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 and typically arises in the setting of chronic liver disease or cirrhosis [1]. While HCC already represents one of the leading causes of cancer-related deaths worldwide, its burden is still expected to increase due to the soaring prevalence of non-alcoholic fatty liver disease in the general population [2, 3]. In addition, hepatocellular carcinoma also remains difficult to diagnose, particularly in its early stages. This is because clinical presentation of the disease is often non-specific and can include abdominal pain, weight loss and jaundice [4]. HCC is thus repeatedly only diagnosed at an advanced stage when curative treatments are no longer feasible and even a tissue diagnosis via biopsy is regularly required for definitive confirmation [5].

As a result, medical practitioners have increasingly started to disobey international screening guidelines [6] and introduced more refined medical imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI) to aid tumor diagnosis. In comparison to ultrasonography (US), these techniques offer higher image resolutions and capture the organ's heterogeneity in greater detail. Especially, the complex contrast enhancement offered by MRI make it the widely considered gold standard for the detection of HCC [7]. However, although such cross-sectional imaging techniques have become essential tools in modern oncology, diagnosis still remains challenging [8]. Even with MRI hepatocellular carcinoma can be difficult to differentiate from other liver lesions and detection is possibly affected by the radiologist's experience [9].

Fortunately, recent advancements in the medical imaging domain have led to the development of several computational approaches that allow for the quantification of textural information through the mathematical extraction of spatial distribution and pixel interrelationships [10, 11]. These approaches, usually coined handcrafted radiomics, have proved promising in many fields of medical imaging [12, 13]. In liver oncology, for instance, studies have shown that predictive models based on handcrafted radiomic features are feasible in tumor

diagnosis while others displayed the potential of handcrafted radiomics to disentangle hepatocellular carcinoma from other liver lesions. One shortcoming of these studies is however that results are highly sensitive to the image segmentation and processing steps involved [14, 15]. Moreover, relying solely on mathematically defined handcrafted features may still miss much of the abstract information captured in medical images [16]. Some have thus advocated for the use of deep learning (DL) based radiomics [17, 18]. Unlike handcrafted radiomics, which require features to be mathematically defined, these approaches learn feature representation from sample images using (convolutional) neural networks (CNN). They have not only been shown to outperform the predictive performance of handcrafted radiomic models, but also to surpass human performance in specific medical imaging tasks. Accordingly, CNNs have gained considerable traction in the medical imaging domain, yet few studies have been reported discussing liver tumor diagnosis and even fewer studied MRI compared to CT [19, 20].

Consequently, this study addresses this gap in the literature, posing the following research questions: (1) Which deep features are identifiable as risk factors for HCC development? (2) What is the added value of deep compared to handcrafted radiomics in early HCC detection? Given that this study also aims to contribute to the clinical utility of deep learning based radiomics, it adopts the methodology proposed by Cho et al. [21] and relates the deep features learned from convolutional neural networks to the handcrafted features separately derived. In doing so, the otherwise abstract information captured by the deep features is augmented with context interpretable to the practising radiologist and, as such, the study does not only contribute to the increasing need for early HCC detection but also to the rising demand for explainable artificial intelligence. Note however that due to the restrictions in data access, this report merely present a proof of concept and any inference drawn from it does not yield generalization.

2 Data Description

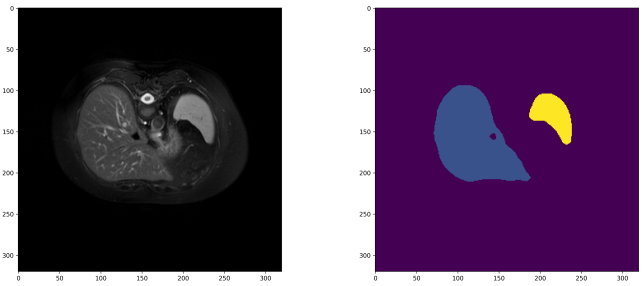
Given that access to the originally intended data source was not yet provided at the point of writing the report, data was drawn from two publicly available data sources. That is, abdominal MR images of 20 healthy individuals were derived from Kavur et al. [22], while magnetic resonance scans of 20 patients diagnosed with hepatocellular carcinoma were drawn from the cancer imaging archive [23]. Both data sources implemented varying imaging modalities to obtain the MR scans, but to enhance comparability only T2-weighted images were 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. Prior to the analysis, all images were converted from DICOM to nifti file format.

