


RESEARCH ARTICLE

Open Access



# Detecting neonatal acute bilirubin encephalopathy based on T1-weighted MRI images and learning-based approaches

Miao Wu<sup>1,4\*</sup> , Xiaoxia Shen<sup>2</sup>, Can Lai<sup>3</sup>, Weihao Zheng<sup>1</sup>, Yingqun Li<sup>3</sup>, Zhongli Shangguan<sup>3</sup>, Chuanbo Yan<sup>4</sup>, Tingting Liu<sup>1</sup> and Dan Wu<sup>1</sup>

## Abstract

**Background:** Neonatal hyperbilirubinemia is a common clinical condition that requires medical attention in newborns, which may develop into acute bilirubin encephalopathy with a significant risk of long-term neurological deficits. The current clinical challenge lies in the separation of acute bilirubin encephalopathy and non-acute bilirubin encephalopathy neonates both with hyperbilirubinemia condition since both of them demonstrated similar T1 hyperintensity and lead to difficulties in clinical diagnosis based on the conventional radiological reading. This study aims to investigate the utility of T1-weighted MRI images for differentiating acute bilirubin encephalopathy and non-acute bilirubin encephalopathy neonates with hyperbilirubinemia.

**Methods:** 3 diagnostic approaches, including a visual inspection, a semi-quantitative method based on normalized the T1-weighted intensities of the globus pallidus and subthalamic nuclei, and a deep learning method with ResNet18 framework were applied to classify 47 acute bilirubin encephalopathy neonates and 32 non-acute bilirubin encephalopathy neonates with hyperbilirubinemia based on T1-weighted images. Chi-squared test and t-test were used to test the significant difference of clinical features between the 2 groups.

**Results:** The visual inspection got a poor diagnostic accuracy of  $53.58 \pm 5.71\%$  indicating the difficulty of the challenge in real clinical practice. However, the semi-quantitative approach and ResNet18 achieved a classification accuracy of  $62.11 \pm 8.03\%$  and  $72.15\%$ , respectively, which outperformed visual inspection significantly.

**Conclusion:** Our study indicates that it is not sufficient to only use T1-weighted MRI images to detect neonates with acute bilirubin encephalopathy. Other more MRI multimodal images combined with T1-weighted MRI images are expected to use to improve the accuracy in future work. However, this study demonstrates that the semi-quantitative measurement based on T1-weighted MRI images is a simple and compromised way to discriminate acute bilirubin encephalopathy and non-acute bilirubin encephalopathy neonates with hyperbilirubinemia, which may be helpful in improving the current manual diagnosis.

**Keywords:** Acute bilirubin encephalopathy, Hyperbilirubinemia, Normalized T1-weighted intensities, Deep convolutional neural networks, ResNet18, Classification, Diagnosis

## Background

Neonatal jaundice, which develops in about 60% of term and 80% of preterm babies during their first week of life, is one of the most common conditions that require medical attention in newborns [1, 2]. It is mainly caused by the

\*Correspondence: wumiao@zju.edu.cn

<sup>1</sup> Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou 310027, China

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

inability of the newborn's immature liver to process the excessive bilirubin that was produced by the accelerated breakdown of red blood cells at this age [3, 4]. Although most jaundice is benign, 8–9% of infants might develop severe hyperbilirubinemia (HB), with approximately 4% affected after 72 h [5]. In even more serious, due to the lack of appropriate diagnoses and delayed treatments, severe HB patients may develop acute bilirubin encephalopathy (ABE) in response to the entry of bilirubin toxicity into the basal ganglia and various brain nuclei. And a long-term outcome of ABE could be kernicterus which is a permanent disabling neurologic condition classically characterized by the movement disorders of dystonia and choreoathetosis, hearing loss caused by auditory neuropathy spectrum disorders, and oculomotor pareses [6]. A recent study indicated that ABE accounts for 3.4% of neonatal admissions with 21.4% of those infants in severe conditions and at least 15% of them died [7]. Therefore, early diagnosis of neonates with a high risk of ABE and timely taking effective intervening measures are very important for pediatricians to minimize the mortality or prevent them from kernicterus.

The total serum bilirubin (TSB) measurement is a traditional and most widely used method for screening and diagnosing HB in neonates, but it needs a blood draw which is invasive and carries a risk of infection and anemia [8]. Meanwhile, as it is not a direct measurement of actual bilirubin level in the brain, TSB alone could not accurately predict the occurrence of ABE [9, 10]. Magnetic resonance imaging (MRI), as a non-radiation and non-invasive imaging technique, is widely used in the ABE diagnosis in newborns [11]. Many MRI studies found that newborns following ABE in the first days to weeks showed an increased signal on T1-weighted images (T1WI) at globus pallidus (GP) and subthalamic nucleus (STN) in most cases, while T2-weighted imaging of these regions was often unremarkable or shows subtle T2-hyperintensity [12–16]. Although the T1 hyperintensity in GP provides an efficient marker for diagnosing ABE neonates, it remains challenging to further separate ABE and non-ABE HB patients since both groups may demonstrate elevated T1 signals. On the other hand, not all ABE patients develop abnormalities in their T1WI at the time. A study from Mao [17] reported that 20 of 36 HB neonates have symmetric hyperintense GP on T1WI; and among these 20 HB neonates, 15 of 20 were ABE neonates. Another study from Wang et al. [14] reported 19 of the 24 ABE patients in their study were observed T1 hyperintensity in the bilateral GP while other 5 of 24 were not; and among these 19 patients, 10 of 19 had high T1 intensities in the STN while others had not. Coskun et al. [12] investigated the GP involvement in 13 neonates with ABE, and 8 of them demonstrated bilateral, symmetric

increased signal intensity in GP on T1WI, while others did not. These studies based on manually radiological reading were subjective and prone to bias since various degree of T1 signal might be involved and there was not a standard for how high the T1 intensity can be a hyperintensity, which may result in different results for different observers. Therefore, improving the sensitivity and specificity of identifying ABE from HB patients based on T1WI alone remains challenging in conventional radiological reading.

