Computer-Aided Diagnosis of Alzheimer's Disease via Deep Learning Models and Radiomics Method

Yin Dai [1,2,*], **Wenhe Bai** [1,2], **Zheng Tang** [1], **Zian Xu** [1,2] **and Weibing Chen** [1,2]

1   College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China; 2071205@stu.neu.edu.cn (W.B.); 20175667@stu.neu.edu.cn (Z.T.); 2171272@stu.neu.edu.cn (Z.X.); 20185488@stu.neu.edu.cn (W.C.)
2   Engineering Center on Medical Imaging and Intelligent Analysis, Ministry Education, Northeastern University, Shenyang 110169, China
*   Correspondence: daiyin@bmie.neu.edu.cn; Tel.: +86-139-4024-3520

Abstract: This paper focused on the problem of diagnosis of Alzheimer's disease via the combination of deep learning and radiomics methods. We proposed a classification model for Alzheimer's disease diagnosis based on improved convolution neural network models and image fusion method and compared it with existing network models. We collected 182 patients in the ADNI and PPMI database to classify Alzheimer's disease, and reached 0.906 AUC in training with single modality images, and 0.941 AUC in training with fusion images. This proved the proposed method has better performance in the fusion images. The research may promote the application of multimodal images in the diagnosis of Alzheimer's disease. Fusion images dataset based on multi-modality images has higher diagnosis accuracy than single modality images dataset. Deep learning methods and radiomics significantly improve the diagnosing accuracy of Alzheimer's disease diagnosis.

Keywords: image classification; muti-modality fusion; neural network

1. Introduction

With the aging problem aggravating in China, the prevalence of Alzheimer's disease (AD) increases year by year. Diagnosis of Alzheimer's disease is critical, and multimodal diagnosis is an indispensable diagnostic method for Alzheimer's disease [1]. Multimodal image fusion techniques combine different modalities of medical images into a single image that can contain more information. Due to the lack of specific diagnostic features, convolution neural networks are trained to assist in the diagnosis of Alzheimer's disease [2].

1.1. Alzheimer's Disease
1.1.1. Status of Alzheimer's Disease

Alzheimer's disease, commonly known as senile dementia, is an insidious onset cognitive and behavioral disorder, of which the morbidity rate keeps increasing. Alzheimer's disease is one of the most common types of dementia [3]. Figure 1 shows a comparison between the normal brain and the brain with Alzheimer's disease. The main clinical symptoms of Alzheimer's disease are psychiatric symptoms and behavioral disorders including progressive memory loss, cognitive impairment, etc. [4]. Alzheimer's disease has been discovered for more than a century, but the current research institutions still lack diagnosis technology and effective treatment to it. Consequently, Alzheimer's disease is treated as a global problem [5]. The pathogenesis of Alzheimer's disease is very complicated, which seriously infects the elder's life quality, and also brings a heavy burden to the family and society. Hence, the diagnosis is of great significance in delaying the process of disease and improving the quality of life. In Figure 1, comparison between a healthy brain and an Alzheimer's disease patient's brain is displayed. The Alzheimer's disease patient's brain is characterized by neural amyloid plaques and neurofibrillary tangles, which result in

severe neurodegeneration, including shrinkage of the hippocampus and other cerebral cortices [6].
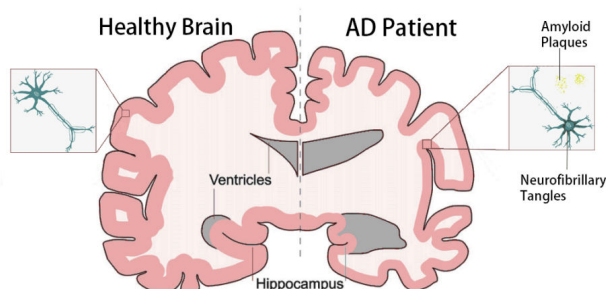

Figure 1. Difference between healthy brain and AD (Alzheimer's disease) patients (adapted from ref [6]).

1.1.2. Multimodality Diagnosis of Alzheimer's Disease

Multimodal diagnosis is an exhibition of the structural and functional imaging of brain images by a variety of imaging devices. By comparing the differences between various images and healthy brain image templates, the result reflects the image characteristics of Alzheimer's disease. It has been shown that in diagnosis and prognosis of Alzheimer's disease, multi-modal use may improve the performance more than a single modal [7]. In 2019, Zhou et al. [8] focused on how to make the best use of multimodal neuroimaging and genetic data for Alzheimer's disease diagnosis. They proposed a three-stage deep feature learning and fusion framework, and proved the superiority of using multimodality data through experiments. In 2021, for Alzheimer's disease diagnosis with incomplete modalities, Liu et al. [9] proposed an Auto-Encoder, which can complement the missing data in the kernel space. This can help solve the problem of incomplete medical data. Such research promotes the development of multimodal technology in the field of Alzheimer's disease.

The project intends to use T2-weighted Magnetic Resonance Imaging (T2-MRI) and positron emission tomography (PET). The lesions of T2-MRI in Alzheimer's disease focus on atrophy of brain regions such as hippocampus, amygdala, and entorhinal cortex. Studies have confirmed [10] that the hippocampus atrophy is earlier than other clinical symptoms of the disease, which is an early critical characteristics performance of Alzheimer's disease and can be used as an early specific and sensitive predictor. The only way that the brain gains energy is glucose uptake. Accordingly, the extent of Alzheimer's disease can be determined according to differences in glucose uptake in various parts of the brain. 18F-flurodeoxyglucose (18F-FDG) PET imaging shows a phenomenon in patients with Alzheimer's disease where the area of both sides of the temporal lobe and parietal lobe demonstrates a symmetrical decrease in glucose metabolism, and with the deteriorating of condition, the phenomenon spreads from the temporal lobe to the other cortical areas [11]. PET uses a unique imaging technique to present metabolic changes of glucose in the brain, which can be used to map the distribution of the lesions. In Figure 2, two different modality images of Alzheimer's disease patients are displayed.
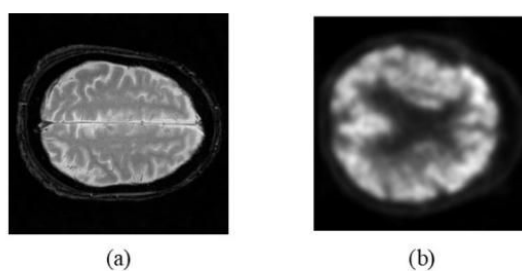

Figure 2. (**a**) T2-weighted MR (Magnetic Resonance) image of AD. (**b**) PET (Positron Emission Tomography) image of AD.

1.2. Multimodal Image Fusion Technology

Multimodal image fusion technology polymerizes the medical images of different modes into a single image set, and the images from different imaging systems are aggregated into a new output image [12]. Different modality medical images focus on different contents, and multimodal fusion technology can achieve the integration of functional image and anatomical image through the designed algorithm. Fusion image contains more useful information and makes the image more suitable for human vision processing, and for this reason it plays an important role in the diagnosis and treatment of diseases.

Meanwhile, multimodal image fusion is divided into three levels: pixel-level fusion, feature-level fusion and decision-level fusion. Among them, pixel-level fusion is the lowest level fusion, which is the direct fusion between pixels on the pixel level. Figure 3 illustrates how the fusion process works and displays an example of fusion. Pixel-level fusion has many advantages. Compared with others, it can restore the previous information of the image and contain more details of the image. So, the pixel-level fusion has the most application among these three fusion layers.
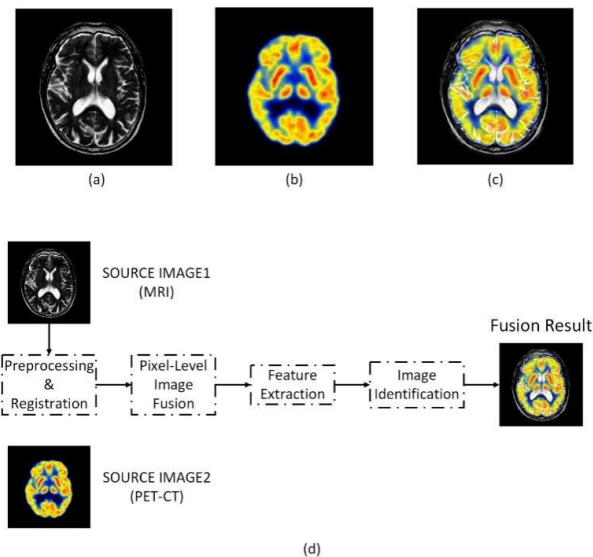
Figure 3. (**a**) T2-weighted MRI image. (**b**) PET image. (**c**) Fusion result of pixel-level fusion method. (**d**) Pixel-level image fusion flow chart.

