Preoperative assessment of the histopathological growth patterns of colorectal liver metastasis on CT using artificial intelligence

MSc Thesis Biomedical Engineering - BM51035
Medical Physics

by

Samuel van Gurp

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended on Friday, July 21, 2023, at 2:00 pm.

Chair:	Dr. F.M. Vos,	Erasmus MC, TU Delft
Supervisors:	Dr. M.P.A Starmans,	Erasmus MC
	Dr. S. Klein,	Erasmus MC
Independent committee member:	Dr. D.A. De Jesus,	Erasmus MC

Abstract

Background Histopathological growth patterns (HGP) are a biomarker for predicting survival and systemic treatment effectiveness in colorectal liver metastasis (CRLM). Currently, HGP assessment in CRLM requires the resection specimen. Predicting the HGP from preoperative medical imaging could allow more personalised care and better outcomes.

Methods. 252 patients underwent CRLM resection between 2004 and 2018 without receiving any systemic treatment. Patients were characterised as having either pure desmoplastic growth (dHGP) or any other type of growth pattern combination (non-dHGP) (21% dHGP; 79% non-dHGP). These categories were chosen because pure desmoplastic growth is predictive of better overall survival. regions of interest were automatically extracted using a UNet based segmentation model. These ROIs were passed to a radiomics model and a deep learning model to classify between dHGP/non-dHGP and predict the fraction of dHGP.

Results. The best-performing classification method was the radiomic approach achieving an AUC of 0.67 (95% CI: 0.58-0.78), whereas the best-performance deep learning model achieved an average AUC value of 0.59 (95% CI: 0.53-0.65). Additionally, regression predicting the fraction of dHGP failed, with the predicted values showing no significant correlation with the actual value.

Conclusions. Radiomics can be used to assess HGP, however further improvements in predictive performance are needed before these methods can be applied.

1 Introduction

Worldwide, colorectal cancer is responsible for 10% of all new cancer cases and 9.4% of cancer-related deaths [1]. 30% of these patients develop colorectal liver metastasis (CRLM) [2]. Patients diagnosed with CRLM have poor outcomes, less than half survive past 5 years after the diagnosis [3].

In the pursuit of improved treatment for these patients, the need has arisen for biomarkers that help personalise care by predicting systemic therapy response and survival. One such biomarker is the histopathological growth pattern (HGP). The HGP characterises the interface between the healthy liver and the cancerous tissue of the CRLM. [4, 5]. The two most common HGP types are desmoplastic (dHGP) and replacement growth (rHGP), in addition to the less common pushing HGP. These HGP types can occur together in a single CRLM.

It has previously been found that patients with dHGP have better survival compared to those with replacement growth [6]. Furthermore, the HGP type has been associated with the effectiveness of systemic treatments [7, 8]. In the past, guidelines suggested a cut-off of 50% of a single HGP type to determine the dominant pattern [9]. However, more recent investigations, show that distinguishing between dHGP and all replacement and mixed cases (together called non-dHGP) is more relevant for predicting survival [3].

Currently, it is not possible to assess the HGP before the CRLM has been surgically resected. Considering the predictive power of HGPs in determining both survival and the effectiveness of systemic treatments, having this information early on in the treatment process would be a valuable tool [10, 11].

One potential approach to pre-operatively assess the HGP type is through medical imaging. However, it is currently challenging for radiologists to visually distinguish the growth patterns of HGPs from a magnetic resonance imaging (MRI) or computed tomography (CT) scan [10]. In order to address such issues and to go beyond visual inspection the field of radiomics has emerged. Radiomics involves the extraction and analysis of a large number of quantitative features from medical images, aiming to find patterns between imaging characteristics and clinical variables.

Promising results have been achieved by applying radiomics for the prediction of HGP type [12–18]. However, most of the existing studies do not address the prediction of the most relevant cut-off for distinguishing between dHGP and non-dHGP. The aim of this work is to address this limitation by predicting the newest dHGP/non-dHGP cut-off and directly predicting the fraction of dHGP, thereby avoiding the use of a specific cut-off value. To this end, this work explores two methods, a deep learning based method and a feature-based radiomics approach. To train and test these models a dataset of CT scans along with corresponding postoperative histological assessments was retrospectively collected. The dataset consists of scans from 252 patients.

2 Background

2.1 Histopathological Growth Patterns

The HGP characterise the interface of the CRLM and the healthy liver tissue. The two most common HGP types observed are desmoplastic (dHGP) and replacement growth (rHGP), in addition to the less common pushing HGP (pHGP). In the case of desmoplastic growth, the CRLM and liver are separated by a rim of desmoplastic tissue which can be seen as a sort of scar tissue (see Figure 1a). In contrast to replacement growth, this reaction is not observed, here the cancerous cells infiltrate the healthy tissue, where they are in direct contact (see Figure 1c). The exact mechanism by which either dHGP, rHGP or pHGP arise is still unknown [19].

Importantly, the different HGP can co-exist in a single CRLM. One region of the interface can exhibit dHGP and another rHGP, this occurs in around 60% of cases in our dataset [6]. This mixing behaviour makes the assessment of HGP substantially different from distinguishing tumour subtypes.

The interest in HGP stems from its association with prognosis and systemic treatment effectiveness. Generally, dHGP has been connected to better outcomes as compared to rHGP. More specifically The stratification of patients based on HGP has evolved over recent years. Initially, patients were divided into dHGP and rHGP categories based on the predominant growth pattern. However, this may not be the most informative approach. Specifically, a study by Galjart et al. (2019) revealed that patients exclusively exhibiting desmoplastic HGP (dHGP) have better overall survival (OS) compared to those with replacement growth HGP or a mixture of HGPs (non-dHGP) [3]. Furthermore, patients with small fractions of non-dHGP growth were shown to have poor survival.

Not only is the OS correlated to the HGP of CRLM but, the HGPs are also predictive for the effectiveness of systemic treatments like chemotherapy and anti-angiogenic therapy [7, 20]. For instance, non-desmoplastic patients have improved OS when treated with neoadjuvant chemotherapy while this is not the case for desmoplastic patients [8]. Next to chemotherapy, bevacizumab is another treatment used in the management of CRLM, often given in conjunction with chemotherapy. Bevacizumab is an anti-angiogenic agent, which means it is a drug that inhibits the formation of new blood vessels. Bevacizumab has been found to induce response more frequently in patients with dHGP compared to those with rHGP [20]. This difference in response is attributed to rHGP using vessel co-option instead of sprouting angiogenesis, which confers resistance to the anti-angiogenic agent since it does not rely on the growth of new vessels to proliferate.

Considering the predictive value of HGP on OS and systemic treatment effectiveness, it is clear that the assessment of the HGPs could be a valuable tool for treating CRLM. However, the current method for evaluating HGP requires pathology slices of the resection specimens to be analysed using a light microscope. Thus, limiting HGP assessment until after resection CRLM.

It would be advantageous if the HGP type could be determined earlier in the disease treatment. One way a tumour can be examined in vivo is by taking a biopsy. However, in the case of CRLM, tumour heterogeneity and the potential risk of complications for the patient, make obtaining a biopsy of the metastasis, not a viable option for the assessment of the HGP type [10]. This raises the need for other ways of HGP assessment. Medical imaging would be a favourable option as it is non-invasive and is already part of clinical care. Therefore, the need for computational methods that can do so is clear [10, 11].

2.2 Histopathological Growth Patterns on Medical Imaging

Imaging of CRLM is done using either contrast-enhanced CT or MRI. Due to the unique dual vascular supply of the liver, there are distinct contrast phases depending on the timing of the scan with respect to the injection time of the contrast agent. First the arterial phase (AP) then the portal venous phase (PVP) and lastly the delayed phase (DP) or the washout phase. The ns are most commonly used for the imaging of metastases like CRLM [21].

In order to assess the HGP research has been conducted to find markers that are unique to one of the HGP on medical imaging. Until now, no conclusive marker has been found to differentiate dHGP from rHGP. However, two markers have shown some association with HGP type [10]. These are, whether or not the CRLM has a clearly defined border and if there is rim enhancement present on contrasted enhanced scans. Here the lack of a clearly defined border is indicative of replacement growth, which is explained by the infiltration of the CRLM into the surrounding tissue. Secondly, rim enhancement is associated the desmoplastic growth this is explained by the inflammatory microenvironment causing vasodilation in the surrounding liver tissue [10].

