Scale-Aware Multi-Instance Learning for Early Prognosis of Subjects at Risk of Developing Hepatocellular Carcinoma

Ananya Jana[1] AJ611@CS.RUTGERS.EDU
Ramanathan Arunachalam[1] RAMANATHAN.ARUN@.RUTGERS.EDU
Carlos D Mincapelli[2,3] MINACACD@RWJMS.RUTGERS.EDU
Kaitlyn Catalano[2,3] KLC297@SCARLETMAIL.RUTGERS.EDU
Carlos Catalano[2,3] CJC398@SCARLETMAIL.RUTGERS.EDU
Vinod Rustgi[2,3] VR262@RWJMS.RUTGERS.EDU
Dimitris Metaxas[1] DNM@CS.RUTGERS.EDU
[1] *Rutgers University*
[2] *Rutgers Robert Wood Johnson Medical School, Division of Gastroenterology and Hepatology*
[3] *Center for Liver Diseases and Masses, Rutgers Robert Wood Johnson Medical School, New Brunswick, New Jersey*

Editors: Under Review for MIDL 2022

Abstract

Hepatocellular carcinoma (HCC) is a type of primary liver cancer which can potentially lead to the subject's death. The subject usually experiences the symptoms when the HCC has grown significantly and needs advanced treatment such as surgery and extensive chemotherapy. Thus it would be beneficial to classify subjects at risk of developing HCC based on non-invasive CT scans. In this work, we propose a novel method for pre-cancer stage classification. To our knowledge, we are the first to propose a Deep Learning based method to classify pre-cancer subjects based on CT images. Our hypothesis is that the subjects at risk of developing HCC may have grown different scales/levels of visual cues on the CT scans. We propose a scale-aware method to facilitate the prediction of a free of HCC subject to develop HCC without additional annotation cost. We show the efficacy of our method on a dataset of 60 subjects[1].

Keywords: Hepatocellular carcinoma, liver cancer, Deep Learning, HCC prognosis, early detection, multi-instance learning

1. Introduction

Hepatocellular carcinoma (HCC) is the most common form of primary liver cancer(Nam et al., 2020) and it is becoming an increasingly prevalent cause of cancer-related death(Rawla et al., 2018). During the early stages of HCC, the subject might be asymptomatic(Ayuso et al., 2018). The diagnosis of HCC usually occurs at a more advanced stage(Sun and Sarna, 2008) which needs more complex therapies while the survival expectancy is significantly reduced(Sun and Sarna, 2008). Early screening of HCC in a clinical setting can lead to more effective therapies and better preservation of the liver function(Sun and Sarna, 2008).

[1] https://github.com/ananyajana/pre_hcc_classification

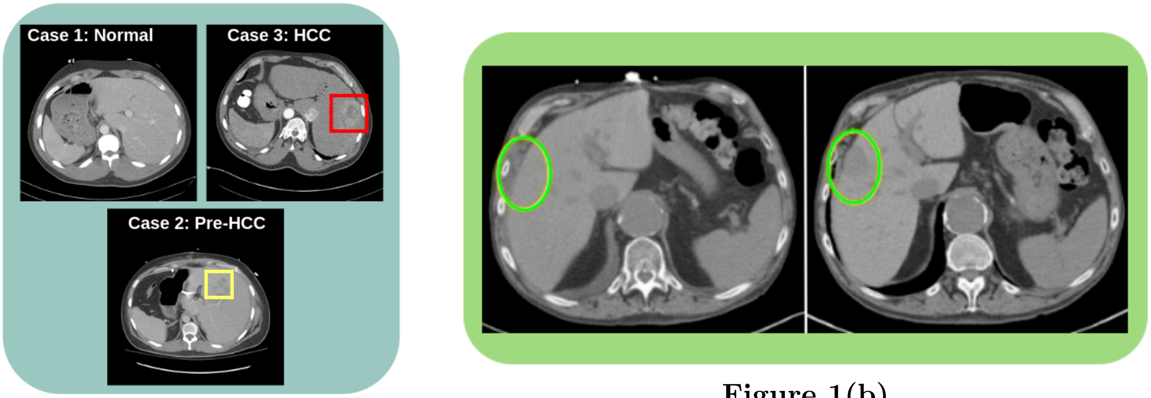

Figure 1(a)

Figure 1(b)

Figure 1: (a)Examples of the Normal, HCC, and pre-HCC CT scans. The red and yellow rectangles mark the cancer and pre-cancer regions respectively. (b) Examples of a preHCC CT scan(left) and later HCC diagnosis CT scan(right) of the same subject. The green oval marks the region where HCC developed.

