Liver Cirrhosis Stage Estimation from MRI with Deep Learning

Jun Zeng[1], Debesh Jha[2], Ertugrul Aktas[1], Elif Keles[1], Alpay Medetalibeyoglu[1], Matthew Antalek[1], Amir A. Borhani[1], Daniela P. Ladner[1], Gorkem Durak[1], Ulas Bagci[1]

Abstract— **Cirrhosis is a severe scarring (fibrosis) of the liver and a common endpoint of various chronic liver diseases. The lack of high-quality, large-scale magnetic resonance imaging (MRI) datasets limits the application of data-intensive image analysis methods, such as radiomics and deep learning, for analyzing cirrhosis. Furthermore, analyzing cirrhotic MRIs is challenging due to the varied patterns of cirrhosis, the overlap between stages, and changes in normal abdominal anatomy. The newly introduced large-scale *CirrMRI600+* dataset contains cirrhotic MRIs from different stages and aims to fill these gaps. Using this dataset, we developed a framework for radiomics and deep learning to predict the stages of cirrhosis. Our findings suggest that both methods help predict the stage of cirrhosis, while deep learning significantly surpasses traditional radiomics-based machine learning techniques. To our knowledge, this is the first study to classify the stages of cirrhosis using deep learning on an MRI dataset. The source code will be available at `https://github.com/JunZengz/CirrhosisStage`.**

I. Introduction

Liver cirrhosis is a prevalent disease worldwide and is associated with a high mortality rate [1]. It is characterized by gradual replacement of healthy liver tissue with irreversible scarring or fibrosis caused by various chronic liver diseases [2]. In the early stages of cirrhosis, the liver typically shows no significant morphological changes. As a result, diagnosing cirrhosis at an early stage is challenging and often missed, even by experienced specialists. If the diagnosis is missed and the disease progresses, it leads to various complications that severely affect patient outcomes [3].

Early detection and accurate assessment of cirrhosis stages are crucial for effective disease management and treatment planning. Traditional methods rely on clinical symptoms, laboratory tests, and ultrasound imaging, which have limitations in detecting subtle tissue changes and clinical conditions. In contrast, MRI provides improved diagnostic accuracy for cirrhosis and its stages by delivering detailed images that allow physicians to evaluate the liver's structure and tissue precisely. However, well-curated MRI datasets remain insufficient, limiting the application of data-intensive advanced image analysis methods like deep learning [4]. To address this gap, a large-scale MRI dataset, CirrMRI600+, has been established for liver cirrhosis research [5]. This dataset aims to improve the application of deep learning in analyzing cirrhosis and its stages, enhance the accuracy of diagnosing cirrhosis at earlier stages in clinical practice, elevate clinical diagnostics and patient outcomes, and promote further research.

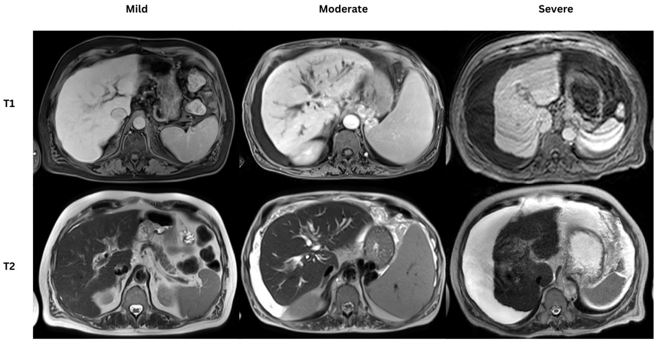

Fig. 1. T1W and T2W MRI scans of cirrhotic patients are shown for mild, moderate, and severe cases.

The cirrhosis MRIs classified into three severity stages by an abdominal radiologist based on severity: mild, moderate, and severe. We utilized radiomics and deep learning to evaluate the stages of cirrhosis. We demonstrated the effectiveness of deep learning techniques applied to extensive MRI datasets. The main contributions of this work are as follows:

- **Cirrhosis stage estimation using MRI:** Existing studies on liver cirrhosis using MRI image analysis are limited by a lack of available MRI datasets, which hinders the advancement of deep learning in this field. By leveraging a large-scale MRI dataset, we demonstrated the effectiveness of radiomics and deep learning in assessing the stages of liver cirrhosis, significantly improving the robustness and reliability of deep learning models.
- **Comprehensive evaluation**: We conducted extensive experiments to assess the stage of cirrhosis and thoroughly evaluated deep learning and radiomics methods. These experiments show that deep learning techniques can assist in automatically recognizing the stages of cirrhosis.

