Deep Learning Radiomics of Shear Wave Elastography Significantly Improved Diagnostic Performance for Assessing Liver Fibrosis in Chronic Hepatitis B: A Prospective Multicenter Study

Supplementary method

Mathematical description of the DLRE method. The DLRE method contains convolution, activation, pooling, forward computation, and back propagation.[1] Its details are explained as following.

***Convolution*.** Assuming there are matrix $A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ and matrix $K=\begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}$, the result of convolution between A and K is to let matrix K slides on the matrix A, in other words, matrix K and all of the 2×2 continuous submatrix of A will perform the operation of "corresponding sum of product of elements". Then we could get the result

$$C=\begin{pmatrix} a_{11}*k_{11}+a_{12}*k_{12}+a_{21}*k_{21}+a_{22}*k_{22} & a_{12}*k_{11}+a_{13}*k_{12}+a_{22}*k_{21}+a_{23}*k_{22} \\ a_{21}*k_{11}+a_{22}*k_{12}+a_{31}*k_{21}+a_{32}*k_{22} & a_{22}*k_{11}+a_{23}*k_{12}+a_{32}*k_{21}+a_{33}*k_{22} \end{pmatrix}$$

Therefore, assuming the size of matrix A is $h_a \times w_a$, the size of matrix K is $h_k \times w_k$, where $h_a \geq h_k, \mathrm{w}_a \geq w_k$, then the size of C will be $(h_a-h_k+1)\times(w_a-\mathrm{w}_k+1)$. The matrix K is called filter, while the matrix C is called feature map. The convolution operation between matrix A and matrix K will be presented as $C=conv2(A,\ K)$.

Activation. After the operation of convolution, the result will be activated by an activation function, here we adopted the "ReLU" function $f(x)=\max(0,x)$, when the input is negative, the output of the activation function will be zero, and when the input is positive, the result will be equal to the input. This property helps speed up training process. The activation operation will be presented as $a=f(x)$.

Pooling. Assuming there is a matrix $C=\begin{pmatrix}1&2&3&4\\1&2&3&4\\5&6&7&8\\5&6&7&8\\9&0&1&2\\9&0&1&2\end{pmatrix}$, whose size is 6×4, and a matrix P, whose size is 2×2. The pooling operation will divide the matrix C into six disjoint 2×2 small matrixes, each maximum value of the small matrix will be extracted to form the result matrix $S=\begin{pmatrix}2&4\\6&8\\9&2\end{pmatrix}$. The matrix P is called pooling window, while the matrix S is called pooled map. The pooling operation on matrix C with pooling window P will be presented as $S=pooling(C,\ P)$.

Forward computation. The input of the first convolutional layer is the raw data matrix A with a size of 250×250, after an operation of convolution which contains a number of 16 filters with the size of 3×3, the output of this convolutional layer will be a number of 16 feature maps with the size of 248×248, and then the result will be activated with the activation function of "ReLU".

After the first operation of convolution, there will be a pooling operation. The size of the pooling window is 2×2, and we will adopt the max-pooling strategy here. Then the 248×248 feature map will be transferred to a 124×124 pooled map.