In recent years, computer-aided diagnosis (CAD) technology was widely used to improve the radiologist's performance [18–20]. One of the important CAD methods named deep convolutional neural networks (CNN) was applied in our study, which demonstrated remarkable ability in diagnosing a variety of neurological diseases [21–25]. In this study, we compared 3 approaches to differentiate the ABE and non-ABE neonates from a cohort of HB babies based on routine clinical 2D multislice T1WI, including (1) radiological reading, (2) semi-quantitative analysis with normalized T1 intensities, (3) CNN-based classification with ResNet18 [26]. We systematically evaluated the diagnostic accuracy of 3 approaches in a retrospective study of 79 HB patients including 47 ABE and 32 non-ABE neonates. To the best of our knowledge, this is the first work to use semi-quantitative assessment and CNN in ABE diagnosis.

## Methods

### Study subjects

All procedures performed in this study involving human participants were following the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. All the MR images were collected retrospectively from routine clinical scans at the Children's Hospital of Zhejiang University School of Medicine between the years of 2009 and 2018. A total of 79 HB patients (ABE/non-ABE = 47/32, male/female = 52/27), who had MRI examinations at chronological age from 1 to 18 days during their hospitalization, were selected. The diagnostic criteria for ABE positive cases met either of the following clinical diagnosis criteria, including (1) severe hyperbilirubinemia; (2) at least one of the ABE-related clinical symptoms with bilirubin-induced neurologic dysfunction (BIND) score  $\geq 1$  point, where 1, 2, or 3 points corresponding to mild, moderate, or severe symptoms based on the severity of the crying pattern, behavioral and mental status, and muscle tone (a total of 9 points). Overall BIND score of 1–3 points, 4–6 points, 7–9 points represented subtle signs of mild ABE, moderate ABE, and

advanced ABE, respectively. At last, 47 ABE and 32 non-ABE (HB) cases were confirmed by 2 experienced pediatric radiologists. (S.XX and C.L.).

### MRI Acquisition

T1WI was acquired on a 3.0 T Achieva scanner (Philips Healthcare, Best, the Netherlands) using a 2D multislice T1-weighted fast field-echo sequence in the axial direction with the following parameters: echo time of 2.14 ms, repetition time of 200 ms, flip angle of 80°, field-of-view of 330 × 330 mm, resolution of 0.45 mm, and 18 slices with a thickness of 4.5 mm. All slices were visually examined by the radiologists with high image quality and none of them were excluded.

### Visual inspection of the MR images

We invited three pediatric radiologists, including a senior radiologist with 12 years of experience (C.L.) and 2 fellows (L.Y. and Z. S.) with 7 years of experience, to independently review the MR images. Since there are currently no radiological standards for diagnosing ABE, the raters were first trained by reviewing all the MR images with true labels (ABE or non-ABE) and the corresponding clinical information, such as age, sex, TSB, etc. One week after the training, they were asked to review the images again without the labels nor the clinical information to make a diagnosis decision based on T1WI only. The images were shuffled with re-assigned identification numbers at the training and testing sessions.

### Semi-quantitative diagnosis with normalized T1 Intensity

We chose a slice from T1WI of each HB neonate covering the largest area of GP and STN for analysis. Region of interest (ROI) was manually delineated on the selected slice, including the anterior subcortical white matter (WM), GP, and STN, as shown in Fig. 1a. WM was used as the reference region to normalize the T1-signals in GP or STN, because there are no known T1-signal changes in WM between the ABE and non-ABE patients. The normalized T1 intensities of GP and STN were defined as

$$GP_{\text{norm}} = \frac{\overline{GP}}{\overline{WM}} \quad (1)$$

$$STN_{\text{norm}} = \frac{\overline{STN}}{\overline{WM}} \quad (2)$$

where  $\overline{GP}$ ,  $\overline{STN}$ , and  $\overline{WM}$  denote the averaged T1WI intensities in the GP, STN, and WM ROIs, respectively. We then applied the Youden Index [27–29] in the software of IBM SPSS Statistics 21 to determine the optimal cut-off threshold of  $GP_{\text{norm}}$  and  $STN_{\text{norm}}$  for separating ABE and non-ABE patients, respectively.

### Deep learning framework

In this section, we describe the deep learning procedure for classification work. 2 or 3 continuous T1WI slices covered the GP from each patient were selected as inputs of the CNN. A total of 190 slices were collected, including 95 randomly selected slices from ABE patients (approximately 2 continuous slices per patient) and 95 slices from non-ABE patients (approximately 3 continuous slices per patient). All selected images were normalized between 0 and 1 with a min–max normalization algorithm.

As our classifier was performed based on a pre-trained CNN in Matlab 2019a (<https://www.mathworks.com>), which requires 3 channels image input with a size of 224 × 224 × 3 pixels. Consequently, the normalized image was resized into 224 × 224 pixels and then replicated to 3 channels to form a 224 × 224 × 3 image which served as an input of the network. The dataset was randomly split into a training dataset and testing dataset with 80% and 20% split ratio, respectively. Then, fivefold cross-validation was followed to estimate the model's performance.

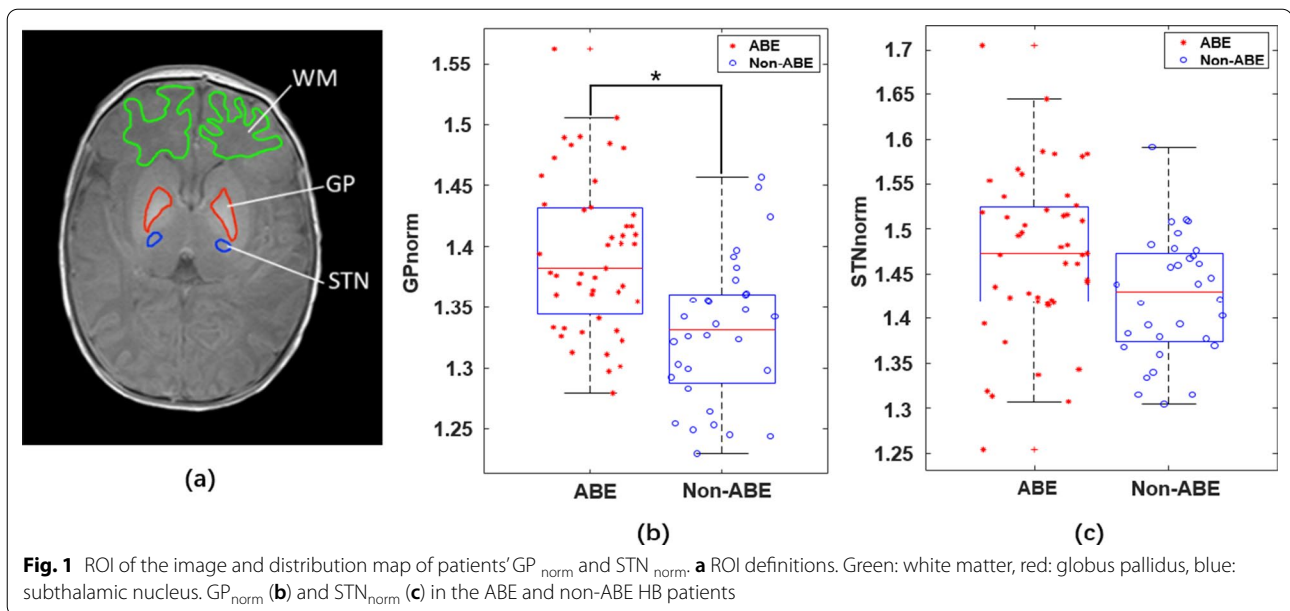
We employed a CNN of Resnet18 [26] which consists of 18 residual blocks, where each residual block is defined as:

