



## OPEN Brain CT image classification based on mask RCNN and attention mechanism

Shoulin Yin<sup>1</sup>✉, Hang Li<sup>1</sup>✉, Lin Teng<sup>1</sup>✉, Asif Ali Laghari<sup>1</sup>✉, Ahmad Almadhor<sup>2</sup>, Michal Gregus<sup>3</sup>✉ & Gabriel Avelino Sampedro<sup>4</sup>

Along with the computer application technology progress, machine learning, and block-chain techniques have been applied comprehensively in various fields. The application of machine learning, and block-chain techniques into medical image retrieval, classification and auxiliary diagnosis has become one of the research hotspots at present. Brain tumor is one of the major diseases threatening human life. The number of deaths caused by these diseases is increasing dramatically every year in the world. Aiming at the classification problem of brain CT images in healthcare. We propose a Mask RCNN with attention mechanism method in this research. First, the ResNet-10 is utilized as the backbone model to extract local features of the input brain CT images. In the partial residual module, the standard convolution is substituted by deformable convolution. Then, the spatial attention mechanism and the channel attention mechanism are connected in parallel. The deformable convolution is embedded to the two modules to extract global features. Finally, the loss function is improved to further optimize the precision of target edge segmentation in the Mask RCNN branch. Finally, we make experiments on public brain CT data set, the results show that the proposed image classification fragrance can effectively refine the edge features, increase the degree of separation between target and background, and improve the classification effect.

**Keywords** Machine learning, Healthcare, Brain CT image classification, Mask RCNN, Attention mechanism, Deformable convolution

In 2018, the relevant documents clearly required the improvement of the infrastructure support capacity of medical institutions, focusing on supporting the universal coverage of high-speed broadband networks for medical institutions at all levels in urban and rural areas. To meet the needs of telemedicine and medical information sharing, telecommunications enterprises are encouraged to provide medical institutions with high-quality Internet dedicated lines, virtual private networks (VPNS) and other network access services, promote the construction of telemedicine private networks, ensure the quality of medical data transmission services, and support medical institutions to choose to use high-speed and reliable network access services. After the 6G edge computing solution is landed, relying on its IT capabilities and self-serving capabilities, it can provide high-quality data connection services for customers in the medical industry. At the same time, based on MEC, it provides customers with a low-level application platform to support various services of the hospital and help the hospital realize intelligence and information.

Medical imaging technology is the application of radioactive material imaging technology to medical treatment. B ultrasound, X-ray, CT imaging, magnetic resonance imaging, digital subtraction angiography, etc., are all applications of medical imaging technology in real life. Medical imaging technology uses advanced medical instruments to take images of lesions. According to the morphological changes of pathological changes, they make diagnosis for the treatment of disease to bring practical help. Medical imaging technology is closely related to clinical practice and is an indispensable auxiliary discipline in clinical practice. Computerized Tomography (CT) is widely used in medical imaging to help doctors diagnose cancer, tumors and bone necrosis earlier, because it is noninvasive, inexpensive and convenient<sup>1</sup>. CT is also the preferred imaging tool for diagnosing brain diseases, enabling early detection of potential brain disorders such as Alzheimer's Disease (AD) and timely

<sup>1</sup>Software College, Shenyang Normal University, Shenyang, China. <sup>2</sup>Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jof University, 72388 Sakaka, Saudi Arabia.

<sup>3</sup>Faculty of Management, Comenius University in Bratislava, Bratislava, Slovakia. <sup>4</sup>Department of Computer Science, University of the Philippines Diliman, 1101 Quezon City, Philippines. ✉email: yslin@synu.edu.cn; lihag@synu.edu.cn; tenglinheu@163.com; asiflaghari@synu.edu.cn; michal.gregus@fm.uniba.sk

treatment of patients. CT and MRI are the main auxiliary equipment for diagnosing such diseases<sup>2</sup>. Currently, many hospitals have accumulated a large number of rich brain medical images. Using machine learning, pattern recognition and blockchain technology to mine useful disease association information from such large and rich brain image data sets, and then design auxiliary diagnosis and intelligent classification systems have great practical significance to improve the work efficiency and service quality of hospital radiology department<sup>3,4</sup>.

Recently, the incidence of AD and lesions (brain tumors) has increased significantly, and clinical diagnosis of the disease is a long and complex process. To help doctors detect abnormal brain symptoms in patients early and reduce the difference of artificial diagnosis, many scholars have carried out classification studies on AD, lesions (such as brain tumors) and healthy aging based on brain CT images. Brain CT images contain a lot of noise and have lower resolution compared with liver and gallbladder images, so it is difficult to extract relevant features, resulting in low classification accuracy. Therefore, it is very important to extract distinguishing features from brain CT images for early screening of brain diseases<sup>5</sup>.

As we all know, deep learning (DL) is an important branch of machine learning, which has been adopted widely in various fields of medical image analysis due to its powerful feature learning ability, such as breast cancer, tumor screening<sup>6</sup>, pulmonary nodule classification and diagnosis of congenital heart disease. Nowadays, brain CT and MRI (Magnetic Resonance Imaging) medical images using deep learning algorithms is also gradually popular. Lu et al<sup>7</sup>. proposed a convolutional neural network (CNN) model for automatic segmentation of brain tumors, which adopted a small convolutional kernel to alleviate the over-fitting problem caused by increasing the model depth. Paul et al<sup>8</sup>. extracted local and global features of brain tumors by combining different convolutional neural networks to improve the segmentation of tumors. Charron et al<sup>9</sup>. used an integrated approach to classify medical images. By integrating different CNN models to learn about hierarchical characteristics of medical images such as brain CT, the feature information with strong resolution of various images could be obtained. Reference<sup>10</sup> proposed a semi-supervised multitask learning approach that exploited widely available unlabelled data (i.e., Kaggle-EyePACS) to improve DR segmentation performance. The proposed model consisted of novel multi-decoder architecture and involved both unsupervised and supervised learning phases. Reference<sup>11</sup> proposed a novel fully automatic technique for brain tumor regions segmentation by using multiscale residual attention-UNet (MRA-UNet). MRA-UNet used three consecutive slices as input to preserve the sequential information and employed multiscale learning in a cascade fashion, enabling it to exploit the adaptive region of interest scheme to segment enhanced and core tumor regions accurately. Reference<sup>12</sup> proposed a solution to the limited amount of labeled data by using a multi-task semi-supervised learning (MTSSL) framework that utilized auxiliary tasks for which adequate data was publicly available. Reference<sup>13</sup> proposed a densely attention mechanism-based network (DAM-Net) for COVID-19 detection in CXR. DAM-Net adaptively extracted spatial features of COVID-19 from the infected regions with various appearances and scales. The DAM-Net was composed of dense layers, channel attention layers, adaptive downsampling layer, and label smoothing regularization loss function.

