A Reliable and Interpretable Framework of Multi-view Learning for Liver Fibrosis Staging

Zheyao Gao[1], Yuanye Liu[1], Fuping Wu[2], Nannan Shi[3], Yuxin Shi[3], and Xiahai Zhuang[1(✉)]

[1] School of Data Science, Fudan University, Shanghai, China
zxh@fudan.edu.cn
[2] Nuffield Department of Population Health, University of Oxford, Oxford, UK
[3] Department of Radiology, Shanghai Public Health Clinical Center, Shanghai, China
http://www.sdspeople.fudan.edu.cn/zhuangxiahai/

Abstract. Staging of liver fibrosis is important in the diagnosis and treatment planning of patients suffering from liver diseases. Current deep learning-based methods using abdominal magnetic resonance imaging (MRI) usually take a sub-region of the liver as an input, which nevertheless could miss critical information. To explore richer representations, we formulate this task as a multi-view learning problem and employ multiple sub-regions of the liver. Previously, features or predictions are usually combined in an implicit manner, and uncertainty-aware methods have been proposed. However, these methods could be challenged to capture cross-view representations, which can be important in the accurate prediction of staging. Therefore, we propose a reliable multi-view learning method with interpretable combination rules, which can model global representations to improve the accuracy of predictions. Specifically, the proposed method estimates uncertainties based on subjective logic to improve reliability, and an explicit combination rule is applied based on Dempster-Shafer's evidence theory with good power of interpretability. Moreover, a data-efficient transformer is introduced to capture representations in the global view. Results evaluated on enhanced MRI data show that our method delivers superior performance over existing multi-view learning methods.

Keywords: Liver fibrosis · Multi-view learning · Uncertainty

Z. Gao and Y. Liu—These two authors contribute equally.

X. Zhuang—This work was funded by the National Natural Science Foundation of China (grant No. 61971142 and 62111530195).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_18.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 178–188, 2023.
https://doi.org/10.1007/978-3-031-43904-9_18

1 Introduction

Viral or metabolic chronic liver diseases that cause liver fibrosis impose great challenges on global health. Accurate staging for the severity of liver fibrosis is essential in the diagnosis of various liver diseases. Current deep learning-based methods [24,25] mainly use abdominal MRI and computed tomography (CT) data for liver fibrosis staging. Usually, a square sub-region of the liver instead of the whole image is cropped as input features, since the shape of the liver is irregular and unrelated anatomies in the abdominal image could disturb the training of deep learning models. To automatically extract the region of interest (ROI), a recent work [8] proposes to use slide windows to crop multiple image patches around the centroid of the liver for data augmentation. However, it only uses one patch as input at each time, which only captures a sub-view of the liver. To exploit informative features across the whole liver, we formulate this task as a multi-view learning problem and consider each patch as a view.

The aim of multi-view learning is to exploit complementary information from multiple features [23]. The central problem is how to integrate features from multiple views properly. In addition to the naive method that concatenates features at the input level [5], feature-level fusion strategies seek a common representation between different views through canonical correlation analysis [12,22] or maximizing the mutual information between different views using contrastive learning [1,21]. In terms of decision-level fusion, the widely used methods are decision averaging [18], decision voting [14], and attention-based decision fusion [9]. However, in the methods above, the weighting of multi-view features is either equal or learned implicitly through model training, which undermines the interpretability of the decision-making process. Besides, they are not capable of quantifying uncertainties, which could be non-trustworthy in healthcare applications.

To enhance the interpretability and reliability of multi-view learning methods, recent works have proposed uncertainty-aware decision-level fusion strategies. Typically, they first estimate uncertainties through Bayesian methods such

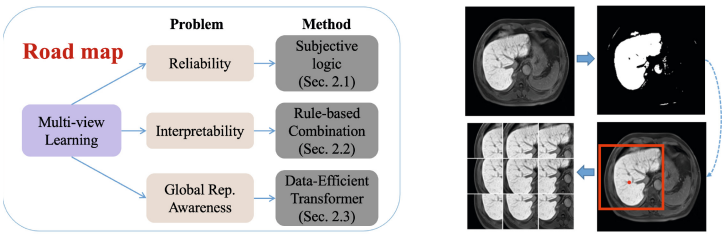


Fig. 1. (a) The road map of this work. (b) The pipeline to extract sub-views of the liver. First, the foreground is extracted using intensity-based segmentation. Based on the segmentation, a square region of interest (ROI) centered at the centroid of the liver is cropped. Then overlapped sliding windows are used in the ROI to obtain nine sub-views of the liver.