$$y = F(x, \{Wi\}) + x \quad (3)$$

where  $x$  and  $y$  were the input and output, and  $F(x, \{Wi\})$  represented the residual mapping to be learned. ResNet18 applied residual learning to every few stacked layers. A residual block was different from conventional CNN architecture in the existence of a shortcut connection between the input and output, serving as an identity projection for alleviating the vanishing gradient issue in deep networks[26], as shown in Fig. 2a. The mapping function  $H(x) = F(x) + x$  was realized as a residual shortcut connection in a feedforward neural network and performs element-wise addition.

The ResNet18 architecture was shown in Fig. 2b, containing 18 learnable layers. The convolutional layers used 3 × 3 filters, and the downsampling was performed for every 4 layers after the input layer by convolutional layers with a stride of 2. Note the number of filters get doubled as a downsample took place. At the end of ResNet18, an average-pooling was applied followed by a fully-connected layer and a softmax layer. Residual shortcut connections denoted as the curves in Fig. 2b were added throughout the network. The solid curves were used when input and output had the same dimensions; while the dotted curves were used when the dimension increased, where the shortcut performed identity mapping with zeros padding for the increased dimension with a stride of 2.

Since the size of our dataset was limited, we applied the transfer learning approach for our classification



model [30]. The weights of ResNet18 were initialized by pre-training on the ImageNet [31] and then fine-tuned with our datasets. Data augmentation was also applied on the training datasets to enhance our model's performance, which included image rotation with a random angle in the range of  $-30^\circ$  to  $30^\circ$ , image vertical flipping with 50% probability, images zooming by a random scale within the range of 0.9 to 1.1, image horizontal and vertical translation with random distance in the range of  $-30$  to 30 pixels. The learning rate was initialized to 0.0003, MaxEpoch=6, Stochastic Gradient Descent Momentum based solver is used with a minibatch size of 10 images for training.

To investigate which brain areas influence our classification results most, we applied the class activation mapping (CAM) to each testing subject. CAM is a technique used to get visual interpretations of the regional contributions to the predications of CNN [32].

The model was implemented under the environment of Matlab 2019a on a computer having specifications of 16 GB RAM and Inter<sup>®</sup> Core<sup>™</sup> i7-8700, CPU@ 3.20 GHz, GPU NVIDIA Geforce GT 730.

### Statistical analysis

The group differences in the sex distribution among groups were evaluated using the chi-squared test, while other clinical features were evaluated by t-test.

The group differences in  $GP_{norm}$  and  $STN_{norm}$  between the ABE and non-ABE patients were accessed by using

analysis of covariance (ANOVA) with age, sex, gestational age at birth, and PMA as covariates.

To evaluate the classification performances of different methods, several performance metrics were applied in this study, including sensitivity, specificity, precision, F1-score, and the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curves [33, 34].  $\chi^2$ -test was applied to determine the significant differences in the classification accuracy by different methods.

All statistical analysis was performed using IBM SPSS Statistics 21 (<https://www.ibm.com/products/spss-statistics>).

### Results

The demographic and clinical characteristics of the HB patients were listed in Table 1, including the patient's sex, age, weight, gestational age, TSB, and albumin. The results were shown in Table 1.

Figure 3 showed several representative T1WI from ABE and non-ABE patients who were diagnosed with HB. The two groups exhibited similar image features with hyperintensity in the GP and large individual variations were observed.

### Results of visual inspection of the MR images

We recorded the 3 experienced radiologists' visual diagnosis results, and their average diagnostic results were 52.48%, 55.21%, 62.95%, 0.5679, 53.58%, 0.5387 for sensitivity, specificity, precision, F1-score, accuracy, AUC, respectively, shown in Table 2. The results indicated the difficulty in separating the 2 groups by conventional