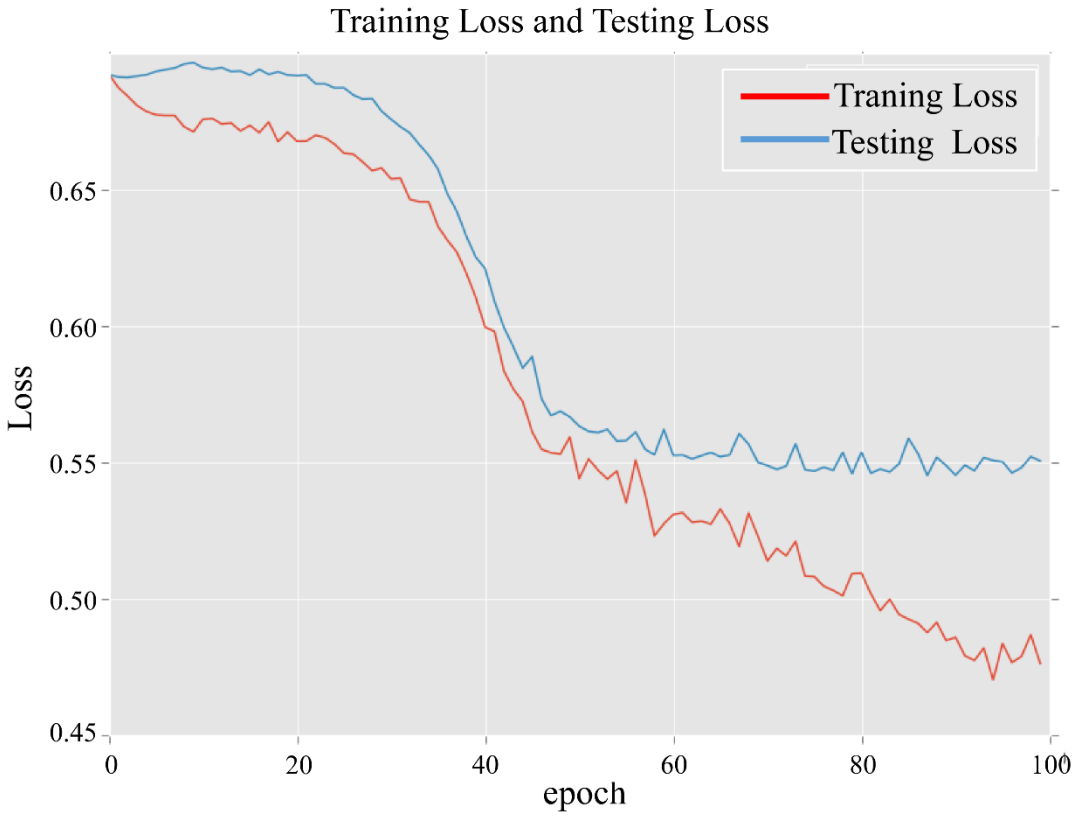
A total of four times of convolution, activation and pooling operations will be executed to complete the computation in turn. When it comes to the last fully connected layer and the output layer, the result will be the possibility.

Back propagation. Assuming the loss function of the whole network is J ,and $J(w,b)=\frac{1}{q}\sum_{j=1}^{q}\frac{1}{2}\left[y^j-p_{(w,b)}(x^j)\right]^2$, where j represents the order of neuron, q means the number of neuron. The most important parameters in the network are the weights w between two neurons and the bias b between two layers. And x is the input of a neuron, $p(x)$ means the actual output of the neuron, while y is the expected output of the neuron.

Therefore, J represents the sum of squared error between the actual output and the expected output. In the end, our task is to make J as small as possible, and to achieve this goal, we need to acquire suitable parameters w and b through learning process from data, here we will use

gradient descent strategy,[1] then $w^l := w^l - \alpha \frac{\partial J}{\partial w^l}$ and $b^l := b^l - \alpha \frac{\partial J}{\partial b^l}$ to fine-tuning w and b ,where l means the order of the layer, α means the learning rate.

The parameters w and b will continue to improve at the end of each iteration of the whole training process. When the loss function tends to decrease and be stable, the CNN model is considered as having completed the training process (Supplementary Fig.1), that means the CNN model is ready to predict new data.



Supplemental Figure 1. Training and testing losses in liver fibrosis classification based on CNN model.

Accessing the diagnostic accuracy of DLRE and 2D-SWE regarding to different ALT, BMI, and inflammation levels. After we performed the liver fibrosis classification using DLRE and 2D-SWE in both validation and training cohorts, all 398 CHB patients were further divided into subgroups regarding to their ALT, BMI, and inflammation levels. Then, the diagnostic performance of each method was compared in different subgroups for each fibrosis stage, so that the factors with significant impact of the diagnostic accuracy of DLRE and 2D-SWE can be revealed, respectively. Patients with ALT > 40 IU/ml and ≤ 40 IU/ml were defined as ALT elevated and normal groups.[2] Patients with BMI > 24 kg/m^2 and ≤ 24 kg/m^2 were defined as BMI high and low level groups.[3] Patients with Metavir score of A0-A2 and A3 were defined as non-severe and severe inflammation groups.

Supplementary results

Supplementary Table 1. Diagnostic performance comparison of DLRE and 2D-SWE for patients with different ALT levels.

