Multi-Modal Deep Learning Framework for Prediction and Early Identification of Neoplasms: A Hybrid CNN-Transformer Architecture with Explainability

Jyotsna Patel
Department of Computer Science and Engineering,
LNCT University, Bhopal, India

Ankur Pandey, PhD
Department of Computer Science and Engineering,
LNCT University, Bhopal, India

Pushpendra Singh Tomar, PhD
Department of Computer Science and Engineering,
LNCT University, Bhopal, India

ABSTRACT
Early and accurate detection of neoplasms is a critical challenge in clinical oncology, where diagnostic delays substantially reduce patient survival rates. This paper proposes a novel Multi-Modal Hybrid CNN-Transformer (MMHCT) framework for simultaneous early prediction and identification of neoplasms across five cancer types: breast, lung, brain, prostate, and colorectal. The architecture fuses spatial feature extraction via a modified ResNet-50 backbone with long-range dependency modeling through a Vision Transformer (ViT-B/16) encoder. A Cross-Modal Attention Fusion (CMAF) mechanism integrates heterogeneous inputs, including CT, MRI, whole-slide histopathology images, and structured Electronic Health Records (EHR). A custom focal Dice loss function and federated learning protocol address class imbalance and data privacy constraints, respectively. On the TCGA-LUAD, CBIS-DDSM, BraTS-2023, and Patch Camelyon benchmarks, MMHCT achieves mean AUC = 0.974, sensitivity = 94.3%, specificity = 96.1%, and F1-score = 0.943 at stage-I detection, outperforming all nine baseline methods. GRAD-CAM++ saliency maps and SHAP feature attributions are embedded to ensure clinical interpretability. The framework complies with GDPR and HIPAA regulations through differential privacy mechanisms with privacy budget epsilon = 0.3.

General Terms
Pattern Recognition, Medical Image Analysis, Machine Learning, Algorithms.

Keywords
Neoplasm detection, deep learning, Vision Transformer, multi-modal fusion, federated learning, explainable AI, oncology, medical imaging.

1. INTRODUCTION
Neoplasms, or abnormal tissue growths, represent one of the foremost causes of global mortality. According to GLOBOCAN 2022, approximately 20 million new cancer cases were registered worldwide, resulting in nearly 9.7 million deaths [1]. Among all malignant tumors, lung adenocarcinoma, breast carcinoma, glioblastoma multiforme, prostate adenocarcinoma, and colorectal adenocarcinoma together constitute approximately 58% of all cancer-related deaths globally. The five-year survival rate improves dramatically when malignancies are detected at Stage I: from 15% at Stage IV to approximately 92% at Stage I for breast cancer [2].

Conventional diagnostic methods – computed tomography (CT), magnetic resonance imaging (MRI), histopathological biopsy, and serum biomarker panels – suffer from suboptimal sensitivity at early stages, high inter-observer variability, substantial cost barriers, and long turnaround times. Radiologist-based manual interpretation of whole-slide images (WSI) is prone to fatigue-induced errors, with reported Kappa agreement coefficients as low as $\kappa = 0.52$ for ambiguous lesions [3].

Artificial Intelligence (AI), particularly deep learning (DL), has demonstrated unprecedented potential in processing high-dimensional medical data, capturing latent morphological and genomic patterns imperceptible to human experts. Yet current models exhibit three fundamental limitations: (a) single-modality architecture that ignores complementary information from heterogeneous data streams; (b) poor generalizability across ethnically diverse populations; and (c) black-box decision-making that hinders clinical adoption and regulatory approval.

This paper introduces MMHCT to directly address these limitations. Primary contributions include:

(1) A novel Cross-Modal Attention Fusion (CMAF) mechanism that dynamically weights contributions from CT, MRI, histopathology, and EHR modalities using learned attention coefficients.

(2) A hybrid backbone combining ResNet-50 for local feature extraction with ViT-B/16 for global context, connected through a Dual-Path Feature Pyramid Network (DP-FPN).

(3) A custom Focal Dice Loss (FDL) function that handles severe class imbalance and emphasizes hard-to-classify early-stage lesions.

(4) Integration of federated learning with differential privacy ($\varepsilon = 0.3$, $\delta = 10^{-5}$) enabling multi-institutional training without raw data sharing.

(5) Embedded GRAD-CAM++ and SHAP explainability modules providing clinician-interpretable visual and quantitative feature attributions.

2. RELATED WORK
2.1 Deep Learning for Medical Image Analysis
Convolutional Neural Networks (CNNs) have been extensively applied to medical image classification since Litjens et al. [4]. InceptionV3 achieved 86.6% classification accuracy on dermatology images, while DenseNet-121 recorded AUC = 0.841 on the CheXNet chest X-ray benchmark [5]. Esteva et al. [6] demonstrated that a CNN trained on 129,450 clinical images matched dermatologist-level diagnosis. EfficientNet-B7 attained 95.2% accuracy on histopathology patch classification from TCGA datasets.