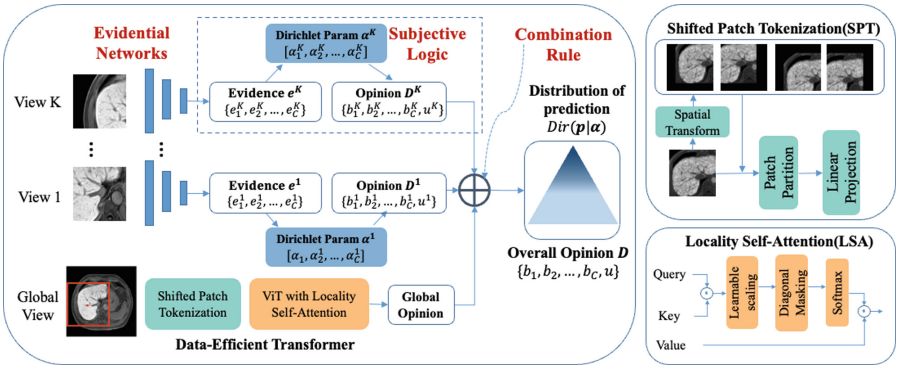


Fig. 2. The left side shows the main framework. Multi-view images are first encoded as evidence vectors by evidential networks. For each view, an opinion with uncertainty u is derived from evidence, under the guidance of subjective logic. Finally, the opinions are combined based on an explicit rule to derive the overall opinion, which can be converted to the distribution of classification probabilities. The right side illustrates the SPT and LSA modules in the data-efficient transformer that serves as the evidential network for the global view.

as Monte-Carlo dropout [20], variational inference [19], ensemble methods [4], and evidential learning [17]. Then, the predictions from each view are aggregated through explicit uncertainty-aware combination rules [7,20], as logic rules are commonly acknowledged to be interpretable in a complex model [26]. However, the predictions before the combination are made based on each independent view. Cross-view features are not captured to support the final prediction. In our task, global features could also be informative in the staging of liver fibrosis.

In this work, we propose an uncertainty-aware multi-view learning method with an interpretable fusion strategy of liver fibrosis staging, which captures both global features across views and local features in each independent view. The road map for this work is shown in Fig. 1(a). The uncertainty of each view is estimated through the evidential network and subjective logic to improve reliability. Based on the uncertainties, we apply an explicit combination rule according to Dempster-Shafer's evidence theory to obtain the final prediction, which improves explainability. Moreover, we incorporate an additional global view to model the cross-view representation through the data-efficient transformer.

Our contribution has three folds. First, we are the first to formulate liver fibrosis staging as a multi-view learning problem and propose an uncertainty-aware framework with an interpretable fusion strategy based on Dempster-Shafer Evidence Theory. Second, we propose to incorporate global representation in the multi-view learning framework through the data-efficient transformer network. Third, we evaluate the proposed framework on enhanced liver MRI data. The results show that our method outperforms existing multi-view learning methods and yields lower calibration errors than other uncertainty estimation methods.