		N（P）	AUC	Sensitivity %	Specificity %	PPV %	NPV %	LR+	LR-
Cirrhosis（F4）									
DLRE	E	189 (25.9%)	0.98$^{\#}$ (0.96-0.99)	91.8 (82.4-95.4)	95.0 (86.4-98.8)	86.5 (80.1-91.5)	97.1 (93.4-99.4)	18.4 (18.0-18.7)	0.1 (0.1-0.1)
	N	209 (23.4%)	0.97$^{\#\#}$ (0.95-0.99)	89.8 (80.1-95.3)	96.9 (88.1-99.2)	89.8 (83.4-95.2)	96.9 (91.5-99.2)	28.7 (27.3-29.9)	0.1 (0.1-0.1)
2D-SWE	E	189 (25.9%)	0.86 (0.80-0.90)	81.6 (68.0-91.2)	75.0 (67.0-81.9)	53.3 (41.4-64.9)	92.1 (85.5-96.3)	3.3 (2.8-3.8)	0.2 (0.1-0.5)
	N	209 (23.4%)	0.88 (0.82-0.92)	89.8 (77.8-96.6)	73.8 (66.2-80.4)	51.2 (40.1-62.2)	95.9 (90.8-98.7)	3.4 (3.0-3.9)	0.1 (0.1-0.3)
Advanced fibrosis（≥F3）									
DLRE	E	189 (58.2%)	0.99$^{\#\#\#}$ (0.97-1.0)	92.7 (83.2-95.7)	98.7 (89.3-99.6)	99.0 (93.4-100.0)	90.7 (83.5-96.2)	73.3 (70.5-76.4)	0.1 (0.1-0.1)
	N	209 (54.5%)	0.98$^{\#\#\#}$ (0.96-1.0)	93.9 (85.7-97.8)	97.9 (88.5-99.5)	98.2 (93.5-99.9)	93.0 (88.5-98.5)	44.6 (41.5-47.7)	0.1 (0.1-0.1)
2D-SWE	E	189 (58.2%)	0.82 (0.75-0.87)	79.1 (70.3-86.3)	72.2 (60.9-81.7)	79.8 (71.1-86.9)	71.2 (60.0-80.8)	2.8 (2.4-3.4)	0.3 (0.2-0.5)
	N	209 (54.5%)	0.82 (0.76-0.87)	79.8 (71.3-86.8)	72.6 (62.5-81.3)	77.8 (69.2-84.9)	75.0 (64.9-83.4)	2.9 (2.5-3.4)	0.3 (0.2-0.5)
Significance fibrosis(≥F2)									
DLRE	E	189 (80.4%)	0.92$^{\#\#}$ (0.88-0.95)	88.1 (79.8-95.1)	81.1 (68.8-91.1)	95.0 (90.5-98.7)	62.5 (59.4-95.4)	4.7 (4.3-5.0)	0.2 (0.1-0.2)
	N	209 (86.6%)	0.85 (0.81-0.89)	79.0 (69.1-83.2)	78.6 (69.3-85.2)	96.0 (90.2-99.6)	36.7 (34.1-38.9)	3.7 (3.5-3.9)	0.3 (0.2-0.3)
2D-SWE	E	189 (80.4%)	0.73 (0.66-0.79)	57.2 (49.0-65.2)	86.5 (71.2-95.5)	94.6 (87.7-98.2)	33.0 (23.8-43.3)	4.2 (3.5-5.1)	0.5 (0.2-1.1)
	N	209 (86.6%)	0.80 (0.74-0.86)	84.0 (77.8-89.0)	64.3 (44.1-81.4)	93.8 (88.9-97.0)	38.3 (24.5-53.6)	2.4 (1.8-3.1)	0.3 (0.1-0.5)

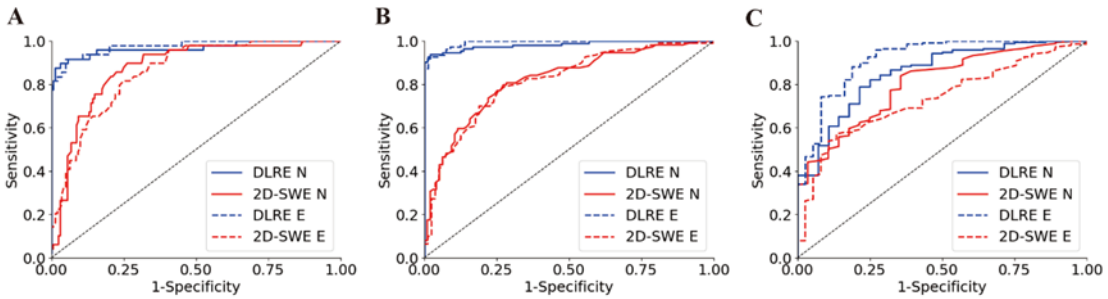
Statistical quantifications were demonstrated with 95% confidence interval, when applicable.

