International Journal of Computer Assisted Radiology and Surgery (2022) 17:1235–1243
https://doi.org/10.1007/s11548-022-02661-9

ORIGINAL ARTICLE

Classification of Alzheimer's disease in MRI images using knowledge distillation framework: an investigation

Yiru Li[1] · Jianxu Luo[1] · Jiachen Zhang[1]

Received: 21 October 2021 / Accepted: 26 April 2022 / Published online: 28 May 2022

Abstract

Purpose Computer-aided MRI analysis is helpful for early detection of Alzheimer's disease(AD). Recently, 3D convolutional neural networks(CNN) are widely used to analyse MRI images. However, 3D CNN requires huge memory cost. In this paper, we introduce cascaded CNN and long and short-term memory (LSTM) networks. We also use knowledge distillation to improve the accuracy of the model using small medical image dataset.
Methods We propose a cascade structure, CNN-LSTM. CNN is used as the function of feature extraction, and LSTM is used as the classifier. In this way, the correlation between different slices can be considered and the calculation cost caused by 3D data can be reduced. To overcome the problem of limited image training data, transfer learning is a more reasonable way of feature extraction. We use the knowledge distillation algorithm to improve the performance of student models for AD diagnosis through a powerful teacher model to guide the work of student models.
Results The accuracy of the proposed model is improved using knowledge distillation. The results show that the accuracy of the student models reached 85.96% after the guidance of the teacher models, an increase by 3.83%.
Conclusion We propose cascaded CNN-LSTM to classify 3D ADNI data, and use knowledge distillation to improve the model accuracy when trained with small size dataset. It can process 3D data efficiently as well as reduce the computational cost.

Keywords Alzheimer disease · MRI · Small samples · Deep learning · Classification · Knowledge distillation

Introduction

Alzheimer Disease (AD) is a progressive neurodegenerative disease characterized by cognitive decline and memory loss. This disease causes irreversible damage to the brain and eventually leads to the death of individuals due to complete brain failure. It is also one of the leading causes of death in the aging population which affects nearly 50 million people worldwide [1], and the socio-economic cost of this is enormous. Due to the irreversibility of AD, early diagnosis plays a critical role in helping to mitigate disease progression. Currently, researchers are using advanced neuroimaging techniques, such as magnetic resonance imaging (MRI), to identify AD. MRI can produce 3D images, which have millions of voxels. Figure 1 shows three slices of patients' scans in different directions.

In medical imaging diagnosis of Alzheimer's disease, lesions are mainly found in MRI images, which are often judged with the help of the physician's experience. Digital image processing technology is used to achieve reconstruction and measurement of the brain, soft tissues and lesions. With the help of computers, doctors can analyze lesions and other areas of interest qualitatively and even quantitatively. AI-enabled medical system can assist doctors and improve the accuracy and reliability of judging lesions. Deep learning is the main technical tool.

With the latest advances in deep learning, convolutional neural networks have good performance in the classification

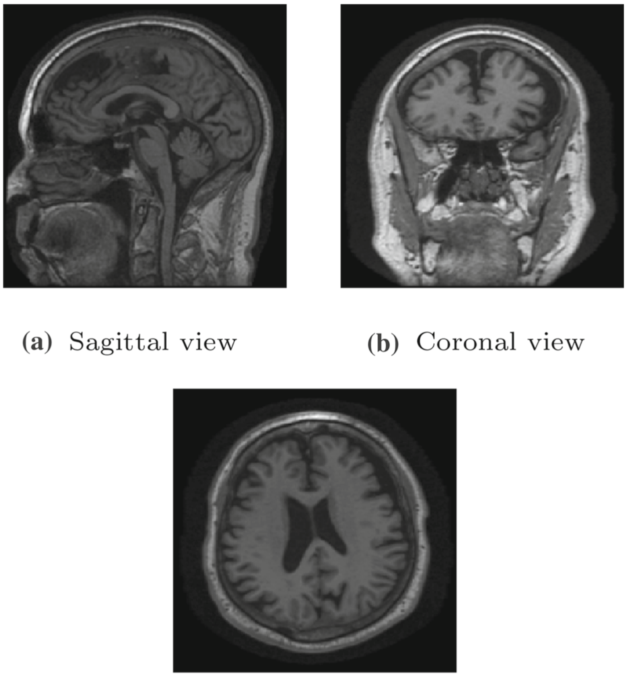
Yiru Li and Jiachen Zhang have contributed equally to this work.

