www.nature.com/pr

Pediatric RESEARCH

CLINICAL RESEARCH ARTICLE

Detecting acute bilirubin encephalopathy in neonates based on multimodal MRI with deep learning

Miao Wu[1,2], Xiaoxia Shen[3], Can Lai[4], Yuqing You[4], Zhiyong Zhao[1] and Dan Wu[1]

BACKGROUND: Differentiating acute bilirubin encephalopathy (ABE) from non-ABE in neonates with hyperbilirubinemia (HB) from routine magnetic resonance imaging (MRI) is extremely challenging since both conditions demonstrate similar T1 hyperintensities. To this end, we investigated whether the integration of multimodal MRI from routine clinical scans with deep-learning approaches could improve diagnostic performance.

METHODS: A total of 75 neonates with ABE and 75 neonates with HB (non-ABE) were included in the study. Each patient had three types of multimodal images taken, i.e., a T1-weighted image (T1WI), a T2-weighted image (T2WI), and an apparent diffusion coefficient (ADC) map. The three types of MRI contrasts and their combination were fed into two deep convolutional neural networks (CNNs), i.e., ResNet18 and DenseNet201. The performance of CNNs was compared with a traditional statistical method named logistic regression.

RESULTS: We demonstrated that diagnostic methods with the multimodal data were better than any of the single-modal data. Both CNN models outperformed the logistic regression method. The best performance was achieved by DenseNet201 with the combination of three modalities of T1WI, T2WI, and ADC, with an accuracy of 0.929 ± 0.042 and an area under the curve (AUC) of 0.991 ± 0.007.

CONCLUSIONS: Our study demonstrated that CNN models with multimodal MRI significantly improve the accuracy of diagnosing ABE.

Pediatric Research (2022) 91:1168–1175; https://doi.org/10.1038/s41390-021-01560-0

IMPACT:

- We proposed an efficient strategy of detecting ABE in neonates based on multimodal MRI with deep learning, which achieved an accuracy of 0.929 ± 0.042 and an AUC of 0.991 ± 0.007.
- We demonstrated the advantage of integrating multimodal MRI in detecting ABE in neonates with HB, using deep-learning models.
- Our strategy of diagnosing ABE using deep-learning techniques with multimodal MRI from routine clinical scans is potentially applicable to clinical practice.

INTRODUCTION

Neonatal jaundice is one of the most common conditions encountered by neonatologists and pediatricians, and occurs in ~60–80% of healthy term newborns during the first days of life.[1,2] Although most jaundice is benign, 8–9% of newborns might develop severe hyperbilirubinemia (HB), defined as total serum bilirubin (TSB) level above the 95th percentile for age in hours (high-risk zone) during the first week.[3] Neonates with HB, when unmonitored or untreated, can develop acute bilirubin encephalopathy (ABE), which can lead to varying degrees of brain damage and neurobehavioral disorders.[4,5] If the TSB concentration is not reduced in time to prevent further neurotoxicity in these neonates, chronic irreversible encephalopathy, known as kernicterus, or even death can occur.[6] Nowadays, ABE remains a significant cause of morbidity and mortality throughout the world, which can account for up to 15% of neonatal deaths in low- and middle-income countries.[4] The incidence of ABE may have decreased in developed countries in recent years, but it still occurs at a rate of 0.4–2.7 cases per 100,000 infants,[7,8] with a higher incidence in Asia, the Middle East, and Africa.[9] In Nigeria, 159 cases of ABE were diagnosed in 1040 patients who were admitted for treatment of jaundice (15.3%)[10] and ~4.8% in China.[11] Early identification for newborns at high risk of ABE for timely treatment is crucial to minimize the incidence of kernicterus or to avoid overtreatment.

The TSB concentration is most widely used for evaluating neonatal jaundice, but it is not a direct index of the actual bilirubin level in the brain and not an accurate predictor of ABE.[12]

[1]Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hang Zhou, China; [2]College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, China; [3]Department of Neonatal Intensive Care Unit, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China and [4]Department of Radiology, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China
Correspondence: Dan Wu (danwu.bme@zju.edu.cn)
These authors contributed equally: Miao Wu, Xiaoxia Shen

Received: 16 November 2020 Revised: 9 April 2021 Accepted: 20 April 2021
Published online: 5 June 2021

Moreover, because the collection of blood is required, TSB measurement has a risk of infection and anemia.[13,14] Therefore, a noninvasive method for direct detection of bilirubin-induced changes in the brain is needed for ABE diagnosis. Magnetic resonance imaging (MRI) has been widely used for diagnosing neurological diseases, including bilirubin encephalopathy.[15–17] Many radiological reports found that in the early stages of ABE, T1 hyperintensity of the globus pallidus (GP) bilaterally is a common characteristic in most cases.[18–20] This might be caused by the relatively high resting neuronal activity in the GP, which makes it particularly vulnerable to the intense, subacute oxidative stresses from mitochondrial toxins such as bilirubin.[6] However, this radiological signature does not hold for all cases. To complicate things further, non-ABE neonates with HB conditions often exhibit T1 hyperintensity in the GP as well, making it difficult to differentiate ABE and non-ABE HB patients only by T1-weighted images (T1WIs).

Previous studies revealed that T1WI, T2-weighted imaging (T2WI), and diffusion-weighted imaging (DWI) all contributed to the diagnosis of ABE and may provide complementary information to improve diagnostic accuracy. Wisnowski et al.[15] and Wang et al.[19] reported that MRI of neonates in the first days to weeks following ABE showed an increased T1 signal in the GP, while T2WI of this region was often unremarkable or showed subtle T2 hyperintensity. The increased T2 signal in the GP was often observed in the chronic stage.[21] In a study of 30 ABE patients and 24 control subjects, Cece et al.[22] found that there was a significant correlation between bilirubin values and DWI-based apparent diffusion coefficients (ADCs) ($r = 0.41$, $p < 0.05$). We speculate that the accuracy of diagnosing ABE could be further improved by combining information from multimodal MRI.

