摘　　要

阿尔兹海默症是最常见的老年疾病，俗称为老年痴呆。其治疗难度大、周期长且费用高昂，目前还未找到有效的治愈方法。在病理学中研究发现，阿尔兹海默症在被确诊之前具有较长的潜伏期，称为轻度认知障碍阶段。此阶段若及时被发现并加以科学的预防干预和规范的辅助治疗，可以极大的减缓其转化为阿尔兹海默症的速度和降低其转变为阿尔兹海默症的几率。因此如何科学且精准的诊断出处于轻度认知障碍阶段的患者对于阿尔兹海默症识别的研究具有重要意义。近年来，深度学习的飞速发展为阿尔兹海默症识别的研究提供了一个新的视角。本文主要基于深度学习技术使用语音模态的 NCMMSC2021 数据集和图像模态的 PET2020 数据集，从以下两个方面来开展多模态表示学习的阿尔兹海默症识别的研究：

（1）基于语音模态的阿尔兹海默症识别研究。本文在前人相关的研究基础上，从语音模态数据对阿尔兹海默症识别进行了更全面的探究，选用声谱图、梅尔声谱图和梅尔频率倒谱系数作为音频特征，针对单一音频特征及其特征融合情况，利用基于不同预训练网络设计的分类模型展开了研究。

（2）基于图像模态和集成学习的阿尔兹海默症识别研究。本文提出了利用语音特征转化为图像的方法和利用预训练模型来研究阿尔兹海默症的识别。在此基础上，对转化得到的图像进行了简单有效的去噪处理。此外，鉴于阿尔兹海默症识别研究中往往数据较难获取，且数据规模偏小，在大规模预训练模型中可能会存在过拟合的问题。本文采用了集成学习的方法来降低过拟合的影响，通过使用投票集成学习，识别性能得以进一步提升。

本文所提出的方法在 NCMMSC2021 数据集中取得了 0.9244 的准确率，远高于数据提供方发布的基线模型 0.7980 的准确率，也优于百度研究院 0.8992 的准确率；该方法在 PET2020 数据集中也取得了 0.9950 的准确率，显著优于数据提供方发布的基线模型 0.8640 的准确率。这些实验结果充分证明了本文所提出方法的优异性和普适性。为加强理论研究与工程实践的紧密联系，考虑到当前阿尔兹海默症的临床诊疗现状，最后本文将所提出的模型设计成阿尔兹海默症识别系统，并对系统的设计理念、实现方法、前端主要界面进行了阐述。

关键词：深度学习；语音模态；图像模态；阿尔兹海默症；集成学习.

Abstract

Alzheimer's disease is the most common senile disease, commonly known as senile dementia. Since its treatment is difficult, long period and expensive, there is no effective cure so far. Studies in pathology have found that Alzheimer's disease has a long incubation period before diagnosis, called mild cognitive impairment stage. At this phase, if it can be timely discovered and handled with scientific prevention intervention and standard adjuvant therapy, it will greatly slow down the speed and probability of its transformation into Alzheimer's disease simultaneously. Therefore, how to scientifically and accurately diagnose patients at the stage of mild cognitive impairment is of great significance for the research of Alzheimer's disease recognition. Recently, the rapid development of deep learning provides a new perspective for the research of Alzheimer's disease identification. This thesis mainly adopts deep learning-based technology to carry out research on Alzheimer's disease recognition via multimodal representation learning on the benchmark NCMMSC2021 data set of speech mode and PET2020 data set of image mode from the following two aspects:

(1) Research on the recognition of Alzheimer's disease based on speech modes. In this thesis, on the basis of previous studies, the recognition of Alzheimer's disease is explored more comprehensively from the modality of speech. Spectrogram, Melspectrogram, and Mel-Frequency Cepstral Coefficients are selected as audio features. For single audio features and feature fusion, classification models based on different pre-training networks are adopted.

(2) Research on Alzheimer's disease recognition based on image modal and ensemble learning. In this thesis, we propose a method of transforming speech features into image and pre-training model driven to study the problem of Alzheimer's disease recognition. Based on these, we adopt a simple but effective method to conduct noise reduction for the converted image. In addition, since it is difficult to obtain Alzheimer's disease corpus and the dataset size is small, there is a overfitting problem in current large-scale pre-training models. In this thesis, ensemble learning is adopted to reduce the influence of overfitting, and the performance is further improved by using voting ensemble learning.

The proposed method achieves the best accuracy 0.9244 on the NCMMSC2021 dataset, which is much higher than the baseline released by the data provider with

accuracy 0.7980 and the accuracy 0.8992 from Baidu Research Institute. The proposed method also achieved the accuracy 0.9950 in the PET2020 data set, which was also much higher than the baseline model released by the data provider with the accuracy 0.8640, which fully verified the effectiveness and universality of the proposed methods in this thesis. In order to strengthen the close connection between theoretical research and engineering practice, considering the current status of clinical diagnosis and treatment of Alzheimer's disease, we finally designed the proposed models into an Alzheimer's disease recognition system, and introduced the design concept, implementation method and front-end interface of the system.

Key words: Deep learning; Speech modality; Image modality; Alzheimer's disease; Ensemble learning.

目　　录

1 绪　论

1.1 研究背景及意义

1.1.1 研究背景

随着经济的发展和时代的进步，人们的生活质量和医疗水平均有了显著提高，对健康问题的关心也达到了前所未有的高度，都渴望能获得高质量的医疗条件。当今人口老龄化问题愈发严重，且这一现象一旦出现将难以逆转，全球不少国家已经处于或即将步入老龄化社会，这意味着将会出现更多的社会难题，最直接的就是预防和治疗老年疾病问题。老年疾病种类繁多且当今还有年轻化的趋势，治疗又通常较为困难，因而会给患者带来很大的经济负担，给家庭会带来沉痛的一击，也会极大地消耗社会公共医疗资源。因此，如何预防和治疗老年人的常见疾病成为当今社会医学研究上的难题之一。

阿尔兹海默症（Alzheimer's Disease，AD）[1]是一种由于大脑的神经细胞死亡而造成的神经变性疾病，普遍发病在老年人群体中，民间俗称为老年痴呆，是目前全球最常见的神经退行性疾病，在六十五岁以上人群中具有较高的发病率，且随着年龄的增长阿尔兹海默症的发病率呈上升趋势[2]。阿尔兹海默症会逐渐损伤患者的大脑神经元，导致其认知、交流和执行能力逐渐下降，同时还极有可能伴随着健忘、失语、失明等外在症状，若没有及时得到有效防范可能会引起并发症导致生命危险[3]。阿尔兹海默症确诊前通常具有很长的潜伏期，又因其是一种多因素疾病，故而患者及其家属在初期难以察觉，这一疾病在老年人群体中覆盖面广且对日常生活影响巨大，严重影响了患者的健康和情绪，甚至在一定程度上是影响社会的不稳定诱因之一。根据调查显示，2021年全球阿尔兹海默症患者数量超过5200万人，预计到2050年全球阿尔兹海默症患者将达到1.05亿人，在阿尔兹海默症的医学研究和治疗费用也将高达1.1万亿美元[4]。由此可见，阿尔兹海默症已经成为非常严重的公共卫生问题，其治疗起来难度大、周期长、费用高昂，对家庭、社会和政府都造成了很大的负担[5]。由于目前对其的病理学研究还较为有限，没有找到权威和有效的治疗手段，因此诊断出阿尔兹海默症的早期阶段并加以干预治疗，对防止或减缓患者病情的进一步恶化能起到至关重要的作用。

研究表明依据阿尔兹海默症的病理学特征可以将其分为三个阶段，分别为健康老年化（Healthy Controls，HC）、轻度认知障碍阶段（Mild Cognitive Impairment，MCI）和阿尔兹海默症阶段。轻度认知障碍阶段是处于健康人和阿尔兹海默症的过渡阶段，此时大脑结构暂未发生明显变化，已经具有客观的认知受损但对日常

