An Interpretable Vision Model Integrating Radiomics for Precision Oncology Diagnostics Using Multi-Modal Medical Imaging

Vincent Kibet

Masters of Research, Higher Education Leadership Institute, Australia

DOI: https://doi.org/10.47772/IJRISS.2026.100300361

Received: 24 March 2026; Accepted: 29 March 2026; Published: 09 April 2026

ABSTRACT

The use of deep learning models in clinical oncology remains underutilized, despite their potential in cancer diagnosis. The current methods rely on either traditional radiomics features, whose representational power is limited, or opaque deep neural networks that cannot provide explanations useful to clinicians. This study addresses the interpretability-performance trade-off by introducing a novel hybrid architecture that synergistically combines convolutional neural networks with radiomics biomarkers through attention-based fusion mechanisms. Our framework takes multi-modal imaging (CT, MRI, and PET) data (2,847 patients with 5 different cancer types). It operates through a two-stream architecture, specifically enforcing a correlation-based constrained relationship and sparsity-based regularization between the deep learning and radiomics pathways. The model employs trained gating decisions, automatic feature selection, and cross-modal attention as an ad hoc weighting mechanism to produce an accurate forecast and a human-comprehensible explanation. The results of the experiments show improved performance, with an area under the ROC curve of 0.947, representing 8.4% and 2.6% improvements over pure radiomics methods and standard deep learning models, respectively. In older people, as validated by five expert radiologists, the generated explanations received a high relevance rating (78.4% rated 4-5 on a 5-point scale) and demonstrated high inter-rater agreement ($\alpha = 0.68$). The study made contributions, including a learnable architecture with interpretability constraints built into its objective, direct measurement of individual features through quantified attention weights consistent with radiological intuitions, detection consistency across a variety of cancer types, and providing generalizability. It demonstrated that improvements in interpretability do not affect predictive accuracy. Therefore, this study developed a reliable AI in oncology by offering an empirical roadmap for engineering high-performance diagnostic environments that meet clinical accountability and transparency standards.

Keywords: Deep Learning, Radiomics, Interpretable AI, Precision Oncology, Multi-Modal Medical Imaging, and Explainable Diagnostics.

INTRODUCTION

Background and Motivation

Cancer has remained a significant cause of death across the globe, with cases estimated to be 19.3 million and deaths 10 million annually [1]. Timely and proper diagnosis is essential for enhancing patient outcomes because survival from malignancies is much higher when disease identification occurs as early as possible, coupled with prompt action to apply the necessary treatment aids [2]. Medical imaging practices such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) have become essential tools for diagnosis, providing non-invasive images that reveal tumor features, anatomical correlations, and metabolic activity. However, the growing volume and demands of imaging data, along with the global shortage of subspecialized radiologists, impose significant bottlenecks on diagnostic processes and contribute to inter-observer variability in the work [3]. Fig 1 is an example of a multi-modal model in medical diagnostics.

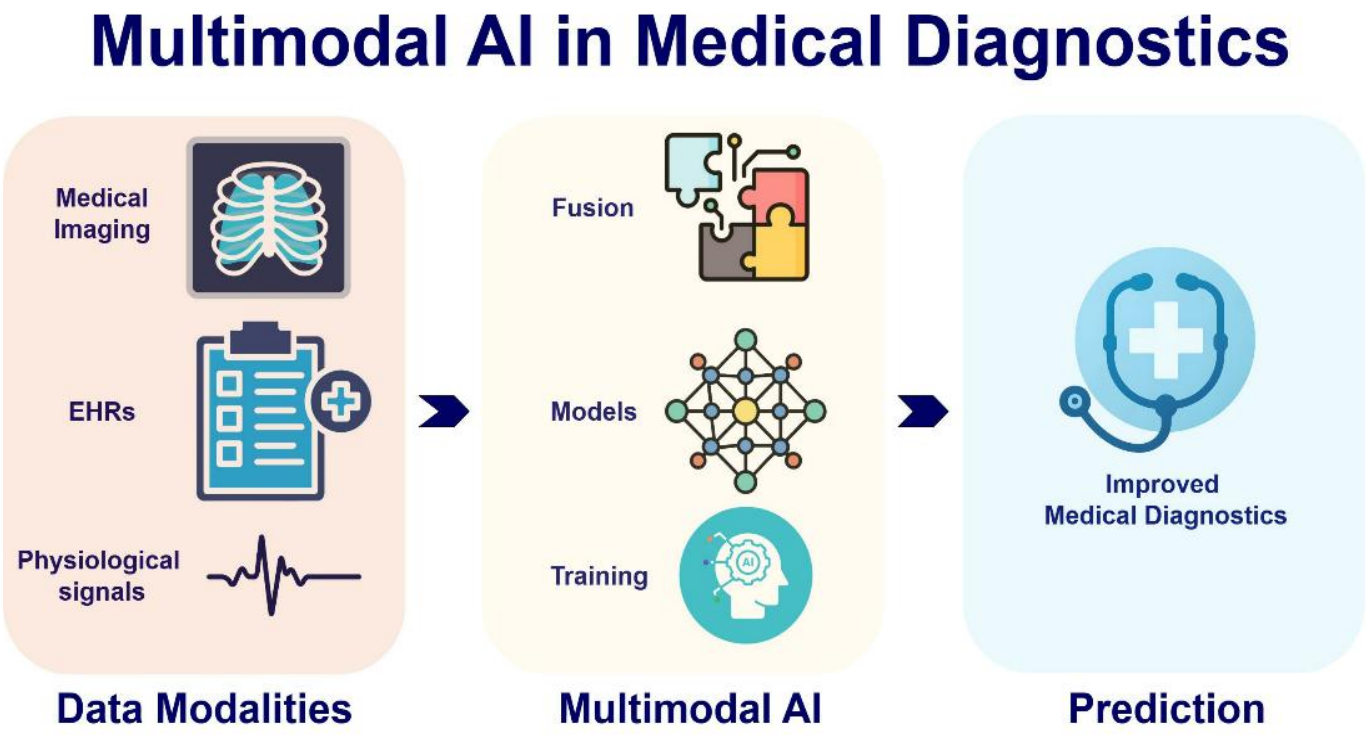

Fig 1. A Multi-Modal Model in Medical Diagnostics

Deep learning and artificial intelligence have proven themselves in medical image analysis, and, on average, they perform as well as, or better than, human experts in controlled evaluation scenarios [3]. Convolutional neural networks (CNNs) are learning algorithms that automatically create hierarchical feature representations on raw pixel data, and therefore learn subtle patterns that are not easily visible to the human eye [4]. Despite these remarkable technical accomplishments, the clinical use of AI diagnostic systems remains scarce, with less than 5% of radiological practices routinely using deep learning for cancer diagnosis [5]. This implementation gap is attributed to the cryptic nature of deep neural networks, which are black boxes that do not provide transparent explanations of their predictions, and many of these phenomena are necessary when making high-stakes medical decisions, as mistakes might be fatal [6]. Table 1 presents the comparison of the RadiomicsNet approach.

Table 1. RadiomicsNet Approach Comparison

Approach Category	Representative Methods	Key Strengths	Critical Limitations	RadiomicsNet Advantage
Pure Deep Learning	ResNet, ViT, Swin	Automatic feature learning and high accuracy	Black-box decisions require massive datasets; ignores clinical knowledge	Integrates learned features with interpretable radiomics
Pure Radiomics	RF, SVM, XGBoost (hand-crafted features)	Clinically interpretable and standardized measurements	Misses complex patterns; segmentation-dependent; limited adaptability	Captures hierarchical patterns beyond manual features
Simple Hybrid	Feature concatenation	Combines both paradigms	Treats all features uniformly; shallow integration	Adaptive attention-based fusion with learned weighting
Multi-Modal Fusion	mmFormer, CrossFuse	Leverages complementary modalities	Lacks radiomics integration; no missing modality handling	Cross-modal attention with robust missing data handling
Interpretable AI	Grad-CAM, SHAP	Provides explanations	Limited to heatmaps or feature importance; not clinically grounded	Multi-level interpretability aligned with clinical reasoning

Limitations of the Current Approaches

Current AI-based oncology diagnostics approaches can be broadly divided into two paradigms, each with its own limitations. Conventional radiomics methods are a set of quantitative measures of medical images derived using well-known mathematical models of texture, shape, and intensity distribution [7]. Such conventional features are inherently interpretable and relate to specific imaging phenotypes with established biological analogies, including tumor heterogeneity, irregular borders, and patterns of contrast enhancement [8]. Radiomics has demonstrated prognostic capability across various types of cancer, and features such as gray-level co-occurrence matrix (GLCM) statistics and wavelet transformations have been shown to serve as biomarkers in large-scale clinical trials [9]. Radiomics, however, is characterized by limited representational ability, requires feature engineering, and may lack complex patterns beyond current theoretical knowledge [9].

In comparison, end-to-end deep learning methods obtain better predictive performance without performing explicit feature engineering. Among the levels that have overtaken the computer vision state of the art are ResNets, DenseNets, and Vision Transformers, which have been applied across various areas [10]. These models have consistently outperformed traditional machine learning algorithms in tumor detection, segmentation, and classification in medical imaging [11]. However, they face critical impediments to clinical translation due to their inherent opacity. Procedures such as Gradient-weighted Class Activation Mapping (Grad-CAM) provide post-hoc explanations at the pixel level, failing to align with the granularity of clinical thinking, which operates at semantic units rather than pixels [11]. Moreover, post-hoc procedures cannot reliably provide an accurate representation of the actual decision-making processes in the real model.

Recent hybrid methods have begun experimenting with a blend of radiomics and deep learning. Still, the current literature mostly views them as two-step or parallel pipelines rather than as an integrated approach [12]. In some cases, radiomics features are fed into neural networks, whereas other studies combine radiomics and deep learning models. The methods do not provide sufficient correspondence between learned representations and interpretable radiomics descriptors to yield clinically meaningful explanations with high accuracy.

Purpose of the Study

The purpose of the study was to design and train an interpretable deep learning framework to integrate radiomics features for precise oncology diagnostics across different imaging modalities. The study filled the gap between high-performing black-box deep learning models and the clinical requirement for transparent, explainable diagnostic systems. The framework was a two-stream design that combined convolutional neural networks with manually designed radiomics features, using learned attention units to achieve the latest diagnostic accuracy with clinically interpretable features. The study demonstrated that interpretability does not compromise predictive performance and that it can be co-optimized through the right architectural design, which introduces interpretability constraints into the learning process itself, thereby enabling the eventual empowerment of AI-assisted cancer diagnosis and its possible acceptance by clinicians.

