Supplementary: HyPCA-Net: Advancing Multimodal Fusion in Medical Image Analysis

Joy Dhar[1,*] Manish Kumar Pandey[2,*] Debashis Das Chakladar[3,*] Maryam Haghighat[4]
Azadeh Alavi[5] Sajib Mistry[6] Nayyar Zaidi[7]

[1]Indian Institute of Technology Ropar, India [2]RoentGen Health, India [3]Lulea University of Technology, Sweden
[4]QUT, Australia [5]RMIT University, Australia [6]Curtin University, Australia [7]Deakin University, Australia

1. Extended Related Works

Advances in both unimodal and multimodal fusion have significantly improved medical–image analysis by exploiting either the depth of a single modality or the synergy between multiple modalities. Nevertheless, three key shortcomings remain:

Inadequate shared feature learning. Representative unimodal architectures—such as `DDA-Net` [5], `POTTER` [33], and `NAT` [9]—rely on *computationally expensive attention operations*. Trained in isolation, they cannot leverage *complementary cross-modal contextual information*. As a result, they tend to overfit to noise and *generalize poorly across heterogeneous modalitites*, a limitation that is magnified in resource-constrained environments.

Information loss in cascaded attention. Many fusion pipelines—for example, `DRIFA-Net` [6]—stack attention blocks in a cascade. During successive hand-offs, *fine-grained, modality-specific details are gradually lost*, diluting the quality of the *common representation* required for reliable multimodal fusion.

High computational overhead. While state-of-the-art fusion models can learn effective shared representations, they do so at the price of *substantial memory and runtime costs*, driven by heavy convolutional backbones and dense attention layers. Such resource demands hinder *scalability and practical deployment* in low-resource environments. To bridge these gaps, we introduce `HyPCA-Net`. At its core is the `HyPCA` block, which combines two novel components—`RALA` and `DVCA`—to learn robust shared features that balance *state-of-the-art performance with economical computation*. Table 1 contrasts these attributes with prior baselines, highlighting how `HyPCA-Net` systematically overcomes each of the above limitations.

*These authors contributed equally.

Table 1. Characteristics of unimodal and multimodal fusion methods.

Type	Model	Shared Representation	Cost Effective	Generalization
Unimodal	DDA-Net [5]		✓	
	POTTER [33]		✓	
	NAT [9]			
	MSCAM [23]		✓	✓
	MFMSA [19]		✓	✓
Multimodal	AsymFusion [30]	✓		
	MMTM [14]	✓		
	GLORIA [11]	✓		
	HAMLET [12]	✓		
	MuMu [13]	✓		
	M³Att [17]	✓		
	DRIFA-Net [6]	✓		
	HyPCA-Net (ours)	✓	✓	✓

2. Hierarchical Channel Fusion (`HCF`)

Figure 1 illustrates `HCF`, which models diverse channel relationships by hierarchically fusing sub-band features.

3. Extended Evaluation Setup

Datasets: We evaluate `HyPCA-Net` on ten public medical-imaging benchmarks: for classification (`D1–D8`) we use `Nickparvar` [20], `IQ-OTH NCCD` [1], `Tuberculosis` [24], `CNMC-2019` [18], `HAM10000` [29], `SIPaKMeD` [21], `CRC` [16], `CBIS-DDSM` [26]; for segmentation (`D9–D10`) we use `COVID-19` lung `CT` [15] and `ISIC2018` skin lesions [4] datasets. Images are resized to $128 \times 128 \times 3$ (classification) or $224 \times 224 \times 3$ (segmentation), split 80/10/10 for train/val/test, with standard on-the-fly augmentations.

`Nickparvar MRI` (D1) Consists of 7,023 T1-weighted axial brain MRI scans collected from Figshare, SARTAJ,

Table 2. Performance comparison of `HyPCA-Net` with SOTA methods on datasets `D1-D8` for classification. Bold and underlined values indicate the best and second-best results, respectively.