However, conflicting reports exist regarding these findings. Rim enhancement on both AP and PVP has been associated with dHGP in some studies [16, 17, 22]. On the other hand, Han et al. [14] found that this relationship is not significant, while Li et al. [15] reported the opposite relationship.

With respect to a clearly defined border being predictive of growth patterns similar ambiguity exists. Correlations have been reported [23]. However, other works that have investigated this report only weak correlations between growth patterns and HGP [15–17].

Taken together, the predictive value of these imaging markers for differentiating HGP types remains unclear. One possible reason for this uncertainty is the subjective nature of assessing rim enhancement and border sharpness, which may not be consistent between studies. Moreover, the varying cut-offs used in

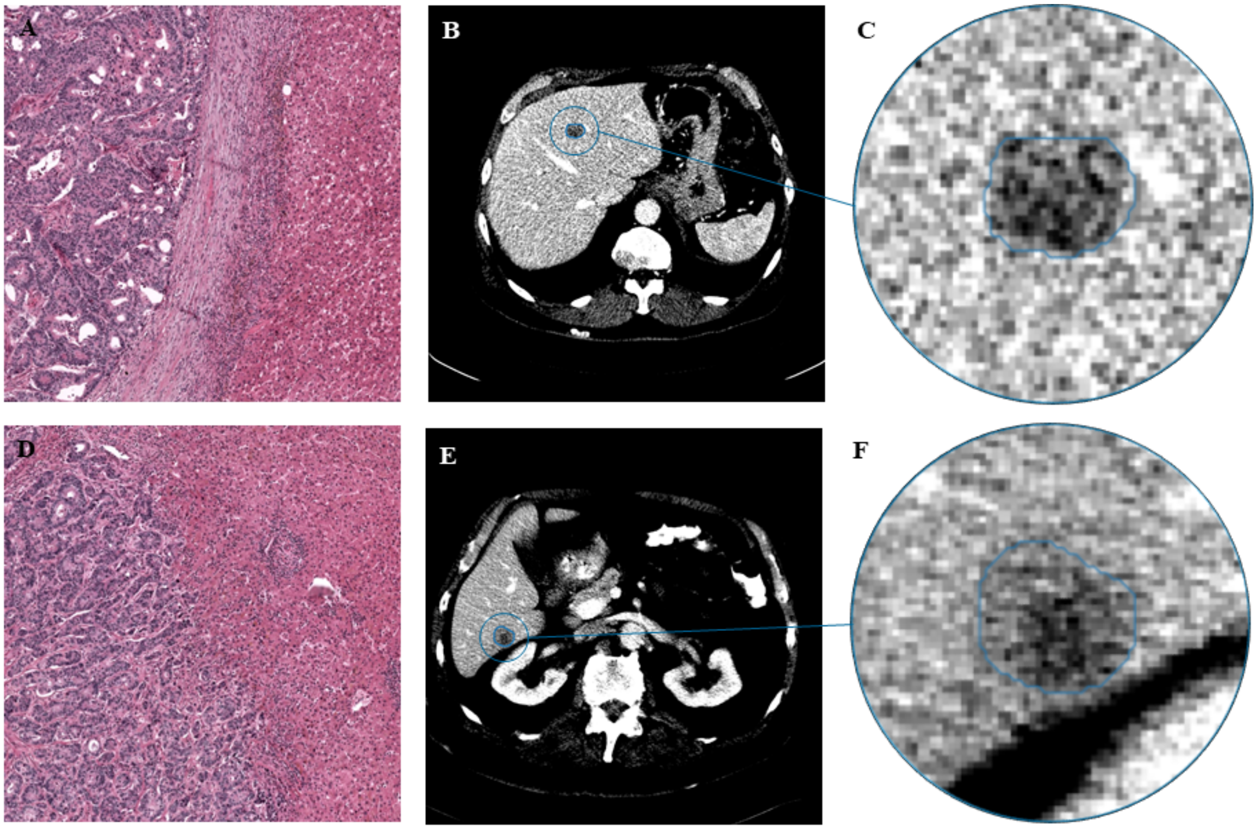


Figure 1: Colorectal liver metastases on H&E stained histology slides and PVP CT scan. Figure 1 a-c show desmoplastic growth pattern, while Figure 1 c-e show replacement growth pattern. The intensity range of the CT scans has been clipped to between 30-150 HU. to improve the visibility of the CRLM.

different studies to categorise dHGP and rHGP further obfuscate the results.

2.3 Radiomics for Histopathological Growth Patterns

As discussed in the previous section it is not possible to distinguish HGP types from medical imaging by eye. Therefore, computational methods have been developed [12–17]. These methods can be grouped under radiomics. This is a method where a wide range of computational imaging features are extracted from a region of interest (ROI) in the scan. These are then used to train a classifier to predict a clinical variable such as the HGP type.

These methods consistently report excellent predictive performance of their models across diverse populations worldwide. However, it is important to note that this line of research is relatively new, with all work being published within the last four years. As such, there is still room for further development in methodologies and evaluations. Some of the major points that still could be improved are; More clinically relevant predictions, more rigorous validation of model performance, and more reproducible methods.

Regarding clinical relevancy, the grouping of HGP types is an important issue to consider Since most studies pose HGP prediction as a binary classification problem this necessitates the selection of a cut-off point. However, the placement of this cut-off point significantly impacts patient group outcomes. The consensus on the most predictive cut-off has evolved over time. Recent evidence highlights the importance of distinguishing between dHGP and non-dHGP [3]. This new finding limits the usefulness of earlier work which predicts the predominant growth pattern [13, 14, 16]. The latest research on HGP prediction has adopted this new cut-off [17]. Future work should adopt this new dHGP/non-dHGP cutoff or predict the fraction dHGP/rHGP directly.

One notable concern is the issue of reproducibility in the existing literature. Many studies lack the publication of their code or data, which is considered the gold standard for ensuring transparency and reproducibility. Furthermore, the methods employed are often described only superficially, limiting the ability of others to replicate and validate the findings.

Another aspect that contributes to the challenge of reproducibility is the process of deriving regions of interest (ROIs). Manual annotation, which is commonly used for this purpose, can introduce inconsistencies due to inter-observer variability [12]. To address this issue, automatic segmentation techniques have been proposed as a preferable alternative. Automating the segmentation of CRLM would not only improve reproducibility but also facilitate the eventual clinical implementation, as manual segmentation of CRLM is not feasible in routine clinical practice.

By addressing these concerns, such as providing code and data, offering detailed descriptions of methods, and exploring automated segmentation approaches, the field can enhance reproducibility and promote more reliable and clinically applicable assessments of HGP in CRLM.

3 Methods

3.1 Dataset

This study was conducted in accordance with the Dutch Code of Conduct for Medical Research of 2004 and received approval from the local institutional review board, "Medische Ethische Toetsings Commissie" (METC) under the reference MEC-2023-0016. As the study used retrospectively collected and anonymized data, informed consent was waived.

The study enrolled patients who underwent surgical treatment for CRLM at Erasmus MC. Subsequently, the resected CRLM specimens were evaluated for HGP following the most recent guidelines [9]. The assessment of HGPs was conducted for patients treated between 1999-2018. Only chemotherapy-naive patients were included in the study because chemotherapy can influence the HGPs [24]. Furthermore, subjects needed to have a preoperative PVP CT scan available. Pre-contrast and arterial phase CT scans were excluded due to limited availability in a minority of patients (161 had an AP scan available and only 115 had a non-contrast scan).

Table 1: The median value together with the interquartile range of patient and imaging characteristics of the 252 individuals included in the study. Statistical analysis was performed using a Mann-Whitney U test for continuous variables and a chi-square test for categorical variables to calculate the corresponding p-values.