3 Experimental Setup

3.1 Feature Engineering

3.1.1 Handcrafted Features

Handcrafted features were derived in a sequential manner. In the first step, the study's region of interest (ROI), i.e. the liver, was delineated using nnU-Net [24], which is a self-configuring, multi-organ segmentation algorithm closely based on the original U-Net architecture [25] and its 3D counterpart [26]. To illustrate this, figure 1 presents an exemplary MRI slice along with its corresponding segmentation mask.



(a) Raw Image (b) Segmentation Mask

Figure 1: Exemplary MRI Scan with corresponding Segmentation Mask

Following the image segmentation, the raw images as well as their corresponding segmentation masks were given as input to derive 14 shape, 18 first-order, and 40 second- and higher-order radiomic features. No filters were applied to the raw images. The step was implemented using the *pyradiomics* library (citation).

3.1.2 Deep Features

Extraction of the deep features was implemented using transfer learning, which refers to the adoption of standard large-scale deep learning models with corresponding pretrained weights [27, 28]. Due to the limited sample size encountered in most clinical studies, the technique has become the de facto method for deep learning based applications in medical image analysis [29]. Henceforth, this study adopted a slightly adjusted 3-dimensional extension of the original ResNet50 [30] architecture. That is, as depicted in figure 2, an intermediate layer was added between the network's average pool and its fully connected layer. The step serves a preliminary feature selection step since it maps the network's high-dimensional feature space (2048 features) to a less dimensional one (128 features).

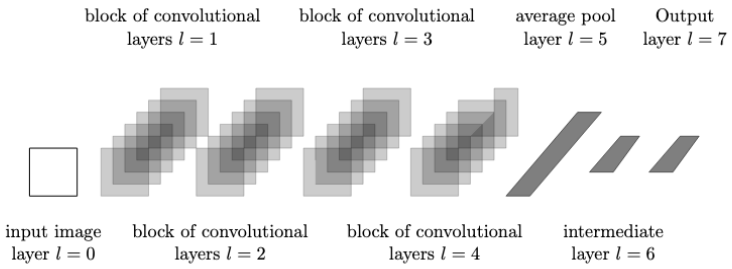


Figure 2: Adjusted ResNet50 Network Architecture

Prior to model training, the data set was randomly divided into a training and a test set using a 80:20 split. The network's parameters were initialized using pretrained weights from the MedicalNet project [31] and retrained over a span of 300 epochs using stochastic gradient descent with batches of size 8 and a learning rate of $\eta = 0.01$. Note, here, that to avoid overfitting, only a set of model parameters were updated. That is, finetuning was only applied to the network's fully connected layer, its newly introduce intermediate layer and the layer before the newly introduced layer. All other model weights were not updated. After model training, the model configuration with the best performance, i.e. with the highest classification accuracy on the test data, was selected. If the current highest classification accuracy was equalled during an epoch, the model configuration with the lower loss on the validation set was selected. Given that this report solely presents a proof of concept, model hyperparameters were not optimized and data augmentation was also not performed. The analysis was implemented using the *pytorch* (citation) based *MONAI* framework (citation).

Following model finetuning, the deep features were extracted under the assumption that the features from the last layer before the fully connected layer can be considered representative of the latent characteristics captured in medical images. Correspondingly, the fully connected layer was removed from the network and the inputs of the newly introduced intermediate layer were extracted, yielding a total of 128 deep features.

3.2 Preprocessing

Before model building, multiple data preprocessing steps were required to enhance the sustainability of prediction for both handcrafted and deep radiomic features. As such, constant (i.e. $\sigma = 0$) as well as quasi-constant (i.e. $\sigma \leq 0.001$) features were removed from both data sets and all features were z-score standardized. In addition, for 6 of the patients diagnosed with HCC handcrafted radiomic features could not be computed. The observations were thus excluded from the set of handcrafted features.

3.3 Feature Selection

After preprocessing, pairwise correlations between all pairs of deep and handcrafted features were computed. The step was performed to ensure the clinical utility of the selected deep features since it allows them to be interpreted in the context of handcrafted radiomics. As such, deep features were classified according to their highest absolute correlation $|r|$ with the set of handcrafted features. Specifically, deep features with an absolute correlation of $|r| \geq 0.4$ with at least one of the handcrafted features were classified as explainable deep features, while the others were considered non-interpretable. In total, three set of features were thus retained: handcrafted features, explainable deep features, and non-interpretable deep features.

