Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Jakob Nolte (7587570)
TITLE: Early Stage Hepatocellular Carcinoma Diagnosis Using a 3D Convolutional Neural Network Ensemble
May 2023

Supervisors:
Dr. Stephanie van den Berg (University of Twente)
Dr. Maryam Amir Haeri (University of Twente)

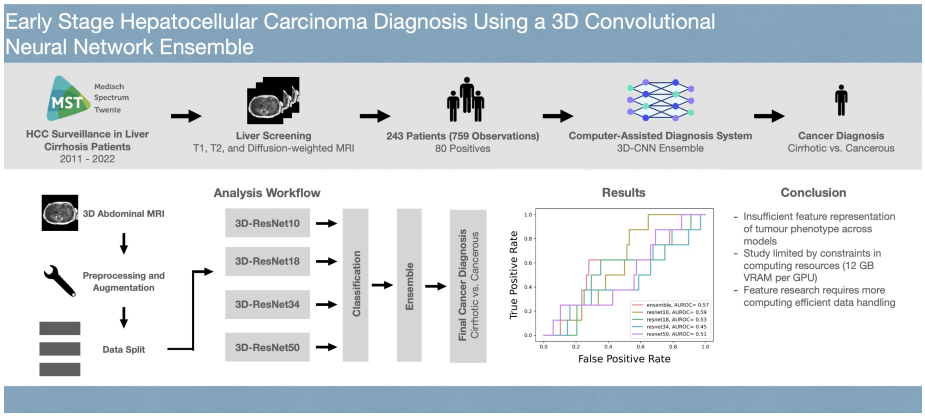
Second grader:
Prof. Dr. Herbert Hoijtink (Utrecht University)

Preferred journal of publication: Medical Image Analysis
Word count: 5105

Graphical Abstract

Early Stage Hepatocellular Carcinoma Diagnosis Using a 3D Convolutional Neural Network Ensemble

Jakob Nolte



Highlights

Early Stage Hepatocellular Carcinoma Diagnosis Using a 3D Convolutional Neural Network Ensemble

Jakob Nolte

- Training convolutional neural networks on 3-dimensional magnetic resonance images can enhance accurate liver tumor diagnosis but comes at a high computational expense.
- The loss of information induced by down-sampling 3-dimensional whole volume images proves to substantially subvert accurate representation of liver tumor's phenotype.
- Combining object detection and classification networks may yield more precise tumor diagnosis in the context of limited computational resources.

Early Stage Hepatocellular Carcinoma Diagnosis Using a 3D Convolutional Neural Network Ensemble

Jakob Nolte[a]

[a] *Utrecht University, Heidelberglaan 8, Utrecht, 3528 CS, Utrecht, The Netherlands*

Abstract

For hepatocellular carcinoma (HCC) patients, early-stage diagnosis continues to be the most important predictor of survival. However, lesion detection typically only occurs at an advanced stage and continues to be largely affected by radiologists' experience. To facilitate patient survival, this study proposed a novel computer-aided diagnosis (CAD) system using data on 243 patients (759 observations) with compensated liver cirrhosis. Particularly, four individual 3-dimensional convolutional neural networks were trained, and their predictions aggregated in an ensemble to further enhance accurate tumor diagnosis. The models were validated on an independent internal test set, and their performance was evaluated with respect to a number of standard evaluation metrics. Despite the promising findings attained in our preliminary investigation, the results showed that all four individual models displayed limited diagnostic capabilities, with their predictions remaining largely inferior to the naive prediction of the study's majority class. Meanwhile, the ensemble of the four models displayed greater discriminatory power between HCC and common liver cirrhosis but still failed to accurately infer cancer prevalence. While the proposed CAD thus yielded limited diagnostic value of HCC, its accurate evaluation may have been severely constrained by the study's limitations in computing resources, thus highlighting the need for future research to reevaluate the CAD's viability.

Keywords: Hepatocellular Carcinoma, 3D Convolutional Neural Networks, Ensemble Learning, Magnetic Resonance Imaging

1. Introduction

Hepatocellular carcinoma (HCC) is an aggressive primary liver cancer that predominantly arises in the context of liver cirrhosis. With an estimated global incidence of approximately 500.000 cases per year, the disease represents one of the leading causes of cancer-related deaths worldwide (Bray et al., 2018; McGlynn et al., 2021). However, increasing prevalence rates of liver cirrhosis preceding conditions like nonalcoholic fatty liver disease are expected to further exacerbate the disease's burden on the general population (Estes et al., 2018). The trend is particularly worrisome given that, despite recent advancements in curative treatment, prognosis of the disease remains poor (Chen et al., 2020).

In fact, the tumor's non-specific clinical presentation often delays diagnosis such that definite confirmation typically only occurs at an advanced stage, and the efficacy of curative regimens is substantially impeded (McGlynn et al., 2021).

To improve early-stage tumor detection for cirrhosis patients, international guidelines have therefore increasingly emphasized frequent ultrasonography (US) screening (European Association For The Study Of The Liver, 2018; Heimbach et al., 2018). However, even with contrast-enhanced imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI), HCC diagnosis remains strenuous as the tumor's heterogeneous phenotype, atypical radiological imaging artifacts, and the existence of other malignant tumors, such as intrahepatic cholangiocarcinoma (ICC), commonly subvert accurate diagnosis (Wolf et al., 2021). For most patients suffering from compensated cirrhosis, timely HCC diagnosis and initiation of curative measures thus continue to depend on the radiologist's experience (Mitchell et al., 2015), highlighting the need for computational decision-support systems independent of inter-operator variability.

Meanwhile, recent advancements in the medical imaging domain have led to the development of several computational approaches that aim to facilitate tumor diagnosis. Among these approaches, convolutional neural networks (CNN) have emerged as the state-of-the-art modeling technique in medical image analysis (Chen et al., 2022; Yu et al., 2021). In short, these models iteratively deduce a set of nondeterministic feature representations unique to the sample images by distorting the images through convolutional operations. In liver oncology, CNN-based modeling approaches have yielded promising results. Applied to ultrasound and computed tomography images, for instance, CNNs have been shown to accurately segment HCC from surrounding cirrhotic liver tissue (Brehar et al., 2020), to correctly differentiate benign from HCC and other malignant non-HCC liver lesions (Schmauch et al., 2019; Yang et al., 2020; Yasaka et al., 2018; Todoroki et al., 2019; Shi et al., 2020; Zhou et al., 2021; Ponnoprat et al., 2020), and to precisely classify HCC according to its severity (Li et al., 2020a). Yet, due to the often inadequate level of detail in US and CT images, more recent studies investigating HCC have increasingly focused on MRI as their primary source of imagery data. Here, CNNs have been equally shown to correctly differentiate HCC from other focal liver lesions (Hamm et al., 2019; Zhen et al., 2020; Oestmann et al., 2021), to distinguish benign from malignant liver lesions (Zhen et al., 2020), to classify the lesions based on the Liver Imaging Reporting and Data System (LI-RADS) grading system (Wu et al., 2020), and to accurately detect HCC affected regions in single MRI slices.