生活能力影响甚微，具有很强的隐蔽性，即使患者和家属有所察觉，但可能往往以为是自然衰老带来的正常变化，不同阶段阿尔兹海默症患者的大脑正电子发射断层扫描影像(Positron emission tomography，PET)见图1-1，从左到右分别是健康老年化、轻度认知障碍阶段、阿尔兹海默症阶段，可以清楚的看到不同阶段大脑的内部变化。据医学调查显示，在65岁以上群体中，健康人平均每年转化为阿尔兹海默症的几率仅仅为1%左右，但处于轻度认知障碍阶段患者平均每年转化为阿尔兹海默症的几率可高达15%，5年内转化为阿尔兹海默症的几率甚至高达32%。由此可见及时诊断出处于轻度认知障碍阶段的老年人，并对其加以科学的预防干预和规范的辅助治疗，可以有效的减缓其转化为阿尔兹海默症的速度和降低其转化为阿尔兹海默症的几率[6]。因此如何科学且准确的诊断出处于轻度认知障碍阶段的患者对于阿尔兹海默症的识别研究具有重要意义。

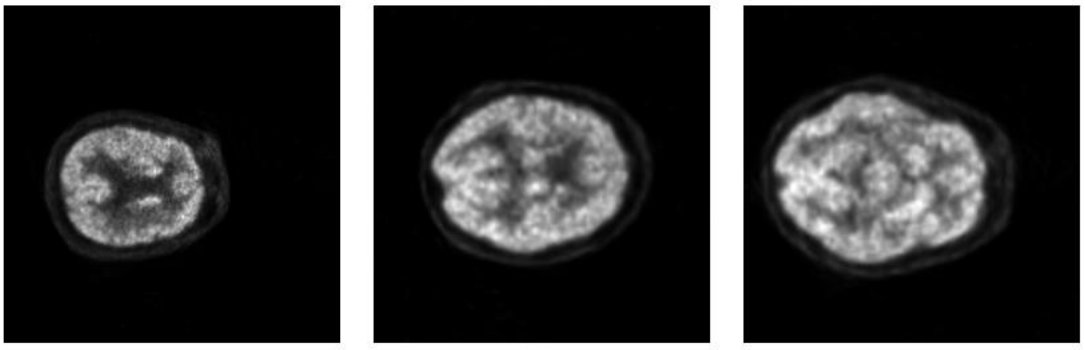

图1-1 不同阶段大脑PET影像图

目前阿尔兹海默症临床上对患者的识别方式主要是通过医护人员进行神经心理学测试、查看脑部影像图和大脑生物标志检测来判断。神经心理学测试是检测患者的认知能力、语言流利性、逻辑连贯性、说话停顿以及重复词的出现频率。脑部影像图就是利用已经成熟的医学成像技术辅以识别，此类技术已经成功运用在多种疾病的检测，在阿尔兹海默症检测中最常见的医学成像技术有大脑正电子发射断层扫描、核磁共振成像（Magnetic resonance imaging，MRI）和弥散张量成像 (diffusion tensor imaging，DTI)等，通过这些技术手段，可以直观的查看大脑的组织结构是否发生改变。大脑生物标志检测主要就是通过提取脑脊液(Cerebrospinal fluid，CSF)中与阿尔兹海默症相关的蛋白质，并查看其积累程度，该技术在当前阿尔兹海默症检测中主要查看的是 Aβ 和 Tau 蛋白的积累量[7]。目前最常见的就是利用核磁共振成像和正电子发射断层扫描来检测阿尔兹海默症，因其技术较为成熟、费用相对低廉且检测速度快，被广泛地应用在临床阿尔兹海默症检测中。

1.1.2 研究意义

本文利用阿尔兹海默症的相关语音和图像数据，对其识别问题展开了细致和

深入的研究，取得了一定的研究成果。研究内容不仅仅丰富了阿尔兹海默症识别的理论基础，对现如今的阿尔兹海默症医学实践也有一定的指导意义，从以下两个方面展开简述：

（1）理论方面

语音作为可以体现阿尔兹海默症患者的特征之一，其获取相对简单且可展开的研究方法众多，本文先从语音数据对阿尔兹海默症识别展开了研究，补充了当前基于语音方面研究阿尔兹海默症识别的理论基础，可以为后面的广大科研工作者提供一条研究思路。本文在特征提取及转换上依托数字信号处理的科学方法，通过对原始语音不同特征的提取过程及实验结果对比分析，找到了从本质上去研究阿尔兹海默症识别的办法。例如，可以通过语音特征转化为图像去展开研究，为阿尔兹海默症的识别研究开辟了一条新的理论探索思路。

（2）医学实践方面

从医学实践的视角出发，本文的研究对计算机辅助诊断技术（Computer Aided Diagnosis，CAD）[8]的发展具有重要的价值，为开发临床阿尔兹海默症识别系统起到了推动作用。除此之外，就目前人口老年化带来的公共医疗资源短缺问题，也将会起到一定的缓解作用。最后，本文可以进一步结合现有的阿尔兹海默症识别研究方法，通过广大一线医疗工作者和科研工作者的共同努力，加快智慧医疗体系的建设，增加我国的医疗卫生资源。

1.2 国内外研究现状

近年来，随着互联网公司的大量出现，它们为数以亿计的用户提供在线服务，其中不乏全球知名的互联网公司。基于此背景下，使得大规模数据的获取不再困难[9]。另外，硬件技术的快速发展，如低价且高质量的传感器、便宜的数据存储以及廉价计算的普及，尤其是图形处理器的普及，使大规模算力不再是遥不可及，正是因为数据的井喷涌现与算力的快速发展，近些年极大推动了人工智能(Artificial Intelligence，AI)[10]在众多领域的快速发展。如人工智能在计算机辅助诊断技术领域也取得了突破性的进展，在阿尔兹海默症识别中，现今最前沿的研究主要是依托机器学习(Machine Learning，ML)和深度学习(Deep Learning，DL)来展开[11]。本节介绍了基于这两种技术下当前国内外相关研究现状，主要可以分为两个方面：基于机器学习的阿尔兹海默症识别研究现状，基于深度学习的阿尔兹海默症识别研究现状。

1.2.1 基于传统机器学习的阿尔兹海默症识别研究

传统机器学习方法因其理论完备、普适性强、便于实现等优点，曾经被广泛应用在阿尔兹海默症识别中。一般通过人工或半人工的方式从数据集中提取特征，再将所得到的特征输入机器学习模型中进行训练。如图1-2所示为基于机器学习

方法的阿尔兹海默症识别具体流程，主要包括特征提取和分类两个阶段，针对不同数据或特征有时也包括特征选择和特征融合等操作。

图1-2 传统机器学习方法的阿尔兹海默症识别算法流程

应用机器学习方法的前提是需要手动设计特征，针对不同的任务和数据，需要凭借专家经验或先验知识作为特征设计的指导。一般流程是使用标准的特征提取算法，从不同类型的数据提取到相应的特征后，然后将特征输入到选定的模型中，再得到最终结果。其中，在使用多模态数据作为样本进行研究时，为了获取更多信息，可能还通常包含相应的特征选择或者特征融合的操作。在实际操作的过程中，为了便于运算，通常针对数据还有数据预处理的过程。在这之后，才算完成了所有的前期准备工作，可以将准备好的特征输入到机器学习模型中。

机器学习方法在阿尔兹海默症识别中的特征提取方法针对各类数据也各有不同。例如对于医学影像图，常见的特征提取往往是从区域或者全局体积亦或是形态上入手。前者可以是针对全脑的，也可以是针对部分区域的，比如海马体，后者则是侧重于影像形状。在对医学图像数据进行特征抽取前，研究人员通常会使用一些算法工具对核磁共振成像图进行分割处理，如将完整的图像切分为灰质、白质、脑脊液三部分，然后选择三者中某一部分或者不同组合，以其体积作为分类特征输入模型训练。又或者使用脑分区模板，例如自动解剖标记（Anatomical Automatic Labeling，AAL）模板将大脑分成若干个脑区，然后选择部分蕴含信息丰富区域（Regions of Interest，ROI）的体积作为特征[12]。而基于形态学的特征提取则是以核磁共振成像全脑或部分区域的形状信息作为分类特征，由于阿尔兹海默症患者通常伴随大脑皮质一定程度的形变，例如海马体萎缩或者脑室扩大等症状，所以时常选择 ROI 的皮质厚度值或者海马体的灰质密度值等形态学特征。除了使用简单的形态学和纹理信息特征来研究阿尔兹海默症的识别之外，还可以使用图论的方法对正电子发射断层扫描图或核磁共振成像图进行特征提取，再将特征输入模型中来展开阿尔兹海默症的识别研究。

