Modality-Specific Encoding with a Shared Backbone for Early HCC Prediction from Multi-Modal Liver MRI

UNIVERSITY OF TWENTE

Medisch Spectrum Twente
Enschede, The Netherlands

Axel Faes[1], Maureen M. J. Guichelaar[2], Donald E. Bouman[3], Stéphanie M. van den Berg[1], Maryam Amir Haeri[1]

[1] Dept. of Learning, Data Analytics and Technology, University of Twente [2] Dept. of Gastroenterology and Hepatology, MST [3] Dept. of Radiology, MST

BACKGROUND

Hepatocellular carcinoma (HCC) is the most common primary liver cancer and a leading cause of cancer-related death. Surveillance relies on multi-modal liver MRI: DWI, T1 in-phase/opposed-phase, T2-weighted, and dynamic contrast-enhanced sequences, each measuring different tissue properties. Current deep learning approaches either use a single modality or naively concatenate all sequences, forcing a single network stem to handle physically distinct signals simultaneously.

> **Hypothesis:** Modality-specific encoder heads, feeding into a shared backbone that learns modality-invariant spatial patterns, will outperform naive concatenation for HCC prediction, particularly when combined with per-modality self-supervised pretraining.

MRI MODALITIES

Diffusion-Weighted
Measures water molecule diffusion. Restricted diffusion indicates high cellularity — a hallmark of malignancy.
`b0 · b150 · b400 · b800`

T1W IP / OOP
Signal dropout between IP and OOP reveals fat–water content. Quantifies hepatic steatosis.
`T1W_IP · T1W_OOP`

T2-Weighted
Sensitive to fluid content and iron deposition. Signal decay between TEs reflects iron overload.
`T2W_TES · T2W_TEL`

Dynamic Contrast-Enh.
Captures vascular perfusion. Arterial hyperenhancement and washout are key HCC features.
`T1_Pre · Art · Del · Ven`

DATASET

Retrospective single-centre cohort from Medisch Spectrum Twente, Enschede.