2 Methods

The aim of our method is to derive a distribution of class probabilities with uncertainty based on multiple views of a liver image. The pipeline for view extraction is shown in Fig. 1(b). A square region of interest (ROI) is cropped based on the segmentation of the foreground. Then nine sub-views of the liver are extracted in the ROI through overlapped sliding windows. The multi-view learning framework is shown in Fig. 2. Our framework mainly consists of three parts, *i.e.*, evidential network, subjective logic, and combination rule. The evidential networks encode local views and the whole ROI as global view to evidence vectors $\boldsymbol{e}$. For local views, the networks are implemented with the convolutional structure. While for the global view, a data-efficient vision transformer with shifted patch tokenization (SPT) and locality self-attention (LSA) strategy is applied. Subjective logic serves as a principle that transforms the vector $\boldsymbol{e}$ into the parameter $\boldsymbol{\alpha}$ of the Dirichlet distribution of classification predictions, and the opinion $\boldsymbol{D}$ with uncertainty u. Then, Dempster's combination rule is applied to form the final opinion with overall uncertainty, which can be transformed into the final prediction. The details of subjective logic, Dempster's combination rule, the data-efficient transformer, and the training paradigm are discussed in the following sections.

2.1 Subjective Logic for Uncertainty Estimation

Subjective logic, as a generalization of the Bayesian theory, is a principled method of probabilistic reasoning under uncertainty [10]. It serves as the guideline of the estimation of both uncertainty and distribution of predicted probabilities in our framework. Given an image x_k from view k, $k \in \{1, 2, \cdots, K\}$, the evidence vector $\boldsymbol{e}^k = [e_1^k, e_2^k, ..., e_C^k]$ with non-negative elements for C classes is estimated through the evidential network, which is implemented using a classification network with softplus activation for the output.

According to subjective logic, the Dirichlet distribution of class probabilities $Dir(\boldsymbol{p}^k|\boldsymbol{\alpha}^k)$ is determined by the evidence. For simplicity, we follow [17] and derive the parameter of the distribution by $\boldsymbol{\alpha}^k = \boldsymbol{e}^k + 1$. Then the Dirichlet distribution is mapped to an opinion $\boldsymbol{D}^k = \{\{b_c^k\}_{c=1}^C, u^k\}$, subject to

$$u^k + \sum_{c=1}^{C} b_c^k = 1, \tag{1}$$

where $b_c^k = \frac{\alpha_c^k - 1}{S^k}$ is the belief mass for class c, $S^k = \sum_{c=1}^{C} \alpha_c^k$ is the Dirichlet strength, and $u^k = \frac{C}{S^k}$ indicates the uncertainty.

The predicted probabilities $\tilde{\boldsymbol{p}}^k \in \mathbb{R}^C$ of all classes are the expectation of Dirichlet distribution, *i.e.*, $\tilde{\boldsymbol{p}}^k = \mathbb{E}_{Dir(\boldsymbol{p}^k|\boldsymbol{\alpha}^k)}[\boldsymbol{p}^k]$. Therefore, the uncertainty u^k and predicted probabilities $\tilde{\boldsymbol{p}}^k$ can be derived in an end-to-end manner.

2.2 Combination Rule

Based on opinions derived from each view, Dempster's combination rule [11] is applied to obtain the overall opinion with uncertainty, which could be converted to the distribution of the final prediction. Specifically, given opinions

$\boldsymbol{D}^1 = \{\{b_c^1\}_{c=1}^C, u^1\}$ and $\boldsymbol{D}^2 = \{\{b_c^2\}_{c=1}^C, u^2\}$, the combined opinion $\boldsymbol{D} = \{\{b_c\}_{c=1}^C, u\} = \boldsymbol{D}^1 \oplus \boldsymbol{D}^2$ is derived by the following rule,

$$b_c = \frac{1}{N}(b_c^1 b_c^2 + b_c^1 u^2 + b_c^2 u^1), u = \frac{1}{N} u^1 u^2, \tag{2}$$

where $N = 1 - \sum_{i \neq j} b_i^1 b_j^2$ is the normalization factor. According to Eq. (2), the combination rule indicates that the combined belief b_c depends more on the opinion which is confident (with small u). In terms of uncertainty, the combined u is small when at least one opinion is confident.

For opinions from K local views and one global view, the combined opinion could be derived by applying the above rule for K times, *i.e.*, $\boldsymbol{D} = \boldsymbol{D}^1 \oplus \cdots \oplus \boldsymbol{D}^K \oplus \boldsymbol{D}^{Global}$.