Sukkar 等人[13]采用隐马尔科夫模型来识别阿尔兹海默症，具体是通过模型的不同状态表示受试者所处患病阶段来展开研究。Christian 等人[14]在 MRI 数据上采用支持向量机 (Support Vector Machine，SVM) [15]模型并多次配置参数，进一步探究了不同轻度认知障碍阶段的差异。Yang 等人[16]同样采用 SVM 模型并结合粒子群优化算法来对阿尔兹海默症进行识别。Querbes 等人[17]使用皮层厚度

作为特征来识别阿尔兹海默症。Plocharski 等人[18]创新性的利用脑沟深度、长度、曲率和表面积等数据作为阿尔兹海默症识别的特征，除此之外 Akhila 等人[19]在研究中使用 MRI 图的纹理信息也作为阿尔兹海默症的识别特征并取得了一定效果。Liu 等人[20]从 PET 图和 MRI 图中共提取 80 多个 ROI，提出了一种基于贝叶斯框架的多重贝叶斯核化（Multifold Bayesian Kernelization，MBK）的方法来研究阿尔兹海默症的识别问题。Zhang 等人[21]用标准范式的模板从 PET 图和 MRI 图中提取了 90 多个 ROI，使用 SVM 对多峰特征进行融合后再来展开阿尔兹海默症的识别研究。Moradi 等人[22]先对特征进行筛选后，再将所获得的特征输入到 SVM 来进行分类，最后通过观测者操作特性曲线（Receiver Operating Characteristic，ROC）对阿尔兹海默症识别结果进行评价。Kruthika 等人[23]使用多级分类器来开展阿尔兹海默症的识别研究，其思想是首先将浅层特征输入到分类器中，实现快速分类并得到初步的结果，再将快速分类的结果进行筛选，对于其中置信度较低的样本进行二次分类，根据不同阶段的分类效果，可以选择不同复杂程度和性能的分类器以实现最佳分类。Altaf 等人[24]利用医学图像和临床特征，通过对多种分类器赋权重的方法来进行阿尔兹海默症识别研究。Yao 等人[25]使用优先级不同的方法来进行阿尔兹海默症的识别，具体实现思路是先让分类器将样本分为两大类，然后对两大类样本再次进行精细分类，最终达到将所有样本均分类成功的目的。

在以上这些基于机器学习的方法中，往往涉及到特征提取、特征选择、以及分类多个模块，这在实际处理中是一个很繁琐的过程。除此之外，特征选择和分类算法通常是两类不同的数学模型，涉及不同的理论框架，若处理时稍有不慎将极有可能导致分类过程中丢失重要信息。在传统方法中常用的分类器大都涉及数据转换，例如 SVM 和高斯过程(Gaussian Process，GP)都需要使用核函数来进行数据转换，而且核函数的选择范围是有限的，这也是在选用过程中值得注意的地方。实际应用中核函数的作用就是将数据从当前空间映射到一个新的空间，进行该操作的前提是在新的空间中样本更容易区分，且在特定分类任务中选择有效的相似性进行度量的核函数并不容易，所以该过程也是较为复杂。

由于客观上存在以上问题并难以解决，传统机器学习方法在阿尔兹海默症的识别研究中虽然取得了一些成果，但是目前此类方法的应用趋于减少。另外从实验结果上，例如从识别的准确率、召回率这一类评价指标上来看，应用机器学习方法进行阿尔兹海默症识别的最终效果并不让人满意，且短时间内识别的性能很难再获得较大提升。

1.2.2 基于深度学习的阿尔兹海默症识别研究

近年来深度学习一直都是热门的话题，其为多种任务提供了一套端到端的方法去解决问题，省去了繁琐地手工设计特征的流程。研究者只需要将注意力放在

深度神经网络模型的构建上，而不必去过多关心中间的实现过程，使用起来也更加便捷，被广泛的应用在众多领域，在众多任务中都能超越传统机器学习的性能，现今在阿尔兹海默症的识别研究上也频频见到其身影。

在图像模态的阿尔兹海默症的识别中，Hosseini-Asl等人[26]率先使用建立在自动编码器(autoencoder，AE)之上的三维卷积神经网络来研究阿尔兹海默症的识别问题，并且在ADNI数据集上验证了其有效性。Payan等人[27]在数据集上进行随机小块的抽取，利用稀疏自编码器（Sparse Auto-Encoder，SAE）预训练卷积神经网络(Convolutional Neural Network，CNN)[28]的卷积层参数展开阿尔兹海默症的识别研究，还使用二维卷积神经网络和三维卷积神经网络对效果分别进行了探索。Farooq等人[29]通过预先对MRI图分割成为灰质图像后再进行二维切片操作，将得到的数据利用预训练模型针对阿尔兹海默症的识别任务进行微调，构建了一个四向分类器来识别阿尔兹海默症。Basaia等人[30]采用旋转、拉伸、剪裁等方式进行数据增强（Data Augmentation）[31]后，再构建三维卷积神经网络来识别阿尔兹海默症。Chan等人[32]将数据处理成三维切块，使用多个卷积神经网络组合来提取特征再进行阿尔兹海默症的识别。

在语音模态的阿尔兹海默症的识别中，Luz等人[33]在2020年国际语音大会的ADReSS挑战赛定义了一项共享任务，通过该任务可以比较基于自发语音的阿尔兹海默症自动识别的不同方法。ADReSS挑战赛为研究人员提供了一个基准语音数据集，该数据集经过声学预处理，并在年龄和性别方面进行了平衡，定义了两个认知评估任务，即：阿尔兹海默症语音分类任务和神经心理评分回归任务。在阿尔兹海默症语音分类任务中，ADReSS挑战赛参与者创建将语音分类为痴呆或健康对照语音的模型，并为该任务提供了基线。Jiahong等人[34]利用端到端方法从语音模态数据来展开研究，具体是把阿尔兹海默症识别作为wav2vec2模型的一个下游任务，通过微调wav2vec2来实现有效分类。龚振等人[35]使用语音的声学特征集，利用时序卷积网络（Temporal convolutional network，TCN）和自注意力机制(Self-Attention，SA)进行分类，对样本不均衡问题给出不同解决方法，同时在TCN中使用新的残差块并且使用输出聚合的方式以提升模型对阿尔兹海默症的识别能力，实现了对阿尔兹海默症优秀的分类。惠寅华等人[36]提出基于交叉投票的特征选择算法来展开研究，该特征选择算法有效地降低了识别任务数据引起的过拟合风险，使得识别算法拥有更强的泛化能力，并且取得了较好的性能（参考文献34、35、36均被收录于2021年全国人机语音通讯学术会议论文集中，三篇论文依次为阿尔兹海默症识别竞赛板块的性能前三甲）。Jiahong等人[37]通过语音结合转录本，使用基于Transformer的预训练语言模型来对下游任务进行微调，并使用理论填充编码来研究阿尔兹海默症的识别问题，从语言和认知的角度得到了关于阿尔兹海默症患者在自发语音中“-um”的使用频率远低于“-uh”的结论，

这为采集语音数据样本提供了良好的侧重点。

虽然深度学习在阿尔兹海默症的识别研究中相较机器学习的方法已经取得了较大的进步，但在阿尔兹海默症识别任务上仍然存在一些棘手的问题，如数据形式特殊、数据样本量有限等。现有的深度学习方法应用在阿尔兹海默症识别任务时，依旧存在一定程度上的不足。比如针对四维的图像数据，尚未有可以端到端适用的深度学习模型。由于实际研究中获取的数据量有限，三维卷积神经网络也会因为参数过多出现过拟合现象。另外如何抑制高维医学影像图所包含的冗余信息，使模型更加关注与阿尔兹海默症有关的区域，以及对于不同语言和口音的语音数据也迫切需要去研究。此外，对于深度学习模型的构造，也还有提升的空间，以上几点均是需要进一步关注的工作。