II. Methods

Dataset: The CirrMRI600+ dataset includes 628 high-resolution abdominal MRI scans from 339 patients, each diagnosed with liver cirrhosis. It contains 310 T1W and 318 T2W scans. We split the dataset in an approximate ratio of 8:1:1, with 234 samples for training, 29 for validation, and 28 for testing.

*The project is supported by the NIH funding: R01-CA246704, R01-CA240639, U01 DK127384-02S1, and U01-CA268808.

[1]Jun Zeng, Ertugrul Aktas, Elif Keles, Alpay Medetalibeyoglu, Matthew Antalek, Amir A. Borhani, Daniela P. Ladner, Gorkem Durak and Ulas Bagci are with Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, IL 60611, USA

[2]Debesh Jha is with the Department of Computer Science, University of South Dakota, Vermillion, SD 57069, USA

TABLE I

Performance comparison of deep learning models on CirrMRI600+ T1W MRI dataset.

Dataset	Model	Param(M)	Acc↑	Prec(%)↑	Sens(%)↑	Spec(%)↑	F1(%)↑
CirrMRI600+	VGG-19 [6]	139.58	**0.728**	70.23	**68.80**	**85.86**	**69.36**
	ResNet-50 [7]	23.51	0.611	53.87	56.79	80.44	54.82
	MobileNetV3-S [8]	1.67	0.590	58.93	60.64	80.46	57.29
	ConvNext-B [9]	87.57	0.661	**71.25**	64.09	82.97	64.90
	PVTv2-B2 [10]	24.85	0.652	66.41	62.26	82.42	63.13
	Vim-S [11]	25.44	0.563	57.59	54.05	77.45	54.98
	MedMamba-S [12]	18.62	0.490	48.57	46.97	75.01	47.20
	MambaVision-S [13]	49.37	0.584	54.25	51.57	76.95	51.91

TABLE II

Performance comparison of deep learning models on CirrMRI600+ T2W MRI dataset.

Dataset	Model	Param(M)	Acc↑	Prec(%)↑	Sens(%)↑	Spec(%)↑	F1(%)↑
CirrMRI600+	VGG-19 [6]	139.58	0.574	54.10	49.10	79.37	50.36
	ResNet-50 [7]	23.51	0.574	49.50	48.93	78.72	49.20
	MobileNetV3-S [8]	1.67	0.613	45.88	47.47	77.91	45.36
	ConvNext-B [9]	87.57	0.578	53.90	52.14	80.04	52.54
	PVTv2-B2 [10]	24.85	0.544	49.09	49.97	77.15	49.10
	Vim-S [11]	25.44	0.485	52.56	49.74	75.50	48.83
	MedMamba-S [12]	18.62	0.506	53.04	48.93	76.58	49.10
	MambaVision-T [13]	31.16	**0.638**	**59.13**	**58.01**	**81.38**	**58.43**

Implementation details: In this work, we implemented deep learning models using the PyTorch [14] framework. All experiments were conducted on an NVIDIA A10 GPU. We employed the AdamW optimizer to optimize model parameters, with the learning rate of 1e^{-4} and the batch size of 32. Each model was trained for 100 epochs, with an early stopping patience of 20 epochs. The cross-entropy loss was used to minimize the discrepancy between the predicted output and the ground truth label.

Evaluation metrics: The performance of the models in liver cirrhosis stage estimation was evaluated using standard metrics, including accuracy (Acc), precision (Prec), sensitivity (Sens), specificity (Spec), and F1-score (F1). These metrics provide a comprehensive reflection of the model's capability in accurately estimating different stages of liver cirrhosis.