While biopsy, MRI, and other imaging modalities may provide more information for HCC, CT is a comparatively cheaper, and non-invasive imaging modality. Therefore, it is of great value to identify subjects at risk of developing HCC from CT images. The advent of deep learning methods that automatically extract features has resulted in their recent application to many studies on liver cancer and HCC. Examples include liver cancer/tumor classification(and/or segmentation) from CT, MR, and Ultrasonography data (Christ et al., 2016; Nayak et al., 2019; Christ et al., 2017; Vorontsov et al., 2017; Ghoniem, 2020; Mitrea et al., 2021; Kim et al., 2020), follow-up studies(Vivanti et al., 2017; Morshid et al., 2019) etc. The study of CT subject scans who were later on diagnosed with HCC is less explored with Deep Learning. In this work, we try to address this question as follows- Given a CT scan of a subject free of HCC can we identify if the subject is at risk of developing HCC in the future? We formulate this problem as a multi-instance learning problem. This is consistent with how a medical professional examines each of the 2D slices of a subject's 3D liver volume and then decides if the subject is healthy or has HCC. We will use a term pre-HCC for the subjects suspected of HCC who didn't have HCC diagnosis on the CT scan. These subjects were later diagnosed with HCC. The pre-HCC scans may have grown different scales/resolutions of early signs or visual cues(e.g. textural changes) depending on how far they are from the HCC diagnosis. Recently scale aware networks(Zhang et al., 2021) have been shown to facilitate self-supervision in medical data. Consequently, we enable our method to learn and use these scale changes better by introducing a scale-aware module in the framework of multi-instance learning(MIL). We utilize self-supervision in the scale-aware auxiliary task and thus require no additional annotation .

Contrastive learning methods are a great way for learning representations in a fully self-supervised manner, but they may focus more on the macro details like structure and objects instead of the minute details, as can be seen from the way the positive and negative pairs are built(Chaitanya et al., 2020; Zhou et al., 2021; Taleb et al., 2020). And the minute details hold the key to many interesting discoveries in the medical domain, especially when

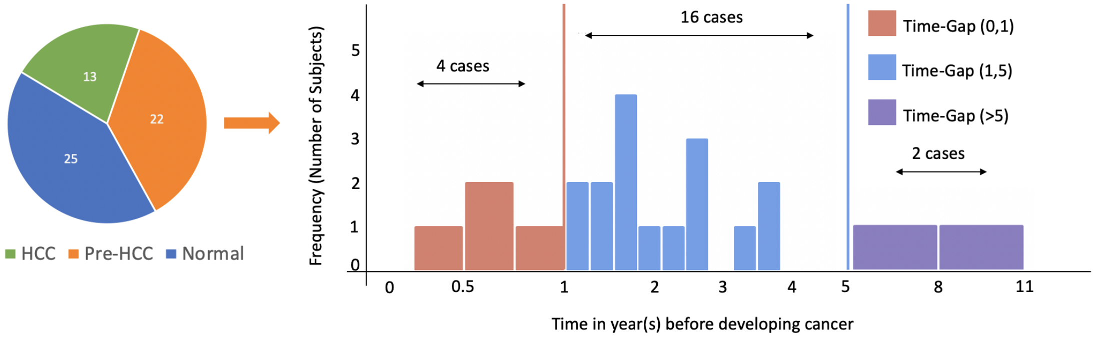

Figure 2: Subjects distribution across the classes(left). Exploring the pre-HCC group more in detail: the time-gap in years between the CT scan and actual HCC diagnosis of the pre-HCC Subjects in our dataset(right).

we consider disease prognosis. From the perspective of contrastive learning, an image and its blurred version may form a positive pair whose representation is similar, but in our application, slight textural changes may mean an entirely different class. Moreover, to focus on only the liver region, we segment out the liver, so there might not be too many structural cues for contrastive learning methods to learn from, unlike the whole axial slice where multiple organs are present and provide a rich structural signal. Thus we choose a scale-aware method for improved self-supervision for this task which in turn makes our proposed method achieve better performance. In summary, our main contributions are: (1) Formulation of a classification problem for identifying subjects at risk of developing HCC. (2) Introduction of a novel Scale-Aware Multi-Instance Learning method for this task, and (3) Demonstration that our method can achieve superior performance on this problem compared to other existing multi-instance learning methods.

