Detecting hepatocellular carcinoma in non-alcoholic fatty liver disease patients using deep learning based radiomics

Research Report

Jakob Nolte

Supervisors: Stephanie van den Berg (Twente University) and
Maryam Amir Haeri (Twente University)

Utrecht University
Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
January 2023

1 Introduction

Hepatocellular carcinoma (HCC) is the most common type of liver cancer in adults and typically arises in the setting of chronic liver disease or cirrhosis [1]. Today, the disease represents one of the leading causes of cancer-related deaths worldwide [2]. However, due to the surging prevalence of non-alcoholic fatty liver disease in the general population, its burden is still expected to increase [3]. Besides that, HCC remains challenging to diagnose. This is particularly because, in its early stages, when curative treatments are still feasible clinical presentation of the disease is often non-specific and can include abdominal pain, weight loss, and jaundice [4, 5]. For at-risk populations, international screening guidelines thus recommend frequent ultrasonography screening [6].Yet, even with more refined imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI), hepatocellular carcinoma can be strenuous to differentiate from other liver lesions.

Fortunately, recent advancements in the medical imaging domain have led to the development of several computational approaches that aim to aid tumor diagnosis [7, 8]. These approaches, usually coined radiomics, mine statistical properties from the raw images and analyze these handcrafted features using varying machine learning algorithms [9, 10]. Although radiomic models have yielded promising results in liver oncology, one shortcoming of these studies is that results are susceptible to the image segmentation and processing steps involved [11, 12]. Further, some argue that relying solely on handcrafted features may still miss much of the abstract information captured in medical images [13]. They thus advocate for the use of deep learning (DL) based approaches to medical imaging [14, 15]. Unlike handcrafted radiomics, these approaches, often referred to as deep radiomics, learn feature representation from sample images using convolutional neural networks (CNN). In doing so, they have been shown to match [16] or even surpass the performance of experienced radiologists in HCC diagnosis [17], to accurately distinguish the disease from other liver lesions [18], and correctly classify HCC according to its severity [19]. However, no baseline for HCC development in non-alcoholic fatty liver disease patients has been established in the literature, and to date, no comprehensive comparison of handcrafted and deep radiomic approaches has been reported concerning HCC diagnosis. The latter is especially worrisome since, despite evidence from other fields of medical imaging, most studies discussing liver tumor diagnosis still employ handcrafted radiomic models rather than DL [20]. In its scope, this study therefore addresses these limitations in the literature, posing the following research questions: (1) Compared to handcrafted radiomics, does deep learning yield improved HCC diagnosis in non-alcoholic fatty liver disease patients? (2) Which DL features can be identified as risk factors of HCC development?

To investigate the research questions posed, this proof of concept study adopts a pretrained convolutional neural network, comparing the predictive performance of the proposed DL approach to a baseline radiomic model separately derived. In addition, the interpretation of the DL approach is augmented by modeling a set of explainable DL features. As such, the information captured

by the DL model is aided by context interpretable to the practicing radiologist. The code to reproduce all results presented in this study is publicly available.[1]

2 Materials and Methods

2.1 Data Description

Given that access to the initially intended data source was not provided at the point of writing the report, data is drawn from two publicly available data sources. That is, abdominal MR images of 20 healthy individuals are derived from Kavur et al. [21], while magnetic resonance scans of 20 patients diagnosed with hepatocellular carcinoma are drawn from the cancer imaging archive [22]. Both data sources implemented varying imaging modalities to obtain the MR scans. Thus, to enhance comparability across data sources, only T2-weighted images are selected for this study. Images of the healthy individuals were acquired using a 1.5T Philips MRI with varying (5.5 to 9 mm; average 7.84 mm) inter-slice distance (ISD). No information on image acquisition was provided for the data on HCC patients.

2.2 Network Architecture

Concerning the DL model, this study employs transfer learning, adopting a pretrained network architecture. The technique builds upon knowledge that has been accumulated during training on a related problem and thus effectively minimizes the amount of data required to train a deep neural network. As a result, transfer learning has become the de-facto method for DL based applications in medical image analysis, yielding good performance despite the small sample sizes typically encountered in clinical studies [23].