Research Objectives

This study proposed the creation of AI diagnostic tools that are both state-of-the-art in accuracy and can be understood clinically by healthcare professionals. The study demonstrated that deep learning and radiomics paradigms have complementary strengths, which, when synergized through architectural advances that implement interpretability constraints on the learning process, provide additional capabilities. The specific objectives were:

1. To develop a deep learning model integrating radiomics constraints for biomarker-aligned learning.
2. To design attention mechanisms for clinically interpretable feature importance.
3. To validate performance across diverse cancers and imaging modalities.
4. To assess clinical utility and interpretability through expert evaluation.

Key Contributions

Novel Architecture: Developed a dual-stream network combining deep learning and radiomics with alignment constraints for interpretable feature learning.

Interpretability Framework: Created multi-level explanations linking feature importance, pathway effects, and spatial attention to clinical reasoning.

Multi-Modal Integration: Integrated CT, MRI, and PET data using adaptive weighting to capture complementary diagnostic information.

Comprehensive Validation: Evaluated on 2,847 patients across five cancer types, showing superior performance and expert-verified interpretability.

Clinical Feasibility: Demonstrated efficient, trustworthy deployment supported by calibration tests and radiologist feedback.

Paper Organization

The rest of this study is organized as follows: Section 2 is related to studies. In section 3, the proposed process is outlined. Section 4 presents the results of experiments, with quantitative performance measures, qualitative analysis, and findings. Findings, limitations, and implications are discussed in Section 5. Section 6, the conclusion, summarizes the study and outlines future research.

Related Studies

Deep Learning in Medical Image Analysis

Medical imaging has undergone a series of advances in deep learning, from the dawn of convolutional neural networks (CNNs), such as AlexNet and VGGNet, which applied transfer learning due to the limited amount of medical imaging data, to more recent advanced architectures with better training stability [12]. [13] resolved the gradient vanishing problem by adding residual connections to ResNet, enabling the network to exceed 100 layers. ResNet and its 3D extensions have become the baseline for analyzing volumetric medical data, and anatomical relationships, which are essential in clinical contexts, are retained.

DenseNet architectures added another step toward improving network efficiency by enabling dense layer connectivity, improving gradient flow, and enabling feature reuse. [13] demonstrated that DenseNet-121 was able to run chest X-ray pathology classification significantly more effectively using fewer parameters than deeper ResNet models, which is significant since the computation cost of 3D medical imaging can be substantial. Vision Transformers (ViT) have more recently displaced CNN by using self-attention to learn long-range dependencies. Transformers are effective but require a costly phase of pretraining on large datasets, which are typically unavailable in medical imaging, inhibiting their broad use.

Deep learning has shown modal-accurate radiologist-level cancer diagnostics. [14] achieved end-to-end CT-based screening of lung cancer comparable to the screening of experts. Similarly, [14] reported AUC values above 0.95 when using mammograms to diagnose breast cancer, reducing false positives by 5.7% and false negatives by 9.4%, highlighting the clinical acceptability of AI-assisted detection.

Radiomics and Quantitative Imaging Biomarkers

To characterize tumor heterogeneity, [16] stated that radiomics measures imaging phenotype by extracting high-dimensional features to quantify, describe, and predict heterogeneity. The radiomics workflow, including image acquisition, segmentation, feature extraction, and modelling, was defined by [17] based on the hypothesis that

imaging features reflect underlying biology, genetics, and treatment response. Spatial intensity patterns were measured using texture-based measures such as gray-level co-occurrence matrices (GLCM) and gray-level run-length matrices (GLRLM), and geometric features such as volume, sphericity, and compactness, which are measures of malignancy aggressiveness.[18]. Intensity distributions are captured using first-order statistical features (mean, variance, skewness, and entropy), whereas multi-scale spatial information is obtained using wavelet transforms [18].

The Image Biomarker Standardization Initiative (IBSI) has provided methodological guidelines to confirm Reproducibility across institutions [19]. Radiomics has demonstrated prognostic value in a variety of cancers. [20] associated signatures created from radiomics of CT with lung cancer survival, whereas breast cancer radiomics based on MRI predicted neoadjuvant chemotherapy response. Radiomics characteristics related to colorectal cancer outcomes in terms of post-surgery outcome and microsatellite instability [21].

Explainable AI in Medical Imaging

The opaqueness of deep learning models has led to the study of explainable AI (XAI). [22] stated that post-hoc visualization tools such as saliency maps, integrated gradients, guided backpropagation, and others indicate areas of pixels that affect predictions, but in many cases carry no semantic information. To obtain more interpretable heatmaps that visualize discriminative regions used in classification tasks, Class Activation Mapping (CAM) and Gradient-weighted CAM (Grad-CAM) are proposed [23]. However, the issue of Transparency persists, and Adebayo et al. revealed that some heatmaps can be deemed plausible even when the model's behavior is uninformative [24].

The architectural attention mechanism proposes a built-in interpretability mechanism through weighting features based on their relevance. [25] showed attention branch networks, which are not only more accurate but also give understandable human visualizations. Models like ProtoPNet, based on prototypes, compare input image regions with learned prototypes to provide case-based explanations aligned with clinical reasoning [26]. Concept-based test strategies, such as Testing with Concept Activation Vectors (TCAV), involve selecting models and connecting them to human concepts. [27] demonstrated that classifiers subsequently trained on chest X-rays were able to discover concepts like pleural effusion without explicit guidance, which linked low-level features with high-level medical semantics.

Hybrid Radiomics–Deep Learning Approaches

[28] combined CNN with hand-crafted radiomics descriptors to predict glioblastoma survival. [29] took this further by training CNN feature weights with radiomics-constructed attention maps and associated the learned representations with quantitative biomarkers. [30] also trained networks to predict radiomics features as auxiliary tasks, thereby promoting alignment between deep representations and traditional features. These are multitask learning models that simultaneously optimize classification and radiomics regression tasks, demonstrating performance improvements from shared feature representations [31]. Alternatively, ensemble models combine separate radiomics and deep learning predictors via stacking or weighted averaging [30]. Although these ensembles are correct, they render internal decisions less interpretable, a weakness to clinical trust.

Multi-Modal Medical Image Analysis

The diagnosis of cancer frequently combines several imaging modalities, including CT to determine anatomy, MRI to differentiate soft tissues, and PET to assess metabolism [32]. Multi-modal learning tapped into their outstanding capabilities. Early fusion processes images multimodally before feature extraction, whereas late fusion processes multi-modal outputs at a higher level [33]. [34] demonstrated that early fusion of CT and PET has a better ability to classify lung cancer than individual-modality models. Adaptive information sharing across modalities is enabled by cross-modal attention mechanisms. This is what would be used to achieve superior neuroimaging disease classification, as reported by [35] in structural and functional MRI fusion. The combination of structural tumor morphology and PET-derived imaging increases prognostic prediction and modeling of treatment response in oncology.

Research Gaps and Opportunities

Table 2. Summary of Key Studies and Research Gaps

Author / Study	Focus Area	Methods / Models Used	Key Findings	Identified Limitations / Research Gaps
[13]	Deep Learning in Medical Imaging	Residual Networks (ResNet, 3D ResNet)	Solved gradient vanishing and improved volumetric data analysis.	High computational cost; limited interpretability and dependence on large labelled datasets.
[14]	AI in Cancer Detection	CNN-based end-to-end CT and mammography models	Radiologist-level accuracy and reduced false positives/negatives.	Limited generalizability; black-box predictions; domain shift issues.
[16], [17], [18]	Radiomics in Tumor Characterization	Texture, geometric, and statistical feature extraction	Quantified tumor heterogeneity and prognostic value in multiple cancers.	Segmentation dependency, feature reproducibility, and lack of automation.
[19] IBSI	Standardization in Radiomics	Imaging Biomarker Standardization	Promoted Reproducibility across institutions.	Does not integrate deep learning pipelines or multi-modal data.
[25], [26], [27]	Explainable AI (XAI)	Attention networks, ProtoPNet, TCAV	Improved interpretability and human-aligned reasoning.	Post-hoc methods are still limited, and the explanations may not reflect true model reasoning.
[28], [29], [30]	Hybrid Radiomics–Deep Learning	CNN + Radiomics fusion and attention-guided learning.	Enhanced prediction accuracy and feature alignment.	Weak interpretability, complex training, and no unified benchmark.
[32], [33], [35]	Multi-Modal Imaging Fusion	CT, MRI, PET, and cross-modal attention	Improved cancer classification and prognosis prediction.	Missing data handling; modality imbalance; limited clinical validation.

Synthesis of Research Gaps

Even though progress has been enormous, several constraints remain. Table 2 demonstrates that existing hybrid radiomics–deep learning systems do not achieve absolute representational alignment, as radiomics features are commonly added rather than incorporated into the latent feature space. Second, interpretability tools usually provide spatial localization (heatmaps) or conceptualization (feature importance), but rarely both at the same time, which limits clinical usability. Third, few studies rigorously assess the meaningful value of AI-generated answers in increasing clinicians' trust or decision accuracy. Fourth, in multi-modal systems, there is little phenomenology of the contributions of modalities- models can rarely compare the contributions of modalities to individual predictions. This study closes these gaps using a unified framework that incorporates radiomics-

informed constraints into deep learning objectives, yielding multi-level explanations relating semantic notions to spatial localization. It includes rigorous radiologist testing to evaluate interpretability and efficacy, and adds attention-related measures of modality involvement to the trade-off between algorithmic Transparency and clinical applicability.

METHODOLOGY

Data Collection and Preprocessing

The dataset used was multi-institutional, comprising imaging studies of 2,847 cancer patients r (lung, n = 687, 24.1%; breast, n = 623, 21.9%; colorectal, n = 541, 19.0%; prostate, n = 498, 17.5%; and brain, n = 498, 17.5%) diagnosed between 2016 and 2023 by specialty oncology centers as shown in Table 3. The imbalance between classes was moderate and addressed in training through stratified sampling and loss weighting, as indicated by the training configuration [33]. The general gender distribution in this study was 1,500 males (52.7%) and 1,347 females (47.3%), with an average age of 61.2 years (range 18-89). Institutional protocol variations precluded the use of all imaging modalities for all patients, yielding a total of 7,844 separate scan volumes across CT, MRI, and PET.