1.3. Deep Learning and Convolution Neural Networks

Recently, deep learning has become a hot research topic in the field of machine learning, and has made many achievements in computer vision and speech recognition [13–15]. Varied deep learning methods such as deep neural network (DNN), convolution neural network (CNN), and recurrent neural network (RNN) have been implemented to diagnose human neurological disorders [14].

Deep learning is derived from artificial neural network, whose essence is the establishment of simulation structure of human brain neural networks and form a high-level feature representation through the combination of low-level features, in order to achieve efficient processing of complex data [15]. In 1962, Hubel [16] first proposed the concept of a receptive field; in 1980, Fukushima [17] proposed a neural perceptron based on receptive field, and implemented the first CNN model. Subsequently, LeCun et al. [18] applied Back Propagation (BP) algorithm to CNN training, and successfully completed the image recognition. CNN has attracted extensive attention because of its parallelism, distortion, and robustness. One example shows the application of deep learning diagnosis [19]. Additionally, CNN and RNN are found to present better results than other deep learning methods in diagnosing Alzheimer's disease [14].

CNN is a feedforward neural network with local connection and weight sharing. Each layer of neural networks consists of a pair of two-dimensional planes, and each

pair of two-dimensional planes is composed of a convolution layer and a lower sampling layer. Convolutional layer is also called feature extraction layers, which consist of multiple independent neurons. The function of convolutional layer is to extract the input data characteristics, and this approach is closer to the treatment of the human brain vision system. Figure 4 shows the basic structure of convolution neural network.
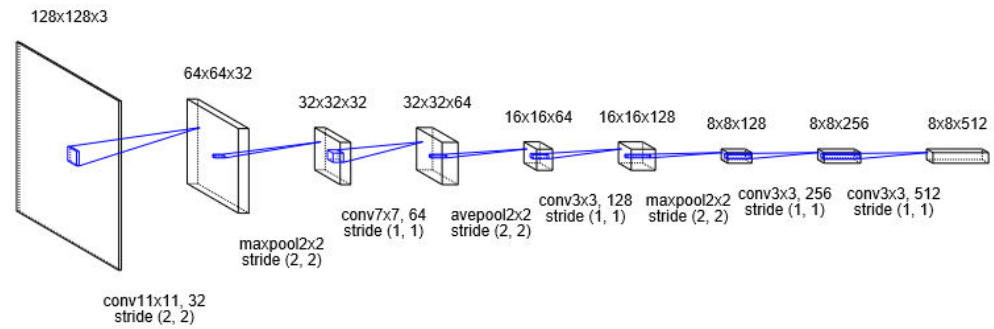

Figure 4. Basic structure of convolution neural network.

2. Research Contents and Methods

2.1. Image Data

A total of 102 Alzheimer's disease patients corresponding to their T2-MRI and PET images, and images of 80 patients with normal cognition as control group were collected from ADNI database, PPMI database. The dataset was composed of the corresponding MRI and PET images of each patients, which were all labeled by professionals in Alzheimer's disease research. Figure 5 shows one of the samples in the project group dataset, T2-MRI and PET image of a patient, respectively.

Figure 5. (**a**) AD patient's PET image. (**b**) AD patient's T2-MRI image.

2.2. Image Preprocessing

We collected images from the ADNI and PPMI database, and preprocessed the images to improve the recognition accuracy. The preprocessing process is divided into three parts, including gray level transformation, histogram equalization and wavelet soft-threshold denoising. The flow chart of the preprocessing is shown in Figure 6. In the following, we will discuss them separately.
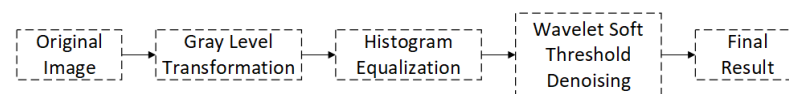

Figure 6. Flow diagram of preprocessing.

Gray level transformation maps image gray-scale from 0 to 255. We histogram equalize the image to adjust contrast after linear gray level transformation, which will increase

the local contrast of the original image. Histogram equalization can be expressed by the following formula:

$$s_k = T(r_k) = (L-1) \sum_{j=0}^{k} p_r(r_j)$$

$$= \frac{(L-1)}{MN} \sum_{j=0}^{k} n_j, k = 0, 1, 2, \ldots, L-1 \tag{1}$$

where L is the number of possible gray levels in the image, MN represents the image pixels number, and n_k represents pixels number in r_k gray scale. Figure 7 demonstrates the effect of histogram equalization.
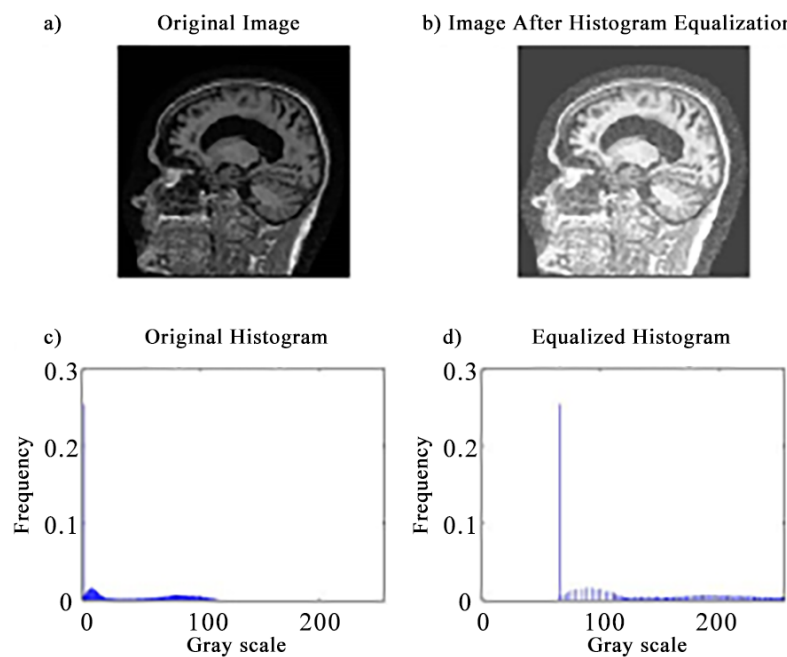
Figure 7. (**a**) Original image. (**b**) Histogram equalization image. (**c**) Original image histogram. (**d**) Histogram after equalization. In the histogram, the abscissa represents the gray scale and the ordinate represents the frequency.

