UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY
& BIOINFORMATICS

**Teresa Sousa Sotto Mayor**
BSc in Medical Biosciences

# CT Radiomics for Hepatocellular Carcinoma Recurrence Risk Prediction Aided by Deep Learning-Based Segmentation and Detection Methods

June, 2024

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
NOVA MEDICAL SCHOOL
NOVA IMS Information Management School
DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL UNIVERSIDADE NOVA DE LISBOA

# CT Radiomics for Hepatocellular Carcinoma Recurrence Risk Prediction Aided by Deep Learning-Based Segmentation and Detection Methods

**Teresa Sousa Sotto Mayor**
BSc in Medical Biosciences

**Adviser:** Nickolas Papanikolaou,
*Principal Investigator, Champalimaud Foundation*
**Co-adviser:** Cládia Soares
*Assistant Professor, NOVA University Lisbon*

**Examination Committee**
**Chair:** Isabel Rocha,
*Principal Investigator, NOVA University Lisbon*
**Opponent:** Ludwig Krippahl
*Investigator, NOVA University Lisbon*
**Adviser:** Nickolas Papanikolaou,
*Principal Investigator, Champalimaud Foundation*

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
NOVA University Lisbon
June, 2024

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to Nikos, for his unwavering support, guidance, and invaluable feedback throughout the development of this thesis and after its completion. A special thank you for welcoming me into his wonderful team and for creating such a positive and collaborative environment. I would also like to sincerely thank Prof. Claudia, for her insightful advice, and continuous support, which were vital to the completion of this work.

I am also immensely grateful to my colleagues Nuno, Ana Carolina, and José, who generously took the time to help me with my work and patiently answered questions. Their willingness to assist and collaborate made a significant difference in my research process. To Mariana, Ana Sofia, João, Miguel, and Raquel, thank you for being such kind and wonderful people, working alongside you has been both a pleasure and a source of inspiration.

A heartfelt thank you goes to my family, Marta and António, as well as my cousins, aunts, and uncles. A massive thank you to Luísa for all her scientific support and knowledge. I could not have done it without you.

Finally, to my friends, both old and new, thank you for being there, whether near or far. Your constant support, jokes, and just being there helped me stay sane and grounded through the ups and downs of this whole process.

This thesis is a reflection of all the support I have received, and I am deeply thankful to each and every one of you.

# Abstract

Liver cancer is the third leading cause of worldwide cancer-related deaths. Hepatocellular carcinoma (HCC) is the most common liver cancer, with recurrence rates reaching up to 70%.

The field of radiomics, which consists of extracting quantitative features from a region of interest (ROI) of a medical image, has been shown to preoperatively predict recurrence risk in HCC patients. However, most studies neglect the time and labour required to manually segment the ROI and fail to use a multi-institutional dataset to comprehensively assess the utility of radiomics-based models across a variety of computed tomography (CT) scanner models.

In this retrospective study, we utilise two datasets composed of around 230 CT examinations of patients eligible for HCC resection, and approximately 40 CT examinations of patients eligible for transarterial chemoembolisation treatment, to perform a comparison of two different deep learning lesion segmentation methods: nnU-Net and MedSam. Using the first dataset, comprised of 26 different scanner models across multiple manufacturers, we demonstrate that the multi-institutional nature of the data can be a limiting factor for the application of radiomics-based models.

Our findings show that, in terms of accuracy, lesion segmentation models generalized well to external test sets. On the second dataset, nnU-Net achieved mean ± standard deviation dice similarity coefficient (DSC) scores of 0.70 ± 0.23 and 0.67 ± 0.27 on arterial and portal venous phase CT scans, respectively. On the same phases of the first dataset, MedSam achieved 0.72 ± 0.13 and 0.70 ± 0.13 DSC, respectively, with no statistically significant difference between the two models. Regarding the prediction of recurrence risk, radiomics-based models outperformed those based on patient-specific clinical characteristics. However, test set performance remained unsatisfactory, with a maximum F1 score of 0.54, across the different data-model combinations.

Therefore, while radiomics provides valuable information over clinical data, challenges remain with respect to its usefulness when differences in image acquisition are present. Conversely, given the positive outcome observed within the lesion segmentation portion of this study, we anticipate the deep learning-based models to greatly enhance clinical workflow. Nevertheless, further work is required, from a clinical perspective, to thoroughly evaluate their real-world impact.

**Keywords:** Hepatocellular carcinoma, lesion segmentation, deep learning, recurrence risk, radiomics

# Resumo

O cancro do fígado é a terceira principal causa de morte relacionada com cancro a nível mundial. O carcinoma hepatocelular (HCC) é o cancro do fígado mais comum, com taxas de recorrência que atingem os 70%.

O domínio da radiómica, que involve a extração de características quantitativas de uma região de interesse (ROI) de uma imagem médica, demonstrou ser capaz de prever, no pré-operatório, o risco de recorrência em doentes com HCC. No entanto, a maioria dos estudos negligencia o tempo e o trabalho necessários para segmentar manualmente a ROI e não utiliza um conjunto de dados multi-institucional para avaliar exaustivamente a utilidade de modelos baseados na radiómica numa variedade de modelos de tomografia computorizada (CT).

Neste estudo retrospetivo, utilizamos dois conjuntos de dados compostos por cerca de 230 exames de CT de pacientes elegíveis para ressecção de HCC e de aproximadamente 40 exames de CT de pacientes elegíveis para tratamento de quimioembolização transarterial, para efetuar uma comparação de dois métodos diferentes de segmentação de lesões: nnU-Net e MedSam. Utilizando o primeiro conjunto de dados, composto por 26 modelos de *scanner* diferentes de vários fabricantes, demonstramos que a natureza multi-institucional dos dados pode ser um fator limitativo para a aplicação de modelos baseados em radiómica.

Os nossos resultados mostram que, em termos de exatidão, os modelos de segmentação de lesões se generalizaram bem para conjuntos de testes externos. No segundo conjunto de dados, nnU-Net alcançou média ± desvio-padrão de coeficiente de similaridade de dados (DSC) de 0,70±0,23 e 0,67±0,27 em exames de TC de fase arterial e portal, respetivamente. Nas mesmas fases do primeiro conjunto de dados, MedSam obteve um DSC de 0,72±0,13 e 0,70±0,13, respetivamente, sem diferença estatisticamente significativa entre os dois modelos. Relativamente à previsão do risco de recorrência, os modelos baseados na radiómica superaram os modelos baseados nas características clínicas específicas do doente. No entanto, o desempenho do conjunto de teste permaneceu insatisfatório, com uma medida máxima de F1 de 0,54, em todas as diferentes combinações de modelos e dados.

Por conseguinte, embora a radiómica forneça informações valiosas sobre os dados clínicos, subsistem desafios no que diz respeito à sua utilidade quando estão presentes diferenças na aquisição de imagens. Inversamente, dado o resultado positivo observado na parte de segmentação de lesões deste estudo, prevemos que os modelos baseados em aprendizagem profunda melhorem muito o fluxo de trabalho clínico. No entanto, é necessário mais trabalho, de uma perspetiva clínica, para avaliar minuciosamente o seu impacto no mundo real.

**Palavras-chave:** Carcinoma hepatocelular, segmentação de lesões, aprendizagem profunda, risco de recorrência, radiómica

# Contents

# List of Figures

# List of Tables

# Acronyms

**2D** Two-Dimensional

**3D** Three-Dimensional

**AFP** Alpha-Fetoprotein

**AI** Artificial Intelligence

**ALP** Alkaline Phosphatase

**ANN** Artificial Neural Network

**AP** Arterial Phase

**ASSD** Average Symmetric Surface Distance

**AUROC** Area Under the Receiver Operating Characteristic Curve

**BCLC** Barcelona Clinic Liver Cancer

**C-index** Concordance Index

**CNN** Convolutional Neural Network

**CT** Computed Tomography

**DSC** Dice Similarity Coefficient

**FN** False Negative

**FP** False Positive

**HCC** Hepatocellular Carcinoma

**HD** Hausdorff Distance

**HU** Hounsfield Units

**IBSI** Image Biomarker Standardization Initiative

**ICC** Intraclass Correlation Coefficient

**KNN** K-Nearest Neighbours

**LASSO** Least Absolute Shrinkage and Selection Operator

**ML** Machine Learning

**MRI** Magnetic Resonance Imaging

**PVP** Portal Venous Phase

**RAVD** Relative Absolute Volume Difference

**RFS** Recurrence Free Survival

**ROI** Region of Interest

**SAM** Segment Anything Model

**SVM** Support Vector Machine

**TACE** Transarterial Chemoembolisation

**TCIA** The Cancer Imaging Archive

**TN** True Negative

**TP** True Positive

**ViT** Vision Transformer

**XGBoost** Extreme Gradient Boosting

# Chapter 1

# Introduction

## 1.1 Context and Motivation

Liver cancer is one of the most common cancers worldwide and the third leading cause of cancer-related deaths [1], with Hepatocellular Carcinoma (HCC) comprising around 90% of all cases [2]. The malignancy has a five-year survival rate of approximately 18% [1], with the figure rising to 70% for correctly detected and treated early-stage HCC [2]. Therefore, early detection of HCC is crucial for the improvement of patient outcomes.

Nevertheless, due to several challenges in screening and surveillance, most tumours continue to be diagnosed at a late-stage, and mortality rates continue to rise [3]. First and foremost, early HCC is asymptomatic by nature, leading to a substantial number of diagnoses arising from the accidental identification of a liver mass during unrelated procedures, or, once the disease has already progressed significantly, following the development of symptoms [2]. Secondly, recommended screening tests rely on ultrasound with or without Alpha-Fetoprotein (AFP) measurement, an HCC tumour marker. The sensitivity and specificity of these tests have been shown to be suboptimal, particularly in patients with cirrhosis or obesity, both of which are related to an increased incidence of HCC [1, 3]. Finally, HCC is known to share many imaging traits with benign liver lesions and non-hepatocellular malignancies, leading to difficulties within the differentiation of the distinct conditions and potential misdiagnosis [4, 5].

Despite accurate detection and curative-intent treatment of HCC, a five-year recurrence rate reaching up to 70% depending on the treatment option, is still of concern [6]. There are two types of recurrence affecting HCC patients, early or late recurrence, which develop within or after two years subsequent to treatment, respectively [6]. Early recurrence is related to tumour aggressiveness as it arises from the presence of invisible intrahepatic metastases. It is the most common type of HCC recurrence and is associated with more severe prognosis [6]. In contrast, late recurrence is related to the severity of the underlying liver disease as it arises from the development of a new and independent tumour [6]. Overall, assessing the risk of recurrence is valuable for patient monitoring and for the appropriate choice and timing of adjuvant treatment.

In recent years, there has been a rapid growth in the application of Artificial Intelligence (AI) within medicine [7, 8]. AI techniques can exploit vast quantities of patient data towards several tasks such as disease characterisation, treatment response prediction, and tumour detection. Machine Learning (ML) is a subset of AI that allows machines to extract knowledge and memorise patterns from large sets of data to learn and make autonomous decisions [9]. In the context of HCC lesion detection and segmentation, ML techniques, particularly models composed of Artificial Neural Networks (ANNs) trained on radiological images, have shown good performance [7]. While not intended to replace human experts, these tools can be used

to reduce the time and labour required to detect and accurately segment suspicious lesions. Not only advantageous for the diagnosis of HCC, lesion segmentation models are useful for several other purposes, such as lesion size assessment for the staging of liver tumours, treatment planning for targeted therapies, a prerequisite for tumour resection, and required for patient postoperative monitoring and care [10].

Radiomics is a process involving the high-throughput extraction of quantitative features from medical images, such as Computed Tomography (CT) scans. Radiomics can be used to discern and quantify tissue characteristics that would be otherwise invisible to a clinician's eye [11]. Coupled with ML, radiomics offers a promising solution to pre-operatively predict the risk of recurrence in HCC patients, in a non invasive and readily available manner.

## 1.2 Related Work

Given the added relevant information that radiomics may provide, several retrospective studies have evaluated its value towards the detection of microvascular invasion, diagnosis, and prognosis of HCC [12]. Of particular relevance to the present study, multiple authors have focused on the use of radiomics for prediction of HCC recurrence risk.

Ji, et al. [13] gathered data from three different institutions to develop models to predict HCC recurrence. The cohort comprised of 295 patients was divided into a train set from institution 1 (n=177) and an external test set from institutions 2 and 3 (n=118). All patients were affected by early stage HCC, were eligible for curative intent resection, and underwent contrast enhanced CT scans, with manual tumour delineations. In this study, radiomic features were extracted from the tumour and its periphery for subsequent radiomics-based model construction. Two models, one incorporating radiomic features alongside pre-operative clinical variables and one alongside post operative clinical variables, were developed via multivariate Cox regression analysis. The authors demonstrated that either model achieved better predictive performance (Concordance Index (C-index) $> 0.77$) than traditional staging systems or models lacking radiomic data.

Shan et al. [14] evaluated the prediction of early recurrence of HCC. The cohort was comprised of 156 patients and was split into a training set (n=109) and a test set (n=47). All patients were eligible for curative intent resection or ablation, and underwent contrast enhanced CT scans, with manual tumour segmentations. Two models, one comprised of tumour extracted radiomic features and one of peritumoural extracted features, were developed using logistic regression and compared. Despite full exclusion of the tumour area, the peritumoural model exhibited greater robustness and generalisability (Area Under the Receiver Operating Characteristic Curve (AUROC) $= 0.79$) compared to the tumour model, which overfitted on the test set.

Zhao et al. [15] also assessed the accurate prediction of HCC early recurrence by exploiting radiomic features. The cohort was composed of 113 patients split into a training set (n=78) and a test set (n=35). All patients were eligible for HCC resection and underwent multiparametric Magnetic Resonance Imaging (MRI), with manual tumour segmentation. Via multivariate logistic regression analysis, the authors developed three distinct models, one including radiomic features, one including clinical features, and one combining both feature types. Ultimately, the combined model (AUROC $= 0.87$) outperformed the remaining models in terms of discriminative ability.

Lastly, Gao et al. [16] sought to stratify patients at high risk of HCC early recurrence. The study included 472 patients, split into a training set (n=378) and a test set (n=94). All patients received curative intent resection and underwent multiphasic contrast enhanced MRI, with manual tumour segmentation. Three models were tested using logistic regression analysis. One model considered radiomic features only, one considered deep features extracted by passing the images through a Convolutional Neural Network (CNN) contracting path, and

one considered a combination of both feature types. Similar to previous studies, the combined model demonstrated improved accuracy when stratifying patients (AUROC = 0.84).

In summary, the reviewed studies suggest the value of radiomics within medicine and, more specifically, within the prediction of HCC recurrence risk. Furthermore, there is a general consensus, both in the above studies and many others [12], that combining radiomic features with clinical or other feature types, such as deep-learning features, leads to improved model performance. Similarly, radiomic features extracted from surrounding tissue rather than the tumour region alone, were revealed to contain valuable information and to enable the development of more robust models.

Nevertheless, these studies suffer from certain limitations. For starters, there is a notable lack of external test sets, that is testing the developed models on data originating from a different patient population than the train set. Only one study of the above included an external test set, and even in this case, all three institutions used the same CT scanner, which does not reflect real-world clinical conditions. Lacking an external validation set could, therefore, lead to an overestimation of model performance and applicability. Furthermore, while some radiomics studies [12, 17] utilise semi automatic methods to generate segmentations of the Region of Interest (ROI) for subsequent feature extraction, most studies, as described above, do not consider the time and manual labour required for this task.

Therefore, given the outlined research gaps and to further build upon the existing studies, the current study addresses the following questions:

- Provide and evaluate semi-automatic approaches for lesion detection and segmentation with the intention of reducing the time and effort radiologists invest on these tasks. This would be advantageous both for clinical practice and for the generation of ROI masks required for radiomics feature extraction.
- Quantify and compare the predictive value of radiomics in estimating HCC recurrence risk using different ROIs and a variety of different CT scanners.

## 1.3 Aim and Contributions

Building upon the outlined pursuits in section 1.2, this study aims to build an end-to-end pipeline for radiomics-based prediction of recurrence risk for HCC patients (Figure 1.1). The process involves semi-automatic tumour segmentation of HCC, expert correction of the generated lesion masks, radiomics feature extraction from the ROI, and prediction of recurrence risk using radiomics-based models.

As an initial step, two different lesion segmentation models were tested and compared. These were a lesion segmentation model built upon an nnU-Net framework [18] trained using CT images and tumour mask pairs of patients with confirmed HCC, and a pre-trained MedSam model [19]. Both models are described in greater detail further in the document (subsection 3.3.1, subsection 3.4.1). In addition to allowing for the semi-automation of the initial part of the pipeline, segmentation of the tumour lesion, the models may also assist towards increased reliability, and decreased time and effort when detecting the malignancy in clinical practice. The models were compared based on the accuracy of their predicted tumour segmentations.

The second step involved the development of radiomics-based HCC recurrence risk prediction models, and the assessment of their diagnostic performance. To this end, classification models were developed using three datasets, one composed of pre-operative clinical variables alone, one composed of tumour-extracted radiomic features alongside clinical variables, and one composed of whole-liver-extracted radiomic features alongside clinical variables. To train the models, radiomic features were extracted from the ground-truth lesion segmentations, or from predicted

whole liver segmentations, and subsequently combined with patient specific clinical data. A feature selection process was carried out for the identification and use of the most relevant and informative clinical and radiomic features. Finally, various classification models were explored and evaluated.

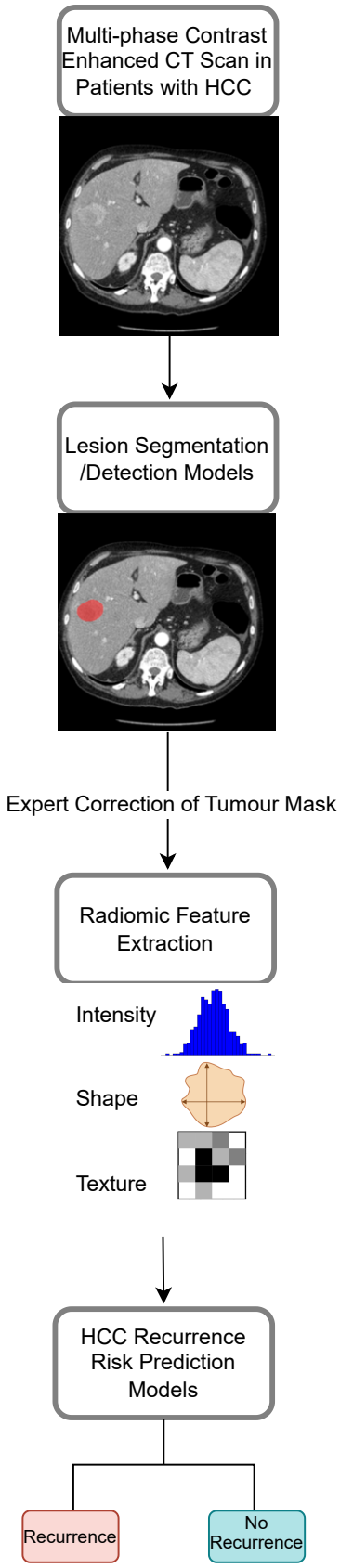



Figure 1.1: Pipeline for radiomics-based prediction of Hepatocellular Carcinoma recurrence risk.

# Chapter 2

# Radiologic Data

## 2.1 CT Imaging

### 2.1.1 Overview

CT is an imaging technique that provides high resolution visualisation of tissues and structures inside the body by exploiting the differential tissue absorption of x-rays. It was the first method to acquire such images in a non-invasive and superimposition-free manner, with its first clinical implementation dating back to 1972 [20]. Nowadays, CT is one of the most used techniques in medical imaging and is recommended for the diagnosis and staging of HCC [21]. The method will be described in further detail in the following subsections.

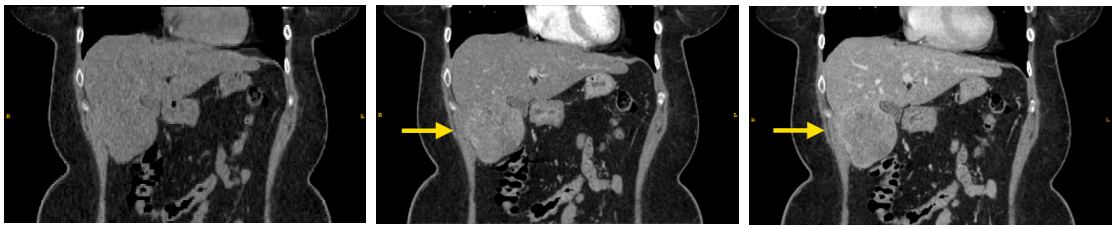
### 2.1.2 Image Acquisition

During CT scans, a motorised x-ray source quickly rotates around a patient's body. The aimed x-rays pass through the patient and produce signals that are detected on the opposite side of the x-ray source. Since different tissue densities result in different levels of x-ray attenuation, the measured signals allow for the construction of an image where each pixel displays the location's attenuation value, usually depicted as grey values [20]. The procedure results in projected Two-Dimensional (2D) slices of the body viewed in different planes that can then be computationally stacked to assemble a Three-Dimensional (3D) reconstruction [22]. The attenuation values are represented on the Hounsfield scale in Hounsfield Units (HU), where the minimum value of -1000 HU is assigned to air, 0 HU to water, and some tissues reaching +3000 HU [20].

### 2.1.3 Contrast-Enhanced Liver CT

Depending on the target tissue, the density difference between a tumour and surrounding tissue is enough to easily distinguish and detect the lesion using standard CT practices. For example, bone is denser than tumours, while the opposite is true for air within the lungs. However, this is not the case for soft tissue organs, such as the liver, where the tumour attenuation is highly similar to that of the liver parenchyma [22, 23]. Therefore, to increase the difference in attenuation between tumour and background tissue, a supplementary intravenous administration of contrast material is routinely used. Nevertheless, for the technique to achieve the desired discrepancy between attenuation values, the contrast substance should predominantly reach either the tumour or the healthy liver tissue, not both simultaneously. This condition can be fulfilled due to the particular dual blood supply of the liver [23]. The liver receives blood from the hepatic artery (20-25% of flow) and hepatic portal vein (75-80% of flow), with the same injected material reaching the liver at two different times due to the distinct circulatory paths of the blood vessels. Additionally, hypervascular tumours, such as HCC, only receive arterial

blood flow, and are not supplied by the portal venous blood flow [23]. As such, during the initial Arterial Phase (AP), HCC will receive contrast material, and present with higher enhancement than the liver parenchyma, as the bulk of the contrast enhanced blood flow, coming from the portal vein, will not yet have arrived (Figure 2.1b). On the other hand, during the Portal Venous Phase (PVP), approximately 30 seconds later, the tumour will present with washout due to the increased proportion of contrast agent in the surrounding liver tissue (Figure 2.1c).