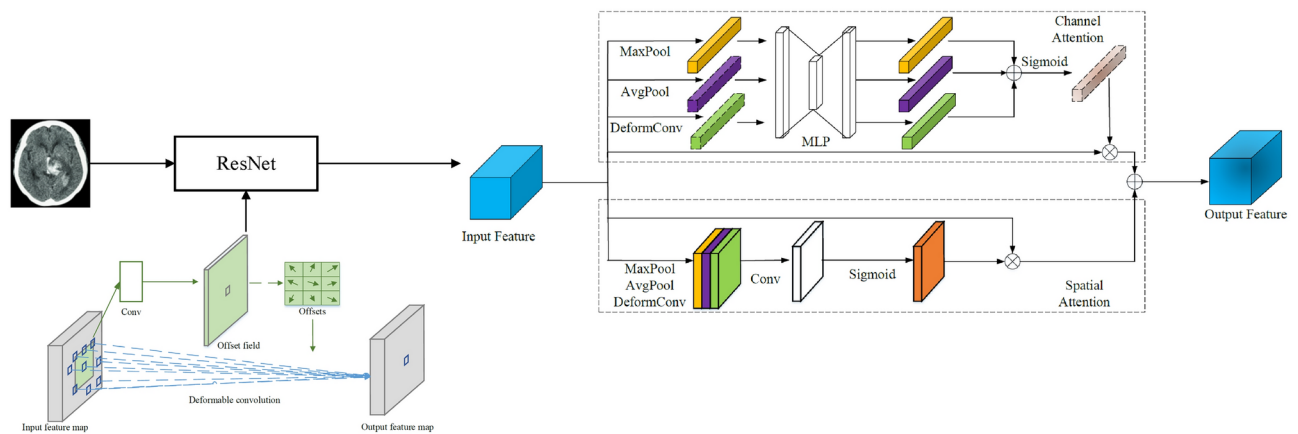
Similarly, Pereira et al<sup>14</sup>. designed four different convolutional neural network models to classify AD, mild cognitive impairment, severe cognitive impairment and healthy brain MRI images, and integrated the classification results through the way of pipeline, achieving better classification effect. It can be seen that studies on DL in brain medical images are becoming more and more successful<sup>15</sup>.

Currently, deep learning-based brain CT classification faces the following challenges:

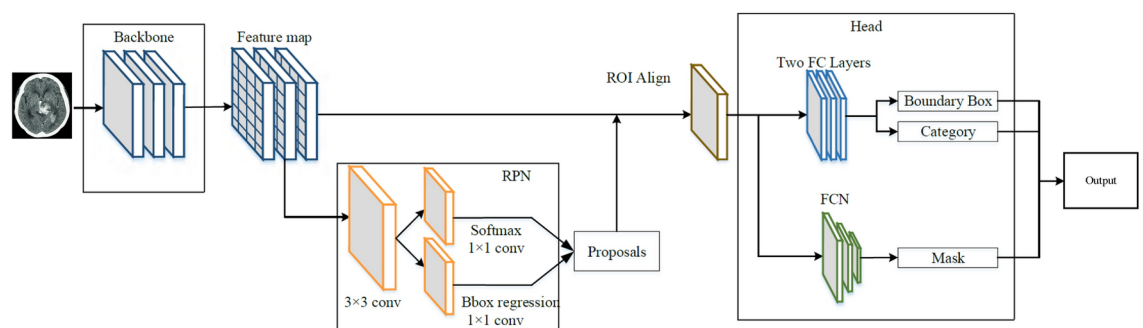
- 1 Data quality and labeling. The quality of brain CT images may be affected by a variety of factors, such as noise, artifacts, etc., which may affect the accuracy of the model. In addition, accurate labeling requires specialized medical knowledge and experience, which can be a challenge.
- 2 Class imbalance problem. In brain CT classification, there may be an imbalance in the number of samples of different categories, which may lead to poor classification performance of the model for a few categories.
- 3 Model complexity and computational resources. Deep learning models usually have high complexity and require a lot of computational resources for training and reasoning. This can be a challenge in some resource-limited environments.
- 4 Model interpretability. The decision-making process of deep learning models is often black-box, which makes it difficult to interpret the model's decisions and results. In the medical field, interpretability of models is very important for clinical applications.
- 5 Domain adaptability. Different medical institutions may use different equipment and scanning parameters, which may lead to differences in brain CT images. Models need to have good domain adaptability to perform well on different data sets.

In fact, deep CNN is used to learn the deep features of medical images, which has better recognition and robustness in image expression<sup>16</sup>. However, most existing DL methods learn features from the entire CT image, which includes not only features related to brain lesions, but also confounding features such as background areas. The features of other tissues contained in the background region are likely to be mixed with the extracted brain features, resulting in poor discrimination of extracted features and inaccurate model classification.

The visual attention mechanism shows its superiority in visual tasks such as facial expression analysis, image classification<sup>17,18</sup> and target detection. This mechanism conforms to the human visual system, which does not attempt to process the entire visual scene information at once, but instead focuses on the salient parts of the entire visual space when they are needed. The visual attention module guides the model to learn from the image region with specific information and helps the model to extract the dynamic features of the most brain-like parts in the image. Wang et al<sup>19</sup>. proposed a residual attention module, which used down-sampling and up-sampling methods to extract attention feature maps<sup>20</sup>. Woo et al<sup>21</sup>. integrated channel and spatial features, proposed a Convolutional Block Attention Module (CBAM), and conducted a classification test in ImageNet data set to



**Fig. 1.** Proposed network in this paper.



**Fig. 2.** Structure of Mask RCNN.

verify the validity of the channel and spatial features combination. However, the attention module mainly used pooling operation to process feature mapping, and it was easy to lose more key information in feature mapping.

Therefore, in this paper, Mask RCNN network is used to construct a classification model to classify brain CT images. For the brain CT image data set used in this paper, Gao et al.<sup>22</sup> designed a simple CNN model using DL method for the first time, which achieved an accuracy of 86.32% for brain CT classification. Zahedinasab et al.<sup>23</sup> proposed a 7-layer CNN model, and combined with different activation functions, the classification accuracy of the model reached 88.67%. Obviously, the classification accuracy of this data set is far from satisfactory.