Datasets →		D1: Nickparvar			D2: IQ-OTHNCCD			D3: Tuberculosis			D4: CNMC-2019			D5: HAM10000			D6: SIPaKMeD			D7: CRC			D8: CBIS-DDSM			Overall	
Models ↓	Backbone ↓	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	#P	#F
NAT	Swin-T	95.5	95.5	95.6	97.5	97.3	97.2	95.1	94.3	95.5	94.1	93.7	93.7	93.1	92.6	93.3	91.2	91.1	91.5	95.9	95.8	96.2	92.1	91.7	91.9	20	1.1
POTTER	ResNet18	95.3	94.7	95.2	97.20	96.6	97.2	95.7	94.6	95.9	93.7	93.5	92.6	91.3	91.2	91.8	92.4	92.2	92.3	95.6	94.9	95.7	91.3	91.1	90.8	**12**	0.95
MMTM	ResNet18	95.9	95.4	95.8	97.5	97.3	97.8	95.3	94.5	95.7	91.9	91.6	92.4	90.9	89.3	90.6	88.7	87.6	90.3	96.3	95.4	96.6	92.7	91.6	90.9	31.6	**0.47**
AsymFusion	ResNet101	96.8	96.2	96.5	98.5	97.8	98.8	96.8	95.7	96.9	92.6	92.1	92.8	94.4	93.8	94.8	91.6	90.7	91.9	95.7	94.6	95.9	91.8	91.5	91.9	118.2	5.26
Gloria	ResNet50	98.1	97.6	97.9	98.5	98.4	98.5	96.6	96.0	96.9	93.3	93.3	93.4	93.8	93.8	94.5	94.2	94.2	94.2	95.9	95.6	95.7	92.5	91.2	91.9	30.8	1.54
MTTU-Net	ResNet50	97.9	97.9	98.0	99.5	99.2	99.5	97.3	96.6	97.6	94.3	93.9	94.1	97.4	96.5	97.2	91.9	92.3	92.6	96.9	96.8	97.0	94.1	93.3	94.5	38.1	6.8
HAMLET	ResNet50	96.3	95.9	96.2	98.0	97.5	98.3	96.8	96.3	96.3	92.8	92.6	92.9	93.5	93.4	93.2	92.8	92.4	93.3	95.7	95.3	96.2	91.9	91.8	91.3	57.3	3.52
MuMu	ResNet50	96.8	96.8	97.2	98.2	97.9	98.7	97.1	96.4	96.8	93.4	93.1	93.9	92.8	92.4	93.2	92.3	91.7	92.9	95.9	95.3	95.9	92.1	91.6	92.5	56.6	2.97
M^3*Att*	Swin-B	97.5	97.4	97.9	98.8	98.7	98.8	96.9	95.6	96.8	94.0	93.6	94.2	95.5	94.9	95.3	92.2	91.5	92.3	96.1	96.2	96.4	93.2	92.7	93.6	183	12.14
DRIFA-Net	ResNet18	98.4	98.4	98.7	99.7	99.5	99.5	98.2	97.5	98.6	96.4	96.3	96.7	98.2	97.9	98.5	95.6	95.5	95.9	97.0	96.8	97.1	95.2	95.1	95.4	53.8	4.83
HyPCA-Net18	ResNet18	98.8	98.7	99.0	99.8	99.8	99.8	98.9	98.0	99.2	97.2	97.1	97.5	99.4	99.3	99.7	95.7	95.7	96.1	98.3	97.9	98.5	96.3	96.2	96.5	14.47	2.25
HyPCA-Net50	ResNet50	99.2	99.1	**99.3**	**99.9**	**99.9**	**99.9**	99.1	98.3	**99.4**	97.5	97.3	97.9	99.6	99.6	99.8	96.7	96.4	97.2	98.5	98.1	98.7	96.9	96.6	97.2	28.4	3.03
HyPCA-Net-IN	Inception-v3	98.5	98.4	98.6	99.8	99.8	99.8	98.4	97.7	98.9	96.9	96.9	97.2	**100**	**100**	**100**	97.2	97.2	97.5	98.7	**98.7**	98.7	97.1	97.1	97.3	26.8	2.76
HyPCA-Net-ViT	ViT-Ti	**99.4**	**99.4**	**99.3**	**99.9**	**99.9**	**99.9**	**99.2**	**98.8**	99.2	**97.9**	**97.7**	**98.1**	**100**	**100**	**100**	**97.7**	**97.7**	**97.7**	**98.8**	**98.7**	**98.8**	**97.8**	**97.5**	**98.2**	22.5	3.42