Patients	**dHGP**	**non-dHGP**	**p-value**
Total	52 (21%)	200 (79%)	
Age	67 [56-71]	67 [61-73]	0.46
Sex			0.68
Male	19	59	
Female	33	141	
Imaging			
Slice thickness (mm)	3 [2.0-5.0]	4 [2.0-5.0]	0.30
Pixel spacing (mm)	0.75[0.70-0.78]	0.73 [0.68-0.78]	0.27
Tube current (mA)	240 [157-340]	269[150-361]	0.42
Peak voltage (kV)	120 [120-120]	120 [120-120]	0.07

The selection of scans based on contrast phases involved categorising them into arterial, portal venous, or non-contrast/delayed phases. Initially, this categorisation was performed based on the series description, where scans were automatically classified into the appropriate categories if the description contained relevant terms indicating the phase. In cases where no series description was available, I manually classified them based on visual inspection, with guidance from an abdominal radiologist. Patients were excluded either due to poor scan quality, characterised by corrupted files or extremely low spatial resolution, such that I was not able to determine the contrast phase. Additionally, patients were excluded if the segmentation model failed to segment any lesion. An overview of the patient selection process is shown in Figure 2.

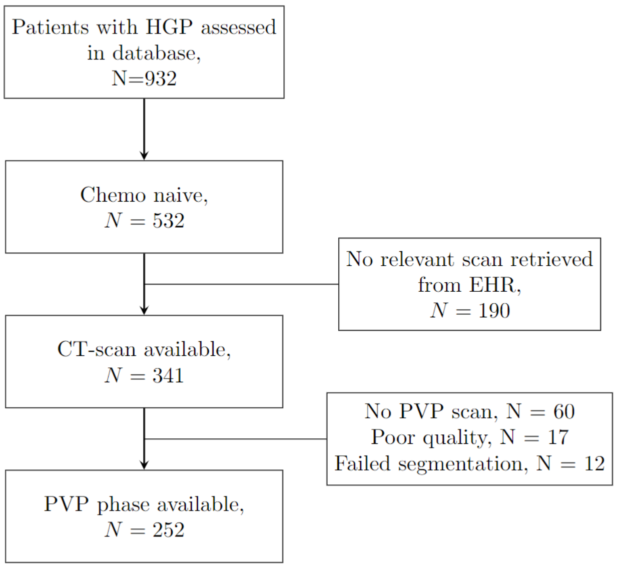

Patients with HGP assessed in database, N=932
Chemo naive, $N = 532$
No relevant scan retrieved from EHR, $N = 190$
CT-scan available, $N = 341$
No PVP scan, N = 60
Poor quality, N = 17
Failed segmentation, N = 12
PVP phase available, $N = 252$

Figure 2: Patient Selection Process for HGP Assessment. The figure depicts the patient selection process. Poor quality includes scans that had either a corrupted file or a very poor spatial resolution, failed segmentations are scans in which no lesion was segmented. EHR: electronic health record, PVP: portal venous contrast phase

The resulting patient population has a similar HGP distribution as earlier work has found on a similar cohort from the Erasmus MC has reported, as shown in Figure 3 [3]. Other patients and scan characteristics are shown in Table 1.

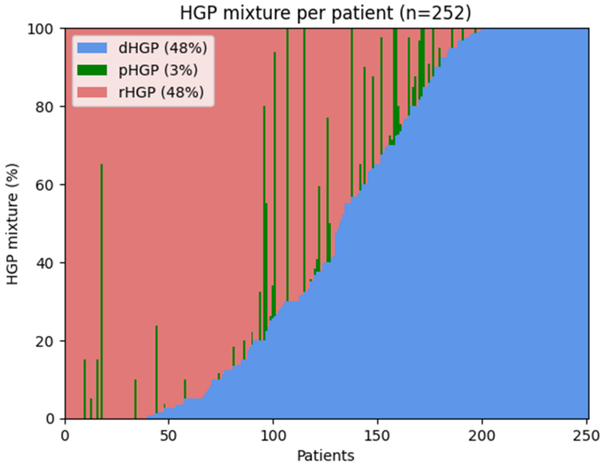


Figure 3: Distribution of histopathological growth patterns, ranked based on percentage dHGP.

3.2 Segmentation

In this study, the regions of interest to be segmented were the liver as a whole and the CRLM. This was done with a pre-trained nnU-Net model that showed the best performance on the liver tumour segmentation (LiTS) challenge [25, 26]. This dataset comprises a wide range of liver tumours, including primary tumours like hepatocellular carcinoma and cholangiocarcinoma, as well as secondary liver tumours originating from colorectal, breast, and lung cancers. Furthermore, the dataset exhibited heterogeneity in terms of imaging protocols, including variations in contrast enhancement time, scanner models, and settings.

The nn-UNet model was able to achieve a dice similarity coefficient (DSC) of 0.74 for tumour segmentation and 0.90 for the liver in the LiTS challenge dataset [26]. Since the authors did not report segmentation performance per lesion type and per contrast phase, it is unclear whether the nn-UNet model performs well on a dataset consisting solely of CRLMs imaged during the portal venous phase.

3.3 Radiomics

For all radiomics experiments, the open-source package WORC (Workflow for Optimal Radiomics Classification) was used [27]. WORC automatically optimises the construction of radiomics workflows based on conventional machine learning. The input to WORC are the CT scans together with the binary segmentation mask of the presumed CRLM. From this segmentation, 564 features are calculated, describing shape, intensity and texture. These features are then used in the data mining component. The data mining component consists of feature processing steps involving imputation scaling and selection. Additionally, the workflow may incorporate dimensionality reduction and resampling. Finally, machine learning algorithms are used to find relationships between image features and clinical variables.

WORC aims to find a radiomics workflow consisting of a selection of the components described above. This is done using a random search of feature processing methods, models and their associated hyperparameters. 1000 radiomics workflows are constructed each with varying models and hyperparameters. From these, a subset of the 100 best-performing models is selected based on the F1 score on a validation set. These are then combined in an ensemble model by averaging their prediction.

WORC was also used for predicting the fraction dHGP using regression. The operations remain the same as for classification only class-based feature and sample preprocessing methods are omitted. In the regression workflow, classification models are replaced by regression models and the selection of the optimal workflows is based on the coefficient of determination (R^2).

3.4 Deep learning

For all deep learning experiments, a 3D ResNet-10 model was used, as implemented in the MONAI framework (see Figure 4) [28, 29]. This relatively shallow implementation of a ResNet was chosen to reduce the model's degrees of freedom and therefore mitigate the risk of overfitting, on the relatively small training set. Additionally, ResNet-10 has shown promising results in studies on liver tumours [30, 31].

For the deep learning experiments, the scans were resampled using b-spline interpolation to match the median value in the axial plane. Additionally, the spacing between slices was adjusted to 1 mm to preserve the high slice resolution found in some of the scans. As a result, the final spacing of the resampled scans was set to 0.74 × 0.74 × 1.0 mm.

The scans were cropped, and a rectangular bounding box was applied around the liver segmentation. This step served two purposes. Firstly, to reduce memory requirements. Secondly, to exclude uninformative areas of the scans. Additionally, an alternative approach was explored, in which a bounding box was placed around the largest CRLM itself. Both bounding boxes were resized to the median shape resulting in a liver bounding box of size 250 × 275 × 40 voxels and 30 × 35 × 25 voxels respectively.

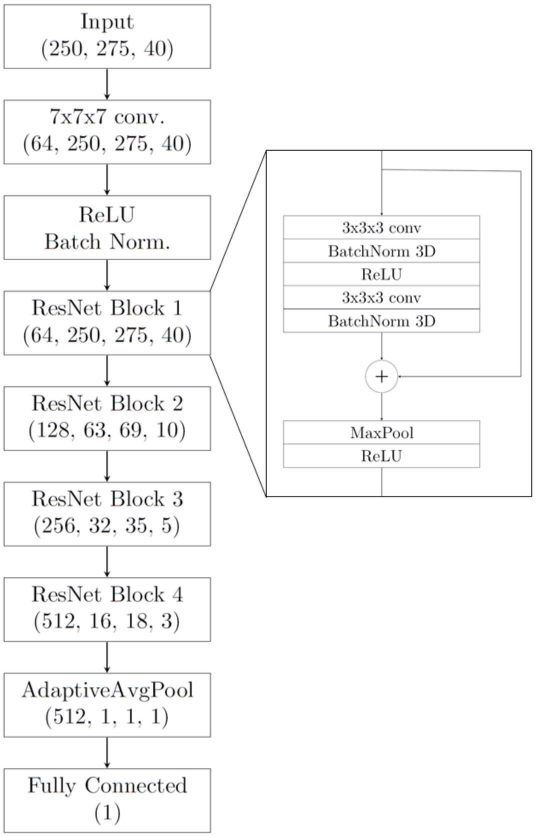


Figure 4: Schematic representation of ResNet-10 model, featuring 4 subsequent ResNet blocks of which one is shown. This is the model configuration for in input image of shape: (250,275,40). Every block represents an operation. The output dimensions are shown in brackets, which indicate the number of convolutional filters along with the width, height, and depth of the image for each layer. The model consists of 3D convolutional layers (n x n x n conv), rectified linear unit (ReLU) activation functions, Batch normalisation (BatchNorm 3D), an adaptive average pooling layer (AdaptiveAvgPool) and a fully connected layer.