While a traditional radiological decision is based on visual inspection, it can be objective, highly empirical, and especially difficult in identifying diseases that do not have a clear radiological standard, such as ABE. To this end, machine-learning-based methods have gained acceptance among radiologists and clinicians.[23] Particularly, deep-learning algorithms, such as convolutional neural networks (CNNs), have been widely used in medical image analyses and achieved great success.[24–27] In this study, we evaluated whether the use of the multimodal MRI and a deep-learning approach can differentiate ABE patients from non-ABE neonates with HB. Two advanced CNN models, namely, ResNet18 and DenseNet201, were tested for classifying ABE and non-ABE patients from a cohort of HB neonates based on T1WI, T2WI, and ADC, and in combination. We also compared CNN-based results with a traditional statistical approach based on normalized T1WI intensity, T2WI intensity, and ADC values in the GP, with logistical regression as the classifier. This study demonstrates that potential and noninvasive diagnostic methods for ABE, which might improve the clinicians' performance and support clinical management, especially for those regions with high ABE incidence.

MATERIALS AND METHODS

Study subjects

The data were collected retrospectively from routine clinical examinations at the Children's Hospital of Zhejiang University School of Medicine between 2016 and 2020. All research protocols were approved by the local Institutional Review Board with a waiver of consent. MRI data were collected from a total of 150 HB neonates who were clinically confirmed with TSB >5 mg/dL,[28] including 75 with ABE and 75 with non-ABE, who underwent MRI during their hospitalization at postmenstrual age (PMA) of 37–41 weeks at the time of the scan. All ABE-positive cases had a bilirubin-induced neurologic dysfunction (BIND) score of ≥1. A BIND score of 1–3, 4–6, and 7–9 represent mild, moderate, and severe ABE, respectively, which is scored based on the muscle tone, cry pattern, and behavioral and mental status with a total of nine points.[29] Non-ABE infants did not exhibit any ABE-related clinical symptoms. The diagnosis was confirmed based on the clinical records by two experienced pediatricians with >8 years of clinical practice (X.S. and C.L.).

MRI acquisition

All images were acquired using a 3.0-T MRI scanner (Achieva, Philips Healthcare, Best, The Netherlands) based on a routine clinical brain MRI protocol with T1WI, T2WI, and DWI. The T1-weighted fast gradient-echo sequence was performed using the following parameters: echo time (TE) of 2.14 ms, repetition time (TR) of 200 ms, flip angle of 80°, a field of view (FOV) of $330 \times 330\ \text{mm}^2$, in-plane resolution of $0.45 \times 0.45\ \text{mm}^2$, and 18 slices with a thickness of 4.5 mm in the axonal direction. T2-weighted turbo spin-echo sequence was performed using the following parameters: TE/TR = 80/3000 ms, FOV of $230 \times 230\ \text{mm}^2$, in-plane resolution of $0.34 \times 0.34\ \text{mm}^2$, and 18 slices with a thickness of 4.5 mm in the axial direction. Diffusion-weighted echo-planar imaging was acquired using the following parameters: TE/TR = 80/2109 ms, FOV of $230 \times 230\ \text{mm}^2$, in-plane resolution of $0.90 \times 0.90\ \text{mm}^2$, 18 slices with a thickness of 4.5 mm in the axial direction, one non-DWI (b_0), and a single DWI at a b value of $800\ \text{s/mm}^2$. All images were visually examined by pediatric radiologists to ensure adequate image quality for further analysis.

Image preprocessing

ADC map was calculated using the following equation: $\text{ADC} = -\log(S_{\text{DWI}}/S_{b0})/b$. In order to combine the three types of images for the CNN models, we first performed image registration between the different image modalities by aligning the T2WI and ADC images to T1WI using the FMRIB's Linear Image Registration Tool (FSL v6.0, FMRIB, Oxford, UK)[30] with a 2D rigid-body transformation since the images were acquired with the same slice center and the same slice thickness. Then, three continuous slices centered around the GP region from the T1WI, T2WI, and ADC images were selected as the inputs to the networks. Thus, 225 slices from 75 ABE patients and 225 slices from 75 non-ABE patients were selected for each MRI modality. We then cropped the images around the brain, resized them uniformly to the size of 224×224 pixels, and normalized the intensities between 0 and 1.

Logistic regression with normalized T1WI, T2WI, and ADC

As the GP is known to be the most vulnerable brain region affected by bilirubin neurotoxicity,[19] we utilized the MR features of the GP for classification using logistic regression.[31] For quantification purposes, we normalized the T1WI, T2WI, and ADC signal intensities of GP to that of the subcortical white matter (WM) as there is no known effect of ABE on the MR properties of the WM. The normalized intensity of the GP was calculated as $\text{GP}_{\text{norm}} = \frac{\overline{\text{GP}}}{\overline{\text{WM}}}$, where $\overline{\text{GP}}$ and $\overline{\text{WM}}$ were averaged intensities in the manually delineated GP and WM regions of interest (ROIs) on a center slice that covered the GP (Fig. 1a).

Logistic regression was performed using a MATLAB toolbox (Mathworks, Natick, MA), with the following input schemes: (1) individual single-modal features of $\text{GP}_{\text{norm, T1}}$, $\text{GP}_{\text{norm, T2}}$, or $\text{GP}_{\text{norm, ADC}}$, (2) combination of any two of these features, and (3) combination of all three features. The maximum Youden index[32,33] was used to determine the optimal cut-off threshold of these features for separating ABE and non-ABE patients.