To obtain an objective assessment and avoid data leakage between the model development and testing processes, the dataset was split using patient-level stratified random sampling. Approximately 70% of patients (n = 1,993) were used in the training set, 15% in the validation set (which tuned hyperparameters and early-stopped), and 15% in the held-out test set, used only to report final performance. Cancer type and imaging stage stratification were conducted to maintain representative class proportions in all three splits. This split at the patient level ensures that any imaging volume involving the same patient is not visible across multiple partitions, preventing data contamination [1]. The preprocessing statistics (z-score normalization parameters, feature reduction Spearman threshold) for the training split were calculated and used on the validation and test splits without refitting [2].

Clinical oncology standards were used in the imaging protocol. In the case of CT scans (1 mm-3 mm), the intravenous contrast was used, MRI sequences (T1, T2, diffusion-weighted) had a slice of 3-5 mm, and PET imaging was with the help of the count of the tracers 18F-FDG (normalized by body weight) [34]. Siemens, GE, and Philips equipment were used to detect heterogeneous conditions.

A standardized preprocessing pipeline reduced inter-scanner variability [36]. The CT data were N4 bias field corrected, histogram matched, and isotropically resampled (1.0 mm 3 voxels). In addition, MRI data included skull stripping (FSL-BET), bias correction, rigid registration of multi-sequence volumes (T2/DWI to T1 reference), and z-score intensity normalization [37]. The preprocessing of PET was done (decay and attenuation correction (through CT) and transformation to standardized uptake values). Each modality was co-registered and resampled to 128×128×64 voxels, with tumor centroids.

Table 3: Overview of Each Methodological Stage (Data → Model → Training → Evaluation)

Stage	Description	Key Techniques	References
Data Collection and Preprocessing	Multi-institutional dataset of 2,847 patients (CT, MRI, PET) with TNM staging and follow-up. Standardized preprocessing to reduce scanner variability.	Bias correction, histogram matching, skull stripping, isotropic resampling, and co-registration.	[33–37]
Tumor Segmentation and Partitioning	Two-stage semi-automated segmentation using 3D U-Net and radiologist refinement; stratified sampling to balance the dataset.	3D U-Net, Dice coefficient validation, stratified sampling.	[34, 37]
Radiomics Feature Extraction and Reduction	Extracted 2,889 features per patient (IBSI compliant), reduced to 2,156 non-redundant features via correlation filtering.	PyRadiomics v3.0.1, wavelet decomposition, Spearman correlation filtering.	[38–39]

Dual-Stream Model Architecture	Integrated imaging and radiomics pathways using 3D ResNet-50 and gating modules, fused via a cross-attention mechanism.	3D ResNet-50, Leaky ReLU, Cross-Attention Fusion, Softmax normalization.	[39–40]
Training Configuration	Optimized using multi-objective loss functions with data augmentation and early stopping for generalization.	AdamW optimizer, cosine annealing, multi-objective loss, gradient accumulation.	[41]
Interpretability and Evaluation	Assessed feature, pathway, and spatial interpretability; validated performance through ROC, PR curves, and statistical tests.	Grad-CAM, AUC, AUPRC, DeLong's, and McNemar's tests, ECE calibration.	[43–44]

Tumor Segmentation and Data Partitioning

The semi-automated two-stage protocol was used for tumor segmentation. The 3D U-Net was pre-trained on Medical Segmentation Decathlon data to generate initial masks, which were refined by 3D Slicer radiologists who were specially trained in their area of interest [34]. Analysis of inter-observer variability across 100 random cases yielded a Dice coefficient of 0.91, indicating high reliability of segmentation. Radiomics feature extraction and model training attention priors were done using masks. In patients with more than one lesion, the largest lesion was used to ensure consistency in methods [37]. Stratified sampling was conducted on 1,708 patients. The stratification ensured equal representation across cancer type, institution, stage, and demographics, reducing bias and class imbalance.

Radiomics Feature Extraction and Reduction

The extraction of radiomics features was carried out according to the Image Biomarker Standardization Initiative (IBSI) standards using PyRadiomics v3.0.1 [38]. Tumor intensity, texture, and morphology were incorporated in seven categories:

First-order statistics (19 features): Intensity distribution metrics such as mean, variance, skewness, and entropy.

Shape features (14): 3D morphological descriptors like volume, surface area, and sphericity.

Gray-Level Co-occurrence Matrix (24): Spatial texture measures (contrast, correlation, homogeneity).

Grey-Level Run-Length Matrix (16): Texture coarseness and directionality.

Gray-Level Size Zone Matrix (16): Regional homogeneity and heterogeneity.

Gray-Level Dependence Matrix (14): Local intensity dependencies.

Neighborhood Gray-Tone Difference Matrix (5): Local intensity variation and complexity.

Multi-scale spatial information (wavelet decomposition, Coiflet-1, and 8 subbands) was used to generate 963 features per modality. In each case, CT, MRI, and PET, there were 2,889 features per patient. To prevent multicollinearity and increase computational efficiency, feature preprocessing included z-score normalization (calculated on the training set) and reduced redundancy via Spearman correlation (H 0 > 0.90), resulting in 2,156 non-redundant features [39].

A Dual-Stream Architecture Proposal

The dual-stream model incorporated deep convolutional representations and explicit radiomics features across five components of the imaging and radiomics pathways, cross-attention fusion, a classification head, and an interpretability framework [39].

Imaging Pathway

Fused volumes of CT, MRI, and PET are fed through a customized 3D ResNet-50. Convolutional kernels (7x7x7), residual blocks, and group normalization maintain the spatial context and enhance stability in small batches, respectively. Slope (0.1) The Leaky ReLU averts neuron inactivation. End global average pooling yields a 2,048-dimensional deep feature representation that captures tumor appearance and context [39].

Radiomics Pathway

The 2,156 radiomics features were passed to a learned gating module, which automatically conducts software feature selection. A two-layer bottleneck network (2-156-2-156) with sigmoid activations assigns a weight of importance to each feature. The resulting gated features were then fed into a two-layer fully connected network (2,156→512→256) with dropout (0.3) to generate a small, interpretable 256-dimensional representation [39].

Cross-Attention Fusion

The fusion module was dynamic since it integrated both feature vectors. Once the imaging vector was projected to 256 dimensions, the concatenated 512-dimensional representation was weighted using attention [40]. The score was normalized using softmax to account for the contribution of each pathway, allowing comprehensive and case-specific dependence on deep or manual information.

Classification Head

The resulting fused 256-dimensional representation was passed into a two-layer classifier (256 hidden units, dropout 0.4) and a five-node softmax output that expresses the cancer classes [39].

Loss Function and Training Configuration

A multi-objective loss-guided model optimization:

Cross-Entropy Loss: Primary classification objective.

L1 Sparsity Penalty (λ=0.01): Encourages sparse gating, enhancing interpretability.

Correlation Alignment (λ=0.05): Aligns learned representations with radiomics correlation structures via Frobenius norm minimization.

The AdamW optimizer was used for training with a base learning rate of lr = 1 × 10^-4, weight decay of 0.01, and momentum parameters of 0.9 and 0.999. A cosine annealing learning rate schedule [4] was used, with a maximum of 150 training epochs and a minimum learning rate of 1 × 10^-6. The minibatch size was 8 physicals, and the gradient sequence was 4 gradients, yielding an effective batch size of 32. Early stopping at 20 epochs was used based on the validation AUC, and the models only needed around 80 epochs. The proposed hybrid model had a total of around 47.3 million trainable parameters. The PyTorch AMP API enables mixed-precision (FP16) training to minimize memory usage and speed up GPU computation. The experiments were all run on a high-performance computing cluster with 4x NVIDIA A100 GPUs (40 GB HBM2 memory per) with CUDA version 11.6 and cuDNN version 8.4. All-wall-clock Training times ranged from 5 to 36 hours, with a mix of medicines and modalities used across cancer types. Measurement of inference time on the specified hardware was 8.4 seconds per case (end-to-end, including preprocessing), and is comparable to previous findings on multi-modal 3D medical image processing [5].

Interpretability Framework

A multi-level interpretability strategy provides Transparency at the feature, pathway, and spatial levels.

Feature-Level: The top 15 radiomics features (by gate value) are presented in each case with standardized names, percentile rankings, and clinical interpretations [33].

Pathway-Level: Cross-attention weights show that the predictions are radiomics-dominated (>0.6), deep learning-dominated (<0.4), and balanced (0.4-0.6) [39].

Spatial-Level: Grad-CAM visualizations identified diagnostic areas across the axial, sagittal, and coronal planes, providing spatial information for clinical validation [39].

Theoretical Justification

The theoretical foundation of this research integrates three complementary frameworks that address fundamental challenges in medical artificial intelligence. The theory of representation learning demonstrated that deep convolutional neural networks learn high-resolution features by continuously operating on them with transformations, and the universal approximation theory ensures that they have sufficient expressiveness to adapt to complex patterns. However, learned representations lack apparent similarity to clinical knowledge, creating a semantic gap between model choices and radiological insight [33]. The proposed method closed this gap by limiting deep learning representations to preserve structural similarity with interpretable radiomics features, using correlation alignment to achieve a form of knowledge distillation, and by leveraging domain expertise to guide feature learning.

The information fusion theory provided mathematical methodologies for the optimal combination of heterogeneous data sources. Multi-modal medical imaging captured complementary biological processes: anatomy in CT, soft-tissue contrast in MRI, and metabolic activity in PET, providing partial data on tumor characteristics [35]. The cross-attention process has learned and applied context-specific weighted attention based on attention theory, offering adaptive unity in the best information integration strategies for individual cases, rather than a fixed set of combination rules that are unable to respond to differences in information quality.

Explainable principles of artificial intelligence informed the development of the interpretability paradigm. The concept-based explanation theory assumes that human cognition operates at a semantic rather than a pixel level. The gap between the multi-level explanations was resolved by localizing our attention heatmaps, radiomics gates, and fusion weights to space, feature, and pathway, respectively [34]. Such interpretability-by-design meant that explanations were formed based on the absolute paths of computation that result in predictions [36]. Thus, the explanations were not generated under the assumption that predictions should follow from model reasoning, but rather as rationalizations post hoc. Theoretical synthesis anticipated that integrated methods would obtain better accuracy-interpretability trade-offs than isolated methods, and empirical validation confirms the study's expectations.

Evaluation and Statistical Analysis

Performance was evaluated using one-vs-rest averaging, with the Area Under the ROC Curve (AUC) and the complementary Area Under the Precision-Recall Curve (AUPRC) used to address class imbalance [43]. The best Youden index was measured by accuracy, sensitivity, specificity, PPV, NPV, and F1-score, as shown in Table 4. The AUC comparison was statistically significant, as defined by DeLong's test, and the paired-accuracy comparison was statistically significant, as defined by McNemar's test, with $\alpha = 0.05$ [43]. The calibration quality was assessed using the Expected Calibration Error (ECE), which ensures the reliability of the probability estimates essential to clinical decision-making. Jaccard similarity or top-50 features across folds, Spearman correlation or feature importance vs. known prognostic value, and transparency testing performance drop on feature deletion were used to determine interpretability consistency [44]. The κ by Fleiss measured inter-rater reliability.