After the gray level transform is the denoising process. The denoising algorithm implemented in this paper is based on the wavelet soft-threshold algorithm [20], which can be expressed in the following piecewise function:

$$f(x) = \begin{cases} \text{sign}(x) \circ (|x| - \alpha), & |x| \geq \alpha \\ 0, & |x| \leq \alpha \end{cases} \tag{2}$$

where x represents the pixel's gray value, α stands for the threshold value of the wavelet soft-threshold algorithm, and this threshold value is defined by the following formula:

$$\alpha = \sigma^2 \sqrt{2 \log(N^2)} \tag{3}$$

where N represents the signal dispersion number, and through the experiment we set $\sigma = 25$. Figure 8 shows the denoising effect with the algorithm.
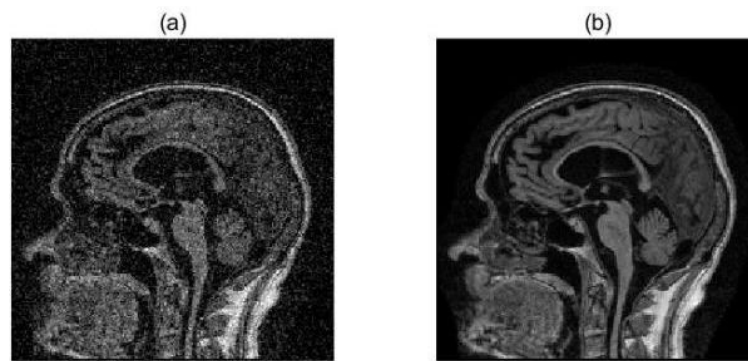
Figure 8. (**a**) Image before denoising. (**b**) Image after denoising.

The next step is contrast stretching. We used three stage contrast stretching methods in this paper. The basic form of piecewise linear function is as follows.

$$f(x) = \begin{cases} \frac{y_1}{x_1}x & x < x_1 \\ \frac{y_2-y_1}{x_2-x_1}(x-x_1)+y_1 & x_1 \le x \le x_2 \\ \frac{255-y_2}{255-y_1}(x-x_2)+y_2 & x < x_2 \end{cases} \tag{4}$$

The specific gray-scale restricting rule is shown in Figure 9, and in Figure 10 we give the image before and after contrast stretching.

Figure 9. Gray-scale restricting rule.

Figure 10. (**a**) Previous image. (**b**) After contrast stretching.

2.3. Image Registration

Image registration is the basis of fusion technology, which matches the same area of two images. It is defined by the following formula:

$$R(x,y) = h[F(f(x,y))] \tag{5}$$

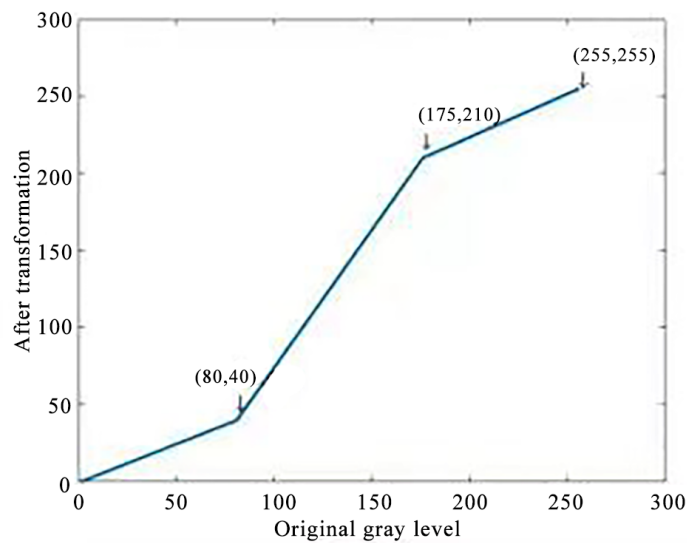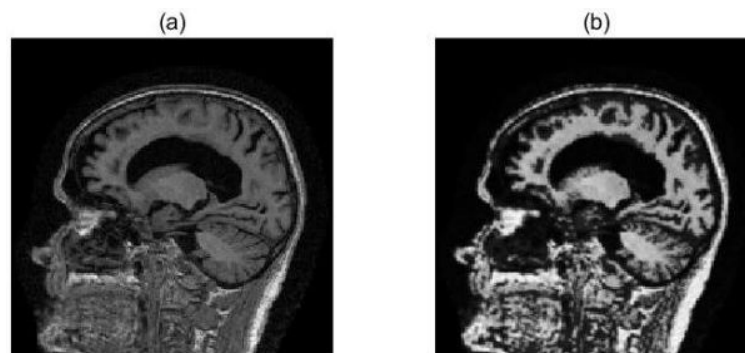where R means reference figure R and floating figure F. After making the transformation of pixels of figure F correspond to pixels of figure R, the matching is completed [21]. Correspondence between two images follows: $f(x,y)$ represents two-dimensional space transformation and $h(x,y)$ represents one-dimensional gray-scale transformation. Image registration is the crucial step to realize image fusion. In order to highlight the diagnosis results, it is necessary to make the previous image registration as accurate as possible.

In this paper, affine transformation was used as spatial geometric transformation of registration. Affine transformation of two-dimensional Euclidean space is defined as:

$$\begin{aligned} S(X) &= T(X) + C, X = [x,y]', \\ T &= [ab,cd]', C = [e,f]' \end{aligned} \tag{6}$$

$a \sim f$ are real numbers. x, y stands for the coordinate of each pixel. Transform S is affine transform, which has the characteristic that a finite pixel is mapped to a finite pixel [22]. Two-dimensional affine transformation is linear, including translation, rotation, and scaling. A coordinate point by affine transformation can be expressed as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = k \times \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{7}$$

where x', y' stands for the coordinate of pixel after affine, $k, \theta, \Delta x, \Delta y$ are registration parameters of the two images. This study divides the registration into four steps according to the requirements, as described below. Step 1: read image. Step 2: initial registration (coarse registration). The optimizer selects Powell algorithm, and similarity measure selects the maximum mutual information method. Step 3: improve registration accuracy by changing the optimizer's step size and increasing iterations times. Step 4: use the maximum mutual information as a reference to adjust the step size and iteration times, in order to improve the accuracy. As an optimization strategy of image registration, the Powell algorithm is a multi-parameter local optimization search algorithm that does not need to compute derivatives [23]. Its essence divides the optimization phase into an iterative process, which consists of $n + 1$ times one-dimensional search. First, an extremum point is obtained after searching n times of different conjugate directions, and the initial value of the search is searched in the direction of the extremum point connection. Then, the latest search direction is replaced by the direction of the extreme points obtained in the previous n times of search and keep iterating until the function stops decreasing. Maximum mutual information method is a registration similarity measure [24]. Based on the gray level of image, mutual information of two images is taken as the similarity measure, and the optimal transform [23] is obtained by maximizing the similarity measure. The formula is as follows:

$$I(A,B) = H(A) + H(B) - H(A,B) \tag{8}$$

$$S* = \operatorname{argmax} I(S(B), A) \tag{9}$$

where $H(A)$ and $H(B)$ are the entropy of the image which indicates the gray distribution of image. $S*$ is the updated transform result.

2.4. Multimodal Pixel-Level Image Fusion

The core of wavelet fusion is multi-resolution fusion, which is like the multi-channel characteristics of spatial frequency of human vision. Therefore, wavelet fusion has become a hot topic. The two input images are decomposed by K-layer wavelet fusion to obtain

$3K + 1$ sub-image, and one sub-image is the low-frequency image of the highest K layer, and the other $3K$ images are high frequency sub-images with different frequency characteristics which is the high frequency image of the original image in the horizontal, vertical and diagonal directions of the K layer. In this paper, we used traditional wavelet weighting algorithm as a standard comparison, and we also proposed a modified wavelet fusion algorithm, which is more suitable for MRI and PET fusion.

2.4.1. Traditional Wavelet Weighting Algorithm

Traditional wavelet algorithm decomposes image after registration by wavelet method and obtains a wavelet coefficients matrix of different frequencies. Then, low-frequency coefficients matrix and high-frequency coefficients matrix are both weighted and averaged by their weighted coefficients that are set according to the characteristics of medical images. Weighting coefficients used in this study is 0.5. After weighted average, low frequency coefficients and high frequency coefficients matrix of the fused image are obtained. Finally, the two coefficients matrixes are transformed, respectively, by wavelet inverse transform to obtain fused image. The formula is as follows:

$$C_k = \omega_k(A) \times C_k(A) + \omega_k(B) \times C_k(B), \tag{10}$$

$$D_k = W_k(A) \times D_k(A) + W_k(B) \times D_k(B) \tag{11}$$

where A means the low frequency image, B is the high frequency image, C_k means low frequency matrix, ω_k means the weight parameter of matrix. W_k means the weight parameter of inverse wavelet transform, D_k means the maximum value of the inverse transform result.