Deep-learning framework

We applied two CNN models, ResNet18[34] and DenseNet201,[35] which were pre-trained on a public database named ImageNet[36] with three-channel (i.e., RGB images) inputs, with a transfer learning strategy for differentiating ABE and non-ABE patients based on multimodal MRI images. Since each single-modal image

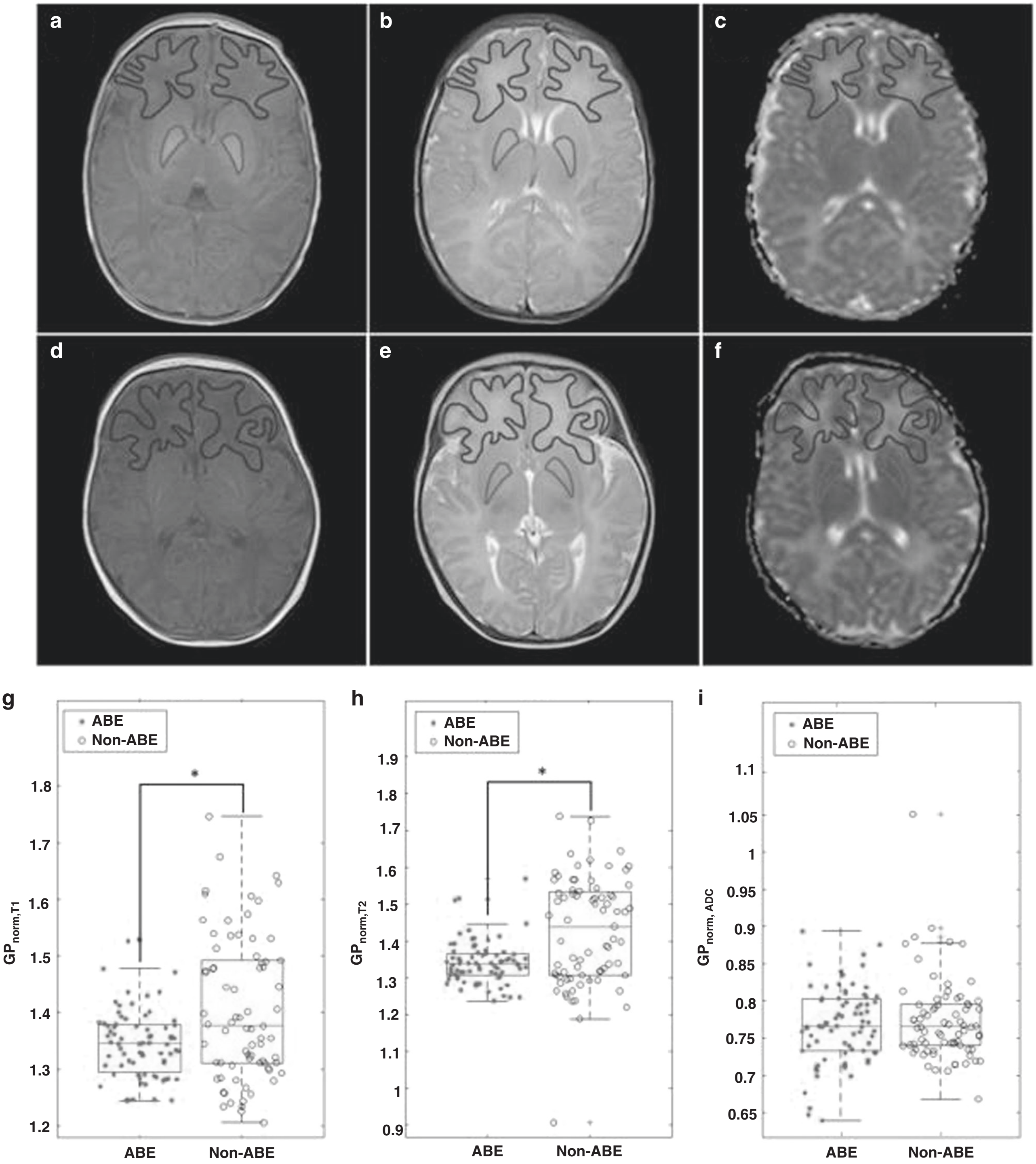

Fig. 1 ROI definition and multimodal MRI measurements in the ABE and non-ABE neonates with HB. a–c T1WI, T2WI, and ADC images of representative ABE neonates. **d–f** T1WI, T2WI, and ADC images of a representative non-ABE neonate. The blue outlines indicate the WM ROI and the red outlines indicate the GP ROI. **g–i** Comparison of the MR features between ABE and non-ABE neonates, in terms of $GP_{norm,T1}$ (**g**), $GP_{norm,T2}$ (**h**), and $GP_{norm,ADC}$ (**i**). $GP_{norm,T1}$ of ABE and non-ABE neonates are 1.345 ± 0.062 and 1.405 ± 0.126. $GP_{norm,T2}$ of ABE and non-ABE neonates are 1.342 ± 0.059 and 1.426 ± 0.146.$GP_{norm,ADC}$ of ABE and non-ABE neonates are 0.767 ± 0.050 and 0.774 ± 0.056.

is a 2D grayscale image, the following strategies were taken to meet the three-channel input scheme: (1) for the single-modal data, we simply duplicated the normalized image to make three identical channels; (2) for the two-modal data, i.e., T1WI + T2WI, T1WI + ADC, or T2WI + ADC, we added an empty image with all zero values as the additional channel; (3) for the three-modal data, T1WI, T2WI, and ADC naturally constituted the three channels. The resulting 225 images were divided into 80% and 20% for the training and testing sets. Data augmentation was applied to the training dataset, which included image rotation with a random angle in the range of −30° to 30°, image zooming by a random scale within the range of 0.9–1.1, and image horizontal and vertical translation with random distance in the range of −30 to 30 pixels. A 5-fold cross-validation was applied to assess the models' generalization performance with metrics of classification accuracy, the area under the ROC curve (AUC), sensitivity, specificity, precision, and $F1$ score. Equations (1)–(5) showed the definition of these performance metrics, where TP, FP, TN, and FN represent the numbers of true-positive, false-positive, true-negative, and false-negative cases, respectively. The performance metrics were presented as mean ± standard deviation from the 5-fold cross-validation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$F1\ \text{score} = \frac{2 \times \text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}} \tag{5}$$