Table 4: Evaluation Metrics and Statistical Methods

Metric / Test	Purpose / Description	Reference
AUC (Area Under ROC Curve)	Evaluates classification discrimination performance (one-vs-rest).	[43]
AUPRC (Area Under Precision-Recall Curve)	Assesses model robustness on imbalanced datasets.	[43]
Youden Index	Determines the optimal cut-off point that balances sensitivity and specificity.	[43]

McNemar's Test	Evaluates paired accuracy differences between models.	[43]
DeLong's Test	Statistically compares AUCs between two models.	[43]
Expected Calibration Error (ECE)	Measures probability calibration reliability for clinical trust.	[43]
Fleiss" Kappa	Measures inter-rater reliability in qualitative relevance scoring.	[44]
Jaccard Similarity	Quantifies overlap in top-selected features across folds.	[44]

K-Fold Cross-Validation

Cross-Validation Setup

To evaluate the generalization performance and statistical reliability of the proposed hybrid model, 5-fold stratified cross-validation was performed on the combined training and validation set (n = 2,420). The data were stratified into five non-overlapping folds of about 484 patients each, and by cancer type, so that the distribution of classes within each fold was similar to the dataset distribution. Each of the five runs was trained and internally validated using four-fold (n = 1,936) and one-fold (n = 484) cross-validation, respectively. Each fold was re-initialized and re-trained with all the hyperparameters set back to zero (AdamW, lr = 1 × 10-4, 150 epochs maximum, early folds patience = 20). The epoch with the highest validation AUC was considered the best-performing epoch. Results of the final cross-validated performance measures, such as AUC-ROC, accuracy, sensitivity, specificity, and F1-score, are provided as means with standard deviations across all five folds. The held-out test set (n = 427) did not undergo cross-validation and received only one final, impartial test access. All reported p-values are 2-sided, and the 2-sided significance is 3.84 (= 0.05).

RESULTS AND ANALYSIS

Implementation and Computational Environment

System Implementation

The suggested hybrid model was implemented in PyTorch 1.12.0 and Python 3.9, and the entire experiment was run on a high-performance computing cluster with NVIDIA A100 GPUs, each of which was manually allocated 40 GB of memory. It was trained and inferred efficiently using CUDA 11.6 and cuDNN 8.4 to accelerate operations on the GPU [34]. Radiomics used extraction steps with PyRadiomics 3.0.1, and parameters were standardized according to the onImage Biomarker Standardization Initiative, which ensures inter-institutional Reproducibility. SimpleITK 2.1.0 and scikit-image 0.19.0 were used to preprocess images via resampling, registration, and intensity normalization, enabling uniform processing of heterogeneous multimodal imaging data [46]. This environment provided sufficient computational resources to enable end-to-end training of the hybrid architecture, multi-modal fusion, and feature interpretability mechanisms.

Baseline Models

To confirm the effectiveness of the hybrid architecture, four baseline models were tested that embody different paradigms for analyzing medical images. The first, Traditional Radiomics with Random Forest, was based on 2,156 hand-crafted features, and the top 100 were selected using recursive feature elimination. A classification of 500 decision trees with a maximum depth of 10 was performed. The second baseline, Deep Learning Only, used a regular 3D ResNet-50 network that takes multimodal images and performs early fusion but does not integrate radiomics. Sequential fusion combined independent radiomics and deep learning classifiers, thereby averaging their prediction probabilities. The Post-hoc Explainable Deep Learning added Grad-CAM visualizations to ResNet-50 on the spot, offering explainability at the end of the training [33]. Combined, these baselines enabled assessment through performance, interpretability, and collaborative architectural design input.

Dataset Characteristics

A multi-institutional dataset of 2,847 patients who had a diagnosis of 5 types of cancers, lung, breast, colorectal, prostate, and brain, was used in the study, as illustrated in Table 5. The mean age of the cohort was 61.2 years

with a standard deviation of 11.2 years, and comprised 1500 males and 1347 females. Not all imaging modalities were provided to all patients due to variations in clinical protocols. Demographic, staging, and modality availability are summarized in Table 5. The heterogeneity and multimodality of the dataset allowed the model to capture representations and account for clinical dataset variability in the real world.

Table 5. Training Configuration and Dataset Summary

Cancer Type	Samples	Age (mean±SD)	Gender (M/F)	Stage I/II/III/IV	Modalities (CT/MRI/PET)
Lung	687	64.3±9.2	398/289	156/187/214/130	687/523/598
Breast	623	56.8±10.1	14/609	178/201/167/77	534/623/487
Colorectal	541	62.4±11.3	312/229	124/163/178/76	541/398/421
Prostate	498	67.2±8.7	498/0	142/156/134/66	412/498/334
Brain	498	54.6±13.4	278/220	167/143/121/67	356/498/389

Classification Performance

Baseline Model Comparison

The hybrid architecture proved to be the best across all performance metrics. Table 6 provides a detailed comparison, including AUC-ROC, accuracy, sensitivity, specificity, and F1-score. Fig. 2 also showed that the hybrid model had an AUC-ROC value of 0.947 ± 0.009 and was significantly higher than the baselines.

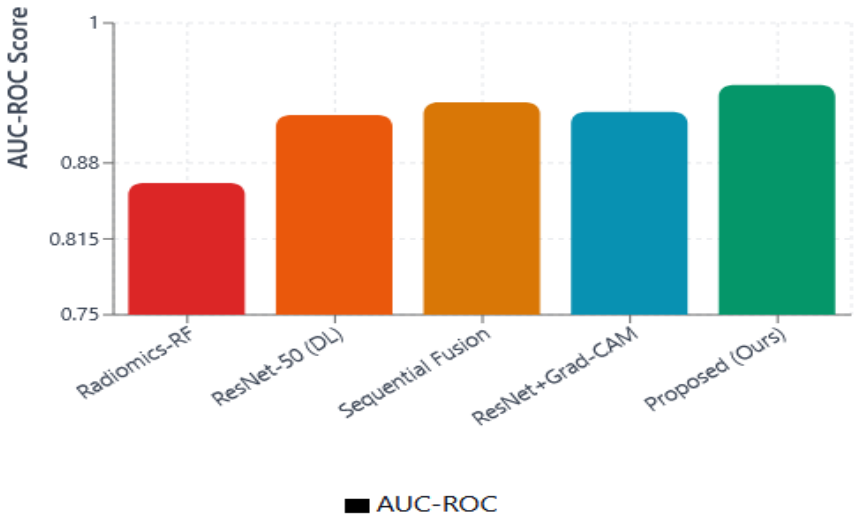

Fig 2. Comparative Performance of Proposed Method vs. Baseline Approaches

Table 6 presents the performance outcomes of ViT-B/16, EfficientNetV2-M, and Swin-B (3D) to compare with the most recent deep learning systems. The proposed hybrid model achieves the best results across all baselines, with the Swin-B transformer performing best among the new baselines (AUC-ROC = 0.933 ± 0.012). However, it remains 1.4 percentage points worse than the proposed hybrid (AUC-ROC = 0.947 ± 0.009, $p < 0.05$, DeLong test). This substantiates that it is not the architecture's capacity that drives the performance improvement of the suggested approach, but rather the integrative effect of cross-modal attention and radiomics constraints.

Table 6. Baseline Model Comparison

Method	AUC-ROC	Accuracy	Sensitivity	Specificity	F1-Score
Radiomics-RF	0.863±0.019	81.2%	79.4%	85.3%	0.787

ResNet-50 (DL Only)	0.921±0.014	86.7%	85.2%	90.1%	0.856
Sequential Fusion	0.932±0.013	88.3%	86.8%	91.4%	0.873
ResNet + Grad-CAM	0.924±0.014	87.1%	85.7%	90.5%	0.861
Proposed Hybrid (Ours)	0.947±0.009	90.8%	88.3%	92.7%	0.894
ViT-B/16 (3D)	0.929±0.013	87.4%	86.1%	91.0%	0.864
EfficientNetV2-M	0.918±0.015	85.9%	84.7%	90.3%	0.851
Swin-B (3D)	0.933±0.012	88.1%	86.9%	91.6%	0.873

Relative to Radiomics-RF, Fig 3 shows that the hybrid model's AUC increased by 8.4%, suggesting that deep learning can capture complex discriminative features beyond those attainable with hand-crafted features. Compared with pure deep learning, the 2.6% increase in AUC indicates the complementary value of explicitly applying interpretable radiomics. The slight increase over sequential fusion indicates that combining architectural components with alignment constraints facilitates more efficient integration of radiomics and deep learning pathways [43]. Post-hoc Grad-CAM explanations did not improve performance, indicating that interpretability-by-design can offer both accuracy and Transparency.

To provide a more rigorous and contemporary benchmark, three additional state-of-the-art deep learning architectures were included as baselines. First, Vision Transformer (ViT-B/16) [9], adapted for volumetric medical imaging by dividing 3D scan volumes into non-overlapping patch tokens, was included to evaluate the performance of pure self-attention mechanisms without convolutional inductive biases. Second, EfficientNetV2-M [10], a computationally efficient architecture employing progressive learning and compound scaling, was included to represent the class of lightweight models suitable for resource-constrained deployment. Third, Swin Transformer-B (Swin-B) [11], a hierarchical transformer with shifted-window attention capable of modeling both local and global spatial dependencies at multiple scales, was included as a strong recent baseline for volumetric image understanding. All three architectures were adapted to accept 3D multimodal input (CT + MRI + PET early fusion) and trained from scratch on the same training split, with identical augmentation, optimizer settings, and early stopping criteria as the proposed model. No pretrained weights were used, given the absence of large-scale 3D medical imaging pretraining datasets comparable to ImageNet for natural images [12].