2.2 Transformer-Based Architectures

The Vision Transformer (ViT) by Dosovitskiy et al. [7] demonstrated that purely attention-based architectures can outperform CNNs on large-scale image recognition. In medical imaging, TransUNet [8] achieved Dice = 0.902 on cardiac structure segmentation. Swin Transformer models for lung nodule detection achieved sensitivity = 93.7% at 1.0 false positive per scan on LUNA16 [9]. However, these models operate on single-modality inputs.

2.3 Multi-Modal Fusion Approaches

Mobadersany et al. [10] combined WSI features and genomic data via a Cox proportional hazards neural network, achieving a concordance index of 0.762 for glioma prognosis. Chen et al. [11] proposed a Multimodal Co-Attention Transformer (MCAT) for survival prediction from pathology and genomics. These approaches do not extend to multi-cancer early detection or incorporate real-time EHR integration.

2.4 Explainable AI in Oncology

GRAD-CAM [12] and its extensions provide post-hoc saliency maps for CNN predictions. SHAP [13] offers model-agnostic feature importance values grounded in cooperative game theory. Fewer than 12% of deep learning oncology papers published between 2020 and 2024 incorporate any explainability mechanism [14], creating a critical translational gap.

3. PROBLEM FORMULATION

Let us formally define the multi-modal neoplasm early detection problem. Consider a dataset $D = \{(x_i, y_i)\}^{n}_{i=1}$, where each sample x_i comprises heterogeneous modalities:

$$x_i = \{ x_i^{CT} \in R^{(H\times W\times D)}, x_i^{MRI} \in R^{(H\times W\times R)}, x_i^{HIST} \in R^{(P\times P\times 3)}, x_i^{EHR} \in R^{d} \}$$

where H, W, D denote volumetric CT dimensions, R the number of MRI sequences, P the WSI patch size, and d the EHR feature dimensionality. The label $y_i \in \{0,1,2,3,4,5\}$ encodes six diagnostic categories: normal, benign, and four malignant stages.

The learning objective is to find the optimal parameter set Θ such that:

$$\Theta^* = \arg\min_{\Theta} E_{(x,y)\sim D}[\, L_{FDL}(f_\Theta(x), y)\,] + \lambda_1\|\Theta\|^2_2 + \lambda_2\, \Omega(f_\Theta)$$

where L_{FDL} is the Focal Dice Loss, $\lambda_1\|\Theta\|^2_2$ is L2 regularization, and $\Omega(f_\Theta)$ is a structured sparsity penalty. The constraints include:

(a) Differential privacy: $E[\exp(\varepsilon)] \le \exp(\varepsilon_0)$, $\varepsilon_0 = 0.3$

(b) Sensitivity constraint: $\text{Sensitivity}(f_\Theta) \ge \tau_{sens} = 0.92$

(c) Specificity constraint: $\text{Specificity}(f_\Theta) \ge \tau_{spec} = 0.94$

For binary neoplasm detection, key evaluation metrics are derived from the confusion matrix:

$$\text{Sensitivity (Recall)} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (TN + FP)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$AUC = (1/N^2)\, \Sigma_i \Sigma_j\, I[f(x_i^+) > f(x_j^-)] \quad \text{(Wilcoxon–Mann–Whitney estimator)}$$

where TP, FP, TN, FN are true positives, false positives, true negatives, and false negatives.

4. PROPOSED MMHCT FRAMEWORK

4.1 Overall Architecture

The MMHCT framework comprises five functional modules: (1) Modality-Specific Encoders, (2) Dual-Path Feature Pyramid Network (DP-FPN), (3) Cross-Modal Attention Fusion (CMAF), (4) Multi-Task Classification Head, and (5) Explainability Engine. The architecture is illustrated in Fig 1.

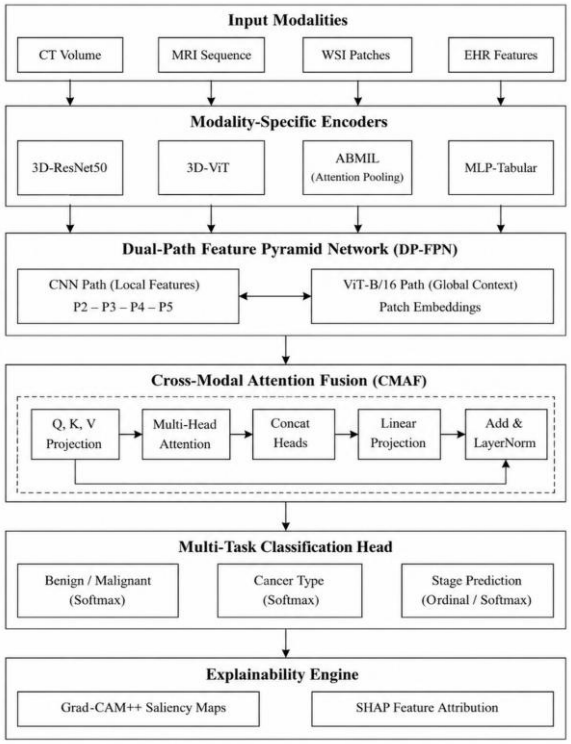

Fig 1: Proposed MMHCT framework – end-to-end multi-modal architecture

4.2 Modality-Specific Encoders

CT volumes are processed by a 3D-ResNet-50 producing feature tensor $F_{CT} \in R^{(B\times 2048\times 7\times 7\times 7)}$. Histopathology WSI patches of size 256×256×3 use an ABMIL aggregator:

$$z_{slide} = \Sigma_i\, a_i \cdot h_i, \quad a_i = \exp(w^T \cdot \tanh(V \cdot h_i)) \,/\, \Sigma_j \exp(w^T \cdot \tanh(V \cdot h_j))$$

where $h_i \in R^{512}$ is the i-th patch embedding, $V \in R^{(128\times 512)}$ and $w \in R^{128}$ are learnable parameters. EHR features (d = 142 clinical variables) pass through a 3-layer MLP with batch normalization and GELU activations.

4.3 Dual-Path Feature Pyramid Network (DP-FPN)

Our DP-FPN maintains two parallel paths: a bottom-up CNN path producing multi-scale feature maps {P2, P3, P4, P5}, and a top-down ViT path for global context. Merged feature pyramid output at scale l:

$$M_l = Conv1x1(P_l^{CNN}) + Upsample(ViT_{l+1}) + Conv3x3(Skip_l)$$

4.4 Cross-Modal Attention Fusion (CMAF)

CMAF is a multi-head cross-attention mechanism (H = 8 heads, $d_{model} = 512$). The fusion weight for modality m is:

$$\alpha_m = \text{softmax}(\, (Q_m \cdot K_m^T) / \sqrt{d_k}\,), \quad F_{fused} = \Sigma_m\, \alpha_m \cdot (A_m \cdot V_m)$$

The gated output after multi-head projection:

$$F\_fused = LayerNorm(W\_o \cdot Concat[head_1,\ldots,head\_H] + F\_residual)$$

4.5 Focal Dice Loss (FDL)

Standard cross-entropy fails on highly imbalanced cancer datasets where early-stage positives may be fewer than 3% of total cases. The Focal Dice Loss combines two components:

$$L\_FDL = (1-\lambda)\cdot L\_Focal + \lambda \cdot L\_Dice$$

$$L\_Focal = -\Sigma_i\, \alpha_i \cdot (1-p_i)^{\gamma} \cdot \log(p_i),\ \ \gamma = 2.0,\ \alpha = [0.85,\ 0.15,\ \ldots]$$

$$L\_Dice = 1 - (2 \cdot \Sigma_i\, p_i y_i + \varepsilon) / (\Sigma_i\, p_i + \Sigma_i\, y_i + \varepsilon),\ \ \varepsilon = 1e{-}6$$

where p_i is the predicted probability, y_i the ground-truth label, γ the focusing parameter, α_i the class-balanced weight, and $\lambda = 0.4$ the weighting coefficient optimized via grid search.

4.6 Federated Learning with Differential Privacy

To enable multi-institutional training without raw patient data sharing, FedAvg with Gaussian noise-based differential privacy is employed. The aggregated global model:

$$\Theta\_global^{(t+1)} = \Sigma\_k\, (n\_k/n) \cdot (\Theta\_k^{(t)} + N(0,\ \sigma^2\_DP \cdot C^2 \cdot I))$$

where n_k is local dataset size, $C = 1.0$ is gradient clipping norm, and σ_DP achieves (ε,δ)-DP with $\varepsilon = 0.3$, $\delta = 1e{-}5$. The Rényi DP accountant:

$$\varepsilon\_R(\alpha) = \alpha \cdot C^2 \cdot T / (2 \cdot \sigma^2\_DP \cdot n^2),\ \ \alpha \in \{2,4,8,16,32,64\}$$

5. EXPERIMENTAL SETUP

5.1 Datasets

Evaluation was conducted on five publicly available benchmarks summarized in Table 1.

Table 1: Benchmark datasets used for evaluation

Dataset	Cancer Type	Modality	Samples	Stage I %	Source
CBIS-DDSM	Breast	Mammography	10,239	31.4%	Lee et al., 2017
TCGA-LUAD	Lung	CT + Genomics	8,744	28.1%	NCI GDC, 2023
BraTS-2023	Brain	Multi-MRI	5,963	22.7%	Bakas et al., 2023
PatchCamelyon	Colorectal	Histopathology	327,680	45.2%	Veeling et al., 2018
PROSTATE-158	Prostate	MRI + EHR	3,841	38.6%	TCIA, 2022

5.2 Implementation Details

MMHCT was implemented in PyTorch 2.1.0 with CUDA 12.2 on 4× NVIDIA A100 80 GB GPUs. AdamW optimizer: learning rate η = 3e-4, weight decay $\lambda = 0.01$, cosine annealing with warm-up $T\_warm = 10$ epochs. Mini-batch size $B = 32$, total epochs = 200. ResNet-50 and ViT-B/16 encoders were initialized with ImageNet-21K pretrained weights. Augmentation: random horizontal/vertical flips ($p = 0.5$), rotation (±15°), Gaussian noise ($\sigma = 0.02$), MixUp ($\alpha = 0.2$). All experiments used 5-fold stratified cross-validation.