1.3 本文的研究内容

本文先阐述了阿尔兹海默症的相关病理学背景，而后介绍了目前关于阿尔兹海默症识别的医学研究现状，以及在阿尔兹海默症的诊断和治疗上面临的巨大挑战，最后说明了对家庭和社会带来的影响。接下来从不同方法的视角综述了近年来阿尔兹海默症识别领域相关的研究方法以及重大创新。在此研究基础上，本文利用深度学习中的三种预训练网络主要做了以下工作：

（1）对语音模态下的阿尔兹海默症识别进行了全面的研究，具体体现在利用不同的音频特征和音频融合特征，分别输入预训练模型，在原始语音以及去噪后的情况下均做了大量实验，为后续的研究开辟了道路。

（2）通过对语音实验的观察，结合预训练模型的特点，采用了语音特征转化为图像的方法来进行阿尔兹海默症的识别研究，并在实验性能上取得了较大进步。针对图像采用了自适应剪裁去噪的办法，并且将该方法迁移到 PET2020 数据集中，性能也有明显提升。将三种预训练模型结合了不同的集成学习（Ensemble Learning）[38]方法，更进一步提升了模型对阿尔兹海默症识别的准确率。

1.4 本文的组织结构

第一章，首先介绍了阿尔兹海默症的现状及对社会的影响，引入了阿尔兹海默症识别研究的背景和意义，其次从基于机器学习和深度学习的视角介绍了国内外阿尔兹海默症识别的研究现状，最后详细的说明了本文的研究内容，以及本文所做出的贡献。

第二章，主要介绍了本文所使用到的相关技术。首先从人工智能理论引入，介绍了机器学习中的支持向量机算法和深度学习中的卷积神经网络模型。其次介绍了集成学习和数据增强的相关背景及理论，最后对本文实验的评价指标进行了介绍，以及介绍了用来可视化分类效果的混淆矩阵（Confusion Matrix）[39]。

第三章，介绍了在阿尔兹海默症识别中常见的音频特征和相应的提取方法，

提出了本章实验的模型框架。然后对三种音频特征在不同预训练模型下的阿尔兹海默症识别情况展开了实验，以及特征进行融合后和采用语音端点检测去噪处理后的对比实验。

第四章，提出了语音特征转换成图像特征的想法，结合集成学习来展开对阿尔兹海默症识别的研究。然后对比了图像进行去噪和数据增强后的性能。将该方法迁移到PET2020脑部影像图数据集中，实验结果充分证明了其有效性。最后使用混淆矩阵对不同数据集下的最佳分类情况进行了可视化。

第五章，为加强深度学习研究成果的落地，将本文的研究成果设计成为阿尔兹海默症识别系统，并对系统的设计理念、实现方法、前端主要界面进行了简要的介绍或展示。

第六章，将本文主要工作、创新点、成果进行了总结，并对本文所提出的方法中存在的不足之处进行了科学的分析。最后对未来阿尔兹海默症识别研究工作的方向进行了展望。

2 相关技术

2.1 人工智能相关理论

人工智能的概念最早来源于图灵测试（The Turing test），于上世纪五十年代被英国数学家、计算机科学家图灵在《Computing machines and intelligence》[40]中提出。当时的定义是指测试者与被试者（人和机器）在充分隔开的环境下，使用输入设备向被试者随意提问，通过不同问题进行多次测试后，如果超过 30%的回答不能使测试者分辨出人和机器的回答，那么这台机器就算通过了测试，并被认为具有人工智能。人工智能在被提出后经历了发展期、低迷期、稳步发展期、蓬勃发展期等曲折的发展历程，在各个发展阶段，对其都有不同的定义。现今被学术界广泛认可的定义是人工智能是研究如何利用计算机模拟人类思维来解决问题的一类方法的统称，具体研究领域涉及数学、语言学、统计学、计算机科学、心理学等，研究成果几乎覆盖社会科学和自然科学的所有学科，并被广泛的应用在当今社会的方方面面。

人工智能的目的是试图掌握对象的本质属性，通过科学的方法使机器可以模拟人类的行为，进而研发出至少在某一方面与人类表现相近的算法，这一机器就被认为拥有人工智能。数据、算法、算力是其发展的三要素，一起支撑着人工智能的发展。在特征提取和模型架构上又具体可以将人工智能细分为机器学习和深度学习[41]，机器学习是实现人工智能的一种重要方式，深度学习是在机器学习的基础上发展而来，图 2-1 呈现了人工智能、机器学习、深度学习三者的关系。

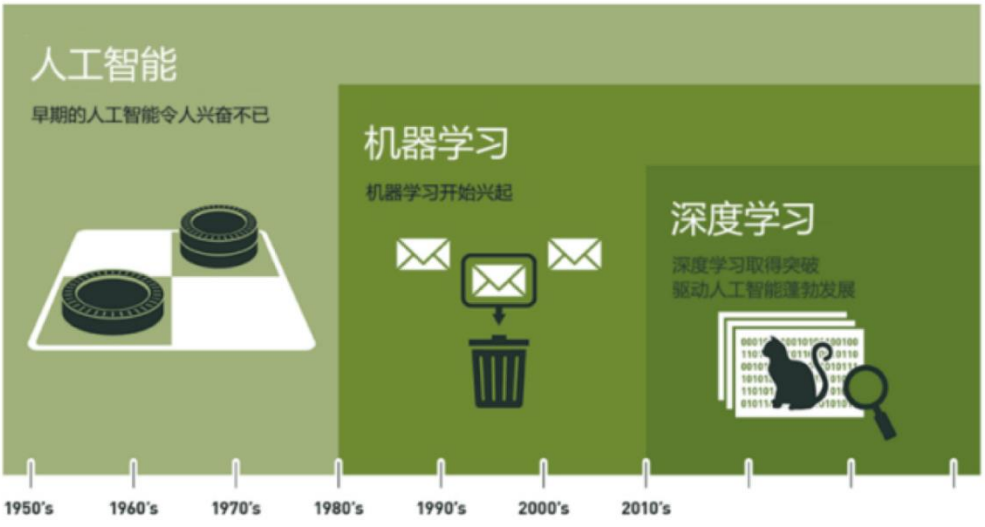

图 2-1 人工智能、机器学习、深度学习三者的关系

机器学习是计算机利用已有的数据，得出某种规律，并利用规律在具有相同属性数据中预测未来的一种范式的统称[42]。通俗来说，机器学习就是一种能够赋予计算机学习的能力，并让该计算机完成通过直接编程上无法完成的功能的方法。按学习方式可以将机器学习划分为有监督学习、无监督学习、半监督学习、强化学习等，其中有无监督就是指数据是否预先设有标签，有标签则是有监督的训练，反之则无。按所实现的任务可以将机器学习划分为回归、分类、聚类等。回归是指对数据进行处理后，通过训练确定所预设函数表达式中自变量和因变量的相关关系。分类是指在已经定义的类别中，按某种属性将数据进行归类，主要是用来实现预测功能。聚类是指在未定义的类别中按特征将数据划分成某几种类别，主要是来实现数据的降维。实现机器学习的代表性算法有线性回归、朴素贝叶斯、支持向量机等，均被广泛应用在各类任务中。

深度学习是从信息处理的角度对人脑神经元网络进行抽象的数学模型，神经元按不同的连接方式组成不同的网络，在工业界与学术界也常直接简称为神经网络或类神经网络[43]。神经网络中的节点（神经元）之间相互连接，每个节点代表一种特定的输出函数，称为激活函数（activation function）[44]。每两个节点间的连接之处具有对通过该处信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依赖于网络的连接方式、权重值和激活函数的区别而有所不同。深度神经网络中最显著的特点是分层学习，其使用大量数据对含多个隐藏层的网络模型进行训练，通过组合浅层网络来形成更加稠密的深层语义抽象，从而学习到数据的本质规律，使模型得到更加精准的预测，分层学习的原理如图2-2所示。