For this study, a 3-dimensional extension of the original ResNet50 architecture [24] is adopted and the network's parameters are initialized using pretrained weights from the MedicalNet project [25].[2] The network is part of the series of residual neural network (ResNet) architectures and one of the most widely employed models for image classification in the medical domain [26]. It consists of four subsequent bottleneck blocks of three convolutional layers and one fully connected output layer. Since the network was originally developed on 2-dimensional images, this study employs a 3D extension of the model, accounting for the volumetric nature of MRI images.

In addition, this study proposes a slightly adjusted network architecture compared to ResNet50's original layout. That is, it introduces a fully connected intermediate layer which is added between the network's output layer and its last bottleneck block. In contrast to the original ResNet50 architecture, where all 2048 activations retained after global average pooling are given as input to

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

[2] MedicalNet comprises a series of pretrained model weights trained on 23 different medical imaging datasets.

predict image labels, the high-dimensional feature space (2048 features) is thus first mapped to a lower dimension (128 features) and then given as input to predict tumor prevalence. The step is introduced to aid the extraction of the explainable DL features and is further discussed in section 2.3.2.

2.3 Experimental Setup

2.3.1 Image Preprocessing

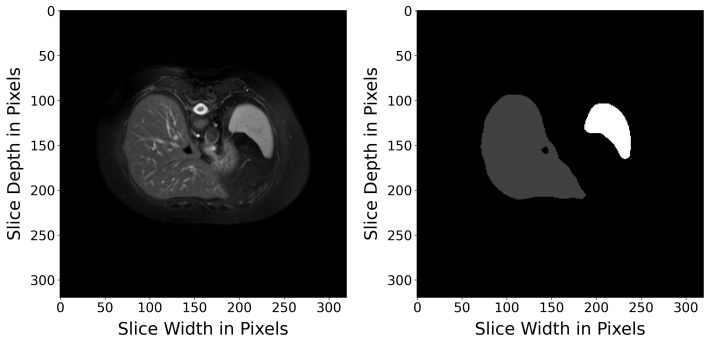
Before model training, minor preprocessing steps are performed on the raw images. That is, all MRI scans are converted from DICOM to nifti file format, the images' dimensions are scaled to 128x128x64 pixels, and the images' greyscale as well as their orientation are normalized. Note that since this study merely presents a proof of concept, no data augmentation is applied to enhance the number of training samples.

2.3.2 Feature Engineering

Handcrafted Features. To compare the performance of the DL approach with a baseline radiomic model, handcrafted features are computed from the raw MRI images. The step requires that, prior to their computation, the study's region of interest, i.e., the liver, is delineated. Therefore, this study employs image segmentation using a pretrained nnU-Net [27] liver segmentation model.[3] An exemplary MRI slice along with its corresponding segmentation mask is presented in figure 1, illustrating the segmentation performed.

Following image segmentation, the raw images and their corresponding segmentation masks are given as inputs to derive 14 shape, 18 first-order, and 40 second- and higher-order radiomic features from the study's region of interest using the *pyradiomics* library [29]. Among these features, shape features capture the shape of the traced region of interest as well as its geometric properties such as volume, maximum diameter, maximum surface, tumor compactness, and sphericity. First-order features, by contrast, are histogram-based properties reporting the mean, median, maximum, and minimum values of the voxel intensities on the image, as well as their skewness, kurtosis, uniformity, and randomness (entropy). Lastly, second- and higher-order features describe the spatial inter-relationships between neighboring voxels, providing a measure of tissue heterogeneity [30, 12].

[3] nnU-Net is a self-configuring, automatic image segmentation algorithm closely based on the original uNet architecture [28].



(a) Raw Image (b) Segmentation Mask

Figure 1: Exemplary MRI Slice with corresponding Segmentation Mask

Note: In the segmentation mask, liver, spleen, and background are denoted in gray, white, and black, respectively.