**Table 1** The demographic and clinical characteristics of the HB patients used in this study

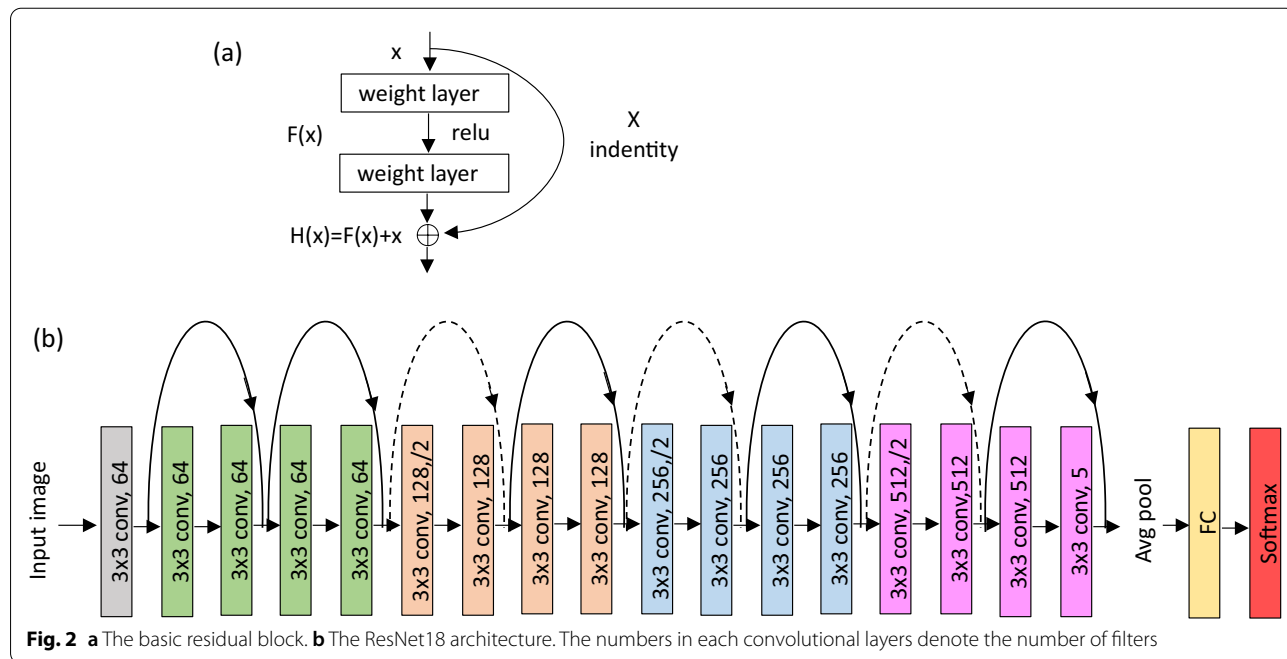
Clinical features	ABE positive (n = 47)	ABE negative (HB) (n = 32)	p value
Sex (male)	29(61.70%)	23(71.88%)	0.349
Age (days)	9.83 ± 3.05	12.15 ± 5.28	0.032
Weight (kg)	3.21 ± 0.48	3.36 ± 0.43	0.162
Gestational age (weeks)	38.47 ± 1.58	38.38 ± 1.47	0.792
TSB (μmol/L)	369.11 ± 114.78	326.13 ± 79.20	0.070
Albumin (g/L)	38.34 ± 2.98	38.45 ± 3.21	0.873

radiological reading in real clinical practice. The overall Fleiss' kappa coefficient for intraobserver reliability is 0.5082 ( $p < 0.05$ ), indicating the agreement of

radiologists' ( $p > 0.05$ ). The ROC curve generated based on the diagnosis results of the senior radiologist, which was the best among the three raters, was shown in Fig. 4d (blue curve).

**Results of Semi-quantitative diagnosis with normalized T1 intensity**

The distribution of normalized T1 intensity in GP and STN were shown in Fig. 1b and Fig. 1c for the ABE and non-ABE patient groups. The t-test indicated that a significant difference between the ABE and non-ABE in the  $GP_{norm}$  ( $1.39 \pm 0.06$  and  $1.33 \pm 0.06$ ,  $p < 0.05$ ), but no significant difference in the  $STN_{norm}$  ( $1.47 \pm 0.09$  and  $1.42 \pm 0.07$ ,  $p > 0.05$ ). The ROC curve based on  $GP_{norm}$  was shown in Fig. 4d (orange curve). The AUC was 0.769 for  $GP_{norm}$ , and 0.678 for  $STN_{norm}$ , respectively. The optimal cut-off thresholds based on the Youden Index were



**Fig. 2** a The basic residual block. b The ResNet18 architecture. The numbers in each convolutional layers denote the number of filters

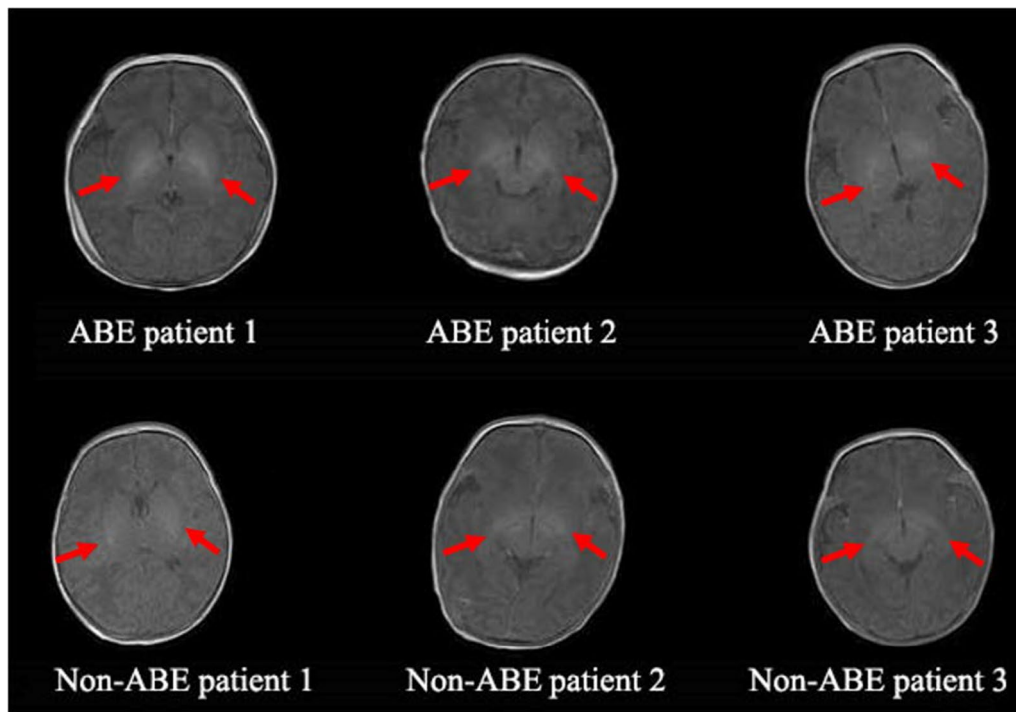
3 radiologists was moderate and not accidental. Furthermore,  $\chi^2$ -test also indicated that there was no statistically significant difference in the results between 3

1.3621 and 1.5118 for  $GP_{norm}$  and  $STN_{norm}$ , respectively.