However, despite recent progress in CNN applications to HCC diagnosis, current studies' focus on liver tumor differentiation still fails to simulate clinical practice conditions. In fact, radiologists utilizing MRI screening for HCC surveillance are typically encountered with liver cirrhosis patients, of whom only a few have already developed HCC. Therefore, a truly effective computer-assisted HCC diagnosis (CAD) system not only requires the ability to differentiate HCC from other liver lesions but also to detect the disease in a population of cirrhosis patients comprising diverse etiologies. Beyond these constraints, most

studies investigating CNN applications to HCC diagnosis have developed their models on small-scale data sets, and all of the models trained on MR images have, to the best of our knowledge, been employed studying single-slice MR images. While this disregards the volumetric nature of an MRI, it also requires labor intensive manual annotations of single MRI slices, and, as a result, effectively discourages their continuous development as well as their later evaluation on additional data from further medical hospitals.

In its scope, this study therefore proposes a novel computer-assisted diagnosis (CAD) system of hepatocellular carcinoma in liver cirrhosis patients that is both accurate and easily applicable to a variety of clinical scenarios. More specifically, it aims first to develop and evaluate four individual 3-dimensional convolutional neural networks of varying complexity, and then aggregate their predictions in a novel ensemble of 3D-CNN models, equally evaluating its performance on an independent internal test set of liver cirrhosis patients who underwent MRI screening for HCC surveillance.

Prior to conducting this study, we tested the proposed methodology in a preliminary investigation. The investigation utilized two small-scale data sets of publicly available abdominal MRI screening data and was conducted to validate the effectiveness of the proposed 3-dimensional modeling approach, before employing it on the more extensive data set of liver cirrhosis patients. All things considered, this paper thus presents three major contributions, which are summarized as follows:

1. Its proposed 3D-CNN architecture minimizes the number of required annotations per imaging sequence to a single annotation per imaging volume, reducing doctors' required engagement to a minimum, and making it easily expandable to imaging data from further medical institutions.
2. It presents, to the best of the authors' knowledge, the first 3D-CNN architecture advanced for HCC diagnosis.
3. Accordingly, its proposed ensemble of CNN models represents the first such model developed in the context of MRI in liver oncology.

The code to reproduce all results presented in this study is publicly available.[1]

2. Materials and Methods

2.1. Data Description

This study has been approved by the ethics review board of the faculty of social and behavioral sciences at Utrecht University, the Netherlands. For the study, abdominal MR images were retroactively collected from a cohort of 325 liver cirrhosis patients who underwent MRI screening for HCC surveillance between March 2011 and September 2022 at Medisch Spectrum Twente (MST)

[1] https://github.com/jmnolte/thesis

in Enschede, the Netherlands. HCC diagnosis was provided by the hospital's senior radiologist and senior hepatologist, and treatment was initiated once the disease was detected. In addition, patients' medical history and general demographics were documented. The number of observations per patient varied between 1 and 11 (median 3 observations), and follow-up time between observations ranged from 1.5 months to 6.5 years (median 1 year).

With respect to the inclusion in this study, observations were selected given the following set of inclusion criteria. That is, (1) they had an unambiguous diagnosis of HCC, and (2) they comprised all of the following set of T1-weighted (T1W), T2-weighted (T2W), and diffusion-weighted (DWI) imaging modalities that are recommended by the LI-RADS 2018 manual on MRI screening for HCC (Mitchell et al., 2015): T1W in-phase, T1W out-of-phase, T1W dynamic contrast, T2W short echo time (80 ms), T2W long echo time (235 ms) MR images, as well as MR images with low ($b = 150$), medium ($b = 400$), and high ($b = 800$) diffusion weighting. Observations were excluded if the scans' quality was severely constrained by imaging artifacts or the practicing doctor's assessment of the lesion remained inconclusive. For this study, observations were considered to include HCC if they were graded as at least a four on the LI-RADS grading scale. All selected observations below that reference standard were considered to not comprise HCC.

Regarding their acquisition, all images were acquired in axial orientation and registered using a Philips Ingenia Elition 3-Tesla MRI scanner with varying inter-slice distances ranging between 2 and 7 mm. Dynamic contrast-enhanced T1W images were acquired utilizing breath-holding techniques and varying acquisition times ranging between 6 and 20 seconds. Conventional unenhanced MRI sequences, by contrast, were obtained utilizing respiratory-triggered techniques with acquisition times ranging from 6 to 23 seconds for the T1-weighted sequences, from 16 to 216 seconds for the T2-weighted images, and from 135 to 509 seconds for the diffusion-weighted images.

2.2. Network Architectures

For the diagnosis of HCC in liver cirrhosis patients, four different pretrained convolutional neural network architectures from the series of residual neural network (ResNet) models (He et al., 2015) were adopted and subsequently aggregated in an ensemble. Due to their effective implementation of residual connections[2], ResNet models are among the most commonly employed CNN architecture in medical imaging. However, being originally developed on 2-dimensional images, their architectural layout does not account for the volumetric nature of an MRI. This study, therefore, extended the aforementioned model archi-

[2]Residual connections enable shortcut paths that bypass one or more intermediate layers, allowing the network to learn an identity mapping and adjust the weights of the intermediate layers during training. They have been shown to be a critical component in deep learning architectures, improving the training efficiency of very deep networks and overcoming the vanishing gradient problem (Yu et al., 2021).

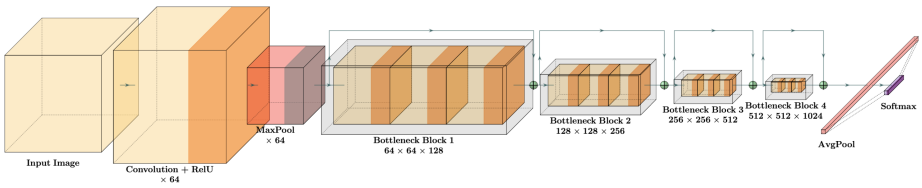


Figure 1: Exemplary 3D-ResNet Network Architecture

Note: Convolutional layers comprising of a convolutional operation, batch normalization, and a recitified linear unit (RelU) activation function are represented in orange. The network's pooling operations are depicted in red, and the architecture's fully connected softmax layer is represented in purple. In addition, the network's residual connections are represented in green.

tectures to 3D, adopting the following set of 3D-CNN models: 3D-ResNet10, 3D-ResNet18, 3D-ResNet34, and 3D-ResNet50.