2.3 Global Representation Modeling

To capture the global representation, we apply a data-efficient transformer as the evidential network for the global view. We follow [13] and improve the performance of the transformer on small datasets by increasing locality inductive bias, *i.e.*, the assumption about relations between adjacent pixels. The standard vision transformer (ViT) [3] without such assumptions typically require more training data than convolutional networks [15]. Therefore, we adopt the SPT and LSA strategy to improve the locality inductive bias.

As shown in Fig. 2, SPT is different from the standard tokenization in that the input image is shifted in four diagonal directions by half the patch size, and the shifted images are concatenated with the original images in the channel dimension to further utilize spatial relations between neighboring pixels. Then, the concatenated images are partitioned into patches and linearly projected as visual tokens in the same way as ViT.

LSA modifies self-attention in ViT by sharpening the distribution of the attention map to pay more attention to important visual tokens. As shown in Fig. 2, diagonal masking and temperature scaling are performed before applying softmax to the attention map. Given the input feature $\boldsymbol{X}$, The LSA module is formalized as,

$$L(\boldsymbol{X}) = \text{softmax}(\mathcal{M}(\boldsymbol{q}\boldsymbol{k}^T)/\tau)\boldsymbol{v}, \tag{3}$$

where $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$ are the query, key, and value vectors obtained by linear projections of $\boldsymbol{X}$. $\mathcal{M}$ is the diagonal masking operator that sets the diagonal elements of $\boldsymbol{q}\boldsymbol{k}^T$ to a small number (*e.g.*,$-\infty$). $\tau \in \mathbb{R}$ is the learnable scaling factor.

2.4 Training Paradigm

Theoretically, the proposed framework could be trained in an end-to-end manner. For each view k, we use the integrated cross-entropy loss as in [17],

$$\mathcal{L}_{ice}^k = \mathbb{E}_{\boldsymbol{p}^k \sim Dir(\boldsymbol{p}^k | \boldsymbol{\alpha}^k)}[\mathcal{L}_{CE}(\boldsymbol{p}^k, \boldsymbol{y}^k)] = \sum_{c=1}^{C} y_c^k(\psi(S^k) - \psi(\alpha_c^k)), \tag{4}$$

where ψ is the digamma function and $\boldsymbol{y}^k$ is the one-hot label. We also apply a regularization term to increase the uncertainty of misclassified samples,

$$\mathcal{L}^k = \mathcal{L}^k_{ice} + \lambda KL[Dir(\boldsymbol{p}^k|\tilde{\boldsymbol{\alpha}}^k)||Dir(\boldsymbol{p}^k|\mathbf{1})], \tag{5}$$

where λ is the balance factor which gradually increases during training and $\tilde{\boldsymbol{\alpha}}^k = \boldsymbol{y}^k + (1 - \boldsymbol{y}^k) \odot \boldsymbol{\alpha}^k$. The overall loss is the summation of losses from all views and the loss for the combined opinion,

$$\mathcal{L}_{Overall} = \mathcal{L}_{Combined} + \mathcal{L}_{Global} + \sum_{k=1}^{K} \mathcal{L}^k, \tag{6}$$

where $\mathcal{L}_{Combined}$ and $\mathcal{L}_{Global}$ are losses of the combined and global opinions, implemented in the same way as $\mathcal{L}^k$. In practice, we pre-train the evidential networks before training with Eq. (6). For local views, we use the model weights pre-trained on ImageNet, and the transformer is pre-trained on the global view images.

3 Experiments

3.1 Dataset

The proposed method was evaluated on Gd-EOB-DTPA-enhanced [25] hepatobiliary phase MRI data, including 342 patients acquired from two scanners, *i.e.*, Siemens 1.5T and Siemens 3.0T. The gold standard was obtained through the pathological analysis of the liver biopsy or liver resection within 3 months before and after MRI scans. Please refer to supplementary materials for more data acquisition details. Among all patients, 88 individuals were identified with fibrosis stage S1, 41 with S2, 40 with S3, and 174 with the most advanced stage S4. Following [25], the slices with the largest liver area in images were selected. The data were then preprocessed with z-score normalization, resampled to a resolution of $1.5 \times 1.5\,\mathrm{mm}^2$, and cropped to 256×256 pixel. For multi-view extraction, the size of the ROI, window, and stride were 160, 96, 32, respectively.