Apart from that, this study employed least absolute shrinkage and selection operator (LASSO) regularization to further reduce the feature spaces' high dimensionality. The LASSO regularization parameter λ was optimised using three-fold cross-validation and the hyperparameter value with the highest average classification accuracy across the three folds was selected. For all sets of features, 11 evenly spaced parameter values on a logarithmic scale between 10^{-5} and 10^{5} were given as inputs. Note here that in *scikit-learn's* implementation of logistic regression the complexity parameter $C = \frac{1}{\lambda}$ is tuned instead of λ. Consequently, the values of λ were converted and C was optimised. The hyperparameter optimisation was implemented using *scikit-learn*.

3.4 Model Building

With respect to model building, three sets of features were modeled: handcrafted features, explainable deep features, and the entire set of deep features (i.e. explainable and non-interpretable). The latter was modeled in addition to the explainable deep features since there is ample evidence to believe that the feature representation learned by convolutional neural networks is more comprehensive than the features that have been solely derived mathematically. To thus gain further insight into the loss of information imposed by removing the non-interpretable features, this study compares them with respect to their predictive performance on the test data.

Consequently, the data was divided into a training and a test data set using a 80:20 split. All three sets of features were trained on the same set of observations and LASSO regularized logistic regression models were fit to them respectively.

3.5 Evaluation Metrics

The diagnostic performance of the three classification models was evaluated given the following set of evaluation metrics: accuracy, sensitivity, and specificity. In addition, the model's respective area under the receiver operating characteristic curve (AUC) was assessed.

4 Results

Summary statistics of the deep and handcrafted radiomic features are presented in table **??** in appendix **??**. Table 1, by contrast, presents the models along with their corresponding performance metrics.

Model	Accuracy	Sensitivity	Specificity	AUC
Handcrafted Features	0.71	x	x	0.58
Explainable Features	0.88	x	x	0.73
Explainable and Non-Interpretable Features	0.88	x	x	0.67

Table 1: Performance Metrics of the Different Models

As indicated, the models fit to the explainable as well as the explainable and non-interpretable features consistently outperform the predictive performance of the handcrafted features. Likewise, removing the non-explainable deep features did not appear to affect to the predictive performance of the deep learning based approach. To illustrate this, the model's respective are under the receiver operating characteristic curve are plotted in figure 3. Note again, however, that given the acquired data these results do not yield generalization.

After the LASSO-regularized feature selection, five explainable features were retained and modeled using logistic regression. Among these features, the results show that var46 exhibits the greatest effect on HCC development ($\beta = 7.626, z = 0.925, p = 0.355$). The feature displays a positive correlation of $|r| \geq 0.4$ with the kurtosis of the image's gray-level distribution, thus implying that the difference in gray-level gradients is indicative of tumor development. Var38 shows the second greatest effect ($\beta = -7.309, z = -0.630, p = 0.529$). In line with the aforementioned var46, the feature is also positively related to the kurtosis of the image's gray-level distribution. Apart from that, it displays a negative correlation with the variance difference, which is a measure of gray-level heterogeneity.

The remaining explainable features underline the importance of the shape of the image's gray-level distribution. That is, all other features are equally related to the kurtosis of the image's gray-level distribution. While var38, var46, and var70, however, display a positive correlation with the kurtosis of the image's gray-level distribution, var6 and var42 exhibit a negative relationship. Apart from that, var42 also shows positive correlations with the image's entropy and its sum entropy, i.e. the sum of the differences in neighborhood intensity values.