Besides their proficiency, ResNet models were selected as the study's backbone architecture to mitigate the challenges posed by the limited sample size in this study. Sparse data is a common issue in deep learning applications in medical imaging, and can significantly hinder model training. To alleviate this challenge, transfer learning has therefore become the de-facto method in CNN applications in medical image analysis, which initializes the network's parameters with weights pretrained on a related task, typically yielding faster convergence and improved performance (Tajbakhsh et al., 2016; Raghu et al., 2019). However, the use of transfer learning in this study was limited by the fact that ResNet models were the only architecture for which weights pretrained on 3D medical images were available. As such, ResNet was selected as the study's backbone architecture and individual model parameters were initialized using weights from the MedicalNet project, which were pretrained on eight different medical datasets comprising varying imaging modalities (Chen et al., 2019).

Likewise, the inclusion of the varying model architectures was motivated by the inconclusive evidence on optimal model complexity. While the literature generally suggests less complex model architectures for the small-scale datasets typically encountered in clinical studies (Chen et al., 2019; Kim et al., 2022), more complex CNN architectures are generally assumed to capture more abstract representations in the training images. In addition, aggregating both less and more complex model architectures in an ensemble combines the feature representations learned by the individual models and has been shown to yield consistently higher performance in image classification tasks (Sagi and Rokach, 2018). This study therefore combined the aforementioned 3D-CNN architectures in an ensemble, feeding the individual models' predictions obtained after training into a new linear classifier.

The individual networks' general architectural layout is depicted in Figure 1. It consists of a common convolutional layer that is followed by a max pooling layer, four subsequent sets of bottleneck blocks (i.e., four subsequent basic blocks in models with less than 50 convolutional layers), a global average pooling layer

and one fully connected softmax layer but varies with respect to the number of convolutional layers represented by the networks' suffix. Note that the figure only represents a general approximation of the individual networks' design. For each of the individual networks, the number of operations denoted below their descriptions thus varies according to the number of convolutional layers.

2.3. Experimental Setup

The modeling approach in this study was twofold. In the first part, the aforementioned individual 3D-CNN architectures were finetuned on the training set of abdominal MRI scans using pretrained weights from MedicalNet (Chen et al., 2019) to initialize the networks' parameters and updating all of the models' parameters over the course of the training process. In the second part, all finetuned models were aggregated in an ensemble, and the ensemble's newly constructed linear classifier was subsequently retrained using the finetuned 3D-CNNs as fixed feature extractors.

2.3.1. Preprocessing

Prior to modeling, minor preprocessing steps were required on the raw images. That is, all eligible MRI sequences were converted from Digital Imaging and Communications in Medicine (DICOM) to Neuroimaging Informatics Technology Initiative (nifti) file format, the resulting volumes were reorientated to have posterior, left, inferior (PLI) orientation, and each imaging modality was rescaled to 96x96x96 voxels. In addition, potential distortions caused by pixel value outliers were eliminated by truncating the pixel intensity values per imaging modality to be within the 5th and 95th percentile, and varying intensity ranges between image modalities were accounted for by standardizing the pixel intensities modality-wise to have zero mean and unit variance. Lastly, the resulting volumes were concatenated to form an 8x96x96x96 tensor, which was used as input to the proposed 3D-ResNet models.

2.3.2. Data Split and Sampling Approach

After preprocessing, the set of multi-modal MR images was randomly partitioned into a train, validation, and test set using a stratified 80:10:10 split to account for the equal class distribution of HCC across data splits. Further, as the number of positive cases only comprised roughly 10.5% of the total number of observations, weighted random sampling was applied to the training set of multi-modal MRI sequences. The technique balances the dataset's unequal class distribution by assigning weights to individual observations, which are then drawn with replacement according to a probability proportional to their assigned class weights (Haixiang et al., 2017). In doing so, the training process ensures that the model is exposed to an equal spread of classes, thereby enhancing its performance in the presence of class imbalance (Bria et al., 2020). However, to warrant optimal generalization to clinical practice conditions, weighted random sampling was only applied to the training set of multi-modal MRI scans. During evaluation and testing, observations from the validation and test sets were thus

drawn without replacement, retaining the dataset's original class distribution of HCC, and hence closely mimicking real-world practice scenarios.

2.3.3. Data Augmentation

To further alleviate the class imbalance issue, data augmentation was applied to the training set of images. The technique synthetically increases the variation in the training data by applying random yet realistic transformations to the data and is a widely employed regularization technique in computer vision, typically yielding improved out-of-sample performance of the algorithm (Mikołajczyk and Grochowski, 2018). For this study, three spatial and two intensity-based image augmentations techniques as well as all of their joint combinations were employed.

With respect to the spatial augmentations, the multi-modal MR images were randomly rotated between -22.5 and 22.5 degrees to account for positional variability in the MRI scan. Furthermore, the scans were randomly flipped along the horizontal axis to represent varying imaging orientations, and random zooms between a factor of 1.1 and 1.3 of the original image were added.

For the intensity-based augmentations, Gibbs artifacts commonly occurring in MR images were simulated by adding random Gibbs noise to the raw images, and the image contrast was randomly adjusted to account for the heterogeneity in imaging modalities. Spatial augmentation techniques were implemented given a marginal probability of $p = 0.2$, while intensity-based augmentation techniques were employed using a marginal probability of $p = 0.1$.

2.3.4. Model Training

Model training was performed distributively using four Nvidia GeForce GTX 1080 Ti graphical processing unit cards (GPU) and pytorch's distributed data parallel framework (Li et al., 2020b). Across all four model architectures, fine-tuning of the pretrained models took approximately 14 minutes per epoch (13 to 28 epochs in total), while training of the ensemble took approximately 15 minutes per epoch (40 epochs in total).

The limited amount of accessible video random access memory (VRAM) per GPU was combated, allocating the maximum possible batch size to each GPU during training (see Table A.3 in Appendix A). Additionally, the still limited total batch size was further enhanced by accumulating the networks' gradients over several batches and only stepping the optimizer after a certain number of batches have been performed. The technique, commonly referred to as gradient accumulation, simulates bigger batch sizes than the processing hardware is able to fit into memory and typically yields more stable model convergence (Hermans et al., 2017). Accordingly, in this study, the number of parameter updates per training epoch were restrained to,

$$\text{Updates} = \left\lfloor \frac{\text{Steps}}{\text{Accumulation Steps}} \right\rfloor + 1, \tag{1}$$

where the numerator denotes the number of steps per training epoch and the denominator represents the number of accumulation steps.

Backward optimization was performed using the Adam optimization algorithm. Furthermore, the models' loss was computed, calculating the cross-entropy loss L_{CE} on the models' Softmax probabilities p_i with respect to the ground truth class t_i,

$$L_{CE} = -\sum_{i=1}^{2} t_i log(p_i), \tag{2}$$

and averaging the sum of losses per network parameter update over the network's batch size and number of accumulation steps.