(a) No contrast (b) AP contrast (c) PVP contrast

Figure 2.1: Liver computed tomography scans. Each arrow points to visible hepatocellular carcinoma.

## 2.2 Radiomics

As previously mentioned, radiomics involves the automatic extraction of numeric features from a ROI of a medical image, such as a tumour or organ [24]. Typically, hundreds or thousands of features are extracted from a single ROI, allowing the data to be mined for the discovery of new patterns and biomarkers for disease progression monitoring, disease characterisation, or the prediction of clinical outcomes, such as treatment response and overall survival [25]. Radiomic features can be mainly divided into shape, statistical, and texture based features [26].

Shape features relate to the size and geometry of the ROI, for example its maximum diameter, volume, compactness, or sphericity. Statistical features include both histogram based and texture based features. First order features are based on properties of the grey values histogram of individual voxels. These characteristics include the mean, maximum, standard deviation, skewness, or kurtosis, the "tailedness", of the grey level distribution within the ROI. Texture based or higher order features, unlike first order features, also consider the spatial relations between voxels. Additionally, different filters can be applied to the original image for the supplementary calculation of all statistical features on a transformed space. Transformation examples include the application of gaussian, square, or exponential filters, for example.

Formerly, one of the main limitations of radiomics was the lack of consensus and reproducibility between different software's definitions and protocols for feature extraction. The issue led to the development of Image Biomarker Standardization Initiative (IBSI) [27] standards for radiomics studies. In this study, *PyRadiomics* [26], an open source python package adhering to IBSI standards, was used for the extraction of radiomic features.

# Chapter 3

# Deep Learning and Computer Vision for the Medical Domain

## 3.1 Overview

Deep learning is a subset of machine learning, that utilizes multilayer ANNs to learn complex patterns from large subsets of data. As previously mentioned, deep learning is the core component of state of the art methods within medical tasks. In this section, ANNs, the foundation of deep learning methods are firstly introduced, followed by the two types of model architecture explored in this study, CNNs and Vision Transformers (ViTs), alongside their specific applications towards segmentation within the medical field.

## 3.2 Artificial Neural Networks

ANNs are computational models inspired by the biological neural networks found in the brain. Artificial neurons receive input, transform it using activation functions, and transmit the output to subsequent neurons. ANNs include an input layer, an output layer, and intermediate hidden layers composed of multiple interconnected stacks of artificial neurons. By adjusting parameters such as weights and biases, the network is able to improve its performance on a specific task.

Activation functions are mathematical functions needed to introduce non linearity to ANNs. The functions are applied to the output of each individual neuron within the network's hidden layers, allowing for the computation and extraction of increasingly complex and abstract features.

One of the most common training schemes of machine learning and deep learning is supervised learning, where labels or target values are provided alongside the input data. In supervised deep learning the computational model is trained with the goal of minimising the difference between the ground truth labels and the model predicted labels. This difference, the loss, is quantified by the loss function. A common loss function used for classification tasks is cross-entropy loss, which measures the difference between the predicted class probabilities and the true labels.

During the training of an ANN, the model's parameters are iteratively adjusted, using an optimisation algorithm, to minimise the loss and improve model performance. The iteration process continues for a predefined number of iterations or when a target performance is reached. Optimisation algorithms are required for parameter updates. Firstly, the network's parameters are commonly initialised at random. Subsequently, optimisers are employed to guide the direction and magnitude of parameter modification, according to a set of hyperparameters. For example, the learning rate controls the step size of parameter updates. A learning rate that is too small will require too many iterations to find a suitable set of parameters. A learning

rate that is too large may be unable to converge towards an optimal solution. Optimisers are therefore crucial to minimise the loss function, while balancing convergence speed and stability, and solution accuracy.

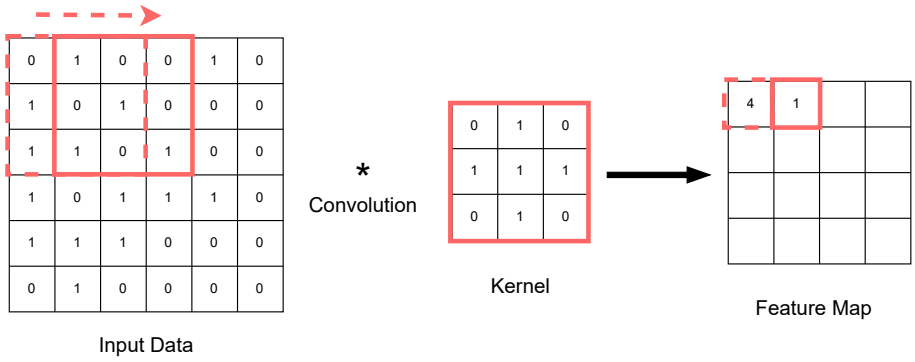
## 3.3 Convolutional Neural Networks

CNNs are a particular type of ANN architecture widely used in pattern recognition tasks requiring imaging data. A CNN's architecture is mainly comprised of two types of stacked layers: convolutional, and pooling layers.

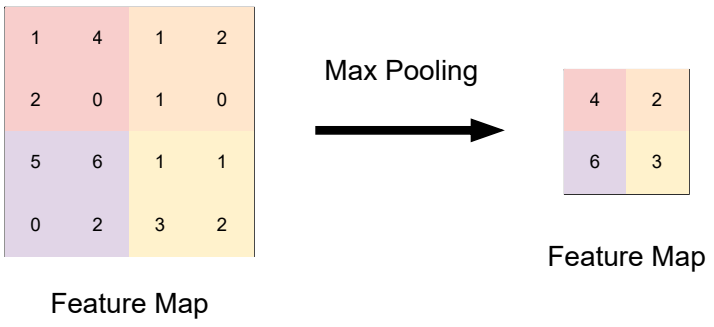
Convolutional operations involve the sliding of a convolutional kernel, or filter, over the image matrix and computing a local weighted sum, representing a single value in the resulting feature map (Figure 3.1a). The filter's weights are learned parameters that are tuned as training progresses but remain constant across different locations of the image matrix. This shared weight characteristic is essential for the network to learn patterns regardless of their spatial location while reducing the number of model parameters [28].

Pooling layers serve to downsample and reduce the dimension of the previous layer. Moreover, pooling improves generalisation of pattern recognition by retaining only the most essential values and broad location of the learned feature [28]. Typical operations include max pooling, computing the maximum value of a region (Figure 3.1b), and, with the same rationale, average pooling.

The alternating process of feature extraction and dimensionality reduction, by convolutional and pooling layers, respectively, render CNNs particularly suited for imaging, array like, data. Additionally, it equips the networks to learn hierarchical visual features, starting with edges and textures and progressing to increasingly complex and abstract patterns.




(a) Convolution




(b) Maximum pooling

Figure 3.1: Operations within Convolutional Neural Networks.

An exemplary CNN is a U-Net, named after its U-like shape, with an architecture developed for biomedical imaging segmentation tasks [29]. Unlike image classification tasks, where the output is a single class label, in segmentation tasks the output involves localisation, where a class label is assigned to each pixel of the input image. The advantage of U-net is that it is able to predict detailed segmentation masks using limited numbers of labelled data, a regular constraint of medical imaging data. The network's architecture consists of a contracting path and an expansive path (Figure 3.2). The contracting path, much like a typical CNN, acts as an encoder, performing downsampling of the input while increasing the number of feature maps to capture contextual information. The expanding path, acts as a decoder, enabling the output and input to have equal dimensions via upsampling. Skip connections concatenate the feature maps of the contracting path to the expanding path, so the learned information is retained, and precise localisation facilitated.

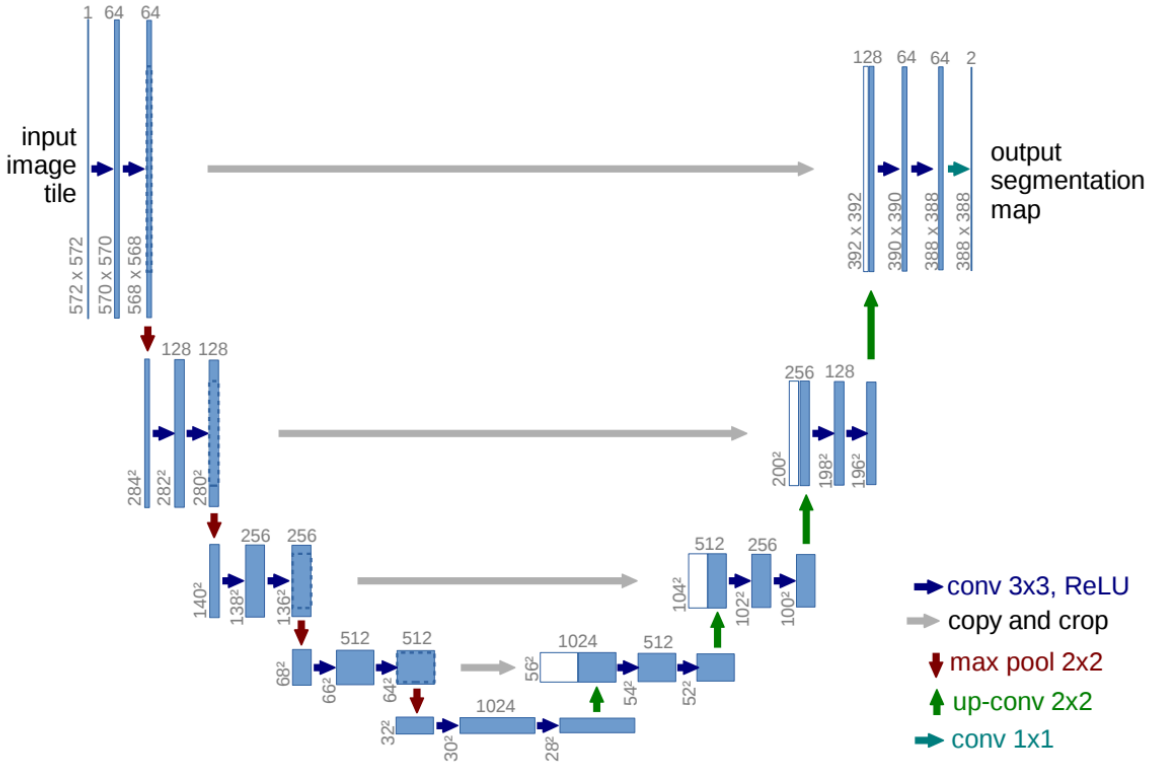



Figure 3.2: 2-dimensional U-net architecture. Boxes represent a multi-channel feature map, with shape given at the box's lower left, and number of channels given at the top. Image taken from Ronneberger, et al [29].

### 3.3.1 nnU-Net

"No new U net" (nnU-Net) is an automatic biomedical image segmentation tool that builds upon the established U net architecture [18]. The tool's value lies in its ability to automatically adapt to new datasets by systematic pipeline optimisation. A pipeline fingerprint conducts automatic design choices, categorised as fixed parameters, inferred rules, or empirical decisions. Fixed and empirical decisions include the fixed selection of a U-Net architecture, the tuning of hyperparameters via cross validation of the training data, or the empirical choice of best performing model or ensemble. Inferred decisions include data augmentation and pre-processing automatically applied according to a dataset's computed fingerprint, which includes image properties such as size, voxel spacing, and modality. Specialist nnU-Net models have shown great performance across a variety of datasets and tasks [18], including liver tumour segmentation,

and are one of the most widely used tools for medical imaging segmentation.

## 3.4 Vision Transformers

Transformer models, a class of deep learning architectures, were firstly developed for natural language processing tasks [30], where they quickly displayed remarkable performance. This success inspired the application of transformers in computer vision tasks, leading to the development of Vision Transformers (ViTs) [31] and their widespread use within medical image segmentation [32].

The main idea behind transformer models is the use of self-attention mechanisms, which allows models to weigh the importance to different elements of a sequence and learn relationships between them. Rather than processing text as a series of tokens, ViTs process images as a series of patches that are flattened and combined with positional encodings to capture spatial relationships. The linear embeddings are then passed through multiple self-attention and feed forward neural network layers. Due to the simple architecture of transformers, the models are highly scalable and efficient. The architecture of a vision transformer is represented in Figure 3.3.

A second key aspect of the models is that of pre-training on large datasets in a supervised or semi supervised manner and subsequent fine tuning of the model using smaller labelled datasets for specific tasks [33].

For example, Segment Anything Model (SAM) is a transformer based promptable image segmentation model, by Meta, trained on over 1 billion masks and 11 million images [34]. It is composed of an image encoder, a prompt encoder, and a mask decoder. The image encoder embeds the input image, and the prompt encoder embeds the user-defined prompt (in the form of text, bounding box, or point). Both inputs are combined and passed through the mask decoder, which outputs the desired segmentation mask.

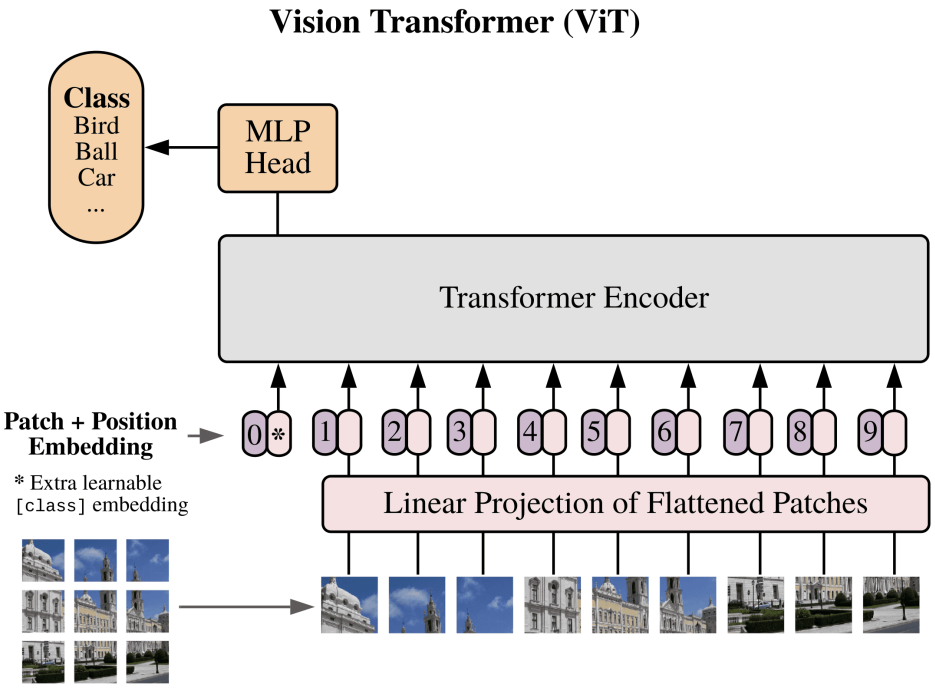



Figure 3.3: Vision transformer architecture. Image taken from Dosovitskiy, et al. [31]

### 3.4.1 MedSam

MedSam is a transformer-based foundation model for medical imaging segmentation [19]. Unlike nnU-net, where a specialist model is typically newly designed and trained for a specific dataset and segmentation task, MedSam aims to be trained once and be able to generalise to a wide range of tasks and imaging modalities. To this end, MedSam was initialised with SAM's architecture and pre-trained weights, and fine tuned on 1 million publicly available image mask pairs covering over 30 cancer types and 15 imaging modalities. MedSam's architecture and its pipeline are represented in Figure 3.4

On certain segmentation tasks, MedSam has been demonstrated to be comparable or even surpass specialist models, such as nnU-Net. However, one of its main limitations is that it requires expert input, commonly in the form of a user defined bounding box. Therefore, while it is not suitable as a detection tool or as a fully automatic segmentation tool [19], it holds significant potential and value for the acceleration of fine segmentation of lesions within clinical practice.

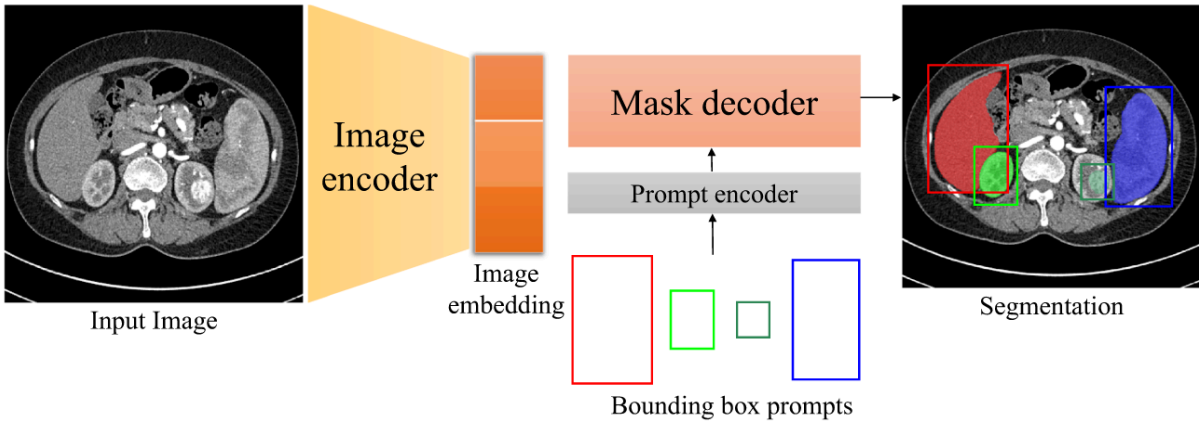



Figure 3.4: MedSam architecture. Image taken from Ma, et al. [19]

# Chapter 4

# Materials and Methods

## 4.1 Datasets' Characteristics

### 4.1.1 Dataset 1 (Beaujon)

The retrospective patient cohort used for training and testing of the models was provided by Beaujon Hospital, Clichy, France, and was comprised of 230 individuals with confirmed HCC who underwent curative intent resection. All patients had no more than two resected HCCs. For each patient, clinical data along with one presurgical CT scan with three phases after injection, acquired between 2011 and 2019, were available. For each of the contrast enhanced scans, AP and PVP contrast, ground truth tumour segmentation masks were provided by one of two abdominal radiologists from the institution, with six and ten years of experience. Throughout the study, different patient data and subsets of the total patient cohort were employed (Figure 4.1).

For training of the lesion segmentation models, scan mask pairs that were incorrect, missing, or poorly acquired were excluded. This was the case for a small quantity of patients, for a final total of 225 and 228 valid cases of AP and PVP mask scan pairs, respectively. The AP dataset was split into 180 patients for training and 45 for testing. The PVP model was split into 182 patients for training and 46 for testing. During data partitioning, stratification by two tumour diameter categories was ensured, with 27% of tumours less than 3 cm and 73% of tumours greater than or equal to 3 cm. The decision to use a threshold of 3 cm was implemented based on Barcelona Clinic Liver Cancer (BCLC) staging guidelines, which state curative treatment for HCC patients is available for single tumours or up to 3 nodules of less than 3 cm each [35].

For training of the radiomics-based classification models, an institutional data partitioning was employed. While most scans were acquired at Beaujon Hospital using one of three CT scanner models, around 30% of scans were acquired across multiple other institutions and belonged to one of 24 scanner models (Figure 5.3). Since, this study endeavours to test the generalisability of radiomics-based models to images acquired using different CT scanners, scans acquired at Beaujon hospital were allocated to the train set, while scans from other locations were assigned to the test set. Comparable proportions of recurrence type in train and test sets were confirmed ( Table A.5).

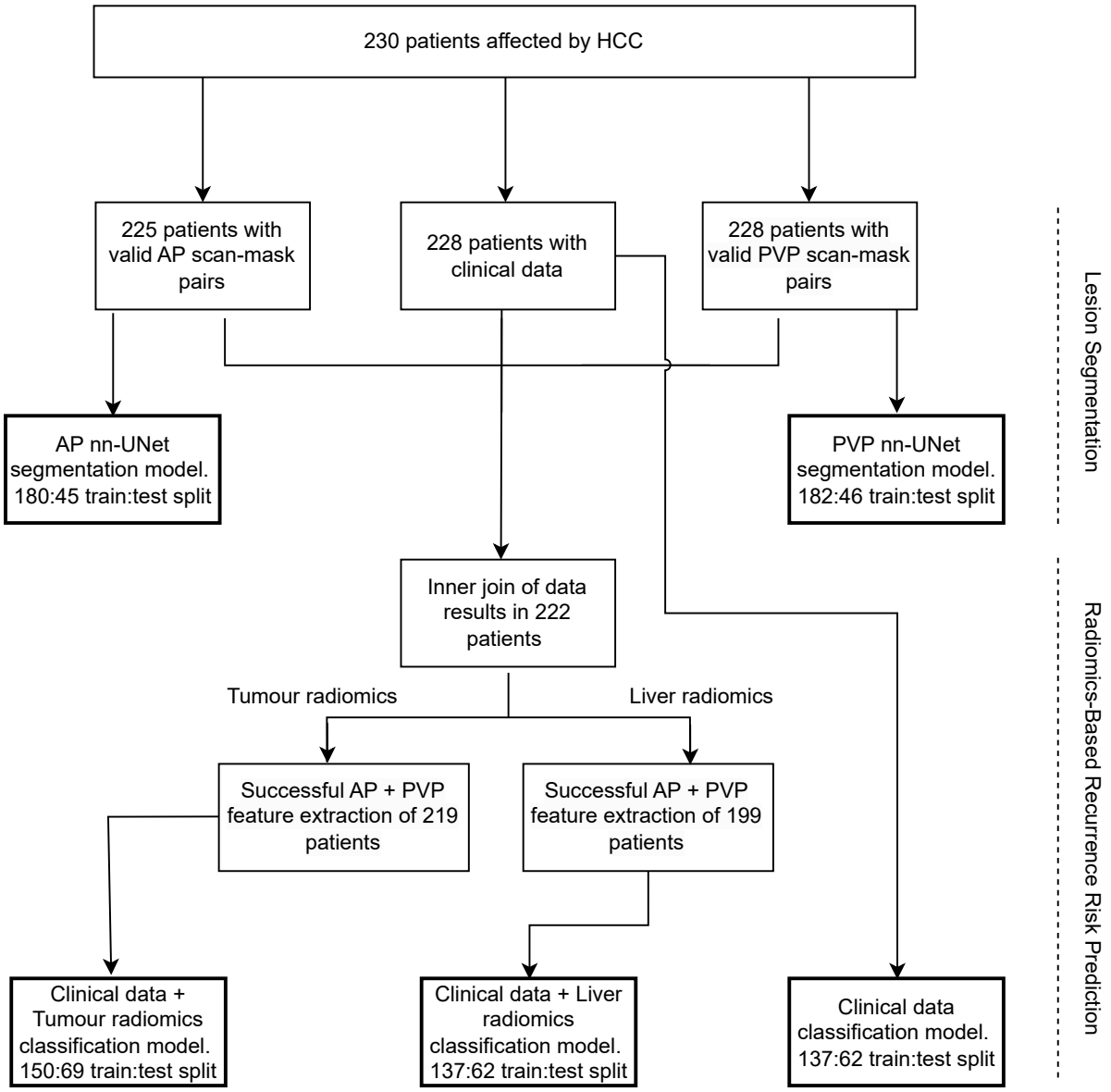



Figure 4.1: Subsets of Beaujon patient cohort used throughout study, along with respective train-test data splits. Trained models are represented in bold boxes. HCC: Hepatocellular Carcinoma; AP: Arterial Phase; PVP: Portal Venous Phase.