The model starts with one convolutional layer followed by 4 ResNet blocks each of which consists of two convolutional layers and a residual connection that adds the input to the block to the output. With every ResNet block the model doubles the number of convolutional filters and uses max-pooling to reduce the spatial dimensions by half. After the last ResNet block the number of features is reduced to 512 by an adaptive average pooling layer. These features are then passed through a fully connected layer and a sigmoid which reduces the output to a single value.

with this prediction, the loss is calculated. For classification, the binary cross entropy loss was used and for the regression experiments, the mean squared error loss was used. Lastly, all models were trained for 200 epochs, with a learning rate of 0.001 for the ADAM optimiser.

3.5 Experimental set-up

3.5.1 Segmentation

To evaluate the nn-UNet model performance on segmentation of CRLMs imaged on PVP CT scans a test was performed on scans from the WORC database [32]. This set comprises 77 PVP CT scans of CRLM together with a manual segmentation of the CRLM made by three different observers. Using the pre-trained nn-UNet the CRLM were segmented. To assess the segmentation performance of the model, the DSC was calculated between the automatic and manual segmentation. This analysis was limited by the fact that not all lesions were manually segmented (lesions that were not surgically resected were excluded). Therefore, the DSC was only calculated for lesions which had some overlap with manual segmentation. So the resulting DSC is an upper bound on the model's performance.

Radiomics

Two radiomics classification experiments were conducted: one including the largest segmented lesion and another involving all lesions consisting of more than 100 voxels. The selection of the largest lesion aimed to explore whether the growth patterns are best visible on the larger CRLMs and that automatic segmentations for larger lesions are more accurate. The experiment including all segmented lesions sought to enhance model performance by increasing the number of training samples. The lower voxel limit of 100 was established to exclude the smallest lesions which may be ill-defined on a CT scan, and run a larger risk of being incorrectly segmented [26].

The experiments with the largest lesion included 252 segmented lesions (the same as the number of patients) while experiments including multiple lesions included 564 segmented lesions.

For the radiomics regression experiments, settings were informed by the earlier classification results. Therefore, only one experiment was performed using only the largest segmented lesion as this performed best. Al predictions of dHGP fraction outside of the range of [0,1] were clipped to either 0 or 1.

Deep learning

Using deep learning both classification and regression experiments were performed. For classification, two types of experiments were performed one where either a bounding box around the liver was passed as input to the model or where a bounding box around the largest tumour segmentation was used. The bounding box around the liver was used to omit uninformative regions. Additionally, to further guide the model to the most salient region a bounding box was placed around the largest tumour.

To improve the model's performance experiments were performed with different settings. A list of these experimental settings is shown in Table 2.

Table 2: Experiment type and abbreviation[†].

Abbreviation	**Experiment type**
DA	Data augmentation
LDA	Light data augmentation
OS	Oversampling of minority class
SM	Segmentation mask in the second channel
LB	Bounding box around the liver
TB	Bounding box around the tumor

[†]For data augmenmation setting see Table 5.

Data augmentation was used to mitigate overfitting by artificially increasing the number of training samples. Here data augmentation (DA) refers to randomly applying a transformation to the training data. These transformations include zooming, rotation, flipping, adding Gaussian noise and elastic transformation. The light data augmentation (LDA) omitted the Gaussian noise and elastic transformation, as Gaussian noise may obscure the detail in the image that is aimed to be captured, and elastic transformations may excessively distort the image.

In the case of the liver bounding box, to guide the model to the region of interest (the CRLM) the segmentation mask of the liver and tumour was passed in a second channel. Lastly, to address the class imbalance (dHGP/non-dHGP $\approx$ 20/80) the dHGP class was over-sampled during training (OS) such that the model is trained on a 50/50 class distribution.

A total of seven classification experiments were conducted. The first experiment (exp.1: LB) used the liver bounding box as input without any of the experimental settings. All the remaining experiments incorporated minority class oversampling (exp.2: LB + OS). For the liver bounding box, two data augmentation approaches were tested (exp.3: LB + OS + DA and exp.4: LB + OS + LDA). Additionally, one experiment was performed where the segmentation mask was passed in the second channel (exp.5: LB + OS + DA + SM). Lastly, there were two variations of experiments conducted using the tumour bounding box as input: one with light data augmentation and one without (exp.6: TB + OS + DA and exp.7: LB + OS + LDA).

3.6 Statistics

Both the deep learning and the radiomics model performance were analysed through cross-validation. However, their specific setups differed. For the radiomics experiments a 100 times random split cross-validation was used. The deep learning experiments used 5-fold stratified cross-validation.

The decision to use 5-fold cross-validation in the deep learning experiments was driven by the longer training time required for deep learning models compared to radiomics models. The reason for stratified cross-validation in the case of deep learning was the low number of folds in combination with a relatively small dataset which results in the risk of larger variation in the distribution of class labels between the folds.

To assess the performance of all classification experiments, the following metrics were used: accuracy, area under the receiver operating characteristic curve (AUC), and F1-score. A threshold of 0.5 was used for both the F1-score and accuracy calculations. For regression experiments, the evaluation metrics consisted of mean squared error (MSE), coefficient of determination (R^2), and Pearson correlation coefficient.

To calculate these metrics the average was taken from the performance on the test set for every cross-validation. For radiomics experiments, 95% confidence intervals were calculated using the corrected resampled t-test, considering that the individual cross-validation results are not independent [33]. For the deep learning experiments, 95% confidence intervals were computed by taking a range of times 1.96 the standard deviation around the average of the 5 cross-validation results.

4 Results

4.1 Segmentation

The performance of the nn-UNet segmentation model was evaluated by computing the DSC for each segmented lesion that had some overlap with the corresponding manual segmentation. This approach was adopted due to the unavailability of manual segmentations for every lesion. This way the model achieved a DSC of 0.73 ± 0.17, similar to the interobserver agreement of 0.69 between the human observers.

The manual annotators segmented 89 tumours (1.2 on average) while the nn-UNet model segmented 229 (3.1 on average). On the dataset introduced in this study, the number of CRLMs per patient found by radiologists was 1.9 on average. This includes both CRLMs that were later on resected and ones that were not. So, the nn-UNet model segmented more lesions than the radiologist finds on average.

4.2 Classification

4.2.1 Radiomics Classification

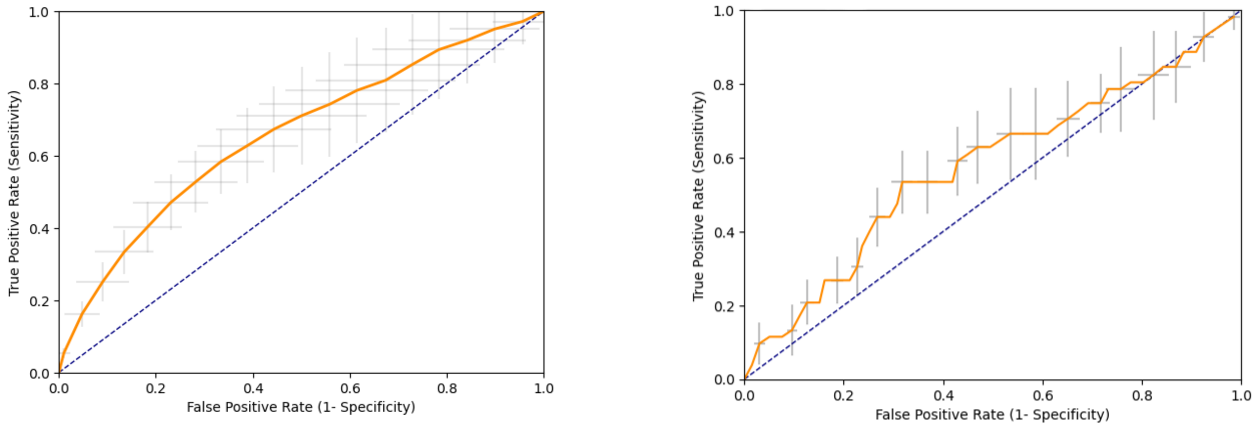
Two radiomics classification experiments were conducted (see Table 3). The experiment using only the largest segmented lesion demonstrated the highest performance on all metrics (AUC = 0.67 (95% CI 0.58-0.76)), also compared to the deep learning methods (see Figure 5). The model using multiple lesions per patient showed worse performance (AUC = 0.53 (95% CI 0.44-0.61)).

