Barwaldt, “Synthetic data generation for explainable ai training in healthcare,” IEEE Access, vol. 14, pp. 7137–7151, 2026