Diagnostic Research of Alzheimer's Disease based on imaging and deep learning

ZhangQi LIYuchen ZhouHan
23065105g 23050743g 23103393g
Data Science & Analytics

Date: July 2, 2024

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disease that affects millions of people worldwide. Early and accurate diagnosis of AD plays a crucial role in timely intervention and improving patient prognosis. In recent years, there has been a growing interest in applying machine learning to medical image data to analyse and diagnose Alzheimer's disease. We pre-processed medical images using grey-scale mapping and denoising techniques to improve data quality. Subsequently, relevant features are extracted from the pre-processed images using feature extraction methods to capture important patterns and anomalies associated with AD. In addition to this, algorithms such as Support Vector Machines, Random Forests and Convolutional Neural Networks have been used to train and evaluate predictive models for extracting features. The aim of this study is to find AD-sensitive biomarkers and screen for signals related to AD pathology by performing data analysis experiments on the hippocampus associated with Alzheimer's disease using the ADNI dataset. The findings have the potential to have a significant impact on Alzheimer's patients and ultimately improve patient outcomes.

Keywords: Alzheimer, Diagnostic, Deep learning, CNN, Neural Network, SVM

1 Research status and background

1.1 Background

Each year, the total number of people visiting various medical institutions in China exceeds 7 billion, and the medical and health industry is facing significant pressure on service demand due to the uneven distribution of medical resources and irrational layout and structure. The use of artificial intelligence technology to support the execution of medical processes and the integration and analysis of data offers new opportunities to improve the capacity of medical and health services and solve the shortage of medical resources. Alzheimer's disease (AD) is an irreversible neurodegenerative disease that progressively destroys cognitive abilities leading to dementia Clinical diagnosis of AD is based primarily on objective cognitive

impairment (usually marked memory impairment). The clinical diagnosis of AD is based primarily on objective cognitive impairment (usually significant memory impairment). In some cases, AD may also present with atypical symptoms, such as impairments in attention, executive function, visual construction practice and language. However, AD shares many clinical features with other neurodegenerative dementias, such as Lewy body dementia, frontotemporal lobe disorders and vascular dementia, making early diagnosis and differential diagnosis difficult, especially in the early stages of the disease. The course of Alzheimer's disease is usually accompanied by a progressive decline in cognitive function. Deep learning and machine learning can be applied to analyse long-term cognitive test data, such as memory tests or cognitive tasks. By building time series models, trends in cognitive function can be detected and predicted to help diagnose and monitor the progression of Alzheimer's disease. This paper discusses the means of diagnosing Alzheimer's disease with the aid of imaging histology, and proposes a new method of assisting diagnosis by combining artificial intelligence with the extraction of its imaging features.

1.2 Research status

Computer-aided diagnosis is an interdisciplinary technology that uses computer vision and artificial intelligence will perform medical image processing to identify diseases.Current research focuses on the diagnosis of ADD using neuroimaging data (e.g., magnetic resonance imaging and positron emission tomography data), which provide high-resolution structural and functional information about the brain and provide rich features for machine learning algorithms.

Here are some common methods and applications for computer-assisted research into Alzheimer's disease:

First,Image analysis.An important task in computer-assisted Alzheimer's disease research is the analysis of brain imaging data. Using computer vision and image processing technology, features and structural information in brain images can be automatically extracted, such as the volume and morphological changes of brain regions. These analysis results can be used to study the early diagnosis of AD, monitor disease progression, and study pathological mechanisms.

Second,Data mining and machine learning.By integrating and analyzing clinical data, genomic data, proteomic data, etc., biomarkers, genetic variations and disease mechanisms related to AD can be discovered, allowing for better treatment.

Third,Database and knowledge graph.By integrating and annotating clinical data, research literature, genetic information, etc., an AD knowledge base containing rich information can be constructed. This can provide researchers with fast data query and knowledge reasoning, accelerating the progress of AD research.

Last,Clinical decision support.Provide doctors with auxiliary diagnosis and treatment suggestions by integrating patient clinical information, imaging data, biomarker data and other information.

It is well known that data preprocessing and feature extraction are key steps when performing analyses.

Preprocessing includes techniques such as grey-scale mapping and denoising for improving image quality and reducing noise interference. Feature extraction methods include both traditional image features (e.g., shape, texture, grey-scale histogram, etc.) and deep learning methods (e.g., convolutional neural networks). With the above methods, we can identify Alzheimer's disease more clearly.

1.3 Data Introduction

The choice of dataset is important for testing deep learning models. When creating a dataset, care must be taken to clean and label the data appropriately, and high quality datasets usually improve the quality of model learning and the accuracy of predictions. This section describes the data types used in this work, details why they were chosen for the study, describes how the data was processed, and explains the rationale behind the processing principles.

1.3.1 Selection of medical data types

In this project, we chose MRI images as our data set. We know that MRI has many advantages.

First, high-quality soft tissue contrast: MRI can provide high-quality soft tissue contrast, allowing doctors to observe and The fine structure that distinguishes different tissues. This is important for detecting lesions, assessing the health of organs and tissues, and guiding treatment decisions.

Second, multi-plane imaging: MRI allows imaging of the human body in different planes (such as transverse, coronal, and sagittal planes). This multi-planar imaging can provide more comprehensive anatomical information and help doctors understand the location and extent of lesions more fully.

Finally, the ability to observe physiological processes: By applying specific MRI sequences and techniques, physiological processes within the human body, such as blood flow, brain activity, and metabolism, can be observed and measured. This gives MRI a unique advantage when studying and diagnosing certain diseases, such as brain function studies and cardiovascular disease assessment.