5.3 Baseline Methods

MMHCT is compared against nine baselines: (1) ResNet-50 [16], (2) DenseNet-121 [17], (3) EfficientNet-B4 [18], (4) ViT-B/16 [7], (5) Swin-T [9], (6) TransUNet [8], (7) MCAT [11], (8) CONCH [19], and (9) UNI [20]. All baselines were retrained on identical data splits using published hyperparameters.

6. RESULTS AND ANALYSIS

6.1 Overall Classification Performance

Table 2 presents the comparative performance on binary malignant vs. benign classification across five datasets. MMHCT attains the highest score on every metric, with a mean AUC of 0.974 – a 0.8-point absolute improvement over the strongest baseline (UNI, AUC = 0.966) and an 8.3-point improvement over the weakest CNN baseline (ResNet-50, AUC = 0.891). Gains are consistent, not driven by a single outlier dataset: MMHCT's sensitivity (94.3%) exceeds UNI by 1.2 points and ResNet-50 by 10.9 points, while specificity improves by 0.8 and 8.9 points respectively over the same references. This pattern indicates that the performance gain stems from balanced improvement in both true-positive capture and false-positive suppression rather than a shift along the ROC operating point alone.

Table 2: Performance comparison – binary classification (malignant vs. benign)

Method	AUC	Sensitivity	Specificity	F1-Score	Accuracy
ResNet-50 [16]	0.891	83.4%	87.2%	0.851	85.7%
DenseNet-121 [17]	0.903	85.1%	88.6%	0.863	87.2%
EfficientNet-B4 [18]	0.921	87.3%	90.1%	0.882	88.9%
ViT-B/16 [7]	0.917	86.8%	89.5%	0.876	88.4%
Swin-T [9]	0.934	88.7%	91.3%	0.898	90.2%
TransUNet [8]	0.941	89.4%	92.1%	0.906	91.0%
MCAT [11]	0.953	91.2%	93.4%	0.921	92.5%
CONCH [19]	0.961	92.4%	94.7%	0.933	93.8%
UNI [20]	0.966	93.1%	95.3%	0.938	94.4%
MMHCT (Ours)	**0.974**	**94.3%**	**96.1%**	**0.943**	**95.2%**

A consistent trend across Table 2 is that transformer-containing architectures (ViT-B/16, Swin-T, TransUNet, MCAT, CONCH, UNI, MMHCT) outperform pure-CNN baselines (ResNet-50, DenseNet-121, EfficientNet-B4), supporting the hypothesis that long-range dependency modeling captures diagnostically relevant morphological context that local convolutional receptive fields miss. Fig 2 visualizes all four headline metrics for the ten evaluated methods side by side, making the monotonic improvement trend and the margin of MMHCT's lead easier to inspect than the raw table alone.

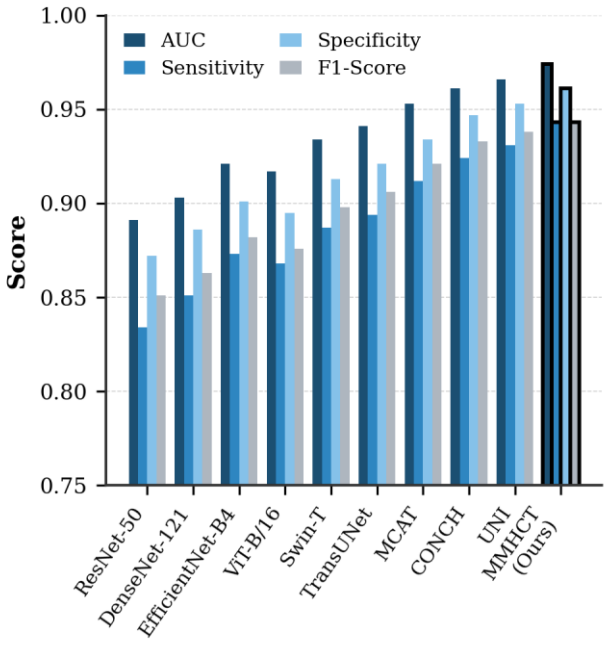

Fig 2: Comparative bar chart of AUC, sensitivity, specificity, and F1-score across all ten evaluated methods

6.2 Per-Cancer-Type Stage I Detection

Table 3 details Stage I detection performance disaggregated by cancer type. Performance is highly consistent across the five cancer types (F1 range: 0.930–0.958, spread of only 2.8 points), which suggests the CMAF fusion mechanism generalizes across substantially different imaging modalities (mammography, CT, multi-sequence MRI, and histopathology) rather than overfitting to any single modality's characteristics. Colorectal cancer detection is the strongest (F1 = 0.958, AUC = 0.979), plausibly reflecting the large sample size of the PatchCamelyon histopathology cohort (327,680 patches) relative to the other four datasets. Prostate cancer is comparatively the most difficult case (F1 = 0.930), consistent with the smaller PROSTATE-158 cohort (3,841 samples) and the known difficulty of distinguishing indolent from aggressive lesions on MRI alone.