2. Problem Formulation

In this work, we formulate the problem at hand as a classification problem into 3 separate classes - Normal subjects who do not have any liver disease, HCC subjects who are already diagnosed with HCC, and pre-HCC subjects who didn't have a HCC diagnosis at the time of the CT scan that we will use, but were diagnosed with HCC at a later point of time. The rationale of such a problem formulation is as follows: (1) it is much easier to get cancer or normal liver score for the entire CT scan rather than actually annotating the cancer region in the entire liver volume of individual subjects, (2) The Normal and HCC (Liver Cancer) scans can still be simpler to annotate compared to the pre-HCC. The pre-HCC subjects pose a challenge for any kind of manual annotation, as we can see from Fig.1(b). The time-gap between the CT scan and actual HCC diagnosis varies widely for the pre-HCC subjects, as can be seen from Fig.2(b), making such an annotation difficult.

3. Methods

In this section, we introduce our Scale-Aware Multi-Instance Learning method. Our proposed method has two steps: (1) Data preprocessing and (2) Classification of Normal, pre-HCC and HCC subjects using a scale-aware network.

3.1. Data Preprocessing

Our data consists of 60 subjects. The subjects distribution is shown in Fig. 2.(a). The 3D CT volume of a subject consists of raw CT slices. We perform Hounsfield windowing on these slices with the range $[-200, 250]$ followed by normalization in the range $[0, 1]$.

3.1.1. 2D liver segmentation (CT data)

The raw CT scans are 3D volumetric data. As we only focus on the liver part in the CT images, we perform 2D liver segmentation to extract liver regions before the classification tasks. All other pixels are set to zero to avoid the negative effect of other organs. We use UNet(Ronneberger et al., 2015) as the segmentation model. The segmentation model is first pretrained on the LiTS(Bilic et al., 2019) dataset. We used this dataset as the images are of similar nature as our dataset. For training the segmentation network on LiTS data, we treated the labels as two class i.e. we treat both the tumor and liver as a single class. Then we annotate the 2D slices of 7 subjects liver volumes in our dataset and fine-tune the segmentation model with the annotated images, utilizing transfer learning. We made a train test split on the annotated liver dataset. The dice score of the segmentation network was 0.9670 on the test liver dataset. After liver segmentation, the slices with an average pixel value below a threshold (5 in our experiments) are discarded [similar to the preprocessing step used by (Jana et al., 2020) for liver Fibrosis classification]. This ensures that slices containing very small liver portions and those slices which do not contain liver portions at all are discarded.

3.2. Scale-Aware Learning

To encourage the network to learn the scale information better, we take twofold measures - first, we introduce the different resolutions of the input image instead of just a single resolution. This enables the network to learn from the rich multi-scale information. Secondly, we add a scale-aware auxiliary module. Our multi-scale image generator creates 3 different scales or resolutions in the following way - the highest resolution image is generated as 224x224 crops of the original 512x512 image, the next resolution is obtained by resizing the 512x512 images to 224x224, and the third resolution is generated by resizing the 224x224 image further to 112x112 and then resizing back to 224x224 using cubic interpolation. Another interesting aspect of this type of multi-scale input is that the lowest two resolutions are actually two different views of the same image at different resolutions. This forces the network to focus on the textural differences between these two images. We want even slight changes in texture and scale to be detected by the network and be treated separately. Hence to push these representations of the same image further, we label these representations as 3 different scale labels and introduce a scale-aware module in our MIL framework. The scale aware module is trained jointly with the original network to classify the 3 different scale labels. The 2D slice-wise training makes our network more robust in detecting changes in texture/scale which helps in turn in the final classification task. This scale-aware module is fully self-supervised i.e., no additional annotations are required. The challenge of this

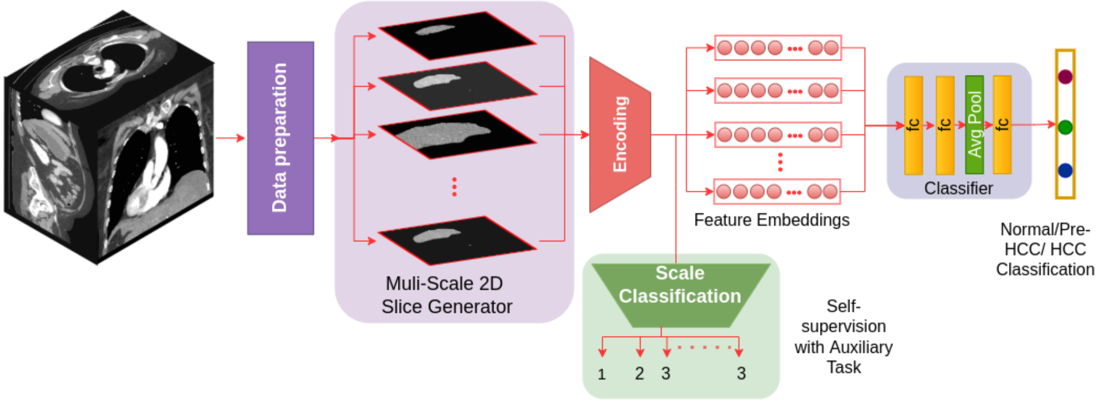

Figure 3: pre-HCC classification: Multi-Scale images are generated from the liver volume, and the feature extractor generates the features from these 2D slices while the scale-aware module jointly predicts the scale labels. The final classifier uses the features to predict a final class label.