Summuary of the studies for the diagnosis of AD using DL methods are as follow.

Table 1: Demographics of subject using MRI modality

Publication	Type of Data	Method	Performance
Oh et al.	MRI from Kaggle	CNN architecture	Accuracy of 84.5% for AD-NC binary classification
Suiaya Murugan et al.	MRI from Kaggle	DL architecture	Accuracy of 92.53% for 4 classes
.....	.....	.....	.....

Based on the results in the above table, we can draw preliminary conclusions:

First,the DL architecture of Suiaya Murugan et al. performs better on MRI image classification tasks, reaching an accuracy of 92.53%. Currently, the CNN architecture of Oh et al. (accuracy of 84.5%) has higher classification performance.

Second,these studies achieved relatively high accuracy when using Kaggle datasets for MRI image classification, indicating the potential of deep learning methods in the diagnosis of Alzheimer's disease.

1.3.2 Demographics of subjects

In this part, the data set is mainly introduced. All data in the data set are divided into four types, namely Non-Demented,Very Mild Demented,Mild Demented,Moderate Demented. The specific information is shown in the table below.

Table 2: Demographics of subject using MRI modality

Research group	Number of Subjects	Age	MMSE
Non-Demented	2560	75.23 (57.5-87.8)	29.11 (27-30)
Very Mild Demented	52	76.01 (62.2-86.6)	27.02 (24-27)
Mild Demented	717	74.54 (58.3-82.2)	24.8 (21-24)
Moderate Demented	1792	75.82 (63.8-88.7)	21.26 (17-24)

*MMSE (Mini-mental State Examination): a neuropsychological examination tool commonly used in clinical, which quantifies the test subject's mentalstate and cognitive function through scoring, and is important for the diagnosis of dementia.

1.3.3 Medical Image file format

In the project, the data format is mainly DICOM format. This format has the following advantages.

First,the pixel value of DICOM format is usually 4k, and the larger the range of gray value, the greater the difference in detail.

Second it can be directly used for deep learning by adjusting the activation function of the network, and does not need to be normalized and then learned.

1.3.4 Data Visualization

Some example images used in the experiment are shown in the figure below. There are four different images, representing four different periods of AD. By comparison, we can see the differences in MRI images at different stages of Alzheimer's disease. In the early stages of AD, extensive brain atrophy has already occurred. In addition, hippocampal atrophy and central lobular thinning are also early structures in Alzheimer's disease

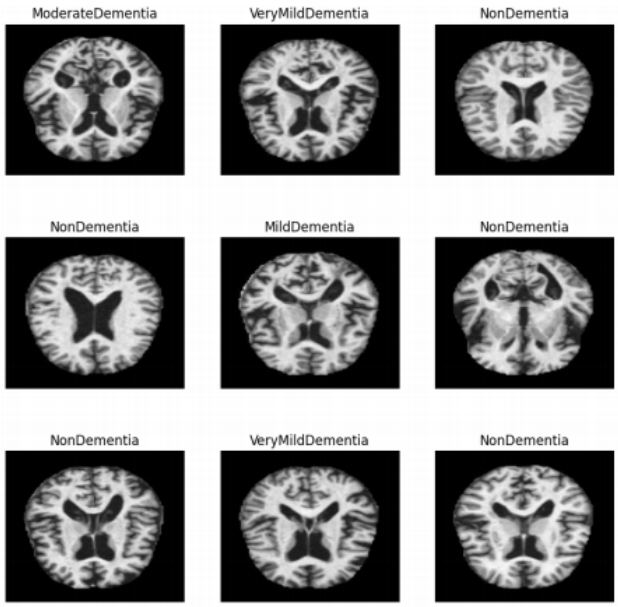

Figure 1: Enter Caption

2 Methodology

2.1 Image Data Feature Extraction

Feature extraction is a concept in image processing and computer vision. The original features are converted into a set of physically meaningful and statistically significant features. Feature description methods are divided into gray scale description, including gray scale statistics, gray scale histogram; shape description, such as area, perimeter, elongation, center of mass; texture description, such as gray scale covariance matrix, gray scale step moment array, etc. In this paper, we will use intensity features, shape features and texture features to extract 17 features.

(1) Intensity feature extraction: mean, median, maximum, minimum, standard deviation, skewness, peak.

(2) Shape feature extraction: area, perimeter, roundness, extension rate, center of mass.

(3) Texture feature extraction: contrast, correlation, entropy.

The selection of intensity features and shape features is derived from the experience of clinical radiologists, and the intensity features are statistically based on the gray-scale image within the mass contour, and the shape features are statistically based on the intersecting-order image within the mass contour. The texture feature extraction method is based on the description of texture features using Gray-Level Co-occurrence Matrix (GLCM) proposed by Haralick et al. Texture is a property of an image region that is related to the gray values of spatially adjacent pixels. Texture features can reflect the spatial information, structural information, and gray-scale statistical information of the image.

2.2 Feature Selection

The purpose of feature selection is to select representative and effective features. When feature extraction is performed, different extraction methods are able to obtain a large number of features.However, there are a large number of redundant and irrelevant features as well as noise in the extracted features. These redundant features add unnecessary burden, reduce the operational efficiency of the machine learning's model and decrease the classification accuracy. Through feature selection, useful features can be selected and the number of feature dimensions can be reduced to avoid overfitting of the machine learning model and improve the performance, classification accuracy and generalization ability of the machine learning model. Feature selection also helps to understand the data characteristics and underlying structure, and enhances the understanding of feature values and between features. Feature selection methods are broadly categorized into three types: filtering methods, packing methods, and embedding methods.