To solve the above problems, we propose a brain CT classification via Mask RCNN and attention mechanism. Our main contributions are as follows.

Firstly, the ResNet-10 feature extraction network is reconstructed by deformable convolution to improve the feature extraction adaptability to the geometric feature diversity of brain CT images and reduce the interference of background information.

Then, an improved attention mechanism is introduced to make the network pay more attention to the features related to the target, so as to reduce the confusion between instances of different similar features.

Finally, the loss function of Mask RCNN mask branch is optimized to improve the segmentation precision of target edge information by the model, wherein, the model can extract more discernibility brain features and improve the accuracy of model classification. The proposed network is shown in Fig. 1.

## Proposed brain CT image classification model

### Mask RCNN

Based on the Faster RCNN<sup>24</sup>, the Mask RCNN algorithm adds parallel fully connected network to target border recognition branches for target mask prediction. Region of Interest (ROI) Align perfectly solves the problem of area mismatch caused by twice quantization in ROI Pooling operation. The improved deep neural network model can not only detect the target, but also produce high quality segmentation mask for each instance. Mask RCNN is a two-stage algorithm. Firstly, the Region Proposal Network (RPN) searches the image and generates proposal boxes. Secondly, proposals are operated with classification, boundary regression and pixel segmentation. The two-stage algorithm uses the anchor frame mechanism to consider regions of different scales and has higher accuracy in positioning and detection than the one-stage algorithm<sup>25</sup>.

The basic framework of Mask RCNN model is mainly composed of CNN, RPN network, ROI Align and result output unit as shown in Fig. 2.

- 1 CNN. The CNN of the Mask R-CNN model uses Residual Network (ResNet) and Feature Pyramid Network (FPN) to extract and fuse multi-scale features of the image and forms corresponding feature maps.
- 2 RPN network. Multiple candidate frames are generated by sliding the feature map through preset anchor frames with various width and height ratios. The generated candidate boxes are then dichotomized (foreground and background) and the position and size of the candidate boxes are corrected. Non-maximum suppression (NMS) is utilized to screen out more accurate candidate boxes.
- 3 ROI Align. It adopts Bilinear interpolation to calculate the pixel value of each sampling point, retain floating point coordinates, and match the extracted candidate target box with the candidate region in the original image pixel by pixel.
- 4 Result output unit. A full convolutional network mask branch is used to achieve pixel-level segmentation of objects, and two fully connection layers are used to classify objects and conduct boundary box regression.

### Deformable convolution

The brain image has geometric diversity, and different texture features, the difference between the target and the background is small, so the edge of the target to be extracted is fuzzy and difficult to separate. In image detection task, the size and shape of the receptive field in the traditional convolution operation are fixed. For the brain, it is expected that the receptive field of the convolution kernel can pay attention to features such as the edge of the target, so as to minimize the influence of background information on feature representation. Deformable convolution<sup>26</sup> adds extra bias to different convolution regions, which makes the convolution operation more adaptable to the geometric deformation of the target and reduces the interference of background information.

The standard two-dimensional convolution operation is a sliding sampling on a regular feature graph by a regular grid  $R$ . The sample values are summed by weighting  $w$ . The grid  $R$  can define the size and weight of the sampling area  $w$ . The output  $y(O_0)$  of  $O_0$  at each position in the input feature graph  $x$  is:

$$y(O_0) = \sum_{O_n \in R} w(O_0) \cdot x(O_0 + O_n) \quad (1)$$

where  $O_n$  is the sampling position in  $R$  centered on  $O_0$ . In the deformable convolution, an offset  $\Delta O_n$  is added to each sampling point  $O_0$ , and the calculation process is as follows:

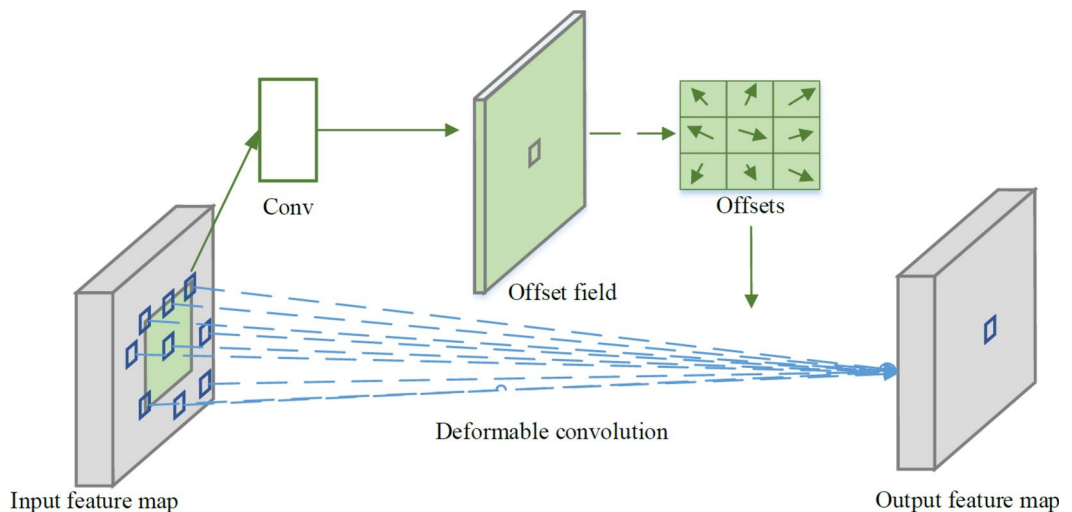
$$y(O_0) = \sum_{O_n \in R} w(O_0) \cdot x(O_0 + O_n + \Delta O_n) \quad (2)$$

The shape of convolution kernel with increasing offset is more flexible and the range of receptive field is optimized. However, when the offset is added to the deformable convolution, the receptive field range is larger than the target range, and interference features are extracted. Therefore, DCNv2 adds weight coefficient  $\Delta m_k$  to each sampling point to ensure effective feature extraction. The calculation process is as follows:

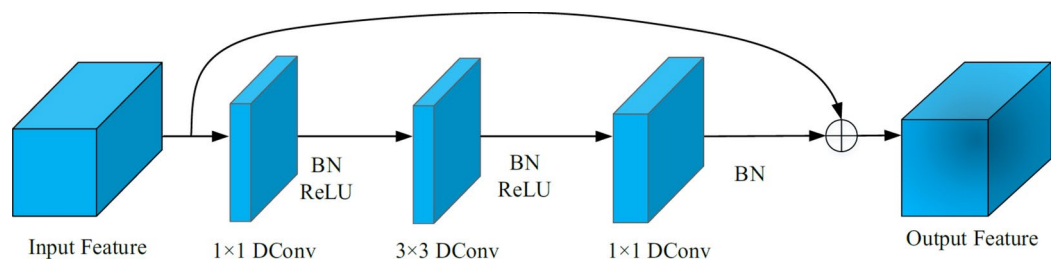
$$y(O_0) = \sum_{O_n \in R} w(O_0) \cdot x(O_0 + O_n + \Delta O_n) \Delta m_k \quad (3)$$

The realization process is shown in Fig. 3.