classification task comes from the fact that based on different time-gaps the pre-HCC CT scans could look closely similar to the Normal or the HCC, with only distinguishing factor being some very subtle visual cues. These cues are not as strong as anatomical structures which is where contrastive learning based methods seem to focus.

3.3. Classification Network

The network architecture of the classification is shown in Fig.3. It consists of the feature extractor, a scale-aware module and a final classifier. The feature extractor consists of pretrained ResNet-18(He et al., 2016) and extracts local features for each individual 2D slice. As we want subject-wise score predictions, the first two fully-connected (fc) layers process the local features, and then the average pooling layer obtains the global feature for each subject from local features. This global feature is used for prediction through the last fc layer. The cross-entropy loss is adopted to train the classification network. To facilitate the network in understanding the different gradations of changes in the liver images, we add a lightweight scale classification task with each encoder. The scale aware module consists of multiple linear layers. The lightweight nature of the scale-aware module helps to avoid over-fitting. The scale-aware module is only for the training phase and helps the model learn multi-level local representations through multi-scale inputs. In the final or test phase, the scale aware module is not used.

3.4. Objective Function

Let N, F and A denote the classification network, the feature extractor and the auxiliary scale-aware module respectively. The ith subject is denoted with p_i and the class (i.e. Normal/pre-HCC/HCC) as c_i. The subject p_i has multiple 2D slices. The kth 2D slice is denoted by x_k and its scale label as y_k. Then the loss function of the network can be expressed in the following way:

$$Loss = \sum_{p_i, c_i \in P} [L_{CE}(N(p_i), c_i) + \alpha \sum_{x_k, y_k \in p_i} L_{CE}(A(F(x_k)), y_k)] \tag{1}$$

Here α is the weight associated with the loss of the auxiliary module and L_{CE} is the cross entropy loss.

Method	AUC
AttnMIL. (Ilse et al., 2018)	69.51±6.96
Gated AttnMIL. (Ilse et al., 2018)	69.64±4.58
Loss AttnMIL. (Shi et al., 2020)	66.29±7.02
Resnet+Avg Pool (Jana et al., 2020)	72.24±2.46
Scale Aware MIL(Ours)	**77.43±1.74**

Table 1: Mean AUC values of classification using different methods (Three-fold cross validation, average of 10 repeated experiments).

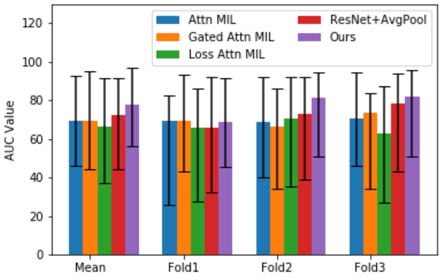

Figure 4: AUC values and 95% confidence intervals for different folds

4. Experiments

4.1. Dataset Evaluation Metrics

The dataset used in our experiments consists of 60 CT volumes with one CT volume per subject. There are 25, 22 and 13 subjects in the categories Normal, pre-HCC and HCC respectively as shown in Fig.2(a). This type of data is usually difficult to acquire as it needs longitudinal observation of the subjects. Only when a subject is diagnosed with HCC, the scans of that subject taken at previous time points are considered pre-HCC. There is no overlap of subjects among these three categories. All data are private and shared by RWJ Medical School after being de-identified and IRB approval. We perform 3-fold cross-validation in all experiments. Area Under ROC Curve(AUC) is used as the performance metric of the models. We also calculate the 95 % confidence interval for AUC.

4.2. Implementation Details

We implement our method using the PyTorch (Paszke et al., 2019) library. In the training phase, we utilize ImageNet (Deng et al., 2009) pretrained ResNet-18 without fc layer as the feature extractor. The network is fully trainable, enabling it to learn lower-level details, which helps in understanding the subtle variations in scale and texture.

For all experiments, we train models with the Adam optimizer and cross-entropy loss for 30 epochs. The learning rate, batch size, weight decay and alpha to be 0.0001, 4, 0.01 and 1.0 respectively. The best model of the 30 epochs is selected for testing. The batch size 4 here means that at a time the network is trained with all the 2D slices from 4 subjects.