✉ Jianxu Luo
jxluo@ecust.edu.cn

Yiru Li
837806012@qq.com

Jiachen Zhang
y30190768@mail.ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China



(a) Sagittal view

(b) Coronal view

(c) Axial view

Fig. 1 Images of MRI sample slices of participants: three views

of natural images and show great potential in medical image diagnosis. Various CNN-based techniques have been proposed for the classification and segmentation of Alzheimer's disease. The classification of AD can be done by studying its underlying features. Thus, feature extraction has an important role in the classification of medical images. Earlier studies on MRI images analysis were performed by extracting the high-level features then going through support vector machine (SVM) [2] and random forest [3] for classification. Manu et al. [4] first used convolutional neural networks for feature extraction and then connected fully connected layers to classify Alzheimer's disease. Ali et al. [5] processed 3D data to obtain slices in three directions and then input them to CNN. Xin et al. [6] transformed 3D MRI image volumes into 2D images as input to CNN. Marcia et al. [7] proposed an intelligent entropy-based technique for selecting training datasets to get a larger amounts of information in a small sample. Hao et al. [8], used transfer learning combined with active learning to classify brain tumor MRI images. Korolev et al. [9] took 3D medical image data as input and used a 3D convolutional neural network for feature extraction. Khvostikov et al. [10] combined multimodal information such as sMRI and PET and proposed a data augmentation method to balance categories of different sizes, enriching the input information of the 3D model and thus improving the impact of the model on the classification results. Huang [11] proposed a multimodal diagnostic system based on T1-MRI and FDG-PET, and only hippocampal areas were used as ROIs.

Due to the volumetric nature that MRI images have, there is a consensus among researchers that slice to slice correlation information should be considered. However, if deep learning 3D models are used, the computational cost of 3D models is higher and the training time is longer due to the high dimensionality of the input. Using a cascade structure enables the network to extract features in each 2D image and then learn slice-to-slice features.

Another problem is that most medical datasets are relatively small at present, with insufficient sample data. When the training sample is not enough, the network does not learn more advanced features, then the model tends to overfit. To overcome the problem of limited image training data, transfer learning is a reasonable way of feature extraction. Our work is mainly as follows:

(1) We propose a cascade structure, CNN-LSTM. CNN is used as the function of feature extraction, and LSTM is used as the classifier. In this way, the correlation between different slices can be considered and the calculation cost caused by 3D data can be reduced.

(2) We use the knowledge distillation algorithm to improve the performance of student models for AD diagnosis through a powerful teacher model to guide the work of student models.

Related work

In recent years, with increasing computational power, deep learning methods have developed extremely rapidly. Among them, convolutional neural networks are widely used in the field of medical imaging analysis since it can automatically learn the feature representation of an image. In order to get a powerful CNN model, a large number of samples are needed for training. However, labeled medical images are usually hard to get , resulting a small data set for CNN training. Transfer learning [12] is an efficient way to deal with small samples. It is the process of transferring the parameters of a trained model (pre-trained model) to a new model to help training the new model. Considering that most of the data or tasks are correlated, transfer learning allows us to share the learned model parameters to the new model through transfer to speed up and optimize the learning efficiency of the model, without learning from scratch as most networks do. It can take advantage of the similarity between the source and target domains to have a good classification effect despite using a small number of samples in the target's task. Pan and Yang [13] showed that the more similar the data distribution between the source and target domains, the closer the learned information is, the better the migration effect will

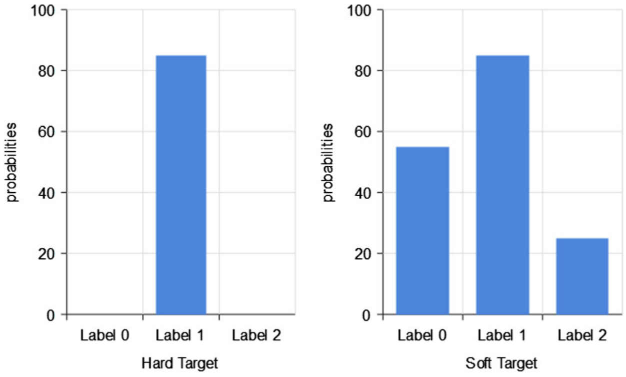


Fig. 2 Information from soft targets and hard targets

be. One such method, knowledge distillation, is a combination of transfer learning and model compression, which is a way to transfer dark knowledge by teaching an untrained student model by a pre-trained teacher model so that the student model also performs well under that task.

Teacher models used softmax to output the probability S_i belong to each class and the logit of knowledge distillation represented the probability prediction value of model output for each class, see Eq. (1)

$$S_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \qquad q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1}$$

where q_i is a soft label for student network learning, z_i is the output probability for each class and T is the hyperparameter representing the distillation temperature. When T is taken as 1, the equation degenerates into softmax, and the probability of each class is output according to logit; when T is close to 0, then it approximates one-hot encoding; if T is lager, the distribution of the output results will be flatter, which serves to retain similar information.

Suppose we have a teacher model with strong generalization ability, we can use the teacher model distillation to train the student model directly to learn the generalization ability of the teacher model. A straightforward approach is to use the probabilities of the output categories of the “Softmax” layer as “soft labels”. The advantage of this is that in addition to the output of positive examples, the negative labels also carry a lot of information. In contrast to the traditional “hard labels”, all one-hot negative labels are treated uniformly. Figure 2 illustrates the probability of getting different labels after softening the hard labels, which enriches the information brought by the labels.