### 4.1.2 Dataset 2 (TCIA)

An additional dataset was used for external validation of the developed models. The dataset [36] is publicly available and was collected from The Cancer Imaging Archive (TCIA) [37]. It was composed of 105 individuals affected by unresectable HCC who underwent Transarterial Chemoembolisation (TACE) treatment. For each patient, a CT scan with three phases after injection, acquired between 1998 and 2006, are available. Although both pre- and post-TACE treatment scans are included in the dataset, only pre-procedural AP and PVP CT scans were utilised in this study. Unlike the first dataset, only one tumour segmentation mask was provided per patient in the TCIA dataset. According to the metadata within the tumour mask file and scan file, scan-mask pairs that did not perfectly match with respect to size, orientation, origin, and spacing information were excluded. A total of 33 and 53 cases remained of AP and PVP scan-mask pairs, respectively. Finally, although various clinical characteristics were available, future recurrence of patients was not.

## 4.2 Lesion Segmentation and Detection

### 4.2.1 Manual Pre-Processing and Data Preparation

To accelerate training, the Beaujon dataset's scans and masks were manually cropped to approximately 5 cm above and below the liver, thus removing the unimportant portion of the image. Thereafter, as was the case for the external TCIA dataset, rather than manual cropping, automatic cropping to the aforementioned range was implemented using predicted liver masks outputted by *TotalSegmentator* [38], an nnU-Net based organ segmentation tool for CT images.

Manual cropping and visualisation of the scan mask pairs was conducted using *ITK SNAP*, a software application for the visualisation and delineation of anatomical structures [39]. Corresponding image properties between scans and masks were confirmed using *simpleITK*, a Python medical imaging processing toolkit [40].

### 4.2.2 nnU-Net Model Specification

As previously mentioned, two distinct tumour masks were provided per patient within the Beaujon dataset. This provides the correct delineation of the tumour at both stages, as it accounts for patient movement between the acquisition of the different CT phases. For this reason, two distinct lesion segmentation models were trained, one on AP CT images and the other on PVP CT images.

Both deep learning models were established upon the nnU-Net framework, using *PyTorch* a Python deep learning library [41]. Pre-processing was automatically conducted by nnU-Net based on the dataset's fingerprint (image size, image spacing, channel input, intensity values of voxels, number of training cases, etc). A CT-specific intensity normalisation scheme was used for both models. The scheme clips all images to the global 0.5% and 99.5% percentiles of the intensity values and normalises all images by subtracting the global mean and dividing by the global standard deviation. Further pre-processing included resampling all images to the median spacings of $1.25 \times 0.78 \times 0.78$ mm and $1.25 \times 0.79 \times 0.79$ mm for the AP and PVP models, respectively. The models were implemented using nnU-Net's 3D full resolution configuration, which involves training the networks for 1000 epochs, with 250 mini-batches per epoch. Stochastic gradient descent was used with Nesterov momentum of 0.99, a starting learning rate of 0.01, and polynomial learning rate decayed. The loss function is the sum of cross-entropy and dice loss.

### 4.2.3 nnU-Net Predicted Mask Post-Processing

Post-processing of the model predicted tumour masks was performed by removing regions outside of the liver. Liver masks were produced by *TotalSegmentator* [38]. Connected regions of the lesion array were separated using *scikit-image*, a Python library for imaging processing [42]. Overlap between whole-liver and lesion regions was calculated using the Dice Similarity Coefficient (DSC), described in detail in the subsection 4.2.5.

### 4.2.4 MedSam Model Specification

The MedSam model, originally trained on the TCIA dataset, was tested on the Beaujon hold-out test set. As MedSam requires expert input, commonly in the form of a user-defined bounding box, the exact bounding box of each tumour region was computed using *scikit-image* [42] and passed as input to the model for subsequent mask inference. This step was performed by José Almeida, a Postdoctoral Researcher at Champalimaud Foundation.

### 4.2.5 Lesion Segmentation Model Evaluation

The trained nnU-Net models were evaluated using 5-fold cross validation, an internal hold out test, and an external test set (Figure 4.2). The MedSam model was evaluated using the Beaujon dataset, as the model was originally trained on many datasets, including the TCIA dataset (Figure 4.2). To reiterate, the TCIA testing data included 33 arterial phase and 53 portal venous phase tumour-mask pairs. The Beaujon training data comprised of 180 arterial phase and 182 portal venous phase tumour-mask pairs. The Beaujon testing data included 45 arterial phase and 46 portal venous phase tumour-mask pairs. Several metrics were calculated on the test sets to evaluate model performance in segmenting and detecting HCC. These included overlap, distance-based, and volume-based metrics, in addition to a measure of successful detection.

The DSC is one of the most used metrics in assessing pairs of medical segmentations [43]. It measures the overlap between the original and the model predicted segmentations. The DSC ranges from 0 to 1, with 1 indicating a perfect predicted segmentation and zero indicating no overlap.

The Hausdorff Distance (HD) measures the spatial distance, in millimetres, between the ground truth and predicted mask volumes [43]. It is defined as the maximum Euclidean distance between the surfaces of both objects. An HD of 0 indicates a perfect segmentation.

The Average Symmetric Surface Distance (ASSD) measures the spatial distance, in millimetres, between the surface voxels of the ground truth and predicted mask [44]. It is calculated by measuring the Euclidean distance between the nearest neighbour of each surface voxel of the former segmentation to the surface voxels of the latter, and vice versa. The final metric is defined as the average of all measurements. An ASSD of 0 indicates a perfect segmentation.

The Relative Absolute Volume Difference (RAVD) measures the volume difference, as a percentage, between a set of two segmentations [44]. The value ranges from -1 to infinity, with 0 denoting two segmentations of equal volume. The RAVD provides no indication on the parity of segmentations or of the overlap between them. However, alongside the previous metrics, it does provide information on whether an object was under or over segmented.

The recall was used as a lesion detection metric. In this study, a true positive prediction was defined as a predicted segmentation having a DSC of greater than 0.5 with the ground truth segmentation. Since all patients were affected by HCC, there were no true negative cases. Therefore, the recall was defined as the ratio of true positives to the total number of patients in the test set.

All metrics, excluding recall, were calculated using *MedPy*, a Python library for medical imaging analysis [45].

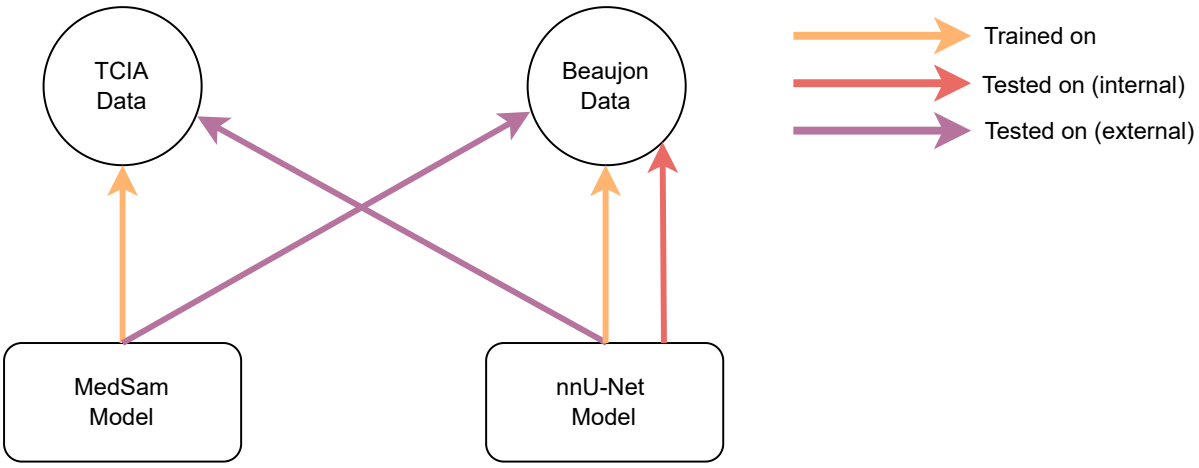



Figure 4.2: Diagram of lesion segmentation models with respective train and test sets.

### 4.2.6 Clinical Variables Analysis

To provide insight into the association between clinical variables and model performance, model predicted segmentations were divided into true positives (DSC $\geq 0.5$) and false negatives (DSC $< 0.5$). Comparisons between both groups were conducted for all available clinical variables. The analysis was performed, in the same manner, to the AP and PVP models independently.

### 4.2.7 Statistical Analysis

Throughout this study, several statistical tests were employed. To compare proportions between two independent groups, a z-test was employed. To assess differences in the distribution of continuous variables between two independent groups, a Mann-Whitney U test was chosen, as it is appropriate for non-normal distributions, a common characteristic of clinical variables such as biomarkers concentrations or of performance metrics of tumour segmentation models. For the same reason, the non-parametric paired sum rank test was used to test differences in the distribution of paired observations, particularly for assessing differences in segmentation models performance tested on the same data. Finally, to test the independence of categorical and continuous variables across two groups, a chi-squared test of independence or a Mann-Whitney U test were employed, respectively. For all statistical tests, a p-value $< 0.05$ was established as statistically significant. Analyses were conducted with Python package *Scipy* [46].

The median Recurrence Free Survival (RFS) of the Beaujon patient cohort was calculated using a Kaplan-Meier estimator.

## 4.3 Radiomics-based Prediction of Recurrence Risk

### 4.3.1 Explored Regions of Interest

To further explore and validate the value of radiomics in predicting HCC recurrence risk, three datasets were tested. To construct the different datasets, different ROIs were used for the extraction of radiomic features. The first dataset, with one of the most commonly used data combination, involved the extraction of radiomic features using the tumour as the 3D ROI alongside clinical variables. However, since recurrence is related to occult lesions in the liver

or to a patient's underlying liver condition, a second data combination, using the extraction of radiomic features using the whole liver as the 3D ROI alongside clinical variables, was also included in this study. A final dataset, using solely clinical variables, was employed to establish the baseline performance. Therefore, in total, three different datasets were considered:

- **Clinical**: Patient-specific clinical characteristics;
- **Tumour**: Tumour-extracted radiomic features alongside clinical characteristics;
- **Liver**: Whole-liver-extracted radiomic features alongside clinical characteristics;

### 4.3.2 Feature Extraction

All pre-processing and feature extraction steps were performed using *PyRadiomics*. For the tumour-ROI, the available ground-truth tumour segmentation masks were used. For the liver-ROI, the liver segmentation masks produced by *TotalSegmentator* were employed.

As an initial step, images were resampled to an isotropic voxel spacing of 1 mm. Subsequently, several classes of features were extracted from the original image and from transformed images after filter application. The details of extracted feature classes and applied filters can be found in Table 4.1.

Table 4.1: List of extracted radiomic features and applied image filters.

| **Feature Classes** | **Image Filter** |
|---|---|
| • First Order Statistics,<br>• Shape based 3D<br>• Gray Level Co occurrence Matrix<br>• Gray Level Run Length Matrix<br>• Gray Level Size Zone Matrix<br>• Neighbouring Gray Tone Difference Matrix<br>• Gray Level Dependence Matrix | • Wavelet<br>• Laplacian of Gaussian with sigma of 1 mm, 2 mm, 3 mm, 4 mm, and 5 mm<br>• Square<br>• Square Root<br>• Logarithm<br>• Exponential<br>• Gradient,<br>• Local Binary Pattern in 3D |

For First Order Statistics, grey values were discretised according to a fixed bin width, rather than a fixed bin number, as this approach has been shown to improve feature reproducibility in quantitative imaging modalities [47]. A bin width where the resulting number of bins falls between 16 and 128 has been demonstrated to have good reproducibility and performance [48]. Therefore, for each image type, the optimal bin width was calculated by minimising the number of training set patients with a bin number falling outside the ideal range. Furthermore, as -1000 is the minimum HU, a voxel array shift of +1000 was applied to prevent squaring of negative values when deriving First Order Statistics.

The same steps were carried out for both AP, and PVP scans, independently. In total, the combination of clinical data, and radiomic features extracted from AP and PVP CT scans, resulted in over 4000 features per patient.

### 4.3.3 Feature Pre-Processing and Selection

Due to the large number of variables, feature selection was carried out for retainment of the most informative features and exclusion of irrelevant features that could otherwise introduce noise into the model. This step is crucial for dimensionality reduction and helps prevent overfitting, thus promoting model robustness to unseen data.

As a first step, an iterative imputer was employed to deal with missing clinical data. Subsequently, unsupervised feature selection was performed. The process involved removal of zero variance features, and removal of highly correlated features (Spearman correlation $> 0.8$), where for each highly correlated feature pair, the feature with highest overall correlation with other variables was removed. Finally, features were standardized by subtracting the mean and dividing by the standard deviation.

Supervised feature selection was also conducted, using Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression, with the regularisation parameter ($\lambda$) tuned via 5-fold cross validation. A stability analysis constructed from 1000 LASSO runs, using the best performing parameter, was exercised to select the final set of highly stable features, that is, features with non-zero coefficients in over 90% of runs.

The test set was imputed, scaled, and reduced according to the transformations applied to the training set.

### 4.3.4 Radiomics-Based Model Evaluation

All models were evaluated using 5-fold cross-validation and an internal hold-out test set.

Several metrics exist to quantify the performance of classification models, and for the most part, can be derived from a confusion matrix. A confusion matrix enables the visualisation of all scenarios of correct and incorrect classifications (Table 4.2). A True Positive (TP) occurs when the model correctly predicts a positive label. A False Positive (FP) occurs when the model incorrectly predicts a positive label. A True Negative (TN) occurs when the model correctly predicts a negative label. A False Negative (FN) occurs when the model incorrectly predicts a negative label.

Table 4.2: Confusion matrix.

| | | **True value** | |
|---|---|---|---|
| | | Positive | Negative |
| **Predicted value** | Positive | $TP$ | $FP$ |
| | Negative | $FN$ | $TN$ |

TP: true positive; FP: false positive; FN: false negative; TN: true negative.

From the confusion matrix, precision and recall can be calculated. Precision (Equation 4.1) is the proportion of true positives out of all positive predictions, namely the accuracy of positive predictions. Recall or sensitivity (Equation 4.2) is the proportion of true positives out of all positive cases, in other words, the completeness of positive predictions. While high recall ensures that the majority of positive events are correctly identified, high precision guarantees the identified positive events remain correctly classified. A balance of both, is therefore required, especially within medical applications, like estimation of recurrence risk. In such scenarios, a false negative is equivalent to a patient being characterised as having low risk of future recurrence when, in fact, they are at high risk. The circumstance could result in inadequate preventive measures or monitoring, potentially leading to worsened patient outcomes. On the other hand, a false positive incorrectly indicates a high risk of recurrence, which could result in unnecessary adjuvant treatment, monitoring, and improper resource allocation.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4.1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4.2}$$

The F1 score quantifies the balance between precision and recall, being defined as the harmonic mean of both metrics, as shown in Equation 4.3. Thus, maximisation of F1 score ensures the reduction of both false negatives and false positives.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.3}$$

The AUROC (Area Under the Receiver Operating Characteristic Curve) measures a model's ability to separate positive and negative classes. It is calculated by plotting the true positive rate, the sensitivity or recall (Equation 4.2), against the false positive rate (Equation 4.4), the proportion of false positives out of all actual negative cases, for different classification thresholds. The area under the plotted curve is then calculated with 1 indicating a perfect classifier, and 0.5 indicating a random classifier. Therefore, while the AUROC is useful to determine a model's discriminatory ability, it does not consider the required balance between precision and recall.

$$False\ Positive\ Rate = \frac{False\ Positives}{True\ Negatives + False\ Positives} \tag{4.4}$$

In this study, both F1 score and AUROC are reported, with models' hyperparameters tuned to maximise F1 score, due to the previously described clinical considerations.

### 4.3.5 Radiomics-Based Model Specification

Several different classification models, of increasing complexity were tested. These included linear, non-linear, and ensemble models. The employed classifiers were:

- Logistic regression
- K-Nearest Neighbours (KNN)
- Support Vector Machine (SVM)
- Decision tree
- Random forest
- Extreme Gradient Boosting (XGBoost)

For each dataset, model hyperparameters were tuned according to an exhaustive grid search with 5-fold cross-validation, with the best model defined as that which maximised F1 score (Equation 4.3). Details of the explored and chosen hyperparameters can be observed in Table 4.3. All models were implemented using the Python package *Scikit-learn* [43].

Table 4.3: Hyperparameter search of classification models. The same search was carried out for each dataset (clinical, tumour, liver).

| **Classifier** | **Parameter** | **Search Space** | **Chosen Parameter (clinical, tumour, liver)** | | |
|---|---|---|---|---|---|
| Logistic Regression | C | np.logspace(-3, 5) | 0.28 | 8.29 | 2.68 |
| K-Nearest Neighbours | # neighbours | np.arange(1, 31) | 1 | 3 | 3 |
| Support Vector Machine | C | np.logspace(-3, 5) | 12.01 | 1.84 | 12.07 |
| Decision Tree | Max depth | [5, 10, 20, 30, 40, 50] | 10 | 10 | 5 |
| Random Forest | Max depth<br># estimators | [5, 10, 20, 30, 40, 50]<br>[50, 100, 200, 500, 1000] | 20<br>1000 | 10<br>100 | 10<br>500 |
| XGBoost | Max depth<br># estimators | [5, 10, 20, 30, 40, 50]<br>[50, 100, 200, 500, 1000] | | | |

C: inverse of regularisation strength; np.logspace(-3, 5): 50 evenly spaced numbers on the log scale between $10^{-3}$ and $10^{5}$, np.arange(1, 31): evenly spaced integers in the range [1, 31[ with step size 1.

# Chapter 5

# Results

## 5.1 Data

### 5.1.1 Patient Characteristics

The Beaujon patient cohort consisted of 180 men and 49 women with mean ± std. age of 62.3 ± 12.7 years. The TCIA patient cohort consisted of 37 men and 21 women with mean ± std. age of 69.0 ± 12.0 years. As previously mentioned, Beaujon cohort patients are eligible for curative-intent resection while TCIA cohort patients are eligible for TACE treatment, intended for palliative care. These differences in disease burden and progression can be observed by the respective patient characteristics (Table 5.1). Compared to Beaujon, the TCIA cohort has, overall, increased proportions of liver disease, increased AFP, larger tumours, and a higher occurrence of multinodular tumours. The full list of Beaujon and TCIA patient characteristics, can be found in the supplementary materials (Table A.1, Table A.2, and Table A.3).

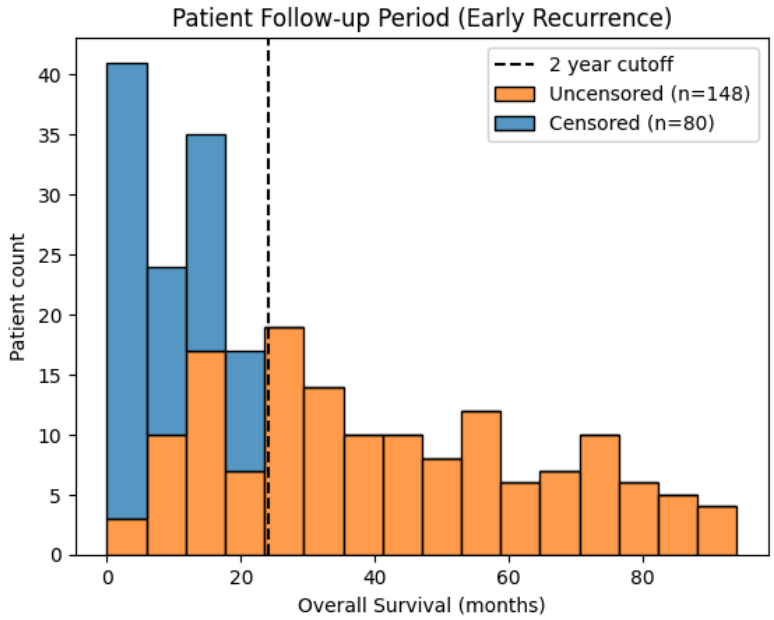
Table 5.1: Patient characteristics of Beaujon and TCIA cohorts.

| **Characteristics** | **Beaujon Patients** | **TCIA Patients** | **p-value** |
|---|---|---|---|
| # Patient Scans | 225 Arterial, 228 Venous | 33 Arterial, 53 Venous | |
| *Demographic Features* | | | |
| Age (years) | 64.0 (55.0, 70.0) | 70.0 (58.0, 76.0) | **0.006** |
| Sex (Men) | 180 (78.6%) | 37 (63.8%) | **0.029** |
| *Chronic Liver Disease* | | | |
| Hepatitis B | 61 (26.7%) | 18 (31.0%) | 0.616 |
| Hepatitis C | 56 (24.5%) | 27 (46.6%) | **0.002** |
| Alcohol Abuse | 41 (18.0%) | 32 (55.2%) | **<0.001** |
| Cirrhosis | 74 (32.5%) | 45 (77.6%) | **<0.001** |
| $\alpha$-Fetoprotein (ng/mL) | 8.0 (3.6, 167.8) | 41.4 (7.2, 2143.5) | **0.001** |
| *Histopathological Characteristics* | | | |
| Tumour Diameter (cm) | 4.1 (2.8, 7.3) | 6.0 (4.0, 10.0) | **<0.001** |
| Uninodular Tumour | 218 (95.2%) | 27 (46.6%) | **<0.001** |

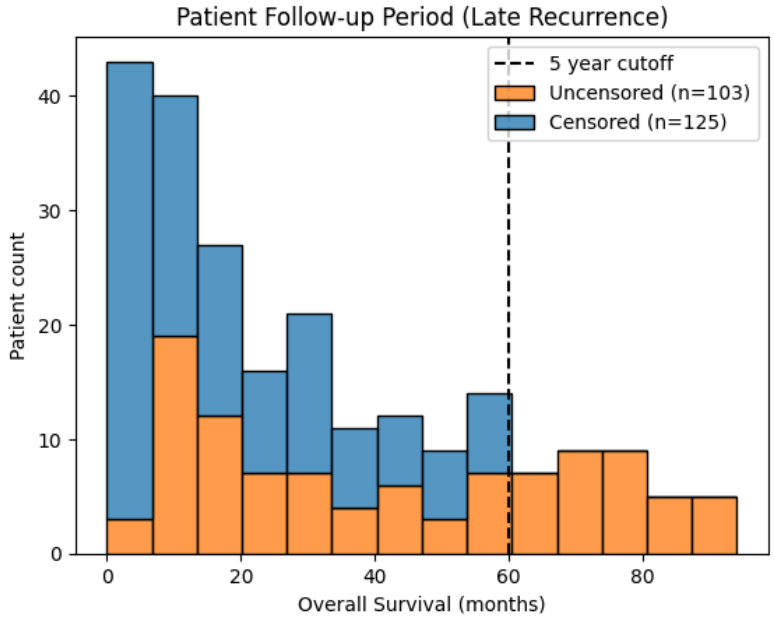
Continuous variables are displayed as medians and 25% and 75% percentiles in parathesis, with p-values calculated for a Mann-Whitney U test. Discrete variables are displayed as count and percentages in paranthesis, with p-values calculated for a z-test.