A. *Results*

To evaluate the performance of deep learning models in estimating the stages of liver cirrhosis, we trained and tested eight state-of-the-art deep learning architectures, first and the largest one in the literature to our best of konwledge. The detailed results are presented in Table I and Table II. Among these models, VGG-19 [6] achieved the highest overall accuracy with a score of 0.728, outperforming the others across various evaluation metrics. Furthermore, we assessed the models' capabilities in estimating each specific stage of cirrhosis, as shown in Table III. The results reveal that while most models attain over 50% precision in estimating mild and severe cirrhosis stages, their effectiveness significantly decreases in moderate cases, with precision dropping below 20% for most models. This discrepancy underscores the increased challenge that severe cases of cirrhosis pose to current deep learning approaches.

To evaluate the impact of network depth and parameter count on liver cirrhosis stage classification, we trained multiple variants of ResNet [7] with different complexities. The results, presented in Table IV, reveal that an increase in parameters does not consistently lead to improved model accuracy. Notably, the pretrained ResNet-50, which has fewer parameters than ResNet-152, demonstrated superior performance. Conversely, the larger pretrained ResNet-50 outperformed ResNet-34. These observations indicate that optimal performance in classification models requires a judicious selection of network architecture and parameter count, rather than simply increasing model size.

We investigated the impact of initial model weights on liver cirrhosis stage estimation by training both from-scratch and pretrained versions of the ResNet and MambaVision architectures. The results are presented in Table IV and Table V. Notably, the from-scratch ResNet-18 demonstrated superior accuracy compared to its pretrained counterpart. Additionally, as shown in Table V, only one of the five pretrained MambaVision models exhibited a performance improvement compared to its from-scratch equivalent. These findings suggest that feature representations learned from the large-scale ImageNet dataset do not significantly enhance performance for this task. This is likely due to the domain discrepancy between natural images and liver cirrhosis images, as well as the high similarity among cirrhotic tissues at different stages. Consequently, estimating liver cirrhosis stages remains challenging, as the models struggle to learn effective discriminative features.

To ensure a fair comparison, radiomics features are also extracted from 2D images and utilized as input features for the machine learning algorithms. The results are presented in Table VI, Table VII and Table VIII, respectively. The features across the three stages of cirrhosis are remarkably similar,

TABLE III

Performance of deep learning models for different liver cirrhosis stages on CirrMRI600+ T2-W MRI dataset.

Liver cirrhosis stage	Model	Prec (%)↑	Sens(%)↑	Spec(%)↑	F1(%)↑
Mild	VGG-19 [6]	82.03	78.68	81.61	80.32
	ResNet-50 [7]	78.78	76.80	77.93	77.78
	MobileNetV3-S [8]	70.73	**94.67**	58.19	**80.97**
	ConvNext-B [9]	**84.82**	68.34	**86.96**	75.69
	PVTv2-B2 [10]	72.73	60.19	75.92	65.87
	Vim-S [11]	71.65	43.57	81.61	54.19
	MedMamba-S [12]	77.33	54.55	82.94	63.97
	MambaVision-T [13]	77.71	76.49	76.59	77.09
Moderate	VGG-19 [6]	19.32	30.08	65.57	23.53
	ResNet-50 [7]	17.24	18.80	75.26	17.99
	MobileNetV3-S [8]	15.00	6.77	**89.48**	9.33
	ConvNext-B [9]	16.11	21.80	68.87	18.53
	PVTv2-B2 [10]	13.16	15.04	72.78	14.04
	Vim-S [11]	22.05	**43.61**	57.73	29.29
	MedMamba-S [12]	22.35	42.86	59.18	29.38
	MambaVision-T [13]	**30.77**	36.09	77.73	**33.22**
Severe	VGG-19 [6]	60.95	38.55	**90.93**	47.23
	ResNet-50 [7]	52.47	51.20	82.96	51.83
	MobileNetV3-S [8]	51.91	40.96	86.06	45.79
	ConvNext-B [9]	60.77	66.27	84.29	63.40
	PVTv2-B2 [10]	61.39	**74.70**	82.74	**67.39**
	Vim-S [11]	63.98	62.05	87.17	63.00
	MedMamba-S [12]	59.42	49.40	87.61	53.95
	MambaVision-T [13]	**68.92**	61.45	89.82	64.97