4.2.2 Deep Learning Classification

Although the confidence intervals of the deep-learning experiments all overlap, some trends can still be observed (see Table 3). Specifically, the inclusion of oversampling of the minority class (OS) resulted in improved performance, as did the use of data augmentation techniques, going from an AUC of 0.54 (95% CI 0.44-0.63) to 0.58 (95% CI 0.39-0.77). Notably, light data augmentation outperformed the stronger variant, which included additional techniques such as elastic deformation and the addition of Gaussian noise. This improved the AUC from 0.52 (95% CI 0.40-0.64) to 0.59 (95% CI 0.53-0.65). Furthermore, using the bounding box around the tumour (TB) as input performed worse than the around the liver, not scoring higher than an AUC of 0.54 (95% CI0.48-0.60). For all of the experiments, the model exhibited overfitting (see Figure 7 in appendix B.2).

4.3 Regression

The results of both the radiomics and deep learning regression models predicting the fraction of dHGP are shown in Figure 6 and Table 4. Figure 6 shows the predicted dHGP fractions plotted against actual value Table 4. Both methods are not able to predict the fraction dHGP, deep learning method and the radiomics method achieve R^2 of -0.03 (95%CI -0.32,0.26) and -0.54 (95%CI -0.76,-0.32) respectively. From Figure 6 it can be seen that the radiomics model has only learned to predict around the average value of dHGP in the dataset (48%). The deep learning model, on the other hand, predicts uniformly distributed random values between 1 and 0. Both models are thus unable to predict dHGP, this is also reflected in the metrics (see Table 4). Again the performance of the deep learning model has been impacted by overfitting in the training set (see Figure 8 in appendix B.2).



(a) Radiomics, AUC: 0.67 (95% CI 0.58-0.760)

(b) Deep learning, AUC: 0.59 (95% CI 0.53-0.65)

Figure 5: Receiver operating characteristic (ROC) curves for radiomics and deep learning classification of dHGP and non-dHGP. The Radiomics model was trained and tested on the largest segmented lesion. The deep learning model was trained using cropped scans around the liver (LB), minority class over sampling OS and light data augmentation (LDA). Error bars for the radiomics ROC-curve correspond to 96% CI and for deep learning to 68% CI.

5 Discussion

This work aimed to develop a fully automatic radiomics and deep learning method to predict the HGP based on CT scans from clinical care. The prediction was done in two ways, one using binary classification between dHGP and non-dHGP and regression to predict the fraction of dHGP directly. The best-performing classification method was the radiomics approach achieving an AUC of 0.67 (95% CI 0.58- 0.76), whereas the best-performing deep learning model achieved an AUC of 0.59 (95% CI 0.3-0.65). Neither the radiomics method nor the deep learning method were able to predict the fraction of dHGP directly, showing no correlation between the predicted and the true dHGP fraction.

The assessment of HGP type is a challenging task, this is demonstrated by the fact that this is not done by radiologists in either clinical practice or in a research setting, despite clinical relevance [10, 11, 19]. With respect to distinguishing the most clinically relevant groups of dHGP/non-dHGP, a specific challenge arises. Only a small fraction of rHGP can change the classification from dHGP to non-dHGP. These small fractions may be hard to detect on medical imaging, such cases are thus prone to be misclassified. As a result, models predicting HGPs will be limited by accurately identifying these small fractions.

Considering these factors, the classification performance remains unsatisfactory. The classification radiomics method, WORC, has been extensively validated in various clinical applications [27]. This indicates that the features used may not capture the differences between HGPs. Therefore, it is worth exploring more tailored features designed to distinguish HGP types. One such feature may be one aiming to capture rim enhancement which has shown an association with dHGP (see subsection 2.2). To avoid the need for manual feature engineering, a deep learning approach was used to learn these tailored features directly from the scans. However, the model overfitted on the training data. Further regularisation and shrinking model complexity could mitigate this. However, it is not clear that this will resolve this issue.

While earlier studies have shown promising results predicting the HGP type using radiomics [12–16], the differences are too large to compare the previous studies to ours. These differences include the use of MRI instead of CT scans and the use of the predominant cut-off to distinguish between dHGP and rHGP instead of classifying based on dHGP/non-dHGP. This precludes a direct comparison.

Only the work by Sun et al. [17] is similar enough to compare. In their work, the authors also use PVP CT scans to predict the HGP based on the dHGP/non-dHGP cut-off. Their radiomics model demonstrated strong performance achieving an AUC of 0.88. While their approach shares similarities with our work, there are some differences. The main difference is in the imaging data used. Sun et al. [17] used a more consistent and higher-quality dataset than our study. They used scans collected over 4 years on two types of scanners in a single institution using a consistent protocol with a 2 mm slice thickness. In comparison, our work involved scans collected over 14 years from different hospitals and with a large variety of acquisition settings and vendors. One notable dif-

Table 3: Performance of radiomics and deep learning classification experiments. For deep learning, the mean and 95% confidence intervals for each metric are calculated over 5 × stratified cross-validation (see Table 2 for experiment abbreviations). For the radiomics method, the averages and 95% confidence intervals for each metric are calculated over 100 × random-split cross-validation. The ”Largest Tumour” experiment used only the largest segmented lesion for training and testing, while the ”Tumour > 100 Voxels” experiment included all segmented lesions consisting of more than 100 voxels.

Radiomics

Experiment	**Accuracy**	**AUC**	**F1-score**
Largest Tumour	0.78 [0.76, 0.81]	0.67 [0.58, 0.76]	0.71 [0.68, 0.75]
Tumour > 100 Voxels	0.82 [0.77, 0.86]	0.53 [0.44, 0.61]	0.74 [0.68, 0.79]

Deep learning

Experiment	**Accuracy**	**AUC**	**F1-score**
1: LB	0.72 [0.67, 0.77]	0.54 [0.44, 0.63]	0.10 [-0.02, 0.22]
2: LB + OS	0.73 [0.68, 0.78]	0.58 [0.39, 0.77]	0.15 [-0.01, 0.31]
3: LB + OS + DA + SM	0.70 [0.64, 0.76]	0.52 [0.46, 0.58]	0.11 [-0.01, 0.23]
4: LB + OS + DA	0.70 [0.63, 0.77]	0.52 [0.40, 0.64]	0.23 [0.07, 0.39]
5: LB + OS + LDA	0.73 [0.68, 0.78]	0.59 [0.53, 0.65]	0.20 [0.06, 0.34]
6: TB + OS	0.75 [0.70, 0.80]	0.52 [0.43, 0.61]	0.03 [-0.06, 0.12]
7: TB + OS + LDA	0.68 [0.64, 0.72]	0.54 [0.48, 0.60]	0.24 [0.19, 0.29]

Table 4: Performance of radiomics and deep learning regression experiments. The mean value of cross-validation together with a 95% confidence interval are reported. For radiomics, the mean and 95% confidence intervals for each metric are calculated over 100 × random-split cross-validation. For deep learning,the mean and 95% confidence intervals for each metric are calculated over 5 × stratified cross-validation