The loss function of Teacher-Student Network consists of two parts as Eq. (2) :

$$\text{Loss} = (1-\alpha) * \text{loss}_{\text{hard}} + \alpha * \text{loss}_{\text{soft}} \tag{2}$$

$$\text{loss}_{\text{hard}} = -\sum_{j}^{N} c_j^T \log(q_j^T) \tag{3}$$

$$\text{loss}_{\text{soft}} = \sum_{j}^{N} p_j^T \log\left(\frac{p_j^T}{q_j^T}\right) \tag{4}$$

where $\text{loss}_{\text{hard}}$ represents the cross entropy of real labels and student model predictions. Using real labels can effectively reduce the possibility of errors being propagated to student models. $\text{loss}_{\text{soft}}$ represents the relative entropy of soft label prediction of teacher and student models. Students are required to imitate the teacher model's learning ability as much as possible. α represents the weight to balance the relationship between the two. c_j is ground truth, q_j is the output of student model and p_j is the output of teacher model.

Materials and methods

In this section, the dataset and the data preprocessing steps are introduced. Then, the model architecture and it's super parameters are discussed.

Dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and follow-up of Alzheimer's disease. We have evaluated our system with MRI data from ADNI because it is the largest publicly available dataset on Alzheimer's Disease. We use 280 samples with Alzheimer's disease, 249 samples with mild cognitive impairment, and 251 cognitively normal samples from the dataset. All samples were pre-treated with GradWarp, (a correction of image geometry distortion due to gradient model), B1 Correction (a correction that uses B1 calibration scans to correct image intensity nonuniformity), and N3 (a processing is the N3 histogram peak sharpening algorithm to reduce intensity non-uniformity of images). To ensure the consistency of the samples, the dimensionality of each data was standardized from the original data dimension to 20*224*224 dimensions. Since hippocampal volume is a good predictor for classifying AD, the dimension reduction was made by scaling rather than cropping, avoiding hippocampi information was lost in the reduced data.

We selected T2 sequences from 780 samples as the input sequences. Middle slices are selected from each sample, and each slice is taken one at a time. As shown in Fig. 3, a total of 20 slices are selected, and the size of each slice is 224*224. We set the sample with Alzheimer's disease as Label 0, the sample with mild cognitive impairment as Label 1, and the

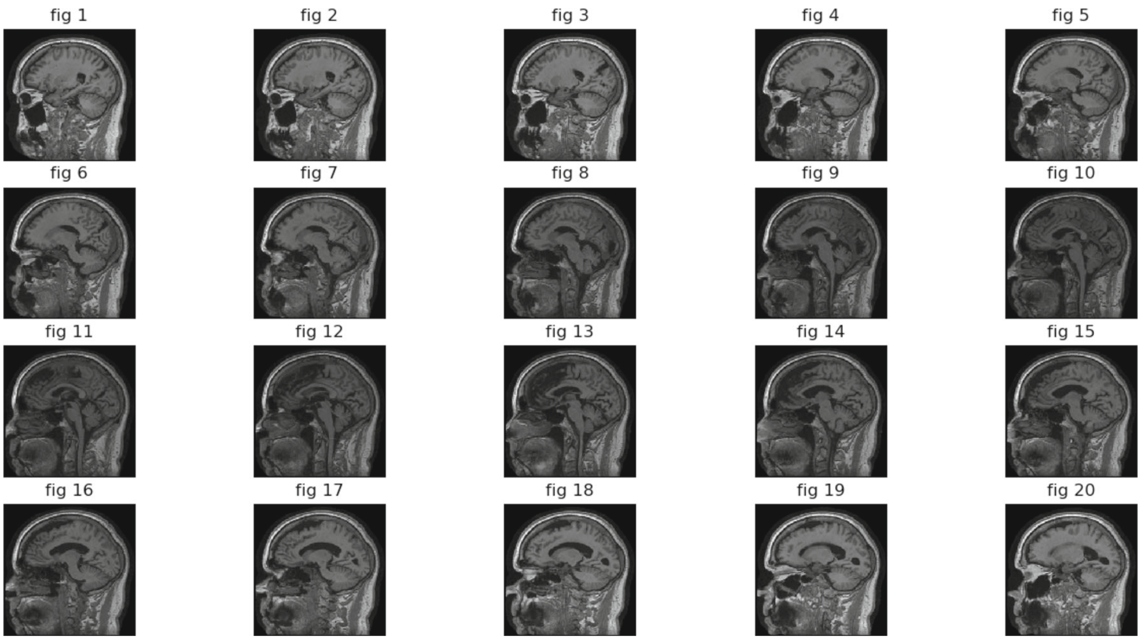


Fig. 3 An input with volume samples. Take 20 slices in the middle of the whole sample to learn the deep image features

Table 1 Number of samples for specific data distribution

	Train (70%)	Test (30%)	All
AD	196	84	280
MCI	174	75	249
CN	175	76	251
All	545	235	780

Table 2 Statistical distribution of demographic information

	Gender		Age groups				
	Male	Female	50–60	60–70	70–80	80–90	90–100
AD	137	143	6	30	138	100	6
MCI	108	141	5	58	105	81	0
CN	110	141	0	0	142	109	0

sample with normal cognition as Label 2. Hence, it's a triple classification problem. We use 70% of the total as training and the remaining 30% as testing. The specific data distribution is shown in Table 1. We show the different classes of gender and age groups in Table 2.