### 5.1.2 Beaujon Patient Cohort Follow-up, Scanner Models, and Recurrence Rates

As previously mentioned, Beaujon data was comprised of 230 patients in total. However, due to missing recurrence information, the clinical dataset was reduced to 228 patients in total. In this cohort, the median recurrence free survival was 36.0 months. When employing the 2-year cutoff for type I or early recurrence, a total of 80 observations (35% of data) are censored (Figure 5.1a). When employing the 5-year cutoff for type II or late recurrence, the number increases to a total of 125 censored observations (55% of data) (Figure 5.1b). Censored observations were defined as those who did not develop recurrence and who died or were lost to follow-up before the cutoff.




(a) 2 year cutoff, applicable to early recurrence only.




(b) 5 year cutoff, applicable to early and late recurrence.

Figure 5.1: Beaujon patient cohort follow-up. Censored patients are defined as those who did not develop recurrence and who died or were lost to follow-up before the cutoff.

Furthermore, due to missing masks of the area of interest or unsuccessful radiomics feature extraction, the tumour and liver datasets were further reduced to 219 and 199 patients respectively. Nevertheless, the proportion of patients that did not develop recurrence is equal to 39% across the clinical, tumour, and liver datasets (Figure 5.2).

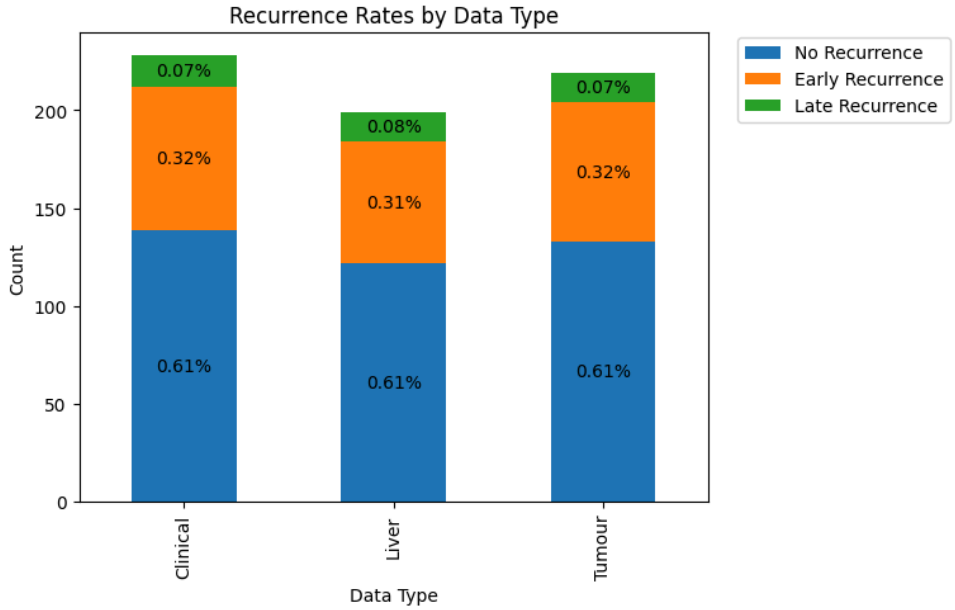
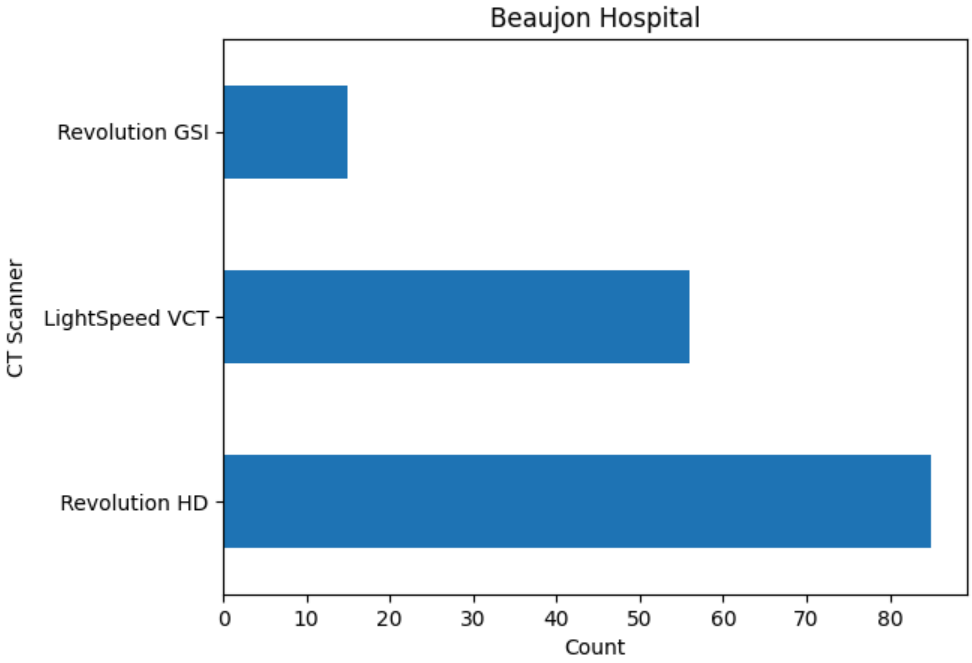
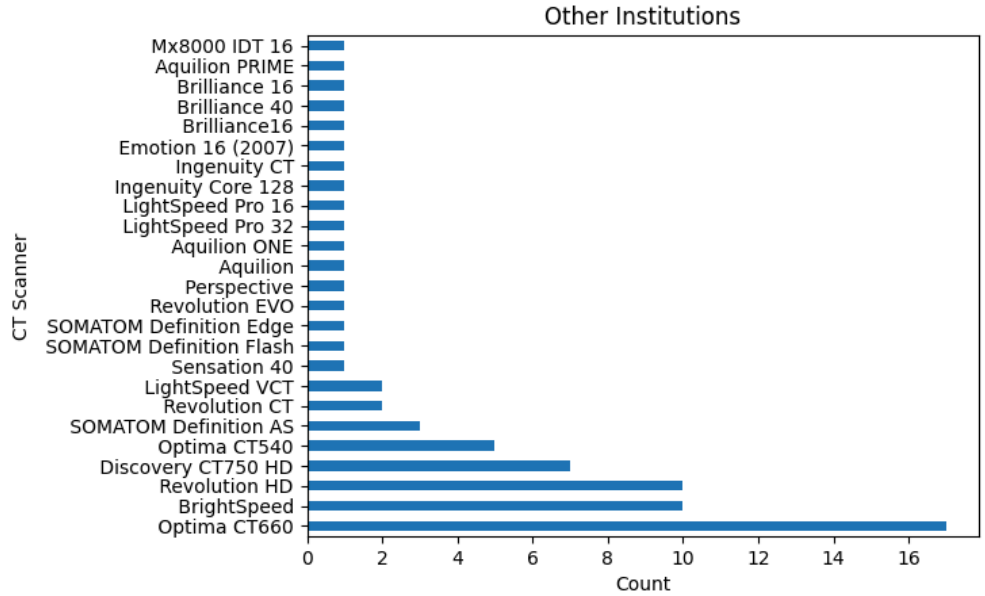



Figure 5.2: HCC recurrence rates of Beaujon cohort by dataset.

To build the recurrence prediction models, a train-test partitioning based on the institution of the CT scan was employed. The train set was composed of CT scans acquired at Beaujon Hospital, belonging to one of three different CT scanner models (Figure 5.3a). The test set was composed of CT scans acquired at other institutions, belonging to one of 24 different CT scanner models Figure 5.3b. For all three datasets, comparable proportions of recurrence were confirmed across the train and test sets (Table A.5).




(a) Scans acquired at Beaujon Hospital (n=137).




(b) Scans acquired at other institutions (n=62).

Figure 5.3: Count of CT scanner types in Beaujon patient cohort. CT scans acquired at Beaujon Hospital were used for training of the HCC recurrence risk classification models. CT scans acquired at other institutions were used for testing of the same models.

## 5.2 Lesion Segmentation and Detection Models

### 5.2.1 nnU-Net Model Validation Without Post-Processing

#### Internal Validation

The segmentation and detection performance metrics of both models during 5-fold cross-validation and hold out test set are summarised in Table 5.2 and visualised in further detail in Figure 5.4.

Table 5.2: Performance metrics of lesion segmentation and detection nnU-Net models on internal test sets of Beaujon data.

| **Cross-validation** | | | | | |
|---|---|---|---|---|---|
| Model | DSC | HD | ASSD | RAVD | Recall |
| AP | 0.65 ± 0.33 | 68.07 ± 86.33 | 14.23 ± 29.90 | 2.66 ± 32.74 | 0.77 |
| PVP | 0.67 ± 0.32 | 59.08 ± 69.47 | 9.63 ± 21.10 | 0.07 ± 0.76 | 0.78 |
| **Hold-out Test Set** | | | | | |
| Model | DSC | HD | ASSD | RAVD | Recall |
| AP | 0.69 ± 0.35 | 41.11 ± 68.33 | 6.4 ± 17.98 | 0.02 ± 0.35 | 0.80 |
| PVP | 0.69 ± 0.34 | 30.0 ± 48.89 | 5.22 ± 13.77 | 0.20 ± 1.20 | 0.80 |

Average metric ± standard deviation is shown. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

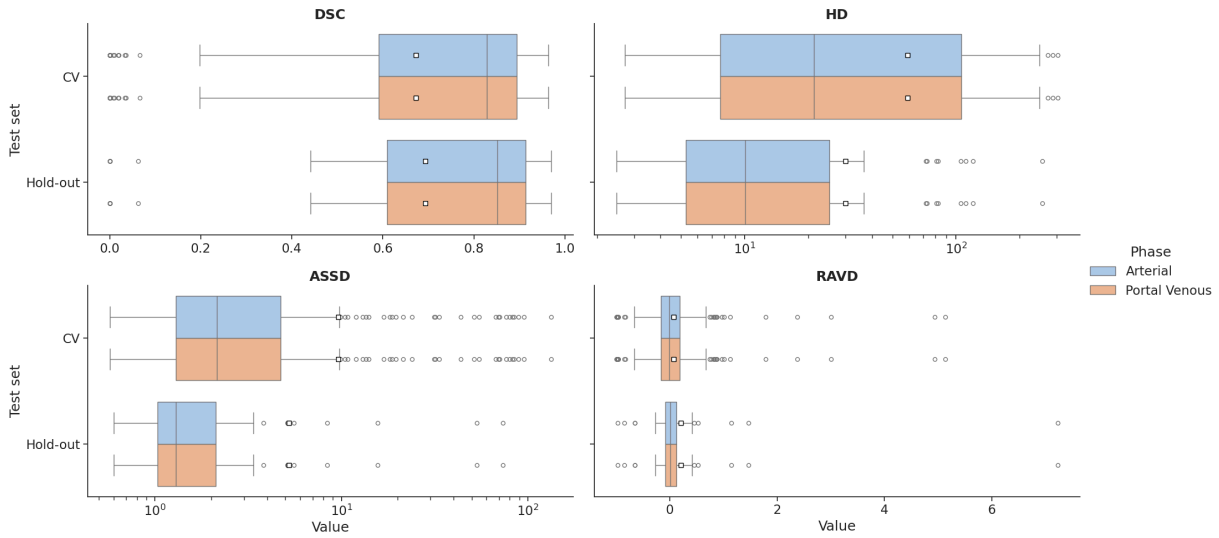



Figure 5.4: Distribution of performance metrics of nnU-Net segmentation models on internal test set of Beaujon data. No post-processing of the model predicted masks was applied. CV: 5-fold cross-validation; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

AP and PVP models demonstrate similar performance across the majority of metrics. For both models, DSC exhibits high variance, due to several instances where the metric is equal to 0 when the models fail to predict any tumour segmentation. This lack of prediction is also partially

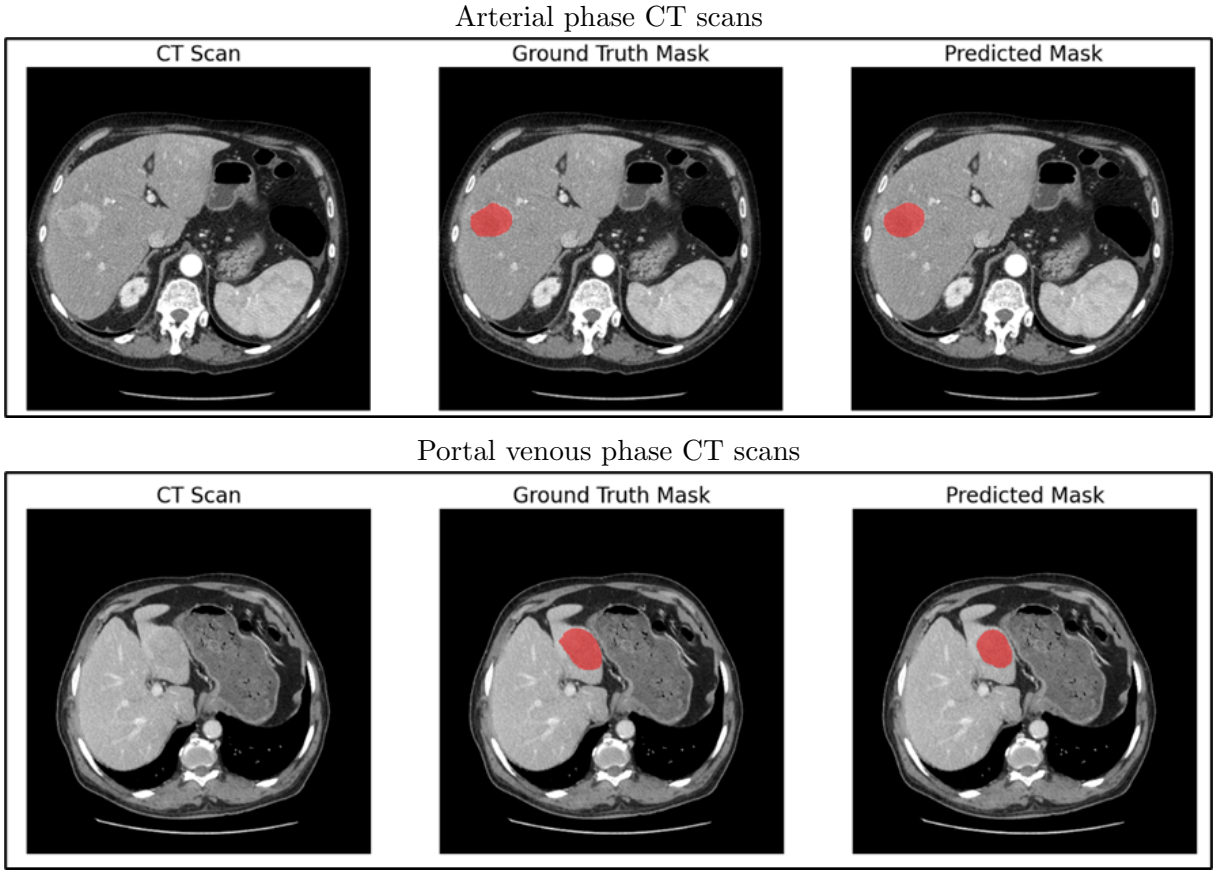



Figure 5.5: Example of nnU-Net models automatic tumour segmentations (Predicted mask) alongside manual segmentations (Ground Truth mask) and original grayscale CT scan. Automatic segmentations of tumours were outputted by the respective nnU-Net models and had a Dice Similarity Coefficient of 0.93 (top) and 0.90 (bottom).

represented in the recall as it contributes to a large portion of false negatives. Even so, the models display strong performance and are robust to an internal hold-out test set, with an average DSC of 0.69 and a recall of around 0.8. However, when examining the additional distance metrics, namely HD and ASSD, a considerable number of outliers are present (Figure 5.4). This suggests that, oftentimes, there is a large discrepancy between the location of the model predicted regions and the ground-truth lesion. Further investigation elucidates that such cases include predicted tumour regions outside the liver (subsection 5.2.2).

Examples of tumour segmentations inferred from the hold out test set by the nnU-Net models are shown in Figure 5.5.

### Association of Model Performance with Clinical Variables

Several clinical variables exhibited a substantial relation with an increased rate of false negatives (Table 5.3). The AP model's performance was negatively associated with increased plasma total bilirubin. The PVP model's performance was negatively associated with the presence of liver cirrhosis, presence of vascular invasion, and absence of HCC capsule. Lastly, the most relevant limitation, as it affects both models, is an association between model performance and tumour size.

Interestingly, for both models, there was no significant difference between the distribution or proportion of the above variables across train and test sets (Table A.1, Table A.2). The full list

of clinical variables, including those lacking a statistically significant association, can be found in Table A.4.

Table 5.3: Association between clinical variables and nnU-Net tumour segmentation model performance.

| **AP Model** | | | |
|---|---|---|---|
| Variable | True Positives | False Negatives | p-value |
| ALP | 103.5 (77.5, 126.0) | 73.0 (71.0, 81.0) | **0.021** |
| Tumour Size (cm) | 4.6 (3.9, 7.5) | 1.50 (1.5, 1.9) | **0.005** |
| **PVP Model** | | | |
| Variable | True Positives | False Negatives | p-value |
| Cirrhosis | 6 (16.2%) | 8 (88.9%) | **<0.001** |
| Vascular Invasion | 19 (51.4%) | 6 (66.6%) | **0.014** |
| Microscopic Capsule | 33 (89.2%) | 4 (44.4%) | **0.009** |
| Macroscopic Capsule | 31 (83.8%) | 4 (44.4%) | **0.041** |
| Tumour Size (cm) | 4.6 (3.2, 7.2) | 2.0 (1.9, 3.2) | **0.003** |

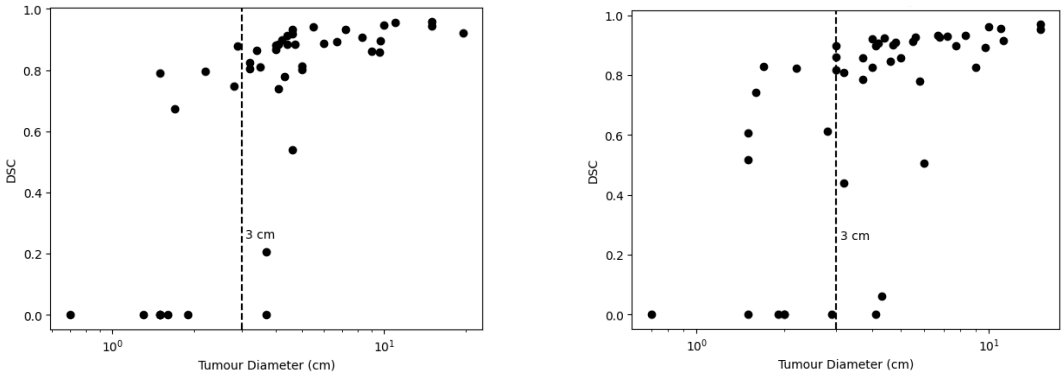
Only statistically significant (p-value $< 0.05$) associations are shown. AP: Arterial Phase; PVP: Portal Venous Phase; Alkaline Phosphatase (ALP): Alkaline Phosphatase.

To quantify the difference in performance between different tumour sizes, patients were split into two groups. For tumours with a diameter less than 3 cm, the average DSC was $0.32 \pm 0.40$ and $0.34 \pm 0.37$ with a recall of 0.42 and 0.50 for AP and PVP models respectively (Table 5.4). The metrics are vastly distinct for tumours with diameters greater than 3 cm, with average DSC of $0.82 \pm 0.20$ and $0.82 \pm 0.23$ and a recall of 0.94 and 0.91 for AP and PVP models respectively (Table 5.4). In general, all performance metrics radically improve for the tumour group of larger size, with DSC exhibiting the most significant improvement Table 5.4. Furthermore, using DSC as the metric of choice, the positive correlation between tumour diameter and improved model performance can be clearly observed in Figure 5.6. The AP model demonstrated a 0.8 Spearman correlation and the PVP model demonstrated 0.75 Spearman correlation.

Table 5.4: Performance metrics of nnU-Net lesion segmentation and detection models on internal hold-out test sets, stratified by tumour diameter ($< 3$ cm and $\geq 3$ cm).

| AP Hold-Out Test Set | | | | | | | |
|---|---|---|---|---|---|---|---|
| Diameter Group | Diameter (cm) | Count | DSC | HD | ASSD | RAVD | Recall |
| <3 cm | 1.76 ± 0.62 | 12 | 0.32 ± 0.40 | 69.27 ± 96.85 | 19.94 ± 37.16 | -0.12 ± 0.29 | 0.42 |
| ≥3 cm | 6.52 ± 3.93 | 33 | 0.82 ± 0.20 | 35.83 ± 62.27 | 3.86 ± 11.0 | 0.05 ± 0.35 | 0.94 |
| p-value | | | **0.001** | 0.598 | 0.770 | 0.316 | |
| PVP Hold-Out Test Set | | | | | | | |
| Diameter Group | Diameter (cm) | Count | DSC | HD | ASSD | RAVD | Recall |
| <3 cm | 1.86 ± 0.60 | 12 | 0.34 ± 0.37 | 32.35 ± 38.3 | 17.10 ± 29.04 | 0.93 ± 2.62 | 0.50 |
| ≥3 cm | 6.22± 3.24 | 34 | 0.82 ± 0.23 | 29.43 ± 51.62 | 2.34 ± 2.95 | 0.03 ± 0.37 | 0.91 |
| p-value | | | **0.001** | 0.428 | 0.355 | 0.487 | |

P-values calculated using Mann-Whitney U test. Average ± standard deviations of tumour diameter and DSC are shown. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.




(a) AP Hold-Out Test Set; $r_s = 0.8$ (b) PVP Hold-Out Test Set; $r_s = 0.75$

Figure 5.6: Scatterplot of DSC by tumour diameter of nnU-Net model predicted masks. The dashed line identifies a tumour diameter of 3 cm. AP: Arterial Phase; PVP: Portal Venous Phase; $r_s$: Spearman's Correlation.

