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

Jakob Nolte

December 2022

1 Introduction

Hepatocellular carcinoma (HCC) is the most common type of liver cancer in adults. The disease typically arises in the setting of chronic liver disease or cirrhosis [1] and represents one of the leading causes of cancer-related deaths worldwide [2]. Despite that, the burden of the disease is still expected to increase due to the soaring prevalence of non-alcoholic fatty liver disease in the general population [3]. In addition, hepatocellular carcinoma remains difficult to diagnose. This is because, particularly 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]. International screening guidelines thus recommend frequent ultrasonography screening for at risk populations [6], but even with more refined imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI) HCC can be difficult to differentiate from other liver lesions.

Fortunately, recent advancements in the medical imaging domain have led to the development of several computational approaches that aid tumor diagnosis [7, 8]. These approaches, usually coined radiomics, mine handcrafted features from the raw images and analyze them using varying machine learning algorithms [9, 10]. Although the approach has yielded promising results in liver oncology, one shortcoming of these studies is that results are highly sensitive to the image segmentation and processing steps involved [11, 12]. Further, some argue that relying solely on mineable handcrafted features that satisfy mathematical properties 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 in medical imaging [14, 15]. Unlike handcrafted radiomics, these approaches learn feature representation from sample images using convolutional neural networks (CNN). In doing so, CNNs have been shown to achieve diagnostic performance of HCC on par with experienced radiologists [16] or even outperforming them [17]. Likewise, they have been shown to accurately distinguish HCC from other liver lesions [18] and correctly classify HCC according to

its severity [19]. Yet, despite these findings, few studies considered MR images and no study has been reported discussing deep learning approaches to HCC developement in non-alcoholic fatty liver disease patients. Consequently, this study addresses this gap in the literature, posing the following research questions: (1) What is the added value of deep learning compared to handcrafted radiomics in early HCC detection? (2) Which deep features are identifiable as risk factors for 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 deep learning approach to a baseline radiomic model separately derived. In addition, the interpretation of the deep learning model is augmented by modeling a set of explainable deep features. As such, the information captured by the deep learning model is aided with context interpretable to the practicing radiologist. All code to reproduce the results presented in this study is publicly available [1].

2   Data Description

Given that access to the originally intended data source was not yet 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. [20], while magnetic resonance scans of 20 patients diagnosed with hepatocellular carcinoma are drawn from the cancer imaging archive [21]. Both data sources implemented varying imaging modalities to obtain the MR scans. Hence, to enhance comparability across data sources, only T2-weighted images are selected for the purpose of 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.

3   Network Architecture

To assess the deep learning approach's viability, this study adopts a slightly adjusted 3-dimensional extension of the original ResNet50 [22] architecture and initializes the network's parameters using pretrained weights from the Medical-Net project [23] [2]. Due to the limited sample size encountered in most clinical studies, this technique has become the de facto method for deep learning based applications in medical image analysis [24] and is commonly referred to as transfer learning.

The network's proposed architecture is depicted in figure 1. It consists of four subsequent bottleneck blocks of three convolutional layers and two fully

---

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

[2]MedicalNet comprises a series of pretrained model weights trained on 23 different medical imaging datasets.
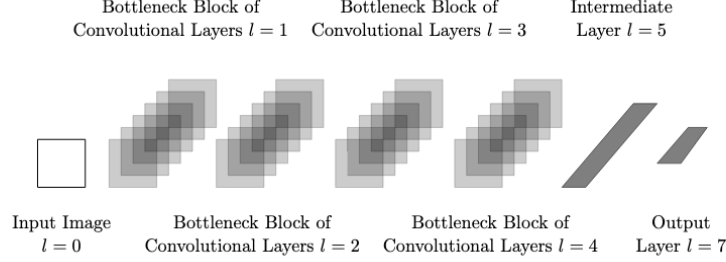
Figure 1: Adjusted ResNet50 Network Architecture

connected layers. As aforementioned, compared to the original ResNet50 architecture, this study thus proposes a slightly adjusted network architecture. That is, instead of only one fully connected output layer, a fully connected intermediate layer is added between the network's output and its last bottleneck block. Doing so, the intermediate layer maps the high-dimensional feature space (2048 features) retained after global average pooling to a lower dimension (128 features). Instead of all 2048 activations, only the feature map comprising the 128 activations of the intermediate layer are thus given as inputs to the network's output layer. The step was introduced to aid the extraction of the explainable deep features and is further discussed in section 4.2.2.