Model architecture

For classifying 3D-MRI images, our method consists of two parts, a pre-trained CNN and an LSTM. The overall structure is sketched in Fig. 4. CNN is used for feature extraction, while the fully-connected (FC) layer is the final layer of the CNN, which transforms the extracted features into a vector of feature sequences. LSTM acts as a classifier.

We use DenseNet [14] and ResNet [15] as the backbone network for feature extraction. Parameter quantity for different models are shown in Table 3.

In the teacher-student network, the teacher model uses the DenseNet model, which uses densely connected blocks, as shown in Fig. 5, and each of its layers is connected to the layers that follow in the forward model. Unlike ResNet, which is add a skip-connection, DenseNet uses direct connection of all inputs to the output layer. For each layer, its input then

Fig. 4 Cascade structure of a CNN-LSTM model

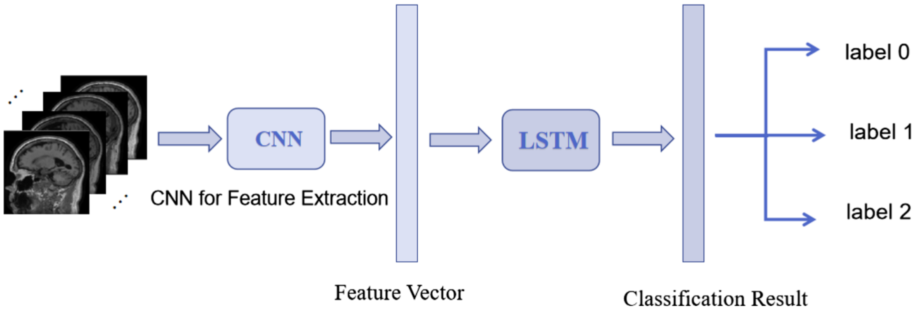


Table 3 Number of parameters of the backbone network we used

Backbone	Params.
DenseNet-169	28,682,000
DenseNet-121	7,978,856
ResNet-50	25,557,032
ResNet-18	11,689,512

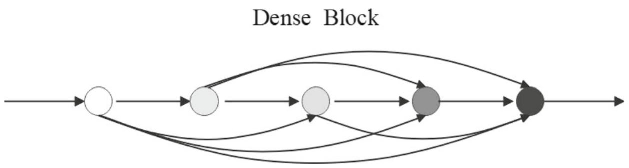


Fig. 5 Dense Block connection diagram. This layer is connected with all subsequent layers in a forward mode

includes the output of the previous layer and the input of all layers before that layer. This connection improves the flow of information and gradients throughout the network, and each layer has access to the gradients from the loss function and the original input signal, making it easier to train the model, and also enhancing the transfer of features. In addition, another advantage of DenseNet is that the network is narrower and has fewer parameters. The growth rate k is a hyperparameter in the network and usually a smaller value of *k* (e.g. *k*=32) has a better result [14]. The network structure of DenseNet is shown in Fig. 6.

The student models use the DenseNet and the ResNet, respectively. The structure of ResNet is shown in Fig. 7a. It consists of two types of network modules, BasicBlock and BottlectBlock, as shown in Fig. 7b and c. They have shortcut connections, which can overcome the gradient disappearance problem caused by deep learning.

LSTM is a special kind of recurrent neural network (RNN) that takes sequence data as input and recurses in the direction of sequence evolution and all nodes are connected in a chain. It is very effective for data with sequential characteristics, and it can mine the temporal and semantic information in the data. LSTM is an extended version of RNN with three gates, i.e., input gate, output gate, and forgetting gate, as shown in Fig. 8. It has two layers of hidden layer neural network, where each layer has 256 nodes. Sequence prediction is performed by LSTM, which uses these gates to learn long-term dependencies between different slices. The LSTM is also effective in dealing with the problem of gradients disappearing during propagation, as it releases some memory that are not helpful