Abbreviations: N, number of patients; P, prevalence; AUC, area under the receiver-operator-characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LR+, positive diagnostic likelihood ratio; LR-, negative diagnostic likelihood ratio; E, ALT elevated group; N, ALT normal group.

AUCs of DLRE and 2D-SWE was statistically compared between ALT elevated and ALT normal groups, respectively, in each fibrosis stage (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

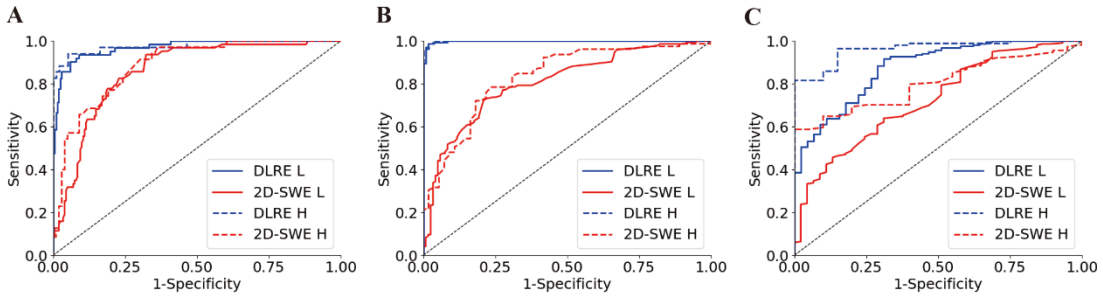
AUCs of DLRE was statistically compared to AUCs of 2D-SWE in the same ALT subgroup in each fibrosis stage ($^{\#}$, $P < 0.05$; $^{\#\#}$, $P < 0.01$; $^{\#\#\#}$, $P < 0.001$).

Comparison of DLRE and 2D-SWE for different ALT levels. Overall speaking, the performance of DLRE was significantly better than 2D-SWE in all fibrosis stage classifications for both ALT elevated and normal groups, except for the ⩾F2 ALT normal group, where DLRE showed higher AUC than 2D-SWE (0.85 vs. 0.80), but no significant difference was found (Supplementary Table 1). Furthermore, neither DLRE nor 2D-SWE showed significant difference in the diagnostic accuracy between ALT elevated and normal groups in all liver fibrosis classifications. ROC curves of the same method (curves with the same color) overlaps each other in Supplementary Fig. 2, which indicates that the ALT level did not affect the performance of DLRE and 2D-SWE for stratifying CHB patients.



Supplementary Figure 2. Comparison of ROC curves between DLRE and 2D-SWE for ALT low and high level groups in three liver fibrosis classifications. (A) F0-F3 versus F4 (F4), (B) F0-F2 versus F3-F4 (≥F3), (C) F0-F1 versus F2-F4 (≥F2). DLRE N, DLRE in ALT normal group; DLRE E, DLRE in ALT elevated group; 2D-SWE N, 2D-SWE in ALT normal group; 2D-SWE E, 2D-SWE in ALT elevated group.

Comparison of DLRE and 2D-SWE for different BMI levels. The performance of DLRE was significantly better than 2D-SWE in all fibrosis stage classifications for both BMI high and low level groups (Supplementary Fig. 3 and Supplementary Table 2). However, no significant difference was found between BMI high and low level groups regardless of using DLRE or 2D-SWE in any fibrosis stage classification. Similar with the ALT level, the BMI level did not affect the performances of DLRE and 2D-SWE neither.



Supplementary Figure 3. Comparison of ROC curves between DLRE and 2D-SWE for BMI low and high level groups in three liver fibrosis classifications. (A) F0-F3 versus F4 (F4), (B) F0-F2 versus F3-F4 (≥F3), (C) F0-F1 versus F2-F4 (≥F2). DLRE L, DLRE in BMI low level group; DLRE H, DLRE in BMI high level group; 2D-SWE L, 2D-SWE in BMI low level group; 2D-SWE H, 2D-SWE in BMI high level group.

Supplementary Table 2. Diagnostic performance comparison of DLRE and 2D-SWE for patients with different BMI levels.