### External Validation

To evaluate the nnU-Net lesion segmentation models' ability to generalise to unseen and external data, the public dataset from TCIA was employed. The segmentation and detection performance metrics of both models evaluated on the external test set are summarised in Table 5.5 and visualised in further detail in Figure 5.7.

Table 5.5: Performance metrics of lesion segmentation and detection nnU-Net models on external test set of tumours with diameter greater than 3 cm.

| **AP Model** | | | | | |
|---|---|---|---|---|---|
| Test set | DSC | HD | ASSD | RAVD | Recall |
| Beaujon ($\geq$ 3 cm) | 0.82 ± 0.20 | 35.83 ± 62.27 | 3.86 ± 11.00 | 0.05 ± 0.35 | 0.94 |
| TCIA | 0.69 ± 0.23 | 102.06 ± 124.51 | 8.52 ± 12.12 | -0.24 ± 0.27 | 0.91 |
| **PVP Model** | | | | | |
| Test set | DSC | HD | ASSD | RAVD | Recall |
| Beaujon ($\geq$ 3 cm) | 0.83 ± 0.23 | 29.43 ± 51.62 | 2.34 ± 2.95 | 0.03 ± 0.37 | 0.91 |
| TCIA | 0.65 ± 0.29 | 174.88 ± 216.53 | 23.04 ± 52.02 | -0.02 ± 1.60 | 0.77 |

Average metric ± standard deviation is shown. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

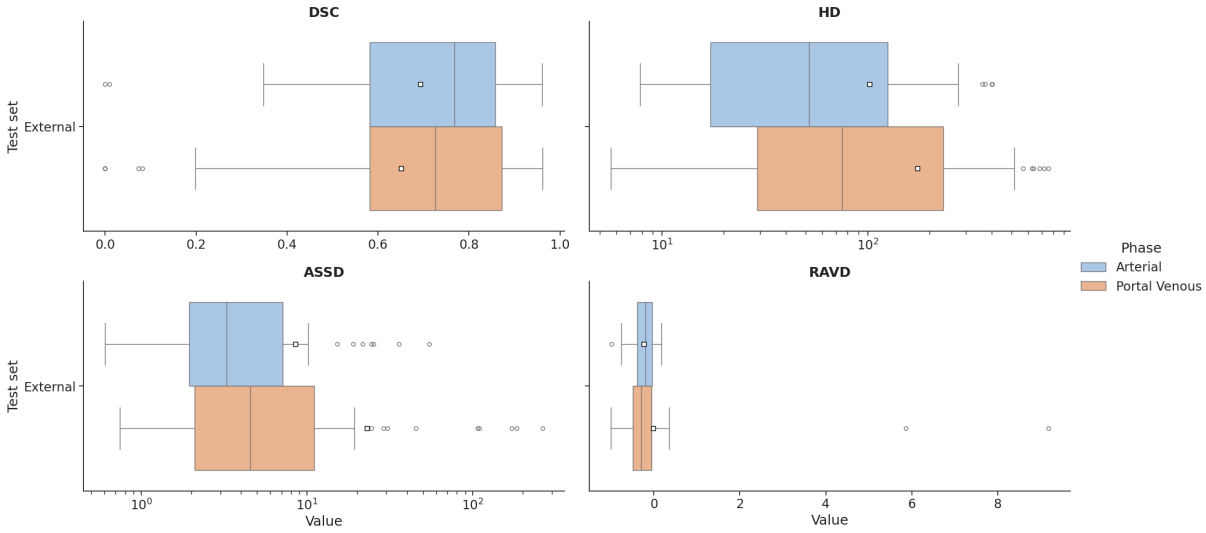



Figure 5.7: Distribution of Performance metrics of nnU-Net segmentation models on external test set of TCIA data. No post-processing of the model predicted masks was applied.AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

The nnU-Net lesion segmentation and detection models demonstrated similar performance on the internal hold out test set Table 5.2 and on external test set Table 5.5. However, considering the TCIA dataset is composed of significantly larger tumours (Table 5.3) and that model performance is positively associated with such tumours, nnU-Net model performance on a subset of the internal test set, strictly of lesions greater than 3 cm in diameter, was included for comparison purposes (Table 5.5). When contrasting between this fairer comparison, the model's ability to generalise becomes less optimistic. Nevertheless, recall remains high with the AP and PVP models' successfully detecting 91% and 77% of lesions, respectively. The results therefore demonstrate that the models are robust to an out of distribution patient population, with the AP model outperforming the PVP model.

Analogous to the models' performance on the internal dataset, several outliers within the distribution of distance- and volume-based segmentation metrics can be observed (Figure 5.7).

### 5.2.2 nnU-Net Model Validation Following Post-Processing

As suggested by the HD and ASSD segmentation metrics in Figure 5.4 and in Figure 5.7, the nnU-Net lesion segmentation models may occasionally delineate a tumour region located considerably distant from its ground-truth counterpart. A potential strategy to enhance model performance and address such dissimilarities involves the exclusion of invalid regions from the model predicted masks, such as HCC lesions situated beyond the liver. This method was carried out as a post-processing step. An example of the segmentation model's predicted mask, previous and subsequent to post-processing, can be observed in Figure 5.8.


Figure 5.8: Example of nnU-Net model predicted tumour mask before and following post-processing. Post-processing involves the removal of predicted tumour regions outside the liver.

### Internal Validation

Subsequent to post-processing, the models were once again tested with 5-fold cross-validation and the internal hold-out test set. The difference between average segmentation and detection metrics prior to and subsequent to post-processing are summarised in Table 5.6. The distribution of the performance metrics, following post-processing, is displayed in Figure A.1a.

Table 5.6: Difference in average performance metrics of lesion segmentation and detection nnU-Net models on internal test sets of Beaujon data, following predicted mask post-processing.

| **Cross-validation** Average Difference | | | | | |
|---|---|---|---|---|---|
| Model | DSC | HD | ASSD | RAVD | Recall |
| AP | +0.01 | -21.24 | -3.54 | -2.62 | +0.01 |
| PVP | 0.00 | -14.90 | -1.83 | -0.04 | +0.01 |
| **Hold-out Test Set** Average Difference | | | | | |
| Model | DSC | HD | ASSD | RAVD | Recall |
| AP | +0.01 | -15.30 | -1.72 | -0.04 | 0.00 |
| PVP | 0.00 | -3.59 | -1.41 | -0.18 | 0.00 |

AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

As expected, the metrics exhibiting the largest improvement include the distance-based values, particularly the HD. Compared to the previously reviewed internal validation, a decrease of 21.24 mm and 14.90 mm is observed for the AP and PVP model cross-validation average HD, and a decrease of 15.30 mm and 3.59 mm for hold-out test set average of the same metric and models. As a distance metric averaging across all surface points, the ASSD is more robust to outliers. Even so, a slight improvement is achieved via post-processing of the model predicted tumour masks, with a resulting decrease ranging from 1.41 to 3.54 mm, depending on model and test set. Conversely, average DSC and recall show minimal to no refinement.

#### External Validation

In the same manner as above, subsequent to mask post-processing, the models were tested using the external TCIA dataset. The differences in segmentation and detection performance are summarised in Table 5.7 with the segmentation performance metrics' distribution displayed in Figure A.1b.

Table 5.7: Difference in average performance metrics of lesion segmentation and detection nnU-Net models on external test sets of TCIA data, following predicted mask post-processing.

| **External Test Set** Average Difference | | | | | |
|---|---|---|---|---|---|
| Model | DSC | HD | ASSD | RAVD | Recall |
| AP | +0.01 | -21.85 | -2.95 | +0.02 | 0.00 |
| PVP | +0.02 | -101.83 | -17.43 | -0.31 | +0.04 |

AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

The difference between model performance prior to and following post-processing using the external test set emulates that of the internal test set, albeit to a greater extent. Once more, the HD and the ASSD benefit the most, with the remaining metrics demonstrating limited enhancement.

### 5.2.3 MedSam Model Validation

#### External Validation

As the MedSam foundation model was trained on TCIA data, an internal hold-out test set was not included. Instead, the model was tested on Beaujon data used, in this case, as an external test set. Performance metrics are summarised in Table 5.9 and visualised in Table 5.8. For the purpose of comparing against MedSam model performance, nnU-Net tumour predicted masks following post-processing were employed.

Table 5.8: Comparison of MedSam and nnU-Net lesion segmentation models performance on Beaujon test sets.

| Beaujon Data (Hold-Out Test Set) | | | | | |
|---|---|---|---|---|---|
| Model | DSC | HD | ASSD | RAVD | Recall |
| MedSAM on AP | 0.72 ± 0.13 | 14.78 ± 10.29 | 2.63 ± 1.87 | -0.03 ± 0.32 | 0.91 |
| nnU-Net on AP | 0.7 ± 0.35 | 25.81 ± 45.92 | 4.67 ± 15.25 | -0.01 ± 0.24 | 0.8 |
| p-value | 0.137 | 0.388 | **0.006** | 0.328 | |
| MedSam on PVP | 0.7 ± 0.13 | 18.38 ± 35.78 | 2.51 ± 1.57 | -0.06 ± 0.32 | 0.93 |
| nnU-Net on PVP | 0.69 ± 0.34 | 26.42 ± 47.6 | 3.81 ± 11.59 | 0.02 ± 0.42 | 0.8 |
| p-value | 0.091 | 0.405 | **0.002** | 0.847 | |

P-values are calculated for a Wilcoxon signed-rank test. Average metric ± standard deviation is shown. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

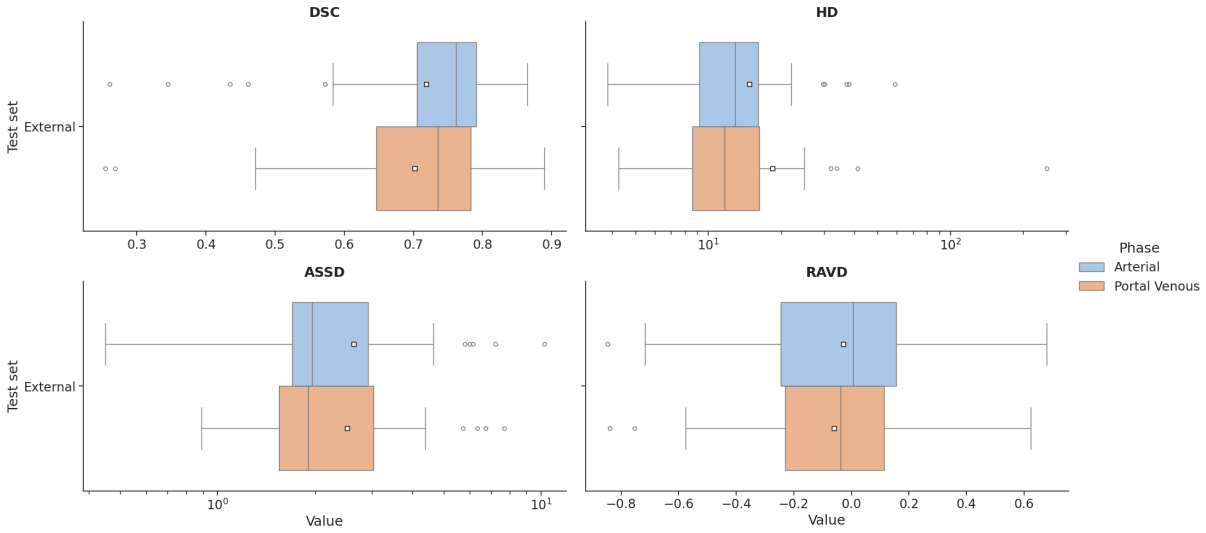



Figure 5.9: Distribution of performance metrics of MedSam segmentation model on external test sets of Beaujon data. DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

Compared to nnU-Net models' tested on Beaujon data, only ASSD and recall exhibited a significant improvement (Table 5.8). It should be noted however that while Beaujon data is employed as nnU-Net training data and therefore within distribution of the hold-out test set, the same is not true for the MedSam model. Therefore, a comparison between MedSam's and nnU-Net's performance metrics on their respective external test sets is also pertinent (Table 5.9).

Table 5.9: Comparison of MedSam and nnU-Net lesion segmentation models performance on their respective external test sets.

| External Test Sets | | | | | |
|---|---|---|---|---|---|
| Model | DSC | HD | ASSD | RAVD | Recall |
| MedSAM on AP | 0.72 ± 0.13 | 14.78 ± 10.29 | 2.63 ± 1.87 | -0.03 ± 0.32 | 0.91 |
| nnU-Net on AP | 0.7 ± 0.23 | 80.21 ± 97.49 | 5.57 ± 6.62 | -0.22 ± 0.27 | 0.91 |
| p-value | 0.357 | **<0.001** | **0.014** | **0.006** | |
| MedSam on PVP | 0.7 ± 0.13 | 18.38 ± 35.78 | 2.51 ± 1.57 | -0.06 ± 0.32 | 0.93 |
| nnU-Net on PVP | 0.67 ± 0.27 | 73.06 ± 99.84 | 5.61 ± 7.4 | -0.33 ± 0.27 | 0.81 |
| p-value | 0.342 | **<0.001** | **0.017** | **<0.001** | |

Beaujon data was used for MedSam model testing and TCIA data was used for nnU-Net model testing. P-values are calculated for a Mann-Whitney U test. Average metric ± standard deviation is shown. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

Using this fairer comparison, we can observe a significant increase in all segmentation performance metrics, apart from DSC. Furthermore, MedSam is able to accurately segment an equivalent or greater number of lesions, as seen by recall, than nnU-Net models.

**Association of Model Performance with Clinical Variables**

In total, solely two clinical patient characteristics demonstrated a significant difference between poor and accurate segmentation groups (Table 5.10). For AP scans, MedSam model performance was negatively associated with absence of HCC capsule, and with decreased tumour sizes. The latter is equally true for the PVP model. Alike the nnU-Net models, tumour size seems to be the strongest indicator of the reduced model performance.

The full list of clinical variables can be found in Table A.4.

Table 5.10: Statistically significant (p-value $< 0.05$) association between clinical variables and MedSam tumour segmentation model performance, tested on Beaujon patient cohort. AP: Arterial Phase.

| **AP Model** | | | |
|---|---|---|---|
| Variable | True Positives | False Negatives | p-value |
| Microscopic Capsule | 34 (82.9%) | 1 (25.0%) | **0.021** |
| Tumour Size (cm) | 4.40 (3.40, 67.0) | 1.55 (1.45, 1.63) | **0.008** |
| **PVP Model** | | | |
| Variable | True Positives | False Negatives | p-value |
| Tumour Size (cm) | 4.30 (3.00, 6.75) | 1.60 (1.15, 1.65) | **0.010** |

In a similar fashion to subsection 5.2.1, patients were split into two groups. For tumours with a diameter less than 3 cm, the average DSC was 0.59 ± 0.17 and 0.54 ± 0.15 with a recall of 0.67 and 0.75 for AP and PVP models respectively (Table 5.11). The metrics are significantly distinct for tumours with diameters greater than 3 cm, with average DSC of 0.77 ± 0.06 and 0.76 ± 0.07 and a perfect recall of 1 and 1 for AP and PVP models respectively (Table 5.11).

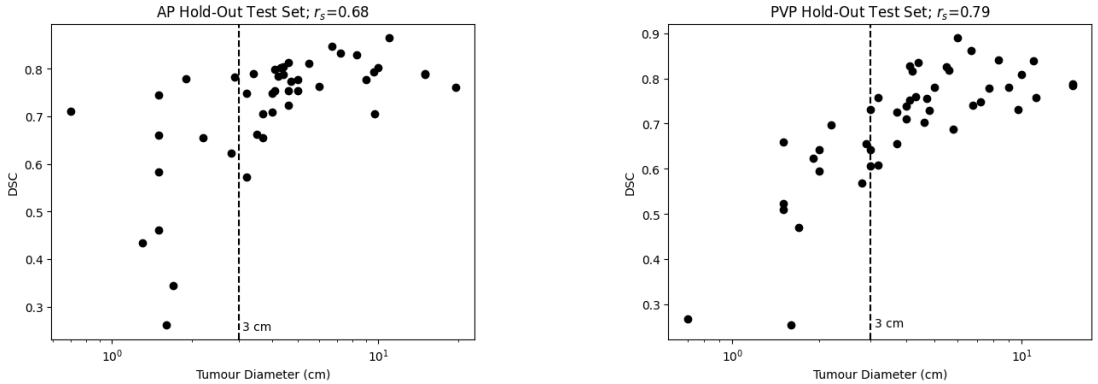
In general, all performance metrics radically improve for the tumour group of larger size, except for the distance-based metrics, HD and ASSD. Nevertheless, MedSam always predicts a tumour region, resulting in more robust predictions for small tumours, as observed by the greater DSC and recall, when compared to nnU-Net models for the same tumour group (Table 5.4).

Table 5.11: Performance metrics of MedSam lesion segmentation model on Beaujon external test sets, stratified by tumour diameter ($< 3$ cm and $\geq 3$ cm).

| AP External Test Set | | | | | | | |
|---|---|---|---|---|---|---|---|
| Diameter Group | Diameter (cm) | Count | DSC | HD | ASSD | RAVD | Recall |
| <3 cm | 1.76 ± 0.62 | 12 | 0.59 ± 0.17 | 8.15 ± 3.55 | 1.48 ± 0.81 | -0.37 ± 0.26 | 0.67 |
| ≥3 cm | 6.52 ± 3.93 | 33 | 0.77 ± 0.06 | 17.2 ± 10.9 | 3.05 ± 1.98 | 0.1 ± 0.23 | 1.00 |
| p-value | | | **<0.001** | **<0.001** | **<0.001** | **<0.001** | |
| PVP External Set | | | | | | | |
| Diameter Group | Diameter (cm) | Count | DSC | HD | ASSD | RAVD | Recall |
| <3 cm | 1.86 ± 0.6 | 12 | 0.54 ± 0.15 | 8.09 ± 3.97 | 1.36 ± 0.41 | -0.41 ± 0.27 | 0.75 |
| ≥3 cm | 6.22± 3.24 | 34 | 0.76 ± 0.07 | 22.01 ± 41.09 | 2.92 ± 1.63 | 0.06 ± 0.23 | 1.00 |
| p-value | | | **<0.001** | **<0.001** | **<0.001** | **<0.001** | |

P-values calculated for Mann-Whitney U test. Average ± standard deviations of tumour diameter and DSC are shown. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

Furthermore, using DSC as the metric of choice, the positive correlation between tumour diameter and improved model performance can be clearly observed (Figure 5.10). The AP model demonstrated a 0.68 Spearman correlation and the PVP model demonstrated a 0.79 Spearman correlation.




(a) AP Hold-Out Test Set; $r_s = 0.68$

(b) PVP Hold-Out Test Set; $r_s = 0.79$

Figure 5.10: Scatterplot of DSC by tumour diameter between MedSam model predicted masks and ground-truth masks. The dashed line identifies a tumour diameter of 3 cm. AP: Arterial Phase; PVP: Portal Venous Phase; $r_s$: Spearman's Correlation.

## 5.3 Radiomics-based Prediction of Recurrence Risk

To predict future recurrence of HCC patients, six different classification models were tested, across three different datasets. The datasets included clinical variables only, clinical variables combined with tumour-extracted radiomic features, and clinical variables combined with liver-extracted radiomic features. Hereafter, the different datasets will be referred to as clinical, tumour, and liver datasets, respectively.

### 5.3.1 Feature Selection

Originally, the clinical dataset was composed of a total of 54 variables, the tumour dataset composed of 4174 features, and the liver dataset composed of 4172 features.

Subsequent to unsupervised and supervised feature selection (Figure 5.11), the datasets were reduced to the following size:

- Clinical data: 27 clinical variables in total
- Tumour data: 73 variables in total (56 radiomics + 16 clinical)
- Liver data: 34 variables in total (6 radiomics + 27 clinical)

Complete details of the composition of each dataset can be found in the supplementary materials (page 68).

### 5.3.2 Radiomics-Based Model Evaluation

Classification performance metrics, F1 score and AUROC, for all models and dataset combinations are displayed in Figure 5.12 and Figure 5.13, respectively. The associated quantitative results are summarised in Table 5.12.

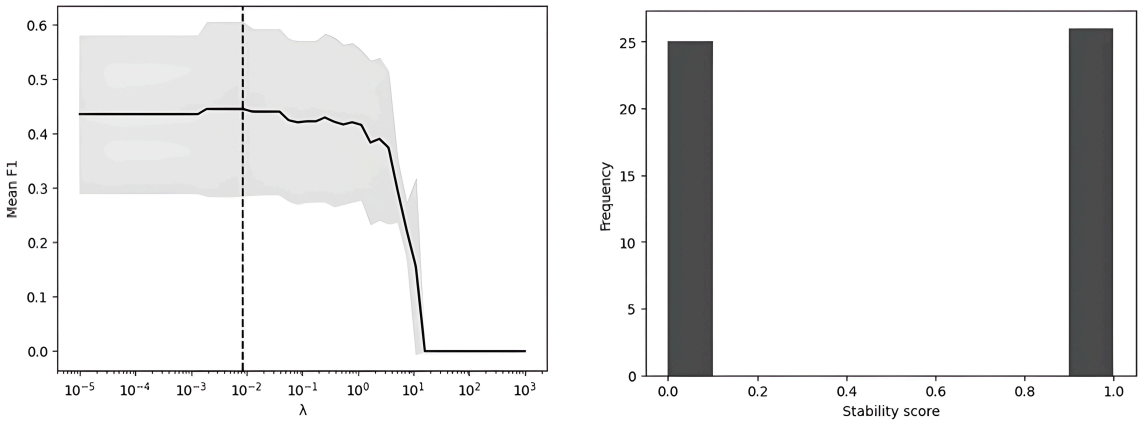
Overall, all models present with unsatisfactory performance, with a common trend of overfitting to the training data. The simpler models like logistic regression have a tendency to overfit

to the more complex tumour dataset. Contrarily, tree-based and ensemble models, which are considered high variance models, tend to overfit more when applied to the simpler clinical and liver datasets (Figure 5.12).