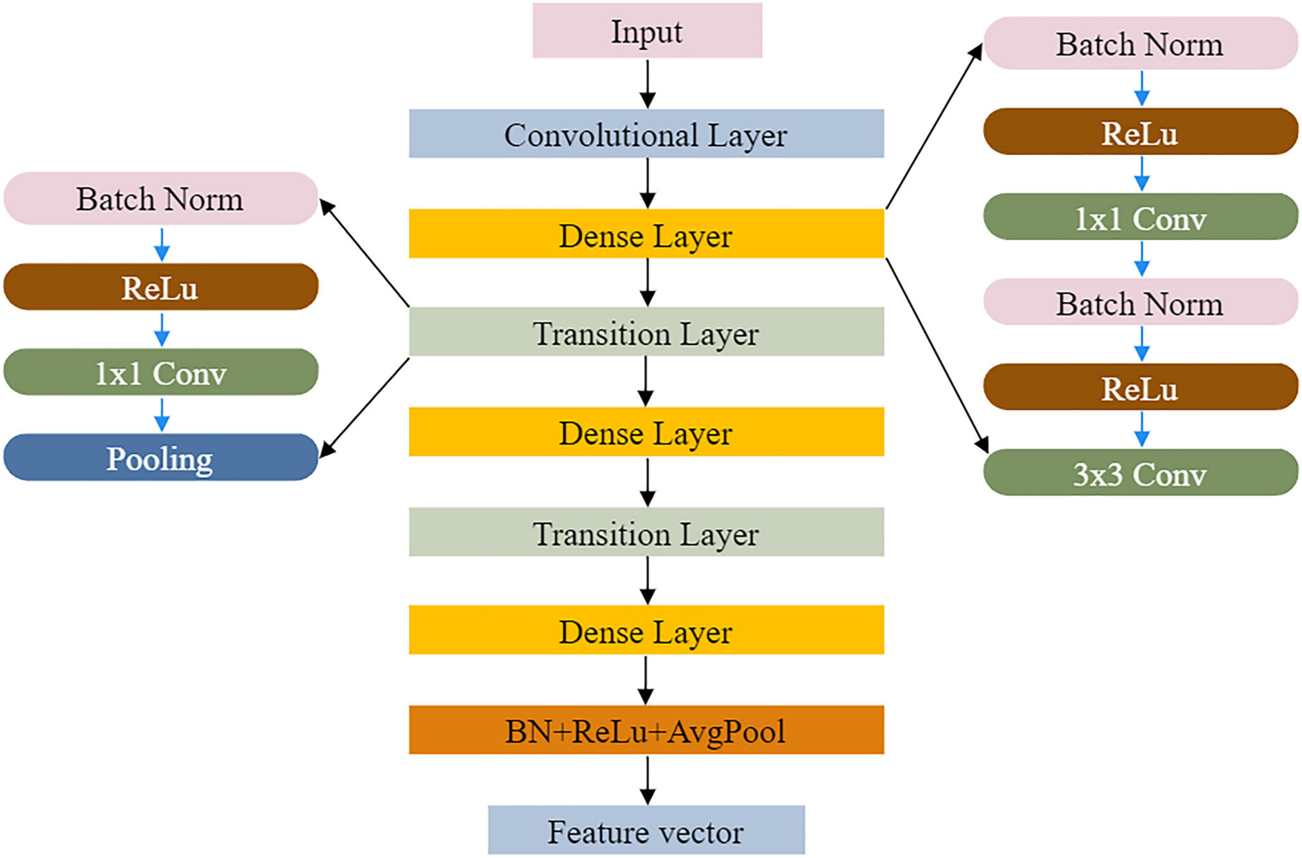
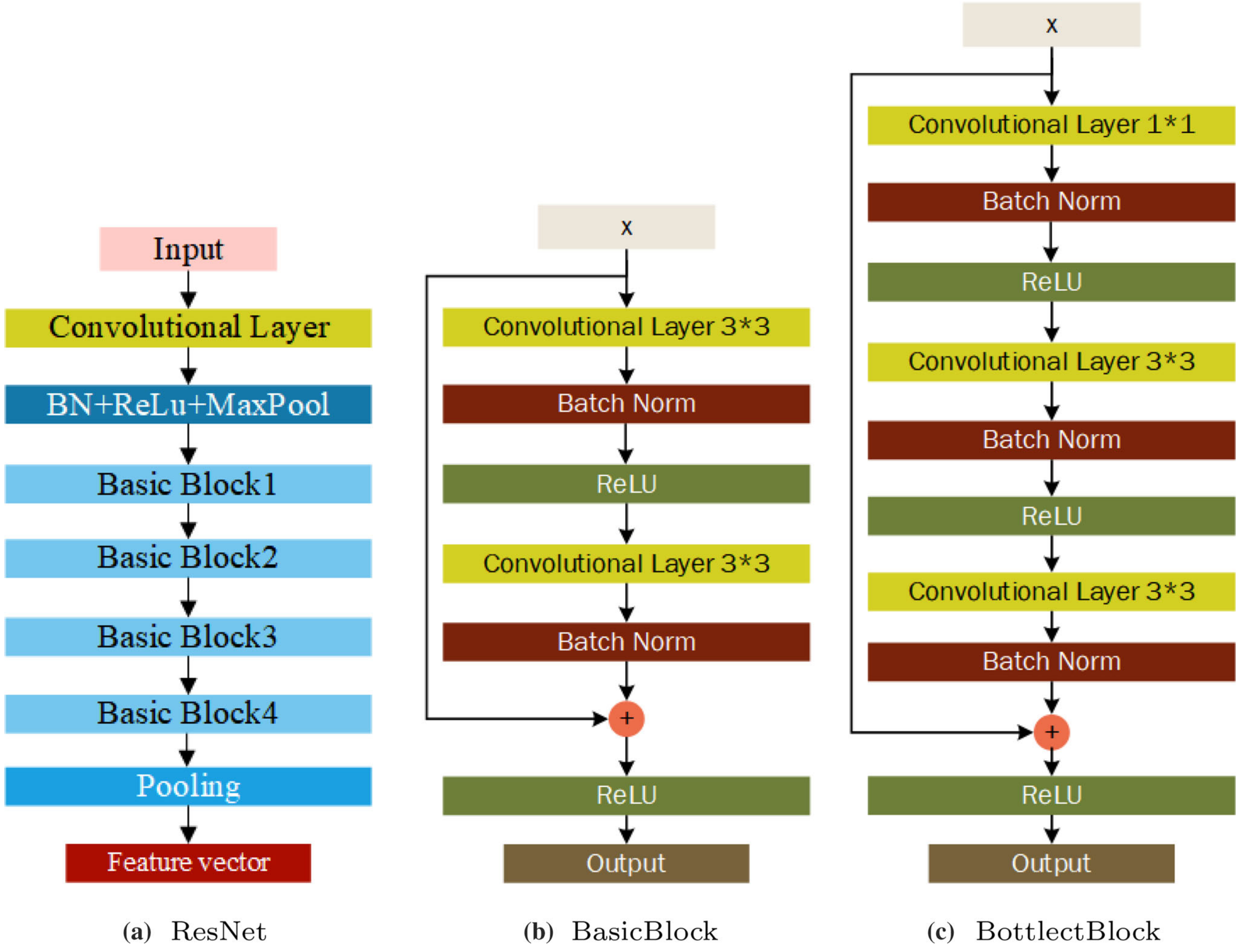