		N（P）	AUC	Sensitivity %	Specificity %	PPV %	NPV %	LR+	LR-
Cirrhosis（F4）									
DLRE	H	134 (26.1%)	0.98$^{\#}$ (0.96-0.99)	94.3 (86.4-97.2)	94.9 (86.5-97.4)	86.8 (80.1-92.3)	97.9 (92.4-99.8)	18.7 (18.5-18.9)	0.1 (0.1-0.1)
	L	264 (23.9%)	0.97$^{\#\#\#}$ (0.95-0.99)	93.7 (85.3-96.8)	91.0 (84.5-94.2)	76.6 (71.5-81.3)	97.9 (92.5-99.8)	10.5 (10.2-10.8)	0.1 (0.1-0.1)
2D-SWE	H	134 (26.1%)	0.89 (0.83- 0.93)	91.4 (76.9-98.2)	69.7 (59.6-78.5)	51.6 (38.5-64.6)	95.8 (88.3-99.1)	3.0 (2.6-3.6)	0.1 (0.0-0.4)
	L	264 (23.9%)	0.86 (0.81-0.90)	93.7 (84.5-98.2)	67.7 (60.7-74.1)	47.6 (38.5-56.7)	97.1 (92.8-99.2)	2.9 (2.6-3.2)	0.1 (0.0-0.2)
Advanced fibrosis（≥F3）									
DLRE	H	134 (59.0%)	0.99$^{\#\#\#}$ (0.97-1.0)	98.7 (92.4-99.8)	98.2 (91.7-99.4)	98.7 (93.5-99.9)	98.2 (93.8-99.9)	54.3 (51.6-57.3)	0.0 (0.0-0.0)
	L	264 (59.4%)	0.99$^{\#\#\#}$ (0.97-1.0)	98.6 (92.3-99.7)	98.3 (91.8-99.5)	98.6 (93.4-99.9)	98.3 (93.1-99.9)	58.7 (55.3-60.4)	0.0 (0.0-0.0)
2D-SWE	H	134 (59.0%)	0.83 (0.76-0.89)	78.5 (67.8-86.9)	76.4 (63.0-86.8)	82.7 (72.1-90.5)	71.2 (57.8-82.3)	3.3 (2.8-4.0)	0.3 (0.1-0.5)
	L	264 (59.4%)	0.81 (0.75-0.85)	72.4 (64.4-79.5)	79.0 (70.6-85.9)	80.8 (72.9-87.2)	70.1 (61.6-77.8)	3.5 (3.0-4.0)	0.4 (0.2-0.5)
Significance fibrosis(≥F2)									
DLRE	H	134 (85.1%)	0.96$^{\#\#}$ (0.93-0.98)	81.6 (72.4-90.3)	100.0 (99.0-100.0)	100.0 (99.0-100.0)	48.8 (46.2-50.2)	-	0.2 (0.2-0.2)
	L	264 (83.0%)	0.88$^{\#\#}$ (0.82-0.93)	91.2 (83.5-97.4)	68.9 (69.2-77.5)	93.5 (88.5-98.1)	63.3 (60.1-66.2)	3.0 (2.8-3.0)	0.1 (0.1-0.1)
2D-SWE	H	134 (85.1%)	0.80 (0.72-0.87)	58.8 (49.2-67.9)	100.0 (83.2-100.0)	100 (94.6-100.0)	29.9 (19.2-42.4)	-	0.4 (0.1-0.5)
	L	264 (83.0%)	0.73 (0.67-0.78)	63.9 (57.2-70.3)	68.9 (53.4-81.8)	90.9 (85.2-95.0)	28.2 (20.0-37.6)	2.1 (1.6-2.6)	0.5 (0.3-0.8)

Statistical quantifications were demonstrated with 95% confidence interval, when applicable.

Abbreviations: N, number of patients; P, prevalence; AUC, area under the receiver-operator-characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LR+, positive diagnostic likelihood ratio; LR-, negative diagnostic likelihood ratio; H, BMI high level group; L, BMI low level group.

AUCs of DLRE and 2D-SWE was statistically compared between BMI high lever and BMI low level groups, respectively, in each fibrosis stage (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

AUCs of DLRE was statistically compared to AUCs of 2D-SWE in the same BMI subgroup in each fibrosis stage ($^{\#}$, $P < 0.05$; $^{\#\#}$, $P < 0.01$; $^{\#\#\#}$, $P < 0.001$).