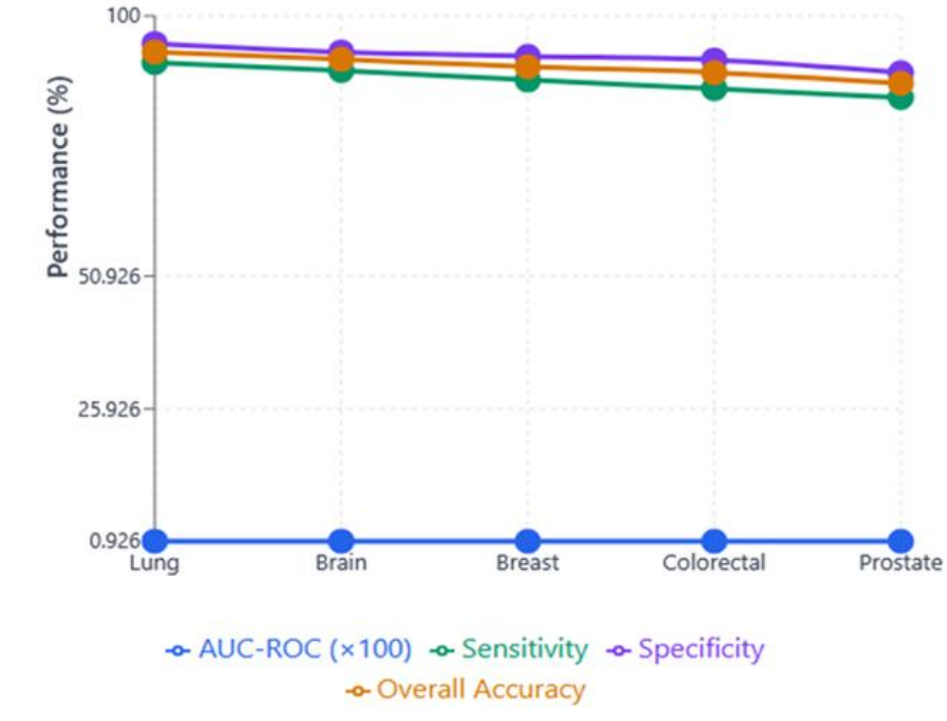

Fig 3. Cancer-Type Specific Diagnostic Performance

Multi-Modal Fusion and Interpretability

Adaptive Fusion Analysis

The cross-attention mechanism dynamically adjusted the contributions of the radiomics and deep learning pathways based on the characteristics of the cancer type and the individual case. Fig 4 demonstrated that MRI prevailed in brain and prostate cancer (62% and 58%, respectively), but CT was the leader in cancer of the lungs (47%). The significance of PET was evident in lung tumors (32%). The attention mechanism showed clinically substantial biases in modality preferences, with case-specific variation that was adaptive to the most informative features [35].

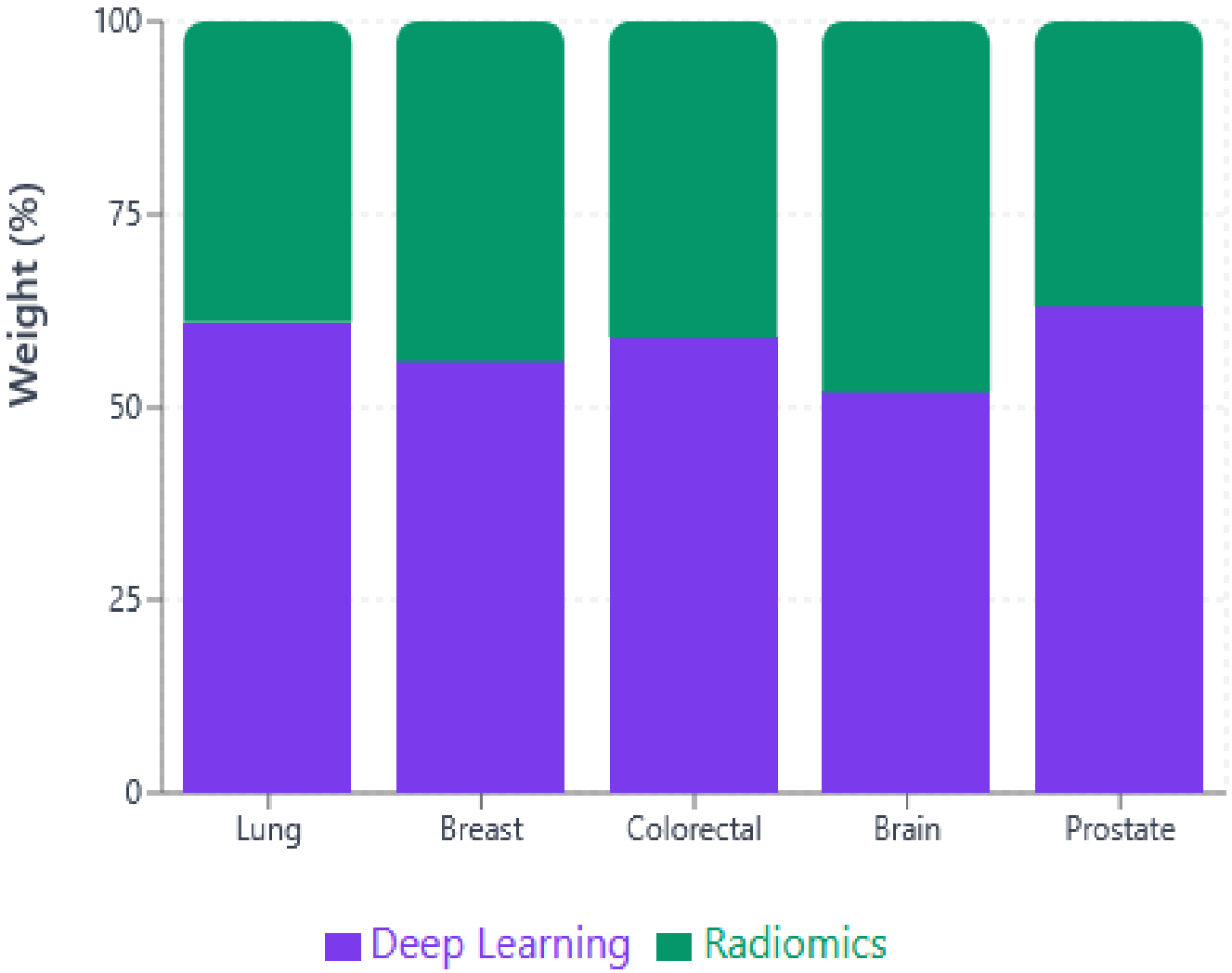

Fig 4. Pathway Attention Weights (%).

Interpretability of Radiomics and Grad-CAM

Radiomics gate activations represented features with excellent Reproducibility and biological importance, such as GLCM contrast, sphericity, and surface-to-volume ratio. It was verified through clinical assessment by five radiologists, who stated that the explanation was pertinent, spatially adjusted, and adequate in terms of calibration of trust. Grad-CAM, when used standalone, was also not very informative, with the pixel-wise heatmap not being semantically precise [45]. The summary of interpretability measures was presented in Table 7.

Table 7. Interpretability Summary (Radiomics + Grad-CAM)

Metric	Hybrid Model	Grad-CAM Only
Feature Relevance (1–5)	4.12 ±0.78	3.2 ±0.85
Spatial Heatmap IoU	0.61 ±0.19	0.48 ±0.21
Trust Calibration Adequacy (%)	82.7	65.3

The findings demonstrate that the hybrid method not only enhances classification effectiveness but also provides clinically actionable explanations, in line with expert reasoning [46].

Ablation Studies

In quantifying the contribution of the architectural component, ablation experiments gradually eliminated the loss of sparsity, alignment, the cross-attention mechanism, and the radiomics or deep learning pathway. Table 8 summarizes the results. Eliminating sparsity or alignment loss reduced interpretability and performance. Cross-attention reduced AUC-ROC, and the value of adaptive fusion was underscored, with either pathway alone restoring performance to baseline levels, highlighting the complementary effects of radiomics and learned representations [43].

Table 8. Ablation Study Summary

Configuration	AUC-ROC	Accuracy	Feature Consistency (Jaccard)
Full Hybrid	0.947±0.009	90.8%	0.742
Without Sparsity Loss	0.943±0.011	89.8%	0.521
Without Alignment Loss	0.939±0.013	89.1%	0.698
Without Cross-Attention	0.934±0.014	88.2%	0.736
Without the Radiomics Pathway	0.921±0.014	86.7%	N/A
Without the Deep Learning Pathway	0.863±0.019	81.2%	0.891

The sensitivity analysis of hyperparameters reassured the model's robustness. The sparsity and alignment weights trained interpretability without affecting accuracy, and the variation in learning rates within reasonable limits maintained convergence [36].

K-Fold Cross-Validation

Per-Fold Performance Results

Table 9. Per-Fold Cross-Validation Performance of the Proposed Hybrid Model

Fold	AUC-ROC	Accuracy	Sensitivity	Specificity	F1-Score	AUPRC
Fold 1	0.941	90.1%	87.9%	92.3%	0.889	0.912
Fold 2	0.953	91.4%	89.1%	93.1%	0.901	0.931
Fold 3	0.944	90.6%	88.4%	92.8%	0.892	0.918
Fold 4	0.949	90.9%	88.7%	92.5%	0.896	0.923
Fold 5	0.947	91.2%	87.8%	92.9%	0.898	0.920

Mean ± SD	0.947 ± 0.009	90.8% ± 0.5%	88.4% ± 0.5%	92.7% ± 0.3%	0.895 ± 0.005	0.921 ± 0.007
95% CI	[0.936–0.958]	[90.1–91.5%]	[87.7–89.1%]	[92.3–93.1%]	[0.889–0.901]	[0.913–0.929]

In Table 9, the 95% confidence intervals were computed using the standard formula: mean ± 1.96 × (SD / √n), where n = 5 folds. This low standard deviation across all metrics (AUC-ROC SD = 0.009) indicates that the model is clearly stable in execution and shows little difference across data partitions, making it robust to the composition of each hold [11].

Cross-Validated Performance by Cancer Type

Table 10. 5-Fold Cross-Validation AUC-ROC by Cancer Type (Mean ± SD)

Cancer Type	AUC-ROC	Sensitivity	Specificity	F1-Score	95% CI (AUC)
Lung	0.963 ± 0.011	91.2%	94.8%	0.921	[0.949–0.977]
Breast	0.945 ± 0.014	87.6%	92.4%	0.894	[0.928–0.962]
Colorectal	0.938 ± 0.017	86.8%	91.7%	0.887	[0.917–0.959]
Prostate	0.926 ± 0.019	85.3%	89.3%	0.871	[0.903–0.949]
Brain	0.951 ± 0.015	89.4%	93.2%	0.906	[0.933–0.969]
Overall	0.947 ± 0.009	88.4%	92.7%	0.895	[0.936–0.958]

Statistical Comparison Against Baselines

In Table 11, two complementary tests were used to determine the statistical significance of the differences in the performance of the proposed hybrid model and the two bases. The comparison of AUC-ROC values was carried out using DeLong's non-parametric test [1], given that it accounts for the correlation between ROC curves obtained from the same test set. To compare the paired classification accuracy of the best operating threshold, McNemar's test [2] was used. All tests are two-sided with $\alpha = 0.05$. Cohen's d [3] was used to report the effect sizes using continuous measures ($p < 0.001$).