Explainable DL Features. To augment the interpretation of the DL approach, this study derives a number of DL features and relates them to the handcrafted features previously discussed. In doing so, this study is subject to the assumption that the feature map of the network's penultimate layer can be considered representative of the latent characteristics captured in medical images. Accordingly, the model's output layer is removed from the network's architecture and the feature map of the model's penultimate layer, i.e., the newly introduced intermediate layer, is extracted, yielding a total of 128 features.

Compared to the original ResNet50 architecture, extraction of the intermediate layer's feature map entails the advantage that the newly introduced intermediate layer already comprises a preliminary feature reduction step. In other words, compared to the extraction of the network's original penultimate feature map, which consists of 2048 possibly highly redundant features, the intermediate layer allows for the extraction of a feature map with lower complexity and thus potentially higher inter-class variance.

Finally, correlations of all possible pairs of DL and handcrafted features are computed to distinguish DL features with explanatory value from DL features that cannot be interpreted in the context of handcrafted radiomics. As such, DL features with an absolute correlation of $|r| \geq 0.5$ with at least one of the handcrafted features are considered explainable, whereas DL features below the given threshold are considered non-informative and are thus removed from the set of explainable DL features.

2.3.3 Feature Selection

Following their derivation, the high dimensional feature space of the handcrafted and the explainable DL features is further reduced using least absolute shrinkage and selection operator (LASSO) regularization.

In addition to LASSO regularization, constant (i.e., $\sigma = 0$) as well as quasi-constant (i.e., $\sigma \leq 0.001$) features are removed from both sets of features, and all features are additionally z-score standardized. Note further that for six of the patients diagnosed with HCC, handcrafted radiomic features could not be computed. The observations are thus excluded from the set of handcrafted features.

2.3.4 Model Training

For model training, the data is randomly divided into a training and a validation set using an 80:20 split. The DL model is trained over a minimum of 50 epochs using batches of size 8, a learning rate of $\eta = 0.001$, and a momentum of $\gamma = 0.9$. The model's loss is computed calculating the binary cross-entropy and early stopping is employed based on the validation loss.[4] After training, the model configuration with the lowest loss on the validation set is selected.

To avoid overfitting on the small number of training samples, only a set of model parameters are retrained. That is, the network's output layer, its newly introduced intermediate layer and the network's last bottleneck block before the newly introduced layer are finetuned. All other model parameters are not updated. Note that since this report merely presents a proof of concept, no model hyperparameter optimization is employed. All analyses are implemented using the *pytorch* based *MONAI* framework.

To obtain performance metrics of the baseline radiomic model and the explainable DL model, LASSO regularized logistic regression models are fit to the training set of handcrafted and explainable DL features, respectively. The LASSO regularization parameter λ is optimized using three-fold cross-validation to ensure optimal classification performance on the validation set. Accordingly, seven evenly spaced parameter values on a logarithmic scale between 10^{-3} and 10^{3} are given as input to the regularization parameter λ and the hyperparameter value with the highest average classification accuracy across the three folds is selected. Note that in *scikit-learn's* implementation of LASSO regularization, the complexity parameter $C = \frac{1}{\lambda}$ is tuned instead of λ. Consequently, the values of λ are converted and C is optimized. Hyperparameter optimization is implemented using *scikit-learn*.

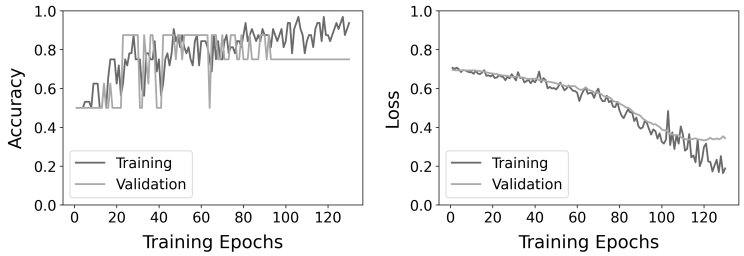
2.3.5 Evaluation Approach

The predictive performance of the respective classification models is reported on the validation set and evaluated given the following set of standard evaluation metrics: accuracy, precision, recall, and F1-score. To capture the uncertainty induced by the datasets small number of observations, all performance metrics are presented along with their bootstrapped 95% confidence intervals, which are calculated using 1000 bootstrap samples of the validation set.