Fig. 6 Structure diagram of DenseNet. It consists of convolutional layer, dense layer and transition layer



(a) ResNet (b) BasicBlock (c) BottlectBlock

Fig. 7 Network structure of ResNet and residual blocks. **a** Acts as the backbone network. The right figure shows the two ways of residual blocks. **b** Is BasicBlock, which consists of two 3*3 convolution blocks; **c** Is BottlectBlock, which consists of one 1*1 and two 3*3 convolution blocks

for predictive classification. Its input is a vector of feature dimensions extracted by the CNN. Each patient slice has a feature vector. Due to the sequence prediction property of LSTM, the last feature information learned by the network will contain information from the shallow layer of the network.

Network and training parameters

Our backbone network uses a fine-tuning learning rate of 0.000015. LSTM use a learning rate of 0.001. A cosine learning rate attenuation mechanism is used to update the learning rate. Adam is used as an optimizer, combining the advantages of both AdaGrad and RMSProp optimization algorithms. The random deactivation value is 0.3, and the 300 dimensional feature vector extracted by CNN is used as the input of LSTM. LSTM has two hidden layers, each layer has 256 neurons. The model is implemented by PyTorch 3.7 and trained on NIVIDA RTX 3090. CPU is i9 10980XE.

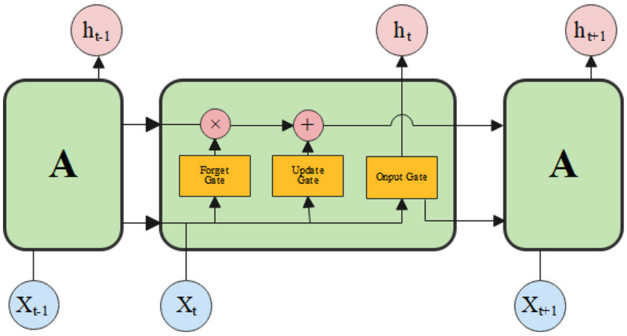


Fig. 8 Structure diagram of LSTM. Input gate remembers part of the present information and then sends the part memories to the output gate together with the present memories. Forgetting gate is to forget part of the past information. Output gate determines the final output

Table 4 Results of the accuracy of CNN-LSTM models with different structures on MRI images

Teacher	DenseNet169	82.12%	
Student	DenseNet121	ResNet18	ResNet50
KD	85.96%	79.57%	75.32%
No-KD	82.13%	71.06%	74.89%

Experimental results and analysis

Result

Our main goal was to investigate the difference between the performance of student models with the guidance of a teacher model, comparing the effect of model size on model accuracy and the effect of having a teacher model for guidance on model accuracy.

DenseNet169 is chosen as the teacher network and three different networks, DenseNet121, ResNet50, and ResNet18, are selected to be student models, respectively. The performances of student models with or without knowledge distillation are shown in Table 4. It can be seen that using the teacher's guidance, the accuracy of DenseNet121 increased by 3.83%, and that of ResNet18 increased by 8.5%, while ResNet50 increased by 0.43%.