2.4.2. Frequency Weighted Wavelet Fusion Algorithm

Frequency weighted wavelet fusion algorithm is similar to the traditional method in the wavelet transform and weighted-average part. However, in the inverse transform part, it calculates the mean square deviation, takes the maximum value to get the corresponding high-frequency coefficients matrix. Finally, the two coefficients matrix is transformed by wavelet inverse transform to obtain the fused image. The formula is as follows:

$$C_k = \omega_k(A) \times C_k(A) + \omega_k(B) \times C_k(B), \tag{12}$$

$$D_k(F) = \max[D_k(A), D_k(B)] \tag{13}$$

where A means the low frequency image, B is the high frequency image, C_k means low frequency matrix, ω_k means the weight parameter of matrix, W_k means the weight parameter of inverse wavelet transform, and D_k means the maximum value of the inverse transform result.

3. Multi-Mode Image Fusion Results and Objective Evaluation

3.1. Multimodal Image Fusion Results

Taking a patient's data from the database as an example, the results of three different fusion images are displayed below in Figure 11.
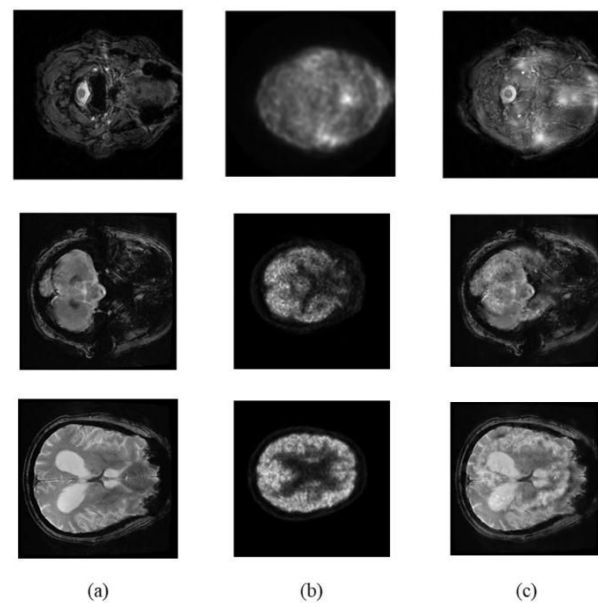
Figure 11. This figure displayed the different modalities space serial images of the same patient and their fusion result. (**a**) The original MRI images. (**b**) The original PET images. (**c**) The fusion results of the corresponding MRI and PET images.

3.2. Evaluation and Analysis of Fused Images

Although fused images cannot directly assist the diagnosis of the image like doctors, it can give basic judgment from the nature of image. In order to comprehensively evaluate the fusion algorithm of this paper, we selected six image evaluation parameters and displayed them in Table 1.

According to the parameter evaluation formula above, the parameters of each algorithm were calculated. Results are shown in Table 2, and the following analysis was made according to the calculated result. We also use the significance test method to observe whether there is a significant difference, the *p*-value is less than 0.05, so a significant improvement is made.

On the basis of the data analysis in the table, the improved wavelet fusion is superior to the traditional wavelet fusion in spatial frequency, mean gradient, information entropy, mutual information, cross entropy, and peak signal to noise ratio. Overall, the improved wavelet fusion implemented in this research has better performance than the traditional wavelet fusion.

Table 1. Objective evaluation principle table.

Parameter	Definition	Formula	Evaluation Criteria
Spatial Frequency (SF)	Spatial frequency reflects the overall activity level of an image space and evaluates the degree of image clarity.	$$SF = \sqrt{RF^2 + CF^2}$$ RF and CF are the row frequency and the column frequency, respectively,	Larger SF refers to more active image and better fusion effect.
Information Entropy (IE)	Information entropy refers to the probability distribution of pixels of different gray levels in space, and describes its detail expressive force.	$$IE = -\sum_{m=0}^{255} P_m \cdot \log(P_m)$$ P_m is the probability that the gray scale m appears in the image.	Bigger IE refers to richer details and better quality of the fused image.
Mean Gradient (MG)	The average gradient represents the mean value of image gradient, which reflects the change of the gray level of image.	$$MG = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{\Delta xF(i,j)^2 + \Delta yF(i,j)^2}$$ $\Delta xF(i,j)$ and $\Delta yF(i,j)$ represent the difference in the x and y directions.	Higher MG refers to higher image contrast.
Cross Entropy (CERF)	Cross entropy is used to measure the difference of information between two images, which reflects the difference between the two-pixel levels.	$$CERF = \sum_{i=0}^{L-1} P_{R_i} \log \frac{P_{R_i}}{P_{F_i}}$$	Smaller CERF difference refers to more information the fusion method extracts
Peak Signal-to-Noise Ratio (PSNR)	The peak signal-to-noise ratio measures the realistic degree of the image.	$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$ MAX_I represents the maximum value of the image gray scale.	Higher PSNR refers to smaller distortion and better fusion effect.

Table 2. Objective evaluation results of fusion algorithm.

	SE	TE	MG	MI	CERF	PSNR
Traditional Wavelet Fusion	6.7249	4.8320	2.8021	0.3543	1.3224	26.5542
Improved Wavelet Fusion	8.3019	4.9741	3.4290	0.3117	1.2041	32.0910
p-value	<0.001	0.011	<0.001	0.005	<0.001	0.017

4. Convolution Neural Network Assisting Diagnose

Figure 12 is the flow chart of data from the end of preprocessing to entering the network and finally realizing classification.
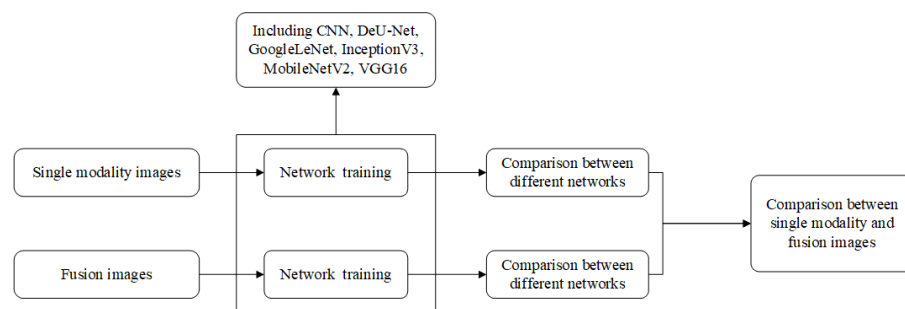

Figure 12. Network training diagram.

4.1. Construction of Convolution Neural Network

In this paper, the project group designed and improved two CNN models and implemented the diagnose model in training. Besides that, four other kinds of image classification deep learning models, including, ResNet, GoogLeNet, Inception Net and AlexNet, were also implemented as reference and comparison. First, the specific model structure and training process of two CNN models we designed were introduced. Then, the MNIST dataset was used to test their performance.

4.1.1. Convolution Neural Network Structure

The first convolutional neural network designed in this paper was based on the classical convolutional neural network model LeNet-5 [18]. This CNN model had 10 layers, and Figure 13 shows the CNN model structure.
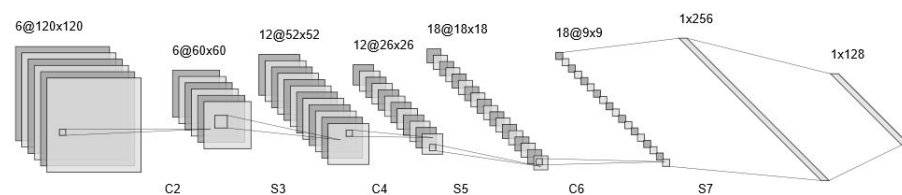
Figure 13. Convolution neural network mode.