4.3. Results

4.3.1. Comparison with state-of-the-art methods

We compare our method with three state-of-the-art multi-instance learning methods (Ilse et al., 2018; Shi et al., 2020). The results are shown in Table 1. Our results outperform all the methods. We show the performance of the models on each of the three folds in Fig. 4

Method	Scale 1	Scale 2	Scale 3	SA Module	Pre-trained	Average AUC
Ablation1	✓	✗	✗	✗	✗	67.34±2.15
Ablation2	✓	✗	✗	✗	✓	72.24±2.46
Ablation3	✗	✓	✗	✗	✗	68.32±1.38
Ablation4	✗	✓	✗	✗	✓	68.07 ±2.36
Ablation5	✓	✓	✗	✗	✗	69.29±2.13
Ablation6	✓	✓	✗	✗	✓	73.84±1.7
Ablation7	✓	✓	✗	✓	✗	68.8±2.25
Ablation8	✓	✓	✗	✓	✓	73.38±2.27
Ablation9	✓	✓	✓	✗	✗	72±1.45
Ablation10	✓	✓	✓	✗	✓	76.47±1.6
Ablation11	✓	✓	✓	✓	✗	71.86±1.19
Ours	✓	✓	✓	✓	✓	**77.43±1.74**

Table 2: Ablation Study. Best results are shown in bold and second-best results in underline.

Method	AUC
Resnet + AvgPool(pred 178)	79.48±2.74
Resnet + AvgPool + Multi-scale input	79.63±2.62
Scale-Aware MIL(ours)	**84.09±1.36**

Table 3: Mean AUC values of Normal Vs pre-HCC prediction using different methods (Three-fold cross-validation, average of 10 repeated experiments).

Compared to the other methods, our method has more discriminative power of different scales and therefore achieves much better performance. Previous works already mentioned are more generic in nature and as a result they may not leverage the disease-specific features to the best possible extent. The generic MIL approaches perform worse in this problem is because their main focus is to improve the aggregation of the local features from individual 2D slices. But in this problem, due to the transitional nature of the features in the pre-HCC category, more attention needs to be placed on the simultaneous strengthening the encoders. This is exactly what our method does which leads to a significant increase in the performance.

4.3.2. Ablation Study

We performed extensive ablation studies to illustrate the effectiveness of the proposed method. The results are shown in Table 2. It can be observed that the introduction of scale-aware self-supervised auxiliary training, and the multi-scale/resolution inputs all have positive effects on the classification performance. Without the multi-scale inputs and scale aware module, the results (Ablation2) are worse compared with those leveraging multi-scale learning. For fair comparison all our models have the same backbone - ImageNet pretrained ResNet-18. We also try our method on the two class classification of Normal Vs. Pre-HCC subjects. We find that our method improves the performance significantly for the two class problem as well and is shown in Table 3.

5. Discussion and Future Directions

Our proposed method incorporates self-supervised scale-awareness in the framework of multi-instance learning. Our encouraging results lead to the realization that the traditional operations(e.g. resizing, resampling) on natural images may not be quite suitable all the time for medical images and could lead to dropping out or tampering the disease cues in the transitional phase. Applying Deep Learning for disease classification/segmentation to predict disease prognosis need a suitable processing of the image data and the design of self-supervision. We want the models to learn meaningful representations while removing noise and diverse multi scale representations of the same image are needed. At the same time we do not want the models to treat valid cues as noise by treating two medically different cases as the same. How we preserve the subtle cues and yet learn powerful representations is one of the open challenges. HCC can still be thought of as a simpler problem because the cues may still form an irregular local group though not as prominent as the fully grown HCC tumor cues. But there are liver diseases where textural cues can give more information e.g. liver Fibrosis and the performance of Deep Learning methods there is still moderate(Jana et al., 2021). Future directions stemming from this work leads to the questions like can super-resolution rather than resizing be a solution and if so, how do we deal with the associated computational challenges.

6. Conclusion

In this work we proposed a scale-aware multi-instance learning method to predict whether a subject has healthy liver or has HCC or is at the risk of developing HCC in the future. Our method can help screen the subjects at risk of developing HCC early and therefore increasing the chances of survival of the subject. The scale-aware nature of our method helps in achieving this by paying attention to the subtle changes in the liver which might be otherwise imperceptible to the normal human vision. This work can serve as a starting point where we formulate a much complex problem of HCC prognosis as a classification problem and obtain interesting results and insights with our novel method.

7. Compliance with ethical standards

This research study includes private data of human subjects from our collaborative partner, and it was approved by the Rutgers University Institutional Review Board (IRB).