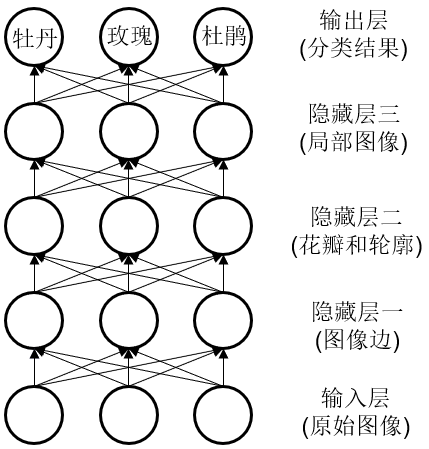

图2-2 深度学习中的分层学习

模型中包含多少层，这称之为模型的“深度”，模型学习了大量层的底层特征，每一层提供各自层次的表示，接近分类输出的层可以用于区分更加抽象的概念。深度学习模型的“深度”够深，相比传统神经网络具备更好的拟合能力。这种分层学习的形式与生物神经学上人脑的学习方式高度相似，是深度学习取得瞩目

成就的重要原因之一。在面临大量的感知信息时，深度学习使用多层非线性模型解决了传统机器学习中需要进行特征工程的问题，为解决不同任务提供了一套统一的工具。深度学习近年被广泛应用于语音识别[45]、计算机视觉[46]、自然语言处理[47]等领域，并取得了优异的成绩，在越来越多的复杂任务上完成度接近甚至超越了人类的水平，且未来将还会有进一步的发展。

2.2 支持向量机

支持向量机算法来源于20世纪70年代迅猛发展的统计学习理论，经过不断地健全和发展后在20世纪90年代被正式作为一种机器学习方法提出，因其理论完备且使用方便，迅速成为机器学习中的热点算法并被应用在各种任务上。支持向量机本质上是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面，因此可以将任务简化为一个求解凸二次规划的问题。通俗来说支持向量机的作用机理就在特征空间中找到一个超级平面尽可能地将两类数据分开，训练过程就是在原始空间寻找两类样本最优分类超平面的过程，图2-3能很直观的看到支持向量机的分类思想，即将最小距离最大化，其中 $\omega^T+b=0$ 表示初始超平面。因支持向量机具有理论完备、健壮性强、性能较好、适合解决小样本问题等优势，目前在计算机视觉、自然语言处理、语义分析等领域仍时常见到其身影。

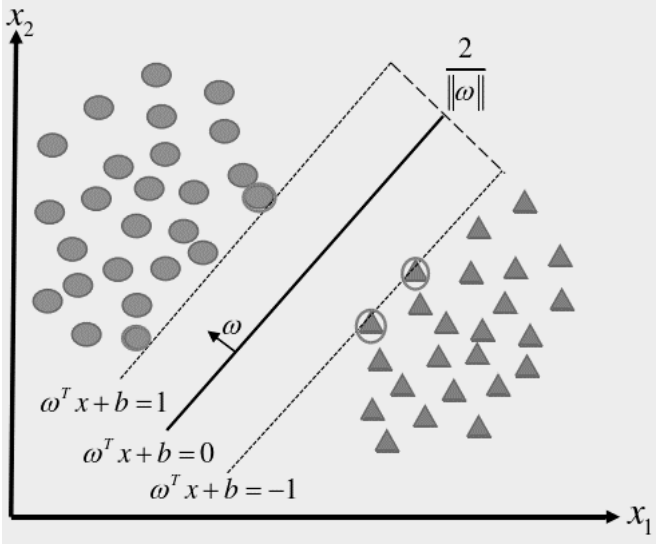

图 2-3 支持向量机二维空间分类示意

2.3 卷积神经网络

卷积神经网络是一类强大的、为处理图像数据而设计的神经网络，其具有数据维度低、权重共享和特征提取能力强的特点，雏形是在解决数字识别任务时提出的LeNet-5[48]。现代卷积神经网络的设计得益于生物学、群论和一系列的补充实验，更加逼近真实的生物神经网络。卷积神经网络的计算复杂度远小于全连接架构的网络，其卷积的并行运算在GPU上十分便捷，因此基于卷积神经网络架构

的模型在计算机视觉领域研究中已经占主导地位，在诸多其他领域也有着优异的表现，成为被广泛应用于解决各类涉及深度学习任务的算法之一。

卷积神经网络的工作方式不同于全连接神经网络，其通过每个隐藏节点映射到上层的数据局部窗口做内积运算，此过程极大地减少了参数量，加快了模型的收敛。卷积神经网络这种首先关注局部数据特性的设计类似于人类视觉，例如人类在观察一张关于猫的图片时，通常首先会关注重要的局部特征如颜色、动作、面部、尾巴等。卷积神经网络主要由卷积层、池化层、激活函数层和全连接层等模块组成，其网络结构模型见图2-4。

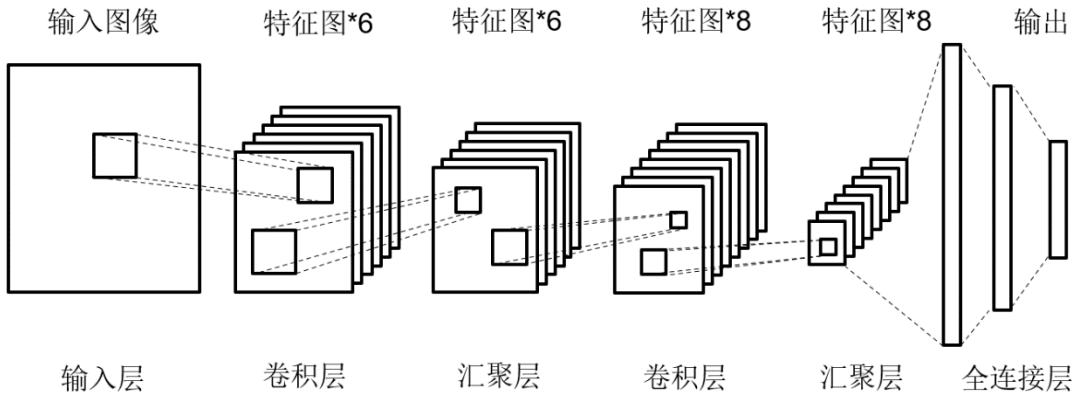

图 2-4 卷积神经网络结构

2.3.1 卷积层

卷积神经网络可以通过端到端的方式来进行训练，网络在训练过程中能自主学习到数据的相关特征，其所依赖的核心是卷积层，卷积层肩负着卷积神经网络中深度特征抽取的重任，还能将非线性可分数据变得线性可分。通常来说，卷积层由一组卷积核（滤波器）组成，卷积核在输入数据上进行有规律的滑动，在此过程中与数据局部窗口做卷积，所谓卷积就是卷积核 w 与数据窗口数据进行内积运算。假设输入数据 X 的规模为 $x*y*z$（高*宽*通道数），在模型构造中卷积核的深度必须与数据的通道数 z 一致。假设输入数据的通道数为3，那么与 X 做卷积操作的卷积核深度必须为3，卷积核大小可根据理论或工程经验进行缩放。例如数据 X 的规模5*5*3经过二维卷积核 w 为3*3*3的卷积层进行卷积运算，且步长和偏置均为1，卷积过程见公式（2-1）：

$$f(x,y)=u(x,y)\times v(x,y)+b \tag{2-1}$$

其中 $u(x,y)$ 表示 X 的窗口数据，$v(x,y)$ 为卷积核的大小，b 为偏置，$f(x,y)$ 表示卷积运算后的窗口值。卷积过程如图2-5所示，从图中可以清晰的看出，卷积过程完成了数据的特征抽取，且生成特征图的大小与原数据相比有所减小。