The first layer is the input layer. The input image is divided into three stages. Each stage consists the convolution layer and the down-sampling layer. The convolution kernel is $9 * 9$, and the down-sampling layer is sampled by $2 * 2$ region. Each stage has the same structure and relative operation. The first stage output is six pieces of $60 * 60$ feature graphs. The second stage output is 12 pieces of $26 * 26$ feature graphs. The third stage finally forms 18 pieces of $9 * 9$ feature graphs after down-sampling, and each of graphs has a corresponding bias term. The eighth layer is a convolution layer, which contains 120 feature graphs. Each element in the feature graph relates to the feature graph of the previous layer. After $9 * 9$ convolution kernel operation, the size of the feature graph in

this layer is $1 * 1$, which constitutes the full connection between the fifth and sixth layers. The ninth full connected layer contains 84 neurons.

The final output layer consists of the Euclidean Radial Basis Function (RBF) unit, and each represents a category. In this project Alzheimer's disease diagnosis is a two-classification problem that is ill or not, so the output layer has two units.

Before CNN training, parameter initialization is needed. Parameter initialization is very important in the gradient descent algorithm. If the error surface is relatively flat, it will lead to a very slow convergence rate. In general, more initialization weights are distributed as follows:

$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{P^{(l)} + P^{(l-1)}}}, \frac{\sqrt{6}}{\sqrt{P^{(l)} + P^{(l-1)}}}\right] \tag{14}$$

This is the representation of Xavier initialization, where $P^{(l)}$ means the number of pixels in layer l, and each parameter will have the value from the interval specified by U.

CNN model training uses the back propagation (BP) algorithm. The main content of the algorithm is divided into forward propagation and back propagation. First, the training data is input into the CNN-10 model, and then the gradient of the weight and bias and inverse calculation error is calculated by the convolution filtering, the down sampling and the activation function, and the weight and bias are updated.

4.1.2. Improved Deformable U-Net (DeU-Net)

U-Net is a special CNN [25] that excels in image segmentation. We use U-net to process data in this research, which is an approach of computer-aided diagnosis of Alzheimer's disease in PET images. Inspired by the network proposed by Dai in 2019 [26], we replace the typical convolution kernel with deformable convolution in the network. Improved structure of network is shown in Figure 14.
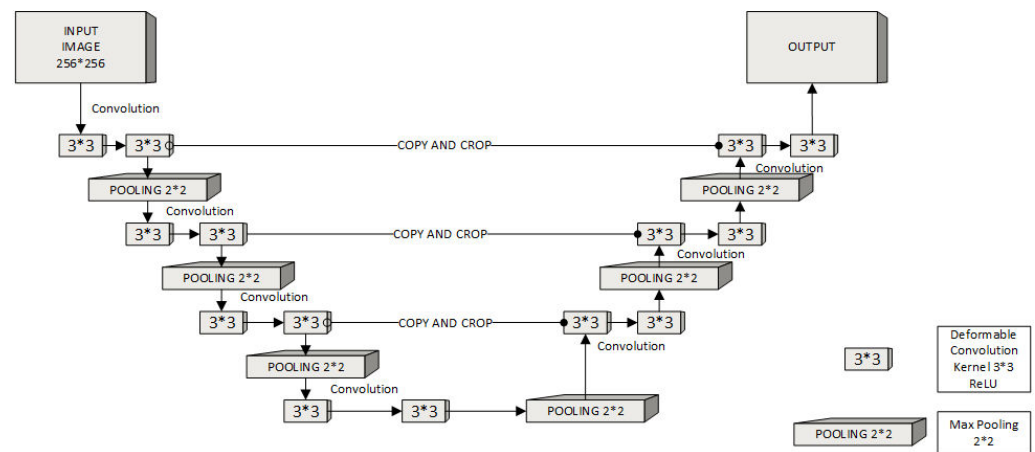
Figure 14. Convolution neural network mode.

This CNN architecture has symmetrical structure and deformable convolution kernels. It consists path of the encoder and decoder, each formed by three layers. The encoder path has two $3 * 3$ deformable convolution kernels and a down-sampling max-pooling operator; similarly, the base layer also formed by two deformable convolution kernels with this structure. Copy and crop between the encoder and the decoder helps the reservation of information for localization [27]. Activation function is Rectified Linear Unit (ReLU). By the Adam-Optimizer based on gradient decent, the parameters are relatively stable and promote their dynamic adjustment.

Definition of deformable convolution kernel:

$$Z_{l+1}(i,j) = \sum_{k=1}^{K1} \sum_{x=1}^{f} \sum_{y=1}^{f} [Z_k^l(s_0 i + x, s_0 j + y) w_k^{l+1}(x + \Delta x, y + \Delta y)], \tag{15}$$

$$Z_{l+1}(p) = \sum_{k=1}^{K1} \sum_{x=1}^{f} \sum_{y=1}^{f} \left[Z_k^l(p_0) w_k^{l+1}(\Delta p_n) \right] \tag{16}$$

where Z^l represents the input of convolution layer l and $Z^{(l+1)}$ represents output. $Z(p)$ is the value of pixels on the feature map. $f, s_0, K1$ represents size, stride, and padding layers size of the convolution kernel, respectively. In this equation $\Delta x, \Delta y$ can be fractional, the bilinear interpolated w_k^{l+1} as follows:

$$w_k^{l+1}(p) = \sum_{x=1}^{q} \sum_{y=1}^{q} u(q_w, p_w) u(q_z, p_z) w(q) \tag{17}$$

where p represents the arbitrary fractional pixels and q represents integral pixels in the feature map. The definition of Kernel u is:

$$u(x, y) = \max(0, 1 - |m - n|) \tag{18}$$

In every layer, we used deformable kernels to work in the same way rather than implementing different kernels for each layer, this will improve network performance. Using the back-propagation method, the network was trained end-to-end with labeled PET/MRI/fusion Images.

Feature map pixels soft-max classifier with cross entropy loss function is the energy function, this classifier is:

$$S_k(x, y) = \frac{e^{A_k(x,y)}}{\sum_{x=1}^{K} \sum_{y=1}^{K} e^{A'k(x,y)}} \tag{19}$$

$A_k(x, y)$ represents the activated feature pixels in layer k, and the pixels number in the feature map is K.

The energy function is:

$$E = \sum_{x=1}^{K} \sum_{y=1}^{K2} \varpi(x, y) \log \left(s_{l(x,y)}(x, y) \right) \tag{20}$$

where $\varpi(x, y)$ is pixel weight map function that previously defined, inspired by Ronnenberger [25]:

$$\varpi(x, y) = w_c(x, y) + \varpi_0 \exp \left(-\frac{(d_1(x, y) + d_2(x, y))^2}{2\sigma^2} \right) \tag{21}$$

To avoid excessive activation of some pixels and reduce errors, the weight of U-net needs to initialize. We set $\sigma = 8$ and $\varpi_0 = 10$ to obtain better performance. Epoch was set at 10,000, learning rate 0.01.

4.2. Performance Testing of Convolutional Neural Networks

After the construction of convolutional neural network, research needs to test the performance of CNN model before formal training diagnosis. Network performance was validated by testing 60,000 handwritten digits downloaded from the MNIST database.

Testing progress was done by the following specific test method: first, 60,000 pieces of $128 * 128$ handwritten numbers and their corresponding labels were read into the CNN model. In view of the large amount of data, the 10-fold cross validation method was used to train the generalization ability of CNN model. One-fold cross validation is a method derived from the k-fold cross validation. By stratified sampling from the data set D, K, exclusive subsets of similar size can be separated mutually. Each cycle took a group as a test set and the rest of k-1 groups were training sets. After k times of training and testing were carried out, the average accuracy rate of k-group test represents the accuracy of the deep learning model [28]. Accuracy rate result of the applied different training tests is

shown in Table 3. CNN and DeU-Net are the networks that the project group built, and the rest of the network is preset architecture as reference.

Table 3. Accuracy table of 10-fold cross validation.