**Table 2** The classification performance of visual inspection,  $GP_{norm}$  and ResNet18 in separating ABE from non-ABE HB patients, as evaluated by sensitivity, specificity, precision, F1-score, Accuracy, AUC

Methods	Sensitivity	Specificity	Precision	F1-Score	Accuracy	AUC
Visual inspection	52.48 ± 13.58%	55.21 ± 7.86%	62.95 ± 3.58%	56.79 ± 8.90%	53.58 ± 5.71%	53.87 ± 4.11%
$GP_{norm}$	68.10%	<b>78.10%</b>	<b>82.05%</b>	<b>74.42%</b>	<b>72.15%</b>	<b>76.90%</b>
ResNet18	<b>78.95 ± 17.85%</b>	45.26 ± 19.19%	59.58 ± 7.09%	67.11 ± 8.28%	62.11 ± 8.03%	68.92 ± 11.06%

The maximum value of performance metrics for each method was marked in bold



**Fig. 3** Representative T1WI from three ABE and three non-ABE neonates who were diagnosed as HB. The arrows pointed to the bilateral areas of the globus pallidus

### Results of ResNet18

The diagnostic performance of ResNet18 as evaluated by five-cross validation was:  $78.95 \pm 17.85\%$ ,  $45.26 \pm 19.19\%$ ,  $59.58 \pm 7.09\%$ ,  $67.11 \pm 8.28\%$ ,  $62.11 \pm 8.03\%$ ,  $68.92 \pm 11.06\%$  for sensitivity, specificity, precision, F1-score, accuracy, AUC, respectively, which were listed in Table 2.

Figure 5b demonstrated four examples of the CAM on the testing samples that were correctly predicted by the network. The red regions in the CAM represented the brain regions that contributed most in predicting the results, and they were mostly located in the center of the brain covering the areas of bilateral GP and STN.

### Comparison of three methods

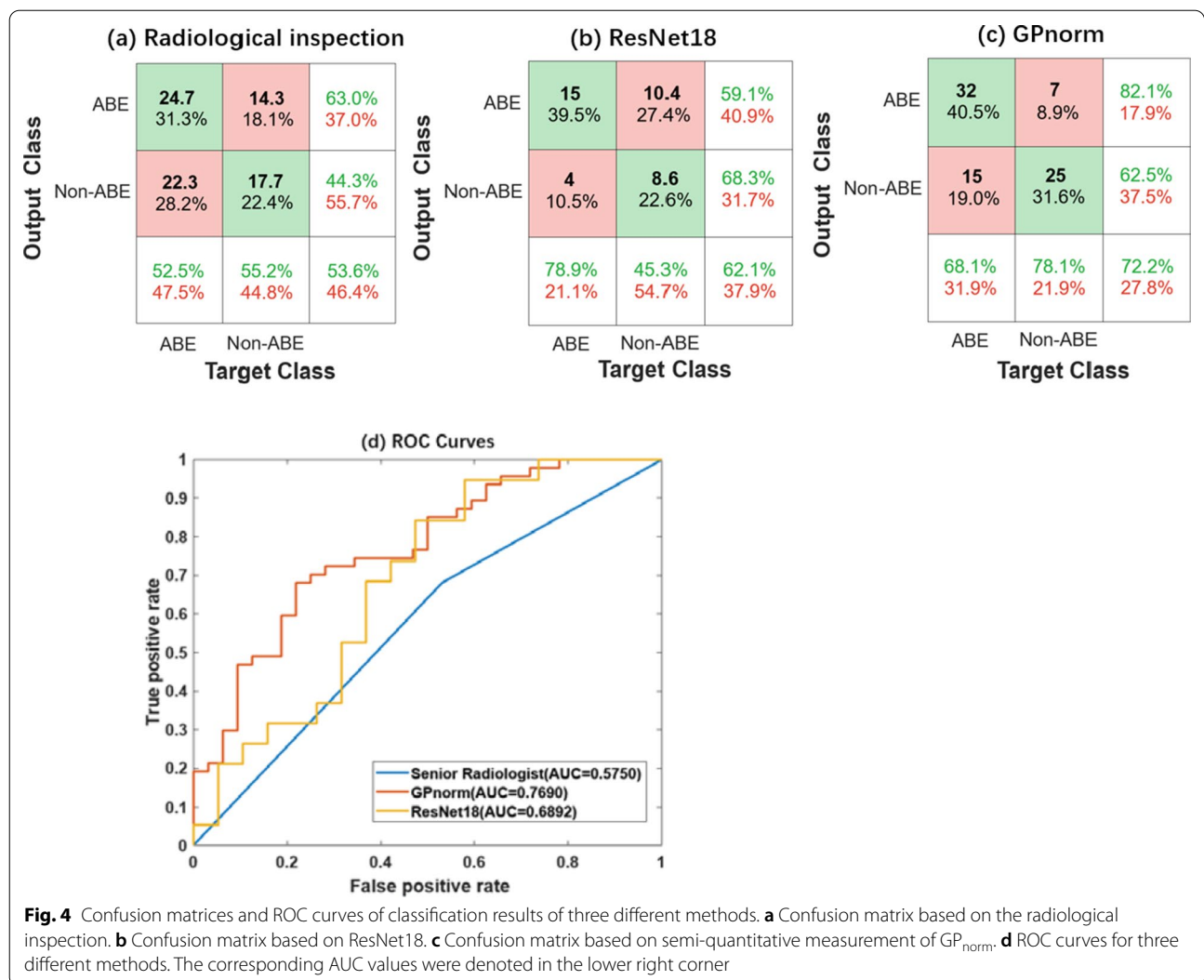
The classification performances of different methods were compared in Table 2. The semi-quantitative method based on  $GP_{norm}$  showed superior performance compared to the other two approaches except for the sensitivity measure. Figure 4a-c was the confusion matrix of classification results of three different methods. Figure 4d showed the ROC curves based on the three methods for direct comparison.  $\chi^2$ -test indicated that the accuracy of  $GP_{norm}$  marker was significantly higher than that of visual inspection ( $p=0.014$ ), but no difference was found

between the results of visual inspection and ResNet18 ( $p=0.234$ ) or between results of ResNet18 and  $GP_{norm}$  ( $p=0.153$ ).

### Discussion

Currently, the radiological finding of ABE is T1 hyperintensity in the areas of GP and STN since they were affected by the bilirubin [14, 35]. A previous study [36] indicated that the relatively high resting neuronal activity in the GP and STN are postulated to make them more vulnerable to oxidative stresses from mitochondrial toxins, such as bilirubin, or genetic mitochondrial disorders. Such damages to the GP and STN of the ABE infants can be often observed on their T1WI, resulting in T1 hyperintensity in various degrees [14, 35]. However, the sensitivity and specificity of this radiological feature are only moderate since only T1WI is studied without any other complementary and useful information [12] [14] [17]. A future study including multi-modal MRI and clinical information of the patients is expected to improve the diagnostic performance.