4 Experimental Setup

For the empirical evaluation, the deep learning approach proposed in this study is compared to the baseline radimomics model separately derived. In addition, the interpretation of the deep learning approach is aided by solely modeling the deep features that can be interpreted in the context of handcrafted radiomics. The corresponding experimental setup as well as the derivation of these features is discussed below.

4.1 Preprocessing

Before model training, minor preprocessing steps are performed on the raw images. That is, all MRI scans are are converted from DICOM to nifti file format, the images are rescaled to 128x128x64 pixels and the images' grey-scale 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.

3

4.2 Feature Engineering

4.2.1 Handcrafted Features

To compare the performance of the deep learning approach with a baseline radiomics model, handcrafted features are computed from the raw MR images. Since, prior to their computation, the step requires the study's region of interest, i.e. the liver, to be delineated, this study employs image segmentation using a pretrained nnU-Net [25] liver segmentation model [3]. To illustrate the segmentation performed, an exemplary MRI slice along with its corresponding segmentation mask is presented in figure 2.
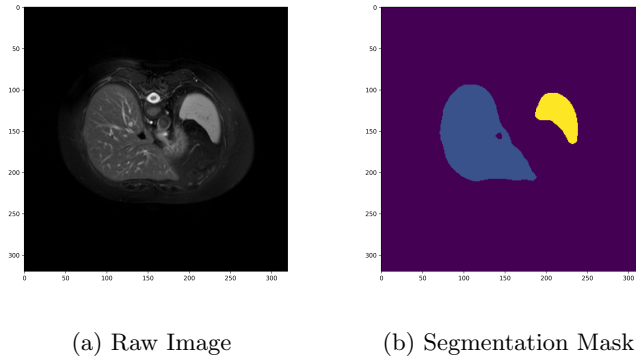

(a) Raw Image   (b) Segmentation Mask

Figure 2: Exemplary MRI Slice with corresponding Segmentation Mask

Following image segmentation, the raw images as well as 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 [27]. Among these features, shape features describe the shape of the traced ROI and its geometric properties such as volume, maximum diameter along different orthogonal directions, maximum surface, tumour compactness, and sphericity. First-order features, by contrast, capture the distribution of individual voxel values without concern for spatial relationships. They are histogram-based properties reporting the mean, median, maximum, minimum values of the voxel intensities on the image, as well as their skewness, kurtosis, uniformity, and randomness (entropy). Finally, second- and higher order features are obtained calculating the statistical inter-relationships between neighboring voxels. They provide a measure of the spatial arrangement of the voxel intensities, and hence of intra-lesion heterogeneity.

---

[3]nnU-Net is a self-configuring, automatic image segmentation algorithm closely based on the original uNet architecture [26].

4.2.2 Explainable Deep Features

To enhance the interpretability of the deep learning approach, this study derives a number of deep 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 original network's 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, to allow for the interpretation of these newly derived deep features, pairwise correlations between all pairs of deep and handcrafted features are computed. Hereby, deep features with an absolute correlation of $|r| \geq 0.5$ with at least one of the handcrafted features are classified as explainable deep features, allowing for their interpretation in the context of handcrafted radiomics.

4.3 Feature Selection

Following their derivation, the high dimensional feature space of the handcrafted and the explainable deep 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 besides that for 6 of the patients diagnosed with HCC handcrafted radiomic features could not be computed. The observations are thus excluded from the set of handcrafted features.

4.4 Model Training

For model training, the data is randomly divided into a training and a validation set using a 80:20 split. The deep learning 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 loss 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 as the best model.

---

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

Note that 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 introduce intermediate layer as well as the network's last bottleneck block before the newly introduced layer are finetuned. All other model parameters are not updated. Note also that since this report merely presents a proof of concept, model hyperparameters are not optimized. All analyses are implemented using the *pytorch* based *MONAI* framework.

To obtain baseline metrics of the handcrafted radiomics model as well as performance metrics of the model fit to the explainable deep features, LASSO regularized logistic regression models are fit to the training set of handcrafted and explainable deep features respectively. To ensure optimal classification performance on the validation set, the LASSO regularization parameter λ is optimised using three-fold cross-validation. 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 optimised. The hyperparameter optimisation is implemented using *scikit-learn*.

4.5 Evaluation Metrics

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.

5 Results

Figure 3 presents the model's accuracy and learning curves plotted over the number of training epochs. As depicted in figure 3b, the model continuously improves the loss on the validation set until plateauing after approximately 120 training epochs. Interestingly enough, the model's accuracy on the validation set also appears to slightly decrease or at least stagnate right after the model starts overfitting on the training data. Despite achieving higher accuracy estimates during previous epochs, the model's best configuration on the validation set thus yields a classification accuracy of 75.2%.