In this experiment, the extracted features are ranked by statistical methods, and the commonly used statistical methods are chi-square test, μtest, $t$test, and Wilcoxon rank sum test. The test requires the data to satisfy the assumptions of normality and chi-square, while the Wilcoxon rank sum test is a non-parametric method and is chosen when the data does not satisfy the normal distribution.The Wilcoxon rank sum test is a method of hypothesis testing that utilizes the rank sum as a statistical measure.In this paper, the features with p-value less than 0.05 are considered as features with significant difference, and those with p-value greater than 0.05 are considered as non-significant features, and the 17 features are ranked in ascending order according to the value of p-valve.

As shown in Table3, among these 17 features, 12 features have p-value less than 0.05, except for skewness. Among these 17 features, 12 features have p-value less than 0.05, and except for skewness, the other 11 features have p-value less than 0.01. Among them, 5 features belong to intensity features, 5 features belong to shape features, and 2 features belong to texture features.

2.3 LeNet-5 CNN

2.3.1 Basic structure

LeNet-5 is one of the most classical convolutional network models, with seven layers and a 32*32 size image as input. The pooling layer operates as a region summation multiplied by a parameter plus a bias term, and both the parameter and bias can be trained. The network activation function is Sigmoid function. The structure of each layer of the network is shown in the table4.

2.3.2 CNN training and testing

(1) Parameter initialization and training algorithm settings

a.Parameter initialization

Table 3: Feature Sorting

Feature	p-value	Feature sorting
s_perimeter	0.0000000001	9
s_form	0.0000000013	14
s_circularity	0.0000000018	12
t_entropy	0.0000176943	17
s_area	0.0000269112	8
i_median	0.0000607978	2
i_mean	0.0001037326	1
t_correlation	0.0001659581	16
i_maximum	0.0016030480	4
i_standard_deviation	0.0018924330	3
s_y_center_mass	0.0022418770	11
i_skewness	0.0058760510	7
t_contrast	0.0646476000	15
s_elongation	0.1837025000	13
i_kurtosis	0.3132612000	6
i_minimum	0.5998737000	5
s_x_center_mass	0.7616871000	10

Since this paper uses a gradient descent algorithm in backpropagation, parameter initialization before training starts is very important. If the initialization values are at the smoother part of the error surface, this will lead to slow convergence of the error. In this paper, the initialization weights are distributed as follows.

$$U\left[\frac{\sqrt{6}}{\sqrt{p^{(l)}+p^{(l-1)}}}\frac{\sqrt{6}}{\sqrt{p^{(l)}+p^{(l-1)}}}\right] \tag{1}$$

b.Forward Propagation

Forward propagation is mainly divided into two stages: through the convolutional layer and through the pooling layer.

Forward propagation through convolutional layer:When the input data passes through the convolutional layer, the output feature maps of the previous layer are convolved with the convolutional kernel of this layer for the convolution operation, and all the feature maps produced by the convolution are combined and summed up, and then the final result is passed through an excitation function, and a new feature map can be produced.

The formula for forward propagation is as follows.

$$x_j^l = s(\sum_{i\in M_j} x_i^{l-1}\cdot k_{ij}^l + b_j^l) \tag{2}$$

The s in the formula is the excitation function of this CNN. The excitation function is a nonlinear factor,

Table 4: Structure of LeNet-5

Layer	Input	Kernal size and number	Stride	Output
Convolutional layer C1	128*128*1	13*13*16	1	116*116*16
Pooling S2	116*116*16	2*2	2	58*58*16
Convolutional layer C3	58*58*16	15*15*16	1	44*44*16
Pooling S4	44*44*16	2*2	2	22*22*16
Convolutional layer C5	22*22*16	15*15*64	1	4*4*64
Fully connected layer FC6	4*4*64	84 nodes in total		1*1*84
Fully connected layer FC7	1*1*84	84 nodes in total		1*1*2

and the inclusion of the excitation function can enhance the expressive ability of the CNN's nonlinearity. The ideal CNN excitation function is a step function, where the input value is passed through the excitation function to get an output value of "0" or "1", where "0" represents that the neuron is inhibited, and "1" represents that the neuron is inhibited. where "0" means the neuron is inhibited and "1" means the neuron is excited. This excitation function acts most like the transmission of neurons in the brain.

However, the non-smooth and discontinuous nature of the step function makes its derivatives infinite, and CNN backpropagation is hindered, reducing the performance of the neural network. So it is not possible to use the step function as an excitation function. In this paper, we use Sigmoid function as the excitation function of CNN, and the formula of Sigmoid function is as follows.

$$s(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

The function of Sigmoid function is that it can compress all the input values into the range of 0-1, so Sigmoid function is also called squeezing function. Figure 2 shows the shape of the curve of the Sigmoid excitation function of the CNN. Compared with the step function, the Sigmoid function is conductible everywhere, so there is no case of infinite derivatives.