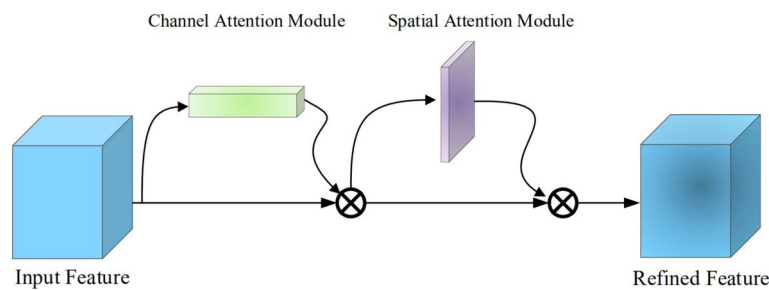
The original ResNet10 model consists of four blocks, and each block consists of several residual modules. In this paper, the ResNet10 model is improved by replacing the standard convolution layer of residual modules



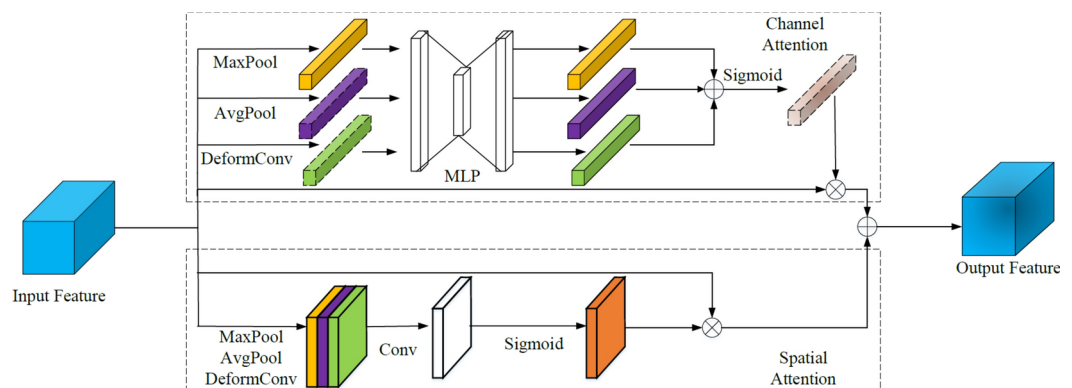
**Fig. 3.** Process of deformable convolution.



**Fig. 4.** Modified residual module.



**Fig. 5.** CBAM structure.



**Fig. 6.** TCCBAM structure.

in the last three blocks with a deformable convolution layer, so as to increase the precision of target feature extraction by the model. The first block is reserved as the standard convolution layer to control the network parameter number brought by deformable convolution. The improved residual module is shown in Fig. 4. Where DConv stands for deformable convolution.

### Residual module based on attention mechanism

The attention mechanism enables the model to select the information in the training process. Convolutional Block Attention Module (CBAM) network is a series combination of channel attention and spatial attention (Fig. 5). In brain CT images, the contrast between background and target is low, and the recognition between different instances can only rely on contour features. Therefore, adding attention mechanism to the network can improve the ability of extracting valid features.

Since CBAM is a serial connection structure, the channel attention module has a certain degree of influence on the features learned by the spatial attention module. In addition, the front-end of the two attention modules uses GAP and GMP to integrate the spatial information of feature mapping, which will lose the accurate spatial relative relationship between image objects. Therefore, the two-channel convolutional block attention module (TCCBAM) is proposed to extract salient feature information, the structure of TCCBAM is displayed in Fig. 6.

In order to avoid interference of channel attention to spatial attention in CBAM serial connection, TCCBAM adopts parallel connection and features fusion of the outputs of the two modules according to elements, thus it is unnecessary to pay attention to the sequence of spatial and channel attention module. In addition, CBAM only focuses on its attention mechanism and does not comprehensively consider the context information of feature



maps. Therefore, the deformable convolution is added to enhance the representation ability of local features, optimize the receptive field and extract rich original information.

It inputs  $f^{ij} \in R^{C \times H \times W}$  through TCCBAM, the output feature graph  $f_{out}^{ij}$  is obtained. Where  $f^{ij}$  is the feature graph of the  $j$ -th residual module in the  $i$ -th block. The calculation process of  $f_{out}^{ij}$  is as follows:

$$f_{out}^{ij} = f^{ij} \otimes M_c^{ij}(f^{ij}) + f^{ij} \otimes M_s^{ij}(f^{ij}) \quad (4)$$

where  $M_c^{ij}(f^{ij})$  is the output of channel attention module. Feature graph  $f^{ij}$  parallel passes global maximum pooling  $F_{cmax}^{ij}$ , global average pooling  $F_{cavg}^{ij}$ , and deformable convolution  $F_{cdeform}^{ij}$  and obtains three feature graphs with size  $C \times 1 \times 1$ . After that, the feature graph enters the full connection layer  $F_c^{ij}$  with the shared weight to fully fuse the information. The output  $M_c^{ij}(f^{ij})$  is obtained by Sigmoid activation function  $\sigma_c^{ij}$  after the weighted three feature graphs:

$$M_c^{ij}(f^{ij}) = \sigma_c^{ij}(F_c^{ij}(F_{cmax}^{ij}(f^{ij})) + F_c^{ij}(F_{cavg}^{ij}(f^{ij})) + F_c^{ij}(F_{cdeform}^{ij}(f^{ij}))) \quad (5)$$

where  $M_s^{ij}(f^{ij})$  is the output. Feature graph  $f^{ij}$  parallel passes global maximum pooling  $F_{smax}^{ij}$ , global average pooling  $F_{savg}^{ij}$ , and deformable convolution  $F_{sdeform}^{ij}$  and obtains three feature graphs with size  $1 \times H \times W$ .