~240	~37	15.4%
PATIENTS	HCC-POSITIVE	PREVALENCE

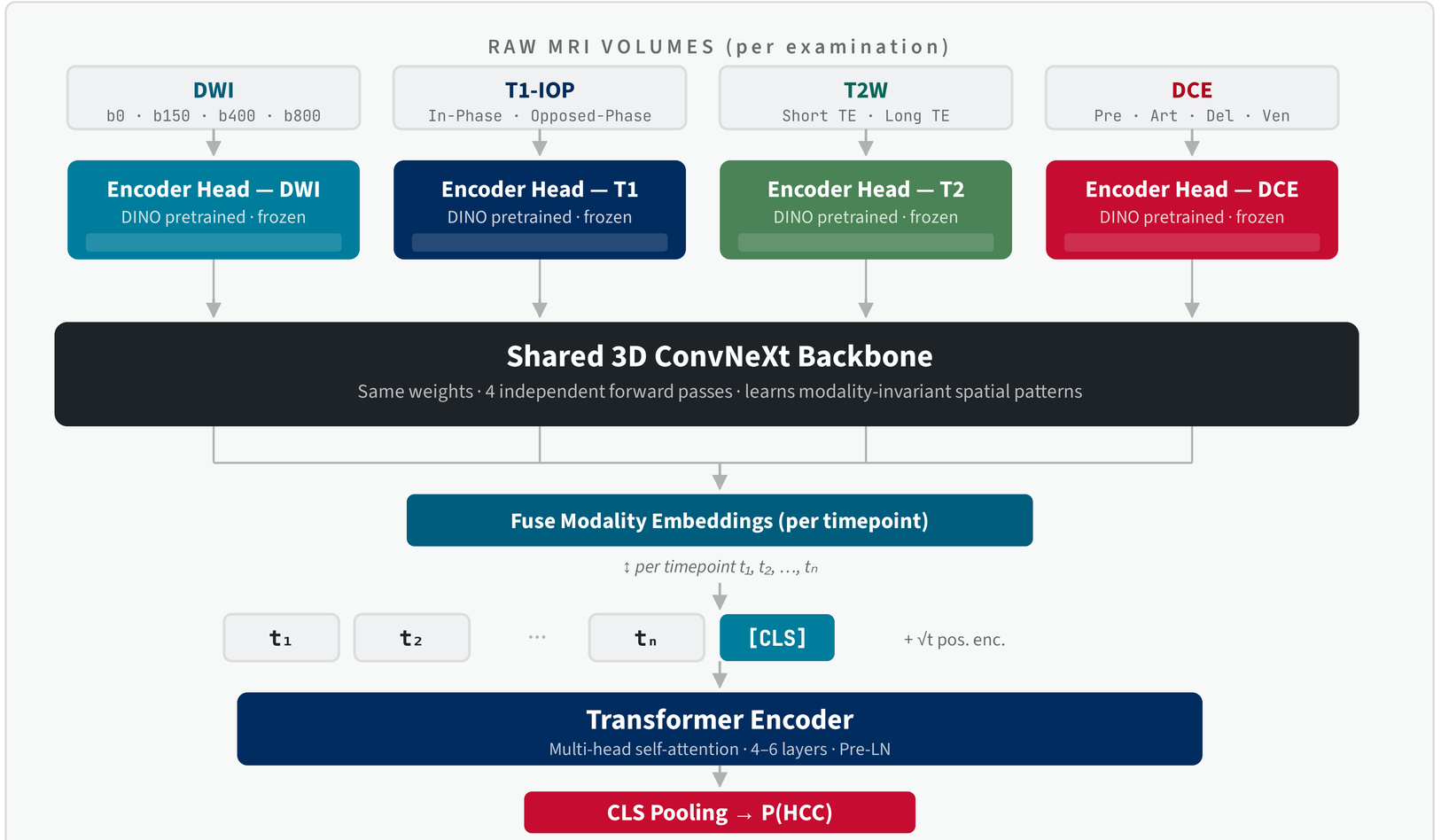
TRAINING STRATEGY

Per-modality DINO pretraining: contrastive views are only constructed *within* a modality group, never across:

1. **Per-Modality DINO** — Each encoder head pretrained on its modality, then frozen
2. **Backbone + Trans.** — Shared backbone trained supervised on P(HCC)
3. **End-to-End** — Full system, all parameters unfrozen

ARCHITECTURE

Modality-specific encoder heads are pretrained via DINO then frozen; the shared backbone processes each modality's features independently using the same weights. Fused embeddings feed a Transformer for longitudinal modelling.



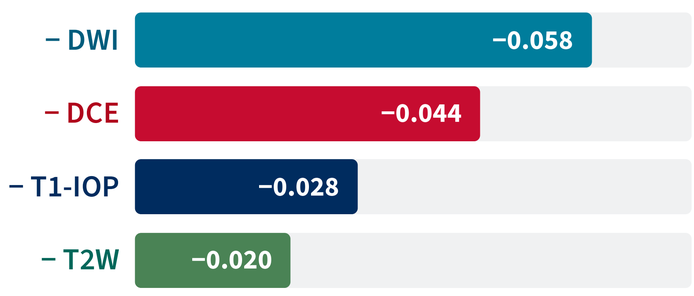
RESULTS

Architecture	Pre.	DWI	T1-IOP	T2W	Multi
Naive concat	✗	0.718	0.720	0.705	0.734
Naive concat	✓	0.795	0.690	0.670	0.772
Enc. heads + shared	✗	0.729	0.731	0.714	0.787
Enc. heads + shared	✓	**0.808**	0.753	0.738	**0.844**

AUROC, 5 folds × 3 seeds. Red: degradation. Green: improvement from pretraining.

MODALITY CONTRIBUTION

AUROC drop when each modality is removed from the full pretrained system:



Relative to full system (0.844). DWI and DCE contribute most, in line with their clinical diagnostic value.

SINGLE MODALITY

AUROC when training with only one modality vs. all four combined:

Multi-modal fusion provides a further gain over any single modality.

KEY FINDINGS

1. Encoder heads mitigate the pretraining disparity: all modalities benefit from DINO when pretrained per modality group.
2. The shared backbone improves multi-modal fusion by +0.072 AUROC over naive concatenation, indicating that modality-invariant spatial patterns are learnable.
3. Without pretraining, the architectural decomposition alone yields +0.053 AUROC. The encoder-backbone separation has value independent of the pretraining strategy.

axel.faes@utwente.nl · m.amirhaeri@utwente.nl

University of Twente · Learning, Data Analytics and Technology