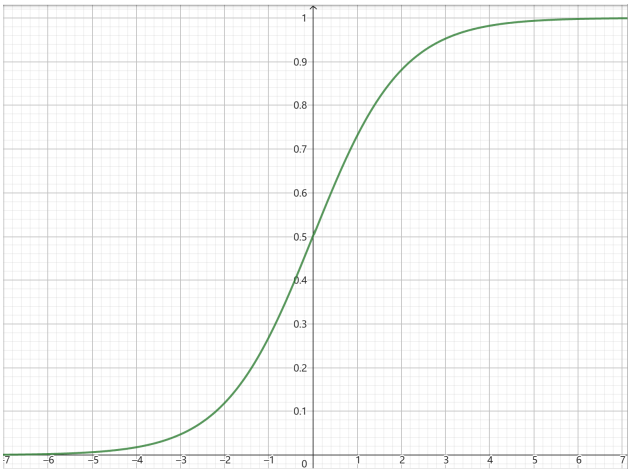

Figure 2: Sigmoid excitation function graph

Forward propagation through the pooling layer:When the output feature graph of the previous layer

passes through the pooling layer, sum the values in the region of the input feature graph and add another bias term, then pass the result through the Sigmoid excitation function, and finally the output feature graph and the input feature graph have the same number of features, but the size of each output feature graph becomes half of the original one, and the expression of this formula is as follows.

$$x_j^l = s(\beta_j^l down(x_i^{l-1}) + b_j^l) \tag{4}$$

$down(\cdot)$ is the pooling function, i.e., summing over mutually non-overlapping 2x2 regions of the input feature map. $s(\cdot)$ is the Sigmoid excitation function.

c.Backward propagation

In this paper, CNN updates the weights and bias by using gradient descent algorithm. And the loss function used in this paper is the common squared error function with the following formula.

$$E^N = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}(t_k^n - x_k^n)^2 \tag{5}$$

where c is the number of labels, N is the number of samples, x_k^n is the nth dimension of the output of the nth sample in the sample, and t_k^n is the kth dimension of the label carried by the nth sample in the sample.

To compute the gradient of the weights of the neurons in the convolutional layer, the residual δ^l of each node is computed first, and the computation of this residual requires the summation of the residuals of the lower layer, δ^{l+1}, because there is a down-sampling layer, and in order to efficiently compute the δ^l, it is necessary to perform the up-sampling operation of the residual graph of the pooling layer. Since the forward computation sums each 2x2 region, each pixel is copied 2 times horizontally and vertically. The upsampling operation done doubles the size of the residual map. Then multiply the weights W between the convolution layer and the neighboring sampling layer, and multiply the derivative of the excitation function of this layer to get the residual value of this layer δ^l. The process function expression is as follows.

$$x_j^l = \beta_j^{l+1}(s'(u_j^l) \cdot up(\delta_j^{l+1})) \tag{6}$$

where up denotes the up-sampling operation.

For each feature map of the convolutional layer, its residuals can be calculated, and the gradient values of its convolutional kernel and bias can be obtained through the residual map, while the weights can be updated through the gradient to complete the learning process. The function expression for calculating the gradient is as follows.

$$\frac{\partial E}{\partial b_j} = \sum_{u,v}(\delta_j^l)_{uv} \tag{7}$$

$$\frac{\partial E}{\partial k_{ij}^l} = \sum_{u,v}(\delta_j^l)_{uv}(p_i^{l-1})_{uv} \tag{8}$$

The u, v denotes the coordinates of the feature map, and $(p_i^{l-1})_u v$ is the region where the elements are multiplied during the convolution operation.

The forward propagation of the pooling layer, because it involves only two bias parameters β and b, and the feature maps of each stage are saved during the forward propagation, its gradient can be calculated directly during the back propagation and updated to the bias parameters,with the following functional expression.

$$\frac{\partial E}{\partial b_j} = \sum_{u,v} (\delta_j^l)_{uv} \tag{9}$$

$$\frac{\partial E}{\partial \beta_j} = \sum_{u,v} (\delta_j^l) \circ down(x_j^{l-1}))_{uv} \tag{10}$$

2.3.3 Training and testing process

In this paper, we use MRI fusion image set to train and test CNN to get the assisted diagnosis of AD. Figure3 shows the flowchart of CNN to diagnose AD.

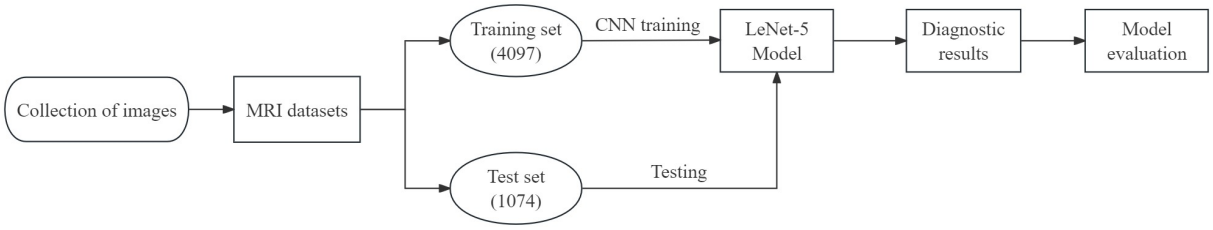

Figure 3: CNN Diagnostic Flowchart

Firstly, the designed CNN suitable for diagnosing AD is duplicated three times to get three models of convolutional networks with the same structure. Then the optimal model for AD diagnosis is selected based on the training test results.

3 Structural Algorithms and Improvement of Neural Networks

This section introduces the two CNN models used in the experiments: AlexNet and GoogleNet.The experiments in this chapter are based on these three models to design, build, and improve the CNN, and train the CNN based on the MRI unimodal images, and then extract the features of the MRI images combined with the clinical scale array and introduce the Support Vector Machines (SVMs) for multi-classification respectively.