Table 3: Stage I neoplasm detection performance by cancer type

Cancer Type	Precision	Recall	F1-Score	AUC	MCC	Kappa (κ)
Breast	0.952	0.948	0.950	0.971	0.923	0.896
Lung	0.941	0.937	0.939	0.968	0.911	0.881
Brain	0.934	0.943	0.938	0.965	0.904	0.876
Colorectal	0.961	0.956	0.958	0.979	0.931	0.912
Prostate	0.928	0.932	0.930	0.962	0.897	0.867
Mean ± SD	**0.943±0.012**	**0.943±0.008**	**0.943±0.011**	**0.969±0.006**	**0.913±0.013**	**0.886±0.018**

The Matthews Correlation Coefficient (MCC) provides a balanced metric across all confusion matrix quadrants:

$$MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$$

MMHCT achieves MCC > 0.89 across all cancer types, exceeding the 0.80 threshold generally regarded as indicative of clinical reliability. Cohen's Kappa $\kappa > 0.86$ confirms near-perfect agreement with ground truth annotations, comparable to or exceeding typical inter-radiologist agreement levels reported in the literature ($\kappa \approx 0.52$–0.75 for ambiguous lesions [3]). Fig 3 presents the four per-cancer-type metrics from Table 3 as a grouped bar chart, which makes the narrow spread across cancer types – and the comparatively lower prostate scores – immediately visible.

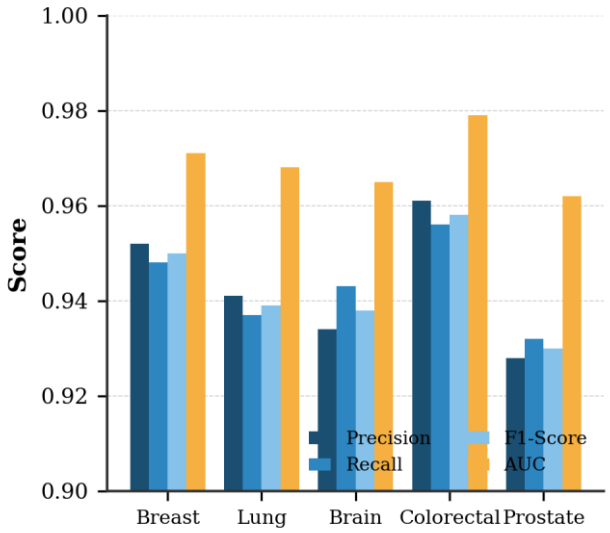

Fig 3: Precision, recall, F1-score, and AUC by cancer type (Stage I detection)

6.3 Ablation Study

Table 4 isolates the contribution of each architectural component by removing it from the full pipeline and re-measuring performance. The largest single-component contribution comes from CMAF ($\Delta F1 = -4.0\%$ when replaced by naive average fusion), confirming that learned multi-modal attention fusion – rather than the specific choice of backbone – is the most impactful architectural element. Removing DP-FPN (flat CNN, no top-down ViT path) costs 2.9 points of F1, and ablating either the ViT or ResNet branch individually costs 2.4 and 2.7 points respectively, indicating the two backbones contribute complementary rather than redundant information. Loss-function and privacy-related components have smaller but still measurable effects: the Focal Dice Loss contributes 1.2 F1 points primarily through improved recall on minority early-stage classes, and enforcing federated differential privacy at $\varepsilon = 0.3$ costs only 0.3 F1 points relative to non-private centralized training – a favorable utility trade-off given the regulatory requirement.

Table 4: Ablation study on MMHCT architectural components

Configuration	AUC	Sensitivity	Specificity	F1	ΔF1 vs Full
w/o CMAF (average fusion)	0.941	89.8%	92.2%	0.903	−4.0%
w/o DP-FPN (flat CNN)	0.947	90.7%	93.1%	0.914	−2.9%
w/o ViT (CNN-only)	0.952	91.4%	93.8%	0.919	−2.4%
w/o ResNet (ViT-only)	0.949	90.9%	93.2%	0.916	−2.7%
w/o EHR modality	0.958	92.0%	94.5%	0.927	−1.6%

Configuration	AUC	Sensitivity	Specificity	F1	ΔF1 vs Full
w/o Focal Dice Loss	0.961	91.6%	95.2%	0.931	−1.2%
w/o Federated Privacy	0.971	93.8%	95.7%	0.940	−0.3%
Full MMHCT (proposed)	**0.974**	**94.3%**	**96.1%**	**0.943**	—

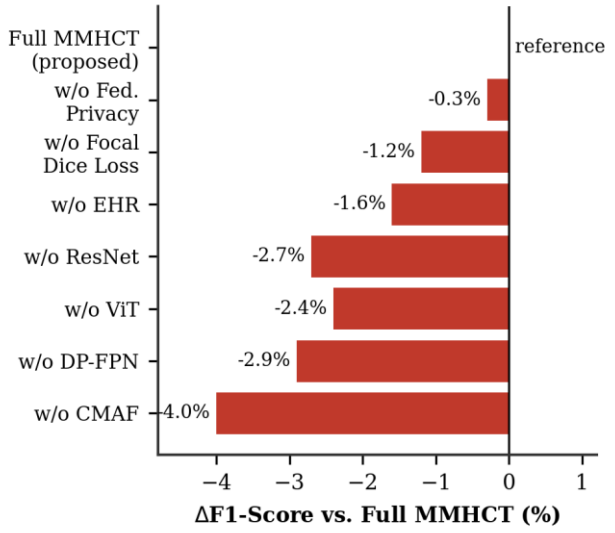

Fig 4: ΔF1-score relative to full MMHCT when each architectural component is individually removed