K-Fold	CNN	DeU-Net	GoogLeNet	InceptionV3	MobileNetV2	VGG16
1	92.35%	92.78%	91.02%	93.04%	92.84%	92.99%
2	95.32%	95.32%	95.66%	95.46%	95.10%	93.56%
3	96.52%	96.52%	96.78%	96.43%	96.32%	96.73%
4	96.97%	96.97%	97.01%	96.99%	96.57%	96.81%
5	97.45%	97.45%	97.62%	97.41%	97.28%	97.39%
6	97.52%	97.74%	97.81%	97.76%	97.86%	97.68%
7	97.76%	97.88%	97.92%	97.81%	97.95%	97.89%
8	97.83%	97.92%	97.92%	97.92%	97.98%	98.00%
9	97.96%	98.01%	98.01%	97.99%	98.04%	98.03%
10	98.03%	98.05%	98.02%	98.01%	98.06%	98.08%

Increasing training times can get better accuracy of the convolutional neural network. High accuracy of the network model test proves that the model has great performance, and the parameter settings of the hidden layer of the model was relatively correct.

4.3. Training and Testing Process

In order to validate the significance of image fusion, single modality images and multi-modal fusion images were trained in both CNN and U-Net models, and the performance of each model were compared to determine which model has the relatively better diagnostic effect. Figure 15 illustrates the constitution of the single modality images dataset and the fusion images dataset.
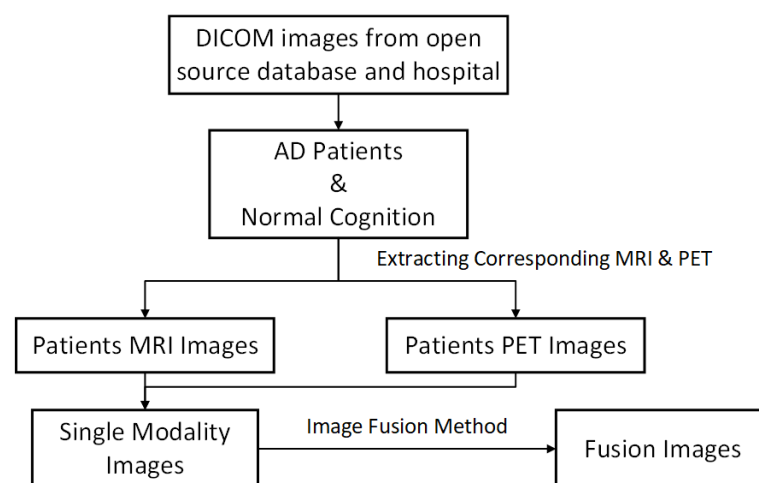
Figure 15. Constitution of the implemented datasets.

4.3.1. Single Modality Image Training Model

T2-MRI and PET image of the same parts of brain from 101 patients with Alzheimer's disease and 80 normal cognitive people as control groups in total 5430 images of each modality were collected from the ADNI database and stored separately in two three-dimensional matrices. Each matrix was read into the network with its corresponding binary classification label. Because of the few training and test data, it was necessary to use the set aside method to verify the generalization ability of CNN.

Set aside was the method that divided data set D into two mutually exclusive sets, one of which was the training set S, and the other one is the test set T. After the model training completed, S and T was used to evaluate the test error [27].

4.3.2. Fusion Image Training Model

Multi-modal fusion algorithm has the highest objective evaluation of all fusion algorithm. T2-MRI and PET images of the same person in the single modality training model are automatically fused. The training and verifying process was the same as the MRI and PET images training model.

5. Diagnosis Results and Analysis

5.1. Convolution Neural Network Diagnostic Results

5.1.1. Single Modality Training Model

By setting aside method, 3008 images were used as training sets, and 1534 images were used as evaluation set. After training and testing, the diagnostic accuracy used MRI only images of CNN reached 80.65% and the U-Net reached 84.17%. By using the similar method, results of training with PET only images were also achieved. Mean and SD of AUC shown as Table 4:

Table 4. Mean and SD of AUC with single modality images.

	CNN	DeU-Net	InceptionV3	MobileNetV2	VGG16
Train	0.914 ± 0.072	0.962 ± 0.075	0.815 ± 0.145	0.970 ± 0.013	0.955 ± 0.012
Test	0.922 ± 0.075	0.926 ± 0.029	0.827 ± 0.151	0.908 ± 0.017	0.917 ± 0.018
Validation	0.899 ± 0.048	0.906 ± 0.028	0.809 ± 0.127	0.920 ± 0.064	0.905 ± 0.019

The detailed training results are shown in Table 5.

Table 5. Performance of different neural network in training with single modality images.

		AUC	ACC (%)	SENS (%)	SPEC (%)	PPV (%)	NPV (%)
CNN	Train	0.914	87.1	90.2	91.4	94.5	89.7
	Test	0.922	89.6	91.4	90.5	91.2	88.7
	Validation	0.899	84.5	88.7	89.1	86.9	84.4
DeU-Net	Train	0.962	89.9	98.1	89.9	92.1	87.1
	Test	0.926	90.4	94.1	83.2	94.3	83.6
	Validation	0.906	89.5	92.6	81.9	97.9	87.3
GoogLeNet	Train	0.852	79.6	83.6	80.1	82.4	84.7
	Test	0.871	77.5	85.4	79.7	81.6	83.5
	Validation	0.839	80.0	78.1	80.5	80.0	79.1
InceptionV3	Train	0.815	78.8	80.8	90.3	80.8	79.0
	Test	0.827	79.7	81.0	89.1	81.0	73.5
	Validation	0.809	77.3	78.7	92.1	88.1	81.2
MobileNetV2	Train	0.970	95.9	98.4	96.5	98.2	94.2
	Test	0.908	88.4	94.7	90.2	94.7	82.5
	Validation	0.920	81.0	96.3	91.9	95.4	87.1
VGG16	Train	0.955	86.4	95.3	97.1	95.3	94.9
	Test	0.917	87.7	91.5	93.2	91.5	86.2
	Validation	0.905	81.1	90.1	91.0	89.7	84.1

In Table 5, AUC means area under the receiver operating characteristic curve. AUC represents the degree of differentiation of different categories, the value is proportional to the probability that the model is correctly classified. *ACC* means accuracy, and its value is proportional to the overall prediction performance. *SENS* means sensitivity. In *SENS*, the numerator is the number of positive samples predicted to be correct, and the denominator is the number of positive samples predicted to be negative, also known as the recall rate, indicating how many positive samples predicted to be correct. *SPEC* means specificity, which is contrary to *SENS*. *SPEC* indicates how many of the negative samples are predicted to be correct. *PPV* means positive predict value, which is based on the result of prediction, and it represents the amount of the samples predicted to be positive were correct. There are two possible ways to predict a positive class, one is to predict a true positive (TP), and

the other is to predict a false positive (FP). *NPV* means negative predict value, which is contrary to PPV. NPV represents how many of the samples predicted to be negative were correct. Test means training cohorts. Validation means validation cohort. Prove means independent prove cohort. Factors above were calculated by the following formulas:

$$ACC = \frac{n_{\text{right}}}{n_{\text{false}} + n_{\text{right}}} \tag{22}$$

$$SENS = \frac{T_{\text{positive}}}{T_{\text{positive}} + F_{\text{negative}}} \tag{23}$$

$$SPEC = \frac{T_{\text{negative}}}{T_{\text{negative}} + F_{\text{positive}}} \tag{24}$$

$$PPV = \frac{T_{\text{positive}}}{T_{\text{positive}} + F_{\text{positive}}} \tag{25}$$

$$NPV = \frac{T_{\text{negative}}}{T_{\text{negative}} + F_{\text{negative}}} \tag{26}$$

where n_{right} is the correctly predicted samples and n_{false} is the falsely predicted samples. T_{positive} is the correctly predicted positive sample. T_{negative} is the correctly predicted negative sample. F_{positive} is the falsely predicted positive sample. F_{negative} is falsely predicted negative sample.