3.1 AlexNet Structure Improvement

In MRI images, regions such as hippocampus and cingulate gyrus are the key areas to distinguish diseases. Because the hippocampus of Alzheimer's patients is severely atrophied compared to normal people, the number of convolutional kernels or the number of convolutional layers are increased according to the pathology and the actual test in order to make a better diagnosis. After reading the relevant references, we

found that the third and fourth layers of the original AlexNet model are more capable of extracting features in this area. Although the parameters of the third and fourth layers are exactly the same, the structure leads to better classification in the fourth layer compared to the third layer.

Therefore, in this paper, we try to add a layer after the fourth layer with exactly the same parameters as it to enhance the ability of the network to extract features. The adding part of structure of the improved AlexNet network is shown in Table5.

Table 5: Structure of improved AlexNet

Layer		Input	Kernal size and number	Stride	Output
3rd layer	Convolutional layer C3	13*13*256	3*3*384	1	13*13*384
4th layer	Convolutional layer C4	13*13*256	3*3*384	1	13*13*384

3.2 GoogLeNet Structure Improvements

GoogLeNet has fewer parameters and higher accuracy than AlexNet. Optimization of GoogLeNet is done using a multi-channel weighting approach.GoogLeNet outputs the final diagnostic result using only a single classifier in the top output layer and discarding the two bottom auxiliary classifiers. In some classification problems, relying only on the classification results of the top classifier can lead to a reduction in classification accuracy, where the top classifier results are wrong and the auxiliary classifiers are correct. Moreover, different levels of features are different levels of abstraction of the original data, and the fusion of these different levels of features can improve the accuracy and reliability of the network.

Therefore, we add inception structures to each layer. The inption results are the same for different layers, only the size and number of convolution kernels are different. Each module of inception has four channels, widening the width of the network. inception's structure is shown in Figure4

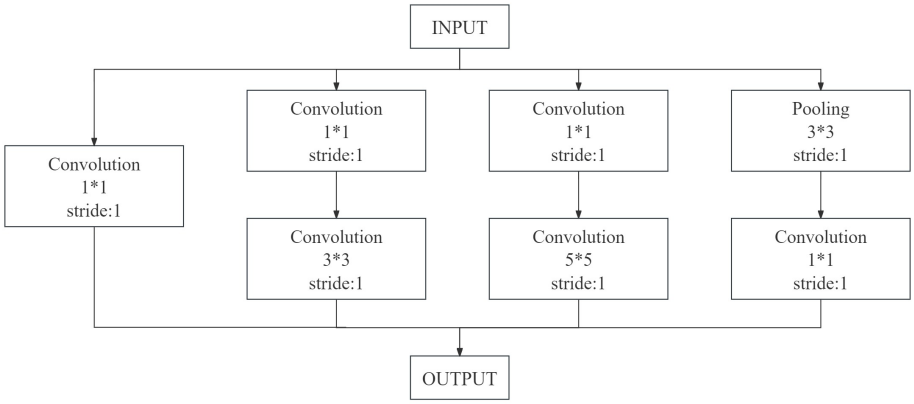

Figure 4: Structure of Inception

4 Classifier Construction

4.1 Support Vector Machine

Support vector machine is a supervised learning method for classification, regression, and outlier detection. Support vector machines are based on statistical learning theory, VC dimension theory and structural risk minimization theory. Support vector machines can overcome the problems of "dimensional catastrophe" and "overfitting", and have advantages in solving small samples, nonlinear classification and high-dimensional pattern recognition. Support vector machines are widely used in pattern recognition, face image recognition, regression analysis, handwriting font recognition, gene classification and other fields.

The basic idea of support vector machine is to map the training dataset nonlinearly to a high-dimensional feature space, so that the linearly indivisible dataset will be linearly divisible in the high-dimensional feature space, so as to find the optimal hyperplane in the high-dimensional feature space, which makes the training samples the farthest distance to the optimal hyperplane.

Taking dichotomous data as an example, the training sample set $(x_i, y_i), i = 1, 2, \ldots, l, x \in R^n, y \in \pm 1$ hyperplane is.

$$(w \cdot x) + b = 0 \tag{11}$$

In order for the classification plane to correctly classify all samples and have a classification interval, it needs to be constrained as follows.

$$y_i[(w \cdot x_i) + b] \geq 1 \tag{12}$$

The classification interval is $\frac{2}{||w||}$,and the problem of constructing the optimal hyperplane can be transformed into solving under the constraint equation:

$$min\phi(w) = \frac{1}{2}||w||^2 = \frac{1}{2}(w' \cdot w) \tag{13}$$

Introducing the Lagrangian function.

$$L(w, b, a) = \frac{1}{2}||w|| - a(y((w \cdot x) + b) - 1) \tag{14}$$

The classification performance of a support vector machine is determined by many influencing factors, two of the key factors are the error penalty parameter C and the kernel function and the parameters of the kernel function.

The error penalty parameter C is a balance between the complexity of the algorithm and the proportion of samples that are misclassified, which can also be interpreted as the tolerance of classification errors. The higher the error penalty parameter C, the less classification error is allowed, which may lead to classification overfitting of SVM; the lower the error penalty parameter C, the more tolerant the classifier is to classification errors, which may affect the classification performance of the classifier on linearly indistinguishable data.