[4]Model training is aborted if for 10 consecutive epochs no improvement on the model's validation loss is registered.

3 Results

Figure 2 presents the model's accuracy and learning curves plotted over the number of training epochs. As depicted in figure 2b, the model continuously minimizes the loss on the validation set until plateauing after approximately 120 training epochs. Interestingly, however, the model's accuracy on the validation set slightly decreases after the model starts overfitting on the training data. While it achieves higher accuracy estimates during previous epochs, the model's best configuration on the validation set thus yields a classification accuracy of 75.2%.



(a) Accuracy over Training Epochs (b) Loss over Training Epochs

Figure 2: Accuracy and Loss on Training and Validation Set over Training Epochs

Concerning its comparison to the baseline radiomic model, table 1 presents the DL model's performance metrics along with its corresponding bootstrapped 95% confidence intervals. As aforementioned, the DL model achieves a prediction accuracy of 75.2% on the validation set. Besides that, it displays a precision of 75.6% and a considerably higher recall than the baseline radiomic model (75.6% recall compared to 50.9% recall). Accordingly, the model is considerably more sensitive to hepatocellular carcinoma detection, which is especially important in a medical context where the misclassification of true positives is considered more severe than wrongly predicting false negative cases.

As for the baseline radiomic model, 75.2% of the validation set's labels are equally correctly predicted. The model precisely classifies tumor patients (precision of 99.2%) but consistently overestimates the number of negative cases (50.9% recall) in the validation set. While it thus effectively minimizes the number of false positive cases, it consistently overestimates the number of negative cases, yielding a lower F1-score than the DL model.

The model solely fit to the explainable DL features achieves a classification performance on par with the DL model. In fact, across all performance metrics, the model fit to the explainable DL features equals the predictive performance of the DL model, although its confidence bounds are principally wider than the DL model's confidence bounds.

Regarding the model's features, 104 DL features display an absolute correlation of $|r| \geq 0.5$ with at least one of the handcrafted radiomic features and three explainable DL features with non-zero coefficients are retained af-

Models	Performance Metrics			
	Accuracy	Precision	Recall	F1-Score
Baseline Radiomic Model	0.752 [0.537, 0.968]	0.992 [0.814, 1.000]	0.509 [0.146, 0.871]	0.654 [0.317, 0.992]
Deep Learning Model	0.752 [0.545, 0.959]	0.746 [0.445, 1.000]	0.756 [0.452, 1.000]	0.740 [0.495, 0.984]
Explainable Deep Learning Model	0.748 [0.525, 0.972]	0.748 [0.427, 1.000]	0.747 [0.422, 1.000]	0.734 [0.466, 1.000]

Note: Bootstrapped performance metrics are presented as single estimates. Their corresponding 95% confidence intervals are presented in square brackets.

Table 1: Performance Metrics of the Different Models on the Validation Set

ter applying LASSO regularization. Among these features, the results indicate that deviations in the image's gray level intensity distribution exhibit the greatest effect on HCC development. In fact, the results show that a one unit increase in var12 [5] ($\beta = -0.325, z = -0.382, p = 0.703$) and var80 ($\beta = -2.096, z = -2.180, p = 0.029$) is associated with a 27.7% and an 87.7% decrease in the odds of HCC development, respectively. The two features share many common radiomic features and are composed of 4 and 10 histogram-based first-order radiomic features, correspondingly. More importantly, their respective associations with the handcrafted radiomic features imply that more uniform gray level distributions with less extreme upper and lower bounds correspond with lower risks of tumor development, albeit var12 is not statistically significant given a significance level of $\alpha = 0.05$. Similarly, the third explainable DL feature, var32, underlines the importance of gray-level variability in the images. It displays an absolute correlation of $|r| \geq 0.5$ with the image's gray level zones homogeneity (i.e., a higher-order radiomic feature) and is equally found to be negatively associated with HCC development ($\beta = -1.081, z = -1.792, p = 0.073$).