8. Acknowledgments

This research was supported by NSF: IIS-1703883, CNS-1747778, CCF-1733843, IIS-1763523, IIS-1849238 – 825536, and MURI-Z8424104 -440149. We thank Dr. Hui Qui and Dr. Zhennan Yan for their useful discussion and suggestions.

References

Carmen Ayuso, Jordi Rimola, Ramón Vilana, Marta Burrel, Anna Darnell, Ángeles García-Criado, Luis Bianchi, Ernest Belmonte, Carla Caparroz, Marta Barrufet, et al. Diagnosis and staging of hepatocellular carcinoma (hcc): current guidelines. *European journal of radiology*, 101:72–81, 2018.

Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

Patrick Ferdinand Christ, Florian Ettlinger, Felix Grün, Mohamed Ezzeldin A Elshaera, Jana Lipkova, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic, et al. Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Rania M Ghoniem. A novel bio-inspired deep learning approach for liver cancer diagnosis. *Information*, 11(2):80, 2020.

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*, pages 770–778, 2016.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Ananya Jana, Hui Qu, Puru Rattan, Carlos D Minacapelli, Vinod Rustgi, and Dimitris Metaxas. Deep learning based nas score and fibrosis stage prediction from ct and pathology data. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 981–986. IEEE, 2020.

Ananya Jana, Hui Qu, Carlos D Minacapelli, Carolyn Catalano, Vinod Rustgi, and Dimitris Metaxas. Liver fibrosis and nas scoring from ct images using self-supervised learning and texture encoding. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1553–1557. IEEE, 2021.

Junmo Kim, Ji Hye Min, Seon Kyoung Kim, Soo-Yong Shin, and Min Woo Lee. Detection of hepatocellular carcinoma in contrast-enhanced magnetic resonance imaging using deep learning classifier: A multi-center retrospective study. *Scientific Reports*, 10(1):1–11, 2020.

Delia Mitrea, Radu Badea, Paulina Mitrea, Stelian Brad, and Sergiu Nedevschi. Hepatocellular carcinoma automatic diagnosis within ceus and b-mode ultrasound images using advanced machine learning methods. *Sensors*, 21(6):2202, 2021.

Ali Morshid, Khaled M Elsayes, Ahmed M Khalaf, Mohab M Elmohr, Justin Yu, Ahmed O Kaseb, Manal Hassan, Armeen Mahvash, Zhihui Wang, John D Hazle, et al. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence*, 1(5):e180021, 2019.

Joon Yeul Nam, Jeong-Hoon Lee, Junho Bae, Young Chang, Yuri Cho, Dong Hyun Sinn, Bo Hyun Kim, Seoung Hoon Kim, Nam-Joon Yi, Kwang-Woong Lee, et al. Novel model to predict hcc recurrence after liver transplantation obtained using deep learning: a multicenter study. *Cancers*, 12(10):2791, 2020.

Akash Nayak, Esha Baidya Kayal, Manish Arya, Jayanth Culli, Sonal Krishan, Sumeet Agarwal, and Amit Mehndiratta. Computer-aided diagnosis of cirrhosis and hepatocellular carcinoma using multi-phase abdomen ct. *International journal of computer assisted radiology and surgery*, 14(8):1341–1352, 2019.

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. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

Prashanth Rawla, Tagore Sunkara, Pradhyumna Muralidharan, and Jeffrey Pradeep Raj. Update in global trends and aetiology of hepatocellular carcinoma. *Contemporary oncology*, 22(3):141, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5742–5749, 2020.

Virginia Chih-Yi Sun and Linda Sarna. Symptom management in hepatocellular carcinoma. *Clinical journal of oncology nursing*, 12(5):759, 2008.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *arXiv preprint arXiv:2006.03829*, 2020.

Refael Vivanti, Adi Szeskin, Naama Lev-Cohain, Jacob Sosna, and Leo Joskowicz. Automatic detection of new tumors and tumor burden evaluation in longitudinal liver ct scan studies. *International journal of computer assisted radiology and surgery*, 12(11): 1945–1957, 2017.

Eugene Vorontsov, An Tang, David Roy, Christopher J Pal, and Samuel Kadoury. Metastatic liver tumour segmentation with a neural network-guided 3d deformable model. *Medical & biological engineering & computing*, 55(1):127–139, 2017.

Xiaoman Zhang, Shixiang Feng, Yuhang Zhou, Ya Zhang, and Yanfeng Wang. Sar: Scale-aware restoration learning for 3d tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 124–133. Springer, 2021.

Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3499–3509, 2021.