The kernel function is critical for support vector machines because samples are difficult to discriminate in low-dimensional space, so they are usually mapped to high-dimensional space. However, the computa-

tional complexity of high dimensional space is very large, which can be reduced by kernel function. Different kernel functions have different classification results, and the same kernel function with different parameters has different classification results. The introduction of the kernel function only requires the inner product operation of the feature space in the input space, and the inner product in this low-dimensional space is equal to the inner product of the higher dimensions, which in turn greatly reduces the amount of computation. The commonly used kernel functions are polynomial kernel function, linear kernel function, radial basis function, two-layer neural network basis kernel function. The polynomial kernel function maps the low-dimensional input space to the high-dimensional feature space with the following formula.

$$K(x, x_i) = e^{(-y \times ||x - x_i||^2)}$$

where y is the kernel function width.

4.2 Artificial neural network

Artificial neural network is a mathematical model for distributed parallel processing of information that mimics the behavior of neural network of biological brain proposed by W. Pitts and W. Mculoch in 1943, with self-learning ability and memory ability. This arithmetic model consists of a large number of nodes interconnected with each other, and each node represents a specific output function called the excitation function (Activation Function). The connection between every two nodes represents the weighted value of the signal passing through the connection. Artificial neural networks are divided into multi-layer and single-layer. Artificial neurons are shown in Figure 5.

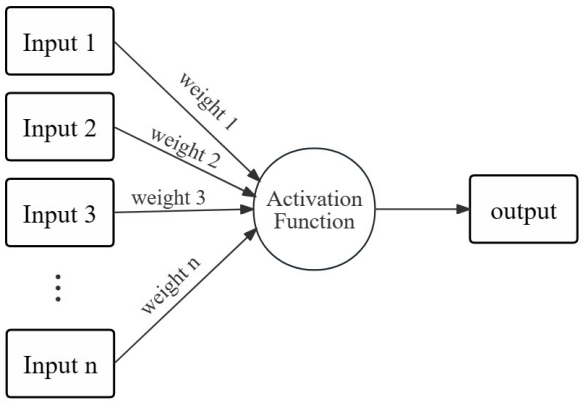

Figure 5: Artificial Neurons Diagram

The artificial neural network network feeds the signal to the neurons located in the input layer, the output of the neurons in this layer is weighted to become the input of the neurons in the next layer, until finally the neurons in the output layer of the network output the result. Afterwards, the error between the output value and the expected value is calculated, and the weights of the network are continuously adjusted through the

back propagation algorithm to realize network training.

Artificial neural networks have four basic characteristics.

(1) non-linear, manifested as the activation state and inhibition state of artificial neurons are a mathematical non-linear relationship; (2) non-linear, manifested as the activation state and inhibition state of artificial neurons are a mathematical non-linear relationship; (3) non-linear.

(2) non-limitation, artificial neural network consists of many small processing units with simple functions connected to each other, the overall behavior of the neural network does not depend on a single neuron but on a large number of small neuron interconnections and interactions; (3) non-limitation, artificial neural network consists of many small processing units with simple functions connected to each other.

(3) very qualitative, artificial neural networks are not fixed information processed, the system itself is constantly changing, iterative process is used to describe the system evolution process; (4) non-qualitative, artificial neural networks are not fixed information processed, the system itself is constantly changing.

(4) Non-convexity, as the function has multiple extremes, the system has multiple smooth states.

In recent years, artificial neural networks have been widely used in artificial intelligence, pattern recognition, signal processing, computer vision and other fields.

4.3 Flowchart of the Diagnostic Process

Therefore, the flowchart of the whole diagnostic process is shown in the figure6.

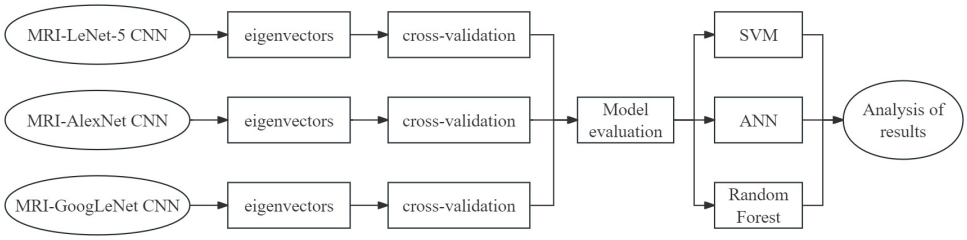

Figure 6: Flowchart of the Diagnostic Process

In the end, five rounds of 5-fold cross-validation were performed for each CNN test. The specific results will be analyzed in the next section.

5 CNN Algorithm Improvement

5.1 Dropout Layer

Since it is currently difficult to obtain large amounts of medical data of high quality that have been classified by alignment, i.e., it is not possible to sample the complete heterogeneous medical information of each individual on a large scale, the addition of a Dropout layer prevents the network from rapidly overfitting.

In machine learning, variational inference can be used to approximate the problem of estimating some complex distribution using some simple form of distribution. The more common variational Bayesian obtains an optimal solution through multiple iterations of a set of interdependent equations.

In a Bayesian network the function is defined by the neural network weights, let the sufficient statistics be.

$$w = (W_i)_{i=1}^{L} \tag{15}$$

Then the weighted posterior probability after assignment is.

$$X, Y : p(w|X, Y) \tag{16}$$

This posterior is more manageable for Bayesian networks using variational inference approximation. To relate approximate inference in Bayesian neural networks to dropout training, the approximate variational distribution $q(Wi)$ for each layer is defined as.

$$W_i = M_i \cdot diag([z_{i,j}]_{j=1}^{K_i}) \tag{17}$$

$$Z_{i,j} \backsim Bernoulli(p_i) \quad for \quad i = 1, ..., L, j = 1, ..., K_{i-1} \tag{18}$$

$Z_{i,j}$ is a Bernoulli distributed random variable with some probability p_i, and M_i is the variational parameter to be optimized. The $diag$ operator maps vectors to diagonal matrices whose diagonals are elements of the vectors.