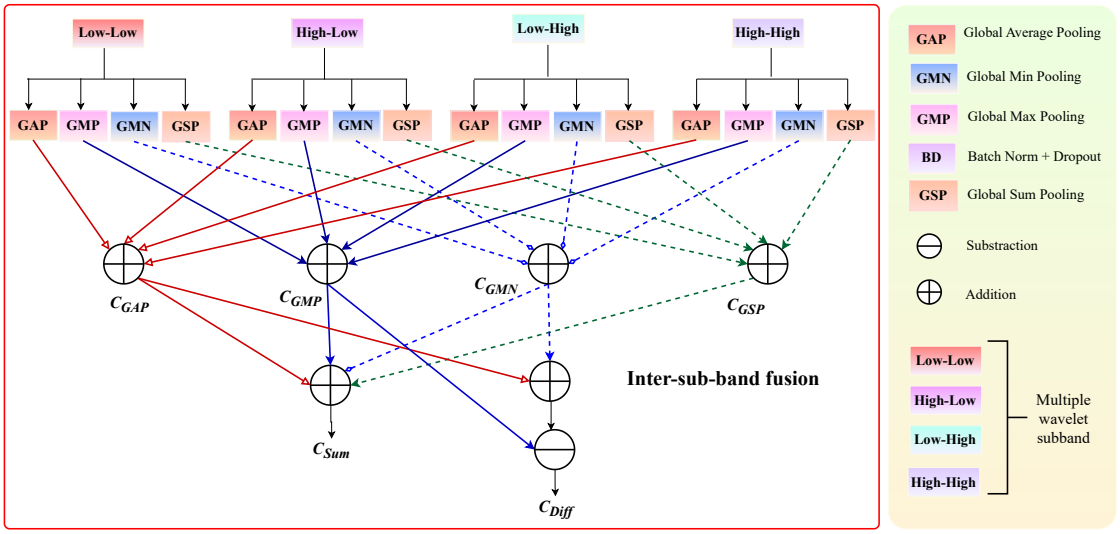

Figure 1. Overview of hierarchical channel fusion. It captures diverse channel dependencies, C_{Sum} and C_{Diff}, via an inter-sub-band fusion mechanism.

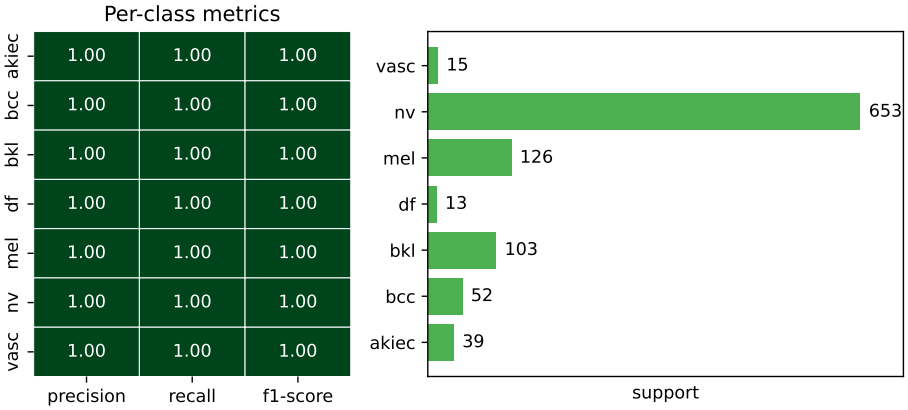

Figure 2. Classification report for HAM10000 test dataset.

and BrH35, covering four categories: Glioma (1,621), Meningioma (1,645), Pituitary (1,757), and No Tumor (2,000). Images are provided in DICOM format with 5 mm slice thickness and an in-plane resolution of 256×256 pix-