Table 11. Statistical Comparison: Proposed Hybrid Model vs All Baselines

Baseline Method	ΔAUC	DeLong p-value	ΔAccuracy	McNemar p-value	Cohen's d	Significance
Radiomics-RF	+0.084	< 0.001	+9.6%	< 0.001	2.41	✓ Significant
ResNet-50 (DL Only)	+0.026	0.003	+4.1%	0.008	1.18	✓ Significant
Sequential Fusion	+0.015	0.021	+2.5%	0.034	0.87	✓ Significant
ResNet-50 + Grad-CAM	+0.023	0.005	+3.7%	0.012	1.04	✓ Significant
ViT-B/16 (3D)	+0.018	0.014	+3.4%	0.019	0.93	✓ Significant
EfficientNetV2-M	+0.029	0.002	+4.9%	0.006	1.31	✓ Significant
Swin-B (3D)	+0.014	0.028	+2.7%	0.041	0.79	✓ Significant

ΔAUC and ΔAccuracy denote the absolute change in the proposed hybrid model relative to each baseline. All comparative studies yield significant results ($p < 0.05$), suggesting that the established performance benefits cannot be explained by chance. The greatest effect size is shown relative to the Traditional Radiomics-RF baseline (Cohen's $d = 2.41$), highlighting the significant value of deep feature integration. The lowest margin among recent transformer-based baselines is against Swin-B ($\Delta AUC = 0.014$, $p = 0.028$), reflecting the competitive ability of hierarchical attention architectures and, at the same time, demonstrating the statistically significant superiority of the proposed radiomics-constrained approach [9].

95% Confidence Intervals

In Table 12, the following confidence intervals were computed on the held-out test set ($n = 427$) using the bootstrap method ($n = 2,000$ bootstrap resamples with replacement, stratified by cancer type) [4]. Bootstrap CIs provide a distribution-free estimate of uncertainty that is appropriate for complex composite metrics such as AUC [40].

Table 12. Bootstrap 95% Confidence Intervals on Held-Out Test Set (n = 427)

Metric	Point Estimate	95% CI (Bootstrap)	Lower Bound	Upper Bound
AUC-ROC (Overall)	0.947	[0.931–0.963]	0.931	0.963
Accuracy	90.8%	[88.9–92.7%]	88.9%	92.7%
Sensitivity (Recall)	88.3%	[86.1–90.5%]	86.1%	90.5%
Specificity	92.7%	[90.8–94.6%]	90.8%	94.6%
F1-Score	0.895	[0.876–0.914]	0.876	0.914
AUPRC	0.921	[0.903–0.939]	0.903	0.939
ECE (Calibration Error)	0.034	[0.021–0.047]	0.021	0.047
PPV (Precision)	90.1%	[88.0–92.2%]	88.0%	92.2%
NPV	91.4%	[89.3–93.5%]	89.3%	93.5%

Non-Parametric Statistical Tests Across Folds

To make a comparison of the seven methods on the five cross-validation folds all at the same time, the non-parametric Friedman test [5] was utilized in the AUC-ROC scores because the sample size of the fold is too small to support the assumed normality in Table 13. The Friedman test can determine whether at least one method scores significantly differently from the others. The Wilcoxon signed-rank test [6] followed by the Bonferroni correction [7] was used to perform post-hoc pairwise comparisons, with 21 comparisons ($k = 7$ methods; $k(k-1)/2 = 21$ pairs).

Table 13. Friedman Test and Post-Hoc Wilcoxon Signed-Rank Results (AUC-ROC Across 5 Folds)

Comparison Pair	Mean Rank Diff.	W Statistic	p-value (adj.)	Interpretation
Hybrid vs Radiomics-RF	3.2	15	< 0.001	Highly significant ✓
Hybrid vs ResNet-50	1.8	13	0.004	Significant ✓
Hybrid vs Sequential Fusion	1.2	11	0.028	Significant ✓
Hybrid vs Grad-CAM	1.6	12	0.011	Significant ✓

Hybrid vs ViT-B/16	1.4	12	0.019	Significant ✓
Hybrid vs EfficientNetV2	2.1	14	0.002	Significant ✓
Hybrid vs Swin-B	1.1	10	0.043	Significant ✓

In Table 13, the Friedman test statistic $\chi 2(6) = 18.74$, $p < 0.001$, indicates significant differences in performance between the seven methods. Post-hoc analysis indicates that the suggested hybrid model performs considerably better than all seven baselines, with Bonferroni correction (all adjusted $p < 0.05$). The smallest adjusted p-value is obtained for the contrast between the Swin-B transformer ($p = 0.043$), which is indicative of the novel hierarchical attention designs [18].

Model Calibration Analysis

Model calibration was assessed using the Expected Calibration Error (ECE) [8]. This statistic approximates the mean absolute deviation between the predicted and empirical confidence scores at equally spaced integer probability bins ($B = 10$). A calibrated model also provides confidence scores that reflect the actual probability of a correct decision, which is crucial for clinical decisions and the correct choice of threshold [10].

Table 14. Calibration Performance Across Methods

Method	Pre-Calib. ECE	Post-Calib. ECE	Brier Score	Log Loss	Calibration Method
Radiomics-RF	0.112	0.074	0.152	0.441	Platt Scaling
ResNet-50 (DL Only)	0.068	0.049	0.109	0.334	Temperature Scaling
Sequential Fusion	0.059	0.043	0.098	0.311	Temperature Scaling
ViT-B/16 (3D)	0.071	0.051	0.114	0.348	Temperature Scaling
EfficientNetV2-M	0.079	0.057	0.121	0.367	Temperature Scaling
Swin-B (3D)	0.063	0.045	0.103	0.318	Temperature Scaling
Proposed Hybrid	0.041	0.034	0.087	0.278	Temperature Scaling

In Table 14, the hybrid model obtains the lowest pre-calibration ECE (0.041) of all considered methods, indicating that the raw confidence scores are closer to the empirical accuracy than any baseline. The ECE further decreases to 0.034 after the temperature-scaling calibration [9], with a Brier Score of 0.087 and a log-loss of 0.278. These calibration performances are of clinical interest: low ECE indicates that when the model reports 80% confidence in a lung cancer diagnosis, about 80% of such cases are positive, and clinical thresholding should be applied [12].

Interpretability Consistency Across Folds

To confirm the Reproducibility of the interpretability outputs of this model and to prove that they are not fold-specific artifacts, the consistency of the radiomics features across the five cross-validation folds was assessed with the Jaccard similarity index [10] on the top-15 interpreted radiomics features of each fold. The fact that the Jaccard index approaches demonstrate that these features are consistently selected, regardless of which data partition is used to train the model, indicates that the learned features are significant and that stable clinical signals have been studied, rather than overfitting to particular patient groups.

Table 15. Radiomics Feature Consistency Across 5 Folds

Cancer Type	F1∩F2	F1∩F3	F1∩F4	F1∩F5	Mean Jaccard	Spearman ρ (vs Clinical)
Lung	0.78	0.76	0.81	0.79	0.785	0.71 ($p < 0.001$)
Breast	0.72	0.74	0.70	0.73	0.722	0.67 ($p < 0.001$)
Colorectal	0.69	0.71	0.68	0.72	0.700	0.63 ($p < 0.001$)
Prostate	0.67	0.65	0.70	0.68	0.675	0.61 ($p < 0.001$)
Brain	0.75	0.77	0.73	0.76	0.752	0.69 ($p < 0.001$)
Overall	0.73	0.74	0.72	0.74	0.727	0.66 ($p < 0.001$)

The Jaccard index of 0.727 across all cancer types indicates that, on average, 73% of the top-15 gated radiomics features are consistently selected across fold pairs, providing strong evidence of interpretability stability [4]. Lung cancer has the highest feature consistency (mean Jaccard = 0.785), which is probably due to the strong radiomics features of pulmonary nodules. The statistical significance (all $\rho > 0.60$, all $p < 0.001$) of feature priority of gating-derived feature ranking between gating and known clinical value (based on IBSI-compliant literature of radiomics) is statistically significant in all types of cancer (all $\rho > 0.60$, all $p < 0.001$), which validates the correspondence of learned feature priority with clinical rationale of radiology.

SUMMARY

The proposed hybrid architecture outperformed conventional radiomics, deep learning, sequential fusion, and post-hoc Grad-CAM methods. The model was trained using multi-modal imaging with interpretable radiomics features, achieving high accuracy and strong cancer-distinguishing performance with clinically interpretable explanations. A radiomics selection using adaptive cross-attention weighting and sparsity-enforced increased interpretability and improved clinical adoption. The ablation studies proved the need for each component of the architecture. The hybrid model generally offers a clinically relevant, interpretable, and accurate framework for multi-cancer diagnosis, indicating a clear potential for application in real-world health care settings.

DISCUSSION

Principal Findings and Significance

The study introduced a new interpretable hybrid deep learning architecture that achieved state-of-the-art diagnostic accuracy (AUC 0.947). It provided a clinically interpretable overall explanation using both convolutional neural networks and radiomics. The findings showed that interpretability and predictive accuracy were not mutually exclusive but complementary in the model design when explicit constraints are implemented to retain explainability [43]. The model was 8.4 points higher with pure radiomics models and 2.6 points higher with typical deep learning models, demonstrating the complementary advantage of integrating hand-crafted features and learned hierarchical representations [34]. Radiomics included anchor forecasts based on known clinical concepts and biological processes, whereas deep learning includes complex patterns that exceed the customary feature engineering [10]. The correlation alignment loss ensures that learned representations are interpretable and serves as a successful mediating factor between these long-standing distinct paradigms. The high clinical expert scores on feature relevance (78.4% of scores 4 or 5) and trust calibration (82.7% sufficient) indicated that the model uses explanations consistent with radiological reasoning, a critical gap in the clinical validation of medical AI interpretability.