4 Discussion

In this study, a deep learning approach to the classification of hepatocellular carcinoma was compared to a baseline radiomics approach. In addition, the interpretation of the DL approach was aided by modeling a set of explainable DL features and interpreting them in the context of handcrafted radiomics. The results show that, despite lower precision, the DL model is considerably more sensitive to lesion detection than the baseline model. Given that model sensitivity (i.e., recall) presents the arguably most important metric in tumor classification, it can thus be concluded that the DL model provides higher clinical utility than the baseline model. The results appear especially promising, since

[5] The numbers appearing as the postfix denote the features's index in the network's intermediate layer.

no hyperparameter optimization was performed for this study. In a future extension, optimizing the model's hyperparameters should thus yield even higher predictive performance of the DL model, further underlining the advantageous applicability of the approach.

As for the interpretation of the DL approach, performance metrics of the model solely fit to explainable DL features were on par with the performance metrics of the DL model. The results indicate that the loss of information induced by solely modeling the explainable features is negligible compared to the DL model. However, only one correlation threshold (i.e., $|r| \geq 0.5$) was assessed in this study, and no other measure of association was taken into consideration. Future extensions of this study should thus investigate varying correlation thresholds and potentially other measures of association to further support the results presented here.

Likewise, future extensions of this study should also explore other model architectures with corresponding weights pretrained on large scale medical imaging dataset. This is especially important since there is ample evidence that less complex model architectures than ResNet50 may yield better predictive performance on simple binary classification problems like the given [31]. Consequently, all model architectures pretrained on medical imaging tasks shall be assessed.

Lastly, future extensions of this study will also consider greater and more harmonized data sources. In fact, for the data sources used in this study, it must be considered that the varying imaging modalities used to obtain the images may largely affect the results presented in this proof of concept and thus greatly compromise its generalizability. In a future extension, images from patients diagnosed with and without hepatocellular carcinoma will thus be drawn from a single data source and the model's performance will ultimately be validated on an external validation set.

References

[1] J. D. Yang, P. Hainaut, G. J. Gores, A. Amadou, A. Plymoth, and L. R. Roberts, "A global view of hepatocellular carcinoma: Trends, risk, prevention and management," *Nature Reviews Gastroenterology & Hepatology*, vol. 16, pp. 589–604, Oct. 2019.

[2] H. B. El-Serag, "Epidemiology of Hepatocellular Carcinoma," in *The Liver*, ch. 59, pp. 758–772, John Wiley & Sons, Ltd, 2020.

[3] C. Estes, H. Razavi, R. Loomba, Z. Younossi, and A. J. Sanyal, "Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease," *Hepatology*, vol. 67, no. 1, pp. 123–133, 2018.

[4] J. Hartke, M. Johnson, and M. Ghabril, "The diagnosis and treatment of hepatocellular carcinoma," *Seminars in Diagnostic Pathology*, vol. 34, pp. 153–159, Mar. 2017.

[5] L. Kulik and H. B. El-Serag, "Epidemiology and Management of Hepatocellular Carcinoma," *Gastroenterology*, vol. 156, pp. 477–491.e1, Jan. 2019.

[6] European Association For The Study Of The Liver, "EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma," *Journal of Hepatology*, vol. 69, pp. 182–236, July 2018.

[7] 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, p. 91, Aug. 2020.

[8] 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, Apr. 2020.

[9] M. Avanzo, L. Wei, J. Stancanello, M. Vallières, A. Rao, O. Morin, S. A. Mattonen, and I. El Naqa, "Machine and deep learning methods for radiomics," *Medical Physics*, vol. 47, pp. e185–e202, June 2020.

[10] T. Wakabayashi, F. Ouhmich, C. Gonzalez-Cabrera, E. Felli, A. Saviano, V. Agnus, P. Savadjiev, T. F. Baumert, P. Pessaux, J. Marescaux, and B. Gallix, "Radiomics in hepatocellular carcinoma: A quantitative review," *Hepatology International*, vol. 13, pp. 546–559, Sept. 2019.