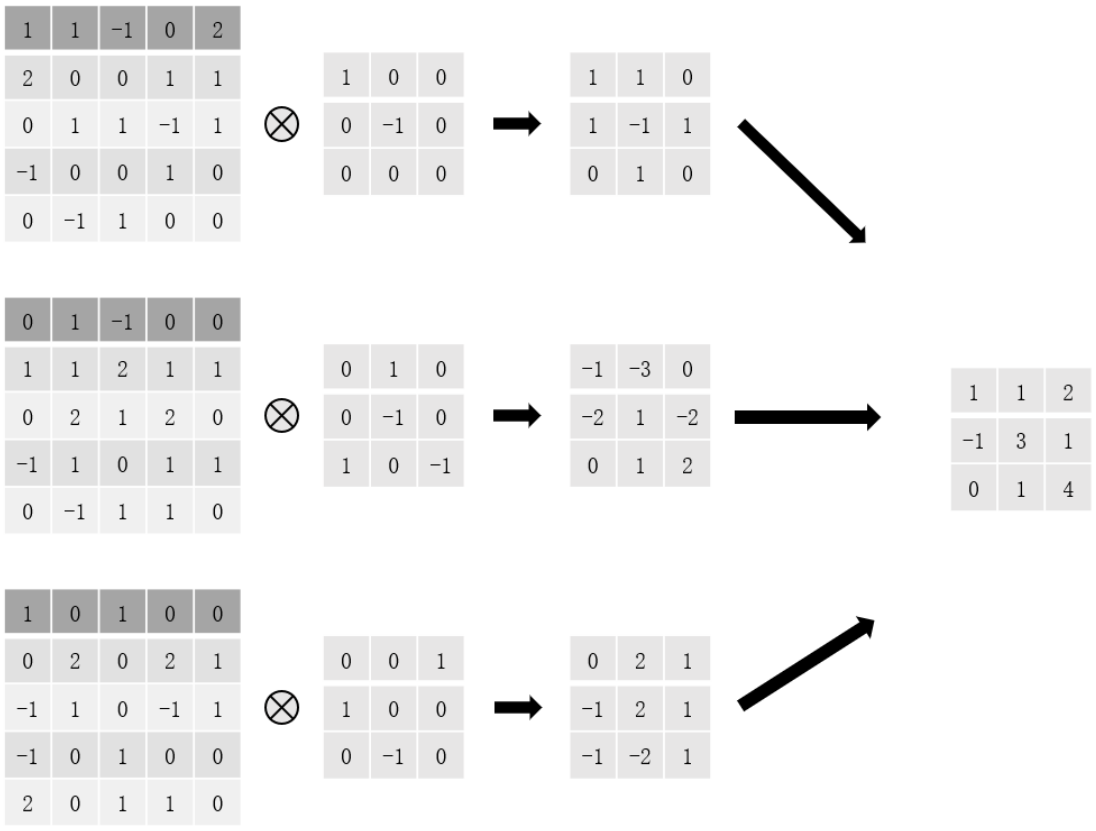

图 2-5 二维卷积计算过程

2.3.2 池化层

池化层[49]是对上层输出特征图进一步处理，池化过程是将特征图上某一位置相邻区域的总体统计特征作为该位置的输出，通俗来讲就是在该区域上指定一个值来代表这个区域，实现提取出该区域的主要特征。池化层与卷积层类似，其运算过程也是通过滑动窗口机制进行，该窗口根据步幅大小在输入数据的区域上滑动，依次遍历每个位置计算后得到输出，这一过程既减小了网络的计算量，又能保持特征的旋转、伸缩、平移的不变性。池化方法主要有平均池化（Average Pooling）、最大池化（Max Pooling）等，平均池化是提取滑动窗口所覆盖特征图的均值，最大池化是提取滑动窗口中所覆盖特征图的最大值。例如上层输出的特征图Y大小为3*3，假设池化核大小为2*2，滑动窗口的步长为1，其池化过程如图2-6所示。

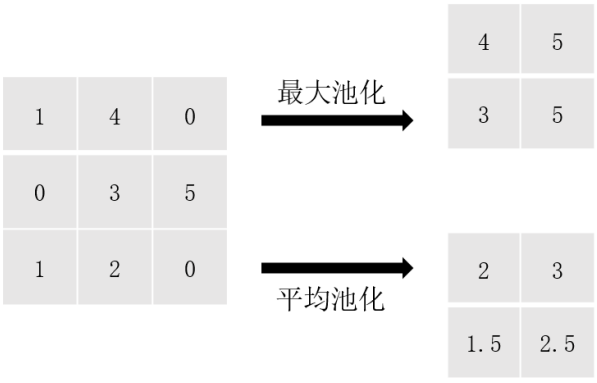

图 2-6 两种池化计算过程

2.3.3 激活函数

全连接层、卷积层、池化层的一个共同特征就是只能对数据进行线性操作，在面对非线性可分的数据时就极大减弱了其网络的特征抽取能力，此时无论如何继续加深网络，其结果都是线性叠加，很难再进一步提升网络的表征能力，让神经网络的“深度”失去意义。激活函数的引入则较好地解决了这个问题，使网络增加了非线性特性，进而提高了模型的学习能力。常见的激活函数包括：Sigmoid函数[50]、Tanh函数[51]、ReLU函数[52]。

Sigmoid函数的表达式（2-2）如下：

$$f(x)=\frac{1}{1+e^{-x}} \tag{2-2}$$

Sigmoid函数是神经网络早期研究中被应用的激活函数之一，其函数图像如图2-7所示(黑色曲线为其函数图像，蓝色曲线为其导函数图像)，酷似字母S的形状，因此也被称为s形函数。从函数图像可以看出，Sigmoid函数可以将神经元的输入变换成值域区间为（0,1）的输出，且只在逼近0的区间内其值变化很大，其余区间变化极其缓慢，不难发现Sigmoid函数具有很大的饱和区，这种特性会在误差反向传播(Backpropagation，BP)中极易发生梯度消失，使模型难以训练。

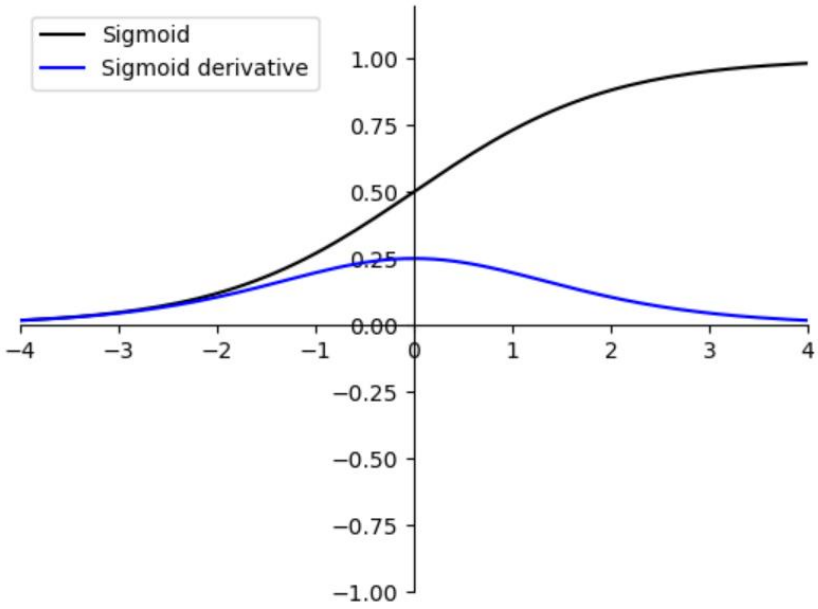

图2-7 Sigmoid函数图像

Tanh函数的表达式（2-3）如下：

$$f(x)=\frac{e^{x}+e^{-x}}{e^{x}+e^{-x}} \tag{2-3}$$

Tanh函数也被称之为双曲正切函数，其可以将神经元的输入值变换成值域区间为（-1,1）的输出，其函数图像如图2-8所示(黑色曲线为其函数图像，蓝色曲线为其导函数图像)，从图像中可以发现其形状类似于Sigmoid函数，因此也同样存在较大饱和区，依旧无法避免训练过程中的梯度消失现象。

