

HCCNet: A Self-Supervised CNN-Transformer for Disease Prognosis from Longitudinal 3D MRI, with a Case Study on Hepatocellular Carcinoma

Jakob Nolte^{a,c,1}, Axel Faes^{c,1,*}, Maureen M. J. Guichelaar^a, Donald E. Bouman^b, Stéphanie M. van den Berg^c, Maryam Amir Haeri^c

^a*Department of Gastroenterology and Hepatology, Medisch Spectrum
Twente, Enschede, The Netherlands*

^b*Department of Radiology, Medisch Spectrum Twente, Enschede, The Netherlands*

^c*Department of Learning, Data Analytics and Technology, University of
Twente, Enschede, The Netherlands*

Abstract

Background: Longitudinal MRI analysis is important for predicting disease outcomes, particularly in chronic conditions such as hepatocellular carcinoma (HCC), where early detection can considerably influence treatment strategies and patient prognosis. Because of limited data availability, subtle parenchymal changes, and the irregular timing of medical screenings, current approaches have so far focused on cross-sectional imaging data.

Methods: We propose HCCNet, which integrates a 3D adaptation of the ConvNeXt CNN architecture with a Transformer encoder to capture the fine-grained spatial features of 3D MRIs and the temporal dependencies across time points. A two-stage pre-training process is tailored to longitudinal MRI: the CNN backbone is pre-trained with a self-supervised framework adapted for 3D MRIs, and the Transformer encoder with a sequence-order-prediction task that strengthens its understanding of disease progression. We apply HCCNet to a cohort of liver cirrhosis patients undergoing regular MRI screenings for HCC surveillance.

Results: HCCNet substantially improves predictive accuracy and reliability over baseline models. With diffusion-weighted MRI, the average across model architectures reaches an AUPRC of 0.614 and an AUROC of 0.901. An exploratory naive fusion of imaging modalities yields only marginal gains over the strongest single modality.

Conclusion: HCCNet provides a reliable tool for personalized HCC surveillance. The approach is versatile and can be adapted to various longitudinal MRI screening applications. Its ability to handle varying patient record lengths and irregular screening intervals makes it a framework for monitoring chronic dis-

*Corresponding author: Axel Faes

Email address: axel.faes@utwente.nl (Axel Faes)

¹These authors contributed equally to this work and share first authorship.

eases, where timely and accurate prognosis is critical for effective treatment planning.

Keywords: Longitudinal Medical Imaging, Self-Supervised Pre-Training, Convolutional Neural Network, Transformer, Hepatocellular Carcinoma

1. Introduction

Predicting disease outcomes based on longitudinal MRI scans is important for improving patient management. This is especially relevant in conditions like hepatocellular carcinoma (HCC), where early detection of tissue changes and subsequent monitoring can considerably impact treatment decisions and prognosis (Parikh et al., 2020). Longitudinal MRI analysis enables the capture of disease progression over time, providing a detailed view of how pathological changes evolve (Gong et al., 2024). However, despite its potential, the development of effective predictive models that can fully exploit the temporal information in longitudinal MRI data remains a significant challenge (Jin et al., 2021).

One of the major challenges lies in accurately detecting early signs of disease progression from MRI scans, particularly in identifying subtle changes in the underlying parenchyma that may eventually lead to a malignancy. These challenges are further compounded by the need to accurately capture complex temporal dependencies across multiple time points while preserving the spatial integrity of the 3D MRI data (Tang et al., 2022). Traditional approaches often simplify this task by extracting 2D slices from 3D MRI scans and applying convolutional neural networks (CNNs) pre-trained on large-scale natural image datasets (Kumar and Marttinen, 2024). However, this method has limitations, particularly in the medical domain where datasets are often small, and the differences between natural and medical images, such as texture, contrast, and noise, considerably impact the transferability of pre-trained models (Zeng et al., 2021).

The scarcity of large, labeled datasets in medical imaging exacerbates these challenges, making it difficult to train models that generalize well across different patient populations. As a result, pre-training approaches have become essential in the medical domain, helping to improve model performance by enabling the network to learn reliable features from limited data (Tang et al., 2022). However, many existing approaches have not fully explored the potential of these techniques in the context of longitudinal MRI analysis, particularly for diseases that require monitoring over extended periods (Parikh et al., 2020). Likewise, most deep learning applications are aimed at detecting a malignancy, resulting in a lack of studies focusing on detecting tissue changes before the tumor actually develops.

In response to these limitations, we propose HCCNet, a novel model architecture designed specifically for longitudinal 3D MRI analysis. HCCNet integrates a 3D adaptation of the ConvNeXt architecture with a Transformer encoder,

allowing it to capture both the spatial features of 3D MRIs and the temporal dependencies between different scans (Gong et al., 2024). Unlike previous approaches that rely on 2D slices or disregard the irregular intervals between scans, HCCNet is designed to handle the complexities of longitudinal data, making it better suited for tasks like predicting disease progression (Jin et al., 2021).

To address the challenges of early detection and small datasets, HCCNet incorporates an effective pre-training strategy. The CNN backbone is pre-trained using an extended version of the DINO framework, adapted to 3D MRI, which treats different MRI sequences as natural augmentations of the same image (Tang et al., 2022). This self-supervised learning approach allows the model to learn from the inherent variability in MRI data, even when labeled examples are scarce (Kumar and Marttinen, 2024). Additionally, the Transformer encoder is pre-trained with a sequence-order-prediction task inspired by techniques used in natural language processing, which helps the model to understand the progression of disease over time and to make accurate predictions despite the irregular timing of medical screenings (Zeng et al., 2021).

Our study specifically applies HCCNet to the task of predicting HCC development in liver cirrhosis patients, based on longitudinal MRI scans. By fine-tuning the pre-trained model on this dataset, we demonstrate the model’s ability to predict whether a patient will develop HCC at the next examination (Parikh et al., 2020). This application not only demonstrates HCCNet’s capabilities but also addresses a key need in personalized HCC surveillance of cirrhotic patients, where timely and accurate predictions can considerably influence clinical outcomes. The methodological approach presented in this paper can also be considered for predicting other conditions that, similar to HCC, benefit from monitoring changes over time. Thus, HCCNet provides a flexible framework that can be adapted to various longitudinal medical imaging tasks, potentially improving predictive accuracy and patient outcomes across a range of diseases. All code for the proposed method is available at: github.com/TheAxeC/hccnet.

2. Related Work

2.1. Modeling Approaches in Longitudinal Medical Image Analysis

To date, deep learning applications to medical imaging still predominately consider cross-sectional instead of repeated imaging data. However, physicians usually assess the progression of disease or the effect of treatment by comparing patients’ current to previous imaging records. As a result, longitudinal imaging has long been established as the gold standard in many branches of oncology, enticing recent studies to extend the analysis of medical imaging to the analysis of imaging data across time.

Three main methodological frameworks have emerged from these studies. Among them, fully convolutional neural networks (CNNs) model the within person change across time by stacking the image screenings from multiple time points into a multidimensional image and feeding the single joint image through the network (Gao et al., 2021; Dammu et al., 2023). While the approach is

relatively simple to implement, it has been shown to not fully explore the relationship between time points (Dammu et al., 2023). Studies investigating treatment effects have therefore increasingly employed Siamese networks (Jin et al., 2021; Dammu et al., 2023). These networks constitute a parallel architecture, where both parts receive different inputs but share the same set of parameter weights. Applied to treatment response analysis, pre- and post-treatment images are thus fed independently through the parallel network architecture and the final embedding is obtained by concatenating the individual vector representations. Despite promising results, Siamese networks are only applicable to the comparison of pairs of time points, making them inadequate for the analysis of long-term disease trajectories.

To mitigate this shortcoming, our study uses a hybrid model architecture, which is the third most common framework in longitudinal medical image analysis. Contrary to the other approaches, these models not only rely on a CNN backbone to embed the raw image screenings in high-dimensional feature space but also employ a sequential model to process the sequence of embedded image representations. Hybrid models can be trained end-to-end, consecutively, or using a pre-trained CNN feature extractor and typically employ a recurrent neural network (RNN), long-short-term memory (LSTM) or Transformer model to process the sequence of image embeddings. Compared to the aforementioned approaches, hybrid models can handle patient records of varying time points, making them especially suited for the continued surveillance of patients at high-risk of disease or after the onset of treatment. As such, they have been successfully employed in response prediction to radiotherapy (Lee et al., 2022; Wang et al., 2019, 2020; Xu et al., 2019) and endothelial therapy (Lu et al., 2021) as well as in breast cancer risk prediction (Dadsetan et al., 2022). Analyzing ultrasound (US) images from 619 cirrhotic subjects, Zhang et al. (2022) even applied them to HCC risk prediction, obtaining a moderate area under the receiver operating characteristic curve (AUC) of 0.775.

However, all studies employing hybrid models rely, to the best of our knowledge, on 2D imaging modalities like ultrasound or the extraction of 2D slices from 3D MRI or CT volumes. Likewise, many of them disregard the irregular time intervals between image screenings. In this study, we therefore propose a spatio-temporal 3D CNN-Transformer hybrid model that can capture the long-range dependencies among longitudinal MRI scans and uses time interval based positional encodings to account for the irregular examination intervals between image screenings.

2.2. Pre-Training Approaches in Longitudinal Medical Image Analysis

While the temporal relationship across medical image screening is increasingly investigated, pre-training approaches for longitudinal medical image analysis remain scarcely explored. As a result, many deep learning studies analyzing medical images over time continue to extract 2D slices from 3D medical images to use CNN model weights pre-trained on large scale natural image datasets. However, the usefulness of transfer learning from natural images remains fiercely disputed (Zhang et al., 2023; Huang et al., 2023; El-Nouby et al., 2021) and our

study also employs a 3D rather than a 2D CNN. Using the potential abundance of unlabeled data, some longitudinal medical imaging studies have therefore started to apply self-supervised pre-training approaches to enhance downstream model performance.

Among them, contrastive learning approaches remain the most commonly applied pre-training framework. These approaches build on the assumption that alterations caused by image transformations do not alter the image’s semantic meaning. As such, they randomly derive different views from the same input image (Chen et al., 2020) or from the same patient at different time points and learn rich contextual feature embeddings by matching the view’s embedded representation to the one obtained with the other part of a Siamese network or a momentum encoder (Grill et al., 2020). Jamaludin et al. (2017), for example, pre-trained a Siamese CNN on baseline and follow-up image pairs by minimizing the distance between view pairs from the same patient and maximizing the distance between images from different patients. Rivail et al. (2019) extended this approach such that it accounts for irregularly sampled data. Instead of simply comparing view pairs from different patients, they additionally considered the time interval between image screenings from the same patient for their loss calculation.