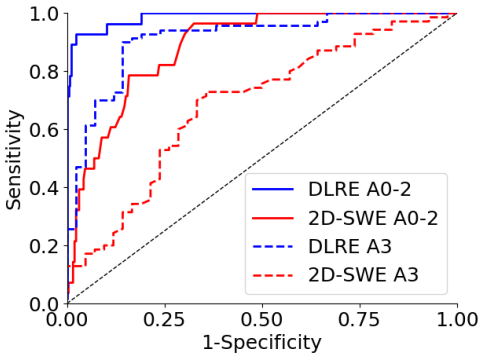
Comparison of DLRE and 2D-SWE for different inflammation levels. For the severe inflammation (A3) group, only two patients were F0-1, and only eight patients were F0-2, which is reasonable, because severe inflammation is normally accompanied with highly developed fibrosis

in liver.[4] Therefore, for classifying ≥F2 and ≥F3, the data distribution is too bias to statistically compare the performances of DLRE or 2D-SWE between severe and non-severe inflammation groups.

Supplementary Table 3. Diagnostic performance comparison of DLRE and 2D-SWE for patients with different inflammation levels.

		N（P）	AUC	Sensitivity %	Specificity %	PPV %	NPV %	LR+	LR-
Cirrhosis（F4）									
DLRE	A0-2	286 (9.8%)	0.99 (0.97-1.0)	92.9 (83.8-97.2)	97.7 (90.2-99.1)	81.2 (76.4-86.2)	99.2 (97.1-100.0)	39.9 (36.4-43.0)	0.1 (0.1-0.1)
	A3	112 (62.5%)	0.91# (0.87-0.94)	90.0 (81.6-95.4)	85.7 (78.6-91.2)	91.3 (84.1-97.5)	83.7 (78.2-88.3)	6.3 (6.0-6.5)	0.1 (0.1-0.1)
2D-SWE	A0-2	286 (9.8%)	0.88*** (0.84- 0.92)	96.4 (81.7-99.9)	67.4 (61.4-73.1)	24.3 (16.7-33.4)	99.4 (96.8-100.0)	3.0 (2.7-3.3)	0.1 (0.0-0.4)
	A3	112 (62.5%)	0.69 (0.59-0.77)	72.9 (60.9-82.8)	64.3 (48.0-78.4)	77.3 (65.2-86.8)	58.7 (43.2-73.0)	2.0 (1.6-2.7)	0.4 (0.2-0.7)

Statistical quantifications were demonstrated with 95% confidence interval, when applicable.
Abbreviations: N, number of patients; P, prevalence; AUC, area under the receiver-operator-characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LR+, positive diagnostic likelihood ratio; LR-, negative diagnostic likelihood ratio; A0-2, non-severe inflammation group; A3, severe inflammation group.
AUCs of DLRE and 2D-SWE was statistically compared between no-severe and severe inflammation groups, respectively, in each fibrosis stage (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$). AUCs of DLRE was statistically compared to AUCs of 2D-SWE in the same inflammation subgroup in each fibrosis stage (#, $P < 0.05$; ##, $P < 0.01$; ###, $P < 0.001$).



Supplementary Figure 4. Comparison of ROC curves between DLRE and 2D-SWE for non-severe and severe inflammation groups in F0-F3 versus F4 (F4). DLRE A0-2, DLRE in non-severe inflammation group; DLRE A3, DLRE in severe inflammation group; 2D-SWE A0-2, 2D-SWE in non-severe inflammation group; 2D-SWE A3, 2D-SWE in severe inflammation group.

For classifying F4, DLRE provided larger AUC than 2D-SWE in both inflammation subgroups (Supplementary Fig. 4), but the difference was only significant ($P < 0.05$) for severe inflammation

group (Supplementary Table 3). Moreover, AUC of 2D-SWE in non-severe inflammation group was significantly larger than that in severe inflammation group (0.88 vs. 0.69, $P < 0.001$), whereas no significant difference was found between AUCs of DLRE in both inflammation subgroups, even though the AUC of DLRE in the non-severe group was larger than that in the severe group (0.99 vs. 0.91). This indicates that for staging F4, inflammation had significant impact on the performance of 2D-SWE, but DLRE may potentially overcome this problem.

Supplementary references

1. Yann L, Léon B, Yoshua B, et al. Gradient-based learning applied to document recognition. *P IEEE* 1998;86:2278–324.
2. Tan YW, Zhou XB, Ye Y, et al. Diagnostic value of FIB-4, aspartate aminotransferase-to-platelet ratio index and liver stiffness measurement in hepatitis B virus-infected patients with persistently normal alanine aminotransferase. *World J Gastroenterol* 2017; 23:5746-54.
3. Chen CM, Zhao WH, Yang XG, et al. Criteria of weight for adults. *The National Health and Family Planning Commission of China* 2013;3.
4. Seki E, Schwabe RF. Hepatic inflammation and fibrosis: functional links and key pathways. *Hepatology* 2015;61:1066-79.