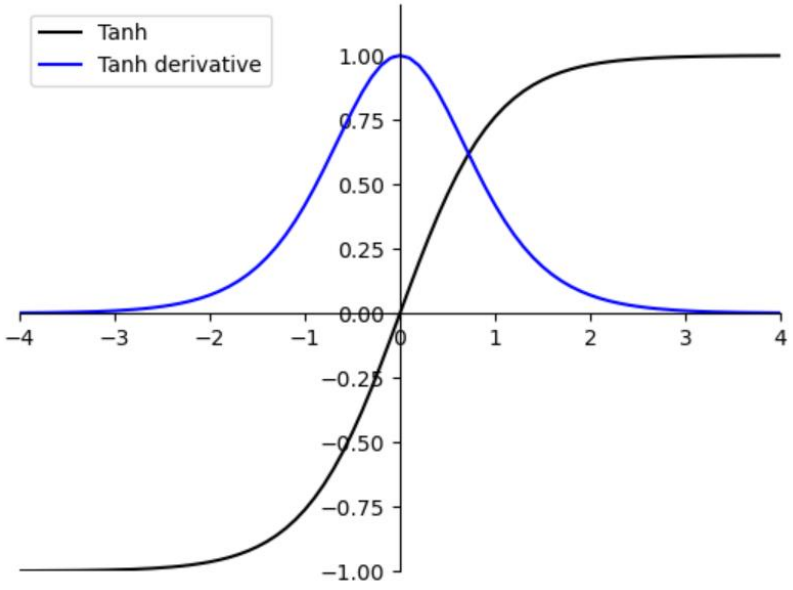

图2-8 Tanh函数图像

ReLU函数的表达式（2-4）如下：

$$f(x)=\max(0,x) \tag{2-4}$$

ReLU函数可以让特征值中的负值置零和正值保持不变，且正值区间的斜率也不变，避免了神经网络在训练中梯度消失的问题（正值区间），而且分段线性运算具有较快的运算速度，训练过程中能加快模型的收敛，现今被广泛应用于深度神经网络中，其函数图像如图2-9所示(黑色折线为其函数图像，蓝色折线为其导函数图像)。

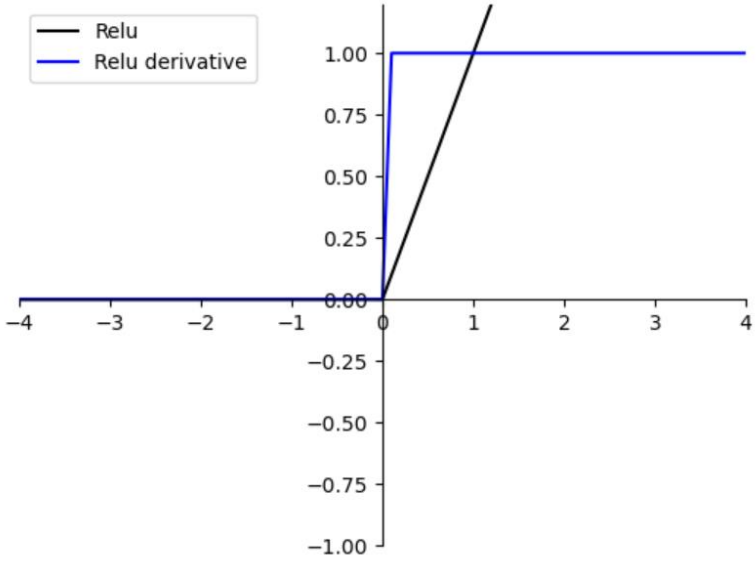

图 2-9 ReLU 函数图像

2.3.4 过拟合及其解决办法

在模型的训练中，时常会出现在训练集中表现很好，而实际测试时性能却较差的现象，这就是发生了过拟合。过拟合贯穿深度学习的整个研究阶段，在深度学习初期这几乎是大家都有碰到过的情况。如何去避免这一现象，使得模型在测试时性能发挥稳定，这是值得去了解和深入研究的问题。具有一定工程经验后不难发现过拟合现象常发生在数据噪声过多、训练数据不足和模型复杂度过高的几种情况下，对于上述情况，有以下几种办法来缓解过拟合。

针对数据噪声过多这种情况，通常是采用数据清洗的办法来解决，值得注意的是在面对不同任务以及不同模态的数据时，都需要选择不同的数据清洗方法，这就要求操作者了解任务对应数据在深度学习模型进行特征抽取的本质。针对训练数据量不够的问题，我们可以增加样本来辅助训练，但是在实际研究中，数据往往是极其有限的，很难获取更多数据，通常是采取数据增强的办法来尽可能缓解数据量不足的情况。除此之外，可以通过降低模型复杂度、正则化[53]、Drop out、早停法、集成学习等方法来缓解过拟合。降低模型的复杂度本质上就是减少模型的参数，让模型在训练过程中不涉及太多参数的更新，从而保证在已有的数据量下能满足当前参数的更新。正则化通俗来讲就是在损失函数中加入一个正则化项，也称为惩罚项，通过该项来调整模型的参数，使得损失最小的情况即为最佳参数，

依据惩罚项的不同又可将正则化分为 L1 正则化和 L2 正则化。

L1 正则化结合损失函数公式（2-5）如下：

$$L(x,y)=\sum_{i=1}^{n}(y_i-h_\theta(x_i))^2+\lambda\sum_{i=1}^{n}|\theta_i| \tag{2-5}$$

从上式中可以看出 L1 正则化的目的是使权重的绝对值最小，其在训练中受极端数据影响比较小，适合于可解释的模型。

L2 正则化结合损失函数公式（2-6）如下：

$$L(x,y)=\sum_{i=1}^{n}(y_i-h_\theta(x_i))^2+\lambda\sum_{i=1}^{n}\theta_i^2 \tag{2-6}$$

从上式中可以看出 L2 正则化的目的是使权重的平方最小，其在训练中受极端数据影响比较大，适合于特征众多的数据和复杂的模型。

Drop out 通俗来讲就是在训练的过程中，以预设的某一概率来随机舍弃部分神经元，这样虽然会造成一定程度上的信息丢失，但同时也让模型不会过度依赖某些局部特征，从而增强了模型的泛化性能。早停法就是在训练过程中进行验证时若发现性能下降就及时停止对模型的训练，让最后一趟训练中的权重作为模型最终参数，由于在训练的实际验证中（如五折交叉验证）模型性能时常发生震荡，使得模型的停止训练时间非常不好把握，故该方法在操作中想要得到较好参数比较困难。

2.4 集成学习

在机器学习算法中，通常希望训练得到一个鲁棒性较高且泛化能力强的模型，但随着数据规模的指数增长和寻优计算本就难度较大，使得实际训练过程中往往得到的是在某方面具有偏好的模型。由于单一模型时常力不从心，导致最终测试结果不尽人意，在此背景下集成学习应运而生。集成学习是一种标准范式，该思想在 1979 年被首次提出，在此之后便受到众多科研工作者的关注，成为了机器学习的研究热点并被广泛地应用在人工智能及其相关研究领域。集成学习的思想意图通过多种模型结合来提升性能，旨在获得比单一模型更好的泛化性能。近年来，各种基于大数据驱动的深度学习比赛如天池、Kaggle、AI 开发者大赛等频频出现集成学习的身影，集成学习思想的应用也帮助模型取得了耀眼的成绩。

集成学习本质上是一个机器学习的过程，首先通过训练得到若干存在差异的基学习器，再应用不同集成学习策略来综合多个基学习器得到最终结果，其思想见图 2-10。其中针对相同任务，基于同一种算法训练得到的基学习器进行集成的方法称为同质集成，基于不同算法训练得到的基学习器进行集成的方法称为异质集成。集成学习模型在一定程度上弱化了单一模型对结果的影响，具体表现在准确率、稳定性和学习效率等方面，根据不同的学习策略，目前集成学习策略可以分为平均值法、加权平均法、投票法、加权投票法等。