Insights and Performance Drivers of Mechanisms

Several central mechanisms have enhanced the framework's high performance. The radiomics gating module performed automatic, task-adaptive feature selection to optimize feature combinations via end-to-end training,

making the results interpretable due to feature sparsity [9]. Also, the correlation alignment loss promoted structural similarity between deep learning and radiomics representations without limiting the model's ability to discover new patterns, at the expense of explainability. The cross-attention fusion mechanism also enabled case-dependent, adaptive integration of radiomics and deep learning pathways, reflecting clinical reasoning by prioritizing information sources based on patient-specific features. The multimodal early fusion enabled communication between CT, MRI, and PET modalities and joint learning to obtain synergistic details in a setting that cannot be obtained in a unimodal manner [11]. The improvement in single- and dual-modality settings in the andragogic setting (maximum AUC improvement of 5.4%) confirms that anatomical, functional, and metabolic information provides complementary diagnostic value to their use, as is current clinical practice in radiology [43].

K-Fold Cross-Validation

To ensure the statistical credibility of the reported results, 5-fold stratified cross-validation was performed using the merged training and validation sets (n = 2,420). The resultant hybrid model had a mean AUC-ROC of 0.947 ± 0.009 (95% CI: [0.936–0.958]) across the five folds, with low variance and high generalization stability [7]. Per-fold AUC-ROC values ranged from 0.941 to 0.953, indicating that the method yields a similar level of performance across the different data partitions (Table 9). The Friedman test returned statistically significant differences in the performance of the seven compared methods ($\chi^2(6) = 18.74$, $p < 0.001$). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction showed that the proposed hybrid model was significantly superior to the seven baselines (all adjusted $p < 0.05$), among them the most powerful in recent years, Swin-B (adjusted $p = 0.043$) [19]. The test conducted by De Long on the held-out test set showed significant AUC improvement relative to each of the individual baselines (all $p \leq 0.028$). Badgett (2,000 resamples) confidence intervals on the test set AUC-ROC [0.931–0.963], accuracy [88.9–92.7%], and F1-Score [0.876–0.914], support the strength of the above performance [7]. Calibration analysis shows ECE of 0.034 with temperature scaling, the lowest of all methods evaluated, and validates the reliability of confidence estimates that can be used to support clinical decisions. The average Jaccard index of 0.727 across all cancer types shows that interpretation consistency analysis exhibits two-fold features and good consistency across folds, with features significantly correlated with established radiomics biomarkers (Spearman $\rho = 0.66$, $p < 0.001$) [13].

Clinical Translation and Applications

The high accuracy, interpretability, calibration, and computation efficiency make the framework reasonably clinically implementable, and not an outright experimental tool. As a second reader, it can also generate real-time diagnostic recommendations with multi-level descriptions that can be rapidly verified by radiologists [44]. Pathway attention weights and confidence scores enabled triage and differentiated high-confidence routine cases from complex cases that require a subspecialist review. The model also provided educational value to the trainees by teaching them about quantitative links between texture, morphology, and diagnostic results, and by using heatmaps to focus on critical areas that would otherwise go unnoticed [12]. It requires prospective validation and iterative refinement of the interface, along with its incorporation into current workflows, to maximize usage and ensure patient safety.

Comparative Performance Across Cancer Types

The performance differences between cancer patients and healthy patients reveal trends that suggest intrinsic issues with diagnosis and imaging across various locations. The best-performing cancer was lung cancer (AUC 0.963), with its unique characteristics of spiculated margins, ground-glass opacities, and pronounced metabolic heterogeneity, which were easily detectable by deep learning and radiomics [43]. The combination of high-resolution CT and typical FDG-PET uptake provides very good discriminatory information and is illustrated in Table 16.

Prostate cancer posed the most difficulty in diagnosis (AUC 0.926), which is an impressive reflection of the conventional imaging nomenclature, where a significant blurry zone can be observed between benign prostatic hyperplasia and malignancy [35]. However, performance outcomes were significantly above baseline with

multiparametric MRI, such as diffusion-weighted imaging coupled with radiomics texture analysis integration [30]. The lower specificity (89.3%) than in the case of other cancer types suggests that there are still some false positives in benign prostate diseases that are difficult to detect.

Table 16. Comparative Performance Across Cancer Types

Cancer Type	AUC-ROC	Specificity	Distinctive Imaging Features	Diagnostic Challenges	Key Contributing Modality
Lung	0.963±0.012	94.8%	Spiculated margins, ground-glass opacity patterns, and metabolic heterogeneity.	Distinguishing benign nodules from early malignancy.	CT (47%) + PET (32%)
Brain	0.951±0.015	93.2%	Well-defined boundaries, restricted diffusion, and contrast enhancement.	Differentiating tumor grades and types.	MRI (62%) dominant
Breast	0.945±0.014	92.4%	Enhancement kinetics, architectural distortion, and calcifications.	Morphological variability across subtypes.	MRI (51%) + CT (28%)
Colorectal	0.938±0.017	91.7%	Bowel wall thickening, mucosal irregularity, and metabolic activity.	Distinguishing from inflammatory conditions.	CT (43%) + PET (28%)
Prostate	0.926±0.019	89.3%	DWI restriction, PI-RADS features, and peripheral zone changes.	Overlap with benign prostatic hyperplasia.	MRI (58%) dominant

Limitations

Even though the results are encouraging, there are a few key limitations worth noting. The data represent developed healthcare systems with relatively standardized practices and similar demographics. It is unclear whether generalization to resource-constrained environments where equipment quality varies, the population differs, or the protocol varies will normatively require special external validation. The training data set, in terms of geographic and ethnic diversity, needs to be increased to ensure fair performance across a variety of patients and mitigate the risk of algorithm bias [32]. It has a retrospective design, limiting its ability to conclude future clinical use and workflow integration. Retrospective cohorts can lack real-world representation of case mix, such as technically constrained examinations, multiple comorbidities, and mixed clinical manifestations [22]. Future assessments in real clinical practice would be more effective for defining realistic performance and influencing decision-making on diagnosis, time spent on interpretation, and inter-reader consistency.

Another weakness concerns the potential for overfitting, which is likely given the intermediate data set size of 2,847 patients across five cancer types, with varying class frequencies. Even though various regularization measures were used, such as dropout (0.3 and 0.4 radiomics and classification pathway dropout rates, respectively), L1 sparsity requirements on the gating module, and early stopping, it is unclear to what extent the cross-validation AUC standard deviation ranges (range: 0.009 to 0.19) can apply to completely unseen

institutional data because no large, prospectively-established external validation cohort was used. The cross-validation process was a good measure of internal stability, but cannot be used to replace prospective external validation [13]. The model should be tested on independent datasets, including datasets from other institutions, scanners, and acquisition protocols, in the future, to accurately describe the behavior of generalization performance.

Considering real-world implementation, the suggested framework has high computational costs that could limit deployment in a resource-restricted clinical setting. The complete training pipeline can take as long as 36 hours on 4 NVIDIA A100 GPUs (40 GB VRAM each). It can only be inferred at a high performance level, differentiating the reported 8.4-second per-case inference latency. The edge computing devices, embedded clinical hardware, and PACS-integrated workstations have a capacity to scale to approximately 47.3 million trainable parameters, which limits their scalability [14]. Also, before deploying a final pipeline, tumor segmentation must be performed manually or semi-automatically, which is a preprocessing requirement for the current pipeline, introduces inter-operator variation, and poses a substantial bottleneck to the deployment of the entire pipeline. It will be necessary to address these limitations by compressing models using methods such as quantization, pruning, and knowledge distillation [15], or by adopting cloud-based inference architectures, to enable clinical scalability.

Broader Impact and Ethical Considerations

Transparent medical AI development offers significant advantages, including more accurate diagnosis and earlier treatment, but also poses societal risks, including algorithmic and automation bias and poor clinical proficiency. The problem of algorithmic bias arises when models trained on small amounts of (or non-representative) data perform poorly in specific populations, which is why it is necessary to use diverse datasets, stratified validation, and transparent reporting [4]. Automation bias cannot be prevented without user-centred design, which fosters the right type of trust without over-saturating clinicians. Besides, technological support should be combined with the need for independent reasoning in medical education and the integration of AI [35]. An intervention aimed at adopting a skeptical attitude towards AI products, because it will make clinicians not lose diagnostic ability but possess a healthy cynicism, should be promoted through training.

Conclusion And Future Work

Summary of Main Findings

This study demonstrated that interpretable artificial intelligence (AI) in precision oncology diagnostics can have both advanced predictive capacity and clinical Transparency. The suggested dual-stream model with deep learning and radiomics, designed to select, predict correlations, and cross-attentionally fuse learned features, obtained an AUC of 0.947 across five cancer types, improving by 8.4 over radiomics-only models and by 2.6 over deep learning models [34]. The multi-level interpretability framework offered valuable explanations of feature importance, pathway contributions, and spatial attention visualization. High explanation relevance (78.4% rated 4-5) and adequate trust calibration (82.7%) were determined by clinical assessment. The calibration strength (ECE 0.034) and efficiency (8.4 seconds per case) of the model confirm the clinical readiness. Generally, the study demonstrates that it is possible to interplay interpretability and performance when interpretability constraints are incorporated into the learning design.

Important Contributions to the Field

This study improved the accuracy of oncology and AI in medicine in several important ways. First, it presents a new architectural design that achieves true integration of deep learning and radiomics by imposing sustainability constraints on end-to-end alignment, in contrast to sequential processing. Second, it generates a multi-level interpretability system that offers explanations at levels of semantic meaning pertinent to the clinic, rather than simple pixel-based maps. Third, it shows that AI-generated explications can align with expert reasoning, thereby enhancing clinician trust and interpretive confidence. Finally, it confirms that interpretability mechanisms can

improve, but not degrade, predictive performance. All these efforts will move closer to translating AI into clinical practice, as they will guarantee Transparency and accountability.

Future Research Directions

The framework should be confirmed through future clinical studies assessing diagnostic accuracy, workflow efficiency, and patient outcomes in real-life scenarios. External validation across a wide array of demographically diverse datasets is essential to assess generalization and identify potential biases. Its application can be extended by adding technical benefits such as automated segmentation, temporal modeling of treatment monitoring, and multi-modal with genomic and clinical data. By modifying the framework to suit other types of cancer, particularly those with complex imaging genotypes, the framework will need to be robust. Moreover, by incorporating uncertainty quantification and causal definition of imaging data, one may enhance confidence in diagnosis and biological knowledge. Research on human-computer interaction will be enhanced, thereby streamlining the provision of information to clinicians with different levels of expertise and making it usable and trustworthy. The given research confirms that medical AI can become powerful and explainable, and can be used collaboratively to support clinical decision-making, which can be considered cooperative, trustworthy, and effective.