Our study aimed to investigate the utility of T1WI for the diagnosis of ABE in neonates. 3 different diagnostic methods are performed. As shown in Table 2, the accuracy and AUC from low to high are visual inspection ( $53.58 \pm 5.71\%$ ,  $0.5387 \pm 4.11\%$ ), ResNet18



**Fig. 4** Confusion matrices and ROC curves of classification results of three different methods. **a** Confusion matrix based on the radiological inspection. **b** Confusion matrix based on ResNet18. **c** Confusion matrix based on semi-quantitative measurement of GP<sub>norm</sub>. **d** ROC curves for three different methods. The corresponding AUC values were denoted in the lower right corner

(62.11 ± 8.03%, 68.92 ± 11.06%), and semi-quantitative diagnostic method with GP<sub>norm</sub> (72.15%, 0.7690). The study demonstrated that the performance of conventional radiological reading based on T1WI for diagnosing ABE is unsatisfactory. Advanced analytical approaches with deep learning and semi-quantitative measurement outperformed the conventional radiological reading, and therefore, offering the opportunity to improve the diagnostic accuracy of the ABE in clinical practice.

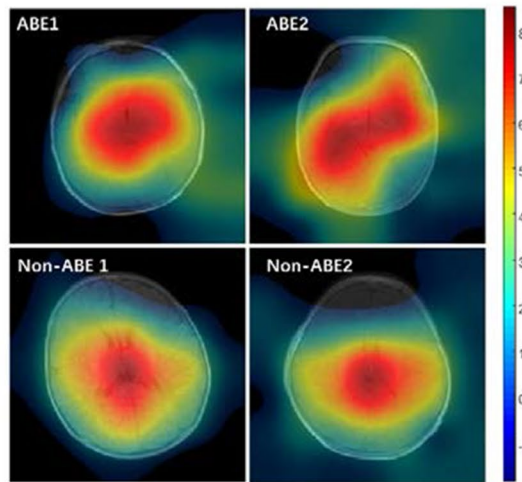
Although MRI has been increasingly applied to investigate the neuropathology induced by ABE in the neonatal clinical practice, the efficiency and accuracy of the conventional radiological reading strategy solely based on visual inspection of GP and STN on T1WI were hardly satisfactory. This is because the T1 intensity of GP and STN in ABE demonstrated a high extent of heterogeneity and the level of T1 hyperintensity may be confounded by

the normal development of GP as well as myelination in the adjacent posterior limb of the internal capsule. There is no clear boundary between ABE and normal neonates, nor to mention the distinction between ABE and non-ABE HB neonates, e.g., some non-ABE cases also showed slight T1 hyperintensity in GP and STN (Fig. 3). Besides, the diagnosis of routine visual inspection is qualitative and subjective, e.g., we observed a relatively high inter-rater variability among the three raters (Fleiss' kappa coefficient = 0.5082, *p* < 0.05).

As the deep learning technology has been wildly used in medical image analysis, we applied the deep learning model ResNet18 to T1WI-based diagnosis of ABE. The CAM map in Fig. 5 indicated the center regions of the brain covering bilateral GP and STN played a critical role in the classification task. This was consistent with our prior knowledge that most ABE patients followed an



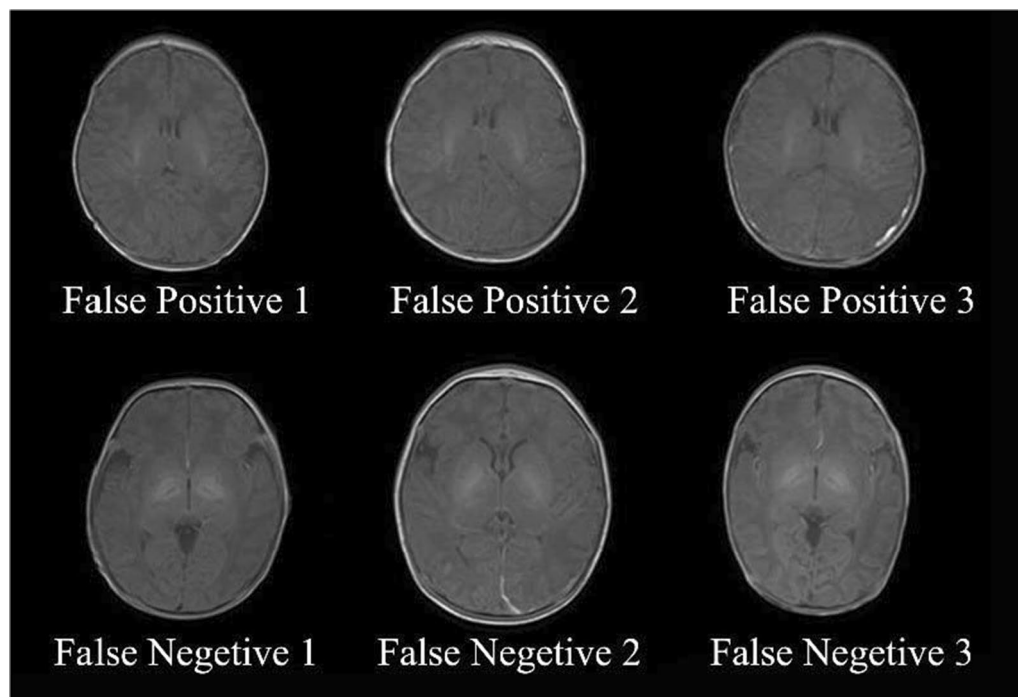
(a)



(b)

**Fig. 5** **a** A training progress of ResNet18 in fivefold cross-validation: the accuracy and loss history. **b** Class activation map of resnet18 for 4 exemplary test samples. The colormap showed the contribution of the voxels in the network in predicting results and the red region contribute most





**Fig. 6** Examples of false-positive cases (non-ABE HB patients who were misclassified as ABE) and false-negative cases (ABE patients who were misclassified as non-ABE HB) by the ResNet18 network

increased T1-signals in GP and STN [12–17]. However, the improvement in diagnostic accuracy was only moderate (from 53.58% for visual inspection to 62.11% for ResNet). It because some image samples applied in our study have not prominent or even inverse manifestation between ABE and non-ABE. The misclassified samples by ResNet18 shown in Fig. 6 indicated that some non-ABE images have a prominent T1 hyperintensity in GP, whereas some ABE images did not show visual abnormalities. It is a common phenomenon in clinical practice as the hyperintensity of GP on T1WI is only a clinical manifestation of ABE that is likely to occur but not necessarily [12, 14], which indicates that ABE diagnosis based only on T1WI is not sufficient in the clinical practice. Thus, additional clinical information is needed to support the diagnosis of ABE, such as TSB, Albumin, etc. [37]. Also, we found that during the training process, the overall tendency of training accuracy did not increase much during the training process, which indicated that ResNet18 cannot perfectly differentiate the samples that did not have instinct features. We noticed that in Fig. 5a validation loss increased after 40 iterations and the training loss kept decreasing, which indicated overfitting took place. This, in turn, means that the complexity of our model was greater compared to the limited training samples. Moreover, the generalizability of our model

is unknown, which is a known caveat of CNN for small sample size data [38]. Additional studies, ideally covering a large number of cases from multiple centers, are needed to further improve the diagnosis of neonatal ABE.