In Figure 16, we also provide the confusion matrix in this case.
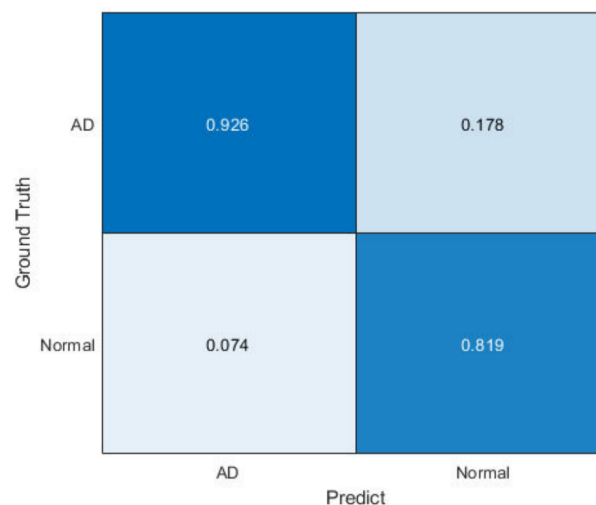
Figure 16. Confusion matrix of DeU-Net with single modality images.

5.1.2. Fusion Image Training Model

According to the objective evaluation of the fused images, the best algorithm is the frequency weighted wavelet fusion algorithm. Therefore, the fused image set was selected to train different models. In order to compare different performance between models in diagnosing Alzheimer's disease, multiple deep learning models were implemented. Mean and Sd of AUC shown are in Table 6.

Table 6. Mean and SD of AUC with fusion images.

|	CNN	GoogLeNet	DeU-Net	InceptionV3	MobileNetV2	VGG16
Train	0.935 ± 0.042	0.916 ± 0.053	0.967 ± 0.022	0.891 ± 0.038	0.945 ± 0.050	0.949 ± 0.040
Test	0.927 ± 0.046	0.909 ± 0.017	0.929 ± 0.037	0.922 ± 0.046	0.909 ± 0.065	0.928 ± 0.010
Validation	0.916 ± 0.053	0.897 ± 0.013	0.941 ± 0.017	0.849 ± 0.043	0.899 ± 0.063	0.856 ± 0.027

Detailed training results were listed in Table 7.

In Table 7, AUC means area under the receiver operating characteristic curve. ACC means accuracy. SENS means sensitivity. SPEC means specificity. PPV means positive predict value. NPV means negative predict value. Test means training cohorts. Validation means validation cohort. Prove means independent prove cohort. We also recorded the training loss curve of DeU-Net, where orange represents the training set and blue represents the test set. It can be seen from Figure 17 that the training loss decreases steadily, which shows the good performance in training.

Table 7. Performance of different neural network in training with fusion images.

		AUC	ACC (%)	SENS (%)	SPEC (%)	PPV (%)	NPV (%)
CNN	Train	0.935	89.4	93.1	95.2	88.2	88.2
	Test	0.927	91.5	92.5	93.4	91.0	93.3
	Validation	0.916	90.9	90.9	90.9	89.5	89.5
DeU-Net	Train	0.967	95.7	97.9	94.8	90.2	88.2
	Test	0.929	91.6	95.1	87.1	95.1	85.6
	Validation	0.941	90.5	92.4	84.0	99.7	84.9
GoogLeNet	Train	0.916	91.1	88.5	86.4	84.9	77.3
	Test	0.909	88.4	86.1	82.3	87.6	71.5
	Validation	0.897	86.0	79.2	81.7	77.5	82.1
InceptionV3	Train	0.891	85.1	78.4	95.1	85.3	85.2
	Test	0.922	89.3	85.0	90.7	82.0	71.5
	Validation	0.849	81.2	82.3	94.0	89.0	91.2
MobileNetV4	Train	0.945	91.4	94.3	96.0	96.9	93.0
	Test	0.909	89.7	92.5	90.2	97.2	86.7
	Validation	0.899	88.1	99.2	88.9	94.9	82.0
VGG16	Train	0.949	89.9	94.2	97.2	98.2	92.9
	Test	0.928	88.5	92.5	93.1	97.3	89.6
	Validation	0.856	84.2	83.9	87.0	84.9	83.9

In Figure 18, we also provide the confusion matrix in training with fusion images.
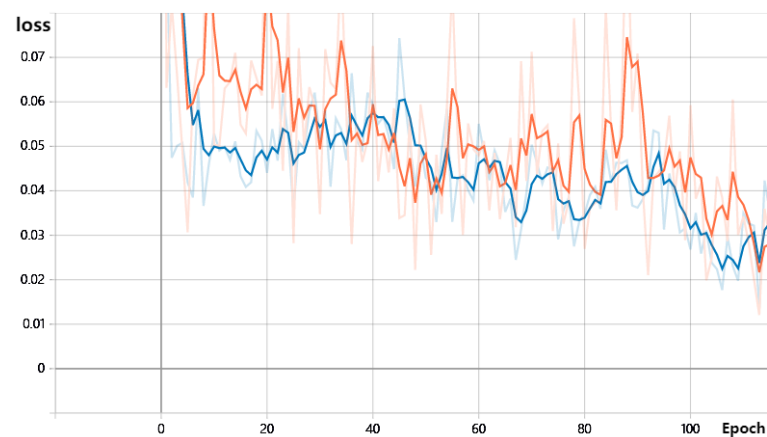
Figure 17. Training loss with fusion images.
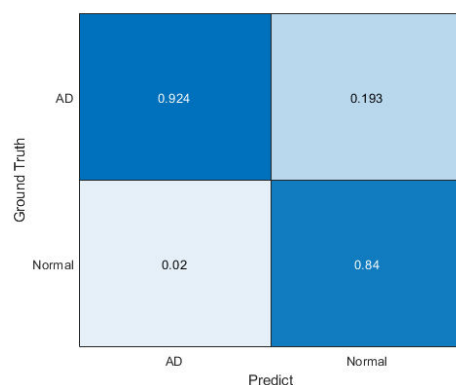
Figure 18. Confusion matrix of DeU-Net with fusion images.

5.1.3. Comparison between Single Modality and Fusion Modality

Table 8 shows comparisons of different neural network accuracy between single modality and fusion images. Notes that the proposed fusion method yields better performance than implementing only a single modality image in training deep learning models. All of the fusion image trained models outperformed the single modality image trained models. Fusion images training can improve accuracy of classification. In training cohort, testing cohort and validation cohort, fusion images significantly raised AUC than training with single modality images.

Table 8. Comparison between single modality images and fusion images implemented in training different deep learning models.

	Dataset Modality	ACC (%)	AUC (%)
CNN	Single	84.5	89.9
	Fusion	90.9	91.6
DeU-Net	Single	89.5	90.6
	Fusion	90.5	94.1
GoogLeNet	Single	80.0	83.9
	Fusion	86.0	89.7
InceptionV3	Single	77.3	80.9
	Fusion	81.2	84.9
MobileNetV2	Single	81.0	92.0
	Fusion	88.1	89.9
VGG16	Single	81.1	90.5
	Fusion	84.2	85.6

6. Discussion

In this study, we developed and validated a pixel-based image fusion method based on brain conventional T2-weighted MRI and 18F-FDG PET-CT images for prediction of Alzheimer's disease. This method showed significantly better diagnostic performances in distinguishing patients with an Alzheimer's disease and cognitive normal patients than any single method.

Diagnostic information in fusion image is more abundant because the fused image integrates the anatomical information and the metabolic function information. Resulting from the above, it can be analyzed that fusion images improved the performance of the two CNN models that were built by the project group, as CNN and DeU-Net have significant improvement in diagnosing accuracy. Meanwhile, DeU-Net has better accuracy or performance in processing medical images than traditional CNN. However, fusion images training in other pretrained models represents irregularity performance.