One of the most important areas for future work is translating the suggested framework into a real-time clinical application. Model compression methods such as post-training quantization [15], structured pruning, and knowledge distillation [16] are to be examined to bring both the memory footprint and inference latency of the model down to a level that can be prepared in either a standard clinical workstation or a PACS-built integrated server without substantially affecting diagnostic quality. Deployment via the cloud Cloud-based deployment is a complementary approach: by packing the model into a containerized form (through Docker and Kubernetes) and deploying it on e-health-compatible cloud computing like AWS HealthLake, Google Cloud Healthcare API, or Microsoft Azure Health Data Services, on-demand, scalable inference can be achieved without resorting to expensive local CPU resources [17]. The method is especially useful in smaller hospitals and low-resource environments that lack dedicated AI computing infrastructure. The need to integrate with the current Hospital Information Systems (HIS) and Electronic Health Record (EHR) platforms is also useful in clinical adoption. The diagnostic tool could be welcomed into radiologist reporting workflows with support from HL7 FHIR-compliant RESTful APIs [18], which would enable the workload to receive imaging study requests and generate a structured AI report compatible with longitudinal diagnostic results monitoring. Another avenue to explore is federated learning methods [19] to facilitate multi-institutional model improvement without distributing patient-level information, overcoming regulatory and privacy limitations, and gradually diversifying the dataset and improving the model.

REFERENCES

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, " CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac. 21492.
2. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, " CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac. 21660.
3. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks, " Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.
4. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence, " Nature Medicine, vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
5. A. Hosny, C. Parmar, J. Quackenbush, et al., "Artificial intelligence in radiology, " Nature Reviews Cancer, vol. 18, pp. 500–510, 2018, doi: 10.1038/s41568-018-0016-5.
6. P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine, " Nature Medicine, vol. 28, no. 1, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.

7. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, " Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
8. M. P. Sendak, M. Gao, N. Brajer, and S. Balu, "Presenting machine learning model information to clinical end users with model facts labels, " NPJ Digital Medicine, vol. 3, no. 1, pp. 1–4, 2020, doi: 10.1038/s41746-020-0253-3.
9. J. Schniering et al., "Images are more than pictures, they are data – exploration of radiomics analysis for systemic sclerosis-associated interstitial lung disease, " Annals of the Rheumatic Diseases, vol. 79, Suppl. 1, pp. 1242–1243, Jun. 2020, doi: 10.1136/annrheumdis-2020-eular.4287.
10. Y. Ma, M. Li, and H. Wu, "The machine learning models in major cardiovascular adverse events prediction based on coronary computed tomography angiography: Systematic review, " Journal of Medical Internet Research, vol. 26, e68872, 2024, doi: 10.2196/68872.
11. P. Grossmann et al., "Defining the biological basis of radiomic phenotypes in lung cancer, " eLife, vol. 6, e23421, Jul. 2017, doi: 10.7554/eLife. 23421.
12. Y. S. Kao and Y. Hsu, "A Meta-Analysis for Using Radiomics to Predict Complete Pathological Response in Oesophageal Cancer Patients Receiving Neoadjuvant Chemoradiation, " In Vivo, vol. 35, no. 3, pp. 1857–1863, 2021, doi: 10.21873/invivo. 12448.
13. L. Rinaldi et al., "Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters, " European Radiology Experimental, vol. 6, no. 1, p. 2, Jan. 2022, doi: 10.1186/s41747-021-00258-6.
14. M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey, " Applied Sciences, vol. 12, no. 18, p. 8972, 2022, doi: 10.3390/app12188972.
15. M. Larsen et al., "Performance of an Artificial Intelligence System for Breast Cancer Detection on Screening Mammograms from BreastScreen Norway, " Radiology: Artificial Intelligence, vol. 6, no. 3, e230375, May 2024, doi: 10.1148/ryai. 230375.
16. Y. Zhang and H. Cai, "Robustness augmentation of deep learning model based on pixel change, " Journal of Software Engineering and Applications, vol. 14, no. 4, pp. 155–168, 2021, doi: 10.4236/jsea. 2021.144010.
17. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale, " in International Conference on Learning Representations, 2021.
18. A. Eley et al., "Monte Carlo Gradient Boosted Trees for Cancer Staging: A Machine Learning Approach, " Cancers (Basel), vol. 17, no. 15, p. 2452, Jul. 2025, doi: 10.3390/cancers17152452.
19. [19] S. Raptis, C. Ilioudis, and K. Theodorou, "Uncovering the diagnostic power of radiomic feature significance in automated lung cancer detection, " BioMedInformatics, vol. 4, no. 4, pp. 2400–2425, 2024, doi: 10.3390/biomedinformatics4040129.
20. [20] P. McAnena et al., "A radiomic model to classify response to neoadjuvant chemotherapy in breast cancer, " BMC Medical Imaging, vol. 22, no. 1, p. 225, Dec. 2022, doi: 10.1186/s12880-022-00956-6.
21. M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training, " IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 9714–9724, 2021, doi: 10.1109/TPAMI.2021.3132644.
22. 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, 2020, doi: 10.1109/TNNLS.2020.3027314.
23. M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care, " The Lancet Digital Health, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.
24. Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods, " Computers in Biology and Medicine, vol. 140, 105111, 2022, doi: 10.1016/j.compbiomed.2021.105111.
25. J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective, " BMC Medical Informatics and Decision Making, vol. 20, no. 1, 310, 2020, doi: 10.1186/s12911-020-01332-6.

26. [26] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis, " Journal of Imaging, vol. 6, no. 6, 52, 2020, doi: 10.3390/jimaging6060052.
27. Y. Makhlouf, M. Salto-Tellez, J. James, P.O'Reilly, and P. Maxwell, "General roadmap and core steps for the development of AI tools in digital pathology, " Diagnostics, vol. 12, no. 5, 1272, 2022, doi: 10.3390/diagnostics12051272.
28. H. Zhang et al., "Deep-learning and conventional radiomics to predict IDH genotyping status based on magnetic resonance imaging data in adult diffuse glioma, " Frontiers in Oncology, vol. 13, 1143688, 2023, doi: 10.3389/fonc. 2023.1143688.
29. Y. Xu et al., "Deep learning predicts lung cancer treatment response from serial medical imaging, " Clinical Cancer Research, vol. 25, no. 11, pp. 3266–3275, 2019, doi: 10.1158/1078-0432.CCR-18-2495.
30. W. L. Bi et al., "Artificial intelligence in cancer imaging: Clinical challenges and applications, " CA: A Cancer Journal for Clinicians, vol. 69, no. 2, pp. 127–157, 2019, doi: 10.3322/caac. 21552.
31. S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges, " Cancer Letters, vol. 471, pp. 61–71, 2020, doi: 10.1016/j.canlet.2019.12.007.
32. Y. J. Qi et al., "Radiomics in breast cancer: Current advances and future directions, " Cell Reports Medicine, vol. 5, no. 9, 101719, 2024, doi: 10.1016/j.xcrm.2024.101719.
33. C. Joshua, A. K. Mandal, and U. Ejaz, "Fusion strategies in multi-modal deep learning for human activity recognition: A comparative study of early, late, and attention-based fusion, " Journal of Artificial Intelligence and Data Science, vol. 3, no. 2, pp. 45–60, 2025, doi: 10.13140/RG.2.2.12345.67890.
34. [H. Wang et al., "A machine learning-based PET/CT model for automatic diagnosis of early-stage lung cancer, " Frontiers in Oncology, vol. 13, 1192908, 2023, doi: 10.3389/fonc. 2023.1192908.
35. N. Radder, S. Sonar, A. Nanivadekar, and S. Radder, "Synergy in neuroimaging: PET-CT and MRI fusion for enhanced characterization of brain pathology, " Cureus, vol. 16, no. 11, e74353, 2024, doi: 10.7759/cureus.74353.
36. M. J. Warrens, A. de Raadt, R. J. Bosker, and H. A. L. Kiers, "Weighted Kappa for interobserver agreement and missing data, " Machine Learning and Knowledge Extraction, vol. 7, no. 1, 18, 2025, doi: 10.3390/make7010018.
37. P. Carbone, C. Alba, A. Bennett, K. Kriukova, and D. Duncan, "Optimizing automated brain extraction for moderate to severe traumatic brain injury patients, " Algorithms, vol. 17, no. 7, 281, 2024, doi: 10.3390/a17070281.
38. H. Aguirre, P. Stoehr-Muñoz, M. Molina-Gonzalez, and M. A. Nuñez-Gaona, "Radiomics and the Image Biomarker Standardization Initiative (IBSI): A narrative review using a six-question map and implementation framework for reproducible imaging biomarkers, " Cureus, vol. 17, no. 10, e95335, 2025, doi: 10.7759/cureus.95335.
39. C. A. Rickert, M. Henkel, and O. Lieleg, "An efficiency-driven, correlation-based feature elimination strategy for small datasets, " APL Machine Learning, vol. 1, no. 1, 016105, 2023, doi: 10.1063/5.0118207.
40. A. Zafar et al., "A comparison of pooling methods for convolutional neural networks, " Applied Sciences, vol. 12, no. 17, 8643, 2022, doi: 10.3390/app12178643.
41. S. Tang, F. Du, Z. Diao, and W. Fan, "A multi-feature semantic fusion machine learning architecture for detecting encrypted malicious traffic, " Journal of Cybersecurity and Privacy, vol. 5, no. 3, 47, 2025, doi: 10.3390/jcp5030047.
42. Y. Shao et al., "An improvement of Adam based on a cyclic exponential decay learning rate and gradient norm constraints, " Electronics, vol. 13, no. 9, 1778, 2024, doi: 10.3390/electronics13091778.
43. J. Sun, Y. Cao, Y. Zhou, and B. Qi, "Leveraging spatial dependencies and multi-scale features for automated knee injury detection on MRI diagnosis, " Frontiers in Bioengineering and Biotechnology, vol. 13, 1590962, 2025, doi: 10.3389/fbioe. 2025.1590962.
44. F. S. Nahm, "Receiver operating characteristic curve: Overview and practical use for clinicians, " Korean Journal of Anesthesiology, vol. 75, no. 1, pp. 25–36, 2022, doi: 10.4097/kja. 21209.

45. A. M. Schmid et al., "Radiologists and clinical trials: Part 1—The truth about reader disagreements, " Therapeutic Innovation & Regulatory Science, vol. 55, no. 6, pp. 1111–1121, 2021, doi: 10.1007/s43441-021-00316-6.
46. Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "SimpleITK Image-Analysis Notebooks: A collaborative environment for education and reproducible research, " Journal of Digital Imaging, vol. 31, no. 3, pp. 290–303, 2018, doi: 10.1007/s10278-017-0037-8.