A direct result of the theoretical developments in the previous section is that approximate variational inference can be achieved by adding dropout layers after certain weight layers in a Bayesian neural network. The Bayesian neural network performs dropout after each layer that has an approximate distribution at training time and evaluates the prediction posterior at testing time using the following formula.

$$p(y^*|x^*, X, Y) \approx \int p(y^*|x^*, w)q(w)dw \approx \frac{1}{T}\sum_{t=1}^{T} p(y^*|x^*, w_t)) \tag{19}$$

In Bayesian neural networks, all weight layers are usually approximated by integrating the weights using the distribution modeling posterior distribution as a regularization parameter, and weight layers without approximated distributions usually lead to overlap.Dropout can also be applied after the CNN convolutional and fully-connected layers, but several studies have shown that applying dropout after convolution does not work as well when tested with the standard dropout However, several studies have shown that applying dropout after convolution is not ideal when testing with standard dropout. In this paper, we address this issue by approximating the predictive distribution of the equation as above.

5.2 Selection of activation function

The role of the activation function is to enable the network to incorporate nonlinear factors to facilitate the solution of analytical processing that cannot be solved by a linear network model. The activation function is categorized into saturated function and unsaturated function, the use of unsaturated activation function can

solve the problem of gradient disappearance and accelerate the convergence speed.

ReLU function and ELU function are non-saturated activation functions, ReLU function means modified linear unit, the function representation is shown in the figure below, ReLU function requires all negative values in the input matrix are set to zero, the rest of the values remain unchanged.

ELU is an exponential linear unit, the function is shown in the figure below, which makes the average value of the activation function as close to zero as possible, thus speeding up the learning efficiency. It also has an output in case of negative input and also has some anti-interference ability.

The training results of ReLU function and ELU function are shown in the figure7, where the X-axis indicates the number of iterations, and the Y-axis shows the training loss.

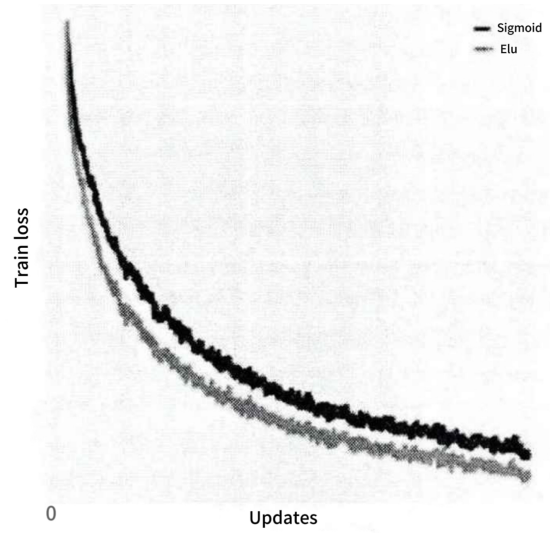

Figure 7: Elu function and ReLU function loss and iteration number relationship

According to the test results, the classification accuracy of ELU is slightly higher than that of ReLU, therefore, this paper chooses to add the ELU function as an activation function to the convolutional layer.

6 Model Training and Evaluation

In this section, we will introduce in detail the model training strategy used in Alzheimer's disease diagnosis research based on imaging and deep learning. Our goal is to extract key features from imaging data that facilitate Alzheimer's disease diagnosis through efficient deep learning methods.

The entire model training process mainly includes data set partitioning, neural network (CNN) construction, loss function and optimization algorithm selection, model training, hyperparameter tuning, and model evaluation and performance analysis.

6.1 Dataset partitioning

We divided the dataset into training and validation sets, allocating 80% of the data to the training dataset and reserving the remaining 20% for the validation dataset.

The number of pictures contained in the training datasets and validation datasets are respectively:

(1) training datasets: 4097 pictures

(2) validation datasets: 1024 pictures

6.2 Model Training

In order to further improve model performance, we used some optimization strategies and also performed system hyperparameter tuning, including adjustments to key parameters such as learning rate and batch size, to optimize the convergence speed and performance of the model.The specific process of model training strategies is as follows:

(1) Use callbacks to adjust the learning rate and stop training after the model has converged.

(2) Save the best weights of the model to a file at the end of each training epoch, and only save the best model.

(3) When the validation loss value during training does not improve in 10 consecutive epochs, training is stopped early and restored to the optimal weights.

(4) Learning rate decay is performed according to the exponential decay learning rate function.

6.3 Visualize Model Metrics

As our dataset is unbalanced, we cannot use accuracy as a metric. Therfore we used ROC AUC(Receiver Operating Characteristic Area Under the Curve)to evaluate our model.The ROC AUC indicator is the area under the ROC curve. The value range is between 0 and 1. The larger the value, the better the performance of the model. Figure8 shows the area under the ROC curve.