Following each training epoch, model performance was monitored on the validation set of images and the training process halted if, for 10 consecutive epochs, no improvement on the model's validation loss was registered. The minimum number of training epochs was set to 10 and, after training, the model configuration with the lowest loss on the validation set was selected. Note here that since model training and validation were performed distributively, losses were computed per processing unit (i.e., four losses in total). For the study's assessment of the model's validation loss, however, only the loss on the processes' main GPU was administered. As such, further improvements of the models' validation loss registered on the remaining processing units may not be ruled out.

Regarding the models' configurations, optimal hyperparameter settings for the four individual as well as for the ensemble of 3D-ResNet models were evaluated on the study's validation set of abdominal MRI scans. For the four pretrained 3D-CNN architectures, the models' learning rates η, number of accumulations steps, and L2-regularization parameter values γ were varied. Specifically, this study assessed three different learning rates η between $1e-2$ and $1e-4$, as well as two different L2-regularization parameters γ values ranging between $1e-4$ and $1e-5$, and three different accumulation step settings, yielding a total of 18 model configurations per CNN architecture. For the ensemble model, hyperparameter tuning was confined to three different learning rates η (i.e., $1e-2$ to $1e-4$) and three different accumulation step settings. All models and their respective optimal configurations are summarized in Table A.3 in Appendix A.

2.3.5. Evaluation

After training, the models' performance was reported on the test set of abdominal MRI scans and their diagnostic power of HCC evaluated given the following set of standard evaluation metrics: accuracy, precision, recall, F1-score, area under the receiver operating curve ($AUROC$), and mean average precision (mAP). Accuracy, precision, recall, and F1-score were calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{N}, \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

Table 1: General Patient Characteristics in Training Set

	Mean	Abs. Frequency	Stand. Deviation	Rel. Frequency
Demographics				
Female	-	85	-	36.5%
Age (in Years)	66.6	-	11.6	-
Etiology				
Nonalcoholic Steatohepatitis	-	89	-	38.6%
Hepatitis B	-	4	-	1.7%
Hepatitis C	-	9	-	3.9%
Alcoholic Liver Disease	-	70	-	30.0%
Haemochromatosis	-	4	-	1.7%
Primary biliary cholangitis	-	5	-	2.1%
Primary sclerosing cholangitis	-	1	-	0.1%
Cryptogenic Cirrhosis	-	5	-	2.1%
Autoimmune hepatitis	-	11	-	4.7%

Note: Continuous variables are presented as means and standard deviations; dichotomous variables are presented using absolute and relative frequencies. Etiologies are also not mutually exclusive. Instead, patients may comprise varying etiologies.

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{6}$$

where TP denotes the model's correct identification of HCC, TN the model's correct prediction of common liver cirrhosis, FP the model's false assessment of cirrhotic observations as cancerous, and FN the model's false prediction of tumorous observations as solely cirrhotic.

2.3.6. Visualization

In addition to the models' quantitative assessment, further intuition into the models' decision process was gained by generating occlusion sensitivity maps of example images for all four 3D-ResNet models. The technique is commonly employed to enhance model interpretability (Jin et al., 2023), and returns a heatmap, visualizing the importance the algorithm assigns to each individual voxel for its predictions, and thus provides an additional diagnostic tool to medical practitioners.

3. Results

3.1. General Patient Characteristics

This study developed and evaluated four individual 3D-CNN architectures as well as their ensemble on a novel cohort of liver cirrhosis patients. From the cohort, 759 observations from 243 patients complied with the set selection criteria. The training, validation, and test set of abdominal MR images comprised 607, 76, and 76 total observations, with 64, 8, and 8 positive cases, respectively.

Table 2: Performance Metrics on Test Set

Models	Performance Metrics					
	Accuracy	Precision	Recall	F1-Score	AUROC	mAP
3D-ResNet10	0.855	0.000	0.000	0.000	0.588	0.130
3D-ResNet18	0.724	0.000	0.000	0.000	0.529	0.127
3D-ResNet34	0.789	0.000	0.000	0.000	0.450	0.112
3D-ResNet50	0.737	0.125	0.250	0.167	0.509	0.137
Ensemble	0.711	0.150	0.375	0.214	0.574	0.143

Note: AUROC = Area under the receiver operating characteristic curve; mAP = Mean average precision.

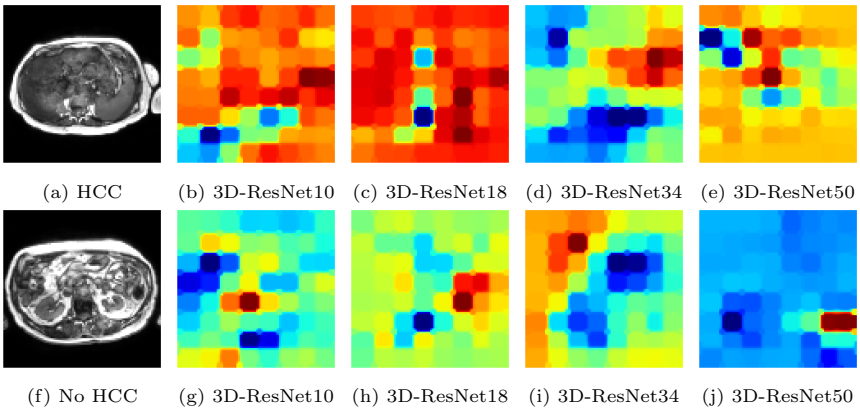
In correspondence with global incidence rates of liver cirrhosis (McGlynn et al., 2021; Yeh and Chen, 2010), the training set of images predominantly comprised male patients (63.5%) at an advanced age (mean age 66.6 years). Meanwhile, nonalcoholic steatohepatitis (NASH) (38.6%) and alcohol-associated liver disease (30.0%) presented the most common causes of liver cirrhosis in the study's training set. Further general patient characteristics for the training set of liver cirrhosis patients are presented in Table 1. Likewise, Table B.4 and Table B.5 in Appendix B present the corresponding general patient characteristics for the validation and test set.

3.2. Individual Models

All four finetuned 3D-CNN architectures as well as the ensemble of 3D-CNN models were evaluated on the study's test set of abdominal MRI scans. Their corresponding evaluation metrics are presented in Table 2.

Overall, the results show that all four individual 3D-ResNet models display limited performance in distinguishing between HCC and non-HCC cases. In fact, despite predicting the correct class for a large proportion of test images, their predictions remain inferior to the naive prediction of the sample's majority class (i.e., Naive Accuracy = 0.895), and three of the four models (i.e., 3D-ResNet10, 3D-ResNet18, and 3D-ResNet34) fail to correctly identify a single true positive observation altogether. Moreover, their low $AUROC$ scores, visually depicted in Figure 3a, indicate that across modeling architectures, the prediction of tumor prevalence remains marginally better than chance, with the 3D-ResNet34 model even displaying a lower than chance performance.