The models were obtained from the Deep Learning Toolbox in MATLAB 2019a. The hyperparameters of the CNN were heuristically set as follows: the learning rate was initialized to 0.0003, maximum epoch number was limited to 6, stochastic gradient descent momentum-based solver was used with a minibatch size of ten images for training. The experiment was implemented in MATLAB 2019a.

Statistical analyses

Differences in the sex distribution among groups were evaluated using the χ^2 test, while other clinical features, which all passed the Kolmogorov–Smirnov normality test, were evaluated by a two-tailed t test with unequal variance. The differences of $\text{GP}_{\text{norm, T1}}$, $\text{GP}_{\text{norm, T1}}$, and GP_{ADC} measurements between ABE and non-ABE groups were also tested using t tests. A p value < 0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics 21 (https://www.ibm.com/products/spss-statistics).

RESULTS

The demographic and clinical characteristics of the patients in our study are listed in Table 1, including sex, age, weight, gestational age (GA), PMA at scan, TSB, and albumin. Significant differences in age ($p = 0.004$, ~2 days off) and TSB ($p = 0.000$, ~5.03 mg/dL difference) were found between the ABE and non-ABE groups, while other features were comparable between the two groups ($p > 0.05$).

Table 1. The demographic and clinical characteristics of the patients.

Clinical features	ABE ($n = 75$)	Non-ABE ($n = 75$)	p Value
Sex (male)	47 (62.67%)	50 (66.67%)	0.61
Age (days)	10.08 ± 3.64	12.37 ± 5.62	0.00
Weight (kg)	3.26 ± 0.43	3.38 ± 0.43	0.08
Gestational age at birth (weeks)	38.54 ± 1.49	38.58 ± 1.26	0.85
PMA at scan (weeks)	39.98 ± 1.52	40.35 ± 1.57	0.14
TSB (mg/dL)	23.45 ± 8.04	18.42 ± 3.71	0.00
Albumin (g/L)	38.47 ± 2.99	37.93 ± 2.94	0.27

Figure 1 shows T1WI, T2WI, and ADC images of representative ABE (Fig. 1a–c) and non-ABE (Fig. 1d–f) patients with HB. Blue and red outlines indicate the manually traced WM and GP ROIs for calculations of normalized intensities. The differences in $\text{GP}_{\text{norm, T1}}$, $\text{GP}_{\text{norm, T2}}$, and $\text{GP}_{\text{norm,ADC}}$ are presented in Fig. 1g–i. T tests indicated a significant difference in $\text{GP}_{\text{norm,T1}}$ (1.345 ± 0.062 versus 1.405 ± 0.126, $p = 0.000$) and $\text{GP}_{\text{norm,T2}}$ (1.342 ± 0.059 versus 1.426 ± 0.146, $p = 0.000$) values, but no significant difference was found in $\text{GP}_{\text{norm,ADC}}$ (0.767 ± 0.050 versus 0.774 ± 0.056, $p >$ 0.05). Considerable overlaps were observed between the two groups for all three measurements, indicating the difficulty of classification based on any of the single modalities.

The performance of logistic regression on identifying ABE and non-ABE infants is shown in Table 2 and Supplementary Fig. S1, with ROC curves shown in Fig. 2a. The combined feature of $\text{GP}_{\text{norm,T2}}$ and $\text{GP}_{\text{norm,ADC}}$ achieved the highest AUC of 0.681, while the highest accuracy of 0.833 was obtained using the combination of $\text{GP}_{\text{norm,T1}}$ and GP_{ADC}. The combination of all three modalities provided the second-highest AUC of 0.677 and the second best accuracy of 0.800. Accuracies of 0.720, 0.773, and 0.600 were found for $\text{GP}_{\text{norm,T1}}$, $\text{GP}_{\text{norm,T2}}$, and $\text{GP}_{\text{norm,ADC}}$, respectively, with optimal cut-off values of 1.439, 1.435, and 0.714, respectively. These results indicated that although ADC alone did not have a good predictive value, but in combination with T1WI or T2WI the prediction accuracy improved. From Supplementary Fig. S1(a) we can see that the sensitivities for the logistic regression are almost 100% for all modalities except ADC, indicating that the logistic regression method has a good capability for predicting true-positive samples (ABE) with no false-negative samples are detected in our experiments. However, since there was no statistically significant difference between the $\text{GP}_{\text{norm,ADC}}$ of ABE and non-ABE (shown in Fig. 1i), the optimal cut-off value of $\text{GP}_{\text{norm,ADC}}$ (0.714) can hardly separate ABE and non-ABE accurately with a lot of false-negative samples and no false-positive samples were detected in the result, which directly leads to the poor sensitivity of 20% and the high precision of 100% and the specificity of 100%.

We then evaluated the performances of the ResNet18 and DenseNet201 CNN models based on the single- or multimodal images through a 5-fold cross-validation. Comparing the results using different classifiers (Table 2), the DenseNet201 achieved the best overall performance, followed by ResNet18, which both outperformed logistic regression. T tests indicated that the classification accuracy of DenseNet201 was significantly higher than ResNet18 when using combined images of T1WI + ADC ($p =$ 0.048, <0.05) and T2WI + ADC ($p = 0.003$, <0.05), but their performance was similar in terms of T1WI, T2WI, ADC, T1WI + T2WI, and T1WI + T2WI + ADC.