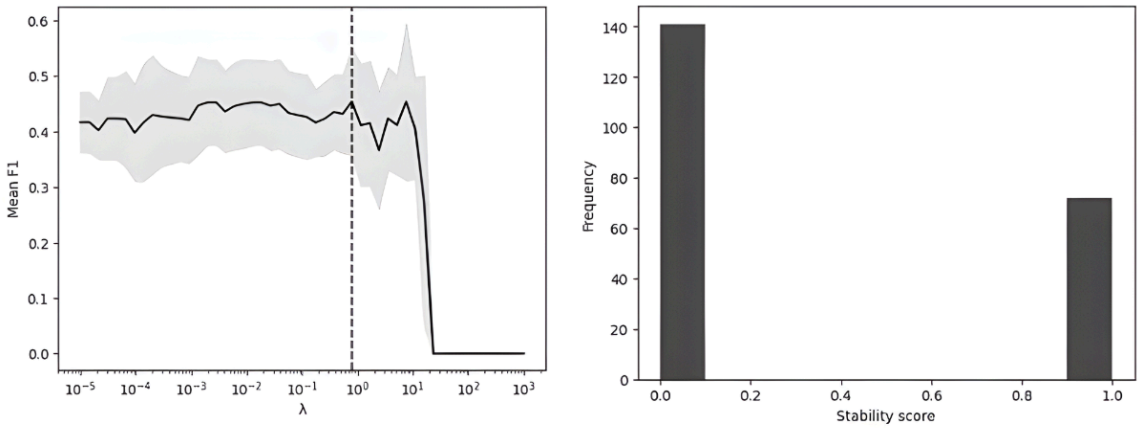
With respect to cross-validation performance, models trained on tumour data consistently emerge as the best performing classifiers. Across all tested models, tumour-based classifiers demonstrate the highest F1 score and, predominantly, the highest AUROC (Table 5.12). More specifically, the logistic regression trained on tumour data achieved the highest F1 score (0.69 $\pm$ 0.13) and AUROC (0.79 $\pm$ 0.05), on the train set. However, these promising results do not translate to the test set, with tumour-based models generally exhibiting the highest degree of overfitting.

Regarding test set performance, models trained on liver data, yielded the most favourable results. Despite the liver dataset containing the smallest number of training cases (Figure 4.1), liver data achieved the highest test set AUROC across all model types (Figure 5.13), and the overall highest F1 score (Table 5.12). The former were achieved with a logistic regression, resulting in an F1 score of 0.54 and a AUROC of 0.66 on the test set.

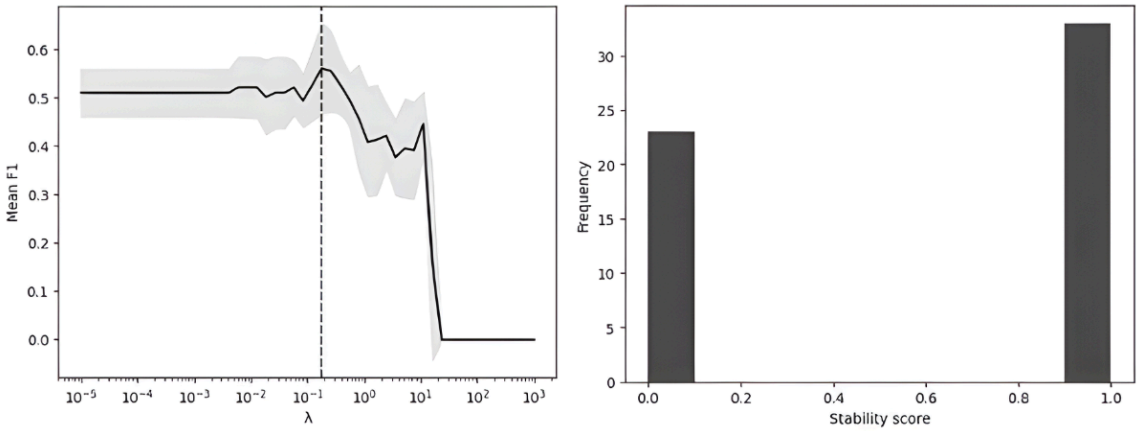
Finally, a closer analysis of models solely comprising clinical data reveals consistent underperformance when compared to radiomics-inclusive models (Figure 5.12). This trend holds true for the vast majority of achieved F1 scores and AUROC across both the training and test sets (Table 5.12).




(a) Clinical dataset.




(b) Tumour dataset.




(c) Liver dataset.

Figure 5.11: Supervised feature selection of clinical and radiomic features. Left: Tuning of $\lambda$, L1 regularisation parameter, for Least Absolute Shrinkage and Selection Operator (LASSO) via 5-fold cross-validation. Right: Stability analysis of features over 1000 LASSO runs using the selected hyperparameter.

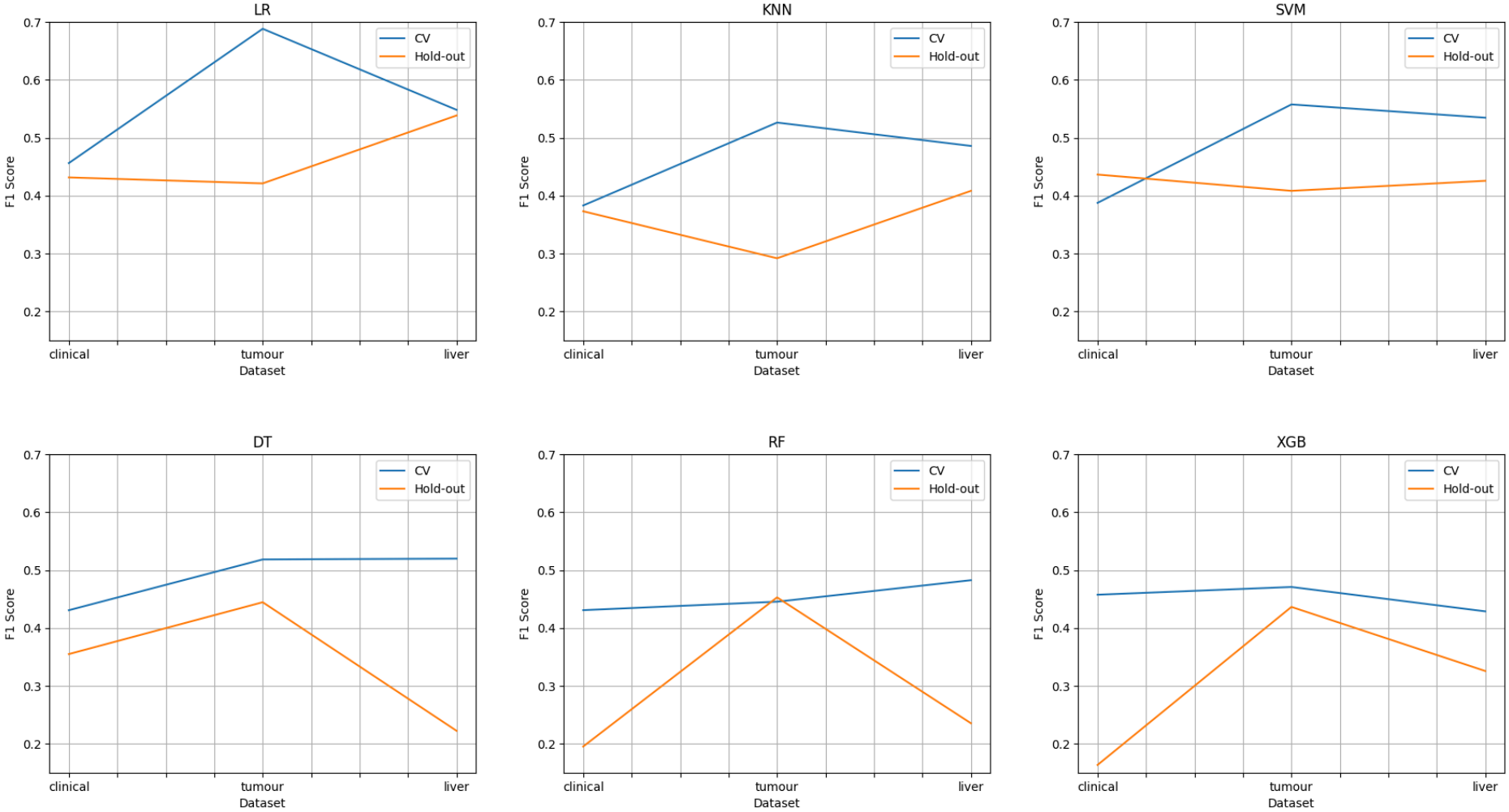



Figure 5.12: Line plots of F1 score for cross-validation (CV) and hold-out test sets of HCC recurrence risk classification models, across clinical, tumour, and liver datasets. LR: Logistic regression; KNN: K-Nearest Neighbours; SVM: Support Vector Machine; DT: Decision tree; RF: Random forest; XGB: Extreme gradient boosting.

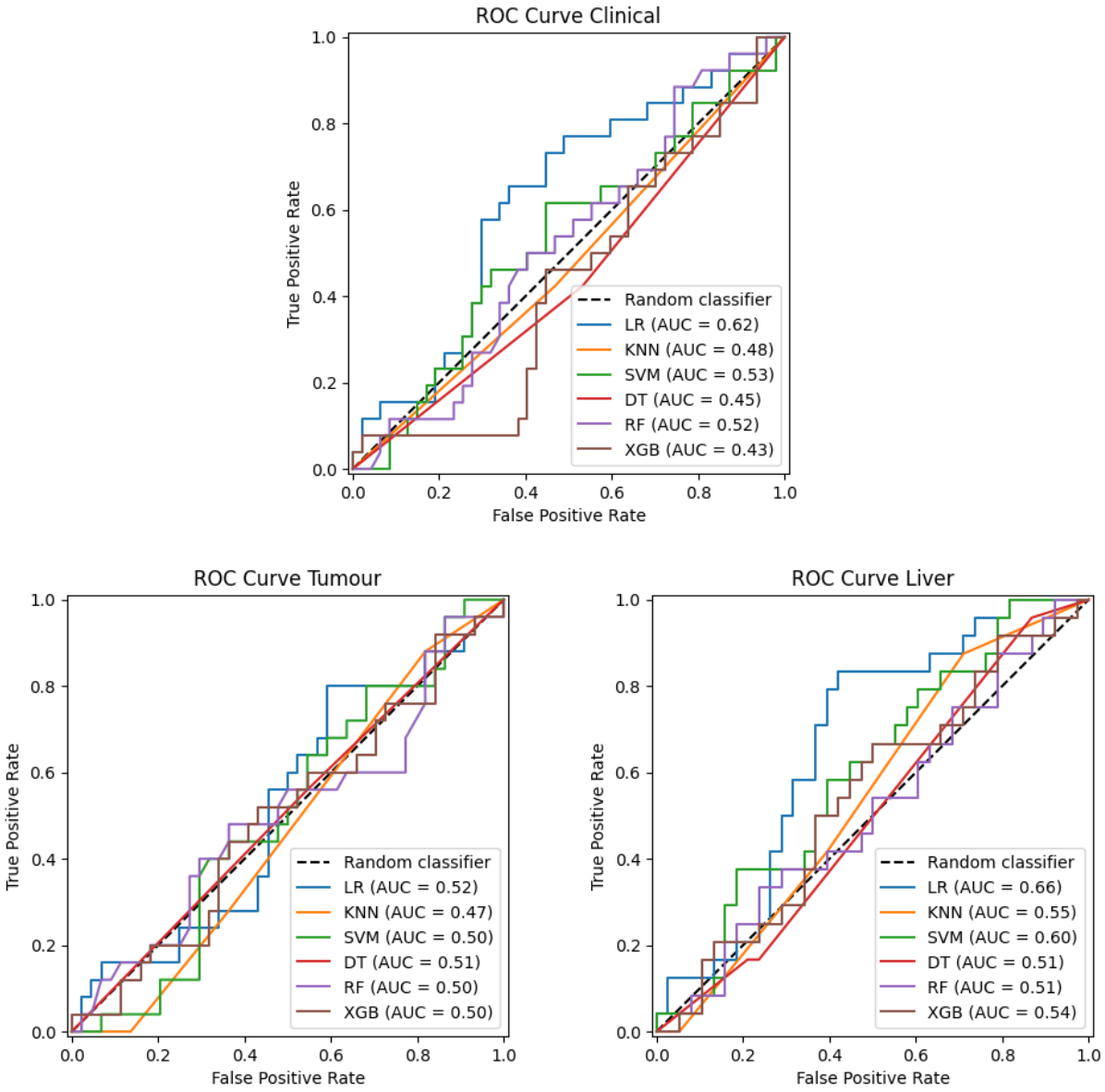



Figure 5.13: ROC curve for hold-out test set of HCC recurrence classification models, using clinical, tumour, and liver datasets. AUC: area under the curve; LR: Logistic regression; KNN: K-Nearest Neighbours; SVM: Support Vector Machine; DT: Decision tree; RF: Random forest; XGB: Extreme gradient boosting.

Table 5.12: Performance metrics of HCC recurrence risk classification models across clinical, tumour, and liver datasets.

| **Clinical** | | | | |
|---|---|---|---|---|
| Model | CV F1 | TS F1 | CV AUROC | TS AUROC |
| LR | 0.46 ± 0.14 | 0.43 | 0.61 ± 0.10 | 0.62 |
| KNN | 0.38 ± 0.07 | 0.37 | 0.48 ± 0.06 | 0.48 |
| SVM | 0.39 ± 0.14 | 0.44 | 0.52 ± 0.12 | 0.53 |
| DT | 0.43 ± 0.11 | 0.35 | 0.50 ± 0.09 | 0.45 |
| RF | 0.43 ± 0.04 | 0.20 | 0.60 ± 0.09 | 0.52 |
| XGB | 0.46 ± 0.08 | 0.16 | 0.59 ± 0.12 | 0.43 |
| **Tumour** | | | | |
| Model | CV F1 | TS F1 | CV AUROC | TS AUROC |
| LR | 0.69 ± 0.13 | 0.42 | 0.79 ± 0.05 | 0.52 |
| KNN | 0.53 ± 0.10 | 0.29 | 0.60 ± 0.13 | 0.47 |
| SVM | 0.56 ± 0.10 | 0.41 | 0.70 ± 0.02 | 0.50 |
| DT | 0.52 ± 0.11 | 0.44 | 0.59 ± 0.09 | 0.51 |
| RF | 0.45 ± 0.11 | 0.45 | 0.63 ± 0.05 | 0.50 |
| XGB | 0.47 ± 0.13 | 0.44 | 0.58 ± 0.08 | 0.50 |
| **Liver** | | | | |
| Model | CV F1 | TS F1 | CV AUROC | TS AUROC |
| LR | 0.55 ± 0.12 | 0.54 | 0.64 ± 0.11 | 0.66 |
| KNN | 0.49 ± 0.13 | 0.41 | 0.60 ± 0.09 | 0.55 |
| SVM | 0.53 ± 0.07 | 0.43 | 0.62 ± 0.06 | 0.60 |
| DT | 0.52 ± 0.10 | 0.22 | 0.61 ± 0.10 | 0.51 |
| RF | 0.48 ± 0.12 | 0.24 | 0.65 ± 0.12 | 0.51 |
| XGB | 0.43 ± 0.07 | 0.33 | 0.61 ± 0.03 | 0.54 |

Mean and standard deviation of 5-fold cross-validation (CV) scores are displayed. TS: hold-out test set; LR: Logistic regression; KNN: K-Nearest Neighbours; SVM: Support Vector Machine; DT: Decision tree; RF: Random forest; XGB: Extreme gradient boosting.

# Chapter 6

# Discussion and Conclusion

## 6.1 Research Objectives

The goal of the present study was to developed an in-house semi-automatic pipeline for radiomics-based recurrence risk prediction. For this end, two main steps were performed. The first included the comparison of two different lesion segmentation models, nnU-Net and MedSam. The second involved the development and testing of HCC recurrence risk prediction models, using a combination of different data and model types. Although progress was made toward both objectives, further work is still required to fully achieve the pipeline's intended goals.

## 6.2 Lesion Segmentation Models

When evaluating the trained nnU-Net models on the internal hold-out test set of Beaujon data, favorable performance for the detection and segmentation of HCC was demonstrated, with models achieving 0.8 recall on AP and PVP scans.

However, assessment of model performance across different patient and tumour characteristics, indicated certain flaws. Poor performance of the PVP model was associated with the presence of liver cirrhosis, presence of vascular invasion, and absence of HCC capsule. The mentioned disease features are all negative factors within the diagnosis or prognosis of HCC. Liver cirrhosis is one of the most relevant risk factors contributing toward the development of HCC [2, 21]. Within the domain of advanced HCC, the presence of tumour capsule is a favorable prognostic factor. Nevertheless, encapsulated tumours are characteristic of progressed disease [49]. Vascular invasion is a well-known poor prognostic factor associated with more aggressive tumours [1, 2]. Therefore, there may be a link between poor PVP model performance and tumour aggressiveness and progression, although further investigation is required to comprehensively assess the association. In contrast to the aforementioned findings, poorer AP model performance was observed with decreased concentrations of Alkaline Phosphatase (ALP), associated with improved prognosis of HCC patients [50]. Overall, although further investigation is required to elucidate the link between segmentation accuracy and patient characteristics, the results suggest that the trained nnU-Net segmentation models may have inherent biases, making them less effective for certain subsets of patients.

Both models demonstrated a decreased performance associated with reduced tumour diameters. However, this is a widely common and inherent limitation of lesion segmentation models. Further aggravating this challenge, patients with small tumours (diameter $< 3$ cm) were underrepresented within the training set, comprising 27% of dataset. It is no surprise, then, that given the data imbalance concerning tumour size, the models exhibited superior performance on the most

prevalent tumour group.

Following, post-processing of the nnU-Net model's predicted tumour masks, which involved removal of tumour predictions outside the liver region, vast improvements were showcased in distance-based segmentation performance metrics. Although, the substantial improvement in the HD and ASSD may appear to contradict the minor difference in DSC and recall, the circumstance can be largely attributed to two main considerations. First and foremost, in cases where the segmentation model predicts a single region outside the liver, the entire predicted region is removed. While DSC remains at 0 prior to and following removal, HD and ASSD decrease from large distances to null values, due to the absence of predicted mask. This is a limitation of the distance-based metrics as, unlike the DSC or recall, they are unable to quantify the absence of a mask prediction. Secondly, in scenarios where the model predicts multiple tumour regions, including one at the correct location and one outside the liver, the DSC may initially be adequate and slightly increase once the secondary incorrect region is removed. On the contrary, HD will be positively impacted to a much greater extent. In summary, while the applied post-processing is suitable for removal of invalid predicted tumour regions and minimisation of the distance-based segmentation metrics, it is less effective for enhancing detection performance.

The generalisability of the trained nnU-Net models was further confirmed with an external test set. As previously highlighted, the patient population of the external TCIA dataset exhibits significant differences compared to the training set. Firstly, the external data was acquired between 1998 and 2006, while the internal data was acquired in 2010 at the earliest. The gap in CT scan technology spanning several years was expected to potentially negatively impact model performance. Additionally, in terms of clinical characteristics, patients in the TCIA cohort tended to have larger tumour sizes, a feature associated with more accurate segmentations. Conversely, TCIA patients also exhibited increased disease progression and burden. This aspect is particularly pertinent considering the identified limitations of the PVP model. Indeed, this concern is reflected in the discrepancy in robust segmentations between the AP and PVP models on the TCIA dataset, with respective recall of 0.91 and 0.81, following post-processing. Despite these challenges, both models remain robust and achieve high recall, thus demonstrating robustness to an out-of-distribution patient population, with variations in scanner technology, institution, and disease progression.

When evaluating MedSam model performance on an external test set, the models demonstrated robustness to unseen and out-of-distribution data. In fact, MedSam exhibited comparable performance to nnU-Net on Beaujon data, despite being an internal dataset for the latter model. Moreover, when comparing the performance of both models on their respective external test sets, MedSam outperformed nnU-Net on every performance metrics other than DSC, where no significant difference was observed. However, it should be noted that MedSam inherently holds the advantage in terms of distance- and volume-based metrics, due to the exact tumour bounding box that is passed as required user-input for model inference.

With regard to patient characteristics, MedSam, like nnU-Net, demonstrated reduced performance with a more pronounced absence of microscopic HCC capsule. Interestingly, MedSam's segmentation accuracy was also negatively affected, to a great extent, by the presence of small tumours, further highlighting lesion size as an inherent limitation of segmentation models, irrespective of their architecture or training process. Nevertheless, MedSam appears to suffer from less bias, and be less sensitive to tumour characteristics. This can be inferred by the reduced number of patient characteristics affecting model performance, and by the adequate performance with lesions lesser than 3 cm. The observed differences, are likely due to the vast amount of data the MedSam model had access to during training and fine-tuning, leading to increased generalisability, and robustness to biases present in a small dataset.

To conclude, both model types demonstrate an ability to accurately segment HCC, in addition

to a good ability to generalise to a completely distinct patient population. However, each model comes with its own advantages and disadvantages. For example, while nnU-Net requires data and training of a new specialist model designed for each specific segmentation task, it is also fully automatic. On the other hand, MedSam, as a foundation model, can immediately be used for most segmentation tasks. However, it is not automatic as it requires a user-provided bounding box of the target lesion. As such, while it is useful for the fine segmentation of lesion, it is much less applicable as a detection model. Lastly, depending on the available computing power, MedSam's inference could take a substantial amount of time during use. For these reasons, further analysis from a clinical perspective is still needed as it would be extremely valuable to uncover and quantify the impact of each model for supporting clinical workflow.

## 6.3 Radiomics-based Prediction of Recurrence Risk

Overall, the performance of the developed models for preoperative recurrence risk prediction in HCC patients was unsatisfactory. Rare instances of good performance on the training set did not generalize well to the test set. For example, the best model on cross-validation demonstrated a significant drop in F1 score from 0.69 to 0.42 and AUC from 0.79 to 0.52 when applied to the test set. Conversely, models achieving the highest performance metrics on the test set did not overfit but only showed adequate performance. For instance, a model with the best performance on the test set scored an F1 of 0.55 and 0.54, and AUROC of 0.64 and 0.66 on the train and test sets, respectively.

Simpler models like logistic regression exhibited a higher degree of overfitting when trained on the tumour dataset. In addition to the large number of features, tumours can be highly heterogeneous, leading to extracted radiomic features with a high degree of noise and complexity. This complexity within tumour data makes it challenging for simple models to identify the true underlying pattern behind recurrence risk. Similar results have been reported not only for liver tumours [14], but for other organs as well [51]. On the other hand, more complex models with high variance, such as decision trees, random forests, and XGBoost, tended to overfit on simpler datasets like clinical and liver data. These models fit the data too closely, capturing not only the underlying pattern but also additional noise and variability, resulting in poor generalisability to new data.

Two main factors contribute to the reduced performance observed in this study compared to the literature. Firstly, recurrence events may go unrecorded for patients lacking adequate follow-up, 2 years for early recurrence and 5 years for late recurrence. The right-censored nature of the data introduces uncertainty regarding the true future recurrence status of patients, potentially undermining model performance. Furthermore, this consideration may also be the reason behind reduced rates of recurrence observed in this study, solely 31%, compared to the 60% following HCC resection, as observed in literature [52]. Secondly, radiomic features are known to be highly sensitive to differences in image acquisition, such as different scanner types or settings [25]. Given the multi-institutional setting employed in this study, these differences in image acquisition could significantly impact model performance. Therefore, our findings highlight the restrictions of the predictive power of radiomic features, when applied to a multi-institutional setting.