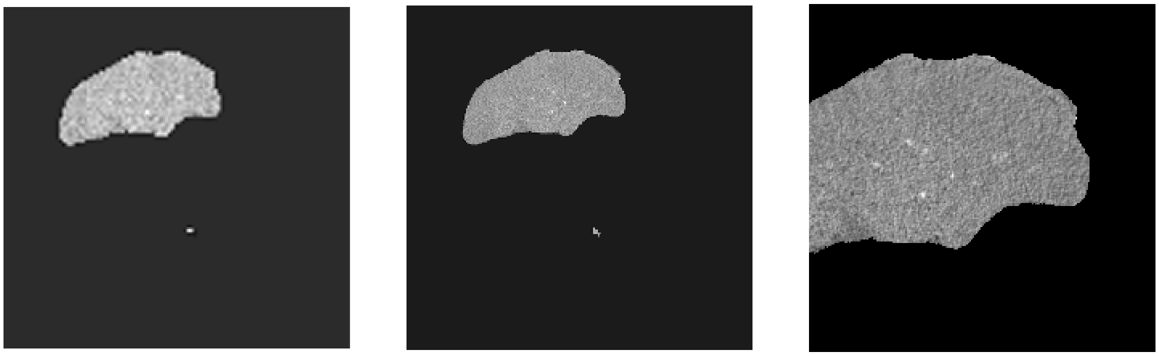

Figure 1(a) **Figure 1(b)** **Figure 1(c)**

Figure 5: Examples of the different scales from the same 2D slice

We show the sample images of the three scales from a 2D slice in Fig. 5.

Appendix A. Additional Implementation Details

Our baseline multi-instance learning network architecture is similar to the network architecture proposed by (Jana et al., 2020) et. al for Fibrosis classification on CT data. There are two reasons behind choosing such a baseline architecture - there is no existing baseline for classifying pre-hcc subjects based on CT images. Logically a good starting point would be to pick up the network architecture for another liver disease classification. Besides, Liver Fibrosis disease involves subtle alterations in liver texture which holds true in case of pre-HCC as well. The liver Fibrosis classification network is also a multi instance learning architecture and our classification problem also needed a multi instance learning framework.

As our dataset is small, we chose to perform k fold cross validation method. Our 3 fold cross validation was performed in the following way - the 60 subjects were shuffled and then split into 3 equal sized groups. We name these groups as split 1, split 2 and split 3. In Fold 1 experiments, the split 1 is used as the test set and split 2 and 3 are used for training. In fold 2, the split 2 is used for test and split 1 and split 3 are used for training. In Fold 3, the split 3 is used as a test set and split 1 and 2 are used for training. The experiments on these three folds are performed separately. The three fold cross validation ensures that the effect of any bias is discarded. We calculate the mean AUC value of an experiment as an average of the fold-wise AUCs. Since our dataset is small we run our experiment 10 times and then take the average of the mean AUC. We report this result in our paper. Out of the 30 epochs, we choose the epoch with the best AUC.
ResNet-18 is the backbone for all the methods including the other multi instance learning(MIL) methods. Implementations of the other models are provided in the code repository provided. The other methods are trained on the same dataset and in similar way.

Method	AUC
AttnMIL. (Ilse et al., 2018)	69.51±6.96
AttnMIL. (Ilse et al., 2018)+M	70.64±3.25
AttnMIL. (Ilse et al., 2018)+SA	72.4±3.31
Gated AttnMIL. (Ilse et al., 2018)	69.64±4.58
Gated AttnMIL. (Ilse et al., 2018)+M	72.05±3.89
Gated AttnMIL. (Ilse et al., 2018)+SA	74.76±2.42
Loss AttnMIL. (Shi et al., 2020)	66.29±7.02
Loss AttnMIL. (Shi et al., 2020)+M	66.14±5.91
Loss AttnMIL. (Shi et al., 2020)+SA	68.19±6.18
Resnet + Avg Pool	72.24±2.46
Scale Aware MIL(Ours)	**77.43±1.74**

Table 4: Mean AUC values of classification using different methods (Three-fold cross validation, average of 10 repeated experiments).

Appendix B. Additional Ablation Study

In this section we provide the results of our extensive experimentation. We add the multi scale input(M) and Scale Aware(SA) modules to the other methods as well. The results are provided in Table 4. As we can see, the attention mechanism in general did not give a good performance in our classification problem. We suspect this is due to the fact that our dataset is size is small and it is not sufficient for the attention mechanism to learn good weights. But we can notice that even with the other methods, our proposed SA module achieves better performance.

We also performed additional 5 ablation study as shown in Table 5. For the sake of completeness, we extend the Table 2 by adding the additional ablation results and present that as Table 5.