Still, among the varying model architectures, the least complex 3D-ResNet10 model achieved the highest overall accuracy (i.e., Accuracy = 0.855). Compared to the other more complex model architectures, the model thus remains most restrictive in its assessment of HCC, predicting tumor prevalence in just 3 out 76 total observations. Contrarily, 3D-ResNet18, 3D-ResNet34, and 3D-ResNet50 display greater lenience in their predictions. Similar to the 3D-ResNet10 model, however, 3D-ResNet18 and 3D-ResNet34 still fail to correctly identify a single HCC patient, and, more importantly, the lower overall $AUROC$ score suggests



(a) HCC (b) 3D-ResNet10 (c) 3D-ResNet18 (d) 3D-ResNet34 (e) 3D-ResNet50

(f) No HCC (g) 3D-ResNet10 (h) 3D-ResNet18 (i) 3D-ResNet34 (j) 3D-ResNet50

Figure 2: Occlusion Sensitivity Maps across Model Architectures

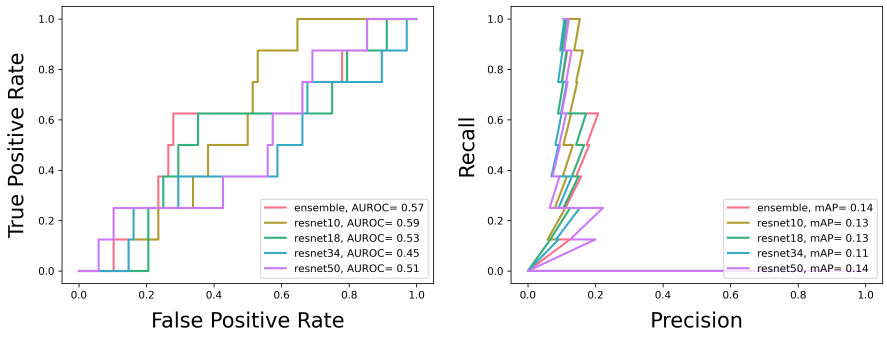
Note: Voxels on which the models assign the greatest relevance for prediction are represented in red. Conversely, less relevant regions are represented in blue.

that 3D-ResNet50's improvement in disease detection is likely at the expense of wrongly predicting a disproportionate number of positive cases (i.e., $AUROC = 0.509$ of 3D-ResNet50 compared to $AUROC = 0.588$ of 3D-ResNet10).

Figure 2 reaffirms the models' limited diagnostic value of HCC visually. The figure presents two example images of a patient with (see Figure 2a) and without diagnosed HCC (see Figure 2f) along with the models' respective occlusion sensitivity mapping for the corresponding images, thereby highlighting those voxel values that the respective model architecture deems most relevant to its prediction. The figure reveals that instead of focusing its attention on the subject's liver, depicted on the left side of the original image, all four model architectures appear to randomly assign relevance to individual voxel values, displaying limited capability to infer the most relevant parts of the image, while also showing little overlap within and between model architectures. Notably, the differences are especially severe among the more complex model architectures (i.e., 3D-ResNet34 and 3D-ResNet50), whereas the differences between 3D-ResNet10 and 3D-ResNet18 are less profound. As depicted in Figure C.4 in Appendix C, one possible explanation for this is that during training, model convergence of the more complex 3D-CNN architectures proved less stable. The increase in network parameters may have thus severely constrained its fit to the study's limited number of training images.

3.3. Ensemble

Despite the individual models' poor distinction between classes, an ensemble of the four 3D-ResNet models was fit to the training set of abdominal MRI scans by feeding the predictions from the four individual CNN architectures as inputs to a newly combined linear classifier. While the ensemble generally displays greater discriminatory power between HCC and common liver cirrhosis,



(a) Receiver Operating Characteristic Curves

(b) Average Precision Curves

Figure 3: Receiver Operating Characteristic and Average Precision Curves across Model Architectures

Note: AUROC = Area under the receiver operating characteristic curve; mAP = Mean average precision.

as evidenced by its comparative improvement in precision, recall, and F1-score, the model still fails to accurately infer cancer prevalence. However, unlike the aforementioned 3D-ResNet50 model, the ensemble's improved $AUROC$ score suggests that the ensemble yields a more cautious balance between positive and negative predictions, thus improving its overall predictive performance.

4. Discussion

In this study, a novel computer-aided diagnosis system of hepatocellular carcinoma in liver cirrhosis patients was proposed. In particular, four pretrained 3-dimensional convolutional neural networks of varying complexity were adopted and finetuned on a series of abdominal magnetic resonance images that were collected from a cohort of liver cirrhosis patients who underwent MRI screening for HCC surveillance. Following their individual training, all four finetuned 3D-CNN architectures were accumulated in an ensemble, and the ensemble additionally retrained, combining the individual models' predictions in a new linear classifier.

Overall, the results show that across modeling architectures generalization of the deduced feature representations proved inadequate. In fact, on the test set of abdominal MR images, all models displayed a diagnostic performance inferior to the naive prediction of the study's majority class, and three out of the four modeling architectures failed to even identify a single true positive observation in the study's test set. In accordance, the models' respective occlusion sensitivity maps presented in Figure 2 further demonstrated the models' limited capability to infer the most relevant parts of the image. That is, instead of focusing their attention on the subject's liver, the heatmaps revealed that all four models failed to display a consistent pattern in their assignment of relevance to

individual voxels while at the same time displaying limited overlap between architectures. The individual models' reasonable predictive performance achieved during training (see Figure C.4 in Appendix C) may thus, in large part, be attributed to the individual models' capability to remember the unique characteristics of the few positive cases provided in the study's training set, and may have been further exacerbated by the study's use of a weighted random sampling approach. However, initial drafts of this study also experimented with cost-sensitive and hard example mining approaches to counteract the class imbalance issue, yielding no improvement over the weighted sampling approach but instead resulting in the model's consistent prediction of the study's majority class. Generally, all four model architectures thus proved insufficient in their representation of the tumor's heterogeneous phenotype, although it shall be noted that the aggregation of their predictions in an ensemble yielded minor but slight improvements in disease detection.

While the overall viability of the study's proposed computer-assisted HCC diagnosis system thus appears limited, it is noteworthy that the study's findings are in stark contrast to the results attained in our preliminary investigation, which was conducted prior to this study. In the preliminary study, the effectiveness of the study's 3-dimensional modeling approach was assessed by adopting the most complex model architecture (i.e., 3D-ResNet50), equally initializing its parameters with pretrained weights from MedicalNet (Chen et al., 2019), and retraining it on two small-scale data sets of T2-weighted abdominal MRI scans comprising 20 HCC patients (Erickson et al., 2016) and 14 healthy individuals (Kavur et al., 2019), respectively. As depicted in Table D.6 in Appendix D, the results on the preliminary study's test set showed that the model accurately distinguished tumours from healthy observations, yielding an accuracy of 0.752 and a F1-score of 0.740. Hence, although the differences in data acquisition between data sets may have contributed to the preliminary study's promising findings, it nevertheless demonstrates the proposed methodology's general feasibility.