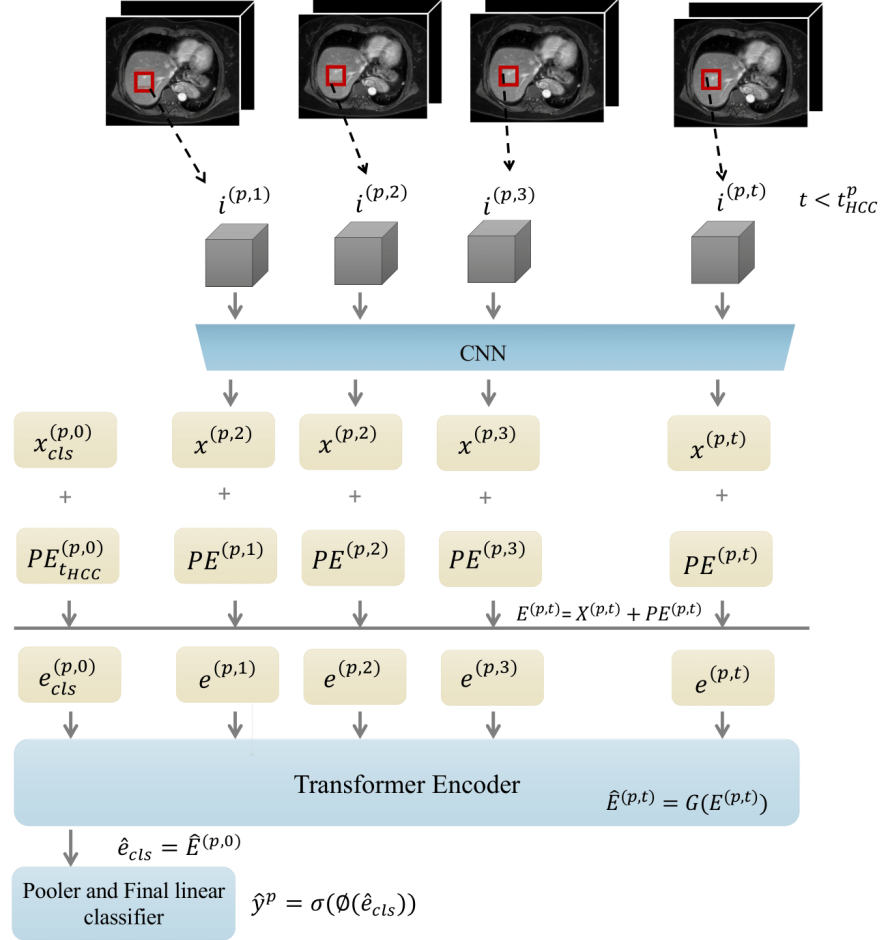
While above approaches proved successful in their respective downstream application, they do not provide a framework for patient records of varying lengths or hybrid model approaches in general. We therefore propose a novel pre-training framework for longitudinal 3D MRI scans that captures both spatial and temporal dependencies both within and between patients. The framework borrows from the contrastive DINO framework proposed by Caron et al. (2021) for pre-training of the CNN backbone and adopts a simple sequence-order-prediction task inspired by Lan et al. (2020) and Ren et al. (2021) for pre-training of the Transformer encoder. Further details on the study’s pre-training implementation are discussed in section 3.4.

3. Methodology

3.1. Problem Formulation

This study aims to define a mapping function f that for each liver cirrhosis patient learns to predict whether the patient will be diagnosed with HCC at the next examination given a set of MR images from past patient visits. Since the frequency of liver screenings is typically determined by the patient’s unique condition, some patients may undergo more frequent screenings than others with more or less regular time intervals between examinations. As such, the data of each patient $p \in \{1, 2, \dots, P\}$ corresponds to a vector of unique length and is denoted by $D^p = \{i^{(p,1)}, i^{(p,2)}, \dots, i^{(p,t)}; t_{HCC}^p\}$, where $i^{(p,t)}$ is the patient’s MRI examination at time point t and t_{HCC}^p is the patient’s date of diagnosis which is set to None if HCC was not diagnosed within t . For each patient with $t_{HCC}^p \neq \text{None}$, patient records are subset during fine-tuning such that the patient’s last considered observation $i^{(p,t)}$ has time point $t < t_{HCC}^p$. Likewise, the label for

Figure 1: Proposed Modeling Approach



Note: The figure presents the modeling approach for an exemplary patient with longitudinal MRI screening. Raw MRI's are fed through the CNN backbone before adding positional encodings to the image embeddings. The resulting embeddings are fed through the Transformer encoder and the last hidden state of the $[cls]$ token used for prediction.

the sequence of patient visits in subset $V^{(p,t)} = \{i^{(p,1)}, i^{(p,2)}, \dots, i^{(p,t)}\} \subseteq D^{(p,t)}$ is set to:

$$\mathbf{y}_p = \begin{cases} 0, & \text{if } t_{HCC}^p = \text{None}, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

3.2. Overall Architecture

To handle the longitudinal nature of the patient records, HCCNet is a CNN-Transformer model based on a 3-dimensional adaptation of ConvNeXt (Liu et al., 2022) and the original Transformer encoder architecture (Vaswani, 2017). We use the CNN backbone to embed the raw MR images in higher-dimensional feature space and the Transformer encoder to process the sequence of embedded image representations. A high level overview of the proposed modeling approach is provided in Figure 1.

CNN Feature Extractor. For the CNN backbone, we adapt the ConvNeXt architecture proposed in (Liu et al., 2022). More specifically, we replace the 2-dimensional depthwise convolutional layers in ConvNeXt with their 3-dimensional counterpart and reduce the proposed kernel size from $k = 7$ to $k = 3$. Additionally, we reimplement the patchify layer in ConvNeXt with a smaller kernel size and stride for the convolutional layer resulting in a 3x rather than 4x downsampling of the input images and thus accommodating for the lower input size typically encountered in volumetric data. More details on the study’s 3D implementation of ConvNeXt are presented in Appendix A.1.

The CNN feature extractor takes the MR images as inputs and returns a high-dimensional vector representation of the raw images. The vector representation can be denoted by:

$$\mathbf{x}^{(p,t)} = F(i^{(p,t)}), \quad (2)$$

where F is the CNN backbone, $\mathbf{x}^{(p,t)} \in \mathbb{R}^H$ is the resulting embedding for patient p at time point t , and $H = \dim(\mathbf{x}^{(p,t)})$.

Transformer Encoder. The Transformer encoder is based on the Pre-LN implementation (Xiong et al., 2020) of the original Transformer encoder architecture (Vaswani, 2017) and is implemented using L encoder layers with a hidden size of $H = \dim(\mathbf{x}_{(p,t)})$ and $A = \frac{H}{128}$ self-attention heads. To model the disease trajectory over time, we concatenate the image embeddings obtained from the CNN feature extractor and prepend a learnable classification token [cls] yielding the following sequence of image representations:

$$\mathbf{X}^{(p,t)} = [\mathbf{x}_{\text{cls}}^{(p,0)}, \mathbf{x}^{(p,1)}, \mathbf{x}^{(p,2)}, \dots, \mathbf{x}^{(p,t)}] \in \mathbb{R}^{(t+1) \times H}, \quad (3)$$

where $\mathbf{x}^{(p,t)}$ is the previously derived image embedding and $\mathbf{x}_{\text{cls}}^{(p,0)}$ is the [cls] token of patient p .

Following that, we embed the input’s positional information using fixed sinusoidal positional encodings. To account for the irregularly spaced intervals between examinations, we replace the conventional sequence position with the elapsed time (in years) between each examination and the *prediction reference point* t_{ref}^p , the screening examination at which the outcome is assessed. The reference point is determined solely from the observed examination dates, identically for patients who do and who do not develop HCC, and never uses the outcome itself; the encoding is therefore well defined at inference for a patient whose status is unknown. Given the examination times $\mathbf{t}^p = (t_1, t_2, \dots, t_i; t_{\text{ref}}^p)$

for patient p , we compute the vector of intervals

$$\Delta \mathbf{t}_i^p = t_{\text{ref}}^p - t_i, \quad (4)$$

prepend $\Delta \mathbf{t}_{\text{cls}}^p = 0$ for the classification [cls] token, and, finally, obtain positional encodings using sine and cosine functions of different frequencies:

$$\begin{aligned} \mathbf{PE}_{(\Delta \mathbf{t}_i^p, 2k)} &= \sin(\Delta \mathbf{t}_i^p / 10000^{2k/H}), \\ \mathbf{PE}_{(\Delta \mathbf{t}_i^p, 2k+1)} &= \cos(\Delta \mathbf{t}_i^p / 10000^{2k/H}), \end{aligned} \quad (5)$$

where k is the dimension and H is the model’s hidden size. After encoding the temporal information, we add the positional encodings to the embedding matrix obtaining the final input to the Transformer encoder:

$$\mathbf{E}^{(p,t)} = \mathbf{X}^{(p,t)} + \mathbf{PE}^{(p,t)} \in \mathbb{R}^{(t+1) \times H}. \quad (6)$$

We feed the input sequence through the encoder G and select the last hidden state of the [cls] token. The operation can be formulated as:

$$\hat{\mathbf{E}}^{(p,t)} = G(\mathbf{E}^{(p,t)}), \quad (7)$$

where $\hat{\mathbf{E}}^{(p,t)} \in \mathbb{R}^{(t+1) \times H}$ is the model’s final hidden state and $\hat{\mathbf{e}}_{\text{cls}} = \hat{\mathbf{E}}^{(p,0)} \in \mathbb{R}^H$ is the last hidden state of the classification token. To enhance the representation embedded in the [cls] token’s final hidden state, we feed it through a pooling layer ϕ consisting of a linear layer followed by a Tanh activation, a normalization layer, and the linear classification layer, before finally applying the sigmoid function to obtain the predicted probability of HCC development. The final model output is thus denoted by:

$$\hat{\mathbf{y}}^p = \sigma(\phi(\hat{\mathbf{e}}_{\text{cls}})), \quad (8)$$

where $\hat{\mathbf{y}}^p$ is the model’s predicted probability, ϕ is the model’s pooling layer, and σ is the sigmoid activation function.

3.3. Architectural Variants

In addition to our base model, called HCCNet-Pico, we introduce four variants of HCCNet, which are versions of about 0.5x, 0.75x, 2x, and 4x the base model’s size respectively. The five model variations are:

- HCCNet-F: $C = 48$, $B = \{2, 2, 6, 2\}$, $H = 384$, $A = 3$, $L = 4$
- HCCNet-C: $C = 56$, $B = \{2, 2, 6, 2\}$, $H = 448$, $A = 4$, $L = 4$
- HCCNet-P: $C = 64$, $B = \{2, 2, 6, 2\}$, $H = 512$, $A = 4$, $L = 4$
- HCCNet-N: $C = 80$, $B = \{2, 2, 8, 2\}$, $H = 640$, $A = 5$, $L = 6$
- HCCNet-T: $C = 96$, $B = \{3, 3, 9, 3\}$, $H = 768$, $A = 6$, $L = 6$

where C represents the number of channels in the ConvNeXt backbone, B are the number of ConvNeXt blocks, H is the Transformer encoder’s hidden dimension, L are the number of encoder layers, and A are the number of self-attention heads. The naming of the femto (F), pico (P), nano (N), and tiny (T) variants follows ConvNeXt’s original implementation in (Liu et al., 2022), comprising 12.4, 22.0, 45.9, and 72.4 million parameters respectively. Since ConvNeXt does not define a configuration between its femto and pico variants, we additionally introduce HCCNet-C (compact), a custom intermediate variant that scales only the channel and hidden dimensions of HCCNet-Pico to approximately 0.75x its size (16.5 million parameters). Deviating from the head dimension of 128 used by the ConvNeXt-derived variants, HCCNet-C retains $A = 4$ self-attention heads at $H = 448$ (i.e., a head dimension of 112), which we found to have no adverse effect on convergence.

3.4. Pre-Training

When training neural networks, task agnostic pre-training has been repeatedly shown to improve performance on downstream tasks. Additionally, subset $V^{(p,t)}$ contains only roughly 75% of observations included in superset $D^{(p,t)}$. To fully use all available data and improve the representations learned during training, we therefore design two (self-)supervised pre-training tasks for both the CNN backbone and the Transformer encoder prior to training on our downstream task.