Figure 2 and Supplementary Fig. S1 show that with the increased number of MR modalities fused in the input image, the AUC gradually improved for both CNN models, and the combination of all three modalities gave the best performance in almost all of the evaluation metrics with high accuracy of 0.929 and an AUC of 0.991 for DenseNet201. Among the single-modal MRI data, T1WI had the best classification performance, followed by T2WI and then ADC for DenseNet201, which is similar to the findings from logistic regression. Interestingly, the sensitivities of DenseNet201 for single-modal MRI from high to low were T2WI, ADC, and T1WI, while their specificities showed approximately an opposite order, which again suggested their complementary roles in the classification task. Among the two-modal data, T1WI + T2WI achieved an accuracy of 0.918 with an AUC of 0.991, which was considerably higher than that for T1WI + ADC and T2WI + ADC.

DISCUSSION

This study evaluated whether multimodal MRI could improve the diagnostic performance compared with using a single modality.

Table 2. The performance metrics of logistic regression, ResNet18 and DenseNet201, on classifying ABE and non-ABE based on single- and multimodal data.

Classifier	MRI modality	Sensitivity	Specificity	Precsion	*F*1 score	Accuracy	AUC
Logistic regression	T1WI	**1.0000**	0.4400	0.6410	0.7813	0.7200	0.6158
	T2WI	**1.0000**	0.5467	0.6881	0.8152	0.7733	0.6612
	ADC	0.2000	**1.0000**	**1.0000**	0.3333	0.6000	0.5012
	T1WI + T2WI	**1.0000**	0.5600	0.6944	0.8197	0.7800	0.6375
	T1WI + ADC	**1.0000**	0.6667	0.7500	**0.8571**	**0.8333**	0.6558
	T2WI + ADC	**1.0000**	0.5867	0.7075	0.8287	0.7933	**0.6807**
	T1WI + T2WI + ADC	**1.0000**	0.6000	0.7143	0.8333	0.8000	0.6766
ResNet18 (5-fold cross-validation)	T1WI	0.5689 ± 0.4198	0.8444 ± 0.1474	0.8425 ± 0.1076	0.5717 ± 0.3493	0.7067 ± 0.1488	0.9084 ± 0.0188
	T2WI	0.8622 ± 0.1373	0.6222 ± 0.1602	0.7052 ± 0.0591	0.7676 ± 0.0429	0.7422 ± 0.0404	0.8217 ± 0.0474
	ADC	0.5644 ± 0.2697	0.8178 ± 0.1309	0.8030 ± 0.1234	0.6090 ± 0.2195	0.6911 ± 0.0829	0.8257 ± 0.0496
	T1WI + T2WI	0.6889 ± 0.1556	**0.9733 ± 0.0186**	0.9653 ± 0.0251	0.7945 ± 0.1120	0.8311 ± 0.0730	0.9702 ± 0.0182
	T1WI + ADC	**0.9244 ± 0.0826**	0.7156 ± 0.1346	0.7747 ± 0.0723	0.8376 ± 0.0199	0.8200 ± 0.0328	0.9320 ± 0.0180
	T2WI + ADC	0.8667 ± 0.1449	0.8578 ± 0.1140	0.8725 ± 0.0800	**0.8593 ± 0.0598**	0.8622 ± 0.0501	0.9680 ± 0.0261
	T1WI + T2WI + ADC	0.7733 ± 0.1567	0.9689 ± 0.0461	**0.9683 ± 0.0419**	0.8504 ± 0.0830	**0.8711 ± 0.0617**	**0.9851 ± 0.0049**
DenseNet201 (5-fold cross-validation)	T1WI	0.5689 ± 0.3033	0.9200 ± 0.0678	0.9009 ± 0.0639	0.6478 ± 0.2386	0.7444 ± 0.1204	0.9209 ± 0.0309
	T2WI	0.7422 ± 0.1263	0.7378 ± 0.0636	0.7407 ± 0.0225	0.7362 ± 0.0631	0.7400 ± 0.0382	0.8261 ± 0.0145
	ADC	0.6089 ± 0.2040	0.7867 ± 0.1182	0.7538 ± 0.0561	0.6524 ± 0.1246	0.6978 ± 0.0600	0.8190 ± 0.0331
	T1WI + T2WI	**0.9022 ± 0.0976**	0.9333 ± 0.0544	0.9356 ± 0.0504	0.9147 ± 0.0453	0.9178 ± 0.0398	0.9907 ± 0.0057
	T1WI + ADC	0.8622 ± 0.0617	0.9067 ± 0.0575	0.9070 ± 0.0484	0.8814 ± 0.0138	0.8844 ± 0.0099	0.9639 ± 0.0233
	T2WI + ADC	0.8089 ± 0.1544	0.9556 ± 0.0272	0.9479 ± 0.0266	0.8664 ± 0.0965	0.8822 ± 0.0744	0.9663 ± 0.0228
	T1WI + T2WI + ADC	0.8756 ± 0.0924	**0.9822 ± 0.0290**	**0.9819 ± 0.0296**	**0.9229 ± 0.0497**	**0.9289 ± 0.0420**	**0.9912 ± 0.0066**

Bold value indicate the maximum value of performance metrics for each classifier.