For the discouraging results obtained in this study, it, therefore, must be taken into account that the study's limitations in computing resources may have severely restrained its diagnostic value. Specifically, it shall be noted that due to the limitations in accessible video random access memory per graphical processing unit (i.e., 12GB VRAM per GPU), the original resolution for some of the imaging modalities used in this study had to be downsampled from approximately 125 millions to roughly 900.000 voxels, reducing the information captured in these images to less than 1% of the original image's information. Moreover, even when fitting the maximum possible image resolution (i.e., 176x176x176 voxels) and a minimum possible batch size of 1 to the processing units memory, images still entailed less than 5% of the original image's total information. Considering that the lesion typically measures only a few millimeters in size, it thus appears reasonable to assume that the CAD's limited diagnostic performance may be largely attributed to the study's constraints in computing resources.

As such, one potential avenue for future research could be to investigate the research question using greater computing resources with GPUs encompassing

more VRAM. However, in most clinical contexts, unlimited computing resources are seldom a realistic scenario, reducing the applicability of the proposed CAD to hardly more than a proof of concept. Therefore, another potential prospect for future research could be to explore HCC diagnosis by splitting the imaging volumes into 2-dimensional MRI slices. However, while the approach reduces the space complexity O from $O(n^3)$ to $O(n^2)$, it would also fail to account for the volumetric appearance of an MRI. More severely, it would require labor-intensive manual annotations of each single MRI slice, thus subverting the distinct benefits of this study's proposed 3D modeling approach.

Consequently, reevaluating the proposed diagnosis system using a more memory efficient data handling approach yields the most promising direction for future research. This is especially since the lesion only occurs in the subject's liver, making large parts of an MRI scan irrelevant to the detection of the disease. Future research should thus investigate the possibility to crop the original images such that they only represent the study's region of interest (i.e., the patient's liver), allowing for a more fine-grained resolution thereof. Accordingly, one possible direction would be to extend the proposed modeling approach to a two-stage segmentation and classification pipeline, where the detection algorithm first segments the patient's liver by drawing a cubic bounding box around it, and the classification network then classifies the images using the previously determined coordinates of the segmentation network's bounding box to crop the image on to the patient's liver. The technique has been employed in previous studies involving 3D medical images (Kern and Mastmeyer, 2021), and, as such, not only ensures the highest possible resolution per imaging volume but also accounts for the vast heterogeneity in liver sizes among cirrhosis patients.

Besides that, another promising direction would be to adopt a weakly supervised learning approach. The technique is especially advantageous since it does not require any additional manual annotations and instead uses image labels at a higher level of abstraction. For the detection of HCC, all images would thus be divided into smaller patches, training a network on the image patches and their corresponding image-level labels and subsequently aggregating its predictions across all patches to obtain a prediction for the entire image. In its simplicity, the technique should be easily applicable to the diagnosis of liver lesions but typically requires a large number of training images to be effective since its prediction may not be as accurate as approaches that use precise annotations. Therefore, future research should also consider combining weakly supervised learning with an additional data reduction technique, such as the object detection approach described earlier.

5. Conclusion

This study developed and evaluated an ensemble of four 3-dimensional convolutional neural networks, proposing a novel computer-aided diagnosis system of hepatocellular carcinoma in liver cirrhosis patients. While the results showed that the proposed approach yielded limited diagnostic capabilities, its

accurate assessment may have been severely constrained by the study's limitations in computing resources. Specifically, the loss of information induced by down-sampling the images may have substantially inhibited the lesion's accurate diagnosis. The study, therefore, emphasizes the need for future research to reevaluate the CAD's viability, discussing multiple potential avenues for further research in HCC diagnosis, and highlighting the need to carefully gauge the trade-off between computational efficiency and predictive accuracy.

Declaration of Competing Interests

Hereby, the authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements

The authors thank Sander Wansing for his invaluable contribution to the study.

Appendix A. Model Configurations

Table A.3: Optimal Model Configurations

Models	Hyperparameter			
	Learning Rate η	Batch Size	Accumulation Steps	Weight Decay γ
3D-ResNet10	$1e-3$	16	4	$1e-5$
3D-ResNet18	$1e-2$	16	8	$1e-5$
3D-ResNet34	$1e-2$	16	8	$1e-5$
3D-ResNet50	$1e-4$	8	8	$1e-5$
Ensemble	$1e-2$	8	4	-

Note: Optimal hyperparameter settings were selected experimentally using three different learning rates η ($1e-2$ to $1e-4$), three different accumulation steps (2 to 2^3 for 3D-Resnet10 to 3D-ResNet34 and 2^2 to 2^4 for 3D-ResNet50 and the ensemble), and two different weight decay values γ ($1e-4$ to $1e-5$). For the ensemble, no L2-regularization was applied.

Appendix B. General Patient Characteristics

Table B.4: General Patient Characteristics in Validation Set

	Mean	Abs. Frequency	Stand. Deviation	Rel. Frequency
Demographics				
Female	-	7	-	29.2%
Age (in Years)	65.3	-	10.6	-
Etiology				
Nonalcoholic Steatohepatitis	-	8	-	33.3%
Hepatitis B	-	0	-	0.0%
Hepatitis C	-	0	-	0.0%
Alcoholic Liver Disease	-	9	-	37.5%
Haemochromatosis	-	0	-	0.0%
Primary biliary cholangitis	-	0	-	0.0%
Primary sclerosing cholangitis	-	0	-	0.0%
Cryptogenic Cirrhosis	-	2	-	0.1%
Autoimmune hepatitis	-	1	-	0.0%

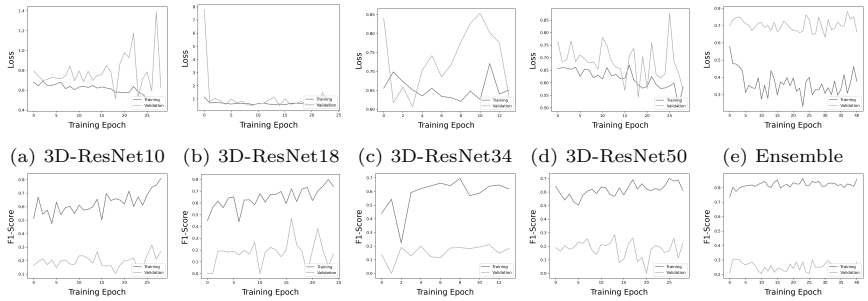
Note: Continuous variables are presented as means and standard deviations; dichotomous variables are presented using absolute and relative frequencies. Etiologies are also not mutually exclusive. Instead, patients may comprise varying etiologies.