. Then, the three feature maps are spliced in the channel dimension to form a  $3 \times H \times W$  feature map. After a convolution layer  $F_s^{ij}$  with a convolution kernel size of 7 and Sigmoid activation function  $\sigma_s^{ij}$ , the output  $M_s^{ij}(f^{ij})$  is obtained. The calculation process is as follows:

$$M_s^{ij}(f^{ij}) = \sigma_s^{ij}(F_s^{ij}[F_{smax}^{ij}(f^{ij}); F_{savg}^{ij}(f^{ij}); F_{sdeform}^{ij}(f^{ij})]) \quad (6)$$

After the second  $1 \times 1$  convolution layer of each residual module in ResNet10 model, TCCBAM is inserted to improve the feature extraction ability of ROI.

### Modified loss function

The loss function  $L$  of the Mask RCNN is composed of three parts: location loss  $L_{box}$ , classification loss  $L_{cls}$ , and segmentation loss  $L_{mask}$ :

$$L = \alpha L_{cls} + \beta L_{box} + \gamma L_{mask} \quad (7)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are the weight parameters of  $L_{cls}$ ,  $L_{box}$  and  $L_{mask}$ .

$L_{cls}$  uses the cross drop loss function to achieve multi-task classification, and the  $L_{cls}$  calculation process is as formula (8):

$$L_{cls} = - \sum_{i=1}^{k+1} (p \log(q_i) + (1-p) \log(1-q_i)) \quad (8)$$

where  $k+1$  represents the number of categories plus the background.  $p$  represents the truth value, and  $q_i$  represents the probability of predicting category  $i$ .

$L_{box}$  calculates the bounding box loss by the Smooth function as follows:

$$L_{box} = \sum_{i \in \{x, y, w, h\}} Smooth_{L1} = \begin{cases} 0.5(t_i - v_i)^2, & \text{if } |t_i - v_i| < 1 \\ |t_i - v_i| - 0.5, & \text{otherwise} \end{cases} \quad (9)$$

where  $(t_x, t_y, t_w, t_h)$  represents the parameters of the real boundary box.  $(v_x, v_y, v_w, v_h)$  represents regression prediction parameters corresponding to real classification.

The calculation process of  $L_{mask}$  through binary cross entropy loss function is as follows:

$$L_{mask} = - \sum_y y \log(y) - (1-y) \log(1-y) \quad (10)$$

where  $y$  represents the real label of pixels in the target region.  $(1-y)$  represents the predicted value of pixels in the target region. The original Mask RCNN using  $L_{mask}$  alone will make the edge loss ignored by the information of the whole region, resulting in poor target edge segmentation accuracy, which will affect the accuracy of brain CT images classification. To solve this problem,  $L_{mask}$  of mask output branch is optimized.

In classification process, the model adopts the IOU between the mask real value and predicted value as the evaluation index of quality. The optimal choice between a given optimization index and the proxy loss function is the optimization index. The Dice coefficient is usually used to calculate the similarity of two samples. When it is designed as a loss function, it improves the performance of the model. Therefore, a loss function similar to IOU calculation method is designed, and Laplacian smoothing is added to reduce over-fitting and prevent gradient explosion and disappearance. The loss function is as Eq. (11):

$$L_{la} = 1 - \frac{2 \times (\sum_{i=1}^H \sum_{j=1}^W P(i, j) \times G(i, j)) + 1}{\sum_{i=1}^H \sum_{j=1}^W P(i, j) + \sum_{i=1}^H \sum_{j=1}^W G(i, j) + 1} \quad (11)$$

where, the input feature graph size is  $H \times W$ .  $P(i, j)$  is the predicted value of  $(i, j)$ .  $G(i, j)$  is the true value of  $(i, j)$ .

To ensure stable convergence of the training process, the original binary cross entropy loss function is fused by  $L_{la}$ , and the optimized  $L_{mask}$  is as follows:

$$L_{mask} = \mu \left( - \sum_y y \log(y) + (1 - y) \log(1 - y) \right) + (1 - \mu) L_{la} \quad (12)$$

where  $\mu$  is the weight coefficient of loss. When  $\mu = 0$ ,  $L_{mask} = L_{la}$ . When  $\mu = 1$ ,  $L_{mask}$  is the binary cross entropy loss function  $L_{bce}$ . The loss function of the optimized Mask RCNN is:

$$L = \alpha L_{cls} + \beta L_{box} + \mu L_{bce} + (1 - \mu) L_{la} \quad (13)$$

## Experiments and analysis

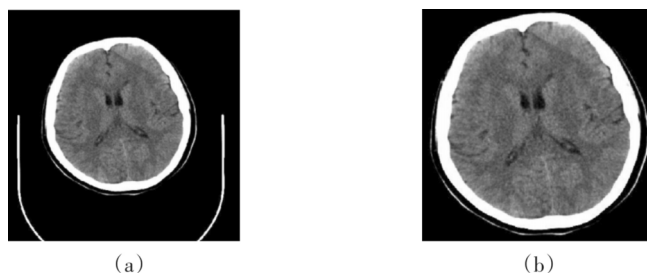
### Image preprocessing

A total of 160 normal and abnormal brain CT images are randomly selected from the labeled database. In addition to the brain regions of interest, there are large areas of un-interest in the original image. And due to the influence of a variety of factors in the shooting process, resulting in dark or bright images, therefore, it is necessary to preprocess each image sample. The specific handling method is as follows. In Matlab 2017a platform, firstly, the original image format (DICOM format) is transformed into BMP format; Secondly, the image details are enhanced by grayscale segmental stretching. Thirdly, Crop Image Tool is used to remove the background area of the original image, so that it contains less background and all the regions of interest. 160 brain images are uniformly output as  $256 \times 256$  pixel. Taking sample No. 1 as an example, its situation before and after treatment is shown in Fig. 7 (a) and b.

### Model training

Because of the same network model, training under different training parameters will get different results. Common training parameters include target error, training step length and learning rate. If the indexes are too high, the over-fitting phenomenon will occur. We has modified the backbone architecture by incorporating deformable convolutions. So we first training the model in used data sets. After training, the result can only identify the trained sample well, and its ability to evaluate the unknown sample is relatively weak, which will increase the error. If the training step length is too long, the convergence time will increase accordingly. On the contrary, if the training indexes are too low, although the required training times are reduced and the time is shortened, the output network testing ability will also be low, and the relationship between various data cannot be well expressed. After many training experiments, training parameters are constantly adjusted. When the maximum iteration number is 1000, the training target error is 0.01 and the learning rate=0.1, the training output network has strong testing ability. The final network training effect is shown in Fig. 8. The training curve basically returns smoothly to reach the specified goal, which shows that the training parameters set are reasonable.