For all experiments, a four-fold cross-validation strategy was employed, and results of two tasks with clinical significance [25] were evaluated, *i.e.*, staging cirrhosis (S4 vs S1-3) and identifying substantial fibrosis (S1 vs S2-4). To keep a balanced number of samples for each class, we over-sampled the S1 data and under-sampled S4 data in the experiments of staging substantial fibrosis.

3.2 Implementation Details

Augmentations such as random rescale, flip, and cutout [2] were applied during training. We chose ResNet34 as the evidential network for local views. For configurations of the transformer, please refer to supplementary materials. The framework was trained using Adam optimizer with an initial learning rate of $1e-4$ for 500 epochs, which was decreased by using the polynomial scheduler.

Table 1. Comparison with multi-view learning methods. Results are evaluated in accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for both tasks.

Method	Cirrhosis(S4 vs S1-3)		Substantial Fibrosis(S1 vs S2-4)	
	ACC	AUC	ACC	AUC
SingleView [8]	77.1 ± 3.17	78.7 ± 4.17	78.2 ± 7.18	75.0 ± 11.5
Concat [5]	80.0 ± 2.49	81.8 ± 3.17	80.5 ± 2.52	83.3 ± 3.65
DCCAE [22]	80.6 ± 3.17	82.7 ± 4.03	83.1 ± 5.30	84.5 ± 4.77
CMC [21]	80.6 ± 1.95	83.5 ± 3.67	83.4 ± 3.22	85.3 ± 4.06
PredSum [18]	78.8 ± 4.16	78.2 ± 4.94	81.1 ± 2.65	84.9 ± 3.21
Attention [9]	76.2 ± 0.98	78.9 ± 3.72	81.4 ± 4.27	84.4 ± 5.34
Ours	$\mathbf{84.4 \pm 1.74}$	$\mathbf{89.0 \pm 0.03}$	$\mathbf{85.5 \pm 1.91}$	$\mathbf{88.4 \pm 1.84}$

The balance factor λ was set to increase linearly from 0 to 1 during training. The transformer network was pre-trained for 200 epochs using the same setting. The framework was implemented using Pytorch and was run on one Nvidia RTX 3090 GPU.

3.3 Results

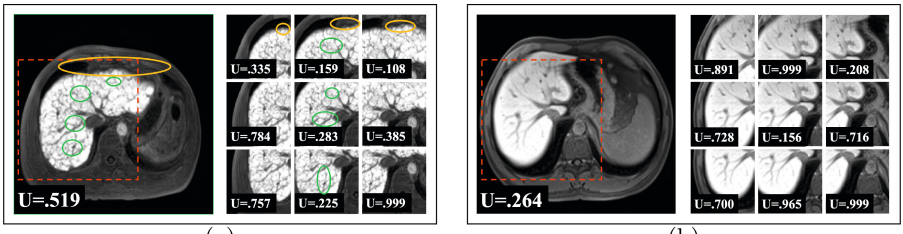
Comparison with Multi-view Learning Methods. To assess the effectiveness of the proposed multi-view learning framework for liver fibrosis staging, we compared it with five multi-view learning methods, including Concat [5], DCCAE [22], CMC [21], PredSum [18], and Attention [9]. Concat is a commonly used method that concatenates multi-view images at the input level. DCCAE and CMC are feature-level strategies. PredSum and Attention are based on decision-level fusion. Additionally, SingleView [8] was adopted as the baseline method for liver fibrosis staging, which uses a single patch as input.

As shown in Table 1, our method outperformed the SingleView method by 10.3% and 12% in AUC on the two tasks, respectively, indicating that the proposed method could exploit more informative features than the method using single view. Our method also set the new state of the art, when compared with other multi-view learning methods. This could be due to the fact that our method was able to capture both the global and local features, and the uncertainty-aware fusion strategy could be more robust than the methods with implicit fusion strategies.