Table B.5: General Patient Characteristics in Test Set

	Mean	Abs. Frequency	Stand. Deviation	Rel. Frequency
Demographics				
Female	-	10	-	40.0%
Age (in Years)	68.9	-	12.8	-
Etiology				
Nonalcoholic Steatohepatitis	-	11	-	44.0%
Hepatitis B	-	0	-	0.0%
Hepatitis C	-	0	-	0.0%
Alcoholic Liver Disease	-	9	-	36.0%
Haemochromatosis	-	0	-	0.0%
Primary biliary cholangitis	-	0	-	0.0%
Primary sclerosing cholangitis	-	1	-	0.0%
Cryptogenic Cirrhosis	-	0	-	0.0%
Autoimmune hepatitis	-	1	-	0.0%

Note: Continuous variables are presented as means and standard deviations; dichotomous variables are presented using absolute and relative frequencies. Etiologies are also not mutually exclusive. Instead, patients may comprise varying etiologies.

Appendix C. Learning Curves



(a) 3D-ResNet10 (b) 3D-ResNet18 (c) 3D-ResNet34 (d) 3D-ResNet50 (e) Ensemble

(f) 3D-ResNet10 (g) 3D-ResNet18 (h) 3D-ResNet34 (i) 3D-ResNet50 (j) Ensemble

Figure C.4: Learning Curves Across Model Architectures.

Note: Figures C.4a through C.4e show the models' respective loss plotted against the number of training epochs. Contrarily, Figures C.4f through C.4j show the models' respective F1-score plotted against the number of training epochs.

Appendix D. Preliminary Study Results

Table D.6: Performance Metrics on Test Set in Preliminary Study

Model	Performance Metrics				
	Accuracy	Precision	Recall	F1-Score	AUROC
3D-ResNet50	0.752	0.746	0.756	0.740	0.800

Note: AUROC = Area under the receiver operating characteristic curve; Results in the preliminary study are based on data derived from Erickson et al. (2016) and Kavur et al. (2019). The model was trained using stochastic gradient descent optimization with batches of size 8, a learning rate η of $1e-3$, and a momentum of $9e-1$.

References

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 68, 394–424. URL: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21492, doi:10.3322/caac.21492. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21492.

Brehar, R., Mitrea, D.A., Vancea, F., Marita, T., Nedevschi, S., Lupsor-Platon, M., Rotaru, M., Badea, R.I., 2020. Comparison of Deep-Learning and Conventional Machine-Learning Methods for the Automatic Recognition of the Hepatocellular Carcinoma Areas from Ultrasound Images. Sensors 20, 3085. URL: https://www.mdpi.com/1424-8220/20/11/3085, doi:10.3390/s20113085. number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

Bria, A., Marrocco, C., Tortorella, F., 2020. Addressing class imbalance in deep learning for small lesion detection on medical images. Computers in Biology and Medicine 120, 103735. URL: https://www.sciencedirect.com/science/article/pii/S0010482520301177, doi:10.1016/j.compbiomed.2020.103735.

Chen, S., Ma, K., Zheng, Y., 2019. Med3D: Transfer Learning for 3D Medical Image Analysis. URL: http://arxiv.org/abs/1904.00625, doi:10.48550/arXiv.1904.00625. arXiv:1904.00625 [cs].

Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y., 2022. Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis 79, 102444. URL: https://www.sciencedirect.com/science/article/pii/S1361841522000913, doi:10.1016/j.media.2022.102444.

Chen, Z., Xie, H., Hu, M., Huang, T., Hu, Y., Sang, N., Zhao, Y., 2020. Recent progress in treatment of hepatocellular carcinoma. American Journal of Cancer Research 10, 2993–3036. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7539784/.

Erickson, B.J., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., Rieger-Christ, K., Lemmerman, J., 2016. The Cancer Genome Atlas Liver Hepatocellular Carcinoma Collection (TCGA-LIHC). doi:10.7937/K9/TCIA.2016.IMMQW8UQ.

Estes, C., Razavi, H., Loomba, R., Younossi, Z., Sanyal, A.J., 2018. Modeling the Epidemic of Nonalcoholic Fatty Liver Disease Demonstrates an Exponential Increase in Burden of Disease. Hepatology 67, 123–133. doi:10.1002/hep.29466.

European Association For The Study Of The Liver, 2018. EASL Clinical Practice Guidelines: Management of Hepatocellular Carcinoma. Journal of Hepatology 69, 182–236. doi:10.1016/j.jhep.2018.03.019.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73, 220–239. URL: https://www.sciencedirect.com/science/article/pii/S0957417416307175, doi:10.1016/j.eswa.2016.12.035.

Hamm, C.A., Wang, C.J., Savic, L.J., Ferrante, M., Schobert, I., Schlachter, T., Lin, M., Duncan, J.S., Weinreb, J.C., Chapiro, J., Letzen, B., 2019. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. European Radiology 29, 3338–3347. URL: https://doi.org/10.1007/s00330-019-06205-9, doi:10.1007/s00330-019-06205-9.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. doi:10.48550/arXiv.1512.03385. issue: arXiv:1512.03385 arXiv: 1512.03385.

Heimbach, J.K., Kulik, L.M., Finn, R.S., Sirlin, C.B., Abecassis, M.M., Roberts, L.R., Zhu, A.X., Murad, M.H., Marrero, J.A., 2018. AASLD guidelines for the treatment of hepatocellular carcinoma. Hepatology 67, 358–380. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.29086, doi:10.1002/hep.29086. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hep.29086.

Hermans, J.R., Spanakis, G., Möckel, R., 2017. Accumulated Gradient Normalization, in: Proceedings of the Ninth Asian Conference on Machine Learning, PMLR. pp. 439–454. URL: https://proceedings.mlr.press/v77/hermans17a.html. iSSN: 2640-3498.

Jin, W., Li, X., Fatehi, M., Hamarneh, G., 2023. Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis 84, 102684. URL: https://www.sciencedirect.com/science/article/pii/S1361841522003127, doi:10.1016/j.media.2022.102684.

Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S., 2019. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. doi:10.5281/zenodo.3431873.

Kern, D., Mastmeyer, A., 2021. 3D Bounding Box Detection in Volumetric Medical Image Data: A Systematic Literature Review, in: 2021 IEEE 8th International Conference on Industrial Engineering and Applications (ICIEA), pp. 509–516. doi:10.1109/ICIEA52957.2021.9436733.

Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T., 2022. Transfer Learning for Medical Image Classification: A Literature Review. BMC Medical Imaging 22, 69. doi:10.1186/s12880-022-00793-7.