The detailed settings of other parameters in this paper are shown in Table 1. In addition, random horizontal flip, random vertical flip and adjustment of brightness, contrast and saturation are used to enhance the training



**Fig. 7.** Original brain CT sample and preprocessed output.

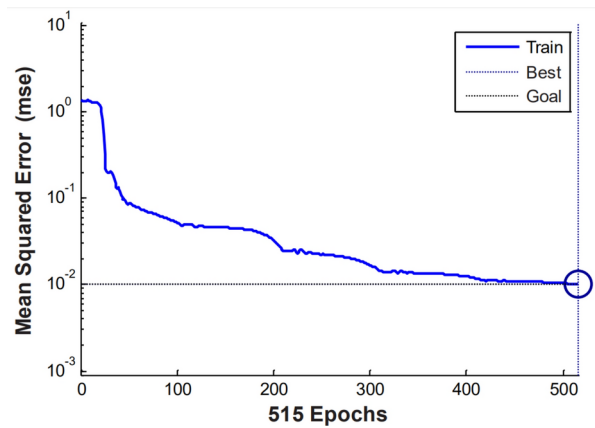


Fig. 8. Network training chart.

Layer	Parameter	Output size
Conv1	7 × 7, 64, stride = 2	100 × 100
Max Pool	3 × 3, stride = 2	50 × 50
Residual Unit 1	3 × 3, 64	50 × 50
Attention model	RHAM	50 × 50
Residual Unit 2	3 × 3, 128	25 × 25
Attention model	RHAM	25 × 25
Residual Unit 3	3 × 3, 256	13 × 13
Attention model	RHAM	13 × 13
Residual Unit 4	3 × 3, 512	7 × 7
Attention model	RHAM	7 × 7
Global average pool	1 × 1	
Dropout	0.5	
FC, softmax	3	

Table 1. Parameters in this paper.

data and improve the generalization ability of the model. All models in this paper are implemented using Pytorch deep learning framework.

All experiments in this paper are built under Windows10 system and Spyder editor. Accelerated computing is performed on Intel(R) Core(TM) i9-9900 K CPU, NVIDIA GeForce RTX2080Ti GPU and 32G memory platform, using Pytorch1.12.1 to build network model and programming language Python3.8 to implement. In this experiment, Adam optimizer is used to calculate and update the network parameters of the optimization model training, and the learning rate and learning rate decay are set to 0.001, the batch size of the model is set to 16, and the input image size is set to 224 × 224.

Experiment and results

In this paper, we adopt accuracy, precision, recall, F1, Kappa and confusion matrix as the evaluation indexes. The calculation formulas are as follows.

Accuracy: The ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set.

accuracy = (TP + TN) / (TP + TN + FP + FN) (14)

Precision: The proportion of the number of positive samples that are correctly classified to the number of samples that the classifier determines to be positive.

Precision = TP / (TP + FP) (15)

The recall rate refers to the proportion of the number of correctly classified positive samples to the number of true positive samples.



Model	Accuracy/%	Precision/%	Recall/%	F1/%	Kappa/%
ResNet50	85.45	75.21	82.55	78.69	76.76
ResNet34	87.38	79.45	82.63	81.06	75.85
ResNet18	89.27	82.57	83.69	83.07	75.97
ResNet10	<b>91.72</b>	<b>88.64</b>	<b>89.52</b>	<b>89.12</b>	<b>77.84</b>

**Table 2.** Classification comparison with different residual network models.

Model	Accuracy/%	Precision/%	Recall/%	F1/%	Kappa/%
DResNet50	87.75	76.32	83.67	79.70	77.87
DResNet34	89.98	80.56	83.74	82.17	76.95
DResNet18	91.23	83.68	84.79	84.18	77.08
DResNet10	<b>93.54</b>	<b>89.75</b>	<b>89.88</b>	<b>89.06</b>	<b>78.95</b>

**Table 3.** Classification comparison with different backbone network models.

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

The F1-Score value is the harmonic average of accuracy and recall.

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \quad (16)$$

The Kappa coefficient is an index used for consistency testing and can also be used to measure the effect of classification.

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (17)$$

$$p_e = \frac{\sum_{k=1}^n (\sum_{i=1}^n x_{ik} \times \sum_{j=1}^n x_{ki})}{(\sum_{j=1}^n \sum_{i=1}^n x_{ij})^2} \quad (18)$$

where TP (true positive) indicates that the label is predicted to be positive. FP (false positive) is a label with negative prediction being positive. TN (true negative) indicates that the label is negative and the prediction is negative. FN (false negative) indicates that the label positive is predicted to be negative.  $p_0$  stands for accuracy.  $n$  stands for category of classification.  $p_e$  represents the sum of the actual and predicted number of products corresponding to all categories in the confusion matrix divided by the square of the total number of samples.  $x_{ij}$  represents the elements of the confusion matrix.

At present, the mainstream residual convolutional neural network models include ResNet10, ResNet18, ResNet34 and ResNet50, and the depth of models increases successively. In order to select models matching the complexity of brain CT data in this paper, the residual network models mentioned above are used to train brain CT image data sets, and the classification results obtained are shown in Table 2. The CT images used in this paper are CQ500 and DeepLesion.

Table 2 shows that ResNet10 has the best classification effect on the brain CT data set, indicating that the model basically matches the complexity of the data set. Therefore, the ResNet10 is selected as the backbone network model in this paper.

Deformable convolution is added into the residual network model to obtain DResNet10, DResNet18, DResNet34 and DResNet50 backbone network models. These backbone network models are used to classify brain CT images respectively, and the classification results are shown in Table 3.

Table 3 shows that although deep convolutional neural network can show good classification effect on large-scale data sets. But for small-scale data sets, it is easy to produce over-fitting phenomenon, which leads to poor classification performance. Therefore, it is very important to design a network model that matches the complexity of the data set.

The proposed TCCBAM and CBAM are added to the above backbone network model, respectively, and classified on brain CT data. The results are shown in Table 4.

By comparing the classification accuracy of the backbone model with the addition of attention module in Table 4, the following two conclusions can be obtained.

- 1 Compared with a single channel domain attention or space domain module, the attention module in this paper is more effective to improve classification results.