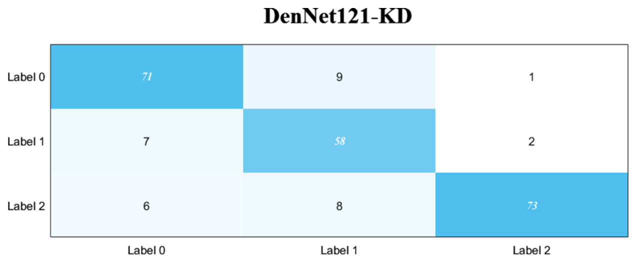
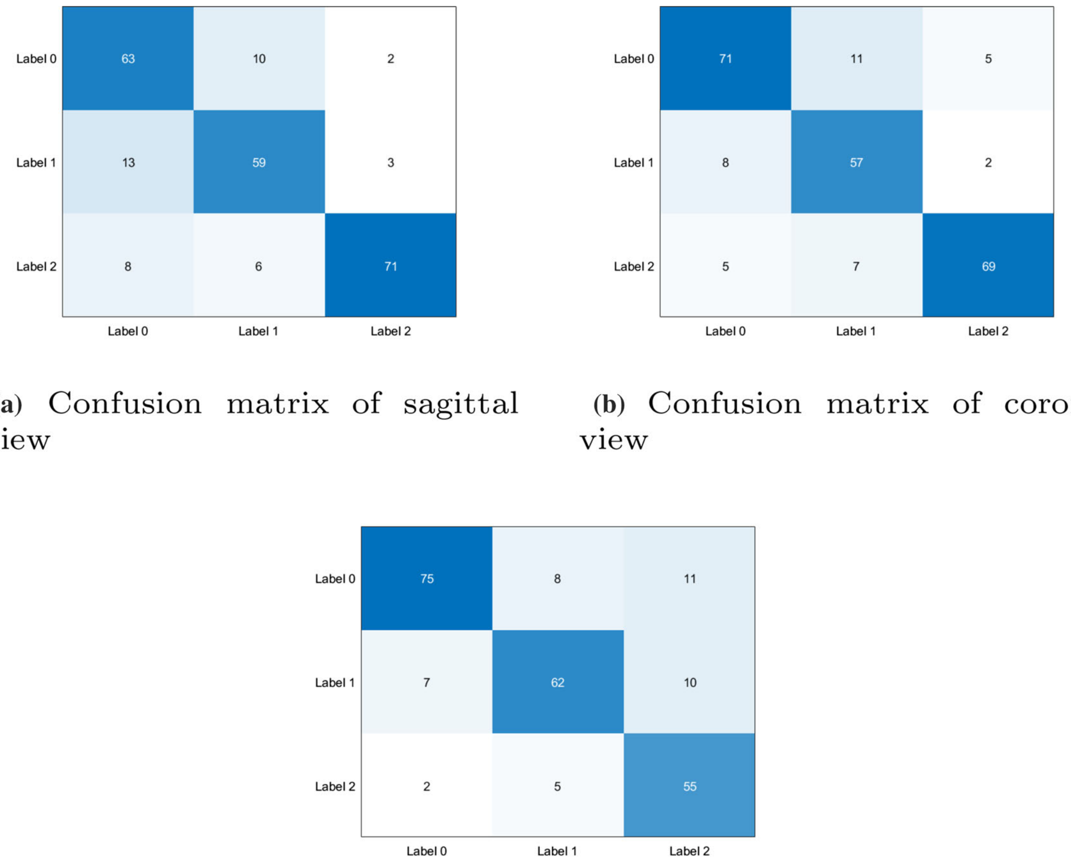


Fig. 10 Confusion matrix for DenNet121 on the test set after knowledge distillation

In Fig. 9, we also noted that our attempts at other 2D slices for view selection (e.g., sagittal, coronal and axial) did not result in significant differences in test accuracy, as long as the slices were selected close to the center of the brain.



(a) Confusion matrix of sagittal view

(b) Confusion matrix of coronal view

(c) Confusion matrix of axial view

Fig. 9 Confusion matrix for three different views of MRI. The three confusion matrices are represented as DenseNet169 as the backbone network and different views as the inputs to the network: **a** Sagittal view; **b** Coronal view; **c** Axial view

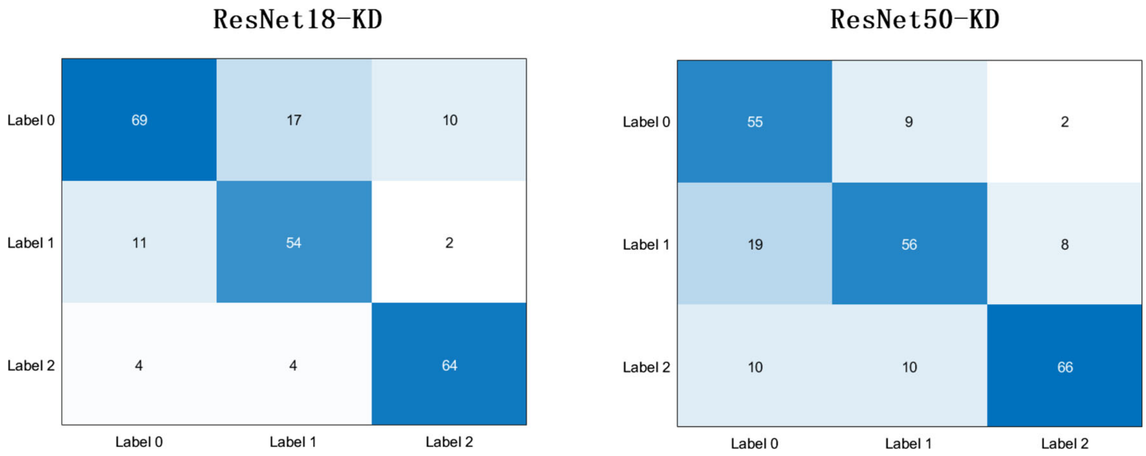


Fig. 11 Confusion matrix of knowledge distillation on test set, with data split by patients