With respect to its comparison to the baseline radiomic model, table 1 presents the models respective performance metrics along with their corresponding bootstrapped 95% confidence intervals. As aforementioned, the deep learning 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
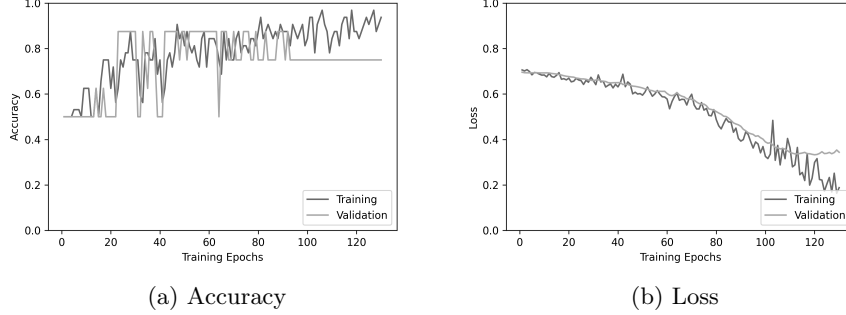
(a) Accuracy   (b) Loss

Figure 3: Accuracy and Loss over Training Epochs

baseline radiomic model (75.6% recall compared to 50.9% recall). Accordingly, the model thus appears to be considerably more sensitive to hepatocellular carcimonia detection, which is especially important in a medical context where the false classification of true positives is considered more severe than predicting false negative cases. As for the baseline radiomic model, an equal 75.2% of the validation set's labels are correctly predicted. The model achieves a very high 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 also consistently overestimates the number of negative cases, yielding a lower F1-score than the deep learning model.

Besides that, it is to be noted that the model solely fit to the explainable deep features achieves a classification performance which is essentially on par with the performance of the deep learning model. In fact, across all performance metrics, the model fit to the explainable deep features equals the predictive performance of the deep learning model, although its confidence bounds are for the most part slightly wider than the deep learning's confidence bounds.

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

Table 1: Performance Metrics of the Different Classifiers

With respect to its features individual effects, the results show that three explainable features were retained after LASSO regularization. Among these features, the results show that var80 exhibits the greatest effect on HCC development ($\beta = -2.096, z = -2.180, p = 0.029$). The feature displays an absolute correlation of $|r| \geq 0.5$ with 10 first-order features. Var32 shows the second

7

greatest effect ($\beta = -1.081, z = -1.792, p = 0.073$). In contrast to the afore-mentioned var80, it is positively (negatively?) related to the image's size zone non-uniformity, which is a higher-order feature measuring the variability of size zone volumes in the image. Finally, var12 displays the smallest effect on HCC development ($\beta = -0.325, z = -0.382, p = 0.703$). In line with var80, the feature is composed of a number of first-order radiomic features, i.e. the intensity value's mean, median, its 90th percentile, and the square-root of the mean over all squared intensity values.

6 Discussion

In this study, the deep learning approach to the classification of hepatocellular carcimonia was compared to a baseline radiomics approach. In addition, the interpretability of the deep learning approach was aided by modeling a set of explainable deep features and interpreting them in the context of handcrafted radiomics. The results show that, despite lower precision, the deep learning model is considerably more sensitive to lesion detection than the baseline model. Since model sensitivity (i.e. recall) presents the arguably most important metric in tumor classification tasks like the given, it can thus be concluded that the deep learning model provides higher clinical utility than the baseline model. The results appear especially promising given the fact that for the purpose of this study no hyperparameter optimization was performed. In a future extension of this study, it thus appears likely that optimizing the model's hyperparameters will yield even higher predictive performance of the deep learning model, further underlining the advantageous applicability of the approach.

As for the interpretation of the deep learning approach, performance metrics of the model solely fit to explainable deep features was on par with the performance metrics of the deep learning model. The results indicate that the loss of information induced by solely modeling the explainable features is negligible when compared to the deep learning model. Note however that in this study only one correlation threshold (i.e. $|r| \geq 0.5$) was assessed 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 that were pretrained on medical images. This is especially important since there is ample evidence that less complex model architectures than ResNet50 can yield better predictive performance on simple binary classification problems like the given [28]. 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 cannot be ruled out that the varying imaging modalities used to obtain the images largely affect the results presented in this proof of concept and thus greatly compromise its generalizability. In a future extension, images from patients di-

agnosed with and without hepatocellular carcimonia 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] "The diagnosis and treatment of hepatocellular carcinoma - ClinicalKey." https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S0740257016301174.

[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. 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.

[21] 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.

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

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.