Model	Accuracy/%	Precision/%	Recall/%	F1/%	Kappa/%
CBAM-DResNet50	89.38	85.26	81.76	83.41	80.83
TCCBAM-DResNet50	89.73	86.78	87.82	87.22	82.19
CBAM-DResNet34	89.96	87.25	88.93	88.05	82.45
TCCBAM-DResNet34	90.51	88.47	89.61	88.97	83.62
CBAM-DResNet18	90.82	89.67	90.57	80.11	84.75
TCCBAM-DResNet18	91.55	90.63	91.49	91.18	84.98
CBAM-DResNet10	95.37	93.69	94.18	93.85	85.31
TCCBAM-DResNet10	<b>98.06</b>	<b>96.85</b>	<b>97.64</b>	<b>97.38</b>	<b>88.79</b>

Table 4. Classification comparison with different attention models.

Model	Accuracy/%	Precision/%	Recall/%	F1/%	Kappa/%
CDL	88.25	87.69	88.64	88.25	79.63
FACNN	90.36	89.64	90.11	89.87	81.65
MMDL	92.51	91.67	92.33	91.94	84.26
HIE	94.77	93.55	94.67	94.12	88.59
Mask RCNN + TCCBAM-DResNet10	<b>98.06</b>	<b>97.29</b>	<b>98.77</b>	<b>97.85</b>	<b>91.35</b>

Table 5. Classification comparison with different methods.

Normal	1899	6	0	0
NSIP	79	1005	76	326
OP	0	105	1065	11
UIP	129	110	0	575
	Normal	NSIP	OP	UIP
CQ500 data set				
Normal	2107	0	0	0
NSIP	0	992	39	203
OP	0	103	1102	28
UIP	0	171	0	681
	Normal	NSIP	OP	UIP
DeepLesion				

Fig. 9. Confusion matrix with different data.

2 Compared with CBAM attention module, this paper puts forward the TCCBAM attention module using convolution operation to replace the pooling operation. In order to avoid the loss of key information and ensure the stability of the model, the residual skip connection structure is introduced. Compared with CBAM, TCCBAM is more effective in helping the backbone model to improve brain CT classification.

The proposed network model in this paper is applied to brain CT image classification task and compared with other methods including CDL<sup>27</sup>, FACNN<sup>28</sup>, MMDL<sup>29</sup> and HIE<sup>30</sup>. The comparison results are shown in Table 5. By comparing the classification accuracy of models in Table 5, the following two conclusions can be drawn.

- 1 Compared with the classification results of CDL and FACNN, it shows that adding attention mechanism to the common CNN model can effectively improve the performance of the model. Also it verifies that the residual attention module can help model focus on the location and content of brain related organizations, it can extract more distinguish sexual features.
- 2 Compared with the classification results of MMDL and HIE, it shows that adding attention mechanism combined with suppression over-fitting method is more effective in improving the classification effect of the model than only modifying the activation function.

In this section, we plot the confusion matrix of different methods under different data. The result is shown in the Fig. 9. There are four types of data in the data set, usual interstitial pneumonia (UIP), nonspecific interstitial pneumonia (NSIP), organizing pneumonia (OP) and Normal control group. As can be seen from Fig. 9, most lesion data can be effectively detected by the method presented in this paper.

To demonstrate the effectiveness of the improved strategy, the ablation experiment is conducted. (1) Mask RCNN backbone network (2) Replace part of convolutional layer with deformable convolutional (DC) layer in Mask RCNN backbone network (3) Add improved attention mechanism (AM) in Mask RCNN backbone network (4) In Mask RCNN backbone network, part of the convolution layer is replaced by deformable convolution, and an improved attention mechanism is added to each residual module, the results are shown in Table 6.

Because deformable convolution can improve the adaptability of the network to the diversity of geometric features of the target and effectively reduce the interference of background information. The attention mechanism

Model	Accuracy/%	F1
Mask RCNN	86.14	85.47
Mask RCNN + DC	90.55	89.76
Mask RCNN + AM	89.61	88.93
Mask RCNN + + DC+ AM	<b>98.06</b>	<b>97.58</b>

**Table 6.** Ablation experiment results.

can enhance the representation ability of target object features. Therefore, it is feasible to use deformable convolution and improved attention mechanism to optimize the model in this study.

Conclusion

To help doctors diagnose abnormalities in the brain early, we propose a Mask RCNN with attention mechanism method in this paper. Deformable convolution can improve the extraction ability of target features. Embedding residual module based on attention mechanism in the model can make the model pay more attention to target-related features, thus reducing the algorithm’s attention to background information. The adopted binary cross entropy loss function by the model mask branch is improved to better solve the difficulty of imprecise classification. In this experiment, we classify brain CT images based on Mask RCNN and attention mechanism. Although some achievements have been made, there are also some limitations.

First, the performance of the model can be affected by the quality and quantity of data. If there are problems in the data set such as noise, missing values, or data imbalance, it may affect the accuracy and generalization ability of the model.

Secondly, the complexity of the model is high, which requires a lot of computing resources and time to train and optimize. This may not be suitable for some resource-constrained application scenarios.

In addition, although the introduction of attention mechanism improves the model’s focus on key areas, it may also lead to the neglect of other areas, thus affecting the comprehensiveness of the model.

Finally, the performance of the model may also be affected by other factors, such as the resolution of the image, contrast, and so on.

In future studies, we will further explore how to improve data quality and quantity, optimize model structure to reduce computational complexity, and improve attention mechanisms to improve model comprehensiveness and accuracy.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 14 March 2024; Accepted: 31 October 2024  
Published online: 26 November 2024

References

1. Mcbee, M. P. & Wilcox, C. Blockchain Technology: Principles and Applications in Medical Imaging. *J. Digit. Imaging* **33**(2), 726–734 (2020).

2. Ruth, V., Kolditz, D., Steiding, C. & Kalender, W. A. Investigation of spectral performance for single-scan contrast-enhanced breast CT using photon-counting technology: A phantom study. *Medical Physics* **47**(7), 2826–2837 (2020).

3. X. Wang, S. Yin, M. Shafiq, A. A. Laghari, S. Karim, O. Cheikhrouhou, W. Alhakami and H. Hamam, “A New V-Net Convolutional Neural Network Based on Four-Dimensional Hyperchaotic System for Medical Image Encryption,” Security and Communication Networks, vol. 2022, Article ID 4260804, 14 pages, 2022. <https://doi.org/10.1155/2022/4260804>

4. P. Kaur and T. Chaira T, “A novel fuzzy approach for segmenting medical images,” Soft Computing, vol. 25, pp. 3565–3575, 2021.