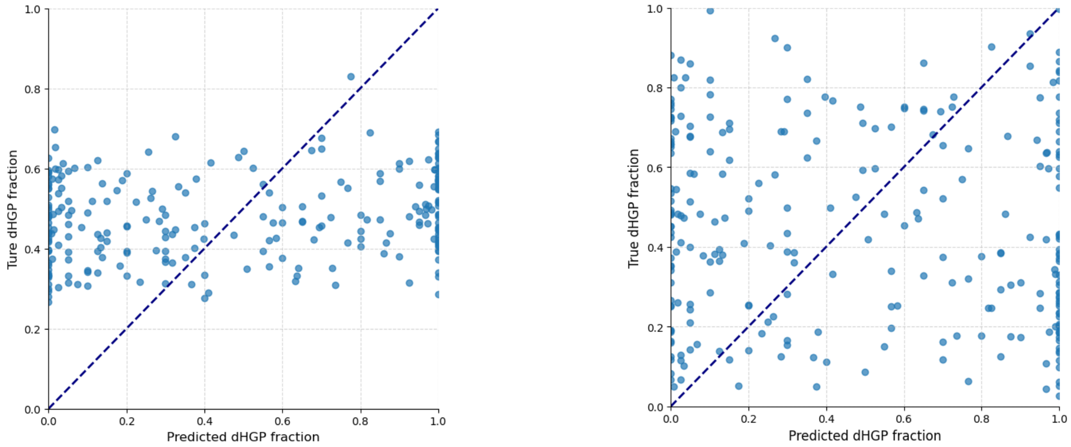
Experiment	**MSE**	$\mathbf{R^2}$	**Pearson Correlation**
Radiomics	0.16 [0.12, 0.21]	-0.03 [-0.32, 0.26]	0.27 (0.16, 0.39)
Deep learning	0.25 [0.21, 0.29]	-0.54 [-0.76, -0.32]	-0.1 [-0.20, -0.00]

ference is that Sun et al. [17] incorporated manual segmentations of the CRLM as a whole, in addition to segmenting the rim. By considering both the entire CRLM and the segmented rim, they were able to calculate additional features to serve as the basis for their prediction. As Sun et al. [17] have not yet externally validated their results, it is hard to assess if their model will generalise. Additionally, The lack of publicly available data and code precludes direct comparison methodologies.

Our work proposed a fully automatic segmentation method that eliminates the need for manual segmentation, resulting in an observer-independent and time-saving approach. Unfortunately, we were not able to rigorously validate the segmentation performance of the nn-UNet model. This was because there were only manual segmentations available for some of the lesions per patient (lesions that were not surgically resected were excluded). For the automatic segmentations that had some overlap with the manual segmentation, the model scored a DSC of 0.74. However, the model segmented more components than the manual observers, this was expected as some CRLMs were excluded from the manual segmentations on purpose. Moreover, on the dataset introduced in this study, the number of CRLMs per patient found by radiologists was 1.9 on average whereas the model segmented 3.1 on average. This points to the possibility that a large number of the segmented lesion are not CRLMs.

These incorrect segmentations could explain why basing the radiomics classification on multiple lesions performed worse instead of using only the largest segmented component. Smaller components are more likely to be incorrectly segmented [26]. An additional explanation could be that the HGP maybe not be as well defined on CT scans for smaller CRLM. The limited evaluation of the segmentation is one of the limitations of this study. To better assess segmentation performance automatic segmentation can be compared to expert manual segmentation by calculating the DSC of all lesions.

The second limitation is with respect to the contrast phases used. Multiple studies have shown that incorporating not only the PVP scans but also the pre-contrast and arterial phase scans can increase performance (increasing the AUC on the order of 10%) [14, 16]. Due to the limited availability of these scans



(a) Radiomics

(b) Deep learning

Figure 6: Predicted vs actual plot showing the relationship between the true fractions of dHGP and the corresponding predicted fractions for both radiomics (a) and deep learning (b). Every point represents the average prediction for that sample over every cross-validation experiment. The dashed diagonal line represents the perfect alignment between predicted and true values in the plot.

in our dataset, we chose to omit them to avoid excluding a significant number of patients.

The major limitation that is probably holding back performance is the heterogeneity and low image quality of the CT scans. Data heterogeneity arose from the fact that we included scans from a 14-year time span (2004 - 2018) from multiple centres without any acquisition protocol restrictions. Not only were scans collected over a long time span, but scans were also old, with a median age of 12 years. Thus scans are not up to modern standards as demonstrated by the large median slice thickness of 4mm (see Table 1). Higher quality scans can improve the classification of HGP types by capturing subtle differences, as there are no clear imaging markers to distinguish them. There is an ongoing effort made within the Erasmus MC to score the HGP of CRLMs resected after 2018. This means that in the future these scores can also be matched to clinical CT scans, and thus more modern scans could be included.

Other than improving the CT scan it would also be of interest to explore other image modalities like MRI. The present work did not explore this due to the limited availability of MRI scans in our cohort. However, MRI maybe better suited to the detection and characterisation of CRLM as it provides superior soft tissue contrast over CT [34].

6 Conclusion

This study presented a radiomics and deep learning approach to predict the HGP of CRLM based on CT scans from clinical care. This was done by using binary classification between dHGP and non-dHGP and regression to predict the fraction of dHGP directly. For classification, both approaches showed limited performance with the radiomics approach achieving an AUC of 0.67 (95% CI 0.58- 0.76), whereas the best-performing deep learning model achieved an AUC of 0.59 (95% CI 0.53-0.65). Additionally, regression predicting the fraction of dHGP failed, with the predicted values showing no correlation with the actual value. The deep learning method and the radiomics method achieve R^2 of 0.15 and -0.54 respectively.

A challenge for this task was the heterogeneity and low quality of the imaging data, which may have limited the ability to capture the subtle differences between the HGP types. Future work should focus on collecting more consistent and high quality imaging data, also including MRI. This could potentially improve the performance and reliability of the predictive models, thereby enabling more personalised care for patients suffering from CRLM.

References

[1] The International Agency for Research on Cancer. Global Cancer Observatory (GLOBOCAN 2020). URL: `https://gco.iarc.fr`, 2020. [Online; accessed 3. Apr. 2023].

[2] Sylvain Manfredi, Côme Lepage, Cyril Hatem, Olivier Coatmeur, Jean Faivre, and Anne-Marie Bouvier. Epidemiology and management of liver metastases from colorectal cancer. *Annals of surgery*, 244(2):254, 2006.

[3] Boris Galjart, Pieter MH Nierop, Eric P van der Stok, Robert RJ van den Braak,

Diederik J Höppener, Sofie Daelemans, Luc Y Dirix, Cornelis Verhoef, Peter B Vermeulen, and Dirk J Grünhagen. Angiogenic desmoplastic histopathological growth pattern as a prognostic marker of good outcome in patients with colorectal liver metastases. *Angiogenesis*, 22(2):355–368, 2019.

[4] Peter B Vermeulen, Cecile Colpaert, Roberto Salgado, Ruben Royers, Hilde Hellemans, Eva Van den Heuvel, Gerda Goovaerts, Luc Y Dirix, and Eric Van Marck. Liver metastases from colorectal adenocarcinomas grow in three patterns with different angiogenesis and desmoplasia. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 195(3): 336–342, 2001.

[5] James S Tomlinson, William R Jarnagin, Ronald P DeMatteo, Yuman Fong, Peter Kornprat, Mithat Gonen, Nancy Kemeny, Murray F Brennan, Leslie H Blumgart, and Michael D'Angelica. Actual 10-year survival after resection of colorectal liver metastases defines cure. *Journal of Clinical Oncology*, 25(29):4575–4580, 2007.

[6] Carlos Fernández Moro, Béla Bozóky, and Marco Gerling. Growth patterns of colorectal cancer liver metastases and their impact on prognosis: a systematic review. *BMJ open gastroenterology*, 5(1):e000217, 2018.

[7] F Buisman, E van der Stok, B Galjart, P Vermeulen, P Allen, V Balanchandran, W Jarnagin, P Kingham, D Grünhagen, B Groot Koerkamp, et al. Histopathological growth patterns as a guide for adjuvant systemic chemotherapy in patients with resected colorectal liver metastases. *European Journal of Surgical Oncology*, 45 (2):e10, 2019.

[8] Florian E Buisman, Eric P van der Stok, Boris Galjart, Peter B Vermeulen, Vinod P Balachandran, Robert RJ Coebergh van den Braak, John M Creasy, Diederik J Höppener, William R Jarnagin, T Peter Kingham, et al. Histopathological growth patterns as biomarker for adjuvant systemic chemotherapy in patients with resected colorectal liver metastases. *Clinical & experimental metastasis*, 37:593–605, 2020.