To circumvent the issues related to small sample size, model complexity, and generalizability, we took a simple and model-free approach using the semi-quantitative based on normalized T1 intensity in GP and STN regions. We found a statistical difference in  $GP_{norm}$  between ABE and non-ABE and there was no distinct line yet; meanwhile, no statistical difference was found in  $STN_{norm}$ . Therefore, an optimal threshold of 1.3621 was determined to separate ABE patients from HB neonates based on  $GP_{norm}$ , which achieved significantly improved diagnostic performance compared to the visual inspection. This semi-quantitative diagnostic pipeline would be expected other datasets given the minimal requirement on preprocessing, computational power, and training data. We deem that GP can be observed at T1WI and its T1-intensity may have a subtle variation when different MRI equipment was applied. However, the value of  $GP_{norm}$  would not be changed as it is a normalized value. Nevertheless, the results of all three experiments demonstrated that it is not enough to make an accurate diagnosis only based on the T1WI alone. Therefore a study combining the information of T1WI and other MRI

modalities, i.e. T2-weighted, diffusion-weighted MRI, etc. is essential in the future work for a more accurate diagnostic result.

## Conclusion

The current study investigates the utility of T1WI in diagnosing ABE conditions through three analytical approaches. The semi-quantitative diagnostic method provided the highest performance followed by ResNet18, which both outperformed the conventional visual inspection strategy. In particular, the semi-quantitative  $GP_{norm}$  achieved the highest accuracy of 72.15% and AUC of 76.90%. Our work showed advanced analytical approaches to make the best use of conventional T1WI which would assist the diagnosis of ABE in real clinical practice.

## Abbreviations

ABE: Acute bilirubin encephalopathy; HB: Hyperbilirubinemia; TSB: Total serum bilirubin; MRI: Magnetic resonance imaging; T1WI: T1-weighted images; GP: Globus pallidus; STN: Subthalamic nucleus; WM: White matter; CAD: Computer-aided diagnosis; CNN: Convolutional neural networks; BIND: Bilirubin-induced neurologic dysfunction; CAM: Class activation mapping; ROI: Region of interest; AUC: Area under the curve; ROC: Receiver Operating Characteristic.

## Acknowledgements

Thanks are due to Dan Wu and Weihao Zheng for assistance with the experiments and paper writing, Xiaoxia Shen and Tingting Liu for data collection, Can Lai, Yingqun Li, and ZhongliShangguan for participating in the experiment, Chuanbo Yan for valuable discussion.

## Authors' contributions

MW proposed methods, performed the computations, and wrote the paper. DW devised and supervised the project. XS and TL collected the data from the hospital. CL, YL, and ZLS performed the experiment of MRI visual inspection. CY solved some detailed problems in the experiment of deep learning. DW and WZ participated in revising the paper.

## Funding

This work was funded by the Ministry of Science and Technology of the People's Republic of China (2018YFE0114600, funder: DW), Natural Science Foundation of China (61801424, 81971606, 91859201, 61801421, and 81971605, funder: DW), Fundamental Research Funds for the Central Universities of China (2019QNA5024, funder: DW). China postdoctoral science foundation (2202M671727, funder: WHZ), postdoctoral science foundation of Zhejiang Province, China (514000-X81901, funder: WHZ), State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Disease in Central Asia Fund (SKL-HIDCA-2020-YG2). DW and WHZ are the authors of this paper.

## Availability of data and material

The datasets generated and/or analyzed during the current study are not publicly available as they contain identifiable and personal information but are available from the corresponding author on reasonable request.

## Declarations

### 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 guardian participants.

## Consent for publication

Written informed consent for publication of their clinical details and clinical images was obtained from the patients' parents. A copy of the consent form is available for review by the Editor of this journal.

## Competing interest

The authors declare that they have no conflict of interest.

## Author details

<sup>1</sup>Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou 310027, China. <sup>2</sup>Department of Neonatal Intensive Care Unit, Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310051, China. <sup>3</sup>Department of Radiology, Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310052, China. <sup>4</sup>State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830011, China.