[11] J. M. Miranda Magalhaes Santos, B. Clemente Oliveira, J. d. A. B. Araujo-Filho, A. N. Assuncao-Jr, F. A. de M. Machado, C. Carlos Tavares Rocha, J. V. Horvat, M. R. Menezes, and N. Horvat, "State-of-the-art in radiomics of hepatocellular carcinoma: A review of basic principles, applications, and limitations," *Abdominal Radiology*, vol. 45, pp. 342–353, Feb. 2020.

[12] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El 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. Socarras 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, May 2020.

[13] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and Opportunities," *IEEE Signal Processing Magazine*, vol. 36, pp. 132–160, July 2019.

[14] Y. S. Sung, B. Park, H. J. Park, and S. S. Lee, "Radiomics and deep learning in liver diseases," *Journal of Gastroenterology and Hepatology*, vol. 36, no. 3, pp. 561–568, 2021.

[15] W. Rogers, S. Thulasi Seetha, T. A. G. Refaee, R. I. Y. Lieverse, R. W. Y. Granzier, A. Ibrahim, S. A. Keek, S. Sanduleanu, S. P. Primakov, M. P. L. Beuque, D. Marcus, A. M. A. van der Wiel, F. Zerka, C. J. G. Oberije, J. E. van Timmeren, H. C. Woodruff, and P. Lambin, "Radiomics: From qualitative to quantitative imaging," *The British Journal of Radiology*, vol. 93, p. 20190948, Apr. 2020.

[16] S.-h. Zhen, M. Cheng, Y.-b. Tao, Y.-f. Wang, S. Juengpanich, Z.-y. Jiang, Y.-k. Jiang, Y.-y. Yan, W. Lu, J.-m. Lue, J.-h. Qian, Z.-y. Wu, J.-h. Sun, H. Lin, and X.-j. Cai, "Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data," *Frontiers in Oncology*, vol. 10, p. 680, May 2020.

[17] C. A. Hamm, C. J. Wang, L. J. Savic, M. Ferrante, I. Schobert, T. Schlachter, M. Lin, J. S. Duncan, J. C. Weinreb, J. Chapiro, and B. Letzen, "Deep learning for liver tumor diagnosis part I: Development of a convolutional neural network classifier for multi-phasic MRI," *European Radiology*, vol. 29, pp. 3338–3347, July 2019.

[18] K. Yasaka, H. Akai, O. Abe, and S. Kiryu, "Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study," *Radiology*, vol. 286, pp. 887–896, Mar. 2018.

[19] Y. Wu, G. M. White, T. Cornelius, I. Gowdar, M. H. Ansari, M. P. Supanich, and J. Deng, "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*, vol. 8, p. 701, June 2020.

[20] A. A. Borhani, R. Catania, Y. S. Velichko, S. Hectors, B. Taouli, and S. Lewis, "Radiomics of hepatocellular carcinoma: Promising roles in patient selection, prediction, and assessment of treatment response," *Abdominal Radiology*, vol. 46, pp. 3674–3685, Aug. 2021.

[21] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data," Apr. 2019.

[22] B. J. Erickson, S. Kirk, Y. Lee, O. Bathe, M. Kearns, C. Gerdes, K. Rieger-Christ, and J. Lemmerman, "The Cancer Genome Atlas Liver Hepatocellular Carcinoma Collection (TCGA-LIHC)," 2016.

[23] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," Oct. 2019.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015.

[25] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer Learning for 3D Medical Image Analysis," July 2019.

[26] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using ImageNet," *Computers in Biology and Medicine*, vol. 128, p. 104115, Jan. 2021.

[27] 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, pp. 203–211, Feb. 2021.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), Lecture Notes in Computer Science, (Cham), pp. 234–241, Springer International Publishing, 2015.

[29] J. J. van 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, Oct. 2017.

[30] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: The facts and the challenges of image analysis," *European Radiology Experimental*, vol. 2, p. 36, Nov. 2018.

[31] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Medical Imaging*, vol. 22, p. 69, Apr. 2022.