The three accuracy rates are: 82.12, 83.83, 81.70%. We also tried slices from all three views as three-channel inputs to the model, and the improvement was little. Performance was relatively improved with teacher model guidance, but the performance was still limited. We think it was over-fitting information from individual patients rather than general differences in the different stages of the brain in Alzheimer's disease [16].

Analysis

From Table 4 we can see the accuracy of predictive classification using DenseNet169 as the teacher network was 82.12%, while the result of transfer learning using DenseNet121 as the student network was 85.96%. The results of the student network are better than the teacher network. We believe it is due to the small size of parameters of the student model, which can fit better for small sample set. The fit is better on small samples, and the process of knowledge distillation enriches the dark knowledge from supervised learning, which allows the student model to learn more information to improve the accuracy of predictive classification. The confusion matrix of DenNet121 is shown in Fig. 10. It shows that Label 1 has the lowest accuracy and is often misclassified into Label 0 and Label 2. Label 1 was considered to be a MCI sample, between AD and CN, with less distinctive feature. Under the guidance of the teacher model DenseNet169, the accuracy of ResNet18 is 4.25% better than that of ResNet50. We think it is the fact that ResNet18 has fewer parameters and can better fit a small number of medical samples. Through the guidance of teachers' model, it can have a significant improvement. The effect of ResNet50 itself is good, and the space for improvement is relatively insignificant. The confusion matrix is shown in Fig. 11.

Conclusion and future work

In this work, we propose cascaded CNN-LSTM to classify 3D ADNI data, and use knowledge distillation to improve the model accuracy in small samples. Instead of using 3D network, it performs sequence prediction classification on 2D data. This method can effectively distinguish the classification situation. Another advantage is that it can process 3D data efficiently as well as reducing the computing pressure on the computer from the data. In future work, we can consider adding an attention mechanism to the backbone for feature extraction, which enables the model to notice changes in the hippocampus. Thus, more characteristic features can be extracted to further improve the accuracy of diagnosis.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Patterson C(2018) The state of the art of dementia research: new frontiers. World Alzheimer Report
2. Wasule V, Sonar P (2017) Classification of brain mri using svm and knn classifier. In: 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), pp. 218–223 . IEEE
3. Moore P, Lyons T, Gallacher J (2019) Random forest prediction of Alzheimer's disease using pairwise selection from time series data. PLOS One 14(2):1–14

4. Subramoniam M, Aparna T, Anurenjan P, Sreeni K (2022)Deep learning-based prediction of alzheimer's disease from magnetic resonance images, 145–151
5. Nawaz A, Anwar S.M, Liaqat R, Iqbal J, Bagci U, Majid M Deep convolutional neural network based classification of alzheimer's disease using mri data. In: 2020 IEEE 23rd International Multitopic Conference (INMIC), 1–6 (2020). IEEE
6. Xing X, Liang G, Blanton H, Rafique M.U, Wang C, Lin A.-L, Jacobs N (2020) Dynamic image for 3d mri image alzheimer's disease classification. In: European Conference on Computer Vision, pp. 355–364. Springer
7. Hon M, Khan NM (2017) Towards alzheimer's disease classification through transfer learning. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1166–1169. IEEE
8. Hao R, Namdar K, Liu L, Khalvati F (2021) A transfer learning-based active learning framework for brain tumor classification. Front Artif Intell 4:635766
9. Korolev S, Safiullin A, Belyaev M, Dodonova Y Residual and plain convolutional neural networks for 3d brain mri classification. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 835–838 (2017). IEEE
10. Khvostikov A, Aderghal K, Benois-Pineau J, Krylov A, Catheline G 3d cnn-based classification using smri and md-dti images for alzheimer disease studies. 2018. In: Computer Vision and Pattern Recognition
11. Huang Y, Xu J, Zhou Y, Tong T, Zhuang X (2019) Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. Front Neurosci 13:509–509
12. Yosinski J, Clune J, Bengio Y, Lipson H How transferable are features in deep neural networks? Advances in neural information processing systems (NIPS) (2014)
13. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Tran Knowl Data Eng 22(10):1345–1359
14. Huang G, Liu Z, Van Der Maaten L, Weinberger K.Q Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778
16. Fung Y.R, Guan Z, Kumar R, Wu J.Y, Fiterau M Alzheimer's disease brain mri classification: Challenges and insights. http://arxiv.org/abs/1906.04231 (2019)

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