Comparison with Uncertainty-Aware Methods. To demonstrate reliability, we compared the proposed method with other methods. Specifically, these methods estimate uncertainty using Monte-Carlo dropout (Dropout) [20], variational inference (VI) [19], ensemble [4], and softmax entropy [16], respectively. Following [6], we evaluated the expected calibration error (ECE), which measures the gap between model confidence and expected accuracy.

Table 2. Comparison with uncertainty-aware methods. The expected calibration error (ECE) is evaluated in addition to ACC and AUC. Methods with lower ECE are more reliable.

Method	Cirrhosis(S4 vs S1-3)			Substantial Fibrosis(S1 vs S2-4)		
	ACC	AUC	ECE	ACC	AUC	ECE
Softmax	77.1 ± 3.17	78.7 ± 4.17	0.256 ± 0.040	78.2 ± 7.18	83.3 ± 3.65	0.237 ± 0.065
Dropout [20]	77.1 ± 4.89	79.8 ± 4.50	0.183 ± 0.063	80.2 ± 5.00	83.8 ± 6.12	0.171 ± 0.067
VI [19]	77.6 ± 2.20	79.5 ± 4.50	0.229 ± 0.020	81.1 ± 2.08	82.2 ± 6.12	0.191 ± 0.023
Ensemble [4]	78.1 ± 1.91	80.8 ± 3.13	0.181 ± 0.040	79.3 ± 5.11	80.4 ± 3.90	0.193 ± 0.031
Ours	$\mathbf{84.4 \pm 1.74}$	$\mathbf{89.0 \pm 0.03}$	$\mathbf{0.154 \pm 0.028}$	$\mathbf{85.5 \pm 1.91}$	$\mathbf{88.4 \pm 1.84}$	$\mathbf{0.156 \pm 0.019}$



(a) (b)

Fig. 3. Typical samples of stage 4 (a) and stage 1 (b). Visible signs of liver fibrosis are highlighted by circles. Yellow circles indicate the nodular surface contour and green circles denote numerous regenerative nodules. Uncertainties (U) of local and global views estimated by our model were demonstrated. Notably, local views of lower uncertainty contain more signs of fibrosis. Please refer to supplementary materials for more high-resolute images (Color figure online)

Table 2 shows that our method achieved better results in ACC and AUC for both tasks than the other uncertainty-ware multi-view learning methods. It indicates that the uncertainty in our framework could paint a clearer picture of the reliability of each view, and thus the final prediction was more accurate based on the proposed scheme of rule-based combination. Our method also achieved the lowest ECE, indicating that the correspondence between the model confidence and overall results was more accurate.

Interpretability. The proposed framework could explain which view of the input image contains more decisive information for liver fibrosis staging through uncertainties. To evaluate the quality of explanations, we compared the estimated uncertainties with annotations from experienced physicians. Views that contain more signs of fibrosis are supposed to have lower uncertainties. According to Fig. 3, the predicted uncertainties are consistent with annotations in local views of the S4 sample (a). In the S1 sample (b), the uncertainty of global view is low. It is reasonable since there are no visible signs of fibrosis in this stage. The model needs to capture the entire view to discriminate the S1 sample.

Table 3. Ablation study for the roles of local and global views, and effectiveness of the data-efficient transformer.

Method	Cirrhosis(S4 vs S1-3)			Substantial Fibrosis(S1 vs S2-4)		
	ACC	AUC	ECE	ACC	AUC	ECE
Global View solely	76.8 ± 2.81	79.4 ± 4.76	0.192 ± 0.071	82.4 ± 3.45	84.9 ± 5.42	0.192 ± 0.071
Local Views solely	84.1 ± 6.47	88.0 ± 8.39	$\mathbf{0.148 \pm 0.086}$	82.0 ± 6.07	86.9 ± 6.68	0.180 ± 0.060
Both views by CNN	82.9 ± 3.17	87.8 ± 3.09	0.171 ± 0.029	82.0 ± 3.54	87.1 ± 3.47	0.174 ± 0.039
Ours	$\mathbf{84.4 \pm 1.74}$	$\mathbf{89.0 \pm 0.03}$	0.154 ± 0.028	$\mathbf{85.5 \pm 1.91}$	$\mathbf{88.4 \pm 1.84}$	$\mathbf{0.156 \pm 0.019}$

Ablation Study. We performed this ablation study to investigate the roles of local views and global view, as well as to validate the effectiveness of the data-efficient transformer.