TABLE IV

Performance comparison of ResNet variants

Dataset	Model	Pretrained	Acc↑
CirrMRI600+	ResNet-18	✗	0.584
	ResNet-18	✓	0.557
	ResNet-34	✗	0.513
	ResNet-34	✓	0.561
	ResNet-50	✗	0.545
	ResNet-50	✓	0.574
	ResNet-152	✗	0.495
	ResNet-152	✓	0.539

TABLE V

Performance comparison of MambaVision variants

Dataset	Model	Pretrained	Acc↑
	MambaVision-T	✗	0.638
	MambaVision-T	✓	0.610
	MambaVision-T2	✗	0.518
	MambaVision-T2	✓	0.500
	MambaVision-S	✗	0.500
	MambaVision-S	✓	0.578
	MambaVision-B	✗	0.568
	MambaVision-B	✓	0.532

making it challenging to effectively segregate these feature regions. Even the GaussianNB, which achieved the highest accuracy among the machine learning methods, only reached an accuracy of 0.540. Its performance is significantly lower compared to most deep learning methods, underscoring the potential of deep learning on liver cirrhosis stage estimation.

III. Discussion and Conclusion

We conduct a comprehensive evaluation of advanced deep-learning models to estimate the stages of liver cirrhosis using a large MRI dataset. This challenging problem remains unresolved due to its morphological complexity and the subtle tissue changes between the cirrhosis stages. Leveraging newer architectures like MambaVision-S and various ResNet variants, we demonstrated that while models like MambaVision-S achieved the highest overall accuracy, there is a marked decrease in performance when estimating severe stages of cirrhosis. This highlights the inherent difficulty in distinguishing advanced disease stages using current deep learning approaches. Our work sets a new benchmark in the field, emphasizing the need for specialized strategies to improve deep learning performance in complex medical imaging tasks like MRI-based cirrhosis staging.

One may argue whether pre-training medical images to computer vision models is suitable for this problem or not. Indeed, fine-tuning pre-trained models on the large-scale dataset Image1k is beneficial for most visual tasks, as it can absorb rich feature representations from existing natural categories. However, the unique characteristics of medical images mean that prior knowledge derived from natural images is not effectively transferable for estimating stages of liver cirrhosis.

Recently, foundation models (FMs) have shown remarkable success in computer vision and medical imaging. Leveraging multiple imaging modalities and clinical reports, large-scale FMs tailored for medical applications hold significant promise for advancing liver cirrhosis stage estimation. However, the data is still limited at the moment and FMs are at suboptimal for the problem tackled in this paper.

Compliance with ethical standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by Northwestern University (No. STU00214545).

TABLE VI

PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS ON CIRRMRI600+ T1W MRI DATASET USING RADIOMICS FEATURES.

Dataset	Model	Acc↑	Prec(%)↑	Sens(%)↑	Spec(%)↑	F1(%)↑
CirrMRI600+	Decision Tree [15]	0.391	42.87	36.90	69.50	37.04
	Random Forest [16]	0.519	**57.68**	49.48	75.35	49.12
	KNeighbors [17]	0.478	50.18	45.87	73.16	46.09
	SVC [18]	0.522	54.18	**50.71**	**75.99**	**50.84**
	GaussianNB [19]	**0.540**	47.18	48.93	75.52	47.39
	Logistic Regression [20]	0.508	53.14	49.05	75.27	49.40
	Gradient Boosting [21]	0.499	54.22	46.91	74.07	47.35

TABLE VII

PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS ON CIRRMRI600+ T2W MRI DATASET USING RADIOMICS FEATURES.

Dataset	Model	Acc↑	Prec(%)↑	Sens(%)↑	Spec(%)↑	F1(%)↑
CirrMRI600+	Decision Tree [15]	0.380	38.88	37.75	69.29	36.31
	Random Forest [16]	0.456	45.62	43.35	72.40	41.88
	KNeighbors [17]	0.434	42.38	40.59	70.34	38.06
	SVC [18]	0.510	**56.86**	48.46	75.21	46.96
	GaussianNB [19]	0.416	42.60	43.00	70.40	41.10
	Logistic Regression [20]	**0.528**	55.12	**49.50**	**75.93**	**48.13**
	Gradient Boosting [21]	0.495	50.20	46.87	74.67	45.92

ACKNOWLEDGMENTS

This project is supported by NIH funding: R01-CA246704, R01-CA240639, U01-DK127384-02S1, and U01-CA268808.