CNN Feature Extractor. For the CNN backbone, this study extends the self-supervised DINO pre-training framework described in (Caron et al., 2021) to 3D MR images. Accordingly, we define a student F_s and a teacher network F_t , where the teacher is built as an exponential moving average of the student and we learn to match the output probability distributions of the networks (i.e., $p_{F_s}(i^{(p,t)})$ and $p_{F_t}(i^{(p,t)})$ respectively) by minimizing the cross-entropy loss w.r.t. the parameters of the student network θ_{F_s} :

$$\min_{\theta_{F_s}} - \sum_{i \in \{i_g^{(p,t)}\}} \sum_{i' \in \{i_g^{(p,t)}, i_l^{(p,t)}\}} p_{F_t}(i) \log(p_{F_s}(i')), \quad (9)$$

where $i_g^{(p,t)}$ and $i_l^{(p,t)}$ are the global and local views, which are cropped from the original image and used as input to the CNN backbone. Note that in our implementation, we use a 3D crop of shape $s = 72^3$ for the global and $s = 48^3$ for the local views. Besides that, our implementation also proposes a novel augmentation strategy specific to MR images. Instead of randomly adjusting the brightness, contrast, saturation and hue of an image as in the original DINO paper (Caron et al., 2021), we exploit the abundance of natural variation in MRI data. Specifically, we treat different series within a contrast-coherent subset of an MRI sequence as natural augmentations of the same image. For a view cropped from a Diffusion-Weighted MRI (DW-MRI), we randomly sample one series among the diffusion coefficients $b \in \{0, 150, 400, 800\}$; for dynamic contrast-enhanced T1-weighted MRIs (T1 DCE-MRI) we sample among the

pre- and post-contrast phases; and for the combined T1 in- and out-of-phase and T2-weighted modality (T1 IOP & T2-MRI) we sample within a single contrast family (the in- and out-of-phase series, or the short- and long-echo T2 series). This teaches the model to identify the same patient under varying acquisition conditions. More details on the study’s data augmentation strategy are provided in Appendix A.2.

Transformer Encoder. To improve the encoder’s understanding of disease progression, this study employs a simple binary sequence-order-prediction (SOP) task inspired by (Lan et al., 2020). Correspondingly, we randomly shuffle 50% of the input sequences² and task the model to differentiate shuffled from non-shuffled sequences. During the SOP pre-training, we keep the parameters of the CNN feature extractor fixed and initialize them using the previously pre-trained weights. We apply the sigmoid function to the output of the model’s pooling layer and learn to identify shuffled sequences by minimizing the binary cross-entropy loss w.r.t. the parameters $\theta_{G'} = \theta_G \cup \theta_\phi$ of the Transformer encoder G and the pooling layer ϕ :

$$\min_{\theta_{G'}} - \sum_{e \in \mathbf{E}^{(p,t)}} t_{G'}(e) \log(p_{G'}(e)), \quad (10)$$

where $p_{G'}(e)$ is the predicted probability of shuffling and $t_{G'}(e)$ is the true value. Note that during pre-training of the Transformer encoder we do not randomly crop multiple views from the same image but instead only crop a single view of shape $s = 72^3$ from the original MRI. Likewise, we concatenate the different series in the sequence of MRI scans to a multi-channel volume instead of randomly selecting a single one. Provided a sequence of DW-MRI’s with diffusion coefficients $b \in \{0, 150, 400, 800\}$, the input $i^{(p,t)}$ to the CNN backbone G is thus a four rather than a one channel volume. Since the backbone is pre-trained on single-channel views, we reconcile the differing input dimensionality by expanding the pre-trained single-channel stem to the required number of input channels: the learned stem filters are replicated across the input series to initialize the first convolutional layer, which is then fine-tuned end to end. This transfers the pre-trained representation to inputs comprising an arbitrary number of series without re-initializing the first convolutional layer from scratch.

3.5. Fine-Tuning

After pre-training, we initialize the parameters of HCCNet using the previously pre-trained weights. Additionally, we append the final pooling layer as discussed in 3.2 and fine-tune the model on our initial task by minimizing the class-balanced binary cross-entropy w.r.t. all parameters θ_{HCCNet} :

$$\min_{\theta_{\text{HCCNet}}} -\omega \sum_{e \in \mathbf{E}^{(p,t)}} \mathbf{y}^p \log(\hat{\mathbf{y}}^p). \quad (11)$$

²Roughly 25% of all sequences comprise only a single MRI screening. Therefore, we actually randomly shuffle 66% of sequences considering sequences with a single observation as negative instances.

where ω is the weighting factor for the positive class, $\hat{\mathbf{y}}_p$ is the predicted probability of HCC development and $\mathbf{y}_p$ is the observed development in patient p .

4. Experiments

4.1. Data Acquisition and Split

We evaluate HCCNet on data that is retrospectively collected from a cohort of liver cirrhosis patients who underwent repeated abdominal MRI screening for HCC surveillance between March 2011 and September 2022 at Medisch Spectrum Twente (MST) in Enschede, the Netherlands. Patients were followed up over 1 month to 11 years with screening intervals varying between 1 month and 6.5 years. In total, 963 MRI examinations were included in the study with longitudinal scans comprising contrast enhanced and non-contrast enhanced T1-, T2- and diffusion-weighted images which were acquired in axial orientation and registered using a Philips Ingenia Elition 1.5- and 3-Tesla scanner respectively. Prior to pre-training and fine-tuning, MR images were anonymized and resampled to 1.5x1.5x1.5mm voxel spacing. All examinations included in the study comprise the complete set of imaging sequences, such that the diffusion-weighted, the dynamic contrast-enhanced T1-weighted, and the combined T1 in- and out-of-phase and T2-weighted modality groups are available for every patient. This complete co-availability allows us to additionally evaluate multi-modal configurations that concatenate series across modality groups (Section 4.4) without discarding patients for whom a modality is missing.

As described in section 3.1, we include all available data in superset D_p during pre-training and subset the data for fine-tuning. During fine-tuning we additionally exclude all patients without a definitive diagnosis of HCC yielding a total of 243 patients of which 37 develop HCC (i.e., 703 and 101 total MRI examinations respectively). To obtain robust performance estimates despite the limited number of positive cases, we evaluate HCCNet using stratified 5-fold cross-validation rather than a single hold-out split. We partition the patient cohort into five mutually exclusive folds at the patient level, such that all examinations of a given patient are assigned to the same fold and no patient appears in more than one fold. The folds are stratified by HCC outcome so that the 37 HCC-positive patients are distributed approximately evenly across folds (i.e., 7–8 positive patients per fold). Each fold serves as the held-out test set D_p^{test} exactly once, while the remaining four folds form the development set D_p^{dev} that is used for both pre-training and fine-tuning. The full modeling pipeline (self-supervised pre-training of the CNN backbone and Transformer encoder, fine-tuning, and baseline training) is repeated independently for each fold using only the data of its development set. Because pre-training is re-run within every fold, the held-out patients are never observed at any training stage, which precludes information leakage between the test fold and either pre-training or fine-tuning. As before, to avoid leakage between pre-training and fine-tuning within a fold, we only fine-tune HCCNet on longitudinal MRIs that are also included in the fold’s development superset D_p^{dev} . All results are reported as the mean and standard deviation across the five folds.

4.2. Implementation Details and Baseline

Pre-Training. As aforementioned, we pre-train both the CNN backbone and the Transformer encoder on development superset D_p^{dev} . We train the CNN over 32000 training steps using the AdamW optimizer (Loshchilov and Hutter, 2019) with an effective batch size of $b = 128$. During the first 5% of training steps (i.e., 1600 warm-up steps), we linearly increase the learning rate η to its base value which is determined using the following square-root scaling rule (Granzio et al.): $\eta = \alpha \frac{\sqrt{b}}{\sqrt{128}}$, where b is the effective batch size and α is a constant scaling factor that is chosen according to b . After reaching its maximum value, we decay the learning rate using a cosine schedule. We employ the same cosine decay without warm-up for the weight decay λ but only apply it to the weights of the parameters excluding normalization layers (Li et al., 2020). Contrary to the original DINO implementation, we also limit the temperature of the teacher network to a constant $\tau_{F_t} = 0.04$ and increase the initial exponential moving average momentum to $m = 0.9995$.

While MRI scans are assumed to be i.i.d. during CNN pre-training, pre-training of the Transformer encoder considers the within-person dependencies across time points. Hence, to mitigate the risk of overfitting due to the reduced number of independent samples, we train the Transformer encoder using an abbreviated pre-training protocol of just 8000 training steps with a learning rate warm-up of 5% or 10% for the smaller and bigger model variants respectively and a reduced effective batch size of $b = 32$. We set the maximal sequence length to $s = 8$, apply dropout with probability $p = 0.2$, and employ the same optimizer, learning rate scaling and cosine schedule as during CNN pre-training.

Fine-Tuning. During fine-tuning, we update HCCNet on the development subset V_p^{dev} . We fine-tune the model over 400 training steps with 5% warm-up using the same optimizer, sequence length, and learning rate schedule as during pre-training of the Transformer encoder but decrease the weight decay to a constant $\lambda = 1e^{-5}$. For the individual MRI sequences, we select the batch size from $b \in \{16, 32, 64\}$ using cross-validation restricted to the development folds and apply no further hyperparameter tuning. Within each of the five outer cross-validation folds, we then train HCCNet on the fold’s development subset V_p^{dev} using the same default hyper-parameter settings as during Transformer encoder pre-training and evaluate it on the corresponding held-out fold.

Baseline. To contrast the proposed modeling approach, we establish a baseline by training HCCNet from scratch using randomly initialized parameters sampled from $\mathcal{N}(0, 0.02)$. To provide an unbiased comparison, we equally train the baseline within each of the five cross-validation folds on the corresponding development subset $V_{\text{dev}}^{(p,t)}$. We employ the same optimizer, learning rate schedule, number of training steps, batch size, and further hyper-parameter settings as during fine-tuning but slightly increase the learning rate warm-up (i.e., from 20 to 40 training steps) to ensure a more stable convergence. A full list of hyperparameter setting is presented in table A.5 in Appendix A.3.

Multi-Modal Fusion. To explore whether combining imaging modalities improves upon the best single modality, we additionally evaluate a set of ex-

ploratory multi-modal configurations. Following the same channel-concatenation strategy already used to combine the series within a single modality, we naively concatenate the series of multiple modality groups into a single multi-channel input volume. Concretely, we evaluate the full fusion of all three modality groups (DW-MRI, T1 DCE-MRI, and T1 IOP & T2-MRI; twelve input channels) as well as the two pairings that include the strongest single modality, namely DW-MRI + T1 DCE-MRI and DW-MRI + T1 IOP & T2-MRI (eight input channels each). As during single-modality pre-training, augmentation views are sampled within a single contrast family. Owing to the limited number of positive cases and our finding that model capacity only negligibly affects downstream performance, we restrict the fusion experiments to the HCCNet-Pico variant.

Pre-Training Ablation. To localize the source of the pre-training effect, we further disentangle the contributions of the two pre-training stages. In addition to the randomly initialized baseline (no pre-training) and the fully pre-trained model (both stages), we evaluate two intermediate configurations: one in which only the CNN backbone is pre-trained using DINO while the Transformer encoder is randomly initialized, and one in which only the Transformer encoder is pre-trained using the sequence-order-prediction task while the CNN backbone is randomly initialized. We conduct this ablation on the HCCNet-Pico variant across all three modality groups using the same five-fold protocol.

4.3. Evaluation Protocol

After fine-tuning, we evaluate HCCNet’s baseline and the fine-tuned predictive capability of HCC on the test set $V_{\text{test}}^{(p,t)}$ on a central crop. Within each of the five cross-validation folds, we train ten independently seeded models (fifty models in total) and average their predicted probabilities to form that fold’s ensemble prediction. Every reported metric is computed on this per-fold ensemble prediction, and we report its mean and standard deviation across the five cross-validation folds. For evaluation, we employ the following performance metrics: accuracy, precision, recall, F1-score, area under the receiver operating characteristic (AUROC), and area under the precision-recall curve (AUPRC). For the comparison across model variations, we also explore the robustness of the model by comparing the mean absolute error (MAE) formulated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\text{gain}'_{\text{obs},i} - \text{gain}_{\text{exp},i}|, \quad (12)$$

where i represents the individual cross-validation fold, ranging from i to n , $\text{gain}^{\text{exp},i}$ denotes the expected cumulative gain equal to the identity line, and $\text{gain}'_{\text{obs},i}$ represents the normalized observed cumulative gain, which is computed over $\text{gain}'_{\text{obs},i} = \frac{\sum_{k=1}^i m_k}{i}$, for $i = 1, 2, \dots, n$, where m_k denotes the metric over which the cumulative gain is computed. Likewise, for the comparison against the randomly initialized baseline, we evaluate the models’ respective calibration considering the expected (ECE), and maximum calibration error (MCE) as well as the Brier score.