[9] Pieter-Jan Van Dam, Eric P Van Der Stok, Laure-Anne Teuwen, Gert G Van den Eynden, Martin Illemann, Sophia Frentzas, Ali W Majeed, Rikke L Eefsen, Robert RJ Coebergh van den Braak, Anthoula Lazaris, et al. International consensus guidelines for scoring the histopathological growth patterns of liver metastasis. *British journal of cancer*, 117(10):1427–1441, 2017.

[10] Emily Latacz, Pieter-Jan Van Dam, Christian Vanhove, Laura Llado, Benedicte Descamps, Núria Ruiz, Ines Joye, Dirk Grünhagen, Steven Van Laere, Piet Dirix, et al. Can medical imaging identify the histopathological growth patterns of liver metastases? In *Seminars in Cancer Biology*, volume 71, pages 33–41. Elsevier, 2021.

[11] Shenglin Li, Zhengxiao Li, Xiaoyu Huang, Peng Zhang, Juan Deng, Xianwang Liu, Caiqiang Xue, Wenjuan Zhang, and Junlin Zhou. Ct, mri, and radiomics studies of liver metastasis histopathological growth patterns: An up-to-date review. *Abdominal Radiology*, 47(10):3494–3506, 2022.

[12] Martijn Starmans, Florian E Buisman, Michel Renckens, François EJA Willemssen, Sebastian R van der Voort, Bas Groot Koerkamp, Dirk J Grünhagen, Wiro J Niessen, Peter B Vermeulen, Cornelis Verhoef, et al. Distinguishing pure histopathological growth patterns of colorectal liver metastases on ct using deep learning and radiomics: a pilot study. *Clinical & experimental metastasis*, 38(5):483–494, 2021.

[13] Vincenza Granata, Roberta Fusco, Federica De Muzio, Carmen Cutolo, Mauro Mattace Raso, Michela Gabelloni, Antonio Avallone, Alessandro Ottaiano, Fabiana Tatangelo, Maria Chiara Brunese, et al. Radiomics and machine learning analysis based on magnetic resonance imaging in the assessment of colorectal liver metastases growth pattern. *Diagnostics*, 12 (5):1115, 2022.

[14] Yuqi Han, Fan Chai, Jingwei Wei, Yali Yue, Jin Cheng, Dongsheng Gu, Yinli Zhang, Tong Tong, Weiqi Sheng, Nan Hong, et al. Identification of predominant histopathological growth patterns of colorectal liver metastasis by multi-habitat and multi-sequence based radiomics analysis. *Frontiers in oncology*, 10:1363, 2020.

[15] Wen-Hui Li, Shuai Wang, Yi Liu, Xin-Fang Wang, Yong-Feng Wang, and Rui-Mei Chai. Differentiation of histopathological growth patterns of colorectal liver metastases by mri features. *Quantitative Imaging in Medicine and Surgery*, 12(1):608, 2022.

[16] Jin Cheng, Jingwei Wei, Tong Tong, Weiqi Sheng, Yinli Zhang, Yuqi Han, Dongsheng Gu, Nan Hong, Yingjiang Ye, Jie Tian, et al. Prediction of histopathologic growth patterns of colorectal liver metastases with a noninvasive imag-

ing method. *Annals of Surgical Oncology*, 26(13): 4587–4598, 2019.

[17] Chao Sun, Xuehuan Liu, Jie Sun, Longchun Dong, Feng Wei, Cuiping Bao, Jin Zhong, and Yiming Li. A ct-based radiomics nomogram for predicting histopathologic growth patterns of colorectal liver metastases. *Journal of Cancer Research and Clinical Oncology*, pages 1–13, 2023.

[18] Shengcai Wei, Yuqi Han, Hanjiang Zeng, Shuai Ye, Jin Cheng, Fan Chai, Jingwei Wei, Jianwei Zhang, Nan Hong, Yudi Bao, et al. Radiomics diagnosed histopathological growth pattern in prediction of response and 1-year progression free survival for colorectal liver metastases patients treated with bevacizumab containing chemotherapy. *European Journal of Radiology*, 142:109863, 2021.

[19] Emily Latacz, Diederik Höppener, Ali Bohlok, Sophia Leduc, Sébastien Tabariès, Carlos Fernández Moro, Claire Lugassy, Hanna Nyström, Béla Bozóky, Giuseppe Floris, et al. Histopathological growth patterns of liver metastasis: updated consensus guidelines for pattern scoring, perspectives and recent mechanistic insights. *British journal of cancer*, 127(6): 988–1013, 2022.

[20] Sophia Frentzas, Eve Simoneau, Victoria L Bridgeman, Peter B Vermeulen, Shane Foo, Eleftherios Kostaras, Mark R Nathan, Andrew Wotherspoon, Zu-hua Gao, Yu Shi, et al. Vessel co-option mediates resistance to anti-angiogenic therapy in liver metastases. *Nature medicine*, 22 (11):1294–1302, 2016.

[21] Gregory T Sica, Hoon Ji, and Pablo R Ros. Ct and mr imaging of hepatic metastases. *American Journal of Roentgenology*, 174(3):691–698, 2000.

[22] Richard C Semelka, Shahid M Hussain, Hani B Marcos, and John T Woosley. Perilesional enhancement of hepatic metastases: correlation between mr imaging and histopathologic findings—initial observations. *Radiology*, 215(1):89–94, 2000.

[23] Junzo Yamaguchi, Ichiro Sakamoto, Toshio Fukuda, Hikaru Fujioka, Kou Komuta, and Takashi Kanematsu. Computed tomographic findings of colorectal liver metastases can be predictive for recurrence after hepatic resection. *Archives of Surgery*, 137(11):1294–1297, 2002.

[24] Pieter MH Nierop, Diederik J Höppener, Florian E Buisman, Eric P van der Stok, Boris Galjart, Vinod P Balachandran, William R Jarnagin, T Peter Kingham, Jinru Shia, Murielle Mauer, et al. Preoperative systemic chemotherapy alters the histopathological growth patterns of colorectal liver metastases. *The Journal of Pathology: Clinical Research*, 8(1):48–64, 2022.

[25] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[26] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

[27] Martijn Starmans, Sebastian R van der Voort, Thomas Phil, Milea JM Timbergen, Melissa Vos, Guillaume A Padmos, Wouter Kessels, David Hanff, Dirk J Grunhagen, Cornelis Verhoef, et al. Reproducible radiomics through automated machine learning validated on twelve clinical applications. *arXiv preprint arXiv:2108.08618*, 2021.

[28] 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.

[29] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[30] Jingwei Wei, Jin Cheng, Dongsheng Gu, Fan Chai, Nan Hong, Yi Wang, and Jie Tian. Deep learning-based radiomics predicts response to chemotherapy in colorectal liver metastases. *Medical Physics*, 48(1):513–522, 2021.

[31] Aisha Goedhart. Classification of primary liver tumors with radiomics and deep learning based on multiphasic MRI. *Delft University of Technology, Student Thesis Repository.*, 2023. URL `https://repository.tudelft.nl/islandora/object/uuid%3Af970b944-7912-4843-a1e5-1d55008d4a90?collection=education`. [Online; accessed 28. Jun. 2023].

[32] Martijn P.A. Starmans, Milea J.M. Timbergen, Melissa Vos, Guillaume A. Padmos, Dirk J.

Grünhagen, Cornelis Verhoef, Stefan Sleijfer, Geert J.L.H. van Leenders, Florian E. Buisman, Francois E.J.A. Willemssen, Bas Groot Koerkamp, Lindsay Angus, Astrid A.M. van der Veldt, Ana Rajicic, Arlette E. Odink, Michel Renckens, Michail Doukas, Rob A. de Man, Jan N.M. IJzermans, Razvan L. Miclea, Peter B. Vermeulen, Maarten G. Thomeer, Jacob J. Visser, Wiro J. Niessen, and Stefan Klein. The worc database: Mri and ct scans, segmentations, and clinical labels for 930 patients from six radiomics studies. *medRxiv*, 2021. doi: 10.1101/2021.08.19.21262238. URL `https://www.medrxiv.org/content/early/2021/08/25/2021.08.19.21262238`.

[33] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[34] Drew Maclean, Maria Tsakok, Fergus Gleeson, David J Breen, Robert Goldin, John Primrose, Adrian Harris, and James Franklin. Comprehensive imaging characterization of colorectal liver metastases. *Frontiers in Oncology*, 11:730854, 2021.