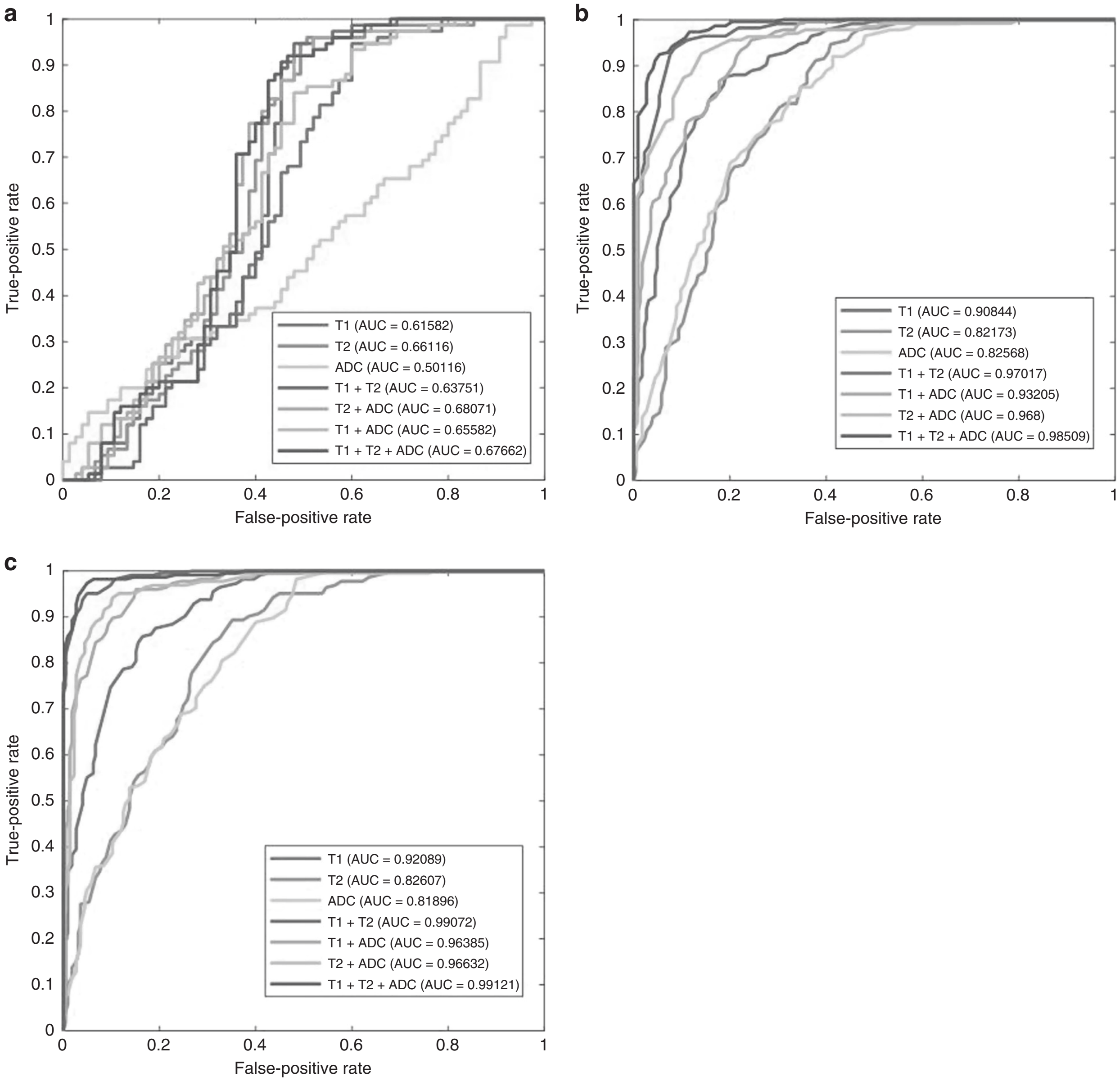

Fig. 2 ROC curves of the single- and multimodal MRI features for differentiating ABE and non-ABE. Single-modal data of T1WI, T2WI, ADC, and multimodal data of T1WI + T2WI, T1WI + ADC, T2WI + ADC, and T1WI + T2WI + ADC were tested, respectively. **a** ROC curves based on logistic regression classifiers using the semi-quantitative MRI measurements in the GP. **b** ROC curves based on ResNet18 using the single- and multimodal images. **c** ROC curves based on DenseNet201 using the single- and multimodal images.

We also demonstrated the advantage of deep-learning networks compared with the traditional statistical methods based on the multimodal MRI markers. This was done in the framework of separating ABE and non-ABE neonates who both had HB, which is known to be particularly challenging with current clinical and radiological examinations. Our results indicated that multimodal MRI plays an important role in the clinical management of ABE, and should be incorporated into the clinical routine whenever MRI is available.

At present, ABE remains one of the most significant causes of neonatal mortality and lifelong disability. The commonly used physiological parameters, such as TSB, albumin need, unconjugated or free bilirubin levels, and bilirubin bound to albumin, do not have sufficient diagnostic power as they do not directly reflect bilirubin toxicity in the brain.[37,38] The clinical manifestations and neurological symptoms could also be absent, subtle, or nonspecific in the early phases of ABE.[39] When an overt clinical sign appears, the bilirubin-induced neurological injuries may have already been present and become irreversible. Although MRI has been increasingly used to investigate the neuropathology induced by ABE in the clinical setting, its diagnostic accuracy is limited and research in this field is relatively scarce. A study by Mao et al.[20] reported that 20 of 36 neonates with HB have symmetric hyperintense GP on T1WI; and among these 20 HB neonates, 15 had ABE. Coskun et al.[18] reported that 8 of 13 (61.54%) ABE patients demonstrated bilateral, symmetric increased signal intensity in the GP on T1WI and these lesions were not apparent on T2WI. Clearly, visual inspection is not sufficient for diagnosing ABE given the subtle and nonspecific differences from single-modality MRI.