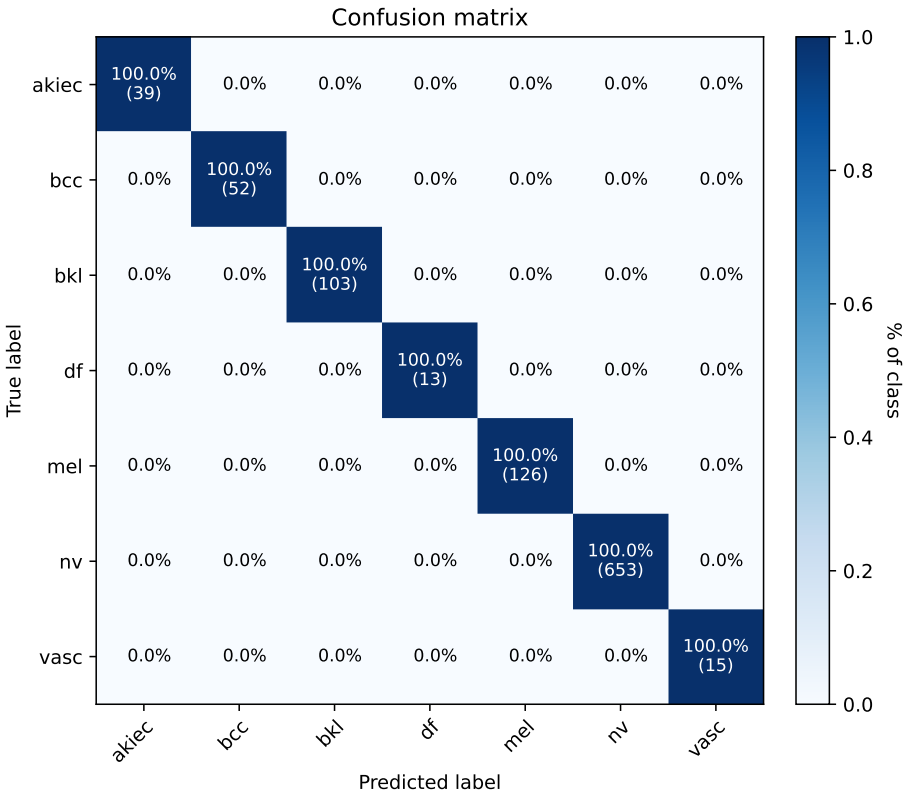

Figure 3. Confusion matrix for HAM10000 test dataset.

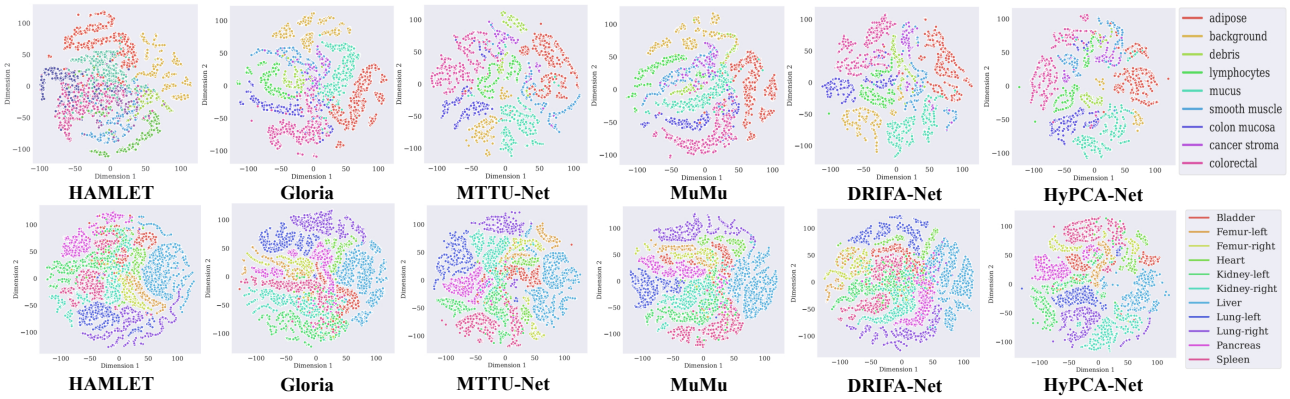

Figure 4. Visual representation of the important regions highlighted by our proposed `HyPCA-Net` and five `SOTA` methods using the `TSNE` technique on two benchmark datasets `PathMNIST` and `OrganAMNIST`.

els.

`IQ-OTHNCCD lung cancer` (D2) Comprises 1,190 CT-slice images from 110 subjects (55 normal, 15 benign, 40 malignant) acquired at the Iraq Oncology Teaching Hospital and National Center for Cancer Diseases over three months in 2019. Scans were performed on a Siemens SOMATOM at 120 kV with 1 mm slice thickness, window width 350–1,200 HU, and window center 50–600 HU; each volume contains 80–200 slices in DICOM format.

`tuberculosis CXR` (D3) Includes 4,200 posterior–anterior chest X-ray images evenly split between normal and tuberculosis cases, sourced from public repositories (e.g., Montgomery County and Shenzhen datasets). Images are provided in PNG format at 1024×1024 pixels

and intensity-normalized for classification tasks.

CNMC-2019 (D4) A challenge dataset of 15,114 single-cell microscopic images from bone marrow smears of 118 subjects (69 B-ALL patients, 49 healthy). All images (2560×1920 px PNG) are stain-normalized and segmented. The public splits include 10,661 training images (7,272 malignant, 3,389 healthy), 1,867 preliminary test images (1,219 malignant, 648 healthy), and 2,586 unlabeled final test images for classification [18].

HAM10000 (D5) Comprises 10,015 dermatoscopic images in JPEG format (600×450 px) across seven pigmented lesion categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesions. Ground-truth diagnoses were confirmed by expert dermatologists and histopathology.