Table 1: Fine-Tuned Model Results for the Prediction of HCC at the Next Observation

		Performance Metrics					
		Accuracy	Precision	Recall	F1-Score	AUPRC	AUROC
DW-MRI	HCCNet-F	0.852 $\pm$ 0.049	0.491 $\pm$ 0.110	0.700 $\pm$ 0.100	0.573 $\pm$ 0.104	0.631 $\pm$ 0.070	0.885 $\pm$ 0.026
	HCCNet-C	0.828 $\pm$ 0.019	0.425 $\pm$ 0.043	0.750 $\pm$ 0.194	0.535 $\pm$ 0.088	0.619 $\pm$ 0.028	0.907 $\pm$ 0.013
	HCCNet-P	0.838 $\pm$ 0.014	0.443 $\pm$ 0.032	0.700 $\pm$ 0.100	0.542 $\pm$ 0.050	0.588 $\pm$ 0.068	0.894 $\pm$ 0.024
	HCCNet-N	0.855 $\pm$ 0.026	0.467 $\pm$ 0.125	0.375 $\pm$ 0.112	0.414 $\pm$ 0.114	0.507 $\pm$ 0.027	0.882 $\pm$ 0.027
T1 DCE-MRI	HCCNet-T	0.897 $\pm$ 0.019	0.591 $\pm$ 0.050	0.850 $\pm$ 0.050	0.695 $\pm$ 0.039	0.725 $\pm$ 0.103	0.936 $\pm$ 0.026
	HCCNet-F	0.814 $\pm$ 0.059	0.436 $\pm$ 0.087	0.775 $\pm$ 0.050	0.550 $\pm$ 0.066	0.543 $\pm$ 0.120	0.792 $\pm$ 0.022
	HCCNet-C	0.730 $\pm$ 0.087	0.322 $\pm$ 0.085	0.675 $\pm$ 0.061	0.426 $\pm$ 0.067	0.329 $\pm$ 0.089	0.734 $\pm$ 0.040
	HCCNet-P	0.828 $\pm$ 0.048	0.456 $\pm$ 0.079	0.750 $\pm$ 0.079	0.558 $\pm$ 0.047	0.575 $\pm$ 0.074	0.857 $\pm$ 0.015
T1 IOP & T2-MRI	HCCNet-N	0.814 $\pm$ 0.036	0.422 $\pm$ 0.055	0.775 $\pm$ 0.050	0.543 $\pm$ 0.042	0.628 $\pm$ 0.068	0.819 $\pm$ 0.021
	HCCNet-T	<u>0.867 $\pm$ 0.018</u>	<u>0.520 $\pm$ 0.049</u>	0.700 $\pm$ 0.061	<u>0.596 $\pm$ 0.051</u>	0.527 $\pm$ 0.071	0.823 $\pm$ 0.029
	HCCNet-F	0.772 $\pm$ 0.043	0.342 $\pm$ 0.048	0.625 $\pm$ 0.112	0.436 $\pm$ 0.043	0.428 $\pm$ 0.046	0.790 $\pm$ 0.059
	HCCNet-C	0.789 $\pm$ 0.037	0.324 $\pm$ 0.082	0.450 $\pm$ 0.127	0.375 $\pm$ 0.094	0.373 $\pm$ 0.074	0.729 $\pm$ 0.076
T1 IOP-MRI	HCCNet-P	0.779 $\pm$ 0.041	0.325 $\pm$ 0.070	0.500 $\pm$ 0.112	0.389 $\pm$ 0.076	0.420 $\pm$ 0.135	0.736 $\pm$ 0.065
	HCCNet-N	0.821 $\pm$ 0.020	0.383 $\pm$ 0.067	0.475 $\pm$ 0.122	0.423 $\pm$ 0.087	0.519 $\pm$ 0.090	0.778 $\pm$ 0.066
	HCCNet-T	0.832 $\pm$ 0.036	0.411 $\pm$ 0.114	0.400 $\pm$ 0.094	0.403 $\pm$ 0.100	0.407 $\pm$ 0.061	0.761 $\pm$ 0.084
	HCCNet-F	0.804 $\pm$ 0.039	0.360 $\pm$ 0.104	0.475 $\pm$ 0.094	0.408 $\pm$ 0.100	0.401 $\pm$ 0.059	0.816 $\pm$ 0.025
T2-MRI	HCCNet-C	0.814 $\pm$ 0.018	0.325 $\pm$ 0.069	0.325 $\pm$ 0.127	0.318 $\pm$ 0.096	0.315 $\pm$ 0.039	0.745 $\pm$ 0.050
	HCCNet-P	0.832 $\pm$ 0.024	0.417 $\pm$ 0.075	0.400 $\pm$ 0.122	0.394 $\pm$ 0.062	0.406 $\pm$ 0.084	0.703 $\pm$ 0.043
	HCCNet-N	0.796 $\pm$ 0.026	0.318 $\pm$ 0.068	0.375 $\pm$ 0.079	0.341 $\pm$ 0.062	0.328 $\pm$ 0.035	0.756 $\pm$ 0.025
	HCCNet-T	0.807 $\pm$ 0.033	0.367 $\pm$ 0.048	0.450 $\pm$ 0.127	0.389 $\pm$ 0.050	0.494 $\pm$ 0.067	0.747 $\pm$ 0.035
	HCCNet-F	0.817 $\pm$ 0.023	0.412 $\pm$ 0.040	0.725 $\pm$ 0.094	0.522 $\pm$ 0.038	0.430 $\pm$ 0.018	0.850 $\pm$ 0.023
	HCCNet-C	0.745 $\pm$ 0.025	0.307 $\pm$ 0.034	0.675 $\pm$ 0.100	0.421 $\pm$ 0.047	0.351 $\pm$ 0.073	0.778 $\pm$ 0.029
	HCCNet-P	0.834 $\pm$ 0.034	0.454 $\pm$ 0.077	0.775 $\pm$ 0.094	0.567 $\pm$ 0.060	0.419 $\pm$ 0.045	0.873 $\pm$ 0.014
	HCCNet-N	0.824 $\pm$ 0.020	0.430 $\pm$ 0.026	<u>0.800 $\pm$ 0.170</u>	0.551 $\pm$ 0.048	0.573 $\pm$ 0.070	0.900 $\pm$ 0.010
	HCCNet-T	0.800 $\pm$ 0.021	0.312 $\pm$ 0.060	0.375 $\pm$ 0.079	0.340 $\pm$ 0.068	0.365 $\pm$ 0.036	0.783 $\pm$ 0.015

Note: Performance metrics range from 0 to 1 or from 0.50 to 1 and from 0.14 to 1 for the AUROC and the AUPRC respectively. All metrics are presented using the mean and standard deviation across the five cross-validation folds. The best model results are emphasised in boldface. Second best results are underlined. DW-MRI comprises diffusion MRI scans with diffusion coefficient $b \in \{0, 150, 400, 800\}$. T1 DCE-MRI contain a pre-contrast and three post-contrast (i.e., late arterial, portal venous, and delayed phase) series. T1 IOP & T2-MRI comprise T1-weighted in- and out-of-phase MRI as well as T2-weighted MRI with short and long echo times. T1 IOP-MRI and T2-MRI report the same combined modality split into its constituent contrasts (i.e., the in- and out-of-phase series, and the short- and long-echo T2 series, respectively).

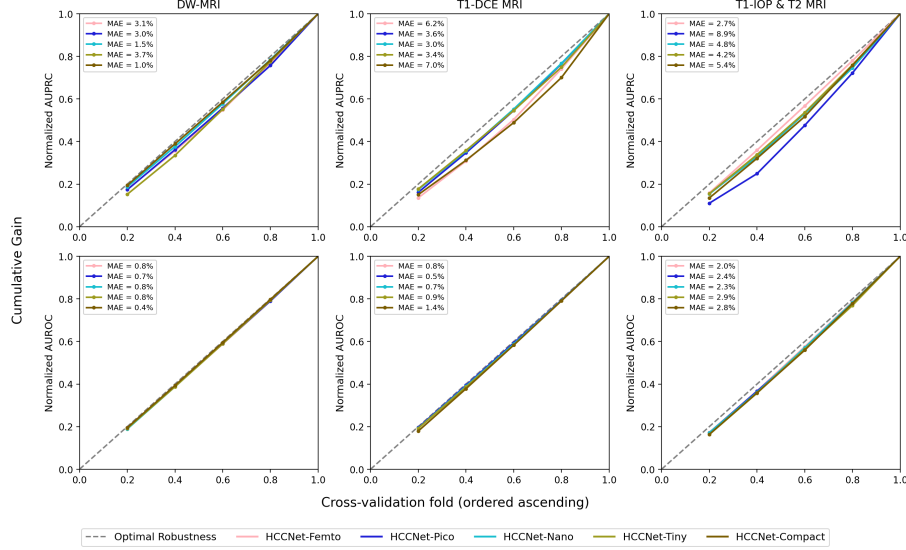
Since all metrics are obtained through cross-validation, we note that the five fold-level estimates are not fully independent: consecutive folds share the majority of their development data, so the reported standard deviations should be interpreted as a measure of fold-to-fold variability rather than as an unbiased estimate of the true standard error, which they are likely to slightly understate. To avoid over-interpreting the limited number of positive cases, we therefore refrain from drawing conclusions from differences that fall within one standard deviation across folds and treat such cases as performance ties.

4.4. Main Results

We report HCCNet’s fine-tuned performance (i) across imaging modalities, (ii) model variations, and (iii) against a randomly initialized baseline. Additional model results are presented in Appendix B.

Comparison across Imaging Modalities. Table 1 presents fine-tuned model results for the prediction of HCC development at the next observation stratified by imaging modality and model architecture. HCCNet trained using diffusion weighted MR images shows excellent predictive capability of tumor development with an average AUPRC = 0.614 and AUROC = 0.901 across model architectures. Contrast enhanced T1 weighted and the combination of

Figure 2: Cumulative Gain Across Cross-Validation Folds



Note: Normalized AUROC and AUPRC values are presented across the five cross-validation folds of different HCCNet model variants (femto, pico, nano, tiny, compact) across three MRI types (DW-MRI, T1-DCE MRI, T1-IOP & T2 MRI). Each colored line represents the cumulative gain in normalized performance. Optimal robustness is presented by the gray dashed line, indicating no sensitivity to the data partitioning. The Mean Absolute Error (MAE) quantifies the deviation of each model’s performance from this optimal line averaged over the five cross-validation folds. Generally, lower MAE values indicate greater robustness to fold-to-fold variation, while higher MAE values signify greater sensitivity to the data partitioning.

T1 weighted in- and out-of-phase and T2 weighted images, on the other hand, appear to display less expressive characteristics for the prediction of HCC development, albeit models trained using contrast enhanced T1-weighted images still maintain a stronger average predictive capability than the combined T1 IOP and T2-weighted MRIs (i.e., AUPRC = 0.520 and AUROC = 0.805 versus AUPRC = 0.429 and AUROC = 0.759). The two contrasts that make up this combined group are also evaluated in their own right as standalone single-modality models, reported as the T1 IOP-MRI and T2-MRI modalities in Table 1. At HCCNet-Pico, the standalone T2-MRI model (AUPRC = 0.419 and AUROC = 0.873) is markedly stronger on AUROC than the standalone T1 IOP-MRI model (AUPRC = 0.406 and AUROC = 0.703), while on AUPRC the two are comparable (around 0.42). The combined T1 IOP & T2-MRI modality reaches AUPRC = 0.420 and AUROC = 0.736, so the T2-MRI model alone exceeds it on AUROC (0.873 versus 0.736), indicating that concatenating the weaker T1 IOP-MRI with T2-MRI dilutes the T2 signal. At the group level, the combined modality nonetheless remains the weakest of the three groups.