Our results demonstrate that for all three methods the performance metrics gradually improved when the input data

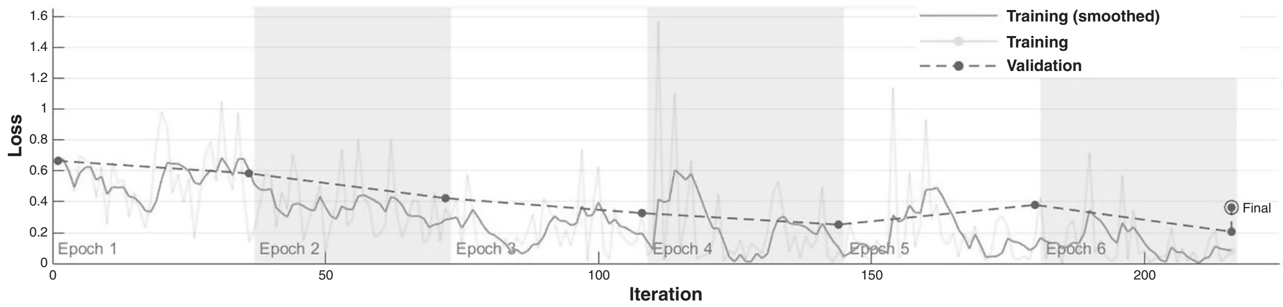

Fig. 3 Traning and validation loss. Training and validation loss on training the DenseNet201 with images fused by T1WI, T2WI, and ADC.

combined more modalities. The combination of three-modal images (T1WI + T2WI + ADC) achieved the best performance with a mean accuracy of 0.929, 0.871, and 0.800 for DenseNet201, ResNet18, and logistic regression, respectively. This was considerably higher than when using a single modality (accuracies < 0.750). This may be because images from the three modalities provide different presentations of the ABE pathology that support each other,[15,19,22] as T1WI and T2WI reflect the chemical components of the tissue, while ADC is associated with tissue microstructures that dictate the water diffusion. The different modality images were complementary in classification sensitivity and specificity, e.g., T2WI had the highest sensitivity and T1WI gave the highest specificity.

Our results also show that the CNN models outperformed the statistical approach of logistical regression, as expected. One reason for this outcome is that the features used for the classification work are much different for CNN and logistic regression. CNN uses the whole image as the input data so that all the information is captured, while the logistic regressor that uses manually defined MRI features of GP_{norm} and other image features that are potentially useful for the classification are ignored. Among the two CNN models, DenseNet201 achieved higher classification accuracy than ResNet18, owing to the more learnable layers, which likely benefited the feature extraction and classification efficacy. However, using a model with a complex architecture has a risk of overfitting, especially for a limited training sample set with little heterogeneity. As shown in Fig. 3, we found that the training loss decreased, whereas the validation loss increased after 160 iterations for DenseNet201, indicating overfitting.

Another limitation of the study is that our data were all collected from one hospital; therefore, the generalizability of the models is unknown. Future multicenter studies are necessary to validate the models' generalizability. Also, in addition to the use of conventional MRI contrasts of T1WI, T2WI, and DWI, the integration of more advanced MRI techniques, such as susceptibility-weighted MRI, perfusion MRI, and spectroscopy, as well as the clinical information, is likely to further enhance the diagnostic power. Moreover, it would be ideal to test the prediction ability to kernicterus, the chronic phase of ABE, which is critical to the clinical management of newborns.

CONCLUSION

Here, we demonstrate the potential of multimodal MRI with machine-learning approaches in identifying ABE in HB patients. The results indicate that the multimodal MRI outperforms the single modalities for all types of classifiers, and the CNN models outperform the logistic regression with predefined features in the GP. The best performance was achieved by DenseNet201 with the fusion images combined by T1WI, T2WI, and ADC, which achieved an accuracy of 0.929 with an AUC of 0.991. The strategy of the multimodal MRI-based diagnosis of ABE is potentially applicable to clinical practice to facilitate clinical management.

DATA AVAILABILITY

The de-identified MRI data are available from the corresponding author upon request.

ACKNOWLEDGEMENTS

This work was funded by the Ministry of Science and Technology of the People's Republic of China (2018YFE0114600), Natural Science Foundation of China (61801424, 81971606, 61801421, and 81971605), and Fundamental Research Funds for the Central Universities of China (2019QNA5024).

AUTHOR CONTRIBUTIONS

M.W. proposed the methods, performed the data analysis, and led the manuscript writing. D.W. designed and supervised the project, and revised the manuscript. X.S. and Y.Y. collected the MRI data and clinical information. C.L. participated in the data analysis and data collection. Z.Z. participated in the data analysis.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41390-021-01560-0.

Competing interests: The authors declare no competing interests.

Consent for publication: Written informed consent for publication of their clinical details and clinical images were obtained from the patients' parents.

Ethics approval and consent to participate: Ethical approval was obtained from the Research Ethics Committee of the School of Medicine, Zhejiang University. Both written consent and verbal consent were allowed according to the Ethics committee. The written informed consent was obtained from the parents or legal guardians.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Keren, R. et al. Visual assessment of jaundice in term and late preterm infants. *Arch. Dis. Child. Fetal Neonatal Ed.* **94**, F317–F322 (2009).
2. Bhutani, V. K. et al. Predischarge screening for severe neonatal hyperbilirubinemia identifies infants who need phototherapy. *J. Pediatr.* **162**, 477–483 (2013).
3. Smitherman, H., Stark, A. R. & Bhutani, V. K. Early recognition of neonatal hyperbilirubinemia and its emergent management. *Semin. Fetal Neonatal Med.* **11**, 214–224 (2006).
4. Usman, F. et al. Acute bilirubin encephalopathy and its progression to kernicterus: current perspectives. *Res. Rep. Neonatol.* **8**, 33–44 (2018).
5. Amin, S. B., Bhutani, V. K. & Watchko, J. F. Apnea in acute bilirubin encephalopathy. *Semin. Perinatol.* **38**, 407–411 (2014).
6. Johnston, M. V. & Hoon, A. H. Possible mechanisms in infants for selective basal ganglia damage from asphyxia, kernicterus, or mitochondrial encephalopathies. *J. Child Neurol.* **15**, 588–591 (2000).
7. McGillivray, A. & Evans, N. Severe neonatal jaundice: Is it a rare event in Australia?. *J. Paediatr. Child Health* **48**, 801–807 (2012).
8. Christensen, R. D. et al. Acute neonatal bilirubin encephalopathy in the State of Utah 2009-2018. *Blood Cells Mol. Dis.* **72**, 10–13 (2018).