Nonetheless, in line with findings from the reviewed studies in section 1.2, radiomics data does provide added value to a model's predictive power. Within the train set, the model achieving the highest F1 (0.69) was trained on data incorporating tumour-extracted radiomic features, while the best performing clinical model achieved a F1 score of 0.46. Similarly, within the test set, the model achieving the highest F1 (0.54) incorporated liver-extracted radiomic features, while the clinical models achieved a maximum F1 score of 0.44. Furthermore, the best performing model on the test set, composed of liver-extracted radiomic features, did not overfit to the training

data. This indicates that whole-liver extracted radiomics may help reduce the heterogeneity and incorporated noise associated with tumour-extracted radiomics.

In summary, while the present study confirms the added value of radiomics within medicine, the observed results suggest that the utility of the features is not universal, as their advantage does not translate to a patient cohort composed of CT images acquired at different institutions and with diverse set of CT scanner models. To fully realize the potential of radiomics for recurrence risk prediction, a different approach must be explored. This could include more robust feature selection strategies, or even incorporating deep learning image embeddings like those generated by MedSam. These steps are necessary to mitigate noise in the data and improve the predictive power of the models. As such, while the groundwork for the pipeline has been established, further refinement is required for its use in clinical practice.

## 6.4 Limitations

While the present study provides promising results into HCC lesion segmentation methods and valuable insight into the restricted applicability of radiomic features for recurrence risk prediction across multiple institutions, several limitations of the study design must be considered.

In this study, the approximate total of 230 patients employed for model training and testing is far from ideal. In the context of machine and deep learning for medical applications, the amount of available data is a common limiting factor. While deep learning foundation models like MedSam offer the advantage of being usable out of the box, other segmentation models such as nnU-Net require previous training. This training, in order to achieve optimal results, inherently requires an extensive and varied dataset. Likewise, in the field of machine learning and radiomics, the availability of an ample dataset is crucial. Insufficient or unrepresentative data, can lead to inherent bias and variance within the model, causing it to overfit or underfit, both resulting in diminished model performance. In summary, lack of data is likely to directly impact a model's generalisability to future cases, thus limiting its use within clinical applications.

During lesion segmentation model testing, the TCIA patient cohort was employed as an external test set of the nnU-Net models, while the Beaujon patient cohort was employed as an external test set of the MedSam model. The tests sets are not analogous to each other. Therefore, the study is missing a suitable test set external to both models. Such a dataset would be beneficial for a more thorough and direct comparison of both models' ability to generalise to unseen data.

Throughout the study, liver segmentation masks were utilised for several purposes. These included, cropping of the CT scans above and below the liver to accelerate training and inference of lesion segmentation models, removing inferred tumour lesions outside the liver boundary, and extracting radiomic features from the liver for radiomics-based recurrence prediction. These masks were predicted by an automatic tool and were not corrected by an expert before being employed in this study, leading to potential errors in the liver delineation that may have gone unnoticed, potentially affecting the accuracy and reliability of downstream analyses. Despite this, the segmentation of large organs such as the liver is typically robust, and the impact of minor inaccuracies is expected to be limited.

A critical limitation of the present study, affecting the comparison of semi-automatic HCC segmentation methods, is the lack of clinical significance analysis. While the methods were compared with respect to accuracy, they were not evaluated in terms of reducing the time and effort required for a radiologist to segment the lesions. Additionally, accuracy alone fails to capture key considerations. For instance, nnU-Net's automatic pipeline may save more time than MedSam's real-time inference, yet MedSam's finer delineation for smaller lesions could reduce the need for correction. For these reasons, evaluating the models' impact on enhancement of clinical workflow is one of the most important factors for determining their real-world applicability and

utility. Without this information, it is impossible to determine and compare the true value of the models when applied to everyday clinical practice.

An additional limitation of this study is the selection bias inherent in the Beaujon cohort, as it only includes patients who underwent surgical resection. This bias affects both the trained nnU-Net lesion segmentation models and the radiomics-based recurrence risk prediction models. While the nnU-Net model demonstrated good generalizability to a different patient population, the TCIA population is outdated. Therefore, further investigation is necessary, as the current evaluation does not ensure robustness for future patients. Conversely, this issue is less pertinent to the radiomics models. Despite the lack of an external cohort, these models failed to accurately predict recurrence risk within the internal cohort, rendering further external validation redundant.

As previously mentioned, one of the limitations of the radiomics-based prediction of future recurrence, is the presence of right-censored data. Right-censored data can significantly undermine a model's performance by introducing error within the training process. This is because, when patients are lost to follow-up, there is incomplete information about their future recurrence status, consequently resulting in potentially incorrect target labels. The model may therefore learn inaccurate patterns, limiting its ability to correctly predict recurrence risk in new patients.

The reproducibility of radiomic features is crucial for ensuring their reliability in predictive modelling. However, since we did not have two masks per lesion available for the entire dataset, we were unable to carry out a stability analysis subsequent to the extraction of radiomic features. This limitation is significant because it could potentially result in the lack of reproducibility across inter-reader mask annotations for some of the tumour-extracted features [53]. The potential variation may lead to inconsistencies or potential biases in the models using tumour-radiomic features, thus limiting their overall performance.

## 6.5 Future Steps

To build upon the findings of this study and overcome the identified limitations, several steps are necessary to fully develop the pipeline for radiomics-based recurrence risk prediction.

First and foremost, the collection of more data is vital to enhance model performance, improve generalisability, and further validate the accuracy of trained models. For lesion segmentation models, additional and varied data would enable further refinement of the nnU-Net models, to promote accurate segmentations of unseen tumour cases, an potentially mitigate the limitations observed with respect to certain patient characteristics. With regard to TCIA data, a possible approach to obtain more data would be the coregistration of both phases, so the single available tumour mask is suitable for both CT examinations. Moreover, further data that is more current would facilitate the creation of an independent test set, external to both MedSam and nnU-Net models, allowing for an unbiased and direct comparison between the two types of segmentation models. For recurrence risk prediction models, more data would be beneficial for improving the models' predictive power and robustness, as it assists in the identification of the true underlying pattern, thus reducing the risk of overfitting.

To address the study's current lack of clinically relevant analyses, a quantitative comparison of time taken for radiologists to segment lesions aided by nnU-Net, MedSam, and without assistance is crucial. Furthermore, assessing inter-reader variability through the Intraclass Correlation Coefficient (ICC) can elucidate the models' potential impact on promoting reproducible segmentations across different readers.

In addition to further data acquisition, implementing stricter inclusion criteria in terms of patient

follow-up would be ideal for enhancing the credibility of radiomics-based prediction of recurrence risk. This approach would minimise the uncertainty for the true recurrence status of patients, helpfully eliminating right-censorship of data as a potential factor hindering model performance.

In relation to the above, to further improve tumour-radiomics feature selection, it would be advantageous to acquire multiple annotated masks per tumour lesion. This step is necessary to identify and remove radiomic features that are not reproducible across distinct expert annotations, thus making the models more reliable and robust to inter-reader variability.

Finally, the recurrence risk prediction models could be developed to an even greater extent. For example, with respect to feature selection strategies, only LASSO was tested. It could be interesting to test additional approaches such as maximum relevance minimum redundancy (mRMR) and recursive feature elimination (RFE). mRMR focuses on features that are most relevant to the target variable, while minimising redundancy among them. RFE iteratively removes the least important features based on model performance. Lastly, treating the correlation threshold as a hyperparameter in need of tuning could help ensure the ideal number of informative features. On the other hand, the study could be further expanded to include alternative feature types. For example, integrating deep learning encodings, such as MedSam's image embeddings could potentially reduce noise within the data. This is because, having been trained on millions of images, MedSam is more likely to focus on the most informative and essential portions of the image, thereby providing a more robust and noise-free representation of the tumour. The approach could, in principle, be leveraged for enhancement of the recurrence risk prediction models.

# Bibliography


[1] A. Vogel, T. Meyer, G. Sapisochin, R. Salem, and A. Saborowski, "Hepatocellular carcinoma," *The Lancet*, vol. 400, p. 1345–1362, Oct. 2022.

[2] J. M. Llovet, R. K. Kelley, A. Villanueva, A. G. Singal, E. Pikarsky, S. Roayaie, R. Lencioni, K. Koike, J. Zucman-Rossi, and R. S. Finn, "Hepatocellular carcinoma," *Nature Reviews Disease Primers*, vol. 7, Jan. 2021.

[3] B. McMahon, C. Cohen, R. S. Brown Jr, H. El-Serag, G. N. Ioannou, A. S. Lok, L. R. Roberts, A. G. Singal, and T. Block, "Opportunities to address gaps in early detection and improve outcomes of liver cancer," *JNCI Cancer Spectrum*, vol. 7, May 2023.

[4] T. K. Kim, E. Lee, and H.-J. Jang, "Imaging findings of mimickers of hepatocellular carcinoma," *Clinical and Molecular Hepatology*, vol. 21, no. 4, p. 326, 2015.

[5] F. Vernuccio, G. Porrello, R. Cannella, L. Vernuccio, M. Midiri, L. Giannitrapani, M. Soresi, and G. Brancatelli, "Benign and malignant mimickers of infiltrative hepatocellular carcinoma: tips and tricks for differential diagnosis on ct and mri," *Clinical Imaging*, vol. 70, p. 33–45, Feb. 2021.

[6] R. Nevola, R. Ruocco, L. Criscuolo, A. Villani, M. Alfano, D. Beccia, S. Imbriani, E. Claar, D. Cozzolino, F. C. Sasso, A. Marrone, L. E. Adinolfi, and L. Rinaldi, "Predictors of early and late hepatocellular carcinoma recurrence," *World Journal of Gastroenterology*, vol. 29, pp. 1243–1260, 3 2023.

[7] J. C. Ahn, T. A. Qureshi, D. Li, A. G. Singal, and J. D. Yang, "Deep learning in hepatocellular carcinoma: Current status and future perspectives," *World Journal of Hepatology*, vol. 13, pp. 2039–2051, 2021.

[8] D.-M. Koh, N. Papanikolaou, U. Bick, R. Illing, C. E. Kahn, J. Kalpathi-Cramer, C. Matos, L. Martí-Bonmatí, A. Miles, S. K. Mun, S. Napel, A. Rockall, E. Sala, N. Strickland, and F. Prior, "Artificial intelligence and machine learning in cancer imaging," *Communications Medicine*, vol. 2, Oct. 2022.

[9] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, and P. N. Ramkumar, "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Current Reviews in Musculoskeletal Medicine*, vol. 13, pp. 69–76, 2 2020.

[10] M. Y. Ansari, A. Abdalla, M. Y. Ansari, M. I. Ansari, B. Malluhi, S. Mohanty, S. Mishra, S. S. Singh, J. Abinahed, A. Al-Ansari, S. Balakrishnan, and S. P. Dakua, "Practical utility of liver segmentation methods in clinical surgeries and interventions," *BMC Medical Imaging*, vol. 22, May 2022.

[11] M. Ni, X. Zhou, Q. Lv, Z. Li, Y. Gao, Y. Tan, J. Liu, F. Liu, H. Yu, L. Jiao, and G. Wang, "Radiomics models for diagnosing microvascular invasion in hepatocellular carcinoma: Which model is the best model?," *Cancer Imaging*, vol. 19, 8 2019.

[12] E. Harding-Theobald, J. Louissaint, B. Maraj, E. Cuaresma, W. Townsend, M. Mendiratta-Lala, A. G. Singal, G. L. Su, A. S. Lok, and N. D. Parikh, "Systematic review: radiomics for the diagnosis and prognosis of hepatocellular carcinoma," *Alimentary Pharmacology and Therapeutics*, vol. 54, pp. 890–901, 10 2021.

[13] G. W. Ji, F. P. Zhu, Q. Xu, K. Wang, M. Y. Wu, W. W. Tang, X. C. Li, and X. H. Wang, "Radiomic features at contrast-enhanced ct predict recurrence in early stage hepatocellular carcinoma: A multi-institutional study," *Radiology*, vol. 294, pp. 568–579, 2020.

[14] Q. Y. Shan, H. T. Hu, S. T. Feng, Z. P. Peng, S. L. Chen, Q. Zhou, X. Li, X. Y. Xie, M. D. Lu, W. Wang, and M. Kuang, "Ct-based peritumoral radiomics signatures to predict early recurrence in hepatocellular carcinoma after curative tumor resection or ablation," *Cancer Imaging*, vol. 19, 2 2019.

[15] Y. Zhao, J. Wu, Q. Zhang, Z. Hua, W. Qi, N. Wang, T. Lin, L. Sheng, D. Cui, J. Liu, Q. Song, X. Li, T. Wu, Y. Guo, J. Cui, and A. Liu, "Radiomics analysis based on multiparametric mri for predicting early recurrence in hepatocellular carcinoma after partial hepatectomy," *Journal of Magnetic Resonance Imaging*, vol. 53, pp. 1066–1079, 2021.

[16] W. Gao, W. Wang, D. Song, C. Yang, K. Zhu, M. Zeng, S. xiang Rao, and M. Wang, "A predictive model integrating deep and radiomics features based on gadobenate dimeglumine-enhanced mri for postoperative early recurrence of hepatocellular carcinoma," *Radiologia Medica*, vol. 127, pp. 259–271, 3 2022.

[17] A. Morshid, K. M. Elsayes, A. M. Khalaf, M. M. Elmohr, J. Yu, A. O. Kaseb, M. Hassan, A. Mahvash, Z. Wang, J. D. Hazle, and D. Fuentes, "A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization," *Radiology: Artificial Intelligence*, vol. 1, 9 2019.

[18] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, p. 203–211, Dec. 2020.

[19] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[20] T. M. Buzug, *Computed Tomography*, p. 311–342. Springer Berlin Heidelberg, 2011.

[21] A. C. of Radiology Committee on LI-RADS® (Liver), "Li-rads® populations: Surveillance, diagnosis, staging, treatment response."

[22] P. J. Withers, C. Bouman, S. Carmignato, V. Cnudde, D. Grimaldi, C. K. Hagen, E. Maire, M. Manley, A. D. Plessis, and S. R. Stock, "X-ray computed tomography," *Nature Reviews Methods Primers*, vol. 1, 12 2021.

[23] R. L. Baron, "Understanding and optimizing use of contrast material for ct of the liver.," *American Journal of Roentgenology*, vol. 163, pp. 323–331, 1994. PMID: 8037023.

[24] N. Papanikolaou, C. Matos, and D. M. Koh, "How to develop a meaningful radiomic signature for clinical use in oncologic patients," *Cancer Imaging*, vol. 20, May 2020.

[25] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. Cook, "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, pp. 488–495, 4 2020.

[26] J. J. V. Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, pp. e104–e107, 11 2017.

[27] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. Cook, C. Davatzikos, A. Depeursinge, M. C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. E. Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh, M. Götz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K. H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. Pfaehler, A. Rahmim, A. U. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. S. Fernandez, E. Spezi, R. J. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E. G. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. van Velden, P. Whybra, C. Richter, and S. Löck, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, pp. 328–338, 5 2020.

[28] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 5 2015.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *Biomedical Signal Processing and Control*, vol. 84, p. 104791, July 2023.

[33] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, 9 2022.

[34] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[35] B. D. Muzio and M. Morgan, "Liver cancer (bclc staging)," *Radiopaedia.org*, 2 2015.

[36] A. W. Moawad, A. Morshid, A. M. Khalaf, M. M. Elmohr, J. D. Hazle, D. Fuentes, M. Badawy, A. O. Kaseb, M. Hassan, A. Mahvash, J. Szklaruk, A. Qayyum, A. Abusaif, W. C. Bennett, T. S. Nolan, B. Camp, and K. M. Elsayes, "Multimodality annotated hepatocellular carcinoma data set including pre- and post-tace with imaging segmentation," *Scientific Data*, vol. 10, Jan. 2023.

[37] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, p. 1045–1057, July 2013.

[38] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, *et al.*, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, 2023.

[39] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, pp. 1116–1128, 7 2006.

[40] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research," *Journal of Digital Imaging*, vol. 31, pp. 290–303, 6 2018.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[42] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 6 2014.

[43] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, 8 2015.

[44] T. Heimann, B. van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Cordova, B. Dawant, M. Fidrich, J. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmuller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Nemeth, D. Raicu, A.-M. Rau, E. van Rikxoort, M. Rousson, L. Rusko, K. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE Transactions on Medical Imaging*, vol. 28, p. 1251–1265, Aug. 2009.

[45] O. Maier, A. Rothberg, P. R. Raamana, R. Bèges, F. Isensee, M. Ahern, Mamrehn, VincentXWD, and J. Joshi, "loli/medpy: Medpy 0.4.0," 2019.

[46] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[47] R. T. Leijenaar, G. Nalbantov, S. Carvalho, W. J. V. Elmpt, E. G. Troost, R. Boellaard, H. J. Aerts, R. J. Gillies, and P. Lambin, "The effect of suv discretization in quantitative fdg-pet radiomics: The need for standardized methodology in tumor texture analysis," *Scientific Reports*, vol. 5, 8 2015.

[48] F. Tixier, C. C. L. Rest, M. Hatt, N. Albarghach, O. Pradier, J. P. Metges, L. Corcos, and D. Visvikis, "Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer," *Journal of Nuclear Medicine*, vol. 52, pp. 369–378, 3 2011.

[49] E.-S. Cho and J.-Y. Choi, "Mri features of hepatocellular carcinoma related to biologic behavior," *Korean journal of radiology*, vol. 16, no. 3, p. 449, 2015.

[50] X. Zhang, Y. Xin, Y. Chen, and X. Zhou, "Prognostic effect of albumin-to-alkaline phosphatase ratio on patients with hepatocellular carcinoma: a systematic review and meta-analysis," *Scientific Reports*, vol. 13, Jan. 2023.

[51] A. Rodrigues, N. Rodrigues, J. Santinha, M. V. Lisitskaya, A. Uysal, C. Matos, I. Domingues, and N. Papanikolaou, "Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness," *Scientific Reports*, vol. 13, Apr. 2023.

[52] W. Abdelhamed and M. El-Kassas, "Hepatocellular carcinoma recurrence: Predictors and management," *Liver Research*, vol. 7, p. 321–332, Dec. 2023.

[53] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into Imaging*, vol. 11, Aug. 2020.

# Chapter A

# Supplementary Material

## A.1 Patient characteristics

Table A.1: Clinical characteristics of entire patient cohort with arterial phase CT scans, alongside data partitioning for train and test sets of arterial phase nnU-Net model.

| **Variable** | **Total** | **Train Set** | **Test Set** | **p-value** |
|---|---|---|---|---|
| Male sex | 177 (79%) | 142 (79%) | 35 (78%) | 1.000 |
| Metabolic syndrome | 77 (34%) | 61 (34%) | 16 (36%) | 0.972 |
| Hepatitis B | 60 (27%) | 48 (27%) | 12 (27%) | 1.000 |
| Hepatitis C | 55 (24%) | 44 (24%) | 11 (24%) | 1.000 |
| HIV | 4 (2%) | 4 (2%) | 0 (0%) | 0.705 |
| Alcohol abuse | 41 (18%) | 35 (19%) | 6 (13%) | 0.463 |
| Genetic pathology | 9 (4%) | 5 (3%) | 4 (9%) | 0.148 |
| Porphyria | 2 (1%) | 2 (1%) | 0 (0%) | 0.777 |
| Liver disease associated with reduced portal blood | 3 (1%) | 3 (2%) | 0 (0%) | 0.884 |
| Healthy liver | 26 (12%) | 23 (13%) | 3 (7%) | 0.375 |
| F1 Fibrosis | 37 (16%) | 31 (17%) | 6 (13%) | 0.740 |
| F2 Fibrosis | 47 (21%) | 37 (21%) | 10 (22%) | 0.819 |
| F3 Fibrosis | 52 (23%) | 39 (22%) | 13 (29%) | 0.446 |
| Cirrhosis (F4 Fibrosis) | 72 (32%) | 62 (34%) | 10 (22%) | 0.177 |
| Macroscopic capsule | 160 (71%) | 127 (71%) | 33 (73%) | 0.854 |
| Nodules satellites | 46 (20%) | 38 (21%) | 8 (18%) | 0.354 |
| Vascular invasion | 124 (55%) | 101 (56%) | 23 (51%) | 0.701 |
| Microscopic capsule | 173 (77%) | 138 (77%) | 35 (78%) | 0.986 |
| Differentiation WHO: | | | | 0.589 |
| I | 69 (31%) | 55 (31%) | 14 (31%) | |
| II | 143 (64%) | 116 (64%) | 27 (60%) | |
| III | 13 (6%) | 9 (5%) | 4 (9%) | |
| Differentiation Edmondson: | | | | 0.781 |
| I-II | 69 (31%) | 55 (31%) | 14 (31%) | |
| III-IV | 156 (70%) | 125 (69%) | 31 (69%) | |
| Alkaline phosphatase | 91 (71.0,123.0) | 89.5 (70.75,125.0) | 98.0 (73.0,118.0) | 0.888 |
| Age (years) | 64.0 (56.0,70.0) | 64.0 (56.75,70.0) | 66.0 (54.0,71.0) | 0.563 |
| Albumin (g/L) | 38.0 (34.0,41.0) | 37.1 (33.0,41.0) | 38.0 (35.0,40.3) | 0.756 |
| AST (IU/L) | 44.5 (32.25,72.0) | 44.0 (33.0,72.0) | 46.0 (31.0,85.0) | 0.955 |
| International normalised ratio | 1.0(0.98,1.1) | 1.0(1.0,1.1) | 1.0(1.0,1.1) | 0.5 |
| Total bilirubin (mg/L) | 9.0 (7.0,12.0) | 10.0 (7.0,12.0) | 8.0 (6.0,11.0) | 0.134 |
| Tumour diameter (cm) | 4.10 (2.8,7.5) | 4.1 (2.8,7.6) | 42.0 (29.0,55.0) | 0.580 |
| Preoperative protrombin time | 94.5 (81.0,102.0) | 94.0 (80.0,102.0) | 95.0 (82.0,102.0) | 0.661 |
| Weight | 74.5 (64.0,84.8) | 75.0 (65.5,87.0) | 70.0 (60.5,79.5) | 0.030 |
| Gamma glutamil transferase | 81.0(45.25,171.5) | 80.0 (47.0,165.0) | 82.0 (42.0,268.0) | 0.827 |
| Height | 172.0 (165.0,177.0) | 173.0 (165.0,178.0) | 168.5 (160.8,173.8) | 0.007 |
| BMI | 25.59 (22.7,29.3). | 25.82 (23.1,29.4) | 24.4 (22.2,28.9) | 0.322 |
| ALT (IU/L) | 43.0 (27.0,67.8) | 43.0 (28.0,67.0) | 41.0(25.0,71.0) | 0.657 |
| Creatinine ($\mu$mol/L) | 80.0 (71.0,95.0) | 79.0 (71.0,95.0) | 80.0 (71.0,96.0) | 0.920 |
| Platelet count ($\times 10^{-3}$ $\mu L$) | 209 (165.5,258.0) | 204 (155.5,247.5) | 241.5 (184.0,289.0) | 0.016 |
| $\alpha$-Fetoprotein (ng/mL) | 9.0 (4.0,179.5) | 8.0 (3.0,209.0) | 13.0 (5.0,26.7) | 0.905 |

Continuous variables are displayed as medians and 25% and 75% percentiles in parathesis, with p-values calculated for a Mann-Whitney U test. Discrete variables are displayed as count and percentages in paranthesis, with p-values calculated for a z-test. HIV: Human Immunodeficiency Virus; WHO: World Health Organisation; AST: Aspartate Aminotransferase; BMI: Body Mass Index; ALT: Alanine Transaminase.