Comparison across Model Variations. In addition to the varying image

Table 2: Baseline vs. Fine-Tuned Model Results

		AUPRC			AUROC		
		Baseline	Fine-Tuned	Rel. Change	Baseline	Fine-Tuned	Rel. Change
DW-MRI	HCCNet-F	0.292 $\pm$ 0.040	<u>0.631 $\pm$ 0.070</u>	115.8%	0.716 $\pm$ 0.024	0.885 $\pm$ 0.026	23.6%
	HCCNet-C	0.293 $\pm$ 0.033	0.619 $\pm$ 0.028	111.2%	0.705 $\pm$ 0.017	<u>0.907 $\pm$ 0.013</u>	28.7%
	HCCNet-P	0.308 $\pm$ 0.032	0.588 $\pm$ 0.068	90.9%	0.722 $\pm$ 0.015	0.894 $\pm$ 0.024	23.8%
	HCCNet-N	0.301 $\pm$ 0.037	0.507 $\pm$ 0.027	68.2%	0.724 $\pm$ 0.024	0.882 $\pm$ 0.027	21.9%
T1 DCE-MRI	HCCNet-T	0.294 $\pm$ 0.029	0.725 $\pm$ 0.103	147.1%	0.722 $\pm$ 0.016	0.936 $\pm$ 0.026	29.7%
	HCCNet-F	0.348 $\pm$ 0.056	0.543 $\pm$ 0.120	55.9%	<u>0.723 $\pm$ 0.027</u>	0.792 $\pm$ 0.022	9.5%
	HCCNet-C	0.312 $\pm$ 0.015	0.329 $\pm$ 0.089	5.6%	0.693 $\pm$ 0.043	0.734 $\pm$ 0.040	5.9%
	HCCNet-P	0.350 $\pm$ 0.053	0.575 $\pm$ 0.074	64.3%	0.703 $\pm$ 0.049	0.857 $\pm$ 0.015	21.9%
T1 IOP & T2-MRI	HCCNet-N	0.375 $\pm$ 0.050	0.628 $\pm$ 0.068	67.2%	0.692 $\pm$ 0.054	0.819 $\pm$ 0.021	18.4%
	HCCNet-T	0.348 $\pm$ 0.059	0.527 $\pm$ 0.071	51.4%	0.692 $\pm$ 0.050	0.823 $\pm$ 0.029	18.9%
	HCCNet-F	0.338 $\pm$ 0.051	0.428 $\pm$ 0.046	26.5%	0.639 $\pm$ 0.024	0.790 $\pm$ 0.059	23.6%
	HCCNet-C	0.313 $\pm$ 0.051	0.373 $\pm$ 0.074	19.0%	0.628 $\pm$ 0.009	0.729 $\pm$ 0.076	16.2%
T1 IOP-MRI	HCCNet-P	0.307 $\pm$ 0.061	0.420 $\pm$ 0.135	36.8%	0.624 $\pm$ 0.024	0.736 $\pm$ 0.065	17.9%
	HCCNet-N	0.251 $\pm$ 0.034	0.519 $\pm$ 0.090	107.2%	0.618 $\pm$ 0.037	0.778 $\pm$ 0.066	25.7%
	HCCNet-T	0.236 $\pm$ 0.030	0.407 $\pm$ 0.061	72.2%	0.592 $\pm$ 0.039	0.761 $\pm$ 0.084	28.5%
	HCCNet-F	0.228 $\pm$ 0.024	0.401 $\pm$ 0.059	75.5%	0.640 $\pm$ 0.051	0.816 $\pm$ 0.025	27.4%
T2-MRI	HCCNet-C	0.247 $\pm$ 0.017	0.315 $\pm$ 0.039	27.4%	0.670 $\pm$ 0.044	0.745 $\pm$ 0.050	11.2%
	HCCNet-P	0.245 $\pm$ 0.022	0.406 $\pm$ 0.084	65.8%	0.669 $\pm$ 0.035	0.703 $\pm$ 0.043	5.0%
	HCCNet-N	0.237 $\pm$ 0.027	0.328 $\pm$ 0.035	38.6%	0.659 $\pm$ 0.048	0.756 $\pm$ 0.025	14.7%
	HCCNet-T	0.234 $\pm$ 0.026	0.494 $\pm$ 0.067	110.9%	0.643 $\pm$ 0.056	0.747 $\pm$ 0.035	16.2%
T2-MRI	HCCNet-F	0.243 $\pm$ 0.065	0.430 $\pm$ 0.018	76.6%	0.618 $\pm$ 0.053	0.850 $\pm$ 0.023	37.6%
	HCCNet-C	0.265 $\pm$ 0.092	0.351 $\pm$ 0.073	32.2%	0.620 $\pm$ 0.038	0.778 $\pm$ 0.029	25.7%
	HCCNet-P	0.257 $\pm$ 0.073	0.419 $\pm$ 0.045	63.4%	0.639 $\pm$ 0.030	0.873 $\pm$ 0.014	36.6%
	HCCNet-N	0.258 $\pm$ 0.077	0.573 $\pm$ 0.070	<u>122.0%</u>	0.619 $\pm$ 0.033	0.900 $\pm$ 0.010	45.5%
T2-MRI	HCCNet-T	0.305 $\pm$ 0.118	0.365 $\pm$ 0.036	19.7%	0.618 $\pm$ 0.038	0.783 $\pm$ 0.015	26.7%

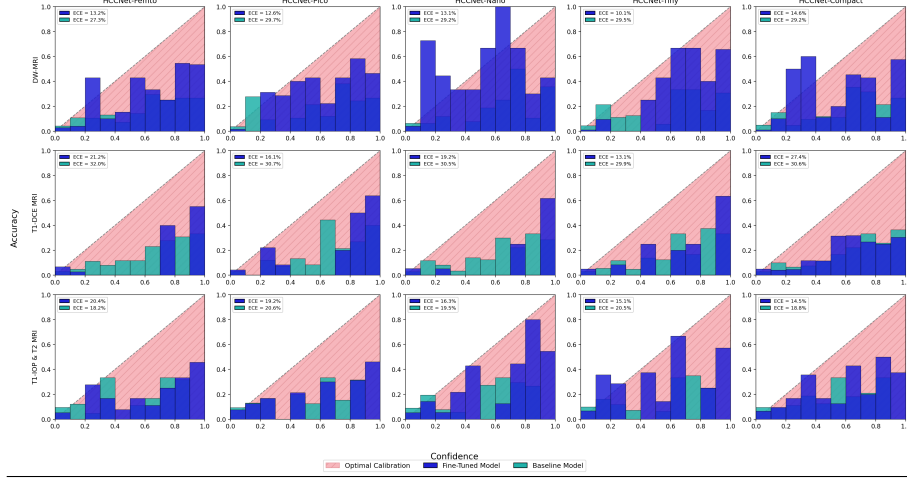
Note: AUPRC and the AUROC values range from 0.14 and 0.5 to 1 respectively. All metrics are presented using the mean and standard deviation across the five cross-validation folds. The best results are emphasised in boldface. Second best results are underlined. DW-MRI comprises diffusion MRI scans with diffusion coefficient $b \in \{0, 150, 400, 800\}$. T1 DCE-MRI comprises a pre-contrast and three post-contrast (i.e., late arterial, portal venous, and delayed phase) series, while T1 IOP & T2-MRI comprise T1-weighted in- and out-of-phase MRI as well as T2-weighted MRI with short and long echo times. T1 IOP-MRI and T2-MRI report the same combined modality split into its constituent contrasts.

modalities, we explore the impact of varying model complexity on HCC risk prediction. As such, Figure 2 plots the cumulative gain in normalized AUPRC and AUROC against the test results obtained over the five cross-validation folds and ordered by ascending cumulative gain. Across model variants and imaging modalities, the per-fold ensemble closely follows the diagonal line, indicating low sensitivity to the data partitioning. This robustness is tightest on AUROC and slightly looser on AUPRC, with the largest, though still modest, fold-to-fold spread observed on the combined T1 in- and out-of-phase and T2-weighted modality. No model variant falls consistently below the diagonal line.

Across imaging modalities, the models are thus robust and not overly sensitive to variation across folds. Likewise, the plots reveal no consistent differences across model architecture, as no model variant is persistently deviating from the optimal diagonal line. This corresponds with the results presented in Table 1, which indicate that compared to the differences between imaging modalities model size only negligibly affects downstream model performance. At least on the data evaluated in this study, variations in model complexity do thus not considerably influence the quality of the models’ predictions nor their robustness.

Comparison against Baseline. Finally, we evaluate the impact of our

Figure 3: Baseline vs. Fine-Tuned Models’ Reliability Diagrams



Note: Calibration plots of HCCNet model variants (femto, pico, nano, tiny, compact) are shown across three MRI modalities (DW-MRI, T1-DCE MRI, T1-IOP & T2 MRI) to evaluate the alignment between predicted confidence and actual accuracy. The x-axis represents confidence levels (0.0 to 1.0), while the y-axis shows accuracy within each confidence bin. Optimal calibration, where predicted confidence matches actual accuracy, is depicted by the pink-striped region. Blue bars show the calibration of fine-tuned models, and green bars represent baseline models. Expected Calibration Error (ECE), shown in the legend as a percentage, quantifies the average mismatch between confidence and accuracy across bins; lower ECE indicates better calibration.

proposed pre-training approach on model performance. Table 2 presents baseline and fine-tuned model results stratified by image modality and model variation. Likewise, Table B.6 in Appendix B.1 presents extended baseline model results. Across imaging modalities, the magnitude of the improvement from our proposed approach varies. That is, although baseline model results vary only gradually across image modalities, diffusion weighted MR images display the largest average improvement of 106.6% in AUPRC and 25.5% in AUROC after fine-tuning, while dynamic contrast enhanced T1-weighted images show smaller but still substantial gains (i.e., 48.9% in AUPRC and 14.9% in AUROC), and the combination of T1-weighted in- and out-of-phase and T2-weighted MRIs displays an average improvement of 52.3% in AUPRC and 22.4% in AUROC after fine-tuning.

To further investigate the apparent disparity in pre-training success, we compare the reliability of the baseline models’ predictions to the reliability of the fine-tuned models’ predictions. Accordingly, Figure 3 presents reliability diagrams across imaging modalities and model variants. Extended reliability metrics are presented in Table B.7 in Appendix B.2. For the fine-tuned models, the plots largely align with the results presented in Table 2. Figure 3 shows consistently lower expected calibration errors among models fine-tuned using dif-

Table 3: Exploratory Multi-Modal Fusion Results for the Prediction of HCC at the Next Observation (HCCNet-Pico)

Configuration	Performance Metrics					
	Accuracy	Precision	Recall	F1-Score	AUPRC	AUROC
DW-MRI (single modality)	0.838 ± 0.014	0.443 ± 0.032	0.700 ± 0.100	0.542 ± 0.050	0.588 ± 0.068	0.894 ± 0.024
DW-MRI + T1 DCE-MRI	0.793 ± 0.032	0.378 ± 0.035	0.725 ± 0.200	0.487 ± 0.071	0.535 ± 0.102	0.873 ± 0.026
DW-MRI + T1 IOP & T2-MRI	0.842 ± 0.031	0.470 ± 0.060	0.750 ± 0.000	0.575 ± 0.047	0.587 ± 0.048	0.891 ± 0.028
Full fusion (all modalities)	0.842 ± 0.065	0.489 ± 0.118	0.650 ± 0.215	0.532 ± 0.144	0.541 ± 0.097	0.894 ± 0.033

Note: All configurations use the HCCNet-Pico variant and are obtained by naively concatenating the series of the listed modality groups into a single multi-channel input volume (full fusion corresponds to twelve input channels). The within-contrast augmentation restriction is retained during pre-training of all fused models. Performance metrics range from 0 to 1 or from 0.50 to 1 and from 0.14 to 1 for the AUROC and the AUPRC respectively. All metrics are presented using the mean and standard deviation across the five cross-validation folds. The best results are emphasised in boldface and second best results are underlined. All differences between configurations fall within one standard deviation across folds.

fusion weighted MR images and the highest expected calibration errors among models fine-tuned on dynamic contrast enhanced T1-weighted MRIs, with the combined T1 in- and out-of-phase and T2-weighted MRIs falling in between. For the randomly initialized baseline, however, the expected calibration error does not display the same negative relationship with model performance. Here, baseline models trained using T1-IOP and T2-weighted MR images consistently show the lowest expected calibration error, while models trained using dynamic contrast enhanced T1-weighted MRIs display the highest expected calibration error.