Li, J., Wu, Y., Shen, N., Zhang, J., Chen, E., Sun, J., Deng, Z., Zhang, Y., 2020a. A fully automatic computer-aided diagnosis system for hepatocellular carcinoma using convolutional neural networks. Biocybernetics and Biomedical Engineering 40, 238–248. URL: https://www.sciencedirect.com/science/article/pii/S0208521619300658, doi:10.1016/j.bbe.2019.05.008.

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., Chintala, S., 2020b. PyTorch Distributed: Experiences on Accelerating Data Parallel Training. URL: http://arxiv.org/abs/2006.15704. arXiv:2006.15704 [cs].

McGlynn, K.A., Petrick, J.L., El-Serag, H.B., 2021. Epidemiology of Hepatocellular Carcinoma. Hepatology (Baltimore, Md.) 73, 4–13. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577946/, doi:10.1002/hep.31288.

Mikołajczyk, A., Grochowski, M., 2018. Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop (IIPhDW), pp. 117–122. doi:10.1109/IIPHDW.2018.8388338.

Mitchell, D.G., Bruix, J., Sherman, M., Sirlin, C.B., 2015. LI-RADS (Liver Imaging Reporting and Data System): Summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology 61, 1056–1065. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.27304, doi:10.1002/hep.27304. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hep.27304.

Oestmann, P.M., Wang, C.J., Savic, L.J., Hamm, C.A., Stark, S., Schobert, I., Gebauer, B., Schlachter, T., Lin, M., Weinreb, J.C., Batra, R., Mulligan, D., Zhang, X., Duncan, J.S., Chapiro, J., 2021. Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver. European Radiology 31, 4981–4990. URL: https://doi.org/10.1007/s00330-020-07559-1, doi:10.1007/s00330-020-07559-1.

Ponnoprat, D., Inkeaw, P., Chaijaruwanich, J., Traisathit, P., Sripan, P., Inmutto, N., Na Chiangmai, W., Pongnikorn, D., Chitapanarux, I., 2020. Classification of hepatocellular carcinoma and intrahepatic cholangiocarcinoma based on multi-phase CT scans. Medical & Biological Engineering & Computing 58, 2497–2515. URL: https://doi.org/10.1007/s11517-020-02229-2, doi:10.1007/s11517-020-02229-2.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding Transfer Learning for Medical Imaging. doi:10.48550/arXiv.1902.07208. issue: arXiv:1902.07208 arXiv: 1902.07208.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery 8, e1249. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249, doi:10.1002/widm.1249. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1249.

Schmauch, B., Herent, P., Jehanno, P., Dehaene, O., Saillard, C., Aubé, C., Luciani, A., Lassau, N., Jégou, S., 2019. Diagnosis of focal liver lesions from ultrasound using deep learning. Diagnostic and Interventional Imaging 100, 227–233. URL: https://www.sciencedirect.com/science/article/pii/S2211568419300592, doi:10.1016/j.diii.2019.02.009.

Shi, W., Kuang, S., Cao, S., Hu, B., Xie, S., Chen, S., Chen, Y., Gao, D., Chen, Y., Zhu, Y., Zhang, H., Liu, H., Ye, M., Sirlin, C.B., Wang, J., 2020. Deep learning assisted differentiation of hepatocellular carcinoma from focal liver lesions: choice of four-phase and three-phase CT imaging protocol. Abdominal Radiology 45, 2688–2697. URL: https://doi.org/10.1007/s00261-020-02485-8, doi:10.1007/s00261-020-02485-8.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging 35, 1299–1312. doi:10.1109/TMI.2016.2535302. conference Name: IEEE Transactions on Medical Imaging.

Todoroki, Y., Iwamoto, Y., Lin, L., Hu, H., Chen, Y.W., 2019. Automatic Detection of Focal Liver Lesions in Multi-phase CT Images Using A Multi-channel & Multi-scale CNN, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 872–875. doi:10.1109/EMBC.2019.8857292. iSSN: 1558-4615.

Wolf, E., Rich, N.E., Marrero, J.A., Parikh, N.D., Singal, A.G., 2021. Use of Hepatocellular Carcinoma Surveillance in Patients With Cirrhosis: A Systematic Review and Meta-Analysis. Hepatology 73, 713–725. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.31309, doi:10.1002/hep.31309. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hep.31309.

Wu, Y., White, G.M., Cornelius, T., Gowdar, I., Ansari, M.H., Supanich, M.P., Deng, J., 2020. Deep learning LI-RADS grading system based on contrast enhanced multiphase MRI for differentiation between LR-3 and LR-4/LR-5 liver tumors. Annals of Translational Medicine 8, 701. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327307/, doi:10.21037/atm.2019.12.151.

Yang, Q., Wei, J., Hao, X., Kong, D., Yu, X., Jiang, T., Xi, J., Cai, W., Luo, Y., Jing, X., Yang, Y., Cheng, Z., Wu, J., Zhang, H., Liao, J., Zhou, P., Song, Y., Zhang, Y., Han, Z., Cheng, W., Tang, L., Liu, F., Dou, J., Zheng, R., Yu, J., Tian, J., Liang, P., 2020. Improving B-mode ultrasound diagnostic performance for focal liver lesions using deep learning: A multicentre study. EBioMedicine 56, 102777. URL: https://www.sciencedirect.com/science/article/pii/S2352396420301523, doi:10.1016/j.ebiom.2020.102777.

Yasaka, K., Akai, H., Abe, O., Kiryu, S., 2018. Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study. Radiology 286, 887–896. URL: https://pubs.rsna.org/doi/full/10.1148/radiol.2017170706, doi:10.1148/radiol.2017170706. publisher: Radiological Society of North America.

Yeh, S.H., Chen, P.J., 2010. Gender disparity of hepatocellular carcinoma: the roles of sex hormones. Oncology 78 Suppl 1, 172–179. doi:10.1159/000315247.

Yu, H., Yang, L.T., Zhang, Q., Armstrong, D., Deen, M.J., 2021. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. Neurocomputing 444, 92–110. URL: https://www.sciencedirect.com/science/article/pii/S0925231221001314, doi:10.1016/j.neucom.2020.04.157.

Zhen, S.h., Cheng, M., Tao, Y.b., Wang, Y.f., Juengpanich, S., Jiang, Z.y., Jiang, Y.k., Yan, Y.y., Lu, W., Lue, J.m., Qian, J.h., Wu, Z.y., Sun, J.h., Lin, H., Cai, X.j., 2020. Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data. Frontiers in Oncology 10. URL: https://www.frontiersin.org/articles/10.3389/fonc.2020.00680.

Zhou, J., Wang, W., Lei, B., Ge, W., Huang, Y., Zhang, L., Yan, Y., Zhou, D., Ding, Y., Wu, J., Wang, W., 2021. Automatic Detection and Classification of Focal Liver Lesions Based on Deep Convolutional Neural Networks: A Preliminary Study. Frontiers in Oncology 10. URL: https://www.frontiersin.org/articles/10.3389/fonc.2020.581210.