A Deep Learning settings

Table 5: Data augmentation techniques as part of the MONAI frame work and their corresponding settings used in the experiment [29]. Each transformation is listed along with its specific settings.

Transformation	**Settings**
Zoom (RandZoom)	prob = 0.5, min_zoom=1.0, max_zoom=1.2
Rotation (RandRotate)	range_z = 0.35, prob = 0.8
Flip (RandFlip)	prob = 0.5
Gaussian noise (RandGaussianNoise)	prob = 0.5, std = 0.05
Elastic deformation (Rand3DElastic)	$\sigma-$range=(5.0, 7.0),magnitude_range=(50, 150),prob=0.5,spatial_size=in_shape

B Supplementary Results

B.1 Multi-Observer Segmentation Evaluation

Table 6: Segmentation agreement measured by Dice Similarity Coefficient (DSC) between human observers (STUD (1st and 2nd time), PhD, RAD,) and two automatic methods (H-DenseUNet, nn-UNet). The table presents the mean and standard deviation of the DSC scores for the inter-observer agreement. The DSC scores quantify the level of agreement in segmentation. The bottom row displays the average values of the mean and standard deviation of the DSC scores for each observer.

Observer	STUD1	STUD2	PhD	RAD	H-DenseUNet	nn-UNet
STUD1	-	0.80 (0.15)	0.73 (0.14)	0.60 (0.18)	0.65 (0.26)	0.78 (0.16)
STUD2	0.80 (0.15)	-	0.77 (0.13)	0.63 (0.18)	0.66 (0.27)	0.79 (0.15)
PhD	0.73 (0.14)	0.77 (0.13)	-	0.69 (0.16)	0.63 (0.25)	0.74 (0.16)
RAD	0.60 (0.18)	0.63 (0.18)	0.69 (0.16)	-	0.58 (0.27)	0.61 (0.18)
CNN	0.65 (0.26)	0.66 (0.27)	0.63 (0.25)	0.58 (0.27)	-	0.75 (0.20)
nnUNet	0.78 (0.16)	0.79 (0.15)	0.74 (0.16)	0.61 (0.18)	0.75 (0.20)	-
Average	0.71 (0.18)	0.73 (0.18)	0.71 (0.17)	0.62 (0.19)	0.65 (0.25)	0.73 (0.17)

B.2 Deep Learning Loss Curves

The model which achieved the highest AUC: 0.59 (95% CI 0.53, 0.65) used oversampling, light data augmentation and used a bounding box around the liver as input (LB + OS + LDA). The loss curve and AUC metric over the training epochs are shown in Figure 7. From this figure, it is clear to see that the model overfits the training set. The train and test loss diverge around epoch 60. Around this time the test and train AUC also split. The AUC ends up being better than guessing but lacks behind the performance achieved with the WORC model.

It can be observed that after the test and test and train loss diverge the test loss increases while at the same time, the AUC stays quite flat.

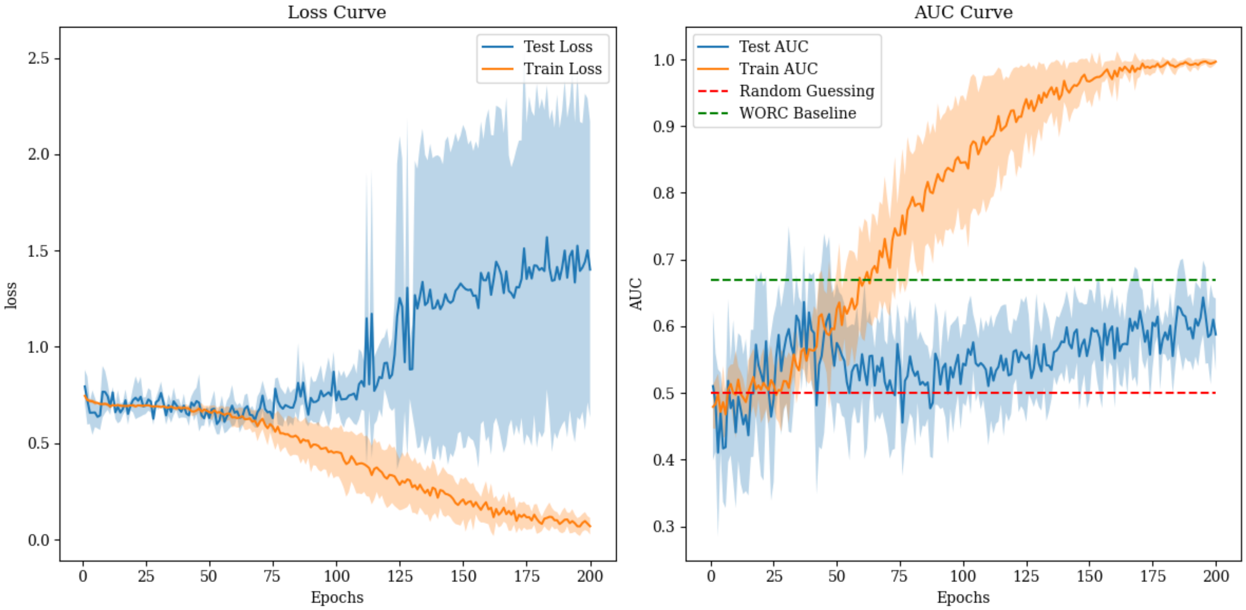


Figure 7: Binary cross entropy loss curve and AUC metric over the training epochs. The model achieving the highest AUC of 0.59, 95% CI (0.53-0.65) used oversampling, light data augmentation, and a bounding box around the liver as input (LB + OS + LDA). In the AUC curve, the red dashed line represents the performance achieved by random guessing, while the green dashed line corresponds to the best AUC achieved by the radiomics model.

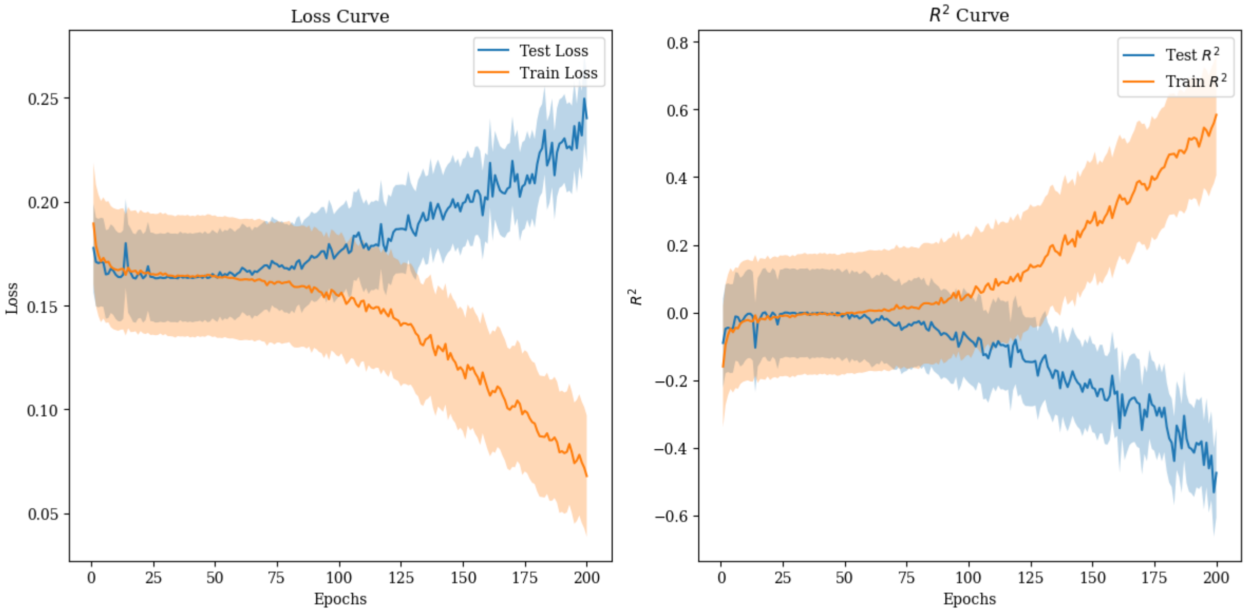


Figure 8: Mean squared error loss curve and R^2 metric over the training epochs of the deep learning regression model.