More interestingly, the comparison between baseline and fine-tuned models' reliability diagrams indicates that the proposed self-supervised pre-training reduces the expected calibration error across all three modality groups relative to the models trained with randomly initialized parameters. This reduction is largest for diffusion weighted MR images at 56.0%, reaches 37.0% for dynamic contrast enhanced T1-weighted MRIs, and is smallest at 12.1% for the already well-calibrated combined T1 in- and out-of-phase and T2-weighted modality. Improved calibration makes the models predicted scores of tumor development more reliable and more dependable for probabilistic interpretation.

Multi-Modal Fusion. As an exploratory analysis, we evaluate whether naively fusing imaging modalities improves upon the strongest single modality. Table 3 reports the fine-tuned HCCNet-Pico performance for the three fused configurations alongside the best single modality (DW-MRI), which reaches $\text{AUPRC} = 0.588$ and $\text{AUROC} = 0.894$. None of the fused configurations exceeds this best single modality. Adding T1 DCE-MRI to DW-MRI lowers performance ($\text{AUPRC} = 0.535$ and $\text{AUROC} = 0.873$), fusing the weaker T1 IOP & T2-MRI modality with DW-MRI leaves performance essentially equal to DW-MRI alone ($\text{AUPRC} = 0.587$ and $\text{AUROC} = 0.891$), and the full fusion of all three modality groups matches DW-MRI alone on AUROC (0.894) while landing slightly below on AUPRC (0.541). As all differences fall within one standard deviation across folds, we interpret these results as indicating that naive channel-concatenation fusion does not meaningfully improve over the best single modality on the present cohort.

Table 4: Pre-Training Stage Ablation for the Prediction of HCC at the Next Observation (HCCNet-Pico)

	Pre-Trained Stage	AUPRC	AUROC
DW-MRI	None (baseline)	0.308 ± 0.032	0.722 ± 0.015
	CNN backbone only	0.672 ± 0.069	0.926 ± 0.013
	Transformer encoder only	0.310 ± 0.039	0.745 ± 0.029
	Both (full pre-training)	<u>0.588 ± 0.068</u>	<u>0.894 ± 0.024</u>
T1 DCE-MRI	None (baseline)	0.350 ± 0.053	0.703 ± 0.049
	CNN backbone only	<u>0.524 ± 0.057</u>	<u>0.773 ± 0.044</u>
	Transformer encoder only	0.361 ± 0.037	0.755 ± 0.029
	Both (full pre-training)	0.575 ± 0.074	0.857 ± 0.015
T1 IOP & T2-MRI	None (baseline)	0.307 ± 0.061	0.624 ± 0.024
	CNN backbone only	0.447 ± 0.088	0.687 ± 0.040
	Transformer encoder only	0.457 ± 0.023	<u>0.735 ± 0.024</u>
	Both (full pre-training)	0.420 ± 0.135	0.736 ± 0.065

Note: All configurations use the HCCNet-Pico variant. “None” corresponds to the randomly initialized baseline, while “Both” corresponds to the fully pre-trained model. “CNN backbone only” pre-trains the CNN feature extractor using DINO while leaving the Transformer encoder randomly initialized, and “Transformer encoder only” pre-trains the encoder using the sequence-order-prediction task while leaving the CNN backbone randomly initialized. AUPRC and AUROC values range from 0.14 and 0.50 to 1 respectively. All metrics are presented using the mean and standard deviation across the five cross-validation folds, with the best result per modality emphasised in boldface and the second best underlined.

Pre-Training Ablation. Finally, to localize the source of the pre-training effect, Table 4 disentangles the contribution of the two pre-training stages for HCCNet-Pico. For diffusion-weighted imaging, pre-training only the CNN backbone recovers essentially all of the full pre-training gain (AUPRC = 0.672 versus 0.588 for both stages and 0.308 for the randomly initialized baseline), with the difference between the two configurations falling within one standard deviation across folds, whereas pre-training only the Transformer encoder yields AUPRC = 0.310, barely above the baseline. The CNN-backbone dominance also holds for dynamic contrast enhanced T1-weighted images, where the CNN backbone alone (AUPRC = 0.524) clearly exceeds the Transformer encoder alone (AUPRC = 0.361). For the combined T1 in- and out-of-phase and T2-weighted modality, however, the two stages contribute comparably (Transformer encoder AUPRC = 0.457 versus CNN backbone AUPRC = 0.447). These results indicate that the self-supervised pre-training of the CNN backbone is the dominant driver of the observed pre-training effects for diffusion-weighted and dynamic contrast enhanced T1-weighted imaging.

5. Discussion

Despite being long established as the gold standard in many branches of oncology, deep learning applications to the study of longitudinal medical images remain notoriously understudied. Therefore, this study developed HCCNet, a spatio-temporal neural network that uses a 3D ConvNeXt backbone combined with a Transformer encoder, to predict future cancer development based on past MRI examinations. To support HCCNet’s capability to capture disease progression over time, a step-wise model training approach was adopted. We first pre-trained HCCNet’s CNN backbone and Transformer encoder using a custom self-supervised pre-training framework tailored for longitudinal medical images and then fine-tuned the model on our downstream task, initializing HCCNet’s parameters with the pre-trained weights.

Overall, fine-tuned model results varied substantially across imaging modalities. Among them, diffusion weighted MR images (DW-MRIs) displayed the highest predictive performance, achieving an average AUPRC of 0.614 and AUROC of 0.901. In comparison, contrast-enhanced T1-weighted MRIs (T1 DCE-MRIs) showed moderate predictive power, while combined T1 in- and out-of-phase and T2-weighted MRIs exhibited the lowest accuracy at the group level, even though their T2-weighted constituent was individually strong. Our proposed pre-training approach improved model performance by up to 147.1% for models fine-tuned using diffusion-weighted MR images and yielded substantial improvements as well for models fine-tuned on the dynamic contrast-enhanced T1-weighted and the combined T1 in- and out-of-phase and T2-weighted modalities.

Inspection of the models’ reliability diagrams further demonstrated that the proposed self-supervised pre-training approach aligned the models closer to the target distribution, reducing the extent to which calibration errors are propagated and amplified during fine-tuning. Thus, even for imaging modalities where

the pre-training approach only yielded moderate improvements, the predicted risks were more reliable, demonstrating the value of our proposed pre-training framework for applications in the medical domain.

The magnitude of the pre-training benefit varied across imaging modalities. Our augmentation strategy exposes the backbone to the natural variation among the series that share a common contrast, and this variation is richest for diffusion-weighted imaging, for which pre-training yielded the largest gains. An ablation that pre-trains only the CNN backbone or only the Transformer encoder further indicates that the self-supervised pre-training of the CNN backbone is the dominant driver of the pre-training benefit, while the sequence-order-prediction task contributes comparatively little. Finally, our exploratory multi-modal experiments show that naively fusing imaging modalities by channel concatenation provides at most marginal improvements over the strongest single modality. We attribute this to the limited number of positive cases and the additional registration burden incurred when concatenating series acquired across separate sequences, and view structured fusion strategies that explicitly account for the distinct characteristics of each modality as a promising avenue for future work.

Despite these successes, several challenges and limitations were identified. The variability in the magnitude of the pre-training benefit across imaging modalities suggests that further refinement of the pre-training process could improve consistency across imaging types. Additionally, while HCCNet demonstrates strong predictive capabilities for HCC development, its application to other diseases will require careful consideration of the specific characteristics and progression patterns of those diseases.

Another important point is the potential for integrating additional clinical data, such as electronic health records (EHR) or molecular biomarkers, into the HCCNet framework. Incorporating these data sources could further enhance the model’s predictive power and provide a fuller understanding of disease progression. In addition, the study’s focus on a single cohort from a specific clinical setting may limit the generalizability of the findings, highlighting the need for external validation in diverse patient populations.

6. Conclusion

In this work, we presented a framework for longitudinal medical image analysis and its application to the study of HCC early detection. Beyond the original framework, an ablation of the two pre-training stages indicated that the self-supervised pre-training of the CNN backbone is the dominant driver of the observed performance gains, while the sequence-order-prediction task contributes comparatively little. Exploratory experiments further showed that naively fusing imaging modalities through channel concatenation yields at most marginal gains over the strongest single modality, suggesting that realizing the full potential of multi-modal fusion will require more structured fusion strategies, which we leave to future work. The methodology is versatile and offers a framework for other longitudinal medical imaging applications where patient monitoring

can lead to better outcomes. Yet, each condition presents its own unique challenges in terms of disease progression and imaging characteristics, which will require further adaptation of the HCCNet framework. Additionally, exploring the integration of multi-modal data, including EHR, genomic information, and other imaging modalities, could provide a fuller view of patient health and improve predictive accuracy, while further external validation is needed to assess the generalizability of the study’s findings. In the future, we therefore plan to validate our findings on an external validation cohort and aim to incorporate multi-modal data to further improve upon our results. Likewise, future research is needed to investigate the versatility of the proposed methodology and its application to diseases beyond HCC.

Acknowledgment

This research was performed within the project entitled *CHARISMA: Predictie van hepatocellulair carcinoom op basis van MRI-radiomics Primair leverkanker (hepatocellulair carcinoma, HCC)*, which is funded by *Pioneers in Health Care* (PIHC) Innovation Fund, Twente region, The Netherlands. This work was conducted within the project "Predicting the Development of Primary Liver Cancer Using the AI-HCC Model", which is funded by ZonMw, The Netherlands Organisation for Health Research and Development.

Data availability

The MRI examinations and HCC diagnoses used in this study come from a retrospective cohort of liver cirrhosis patients who underwent repeated abdominal MRI screening for HCC surveillance between March 2011 and September 2022 at Medisch Spectrum Twente in Enschede, the Netherlands. The images were anonymized before pre-training and fine-tuning. The data cannot be shared publicly.

Code availability

The code of HCCNet is publicly available at <https://github.com/TheAxeC/hccnet>.