Table 3 shows that using the global view solely achieved the worst performance in the staging of cirrhosis. This means that it could be difficult to extract useful features without complementary information from local views. This is consistent with Fig. 3(a), where the uncertainty derived from the global view is high, even if there are many signs of fibrosis. While in Fig. 3(b), the uncertainty of the global view is low, which indicates that it is easier to make decisions from the global view when there is no visible sign of fibrosis. Therefore, we concluded that the global view was more valuable in identifying substantial fibrosis. Compared with the method that only used local views, our method gained more improvement in the substantial fibrosis identification task, which further confirms the aforementioned conclusion. Our method also performed better than the method that applied a convolution neural network (CNN) for the global view. This demonstrates that the proposed data-efficient transformer was more suitable for the modeling of global representation than CNN.

4 Conclusion

In this work, we have proposed a reliable and interpretable multi-view learning framework for liver fibrosis staging. Specifically, uncertainty is estimated through subjective logic to improve reliability, and an explicit fusion strategy is applied which promotes interpretability. Furthermore, we use a data-efficient transformer to model the global representation, which improves the performance.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
2. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

4. Durasov, N., Bagautdinov, T., Baque, P., Fua, P.: Masksembles for uncertainty estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13539–13548 (2021)
5. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
7. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 2551–2566 (2022)
8. Hectors, S., et al.: Fully automated prediction of liver fibrosis using deep learning analysis of gadoxetic acid-enhanced MRI. Eur. Radiol. **31**, 3805–3814 (2021)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
10. Jøsang, A.: Subjective Logic, vol. 4. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42337-1
11. Jøsang, A., Hankin, R.: Interpretation and fusion of hyper opinions in subjective logic. In: 2012 15th International Conference on Information Fusion, pp. 1225–1232. IEEE (2012)
12. Karami, M., Schuurmans, D.: Deep probabilistic canonical correlation analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8055–8063 (2021)
13. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)
14. Liu, X., et al.: Late fusion incomplete multi-view clustering. IEEE Trans. Pattern Anal. Mach. Intell. **41**(10), 2410–2423 (2018)
15. Neyshabur, B.: Towards learning convolutions from scratch. Adv. Neural. Inf. Process. Syst. **33**, 8078–8088 (2020)
16. Pearce, T., Brintrup, A., Zhu, J.: Understanding softmax confidence and uncertainty. arXiv preprint arXiv:2106.04972 (2021)
17. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
18. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
19. Subedar, M., Krishnan, R., Meyer, P.L., Tickoo, O., Huang, J.: Uncertainty-aware audiovisual activity recognition using deep Bayesian variational inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6301–6310 (2019)
20. Tian, J., Cheung, W., Glaser, N., Liu, Y.C., Kira, Z.: Uno: uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 5716–5723. IEEE (2020)
21. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 776–794. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_45

22. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: International Conference on Machine Learning, pp. 1083–1092. PMLR (2015)
23. Yan, X., Hu, S., Mao, Y., Ye, Y., Yu, H.: Deep multi-view learning methods: a review. Neurocomputing **448**, 106–129 (2021)
24. Yasaka, K., Akai, H., Kunimatsu, A., Abe, O., Kiryu, S.: Deep learning for staging liver fibrosis on CT: a pilot study. Eur. Radiol. **28**, 4578–4585 (2018)
25. Yasaka, K., Akai, H., Kunimatsu, A., Abe, O., Kiryu, S.: Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid-enhanced hepatobiliary phase mr images. Radiology **287**(1), 146–155 (2018)
26. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Trans. Emerg. Top. Comput. Intell. **5**(5), 726–742 (2021)