SIPaKMeD (D6) Contains 4,049 high-resolution (1280×960 px PNG) Papanicolaou-stained cervical cell images from Pap smears, labeled into five classes: superficial–intermediate, parabasal, koilocytotic, dyskeratotic, and metaplastic. Images were manually localized and annotated by cytopathologists.

Colorectal Histology MNIST (CRC) (D7) Consists of 5,000 HE-stained colorectal tissue patches (150×150 px PNG) across eight histologic categories: tumor epithelium, stroma, lymphocytes, debris, normal mucosa, muscle, adipose, and mucus. Patches were extracted at 40× magnification and labeled by pathologists.

CBIS-DDSM (D8) The Curated Breast Imaging Subset of DDSM comprises 10,239 digitized mammography images (DICOM/PNG) from 753 studies with pathology-verified labels (benign vs. malignant). Pixel-level region-of-interest annotations are provided for lesion localization.

COVID-19 lung CT (D9) An open classification dataset of 349 RT-PCR–confirmed COVID-19 chest CT images and 463 non-COVID CT images from 216 and 233 patients, respectively. Images are in JPG format with variable resolutions (512×512 to 1024×1024 px), collected from multiple sources and confirmed by clinical testing.

ISIC2018 (D10) The ISIC 2018 Challenge classification dataset comprises 10,015 training, 193 validation, and 1,000 test dermoscopic images in JPEG format (ranging from 540×576 to 1200×1200 px) across seven diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesions. Expert annotations and diagnostic ground truth are provided.

Models We evaluate HyPCA-Net against a comprehensive set of state-of-the-art (SOTA) baselines. For classification (D1–D8), unimodal baselines are POTTER, DDA-Net, MADGNet's MFMSA [19], and EMCAD's MSCAM [23]—denoted as M1–M4; while multimodal fusion baselines comprises Gloria [11], AsymFusion [30], MMTM [14], MTTU-Net [3], HAMLET [12], MuMu [13], M3Att, and DRIFA-Net—denoted as M5–M12. For segmentation (D9–D10), we compare MTTU-Net and DRIFA-Net (with SegNet decoder), alongside SOTA unimodal baselines: UNet[25], UNet++[34], PolypPVT[7], TransFuse[32], MADGNet, EMCAD, PVT-CASCADE[22], PraNet[8], and denoted as M13–M20. Notably, we instantiate HyPCA-Net with four classification backbones—ResNet-18 [10], ResNet-50 [10], Inception-v3 [28], and ViT-Ti [27]—and SegNet [2] and EMCAD [23] decoders for segmentation. These variants are denoted as HyPCA-Net18, HyPCA-Net50, HyPCA-Net-IN, HyPCA-Net-ViT, HyPCA-Net-Seg, and HyPCA-Net-EMCAD.

Training Details. We follow the same training strategy as the DRIFA-Net model [6]. All models were trained for 200 epochs using cross-entropy loss and the Adam optimizer (initial learning rate 1×10^{-3}) on an NVIDIA RTX 4060 Ti GPU. A ReduceLROnPlateau scheduler was used with a minimum learning rate of 10^{-6}.

Table 3. Performance comparison of HyPCA-Net-Seg (SegNet backbone) with SOTA methods on D9–D10 datasets for multimodal segmentation tasks.

	D9: COVID-19		D10: ISIC			
Model	DSC	IOU	DSC	IOU	#P	#F
UNet++	65.7	57.5	87.3	80.3	9.16	34.7
AttnUNet	57.6	48.5	87.8	80.5	34.9	66.6
TransUNet	75.6	68.9	87.3	81.2	105.3	38.5
SwinUNet	80.3	73.6	88.8	81.9	27.2	6.2
TransFuse	77.9	72.5	89.1	82.3	143.7	82.7
DeepLab-v3+	81.4	75.2	83.8	74.5	39.8	14.9
UNeXt	53.4	44.4	61.3	54.5	**1.47**	**0.57**
PraNet	80.6	73.2	88.6	80.6	32.6	6.9
CaraNet	79.8	71.6	90.2	83.4	46.6	11.5
UACANet-L	82.4	76.2	89.8	82.5	69.2	31.5
HyPCA-Net-Seg	88.2	80.9	92.7	85.3	21.7	8.04
HyPCA-Net-EMCAD	**90.3**	**82.5**	**93.8**	**86.4**	18.6	7.65

4. Additional Experimental results

HyPCA-Net demonstrates exceptional classification and segmentation performance across 10 diverse medical imaging datasets, with improvements ranging from 80.9% to 100%. As shown in Tables 2–3, it consistently outperforms SOTA unimodal learning and MFL methods, with perfor-