This result demonstrated that multimodal fusion can improve the diagnostic rate of Alzheimer's disease and confirms that multimodal fusion technology is greatly significant in the diagnosis of Alzheimer's disease. The result also proved that the multimodal image

fusion based on convolutional neural networks has important research value. At the same time, one of the problems in this paper is that the accuracy of the diagnostic models still needs to be improved. The number of samples currently diagnosed by Alzheimer's disease is 182 (102 patients and 80 people without disease) much less than the MNIST set, while the number of sample sets for digital identification is 60,000. One of the reasons for the low accuracy is that the number of training samples is too small, so the extracted features are not obvious.

7. Conclusions

In order to assist doctors in the accurate diagnosis of Alzheimer's disease, this paper conducted an in-depth study of multi-modal image fusion technology. First, the traditional wavelet fusion of T2-MRI and PET modal images of Alzheimer disease patients were studied, and then a wavelet fusion algorithm that is more suitable for medical image fusion was proposed. Based on the evaluation of objective evaluation parameters, we found that the improved wavelet fusion algorithm has better performance than the traditional wavelet fusion algorithm on the testing accuracy.

This paper first improved and implemented the traditional convolution neural network and U-Net in the diagnosis of Alzheimer's disease in order to explore their diagnostic characteristics. It turned out that this modification of CNN has significantly improved the performance of CNN in processing with Alzheimer's disease medical images. Additionally, this paper first implemented a fusion image technique in Alzheimer's disease diagnosing and proved the efficiency of diagnosing using fusion images by implementing different deep learning models, including some pretrained models.

This paper designed two image sets training with six different deep learning structures in the same training process. MRI, PET and fusion images of Alzheimer's disease were trained and tested, respectively. Results of these tests demonstrated that the accuracy of fusion image is higher than that of single mode. In a word, pixel-based multi-modal fusion images applying on convolution neural network can have an outstanding diagnostic effect on the diagnosis of Alzheimer's disease.

Similar to the problems encountered by Liu et al. [9], the author is also actively exploring solutions to the problem of incomplete data, such as using semi-supervised learning or unsupervised learning. The authors also notice that the experiment focuses on binary classification, and the task of multiple classification has not been involved, and plan to focus on it in future research.

Author Contributions: Conceptualization, Y.D.; methodology, Y.D. and W.B.; software, Z.T.; validation, Z.T.; formal analysis, W.B. and Z.T.; investigation, W.B.; resources, Y.D.; data curation, W.B.; writing—original draft preparation, W.B. and Z.T.; writing—review and editing, W.B.; visualization, Z.X.; supervision, W.C. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: According to ADNI protocols, all procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the Helsinki declaration. The ADNI data collection was carried out after obtaining written informed consent from the participants. More details can be found at http://adni.loni.usc.edu. Similarly, in https://www.ppmi-info.org, information about PPMI database can be found (accessed date 25 September 2019).

Data Availability Statement: The data were collected from the open database ADNI and PPMI.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 2021 Alzheimer's disease facts and figures. *Alzheimers Dement.* **2021**, *17*, 327–406. [CrossRef]
2. Wang, S.H.; Phillips, P.; Sui, Y.X.; Liu, B.; Yang, M.; Cheng, H. Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling. *J. Med. Syst.* **2018**, *42*. [CrossRef] [PubMed]
3. Guy, M.M.; David, S.K.; Howard, C.; Bradley, T.H.; Clifford, R.J.; Claudia, R.J.; Claudia, H.K.; Williams, E.K.; Walter, J.K.; Jennifer, J.M.; et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **2011**, *7*, 263–269.
4. Gao, T.L.; Chen, K.P. Common diagnostic methods of Alzheimer's disease. *Mod. Med.* **2016**, *44*, 415–419.
5. Yu, Z.; Fu, W.L. Alzheimer's disease and diagnostic research progress. *J. Clin. Lab.* **2016**, *34*, 49–51.
6. Saraiva, C.; Praça, C.; Ferreira, R.; Santos, T.; Ferreira, L.; Bernardino, L. Nanoparticle-mediated brain drug delivery: Overcoming blood–brain barrier to treat neurodegenerative diseases. *J. Control. Release* **2016**, *235*, 34–47. [CrossRef] [PubMed]
7. Perrin, R.J.; Fagan, A.M.; Holtzman, D.M. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* **2009**, *461*, 916–922. [CrossRef] [PubMed]
8. Zhou, T.; Thung, K.; Zhu, X.; Shen, D. Effective feature learning and fusion of multimodality data using stage wise deep neural network for dementia diagnosis. *Hum. Brain Mapp.* **2019**, *40*, 1001–1016. [CrossRef] [PubMed]
9. Liu, Y.B.; Fan, L.X.; Zhang, C.Q.; Zhou, T.; Xiao, Z.T.; Geng, L.; Shen, D.G. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. *Med. Image Anal.* **2021**, *69*. [CrossRef] [PubMed]
10. Handels, R.; Vermunt, L.; Sikkes, S.; Potashman, M. Predicting the health economic impact of early treatment in pre-dementia alzheimer's disease. *Alzheimer's Dement.* **2017**, *13*, 1457. [CrossRef]
11. Ece, B.; Caldwell, J.Z.K.; Banks, S.J. Current understanding of magnetic resonance imaging biomarkers and memory in Alzheimer's disease. *Alzheimers Dement. Transl. Res. Clin. Interv.* **2018**, *4*, 395–413.
12. Yang, W.; Liu, J. Research and development of medical image fusion. In Proceedings of the 2013 IEEE International Conference on Medical Imaging Physics and Engineering, Shenyang, China, 19–20 October 2013; pp. 307–309. [CrossRef]
13. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical Image Analysis using Convolutional Neural Networks: A Review. *J. Med. Syst.* **2017**, *42*, 226. [CrossRef] [PubMed]
14. Gautam, R.; Sharma, M. Prevalence and Diagnosis of Neurological Disorders Using Different Deep Learning Techniques: AMeta-Analysis. *J. Med. Syst.* **2020**, *44*, 1–24. [CrossRef] [PubMed]
15. Tustison, N.J.; Avants, B.B.; Gee, J.C. Learning image-based spatial transformations via convolutional neural networks: A review. *Magn. Reson. Imaging* **2019**, *64*, 142–153. [CrossRef]
16. Hubel, D.H.; Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **1962**, *160*, 106–154. [CrossRef]
17. Fukushima, K. Nleocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **1980**, *36*, 193–202. [CrossRef]
18. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
19. Islam, M.; Poly, T.N.; Yang, H.C.; Atique, S.; Li, Y.-C.J. Deep Learning for Accurate Diagnosis of Glaucomatous Optic Neuropathy Using Digital Fundus Image: A Meta-Analysis. *Stud Health Technol. Inform.* **2020**, *270*, 153–157.
20. Wang, X.; Zhuang, C. A TE Process Fault Diagnosis Method Based on Improved Wavelet Threshold Denoising and Principal Component Analysis. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018.
21. Ma, J.; Jiang, J.; Zhou, H.; Zhao, J.; Guo, X. Guided Locality Preserving Feature Matching for Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4435–4447. [CrossRef]
22. Uss, M.L.; Vozel, B.; Lukin, V.V.; Chehdi, K. Efficient Rotation-Scaling-Translation Parameter Estimation Based on the Fractal Image Model. *IEEE Trans. Geoence Remote Sens.* **2015**, *54*, 197–212. [CrossRef]
23. Pan, T.-T.; Ji, Z. Research on Medical Image Registration Based on QPSO and Powell Algorithm. In Proceedings of the International Symposium on Distributed Computing & Applications for Business Engineering & Science, Guiyang, China, 18–24 August 2015.
24. Jim, J.; Yan, W.; Yi, W.; Zhao, S.G.; Gao, X. Maximum mutual information regularized classification. *Eng. Appl. Artif. Intell.* **2015**, *37*, 1–8.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing & Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
26. Dai, Y.; Tang, Z.; Wang, Y.; Xu, Z.A. Data Driven Intelligent Diagnostics for Parkinson's Disease. *IEEE Access* **2019**, *7*, 106040–106049. [CrossRef]
27. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef] [PubMed]
28. Zhihua, Z. *Machine Learning*; Tsinghua University Press: Beijing, China, 2016.