9. Bhutani, V. K. et al. Neonatal hyperbilirubinemia and Rhesus disease of the newborn: incidence and impairment estimates for 2010 at regional and global levels. *Pediatr. Res.* **74**, 86–100 (2013).
10. Diala, U. M. et al. Patterns of acute bilirubin encephalopathy in Nigeria: a multicenter pre-intervention study. *J. Perinatol.* **38**, 873–880 (2018).
11. Subspecialty Group of Neonatology, et al. Clinical characteristics of bilirubin encephalopathy in Chinese newborn infants-a national multicenter survey. *Zhonghua Er Ke Za Zhi* **50**, 331–335 (2012).
12. Maisels, M. J. Managing the jaundiced newborn: a persistent challenge. *Can. Med. Assoc. J.* **187**, 335–343 (2015).
13. El-Beshbishi, S. N. et al. Hyperbilirubinemia and transcutaneous bilirubinometry. *Clin. Chem.* **55**, 1280–1287 (2009).
14. Pace, E. J., Brown, C. M. & DeGeorge, K. C. Neonatal hyperbilirubinemia: an evidence-based approach. *J. Fam. Pract.* **68**, E4–E11 (2019).
15. Wisnowski, J. L. et al. Magnetic resonance imaging of bilirubin encephalopathy: current limitations and future promise. *Semin. Perinatol.* **38**, 422–428 (2014).
16. Allen, L. M. et al. Sequence-specific MR imaging findings that are useful in dating ischemic stroke. *Radiographics* **32**, 1285–1297 (2012).
17. Cinar, A. & Yildirim, M. Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture. *Med. Hypotheses* **139** (2020).
18. Coskun, A. et al. Hyperintense globus pallidus on T1-weighted MR imaging in acute kernicterus: is it common or rare?. *Eur. Radiol.* **15**, 1263–1267 (2004).
19. Wang, X. et al. Studying neonatal bilirubin encephalopathy with conventional MRI, MRS, and DWI. *Neuroradiology* **50**, 885–893 (2008).
20. Mao, J. et al. Changes of globus pallidus in the newborn infants with severe hyperbilirubinemia. *Zhonghua Er Ke Za Zhi* **45**, 24–29 (2007).
21. Wu, W. et al. Usefulness of H-1-MRS in differentiating bilirubin encephalopathy from severe hyperbilirubinemia in neonates. *J. Magn. Reson. Imaging* **38**, 634–640 (2013).
22. Cece, H. et al. Diffusion-weighted imaging of patients with neonatal bilirubin encephalopathy. *Jpn J. Radiol.* **31**, 179–185 (2013).
23. Suzuki, K. Overview of deep learning in medical imaging. *Radiol. Phys. Technol.* **10**, 257–273 (2017).
24. Togacar, M., Comert, Z. & Ergen, B. Classification of brain MRI using hyper column technique with convolutional neural network and feature selection method. *Expert Syst. Appl.* **149** (2020).
25. Zong, W. et al. A deep dive into understanding tumor foci classification using multiparametric MRI based on convolutional neural network. *Med. Phys.* **47**, 4077–4086 (2020).
26. Baldeon-Calisto, M. & Lai-Yuen, S. K. AdaResU-Net: multiobjective adaptive convolutional neural network for medical image segmentation. *Neurocomputing* **392**, 325–340 (2020).
27. Bousabarah, K. et al. Automated detection and delineation of hepatocellular carcinoma on multiphasic contrast-enhanced MRI using deep learning. *Abdom. Radiol.* **46**, 216–225 (2021).
28. Porter, M. L. & Dennis, B. L. Hyperbilirubinemia in the term newborn. *Am. Fam. Physician* **65**, 599–606 (2002).
29. Johnson, L., Brown, A. K. & Bhutani, V. K. BIND-a clinical score for bilirubin induced neurologic dysfunction in newborns. *Pediatrics* **104**, 746 (1999).
30. Jenkinson, M. et al. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17**, 825–841 (2002).
31. LaValley, M. P. Logistic regression. *Circulation* **117**, 2395–2399 (2008).
32. Fluss, R., Faraggi, D. & Reiser, B. Estimation of the Youden Index and its associated cutoff point. *Biometrical J.* **47**, 458–472 (2005).
33. Ruopp, M. D. et al. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical J.* **50**, 419–430 (2008).
34. He, K. et al. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778 (IEEE, 2016).
35. Huang, G. et al. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2261–2269 (2017).
36. Deng. J. et al. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255 (IEEE, 2009).
37. Wennberg, R. Unbound bilirubin: a better predictor of kernicterus?. *Clin. Chem.* **54**, 207–208 (2008).
38. Iskander, I. et al. Serum bilirubin and bilirubin/albumin ratio as predictors of bilirubin encephalopathy. *Pediatrics* **134**, e1330–e1339 (2014).
39. Bhutani, V. K. & Johnson-Hamerman, L. The clinical syndrome of bilirubin-induced neurologic dysfunction. *Semin. Fetal Neonatal Med.* **20**, 6–13 (2015).