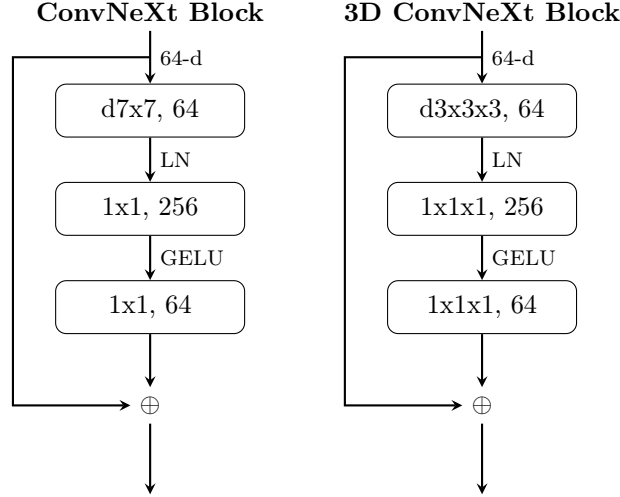
Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a generative-AI assistant (OpenAI ChatGPT) to help refine the English language of the manuscript. After using this tool the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Implementation Details

Appendix A.1. 3D ConvNeXt Block Design

Figure A.4: Original and 3D ConvNeXt Block Designs



Note: The figure illustrates the structure of a ConvNeXt block with an input feature map of dimensionality $d = 64$. The input is passed through a series of convolutional layers, where the leading d denotes the use of depthwise convolutions, before being merged with the block's initial input through a residual connection. The figure's left hand side represents the original implementation in (Liu et al., 2022), while the right illustrates the study's 3D adaptation.

Figure A.4 illustrates the structure of the original ConvNeXt block as implemented in (Liu et al., 2022) as well as the 3-dimensional adaptation proposed in this study. Compared to the original design, we inflate the 2D depthwise and pointwise convolutional layers to 3 dimensions and reduce the depthwise convolutional layer's proposed kernel size from $k = 7$ to $k = 3$, leaving the remaining block design unchanged.

Appendix A.2. Data Augmentation Strategy

During pre-training of the CNN feature extractor, we randomly crop two global views of shape $s = 72^3$ and two local view of shape $s = 48^3$ from the image, randomly selecting a single series for each of the cropped views from within a single contrast family of the provided MRI sequence (e.g., among the diffusion-weighted b -value series, or among the in- and out-of-phase series). Contrarily, during pre-training of the Transformer encoder, we randomly crop a multi-channel sub-volume of shape $s = 72^3$ from the image, concatenating all available series from the sequence of MRI's to a single 4D image. Apart from the alternating cropping approach, further augmentations remain consistent across pre-training. Specifically, we randomly shift and scale the intensity

by a constant v using probability $p = 1.0$, apply random 90 degree rotations with probability $p = 0.5$, and randomly flip the cropped view along the three axes with probability $p = 0.5$. During fine-tuning as well as for the baseline model, we use the same data augmentation strategy as during pre-training of the Transformer encoder.

Appendix A.3. Hyperparameter Settings

Table A.5: Hyperparameter Settings

	Pre-Training		Fine-Tuning	Baseline
	Backbone	Encoder		
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight Initialization	$\mathcal{N}(0, 0.02)$	Partially Pre-Trained	Pre-Trained	$\mathcal{N}(0, 0.02)$
Training Steps	32000	8000	400	400
Warm-up Steps	1600	[400, 800]	20	40
Effective Batch Size	128	32	[16, 32, 64]	[16, 32, 64]
Sequence Length	-	8	8	8
Base Learning Rate	4e-4	1e-4	[5e-5, 1e-4, 2e-4]	[5e-5, 1e-4, 2e-4]
Minimum Learning Rate	1e-6	1e-6	1e-6	1e-6
Weight Decay	5e-2	5e-2	1e-5	1e-5
Maximum Weight Decay	5e-1	5e-1	1e-5	1e-5
Label Smoothing	-	-	0.1	0.1
Dropout	-	0.2	0.2	0.2
Gradient Clipping	1.0	1.0	3.0	3.0
Teacher Temperature	0.04	-	-	-
Student Temperature	0.1	-	-	-
Projection Head Dimensionality	1024	-	-	-
Exponential Moving Average	0.9995	-	-	-
Random Flip and Rotation	0.5	0.5	0.5	0.5
Random Intensity Scaling and Shifting	1.0	1.0	1.0	1.0

Note: During pre-training of the Transformer encoder, model parameters of the CNN backbone are initialized using pre-trained weights, whereas parameters of the transformer encoder are randomly initialized. The learning rate is linearly increased over 400 training steps for model variations HCCNet-F, -C, and -P and over 800 training steps for HCCNet-N and -T. For all phases of training, the base learning rate η is computed as a function of the batch size b using the following formula: $\eta = \alpha \frac{\sqrt{b}}{\sqrt{128}}$, where α is a common scaling factor that is chosen according to b .

Table A.5 presents detailed hyperparameter settings for pre-training, fine-tuning, and the baseline respectively. Note that during CNN-backbone pre-training we slightly deviate from the hyperparameter settings proposed in (Caron et al., 2021). This is due to convergence issues that we noted during initial model runs. Consequently, we drastically reduced the projection head’s output dimensionality to 1024 classes and reduced the temperature of the teacher network to a constant $\tau_{F_t} = 0.04$. For the other parameters, we followed the recommendations in (Caron et al., 2021) and increased the initial momentum of the exponential moving average update to $m = 0.9995$ given our reduced batch size of $B = 128$.

Appendix B. Additional Results

Appendix B.1. Extended Baseline Model Results

Table B.6: Baseline Model Results for the Prediction of HCC at the Next Observation

		Performance Metrics					
		Accuracy	Precision	Recall	F1-Score	AUPRC	AUROC
DW-MRI	HCCNet-F	0.669 ± 0.063	0.251 ± 0.050	0.675 ± 0.100	0.364 ± 0.064	0.292 ± 0.040	0.716 ± 0.024
	HCCNet-C	0.666 ± 0.074	0.251 ± 0.054	0.675 ± 0.100	0.363 ± 0.068	0.293 ± 0.033	0.705 ± 0.017
	HCCNet-P	0.655 ± 0.070	0.248 ± 0.042	<u>0.700 ± 0.061</u>	0.364 ± 0.047	0.308 ± 0.032	0.722 ± 0.015
	HCCNet-N	0.655 ± 0.083	0.256 ± 0.046	0.725 ± 0.050	0.375 ± 0.049	0.301 ± 0.037	0.724 ± 0.024
T1 DCE-MRI	HCCNet-T	0.638 ± 0.081	0.239 ± 0.042	0.700 ± 0.100	0.352 ± 0.048	0.294 ± 0.029	0.722 ± 0.016
	HCCNet-F	0.667 ± 0.105	0.244 ± 0.038	0.600 ± 0.200	0.332 ± 0.053	0.348 ± 0.056	<u>0.723 ± 0.027</u>
	HCCNet-C	0.695 ± 0.069	0.252 ± 0.045	0.600 ± 0.200	0.348 ± 0.072	0.312 ± 0.015	0.693 ± 0.043
	HCCNet-P	0.705 ± 0.034	0.247 ± 0.054	0.550 ± 0.150	0.339 ± 0.079	<u>0.350 ± 0.053</u>	0.703 ± 0.049
	HCCNet-N	0.681 ± 0.050	0.241 ± 0.051	0.575 ± 0.100	0.338 ± 0.064	0.375 ± 0.050	0.692 ± 0.054
	HCCNet-T	0.688 ± 0.037	0.243 ± 0.045	0.575 ± 0.100	0.341 ± 0.058	0.348 ± 0.059	0.692 ± 0.050
T1 IOP & T2-MRI	HCCNet-F	0.758 ± 0.039	0.269 ± 0.030	0.400 ± 0.050	0.318 ± 0.013	0.338 ± 0.051	0.639 ± 0.024
	HCCNet-C	<u>0.758 ± 0.030</u>	0.272 ± 0.016	0.425 ± 0.100	0.327 ± 0.024	0.313 ± 0.051	0.628 ± 0.009
	HCCNet-P	0.747 ± 0.044	0.272 ± 0.034	0.450 ± 0.061	0.335 ± 0.029	0.307 ± 0.061	0.624 ± 0.024
	HCCNet-N	0.758 ± 0.048	0.293 ± 0.045	0.475 ± 0.050	0.359 ± 0.040	0.251 ± 0.034	0.618 ± 0.037
T1 IOP-MRI	HCCNet-T	0.733 ± 0.049	0.258 ± 0.059	0.450 ± 0.061	0.326 ± 0.062	0.236 ± 0.030	0.592 ± 0.039
	HCCNet-F	0.733 ± 0.017	0.218 ± 0.053	0.375 ± 0.137	0.273 ± 0.082	0.228 ± 0.024	0.640 ± 0.051
	HCCNet-C	0.747 ± 0.018	<u>0.283 ± 0.035</u>	0.525 ± 0.094	<u>0.367 ± 0.051</u>	0.247 ± 0.017	0.670 ± 0.044
	HCCNet-P	0.740 ± 0.013	0.269 ± 0.025	0.500 ± 0.079	0.349 ± 0.038	0.245 ± 0.022	0.669 ± 0.035
	HCCNet-N	0.737 ± 0.027	0.253 ± 0.043	0.450 ± 0.100	0.323 ± 0.059	0.237 ± 0.027	0.659 ± 0.048
	HCCNet-T	0.751 ± 0.007	0.266 ± 0.029	0.450 ± 0.100	0.333 ± 0.050	0.234 ± 0.026	0.643 ± 0.056
T2-MRI	HCCNet-F	0.690 ± 0.034	0.185 ± 0.064	0.375 ± 0.137	0.247 ± 0.087	0.243 ± 0.065	0.618 ± 0.053
	HCCNet-C	0.690 ± 0.031	0.225 ± 0.024	0.500 ± 0.000	0.309 ± 0.022	0.265 ± 0.092	0.620 ± 0.038
	HCCNet-P	0.686 ± 0.037	0.216 ± 0.027	0.475 ± 0.050	0.296 ± 0.030	0.257 ± 0.073	0.639 ± 0.030
	HCCNet-N	0.690 ± 0.049	0.225 ± 0.024	0.500 ± 0.079	0.308 ± 0.025	0.258 ± 0.077	0.619 ± 0.033
	HCCNet-T	0.686 ± 0.069	0.201 ± 0.081	0.400 ± 0.146	0.266 ± 0.102	0.305 ± 0.118	0.618 ± 0.038

Note: AUPRC and the AUROC values range from 0.14 and 0.5 to 1 respectively. All metrics are presented using the mean and standard deviation across the five cross-validation folds. The best results are emphasised in boldface. Second best results are underlined. DW-MRI comprises diffusion MRI scans with diffusion coefficient $b \in \{0, 150, 400, 800\}$. T1 DCE-MRI comprises a pre-contrast and three post-contrast (i.e., late arterial, portal venous, and delayed phase) series. T1 IOP & T2-MRI comprise T1-weighted in- and out-of-phase MRI as well as T2-weighted MRI with short and long echo times. T1 IOP-MRI and T2-MRI report the same combined modality split into its constituent contrasts.

Table B.6 presents extended baseline model results for the next observation prediction of HCC stratified by image modality and model variation. Across modalities, baseline models trained using diffusion weighted MRIs consistently display the highest detection rate of HCC as well as the greatest average discriminatory capability. Overall differences between imaging modalities are however far less pronounced than for the models initialized with SSL pre-trained weighted (s. Table 1) and the overall diagnostic performance remains generally poor.