Table A.2: Clinical characteristics of entire patient cohort with portal venous phase CT scans, alongside data partitioning for train and test sets of portal venous phase nnU-Net model.

| **Variable** | **Total** | **Train Set** | **Test Set** | **p-value** |
|---|---|---|---|---|
| Male sex | 179 (79%) | 142 (78%) | 37 (80%) | 0.877 |
| Metabolic syndrome | 79 (35%) | 65 (36%) | 14 (30%) | 0.618 |
| Hepatitis B | 61 (27%) | 46 (25%) | 15 (33%) | 0.414 |
| Hepatitis C | 56 (25%) | 45 (25%) | 11 (24%) | 1.000 |
| HIV | 4 (2%) | 2 (1%) | 2 (4%) | 0.384 |
| Alcohol abuse | 41 (18%) | 34 (19%) | 7 (15%) | 0.740 |
| Genetic pathology | 9 (4%) | 6 (3%) | 3 (7%) | 0.562 |
| Porphyria | 2 (1%) | 2 (1%) | 0 (0%) | 0.775 |
| Liver disease associated with reduced portal blood | 3 (1%) | 3 (2%) | 0 (0%) | 0.879 |
| Healthy liver | 25 (11%) | 21 (12%) | 4 (9%) | 0.774 |
| F1 Fibrosis | 36 (16%) | 33 (18%) | 3 (7%) | 0.100 |
| F2 Fibrosis | 48 (21%) | 37 (20%) | 11 (24%) | 0.702 |
| F3 Fibrosis | 52 (23%) | 38 (21%) | 14 (30%) | 0.276 |
| Cirrhosis (F4 Fibrosis) | 75 (33%) | 61 (34%) | 14 (30%) | 0.861 |
| Macroscopic capsule | 161 (71%) | 126 (69%) | 35 (76%) | 0.465 |
| Nodules satellites | 47 (21%) | 39 (21%) | 8 (17%) | 0.783 |
| Vascular invasion | 127 (56%) | 102 (56%) | 25 (54%) | 0.455 |
| Microscopic capsule | 174 (76%) | 137 (75%) | 37 (80%) | 0.760 |
| Différenciation WHO | | | | 0.563 |
| I | 70 (31%) | 53 (29%) | 17 (37%) | |
| II | 144 (63%) | 118 (65%) | 26 (57%) | |
| III | 14 (6%) | 11 (6%) | 3 (7%) | |
| Différenciation Edmondson | | | | 0.643 |
| I-II | 70 (31%) | 53 (29%) | 17 (37%) | |
| III-IV | 158 (69%) | 129 (71%) | 29 (64%) | |
| Platelet count | 208.5 (165.2,258.0) | 204.0 (165.0,256.0) | 226.0 (174.0,273.0) | 0.158 |
| Alkaline phosphatase | 90.0 (71.0,122.5) | 92.0 (70.0,127.3) | 86.0 (73.0,113.0) | 0.696 |
| Total bilirubin (mg/L) | 10.0 (7.0,12.0) | 10.0 (7.0,14.0) | 9.0 (6.0,11.0) | 0.128 |
| Age (years) | 64.0 (55.0,70.0) | 64.0 (57.0,70.0) | 65.5 (53.0,71.0) | 0.956 |
| Weight | 75.0 (64.0,85.0) | 75.0 (66.0,86.0) | 71.5 (60.8,82.0) | 0.110 |
| AST (IU/L) | 45.0 (32.8,72.3) | 45.0 (32.0,73.5) | 44.0 (33.0,61.0) | 0.815 |
| BMI | 25.6 (22.8,29.4) | 25.8 (23.1,29.5) | 24.34 (22.06,28.3) | 0.144 |
| Creatinine ($\mu$mol/L) | 80.0 (71.0,95.0) | 79.0 (71.0,95.0) | 81.0 (70.0,98.0) | 0.752 |
| Albumin (g/L) | 38.0 (34.0,41.0) | 37.0 (33.0,41.0) | 38.0 (35.3,41.0) | 0.372 |
| ALT (IU/L) | 43.5 (27.8,69.5) | 44.0 (27.0,69.0) | 42.0 (28.0,71.0) | 0.846 |
| $\alpha$-Fetoprotein (ng/mL) | 8.0 (3.4,176.0) | 7.25 (3.0,167.75) | 14.0 (4.0,145.5) | 0.395 |
| Gamma glutamil transferase | 82.0 (45.75,172.8) | 85.0 (48.5,181.5) | 61.0 (39.0,157.0) | 0.258 |
| Height | 172.0 (165.0,177.0) | 172.0 (165.0,177.8) | 170.5 (165.0,176.0) | 0.681 |
| International normalised ratio | 1.0 (1.0,1.1) | 1.0 (1.0,1.1) | 1.0 (1.0,1.1) | 0.510 |
| Tumour diameter (cm) | 41.0 (28.0,73.5) | 41.0 (28.0,75.8) | 41.5 (29.3,64.8) | 0.659 |
| Preoperative protrombin time | 94.0 (81.0,102.0) | 94.0 (81.0,102.5) | 93.5 (79.5,100.0) | 0.440 |

Continuous variables are displayed as medians and 25% and 75% percentiles in parathesis, with p-values calculated for a Mann-Whitney U test. Discrete variables are displayed as count and percentages in paranthesis, with p-values calculated for a z-test. HIV: Human Immunodeficiency Virus; WHO: World Health Organisation; AST: Aspartate Aminotransferase; BMI: Body Mass Index; ALT: Alanine Transaminase.

Table A.3: TCIA patient characteristics of arterial and portal venous test sets.

| **Variable** | **Arterial Test Set** | **Portal Venous Test Set** |
|---|---|---|
| # Scans | 33 | 53 |
| Hepatitis C | 14 (42%) | 22 (42%) |
| Hepatitis B | 10 (30%) | 15 (28%) |
| Male sex | 21 (64%) | 35 (66%) |
| Alcohol | 18 (55%) | 28 (53%) |
| Family history cancer | 22 (67%) | 32 (60%) |
| Family history liver cancer | 3 (9%) | 6 (11%) |
| Diabetes | 11 (33%) | 16 (30%) |
| Personal history cancer | 6 (18%) | 9 (17%) |
| Cirrhosis | 26 (79%) | 40 (75%) |
| Child-Pugh score: | | |
| A | 23 (70%) | 44 (83%) |
| B | 9 (27%) | 8 (15%) |
| C | 1 (3%) | 1 (2%) |
| Uninodular Tumour | 12 (36%) | 24 (45%) |
| Vascular invasion | 7 (21%) | 11 (21%) |
| Metastasis | 3 (9%) | 5 (9%) |
| Lymphnodes | 2 (6%) | 8 (15%) |
| Portal vein thrombosis | 5 (15%) | 8 (15%) |
| CLIP score: | | |
| 0-2 | 26 (79%) | 42 (79%) |
| 3 | 5 (15%) | 8 (15%) |
| 4-6 | 2 (6%) | 3 (6%) |
| BCLC stage: | | |
| A | 0 (0%) | 4 (8%) |
| B | 9 (27%) | 13 (25%) |
| C | 23 (70%) | 35 (66%) |
| D | 1 (3%) | 1 (2%) |
| Age (years) | 71.0 (58.0, 77.0) | 68.5 (59.0, 77.0) |
| Tumour diameter (cm) | 6.0 (4.0, 10.0) | 6.0 (4.0, 10.0) |

Continuous variables are displayed as medians and 25% and 75% percentiles in parathesis. Discrete variables are displayed as count and percentages in paranthesis. BCLC: Barcelona Clinic Liver Cancer.

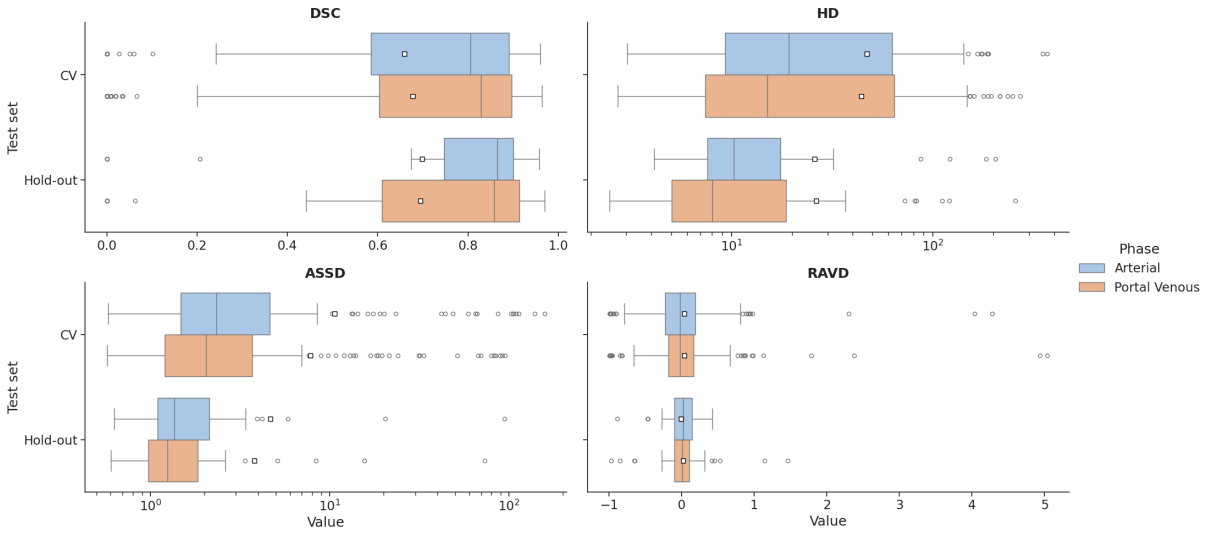
## A.2 Association between lesion segmentation model performance and clinical characteristics of patients

Table A.4: Association between all clinical variables and tumour segmentation models performance.

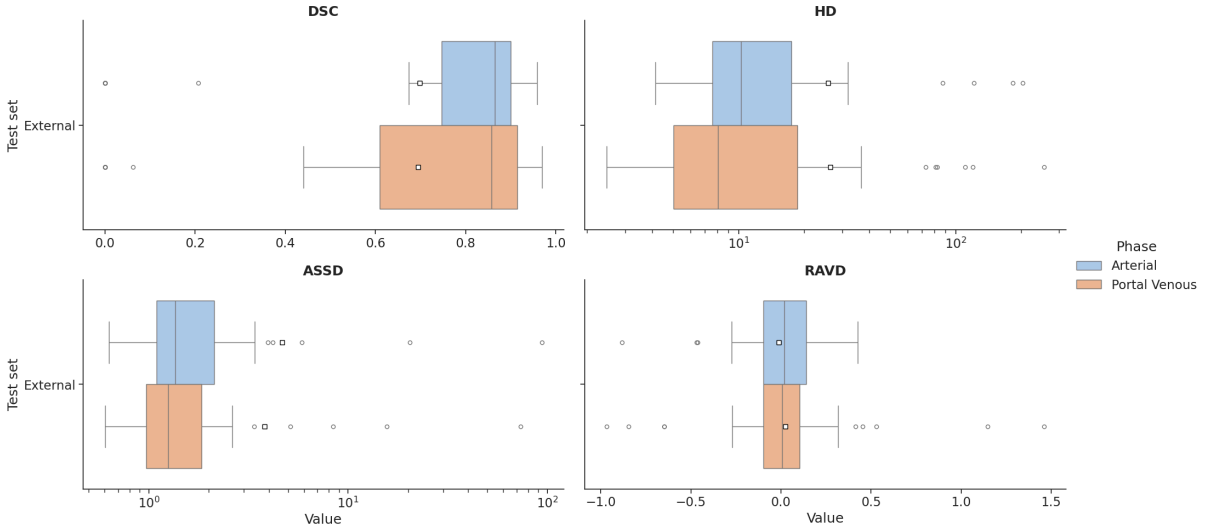
| | **nnU-Net** | | **MedSam** | |
|---|---|---|---|---|
| **Variable** | **AP p-value** | **PVP p-value** | **AP p-value** | **PVP p-value** |
| Male sex | 1.000 | 0.807 | 1.000 | 0.896 |
| Metabolic syndrome | 0.815 | 1.000 | 1.000 | 0.592 |
| Hepatitis B | 0.933 | 1.000 | 0.09 | 1.000 |
| Hepatitis C | 0.795 | 1.000 | 1.000 | 1.000 |
| HIV | 1.000 | 0.843 | 1.000 | 1.000 |
| Alcohol abuse | 0.742 | 1.000 | 0.959 | 1.000 |
| Genetic pathology | 1.000 | 1.000 | 1.000 | 0.462 |
| Porphyria | 1.000 | 1.000 | 1.000 | 1.000 |
| Liver disease associated with reduced portal blood | 1.000 | 1.000 | 1.000 | 1.000 |
| Healthy liver | 0.881 | 0.709 | 1.000 | 1.000 |
| F1 Fibrosis | 0.443 | 0.896 | 0.959 | 1.000 |
| F2 Fibrosis | 0.654 | 0.570 | 0.441 | 0.273 |
| F3 Fibrosis | 1.000 | 0.071 | 1.000 | 0.592 |
| Cirrhosis (F4 Fibrosis) | 0.179 | <0.001 | 0.624 | 1.000 |
| Macroscopic capsule | 0.354 | 0.041 | 0.09 | 0.273 |
| Nodule satellites | 1.000 | 0.810 | 0.772 | 0.65 |
| Vascular invasion | 0.066 | 0.014 | 0.101 | 0.706 |
| Microscopic capsule | 0.104 | 0.009 | 0.021 | 0.092 |
| WHO differentiation | 0.806 | 0.664 | 0.617 | 0.87 |
| Edmondson differentiation | 0.866 | 0.791 | 0.67 | 0.922 |
| Gamma glutamil transferase | 0.733 | 0.660 | 0.842 | 0.682 |
| Platelets | 0.100 | 0.091 | 0.713 | 0.481 |
| ALT (IU/L) | 0.650 | 0.542 | 0.675 | 0.785 |
| Tumor size at CT | 0.000 | 0.003 | 0.008 | 0.01 |
| Total bilirubin (mg/L) | 0.120 | 0.449 | 0.109 | 0.384 |
| BMI | 0.705 | 0.378 | 0.554 | 0.37 |
| Height | 0.205 | 0.988 | 0.898 | 0.25 |
| International normalised ratio | 0.313 | 0.617 | 0.704 | 0.705 |
| Creatinine ($\mu$mol/L) | 0.639 | 0.910 | 0.216 | 0.600 |
| $\alpha$-Fetoprotein (ng/mL) | 0.318 | 0.706 | 0.855 | 0.843 |
| AST (IU/L) | 0.532 | 0.342 | 0.563 | 0.982 |
| Age (years) | 0.125 | 0.463 | 0.590 | 0.306 |
| Preoperative protrombin time | 0.334 | 0.750 | 0.536 | 0.738 |
| Alkaline phosphatase | 0.021 | 0.776 | 0.098 | 0.246 |
| Weight | 0.839 | 0.407 | 0.983 | 0.944 |
| Albumin (g/L) | 0.321 | 0.120 | 1.000 | 0.945 |

P-values were calculated for a chi-square test of independence or a Mann-Whitney U test, for all categorical variables and continuous variables, respectively. AP: Arterial Phase; PVP: Portal Venous Phase; HIV: Human Immunodeficiency Virus; WHO: World Health Organisation; AST: Aspartate Aminotransferase; BMI: Body Mass Index; ALT: Alanine Transaminase.

## A.3 Performance Metrics of nnU-Net Segmentation Models Following Mask Post-Processing




(a) Internal Test Data (Beaujon)




(b) External Test Data (TCIA)

Figure A.1: Distribution of Performance metrics of nnU-Net segmentation models, following predicted mask post-processing. AP: Arterial Phase; PVP: Portal Venous Phase; DSC: Dice Similarity Coefficient; HD: Hausdorff Distance; ASSD: Average Symmetric Surface Distance; RAVD: Relative Absolute Volume Difference.

## A.4 Recurrence proportions

Table A.5: Recurrence proportions across train and test groups of radiomics-based recurrence risk prediction.

| | Recurrence | | No recurrence | | |
|---|---|---|---|---|---|
| Data | Train | Test | Train | Test | p-value |
| Clinical | 62 (40.26%) | 26 (35.62%) | 92 (59.74%) | 47 (64.38%) | 0.226 |
| Tumour | 61 (40.67%) | 25 (36.23%) | 89 (59.33%) | 44 (63.77%) | 0.635 |
| Liver | 53 (38.69%) | 24 (38.71%) | 84 (61.31%) | 38 (61.29%) | 1.000 |

P-values were calculated for a chi-square test of independence.

## A.5 Dataset composition of recurrence risk prediction models

**Clinical:** Age, sex, height, body mass index, metabolic syndrome, Hepatitis B virus infection, Hepatitis C virus infection, Human immunodeficeny virus infection, alcohol abuse, genetic pathology, liver disease associated to reduced portal blood, normal liver, international normalised ratio, Total bilirubin (mg/L), platelets, Albumin (g/L), Creatinine ($\mu$mol/L), AST (IU/L) level, ALT (IU/L) level, gamma glutamil transferase, alcaline phospatase, $\alpha$-Fetoprotein (ng/mL), multinodularity, CT scanner model LightSpeed VCT, CT scanner model Revolution GSI, CT scanner model Revolution HD.

**Tumour:** Original_shape_Elongation_arterial, original_shape_Sphericity_arterial, original_glszm_SmallAreaEmphasis_arterial, wavelet-LLH_firstorder_10Percentile_arterial, wavelet-LLH_glcm_Correlation_arterial, wavelet-LHL_firstorder_Kurtosis_arterial, wavelet-LHL_firstorder_Median_arterial, wavelet-LHL_glcm_MCC_arterial, wavelet-LHH_firstorder_Kurtosis_arterial, wavelet-LHH_firstorder_Median_arterial, wavelet-HLL_glcm_ClusterShade_arterial, wavelet-HLL_glcm_MCC_arterial, wavelet-HLH_glcm_ClusterShade_arterial, wavelet-HHL_firstorder_Median_arterial, wavelet-HHL_firstorder_RootMeanSquared_arterial, wavelet-HHL_firstorder_Skewness_arterial, wavelet-HHL_glcm_Correlation_arterial, wavelet-HHL_glcm_MCC_arterial, wavelet-HHL_glszm_SizeZoneNonUniformityNormalized_arterial, wavelet-HHH_firstorder_RootMeanSquared_a, wavelet-HHH_firstorder_Skewness_arterial, wavelet-HHH_glcm_ClusterShade_arterial, wavelet-HHH_glcm_Imc2_arterial, square_glcm_InverseVariance_a, square_gldm_LargeDependenceHighGrayLevelEmphasis_arterial, square_gldm_SmallDependenceLowGrayLevelEmphasis_arterial, square_ngtdm_Strength_arterial, squareroot_glszm_LargeAreaLowGrayLevelEmphasis_arterial, logarithm_firstorder_Maximum_arterial, logarithm_glcm_Idmn_arterial, exponential_firstorder_Minimum_arterial, exponential_glcm_ClusterShade_arterial, exponential_glcm_Imc1_arterial, exponential_glrlm_LongRunLowGrayLevelEmphasis_arterial, gradient_glcm_Correlation_arterial, lbp-3D-m1_glcm_ClusterShade_arterial, lbp-3D-m2_firstorder_Maximum_arterial, lbp-3D-m2_glszm_SmallAreaHighGrayLevelEmphasis_arterial, lbp-3D-m2_gldm_LowGrayLevelEmphasis_arterial, lbp-3D-k_firstorder_Skewness_arterial, lbp-3D-k_glcm_MaximumProbability_arterial, lbp-3D-k_glszm_GrayLevelNonUniformityNormalized_arterial, lbp-3D-k_glszm_SmallAreaLowGrayLevelEmphasis_arterial, log-sigma-1-0-mm-3D_glszm_GrayLevelVariance_arterial, log-sigma-1-0-mm-3D_glszm_SizeZoneNonUniformityNormalized_arterial, log-sigma-2-0-mm-3D_glszm_SmallAreaEmphasis_arterial, log-sigma-4-0-mm-3D_glcm_Imc2_arterial, log-sigma-4-0-mm-3D_glcm_MaximumProbability_arterial, log-sigma-4-0-mm-3D_glrlm_RunVariance_arterial, log-sigma-5-0-mm-3D_firstorder_InterquartileRange_arterial, log-sigma-5-0-mm-3D_firstorder_Skewness_arterial, log-sigma-5-0-mm-3D_glcm_Correlation_arterial, log-sigma-5-0-mm-

3D_glszm_SmallAreaLowGrayLevelEmphasis_arterial, log-sigma-5-0-mm-3D_gldm_DependenceVariance_arterial, log-sigma-5-0-mm-3D_ngtdm_Complexity_arterial, log-sigma-5-0-mm-3D_ngtdm_Contrast_arterial, age, weight (kg), human immunodeficeny virus infection, genetic pathology, liver disease associated to reduced portal blood, normal liver, Total bilirubin (mg/L), platelets, Albumin (g/L), Creatinine ($\mu$mol/L), ALT (IU/L) level, gamma glutamil transferase, alcaline phospatase, $\alpha$-Fetoprotein (ng/mL), multinodularity, CT scanner model Revolution HD.

**Liver:** Wavelet-LLL_glcm_ClusterShade_arterial, square_glszm_GrayLevelNonUniformityNormalized exponential_glcm_ClusterShade_arterial gradient_glcm_SumSquares_arterial log-sigma-4-0-mm-3D_firstorder_Minimum_arterial log-sigma-5-0-mm-3D_ngtdm_Contrast_pv, age, sex, height (cm) wright (kg), BMI, metabolic syndrome, hepatitis B, hepatitis C, alcohol abuse, genetic pathology, liver disease associated to reduced portal blood, normal liver, preoperatory protrombin time, Total bilirubin (mg/L), platelets, Albumin (g/L), Creatinine ($\mu$mol/L), AST (IU/L) level, ALT (IU/L) level, gamma glutamil transferase, alcaline phospatase, $\alpha$-Fetoprotein (ng/mL), tumour diameter, CT scanner model LightSpeed VCT, CT scanner model Revolution GSI, CT scanner model Revolution HD.

UNIVERSIDADE NOVA
DE LISBOA