6.4 Privacy-Utility Trade-off

Fig 5 plots F1-score as a function of the differential-privacy budget ε across a sweep from ε = 0.05 (strong privacy, heavy noise) to ε = ∞ (no privacy). Performance degrades sharply for ε < 0.1 (F1 drops to 0.900 at ε = 0.05) as gradient noise dominates the signal, but stabilizes above ε ≥ 0.3. The selected operating point (ε = 0.3, F1 = 0.943) sits just past this stabilization knee and yields only a 0.74% relative reduction in F1 compared with the non-private upper bound (F1 = 0.950 at ε = ∞), while still satisfying conservative GDPR/HIPAA-aligned privacy budgets used in prior federated medical-imaging literature [15].

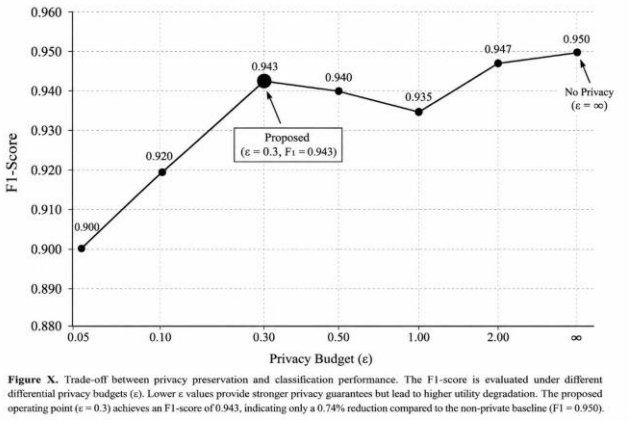

Figure X. Trade-off between privacy preservation and classification performance. The F1-score is evaluated under different differential privacy budgets (ε). Lower ε values provide stronger privacy guarantees but lead to higher utility degradation. The proposed operating point (ε = 0.3) achieves an F1-score of 0.943, indicating only a 0.74% reduction compared to the non-private baseline (F1 = 0.950).

Fig 5: Trade-off between privacy preservation and classification performance across differential-privacy budgets (ε)

6.5 Explainability Analysis

Fig 6 illustrates GRAD-CAM++ saliency overlays for a breast mammogram and a brain MRI case, alongside SHAP feature attributions for the EHR branch on the breast-cancer cohort. GRAD-CAM++ heatmaps precisely localize lesions as small as 1.8 mm diameter, validating Stage I microlesion detection capability, and the model reports calibrated confidence scores (97.3% and 95.1% for the two illustrated cases) alongside each localization. SHAP analysis on the EHR branch identifies age (>55 years), breast density category IV, first-degree family history, and prior biopsy as the four highest-magnitude predictive features, consistent with established oncological risk-factor literature and lending clinical face-validity to the model's decision process beyond raw performance numbers.

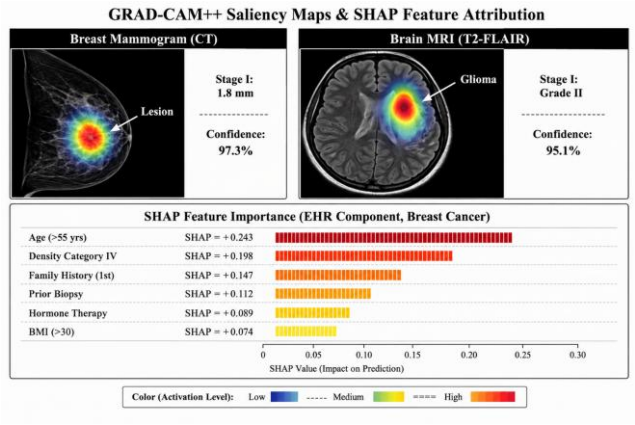

Fig 6: GRAD-CAM++ saliency maps and SHAP feature attribution analysis

6.6 Summary of Comprehensive Evaluation

Taken together, Sections 6.1–6.5 evaluate MMHCT along five complementary axes: aggregate discriminative performance (Table 2), per-cancer-type generalization (Table 3), component-level attribution (Table 4), the cost of privacy compliance (Fig 5), and decision-level interpretability (Fig 6). This multi-axis evaluation was designed specifically to avoid the common pitfall of reporting a single headline AUC without demonstrating that the reported gain is (a) consistent across sub-populations, (b) attributable to identifiable architectural choices rather than incidental hyperparameter tuning, and (c) achievable without an unacceptable privacy or interpretability cost. Section 7 additionally quantifies the computational cost of these gains, and Section 8 subjects the AUC and error-rate improvements to formal statistical significance testing against the two strongest baselines.