5. Yin, S. & Li, H. GSAPSO-MQC:medical image encryption based on genetic simulated annealing particle swarm optimization and modified quantum chaos system. *Evolutionary Intelligence* **14**, 1817–1829 (2021).

6. S. Liu S, Y. Xie, A. Jirapatnakul A and A. P. Reeves, “Pulmonary nodule classification in lung cancer screening with three-dimensional convolutional neural networks,” Journal of Medical Imaging, vol. 4, no. 4, 2017.

7. Lu, F., Wu, F., Hu, P., Peng, Z. & Kong, D. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery* **12**(2), 171–182 (2017).

8. B. D. Paul, V. Axel, S. Niklaus, I. Emmanuel and P. John O, “Automatic lesion detection and segmentation of 18F-FET PET in gliomas: A full 3D U-Net convolutional neural network study,” PLoS ONE, vol. 13, no. 4, pp. e0195798, 2018.

9. Charron, O. et al. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Computers in Biology & Medicine* **95**, 43–54 (2018).

10. Ullah, Z. et al. SSMD-UNet: semi-supervised multi-task decoders network for diabetic retinopathy segmentation[J]. *Scientific Reports* **13**(1), 9087 (2023).

11. Ullah, Z. et al. Cascade multiscale residual attention cnns with adaptive roi for automatic brain tumor segmentation[J]. *Information sciences* **608**, 1541–1556 (2022).

12. Ullah, Z., Usman, M. & Gwak, J. MTSS-AAE: Multi-task semi-supervised adversarial autoencoding for COVID-19 detection based on chest X-ray images[J]. *Expert Systems with Applications* **216**, 119475 (2023).

13. Ullah, Z. et al. Densely attention mechanism based network for COVID-19 detection in chest X-rays[J]. *Scientific Reports* **13**(1), 261 (2023).

14. Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging* **35**(5), 1240–1251 (2016).

15. Shi, Q., Yin, S., Wang, K., Teng, L. & Li, H. Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation. *Evolving Systems* <https://doi.org/10.1007/s12530-021-09392-3> (2021).

16. J. A and S. Yin, "A New Feature Fusion Network for Student Behavior Recognition in Education," *Journal of Applied Science and Engineering*, vol. 24, no. 2, pp.133–140, 2021.
17. Y. Moroto, K. Maeda, T. Ogawa and M. Haseyama, "Tensor-Based Emotional Category Classification via Visual Attention-Based Heterogeneous CNN Feature Fusion," *Sensors*, vol. 20, no. 7, pp. 2146, 2020.
18. Gao, T., Li, H. & Yin, S. Adaptive Convolutional Neural Network-based Information Fusion for Facial Expression Recognition. *International Journal of Electronics and Information Engineering* **13**(1), 17–23 (2021).
19. Wang, Y. & Liu, L. Bilinear Residual Attention Networks for Fine-Grained Image Classification. *Laser & Optoelectronics Progress* **57**(12), 121011 (2020).
20. Si, N., Zhang, W., Qu, D., Luo, X. & Niu, T. Spatial-Channel Attention-Based Class Activation Mapping for Interpreting CNN-Based Image Classification Models. *Security and Communication Networks* **2021**, 1–13 (2021).
21. Woo, S., Park, J., Lee, J., Kweon, I. & "CBAM: Convolutional Block Attention Module", ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, 11211, pp. 3–19., Springer. Cham [https://doi.org/10.1007/978-3-030-01234-2\\_1](https://doi.org/10.1007/978-3-030-01234-2_1) (2018).
22. X. W. Gao and R. Hui, "A deep learning based approach to classification of CT brain images," 2016 SAI Computing Conference (SAI), pp. 28–31, 2016. <https://doi.org/10.1109/SAI.2016.7555958>.
23. Mzoughi, H. et al. Deep Multi-Scale 3D Convolutional Neural Network (CNN) for MRI Gliomas Brain Tumor Classification. *Journal of Digital Imaging* **33**, 903–915 (2020).
24. S. Yin, H. Li, and L. Teng, "Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images," *Sensing and Imaging*, vol. 21, 2020. <https://doi.org/10.1007/s11220-020-00314-2>
25. Yin, S., Li, H., Liu, D. & Karim, S. Active Contour Modal Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation. *Multimedia Tools and Applications* **79**, 31049–31068 (2020).
26. Gurita, A. & Mocanu, I. Image Segmentation Using Encoder-Decoder with Deformable Convolutions. *Sensors* **21**(5), 1570 (2021).
27. X. Gu, Z. Shen, J. Xue, et al., "Brain Tumor MR Image Classification Using Convolutional Dictionary Learning With Local Constraint," *Frontiers in Neuroscience*, vol. 15, 2021.
28. Bacanin, N., Bezdan, T., Venkatachalam, K. & Al-Turjman, F. Optimized convolutional neural network by firefly algorithm for magnetic resonance image classification of glioma brain tumor grade. *Journal of Real-Time Image Processing* **18**, 1085–1098 (2021).
29. Rajasree, R., Columbus, C. & Shilaja, C. Multiscale-based multimodal image classification of brain tumor using deep learning method. *Neural Computing and Applications* **33**(6), 5543–5553 (2021).
30. Ullah, Z., Farooq, M., Lee, S. & An, D. A Hybrid Image Enhancement Based Brain MRI Images Classification Technique. *Medical Hypotheses* **143**, 109922 (2020).

## Acknowledgements

The study was funded by JDRF (1-SRA-2020-973-S-B), Science for Life Laboratory, the Swedish Research Council (2020-02312; OE, 2019-05115; JL, 2019-01415; OK), the Swedish Cancer Society (CAN 2017/649, 20 1090 PjF; JL), Vinnova (2019/00104; JL), Novo Nordisk Foundation, an EFSD/Novo Nordisk grant, the Ernfors Family Fund, Barndiabetesfonden, Diabetesfonden, the Sten A Olssons Foundation, Helmsley Charitable Trust, the Juvenile Diabetes Foundation International. The Preclinical PET/MRI Platform (PPP) and Sofie Ingvast are acknowledged for expert technical assistance. This work was supported by the Slovak Research and Development Agency under the contract No. APVV 19-0581.

## Author contributions

All authors reviewed the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.Y., H.L., L.T., A.A.L. or M.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025