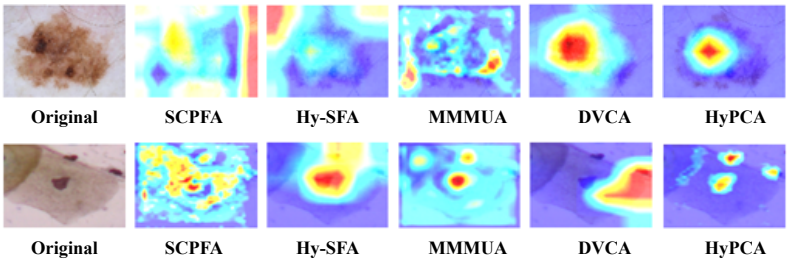

Figure 5. Grad-CAM-based visual comparison of discriminative regions highlighted by our proposed HyPCA block and its constituent attention components—SCPFA, Hy-SFA, MMMUA, and DVCA—on two benchmark datasets: HAM10000 (top row) and SIPaKMeD (bottom row). The visualizations demonstrate the enhanced focus and noise suppression achieved by HyPCA relative to its individual/combined attention mechanisms.

Table 4. Performance comparison of HyPCA-Net with existing methods on D11: PathMNIST and D12: OrganAMNIST [31]. Bold indicates the best result.

Model	D11: PathMNIST		D12: OrganAMNIST	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Gloria	91.94	99.78	95.68	99.80
MTTU-Net	92.28	99.34	95.47	99.85
HAMLET	91.85	99.29	95.28	99.80
MuMu	92.12	99.56	95.70	99.80
M^3Att	91.72	99.15	95.84	99.85
DRIFA-Net	93.10	99.75	96.41	99.90
HyPCA-Net	**93.95**	**99.95**	**97.37**	**99.94**

mance improvements of $\approx 42.3\%$, while reducing parameters by $\approx 92\%$ and FLOPs by $\approx 81.47\%$.

We also evaluated our model's performance on the HAM10000 dataset using a classification report—highlighting class-wise precision, recall, and F1-score (refer to Figure 2)—and a confusion matrix (refer to Figure 3).

We also conducted experiments on two MedMNIST datasets—PathMNIST and OrganAMNIST [31]—where our HyPCA-Net achieves superior performance compared to existing multimodal fusion baselines, as shown in Table 4.

Discussion: HyPCA-Net generalizes effectively across multiple medical imaging modalities by leveraging cross-modal interactions to learn robust shared representations. In contrast, unimodal baselines (e.g., DDA-Net, NAT, POTTER, MSCAM, MFMSA, EMCAD) process each modality independently and thus cannot exploit complementary information. Likewise, existing multimodal fusion methods (e.g., M^3Att, MTTU-Net, HAMLET, MuMu, Gloria) often limited focus that do not fully leverage the potential benefits for learning robust shared representations by capturing complementary information across diverse medical imaging modalities. By integrating these cues holistically, HyPCA-Net captures more discriminative shared representations, leading to superior performance across heterogeneous medical imaging datasets.

To address **Challenge 1**, our HyPCA-Net achieves an effective balance between optimal performance and minimal computational cost by incorporating two efficient and effective modules – RALA and DVCA. Specifically, RALA refines modality-specific representations via capturing multi-scale spatial–channel dependencies, while DVCA captures dual-domain contextual cues through a cascaded hybrid-space mechanism. Together, these modules enable HyPCA-Net to achieve optimal performance with minimal computational overhead, making it well suited for resource-constrained environments (Tables 2–3. Many prior works rely on either a single attention mechanism – such as HAMLET, MuMu – which limits their capacity to learn rich shared representations [6] – or stack multiple attention modules in a purely cascaded fashion (e.g., DDA-Net, MSCAM, MFMSA, MADGNet, EMCAD, DRIFA-Net), which potentially risks progressive information loss during inter-module transitions. Because cascaded designs are inherently *sequential*, they process different aspects in isolation, lack joint optimization, and are therefore *limited in preserving holistic information*, ultimately *constraining the richness of learned representations*. In contrast, our hybrid attention design seamlessly integrates parallel fusion attention with cascaded attention to preserve holistic cues while *jointly* and *cascadingly* modeling spatial–channel dependencies in both hybrid space and dual-domain representation learning. As a result, HyPCA-Net overcomes the intrinsic drawbacks of purely cascaded schemes and directly addresses **Challenge 2**, yielding more robust shared representation learning.