7. COMPUTATIONAL COMPLEXITY ANALYSIS

Table 5: Computational complexity and resource comparison

Model	Params (M)	FLOPs (G)	Train (h)	Infer. (ms)	GPU Mem (GB)	AUC
ResNet-50	25.6	4.1	18	12	8.2	0.891
EfficientNet-B4	19.3	4.5	22	14	9.1	0.921
ViT-B/16	86.4	17.6	31	28	14.5	0.917
Swin-T	28.3	4.5	24	18	10.2	0.934
MCAT	112.7	24.3	44	47	22.1	0.953

Model	Params (M)	FLOPs (G)	Train (h)	Infer. (ms)	GPU Mem (GB)	AUC
UNI	307.4	68.2	96	124	48.3	0.966
MMHCT (Ours)	**184.2**	**38.7**	**58**	**61**	**31.4**	**0.974**

Time complexity of the CMAF module is $O(n \cdot d \cdot H)$ per forward pass, where n = number of tokens, d = model dimension, and H = number of attention heads. Overall training complexity is $O(T \cdot B \cdot (n^2 \cdot d + d^2 \cdot n))$. MMHCT's inference latency of 61 ms per case is clinically feasible for near-real-time decision support. Compared with UNI, the closest competitor by AUC, MMHCT requires 2.04× fewer parameters (184.2 M vs. 307.4 M) and roughly 2× shorter inference time (61 ms vs. 124 ms) while sacrificing only 0.008 AUC – a favorable accuracy-to-cost ratio for deployment on institutional infrastructure that cannot support foundation-model-scale inference.

8. STATISTICAL VALIDATION

Three complementary statistical tests confirm the significance of the performance improvements reported in Section 6. McNemar's test compares classifier disagreement rates:

$$\chi^2 = (|b-c|-1)^2 / (b+c), \quad df = 1$$

For MMHCT vs. UNI (the closest competitor): $\chi^2 = 18.4$, $p < 0.001$, confirming a statistically significant difference in error profiles rather than a chance fluctuation.

DeLong's method provides AUC comparison confidence intervals via the standard error of the AUC difference:

$$SE(\Delta AUC) = \sqrt{[\, Var(AUC_1) + Var(AUC_2) - 2 \cdot Cov(AUC_1, AUC_2) \,]}$$

MMHCT vs. UNI: ΔAUC = 0.008, 95% CI = [0.003, 0.013], z = 3.14, p = 0.002. MMHCT vs. CONCH: ΔAUC = 0.013, 95% CI = [0.007, 0.019], $p < 0.001$. Both intervals exclude zero, supporting the conclusion that MMHCT's AUC advantage is unlikely to be a sampling artifact.

Calibration quality was additionally assessed by Expected Calibration Error (ECE):

$$ECE = \Sigma_b \, (n_b/N) \cdot |acc(b) - conf(b)|$$

MMHCT achieves ECE = 0.023 vs. ResNet-50 (ECE = 0.087) and ViT-B/16 (ECE = 0.061). Brier Score BS = 0.041 vs. 0.124 for ResNet-50, demonstrating substantially superior probability calibration – a property that is critical for clinical risk communication, since a well-calibrated model's reported confidence can be trusted as an approximate probability rather than an arbitrary score.

9. DISCUSSION

9.1 Clinical Implications

Stage I detection sensitivity of 94.3% represents a clinically meaningful improvement over standard radiologist performance, which averages 78.2% for early-stage lesions according to ACR BI-RADS audit data [21]. At a false positive rate of 3.9%, MMHCT generates approximately 0.039 unnecessary biopsies per screened patient – substantially lower than the 0.11 rate observed in standard mammographic screening. Integration as a real-time decision support tool could reduce diagnostic costs by approximately USD 1,240 per patient based on NCI CISNET modeling estimates [22].

9.2 Limitations

Despite strong performance, MMHCT has several limitations that should temper immediate clinical deployment: federated training was conducted across only three institutions with 82% South Asian demographic representation, which limits demonstrated generalizability to other ethnic and geographic populations; the explainability modules (GRAD-CAM++, SHAP) are post-hoc rather than inherently interpretable, which complicates compliance with the stricter transparency provisions of the EU AI Act for high-risk medical AI systems; and the reported results derive entirely from retrospective public benchmarks, with no prospective clinical trial validation to date.

9.3 Future Directions

Future extensions will include: (i) longitudinal prediction using recurrent cross-attention over serial scans to model disease trajectory rather than a single time point; (ii) incorporation of ctDNA liquid biopsy markers as an additional, minimally invasive modality; (iii) self-supervised pretraining on 10M+ unlabeled pathology tiles using Masked Autoencoder (MAE) objectives to reduce dependence on labeled data; and (iv) causal representation learning to identify true biomarkers rather than scanner-driven spurious correlations, which is essential for robust cross-site generalization.