Received: 7 October 2020 Accepted: 4 June 2021

Published online: 22 June 2021

## References

1. Osuorah CDI, Ekwochi U, Asinobi IN. Clinical evaluation of severe neonatal Hyperbilirubinaemia in a resource-limited setting: a 4-year longitudinal study in south-East Nigeria. *Bmc Pediatr*. 2018. <https://doi.org/10.1186/s12887-018-1174-z>.
2. Rennie J, Burman-Roy S, Murphy S. Neonatal jaundice: summary of NICE guidance. *BMJ Br Med J*. 2010;340:c2409.
3. Allen D. Neonatal jaundice. *Nurs Child Young People*. 2016;28(6):11–11.
4. Altuntas N. Is there any effect of hyperbilirubinemia on breastfeeding? If any, at which level? *Breastfeeding Medicine*.
5. Smitherman H, Stark AR, Bhutani VK. Early recognition of neonatal hyperbilirubinemia and its emergent management. *Semin Fetal Neonatal Med*. 2006;11(3):214–24.
6. Watchko FJADotN, 84 – Neonatal Indirect Hyperbilirubinemia and Kernicterus. 2018.
7. Olusanya BO, Osibanjo FB, Mabogunje CA, Slusher TM, Olowe SA. The burden and management of neonatal jaundice in Nigeria: A scoping review of the literature. *Niger J Clin Pract*. 2016;19(1):1–17.
8. Pace EJ, Brown CM, DeGeorge KC. Neonatal hyperbilirubinemia: an evidence-based approach. *J Fam Pract*. 2019;68(1):E4–11.
9. Maisels MJ. Managing the jaundiced newborn: a persistent challenge. *CMAJ*. 2015;187(5):335–43.
10. Wang FQ, Liu XT, Yuan N, Qian BY, Ruan LT, Yin CC, Jin CP. Study on automatic detection and classification of breast nodule using deep convolutional neural network system. *J Thorac Dis*. 2020;12(9):4690–701.
11. Tatli MM, Karadag A, Oedemis E, Sarraoglu S, Yoeruebulut M. The role of magnetic resonance imaging in the prediction of the neurodevelopmental outcome of acute bilirubin encephalopathy in newborns. *Turk J Med Sci*. 2009;39(4):507–11.
12. Coskun A, Coskun A, Yikilmaz A, Yikilmaz A, Kumandas S, Kumandas S, Karahan OI, Karahan OI, Akcakus M, Akcakus M, Manav A, Manav A. Hyperintense globus pallidus on T1-weighted MR imaging in acute kernicterus: is it common or rare? *Eur Radiol*. 2005;15(6):1263–7.
13. Gkoltsiou K, Tzoufi M, Counsell S, Rutherford M, Cowan F. Serial brain MRI and ultrasound findings: Relation to gestational age, bilirubin level, neonatal neurologic status and neurodevelopmental outcome in infants at risk of kernicterus. *Early Hum Dev*. 2008;84(12):829–38.
14. Wang X, Wu W, Hou BL, Zhang P, Chineah A, Liu F, Liao W. Studying neonatal bilirubin encephalopathy with conventional MRI, MRS, and DWI. *Neuroradiology*. 2008;50(10):885–93.
15. Wu W, Zhang P, Wang X, Chineah A, Lou M. Usefulness of H-1-MRS in differentiating bilirubin encephalopathy from severe hyperbilirubinemia in neonates. *J Magn Reson Imaging*. 2013;38(3):634–40.
16. Liao W-H, Wang X-Y, Wu W-L, Jiang X-Y, Liu Y-H, Liu F, Wang R-W. Differentiation of hypoxic-ischemic encephalopathy and acute bilirubin

- encephalopathy with magnetic resonance imaging in neonates. *Zhongguo dang dai er ke za zhi Chin J Contemp Pediatr.* 2009;11(3):181–4.
17. Mao J, Fu JH, Chen L-Y, Wang X-M, Xue X-D. Changes of globus pallidus in the newborn infants with severe hyperbilirubinemia. *Zhonghua er ke za zhi Chin J Pediatr.* 2007;45(1):24–9.
  18. Zhao WJ, Fu LR, Huang ZM, Zhu JQ, Ma BY. Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid nodules on ultrasound A systematic review and meta-analysis. *Medicine.* 2019;98(32):e16379.
  19. Park HJ, Kim SM, La Yun B, Jang M, Kim B, Jang JY, Lee JY, Lee SH. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: Added value for the inexperienced breast radiologist. *Medicine.* 2019;98(3):e14146.
  20. Kawagishi M, Kubo T, Sakamoto R, Yakami M, Fujimoto K, Aoyama G, Emoto Y, Sekiguchi H, Sakai K, Iizuka Y, Nishio M, Yamamoto H, Togashi K. Automatic inference model construction for computer-aided diagnosis of lung nodule: Explanation adequacy, inference accuracy, and experts' knowledge. *PLoS ONE.* 2018;13(11):0207661.
  21. Ceschin R, Zahner A, Reynolds W, Gaesser J, Zuccoli G, Lo CW, Gopalakrishnan V, Panigrahy A. A computational framework for the detection of subcortical brain dysmaturation in neonatal MRI using 3D Convolutional Neural Networks. *Neuroimage.* 2018;178:183–97.
  22. Talo M, Yildirim O, Baloglu UB, Aydin G, Acharya UR. Convolutional neural networks for multi-class brain disease detection using MRI images. *Comput Med Imaging Graph.* 2019;78:101673–101673.
  23. Faturrahman M, Wasito I, Hanifah N, Mufidah R. Structural MRI classification for Alzheimer's disease detection using deep belief network. *IEEE.*
  24. Badža MM, Barjaktarović MČ. Classification of brain tumors from MRI images using a convolutional neural network. *Appl Sci.* 2020;10(6):1999.
  25. Huang YC, Xu JH, Zhou YC, Tong T, Zhuang XH. Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network. *Frontiers Neurosci.* 2019. <https://doi.org/10.3389/fnins.2019.00509>.
  26. He K, Zhang X, Ren S, Jian S. Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition.* 2016.
  27. Reiser BJB. Estimation of the Youden Index and its Associated Cutoff Point. *Biomet J.* 2005;47:458.
  28. Youden WJC. Index for rating diagnostic tests. 1950;3(1):32–5.
  29. Schisterman EF, Faraggi D, Reiser B, Hu J. Youden Index and the optimal threshold for markers with mass at zero. *Stat Med.* 2008;27(2):297–315.
  30. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng.* 2010;22(10):1345–59.
  31. Simonyan K, AZ. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556;* 2014.
  32. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *IEEE.*
  33. Padmanabhan A. ROC Curve. 2019 22 Aug 2019. Available from: <https://devopedia.org/roc-curve>.
  34. Swati ZNK, Zhao QH, Kabir M, Ali F, Ali Z, Ahmed S, Lu JF. Brain tumor classification for MR images using transfer learning and fine-tuning. *Comput Med Imaging Graph.* 2019;75:34–46.
  35. Wisnowski JL, Panigrahy A, Painter MJ, Watchko JF. Magnetic resonance imaging of bilirubin encephalopathy: Current limitations and future promise. *Semin Perinatol.* 2014;38(7):422–8.
  36. Johnston MV, Hoon AH. Possible mechanisms in infants for selective basal ganglia damage from asphyxia, kernicterus, or mitochondrial encephalopathies. *J Child Neurol.* 2000;15(9):588–91.
  37. Zhang F, Chen L, Shang S, Jiang K. A clinical prediction rule for acute bilirubin encephalopathy in neonates with extreme hyperbilirubinemia A retrospective cohort study. *Medicine.* 2020;99(9):e19364.
  38. Tipton E, Hallberg K, Hedges LV, Chan W. Implications of small samples for generalization: adjustments and rules of thumb. *Eval Rev.* 2017;41(5):472–505.

## Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