4.1. Extended Ablation Study

Tables 5–6 present a component-wise ablation study that quantifies the contributions of each HyPCA module—SCPFA, Hy-SFA, and MMMUA—together with their constituent components (MSHC, CHIA, SHIA, FCIF, SMIF, MCBI, TFSI, and FDCA) on the D5 (HAM10000) and D6 (SIPaKMeD) datasets.

Table 5. Ablation analysis of HyPCA-Net's integrated modules and their components contributing to optimal performance.

Integrated components of HyPCA-Net																				
D5: HAM10000										D6: SIPaKMeD										
MSHC	CHIA	SHIA	FCIF	SMIF	MCBI	TFSI	FDCA	Acc	F1	MSHC	CHIA	SHIA	FCIF	SMIF	MCBI	TFSI	FDCA	Acc	F1	#P	#F
✓	×	×	×	✓	×	×	✓	97.7	97.7	✓	×	×	×	✓	×	×	✓	93.5	93.5	**8.73**	**0.81**
✓	✓	×	×	✓	×	✓	×	97.6	97.4	✓	✓	×	×	✓	×	✓	×	93.8	93.7	12.6	1.9
×	×	✓	×	✓	×	×	✓	97.9	97.9	×	×	✓	×	✓	×	×	✓	94.1	94.0	26.7	1.52
✓	✓	✓	×	✓	×	×	✓	98.4	98.4	✓	✓	✓	×	✓	×	×	✓	94.7	94.5	11.6	0.92
×	×	✓	✓	✓	✓	×	✓	98.5	98.3	×	×	✓	✓	✓	✓	×	✓	94.4	94.4	27.5	1.73
✓	✓	✓	✓	✓	✓	×	✓	98.9	98.8	✓	✓	✓	✓	✓	✓	×	✓	95.1	95.0	12.8	0.97
✓	✓	×	✓	✓	✓	✓	×	98.3	98.3	✓	✓	×	✓	✓	✓	✓	×	94.5	94.3	13.1	1.97
×	×	✓	✓	✓	✓	✓	×	97.9	97.8	×	×	✓	✓	✓	✓	✓	×	94.5	94.4	30.7	2.95
✓	✓	×	✓	✓	✓	×	✓	98.6	98.6	✓	✓	×	✓	✓	✓	×	✓	94.9	94.9	11.2	0.94
✓	×	×	✓	✓	✓	×	✓	98.2	98.0	✓	×	×	✓	✓	✓	×	✓	94.7	94.4	9.8	0.87
✓	✓	✓	✓	✓	✓	✓	✓	**99.4**	**99.3**	✓	✓	✓	✓	✓	✓	✓	✓	**95.7**	**95.7**	14.47	0.87

Table 6. Ablation study with each HyPCA attention-driven modules – SCPFA, Hy-SFA, and MMMUA on benchmark D5 and D6 datasets. Here, MSHC remains fixed for all.

Dataset	SCPFA	Hy-SFA	MMMUA	F1	Dataset	SCPFA	Hy-SFA	MMMUA	F1	#P
D5	×	×	×	95.3	**D6**	×	×	×	92.1	**7.42**
	×	×	✓	98.2		×	×	✓	94.5	9.13
	×	✓	×	97.8		×	✓	×	94.1	10.3
	✓	×	×	97.1		✓	×	×	93.6	9.84
	✓	✓	✓	**99.4**		✓	✓	✓	**95.7**	14.5

- **Single-module additions.** Introducing any one component to the lightest configuration (8.73 M parameters, 0.81 GFLOPs) raises $F1$ by roughly 0.7–1.2 points while incurring only 3–30% more parameters and 4–30% more FLOPs.
- **Full integration.** Enabling all eight modules yields the highest scores—**99.3%** (↑ 1.6) on D5 and **95.7%** (↑ 2.2) on D6—using 14.47 M parameters and 0.87 GFLOPs. Relative to the lightest variant, this represents a 66% increase in parameters but only a 7% increase in FLOPs, underscoring an attractive accuracy–efficiency trade-off.

These results demonstrate that each module contributes incremental gains and that their combined effect is synergistic, enabling HyPCA-Net to balance strong performance with computational economy.