Appendix B.2. Reliability Analysis

Table B.7: Baseline vs. Fine-Tuned Reliability Scores

		Baseline			Fine-Tuned		
		ECE	MCE	Brier Score	ECE	MCE	Brier Score
DW-MRI	HCCNet-F	0.273 $\pm$ 0.021	0.785 $\pm$ 0.066	0.230 $\pm$ 0.014	0.132 $\pm$ 0.029	0.579 $\pm$ 0.131	0.117 $\pm$ 0.023
	HCCNet-C	0.292 $\pm$ 0.017	0.761 $\pm$ 0.091	0.236 $\pm$ 0.016	0.146 $\pm$ 0.008	0.805 $\pm$ 0.117	0.123 $\pm$ 0.006
	HCCNet-P	0.297 $\pm$ 0.044	0.830 $\pm$ 0.069	0.237 $\pm$ 0.018	<u>0.126 $\pm$ 0.018</u>	0.656 $\pm$ 0.095	0.119 $\pm$ 0.021
	HCCNet-N	0.292 $\pm$ 0.030	0.758 $\pm$ 0.086	0.233 $\pm$ 0.016	0.131 $\pm$ 0.021	0.895 $\pm$ 0.033	0.119 $\pm$ 0.013
	HCCNet-T	0.295 $\pm$ 0.031	0.830 $\pm$ 0.098	0.236 $\pm$ 0.022	0.101 $\pm$ 0.010	<u>0.606 $\pm$ 0.205</u>	0.088 $\pm$ 0.014
T1 DCE-MRI	HCCNet-F	0.320 $\pm$ 0.078	0.679 $\pm$ 0.167	0.239 $\pm$ 0.034	0.212 $\pm$ 0.050	0.748 $\pm$ 0.101	0.163 $\pm$ 0.044
	HCCNet-C	0.306 $\pm$ 0.070	0.608 $\pm$ 0.144	0.233 $\pm$ 0.025	0.274 $\pm$ 0.059	0.761 $\pm$ 0.102	0.211 $\pm$ 0.049
	HCCNet-P	0.307 $\pm$ 0.055	0.614 $\pm$ 0.107	0.229 $\pm$ 0.020	0.161 $\pm$ 0.040	0.665 $\pm$ 0.078	0.131 $\pm$ 0.029
	HCCNet-N	0.305 $\pm$ 0.059	0.612 $\pm$ 0.135	0.229 $\pm$ 0.022	0.192 $\pm$ 0.032	0.834 $\pm$ 0.015	0.157 $\pm$ 0.028
	HCCNet-T	0.299 $\pm$ 0.055	0.569 $\pm$ 0.113	0.227 $\pm$ 0.019	0.131 $\pm$ 0.013	0.676 $\pm$ 0.205	<u>0.115 $\pm$ 0.009</u>
T1 IOP & T2-MRI	HCCNet-F	0.182 $\pm$ 0.026	0.740 $\pm$ 0.063	0.170 $\pm$ 0.017	0.204 $\pm$ 0.038	0.765 $\pm$ 0.106	0.174 $\pm$ 0.029
	HCCNet-C	<u>0.188 $\pm$ 0.031</u>	0.722 $\pm$ 0.096	<u>0.177 $\pm$ 0.016</u>	0.145 $\pm$ 0.031	0.768 $\pm$ 0.037	0.147 $\pm$ 0.022
	HCCNet-P	0.206 $\pm$ 0.046	0.695 $\pm$ 0.099	0.188 $\pm$ 0.022	0.192 $\pm$ 0.023	0.764 $\pm$ 0.119	0.174 $\pm$ 0.031
	HCCNet-N	0.195 $\pm$ 0.027	0.730 $\pm$ 0.073	0.184 $\pm$ 0.016	0.163 $\pm$ 0.057	0.607 $\pm$ 0.045	0.135 $\pm$ 0.013
	HCCNet-T	0.205 $\pm$ 0.022	0.761 $\pm$ 0.124	0.185 $\pm$ 0.019	0.151 $\pm$ 0.033	0.817 $\pm$ 0.076	0.135 $\pm$ 0.023

Note: Reliability metrics range from 0 to 1. ECE represents the expected, while MCE denotes the maximal calibration error with lower values representing better model performance. All metrics are presented using the mean and standard deviation across the five cross-validation folds. The best model results are emphasised in boldface. Second best results are underlined. DW-MRI comprises diffusion MRI scans with diffusion coefficient $b \in \{0, 150, 400, 800\}$. T1 DCE-MRI contain a pre-contrast and three post-contrast (i.e., late arterial, portal venous, and delayed phase) series. T1 IOP & T2-MRI comprise T1-weighted in- and out-of-phase MRI as well as T2-weighted MRI with short and long echo times.

Table B.7 presents the expected and maximal calibration error as well as the Brier score for baseline and fine-tuned model results stratified by imaging modality and model architecture. As for the comparison across performance metrics depicted in Table 2, variations in reliability scores are predominantly driven by imaging modality rather than model complexity. For the baseline model, models trained using in- and out-of-phase T1-weighted and T2-weighted MRIs consistently display the lowest average reliability scores. That is despite that baseline models trained using diffusion weighted MRIs show the best average discriminatory power between patients of low and high risk of HCC. Contrarily, for the fine-tuned models, preceding pre-training establishes a negative relationship between performance and reliability metrics. Here, HCCNet fine-tuned using diffusion weighted MRIs display the lowest average reliability scores, while models fine-tuned using dynamic contrast enhanced T1-weighted MR images exhibit the highest average reliability metrics.

References

- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. URL: <http://arxiv.org/abs/2104.14294>. arXiv:2104.14294 [cs].
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.
- Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S., 2022. Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. Pattern Recognition 132, 108919. URL: <https://www.sciencedirect.com/science/article/pii/S0031320322004009>, doi:10.1016/j.patcog.2022.108919.
- Dammu, H., Ren, T., Duong, T.Q., 2023. Deep learning prediction of pathological complete response, residual cancer burden, and progression-free survival in breast cancer patients. PLOS ONE 18, e0280148. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0280148>, doi:10.1371/journal.pone.0280148. publisher: Public Library of Science.
- El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E., 2021. Are Large-scale Datasets Necessary for Self-Supervised Pre-training? URL: <http://arxiv.org/abs/2112.10740>, doi:10.48550/arXiv.2112.10740. arXiv:2112.10740 [cs].
- Gao, Y., Ghodrati, V., Kalbasi, A., Fu, J., Ruan, D., Cao, M., Wang, C., Eilber, F.C., Bernthal, N., Bukata, S., Dry, S.M., Nelson, S.D., Kamrava, M., Lewis, J., Low, D.A., Steinberg, M., Hu, P., Yang, Y., 2021. Prediction of soft tissue sarcoma response to radiotherapy using longitudinal diffusion MRI and a deep neural network with generative adversarial network-based data augmentation. Medical Physics 48, 3262–3372. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.14897>, doi:10.1002/mp.14897. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14897>.
- Gong, W., Beckmann, C.F., Smith, S.M., 2024. Individualised prediction of longitudinal change in multimodal brain imaging. Imaging Neuroscience 2, 1–19.
- Granzio, D., Zohren, S., Roberts, S., . Learning Rates as a Function of Batch Size: A Random Matrix Theory Approach to Neural Network Training .
- Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised Learning. URL: <http://arxiv.org/abs/2006.07733>. arXiv:2006.07733 [cs, stat].

- Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S., 2023. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine* 6, 1–16. URL: <https://www.nature.com/articles/s41746-023-00811-0>, doi:10.1038/s41746-023-00811-0. publisher: Nature Publishing Group.
- Jamaludin, A., Kadir, T., Zisserman, A., 2017. Self-supervised Learning for Spinal MRIs, in: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer International Publishing, Cham. pp. 294–302. doi:10.1007/978-3-319-67558-9_34.
- Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., Duan, X., Tang, J., Chang, D.T., Wu, X., et al., 2021. Predicting treatment response from longitudinal images using multi-task deep learning. *Nature Communications* 12, 1–12.
- Kumar, Y., Marttinen, P., 2024. Improving medical multi-modal contrastive learning with expert annotations. *arXiv preprint arXiv:2403.10153*.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. URL: <http://arxiv.org/abs/1909.11942>. arXiv:1909.11942 [cs].
- Lee, D., Hu, Y.c., Kuo, L., Alam, S., Yorke, E., Li, A., Rimner, A., Zhang, P., 2022. Deep learning driven predictive treatment planning for adaptive radiotherapy of lung cancer. *Radiotherapy and Oncology* 169, 57–63. URL: [https://www.thegreenjournal.com/article/S0167-8140\(22\)00098-6/fulltext](https://www.thegreenjournal.com/article/S0167-8140(22)00098-6/fulltext), doi:10.1016/j.radonc.2022.02.013. publisher: Elsevier.
- Li, X., Chen, S., Yang, J., 2020. Understanding the disharmony between weight normalization family and weight decay, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4715–4722. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5904>. issue: 04.
- Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986.
- Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization. URL: <http://arxiv.org/abs/1711.05101>, doi:10.48550/arXiv.1711.05101. arXiv:1711.05101 [cs, math].
- Lu, L., Dercle, L., Zhao, B., Schwartz, L.H., 2021. Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. *Nature Communications* 12, 6654. URL: <https://www.nature.com/articles/s41467-021-26990-6>, doi:10.1038/s41467-021-26990-6. publisher: Nature Publishing Group.

- Parikh, N.D., Mehta, A.S., Singal, A.G., Block, T., Marrero, J.A., Lok, A.S., 2020. Biomarkers for the early detection of hepatocellular carcinoma. *Cancer Epidemiology, Biomarkers & Prevention* 29, 2495–2503.
- Ren, H., Wang, J., Zhao, W.X., Wu, N., 2021. RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation, in: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ACM, Virtual Event Singapore. pp. 3503–3511. URL: <https://dl.acm.org/doi/10.1145/3447548.3467069>, doi:10.1145/3447548.3467069.
- Rivail, A., Schmidt-Erfurth, U., Vogl, W.D., Waldstein, S.M., Riedl, S., Grechenig, C., Wu, Z., Bogunovic, H., 2019. Modeling Disease Progression in Retinal OCTs with Longitudinal Self-supervised Learning, in: Rekik, I., Adeli, E., Park, S.H. (Eds.), *Predictive Intelligence in Medicine*, Springer International Publishing, Cham. pp. 44–52. doi:10.1007/978-3-030-32281-6_5.
- Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740.
- Vaswani, A., 2017. Attention is all you need. *Advances in Neural Information Processing Systems* .
- Wang, C., Alam, S.R., Zhang, S., Hu, Y.C., Nadeem, S., Tyagi, N., Rimner, A., Lu, W., Thor, M., Zhang, P., 2020. Predicting spatial esophageal changes in a multimodal longitudinal imaging study via a convolutional recurrent neural network. *Physics in Medicine & Biology* 65, 235027. URL: <https://dx.doi.org/10.1088/1361-6560/abb1d9>, doi:10.1088/1361-6560/abb1d9. publisher: IOP Publishing.
- Wang, C., Rimner, A., Hu, Y.C., Tyagi, N., Jiang, J., Yorke, E., Riyahi, S., Mageras, G., Deasy, J.O., Zhang, P., 2019. Toward predicting the evolution of lung tumors during radiotherapy observed on a longitudinal MR imaging study via a deep learning algorithm. *Medical Physics* 46, 4699–4707. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13765>, doi:10.1002/mp.13765. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13765>.
- Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T., 2020. On layer normalization in the transformer architecture, in: *International Conference on Machine Learning*, PMLR. pp. 10524–10533.
- Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R.H., Aerts, H.J., 2019. Deep Learning Predicts Lung Cancer Treatment

Response from Serial Medical Imaging. *Clinical Cancer Research* 25, 3266–3275. URL: <https://doi.org/10.1158/1078-0432.CCR-18-2495>, doi:10.1158/1078-0432.CCR-18-2495.

Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y., 2021. Positional contrastive learning for volumetric 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 II* 24, Springer. pp. 221–230.

Zhang, S., Li, Z., Zhou, H.Y., Ma, J., Yu, Y., 2023. Advancing 3d medical image analysis with variable dimension transform based supervised 3d pre-training. *Neurocomputing* 529, 11–22.

Zhang, Y., Hu, C., Zhong, L., Song, Y., Sun, J., Li, M., Dai, L., Zhou, Y., Yang, W., 2022. Spatiotemporal Attention for Early Prediction of Hepatocellular Carcinoma Based on Longitudinal Ultrasound Images, in: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*, Springer Nature Switzerland, Cham. pp. 534–543. doi:10.1007/978-3-031-16437-8_51.