REFERENCES

[1] Yeying Wang, Yang Liu, Yi Liu, Jie Zhong, Jing Wang, Lei Sun, Lei Yu, Yiting Wang, Qinghua Li, Weilin Jin, et al. Remodeling liver microenvironment by l-arginine loaded hollow polydopamine nanoparticles for liver cirrhosis treatment. *Biomaterials*, 295:122028, 2023

[2] Muhammad Ikram Ullah, Ayman Ali Mohammed Alameen, Ziad H Al-Oanzi, Lienda Bashier Eltayeb, Muhammad Atif, Muhammad Usman Munir, and Hasan Ejaz. Biological role of zinc in liver cirrhosis: an updated review. *Biomedicines*, 11(4):1094, 2023.

[3] Madhumita Premkumar and Anil C Anand. Overview of complications in cirrhosis. *Journal of Clinical and Experimental Hepatology*, 12(4):1150–1174, 2022.

TABLE VIII

Performance of machine learning models for different liver cirrhosis stages on CirrMRI600+ T2W MRI dataset.

Liver cirrhosis stage	**Model**	**Prec (%)↑**	**Sens(%)↑**	**Spec(%)↑**	**F1(%)↑**
Mild	Decision Tree [15]	58.11	40.44	68.90	47.69
	Random Forest [16]	63.37	54.23	66.56	58.45
	KNeighbors [17]	57.67	54.23	57.53	55.90
	SVC [18]	68.07	60.82	**69.57**	64.24
	GaussianNB [19]	54.63	38.87	65.55	45.42
	Logistic Regression [20]	**69.26**	**64.26**	**69.57**	**66.67**
	Gradient Boosting [21]	68.50	58.62	71.24	63.18
Moderate	Decision Tree [15]	22.30	45.11	56.91	29.85
	Random Forest [16]	26.36	51.13	60.82	34.78
	KNeighbors [17]	26.85	51.88	61.24	35.38
	SVC [18]	28.36	**58.65**	59.38	38.24
	GaussianNB [19]	27.02	50.38	62.68	35.17
	Logistic Regression [20]	**29.92**	57.14	**63.30**	**39.28**
	Gradient Boosting [21]	27.17	51.88	61.86	35.66
Severe	Decision Tree [15]	36.22	27.71	82.08	31.40
	Random Forest [16]	47.13	24.70	89.82	32.41
	KNeighbors [17]	42.62	15.66	92.26	22.91
	SVC [18]	**74.14**	25.90	**96.68**	38.39
	GaussianNB [19]	46.15	39.76	82.96	**42.72**
	Logistic Regression [20]	66.18	27.11	94.91	38.46
	Gradient Boosting [21]	54.95	**30.12**	90.93	38.91

[4] Hartmut Häntze, Lina Xu, Felix J Dorfner, Leonhard Donle, Daniel Truhn, Hugo Aerts, Mathias Prokop, Bram van Ginneken, Alessa Hering, Lisa C Adams, et al. Mrsegmentator: Robust multi-modality segmentation of 40 classes in mri and ct sequences. *arXiv preprint arXiv:2405.06463*, 2024.

[5] Debesh Jha, Onkar Kishor Susladkar, Vandan Gorade, Elif Keles, Matthew Antalek, Deniz Seyithanoglu, Timurhan Cebeci, Halil Ertugrul Aktas, Gulbiz Dagoglu Kartal, Sabahattin Kaymakoglu, et al. Cirrmri600+: Large scale mri collection and segmentation of cirrhotic liver. *arXiv preprint arXiv:2410.16296*, 2024.

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo

Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 1314–1324, 2019.

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, pages 11976–11986, 2022.

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[11] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[12] Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

[13] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[16] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[17] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[18] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

[19] Irina Rish et al. An empirical study of the naive bayes classifier. In *Proceeding of the IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, page 41–46, 2001.

[20] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.

[21] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 29(5): 1189–1232, 2001.