5. Additional Qualitative Analysis

Figure 5 displays qualitative Grad-CAM heat-maps that pinpoint the image areas most responsible for each model's predictions. In HAM10000, these regions coincide with visible skin lesions; in SIPaKMeD, they mark abnormal cervical cells. The comparison illustrates how the individual and combined attention modules—SCPFA, Hy-SFA, MMMUA, and DVCA—attend to task-relevant features, set against the sharply focused, noise-suppressing maps produced by our HyPCA-based architecture. Specifically:

(i) SCPFA spreads its attention across broad, largely nondiscriminative zones, introducing considerable visual noise and obscuring the class-informative details required for reliable classification. (ii) Hy-SFA achieves tighter localization around key structures but still lights up peripheral, less informative regions, indicating only partial suppression of irrelevant cues. (iii) MMMUA narrows its focus more effectively than SCPFA and Hy-SFA, outlining lesion boundaries with greater precision, yet small peripheral highlights persist, hinting at residual noise. (iv) DVCA further refines the heat-maps, substantially reducing extraneous activations compared with the hyperbolic-only approaches, though faint, diagnostically unhelpful artifacts sometimes remain. (v) Our HyPCA block produces the cleanest and most interpretable maps, crisply isolating crucial regions while nearly eliminating peripheral distractions, thereby providing a clearer window into the model's decision-making process.

We also carried out an additional qualitative analysis by projecting the high-dimensional representations from each model's final embedding layer into two dimensions using t-SNE. Specifically, we compared features produced by top-performing multimodal fusion baselines and our own HyPCA-Net architecture. The resulting 2D scatter plots—shown for the PathMNIST and OrganAMNIST subsets of MedMNIST [31] in Figure 4—highlight how well each method separates distinct tissue classes.

6. Limitations and Planned Clinical Validation

Our evaluation to date is limited to public benchmark datasets, so HyPCA-Net's performance on clinical-grade data remains untested. In future, we plan to bridge this gap by validating the model on proprietary, clinician-curated datasets that reflect real-world workflows. Prior to clinical deployment, we will incorporate rigorous privacy protections and bias-mitigation strategies and conduct comprehensive validation—including prospective clinical trials, regulatory review, and adversarial robustness testing—to ensure safe and trustworthy use in resource-constrained healthcare settings.

References

[1] Hamdalla Alyasriy and A Muayed. The iq-othnccd lung cancer dataset. *Mendeley Data*, 1(1):1–13, 2020. 1

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture

for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 4

[3] J. Cheng, J. Liu, H. Kuang, and J. Wang. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. *IEEE Transactions on Medical Imaging*, 41(6):1520–1532, 2022. 4

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 1

[5] Y. Cui, Y. Tao, W. Ren, and A. Knoll. Dual-domain attention for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 479–487, 2023. 1

[6] Joy Dhar, Nayyar Zaidi, Maryam Haghighat, Puneet Goyal, Sudipta Roy, Azadeh Alavi, and Vikas Kumar. Multimodal fusion learning with dual attention for medical imaging. *arXiv preprint arXiv:2412.01248*, 2024. 1, 4, 5

[7] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021. 4

[8] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 4

[9] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6185–6194, 2023. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[11] S. C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 1, 4

[12] Md Mofijul Islam and Tariq Iqbal. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10285–10292. IEEE, 2020. 1, 4

[13] M. M. Islam and T. Iqbal. Mumu: Cooperative multitask learning-based guided multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1043–1051, 2022. 1, 4

[14] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020. 1, 4

[15] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, et al. Covid-19 ct lung and infection segmentation dataset. *(No Title)*, 2020. 1

[16] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1): 1–11, 2016. 1

[17] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE Transactions on Image Processing*, 32:3054–3065, 2023. 1

[18] S Mourya, S Kant, P Kumar, A Gupta, and R Gupta. All challenge dataset of isbi. 2019. *The Cancer Imaging Archive*, 2019. 1, 4

[19] Ju-Hyeon Nam, Nur Suriza Syazwany, Su Jung Kim, and Sang-Chul Lee. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11480–11491, 2024. 1, 4

[20] Msoud Nickparvar. Brain tumor MRI dataset. Data set, 2021. Accessed on 3rd March. 1

[21] Maria E Plissiti, Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, Orestis Krikoni, and Avraam Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3144–3148. IEEE, 2018. 1

[22] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6222–6231, 2023. 4

[23] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024. 1, 4

[24] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *Ieee Access*, 8:191586–191601, 2020. 1

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4

[26] R Sawyer-Lee, F Gimenez, A Hoogi, and D Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm)[skup podataka]. *The cancer imaging archive*, 2016. 1

[27] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 4

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[29] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1

[30] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3902–3910, 2020. 1, 4

[31] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 5, 6

[32] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*, pages 14–24. Springer, 2021. 4

[33] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1620, 2023. 1

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*, pages 3–11. Springer, 2018. 4