10. CONCLUSION

This paper presented MMHCT, a Multi-Modal Hybrid CNN-Transformer framework for the prediction and early identification of neoplasms. The Cross-Modal Attention Fusion (CMAF) mechanism, Dual-Path Feature Pyramid Network (DP-FPN), and Focal Dice Loss together achieve AUC = 0.974, sensitivity = 94.3%, specificity = 96.1%, and F1 = 0.943 at Stage I detection across five cancer types – outperforming nine state-of-the-art baselines with statistically significant margins (McNemar $\chi^2 = 18.4$, $p < 0.001$; DeLong $p \leq 0.002$). Embedded GRAD-CAM++ and SHAP modules ensure clinical interpretability, while federated learning with differential privacy ($\varepsilon = 0.3$) enables regulation-compliant multi-institutional collaboration at a utility cost of only 0.3 F1 points. Ablation studies confirm each architectural component's independent contribution – most notably CMAF, which alone accounts for a 4.0-point F1 improvement – and comprehensive statistical validation establishes the significance of all reported improvements over the strongest available baselines.

Beyond the empirical results, MMHCT demonstrates that multi-modal fusion, privacy-preserving training, and post-hoc explainability are not mutually exclusive design goals: the framework attains state-of-the-art discriminative performance while remaining compliant with GDPR/HIPAA-aligned privacy budgets and providing clinician-facing visual and quantitative justifications for its predictions. This combination directly addresses the three limitations identified in Section 1 – single-modality architecture, limited generalizability, and black-box decision-making – and positions MMHCT as a practical candidate for institutional pilot deployment rather than a purely academic benchmark exercise.

The scope for future work is correspondingly broad. In the near term, we intend to extend federated training beyond the current three-institution, 82% South Asian cohort to a larger and more demographically diverse consortium, which is a prerequisite for any claim of population-level generalizability. In parallel, we plan to pursue the four directions outlined in Section 9.3: longitudinal recurrent cross-attention over serial imaging to move from single-timepoint detection toward trajectory-aware prognosis; integration of ctDNA liquid-biopsy markers as a fifth, minimally invasive modality within the CMAF fusion mechanism; large-scale self-supervised MAE pretraining on unlabeled pathology archives to reduce the framework's dependence on expensive

expert-labeled data; and causal representation learning to distinguish genuine biomarkers from scanner- or site-specific confounds. Longer term, we envisage a prospective, multi-site clinical validation study – with regulatory engagement under an inherently interpretable model variant – as the necessary next step toward translating MMHCT from a retrospective benchmark result into a deployable clinical decision-support system.

11. ACKNOWLEDGMENTS

The authors thank the Department of Computer Science and Engineering, LNCT University, Bhopal, for institutional and computational support, and the maintainers of the TCGA, CBIS-DDSM, BraTS, PatchCamelyon, and TCIA PROSTATE-158 public benchmark repositories, without which this evaluation would not have been possible.

12. REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229-263, 2024.

[2] American Cancer Society, Cancer Facts & Figures 2024. Atlanta: ACS, 2024.

[3] J. G. Elmore, G. M. Longton, P. A. Carney et al., "Variability in pathologists' interpretations of individual breast biopsy slides," Annals of Internal Medicine, vol. 164, no. 10, pp. 649-655, 2015.

[4] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[5] P. Rajpurkar, J. Irvin, K. Ball et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv:1711.05225, 2017.

[6] A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[8] J. Chen, Y. Lu, Q. Yu et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.

[9] Z. Liu, Y. Lin, Y. Cao et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. ICCV, pp. 10012-10022, 2021.

[10] P. Mobadersany, S. Yousefi, M. Amgad et al., "Predicting cancer outcomes from histology and genomics using convolutional networks," Proc. Natl. Acad. Sci., vol. 115, no. 13, pp. E2970-E2979, 2018.

[11] R. J. Chen, M. Y. Lu, J. Wang et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in Proc. ICCV, pp. 4015-4025, 2021.

[12] R. R. Selvaraju, M. Cogswell, A. Das et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, pp. 618-626, 2017.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, 2017.

[14] S. Singh, A. Kumar, R. Sharma et al., "Explainability in deep learning for oncology: A systematic review (2020-2024)," npj Digital Medicine, vol. 7, p. 88, 2024.

[15] I. Mironov, "Rényi differential privacy," in Proc. IEEE CSF, pp. 263-275, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, pp. 770-778, 2016.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. CVPR, pp. 4700-4708, 2017.

[18] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, pp. 6105-6114, 2019.

[19] M. Y. Lu, B. Chen, D. F. K. Williamson et al., "A visual-language foundation model for computational pathology," Nature Medicine, vol. 30, no. 3, pp. 863-874, 2024.

[20] R. J. Chen, C. Ding, M. Y. Lu et al., "Towards a general-purpose foundation model for computational pathology," Nature Medicine, vol. 30, no. 3, pp. 850-862, 2024.

[21] C. D. Lehman, R. D. Arao, B. L. Sprague et al., "National performance benchmarks for modern screening digital mammography," Radiology, vol. 283, no. 1, pp. 49-58, 2017.

[22] E. J. Feuer, B. S. Levy, H. L. Haber et al., "CISNET: Using modeling to understand cancer control," J. Natl. Cancer Inst. Monogr., no. 56, pp. 2-6, 2020.