Method	Scale 1	Scale 2	Scale 3	SA Module	Pre-trained	Average AUC
Ablation1	✓	✗	✗	✗	✗	67.34±2.15
Ablation2	✓	✗	✗	✗	✓	72.24±2.46
Ablation3	✗	✓	✗	✗	✗	68.32±1.38
Ablation4	✗	✓	✗	✗	✓	68.07 ±2.36
Ablation5	✓	✓	✗	✗	✗	69.29±2.13
Ablation6	✓	✓	✗	✗	✓	73.84±1.7
Ablation7	✓	✓	✗	✓	✗	68.8±2.25
Ablation8	✓	✓	✗	✓	✓	73.38±2.27
Ablation9	✓	✓	✓	✗	✗	72±1.45
Ablation10	✓	✓	✓	✗	✓	76.47±1.6
Ablation11(new)	✓	✓	✓	✓	✗	71.86±1.19
Ablation12 (new)	✗	✓	✓	✓	✓	76.98±1.41
Ablation13 (new)	✗	✓	✓	✗	✓	75.42±1.35
Ablation14 (new)	✗	✗	✓	✗	✓	69.07±2.6
Ablation15 (new)	✓	✗	✓	✓	✓	74.24±2
Ablation16 (new)	✓	✗	✓	✗	✓	72.21±1.91
Ours	✓	✓	✓	✓	✓	**77.43±1.74**

Table 5: Ablation Study. Best results are shown in bold and second-best results in underline.

Method	AUC
Resnet + AvgPool (pred 179)	65.2±6.96
Resnet + AvgPool + Multi-scale input	62.64±6.4
Scale-Aware MIL(ours)	**67.96±3.34**

Table 6: Mean AUC values of HCC Vs pre-HCC prediction using different methods (Three-fold cross-validation, average of 10 repeated experiments).

Method	AUC
Resnet + AvgPool	81.05±4.31
Resnet + AvgPool + Multi-scale input	91.5±2.36
Scale-Aware MIL(ours)	**95.66±2.39**

Table 7: Mean AUC values of Normal Vs HCC prediction using different methods (Three-fold cross-validation, average of 10 repeated experiments).

Appendix C. Additional Two class classification experiments

We perform two two class classification experiments on the Normal Vs preHCC and preHCC Vs HCC. The results are given in the tables Table 7, Table 6. Our experiments show that of all the classification problems, preHCC Vs HCC is the hardest classification problem. this can partly be attributed to the preHCC subjects distribution in our dataset. We have more preHCC subjects with the time-gap 0-5 years and it is likely that these subjects have early visual cues of cancer which causes confusion to the network. This gives a possibility to be able to screen preHCC subjects with that time-gap. Determining the exact time-gap suitable for screening subjects at risk would require exploring this problem with larger datasets. But our work shows the possibility and serves as a starting point for Deep Learning based methods for screening subjects at risk of HCC based on CT images.

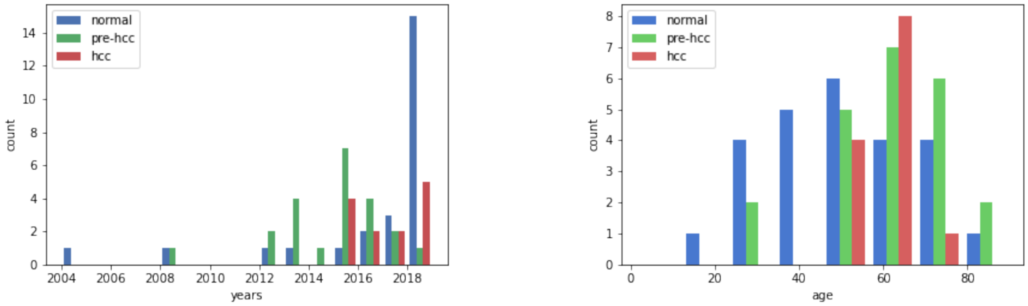

Figure 6(a) **Figure 6(b)**

Figure 6: (a)Distribution of the (a) scan years and (b) subject age among the different groups - normal , pre-hcc an dhcc

Group	Female	Male
Normal	18	7
pre-HCC	9	13
HCC	5	8

Table 8: Number of Female and Male subjects in the different groups - normal, pre-hcc and hcc

Appendix D. Additional Details about the Dataset

In this section we provide the scan years of the subjects across the different classesof subjects - normal, preHCC and HCC in Figure 6. The scans are taken between 2004 and 2019 and maximum scans are taken between the years 2016 and 2019 across the multiple groups. We also show the distribution of the subject ages among the multiple groups. Moreover, we show the distribution of gender among these groups Table 8.