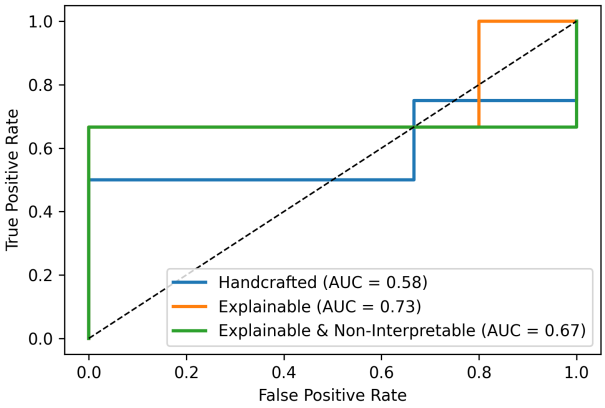


Figure 3: Area under the Receiver Operating Characteristic Curves

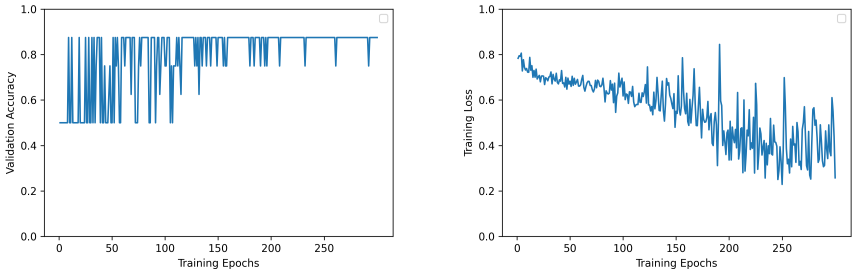
5 Discussion

In this proof of concept study, handcrafted features derived from medical images and deep features learned through convolutional neural networks were compared with respect to their predictive performance of hepatocellular carcimonia. The results show that the deep learning based approach outperformed the handcrafted approach and that the removal of the non-interpretable deep features did not significantly affect the accuracy of the predictions.

Apart from that, five explainable deep features were detected as risk factors of HCC development. The composition of the explainable features clearly underlines the importance of the shape of the image's gray-level distribution. In fact, all five features were consistently correlated (i.e. $|r| \geq 0.4$) with the distribution's kurtosis. Evidently, HCC thus appears to primarily develop in patients with heterogeneous gray-level distributions, indicating the need for radiologist to be especially attentive towards gray-value variations in the liver when assessing patients.

There are a number of major shortcomings to this study. First, the data set was acquired using varying data sources with varying imaging modalities. Second, the data set did not comprise enough observations to effectively train a convolutional neural network. That is, despite only finetuning the final three layers of the network, the model failed to converge at a local minimum (see figure 4b). In a future extension of this study, the data should thus not only comprise more observations, but data augmentation to increase the number of training samples should also be considered. Furthermore, future work should also consider optimization of the network's hyperparameters since it is to be expected that optimized parameters yield faster convergence.

Apart from model convergence, another shortcoming of the study is that it simply adopts a ResNet50 architecture without considering other preexisting



(a) Validation Accuracy (b) Training Loss

Figure 4: Validation Accuracy and Training Loss over Training Epochs

network architectures. This is especially problematic since there is ample evidence that less complex model architectures typically yield better performance for simple binary classification problems like the given. Consequently, in a future extension of this study, all preexisting model architectures, for whom weights that were pretrained on medical imaging data sets exist, should be assessed.

Regardless of feature engineering, modeling the features that were derived from the raw images also comprises major shortcomings. In fact, setting varying random seeds did not only alter the number and size of the model's non-zero coefficients, but in some instances also yielded completely different conclusions. Again, it is, therefore, noted that any inference drawn from this study does not yield generalizations.

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] X. Liu, H. Jiang, J. Chen, Y. Zhou, Z. Huang, and B. Song, "Gadoxetic acid disodium–enhanced magnetic resonance imaging outperformed multidetector computed tomography in diagnosing small hepatocellular carcinoma: A meta-analysis," *Liver Transplantation*, vol. 23, no. 12, pp. 1505–1518, 2017.

[8] R. F. Hanna, V. Z. Miloushev, A. Tang, L. A. Finklestone, S. Z. Brejt, R. S. Sandhu, C. S. Santillan, T. Wolfson, A. Gamst, and C. B. Sirlin, "Comparative 13-year meta-analysis of the sensitivity and positive predictive value of ultrasound, CT, and MRI for detecting hepatocellular carcinoma," *Abdominal Radiology*, vol. 41, pp. 71–90, Jan. 2016.

[9] L. R. Roberts, C. B. Sirlin, F. Zaiem, J. Almasri, L. J. Prokop, J. K. Heimbach, M. H. Murad, and K. Mohammed, "Imaging for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis," *Hepatology*, vol. 67, no. 1, pp. 401–421, 2018.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] H.-h. Cho, H. Y. Lee, E. Kim, G. Lee, J. Kim, J. Kwon, and H. Park, "Radiomics-guided deep neural networks stratify lung adenocarcinoma prognosis from CT scans," *Communications Biology*, vol. 4, pp. 1–12, Nov. 2021.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," June 2016.

[27] 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.

[28] 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.

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

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

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