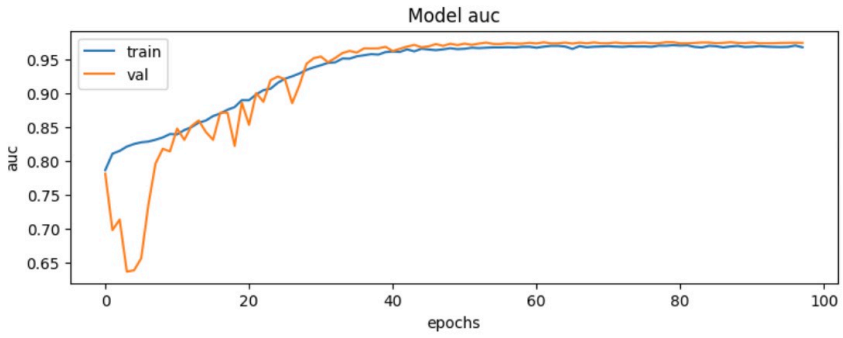

Figure 8: Model ROC AUC

The loss graph is the change of the loss value during the training process. As the training proceeds, the loss value should gradually decrease, indicating that the model is learning the data pattern and improving the accuracy.From Figure9, we can see that the model's loss value has a downward trend, and after roughly 30 epochs, the model's loss value has converged.

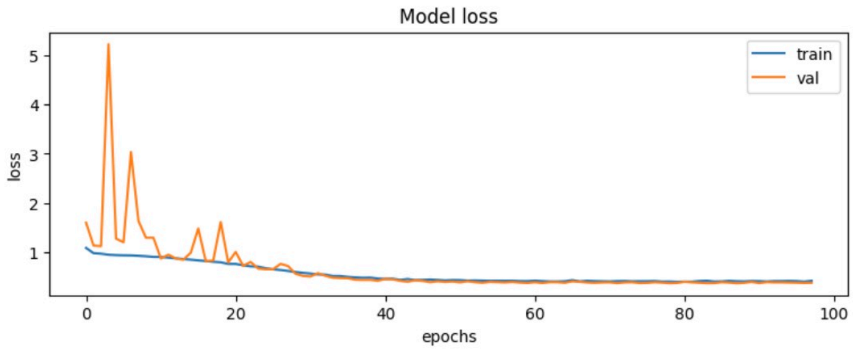

Figure 9: Model Loss

6.4 Model Evaluation

In the model evaluation part, we chose to use cross-validation to test the accuracy of the data set. Through cross-validation, we are able to train and evaluate the model on multiple subsets, which helps to alleviate the model's overfitting to specific data distributions and improve the model's generalization ability.

Table 6: Accuracy Comparison of CNN Models

Training Round	LeNet-5	CNN AlexNet	CNN GoogLeNet
1st round	81.88%	85.58%	87.54%
2nd round	82.46%	84.45%	88.01%
3rd round	81.98%	85.24%	89.54%
⋮	⋮	⋮	⋮
Average Accuracy	82.45%	85.78%	88.64%

Table 5 shows the Accuracy Comparison of CNN Models, and we can clearly find that CNN GoogLeNet has the highest average accuracy.

7 Contribution and Future Directions

7.1 Contribution

Table 7: Contribution distribution

No.	Task		Owner	Due Day
1	Draft outline		LIYuchen	5/10/2023
2	Data download and data preprocessing		ZhangQi,ZhouHan	8/10/2023
3	Classifier and neural network structural modeling		LIYuchen	11/10/2023
4	Programming and code running		ZhangQi	20/10/2023
5	First draft writing section	Background of the project Data presentation and image data preprocessing	ZhouHan	12th week
		Classifier Construction CNN-based neural network construction and model training Improvement of neural network structure	LIYuchen	
		Experimental validation and evaluation Analysis of experimental results	ZhangQi	
6	Beamer production	Complete the corresponding Beamer section based on the first draft writing section		6/11/2023
		Beamer Integration	LIYuchen	

7.2 Future Directions

In this paper, for the optimization of data processing and network there are still a lot of places that can be improved and optimized through more practical experiments, and at this stage, the research work is still in the beginning stage, and there will be the following research directions:

(1) In terms of clinical data, EEG and DTI images can be added to assist in diagnosis, so as to improve the accuracy of diagnosis in all aspects. However, this requires the improvement of personal medical records and the unified planning of electronic medical records. In the experiments conducted in this paper, different electronic medical records and scale data as well as imperfect patient information made it very difficult to filter and process the data.

(2) In terms of network, CNN still has room for improvement. In addition to the improvement of making the network more adapted to medical diagnosis, more performance-enhancing improvements also need to be added. At present, the accuracy of CNNs in medical data diagnosis is still not up to the accuracy based on ImageNet, Cifar-10 and other databases, so further optimization of structure and algorithm is the next research direction.

8 References

[1] Qixiao Zhu. Research on Alzheimer's disease image analysis method based on functional brain network and machine learning.2021.Shandong University,MA thesis.

[2] Zuchen. Brain image analysis based on sparse structural feature learning and its applications.2018.Nanjing University of Aeronautics and Astronautics,PhD dissertation.
[3] Qiao Yingfang. Research on Alzheimer's disease-oriented feature selection and classification methods [D]. Shandong Normal University,2018.
[4] Wang Y. Artificial Intelligence and Multimodal Imaging Histomics in the Analysis of Several Diseases
[5] Northeastern University, 2019.DOI:10.27007/d.cnki.gdbeu.2019.000179.
[6] Shuihua Wang. Research on several techniques of multi-scale brain images based on machine vision [D]. Nanjing University,2017.
[7] AHSAN BIN TUFAIL.Research on early diagnosis of Alzheimer's disease based on neuroimaging and deep learning.2022.Harbin Institute of Technology,PhD dissertation.