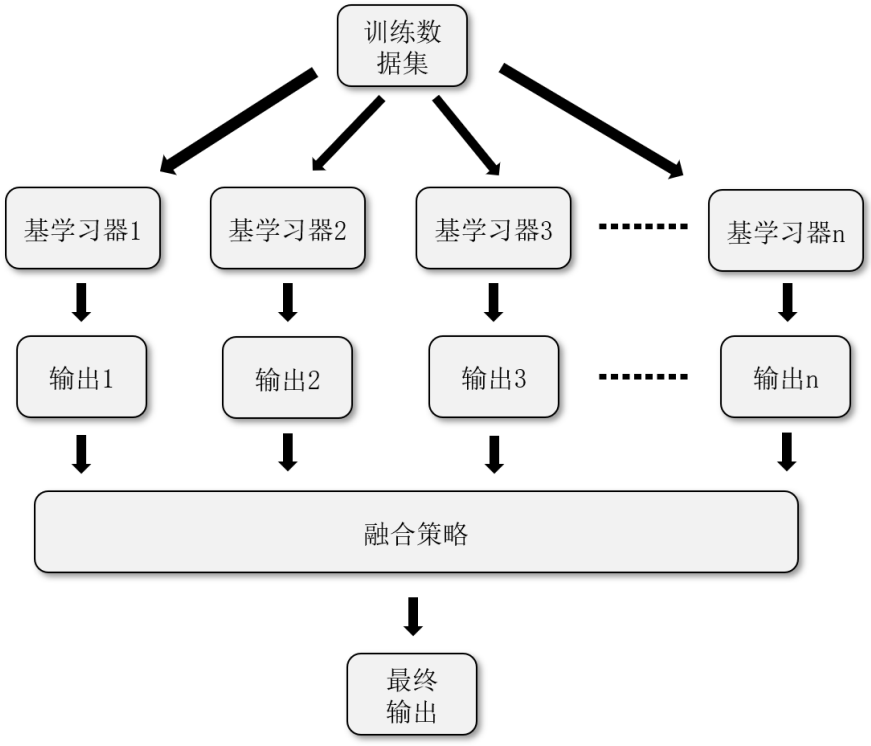

图 2-10 集成学习的基本流程

随着集成学习的不断发展和完善，不少新的集成学习算法争相涌现，但究其根本，大多数集成学习算法均是由经典集成学习算法演变而来，其中四种主要的经典集成学习算法是 voting、Bagging、Boosting、Stacking 等，经过对不同算法的初步应用即可得到最终结果，本文使用的是投票集成学习的方法。

2.5 数据增强

深度学习本质上是以数据和算力来驱动的，数据的规模对模型的训练效果至关重要，理论上若拥有足量的数据，模型可以训练到正确应对各种情况。在深度神经网络模型的飞速发展中，现今模型参数量已经高达几千万级甚至上亿，想要得到合适的参数需要大规模的数据进行训练。一般而言，数据量越充足训练出来的模型泛化性越强，鲁棒性也越高。但目前在很多实际的研究中，数据往往极其稀有难以获取，很难找到足够的数据来展开研究，数据增强为我们在一定程度上解决了数据量不足的困扰。

一般而言，使用某种范式使数据量变多的方法都可以广义的称为数据增强。在机器学习领域，数据增强具有严格的定义，特指对当前的训练数据进行某种处理后得到更多训练数据的过程。在深度学习领域则没有这么严格，把增多数据量或者提高数据质量的方法都可以统称为数据增强，通俗来讲就是非实质性从外部增加数据的情况下，让有限的数据产生等价于更多数据价值的方法称为数据增强，也称为数据集增强。从进行数据增强的阶段可以将其分为两类，一类是离线增强，另一类是在线增强。离线增强就是直接对数据集进行处理，使数据的数量增加；

在线增强就是对输入模型训练的该批次数据进行算法处理来达到增强效果。根据数据类型的不同，数据增强的方法也有所区别。对于文本类型数据，通常主要方法是利用数据分析挖掘技术对文本数据中的同义词进行替换、词序进行调整以及引入噪声词语等方式。对于图像类型数据，可以对图像采用几何变换、颜色变换和基于生成对抗网络（Generative Adversarial Network, GAN）[54]的方法来进行数据增强，图2-11就是考拉图经过常见图像数据增强方法处理后的图像集合。

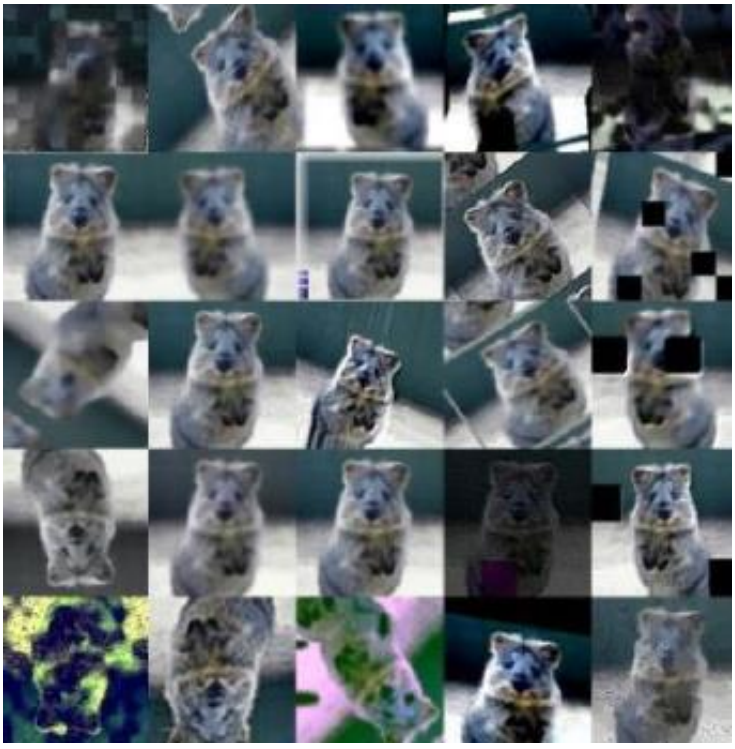

图 2-11 考拉图经过常见图像数据增强方法处理后的图像集合

2.6 评价指标及混淆矩阵

在分类任务中，评价算法性能的标准众多，阿尔兹海默症的识别研究中常见的评价指标有准确率（Accuracy，Acc）、召回率（Recall，R）、精确度（Precision，P）等。准确率即测试数据中分类正确样本与样本总数之比；召回率即患者中诊断正确的概率，也称为敏感度（Sensitivity，SEN）；精确度即衡量预测为患病群体中真正患者的概率。考虑到便于与同样使用本文所选用数据集的研究方法进行对照比较，综合前人的众多研究工作，本文选取准确率作为评价指标来衡量模型的性能，准确率采用公式（2-7）来进行度量：

$$Accuracy = \frac{TP}{TP + FP} \tag{2-7}$$

其中，TP 为正确预测某类别的样本数，FP 为错误预测某类别的样本数。

混淆矩阵是一种分类精度可视化的工具，其又被称为误差矩阵（Error Matrix），主要用来呈现样本数据的真实类别和预测结果之间的关系，是可视化分类模型性

能的方法之一，与具体数值评价指标相比，混淆矩阵能更为直观的呈现分类效果，凸显最佳模型性能，现今被广泛应用于分类任务的结果可视化中。

为了便于理解混淆矩阵的含义，结合实例来进行解读。假设共有90个样本数据被模型预测为A类、B类、C类各30个，分类结束后统计出来得到的混淆矩阵如图2-12所示：

实际				
	C类	1	1	24
	B类	4	29	4
	A类	25	0	2
		A类	B类	C类
		预测		

图2-12 混淆矩阵示例

图中每一列之和表示所属类的预测样本数，每一行表示所属类实际样本数，以第二列为例，A 类样本有 25 个预测正确，4 个错误的预测为 B 类，1 个错误预测为 C 类，其他列的预测情况同理。

2.7 本章小结

本章对人工智能理论进行了简要的介绍，而后介绍了实现人工智能的两种技术：即机器学习和深度学习。接下来介绍了机器学习中的经典算法支持向量机，以及深度学习中的卷积神经网络，并对卷积神经网络的组成模块进行了详细的介绍。为了避免单一模型的偏好性，介绍了利用多模型来提升泛化性能的集成学习理论。为了解决实验中数据样本不够的问题，简要介绍了数据增强的理论及主要方法，最后介绍了本文实验所选择的评价指标和分类结果的可视化工具——混淆矩阵。

3 基于语音模态的阿尔兹海默症识别模型

3.1 引言

现今国内外权威的医学机构和研究单位采用多种方法来研究阿尔兹海默症的识别问题，但万变不离其宗的识别研究标准主要有两个：一是有未出现思维障碍和记忆衰退，二是通过医学影像图对大脑皮层内侧颞页区和中海马区的溃缩程度以及大脑皮层形状进行判断。

语言作为我们日常沟通交流中必不可少的工具，是人脑中所形成思想的外在逻辑表达，自古以来就是人与人沟通交流的重要方式之一，与其他交流方式相比，语言是较为清晰、便捷、也是容易被理解的表达方式。语音是语言的声学表达，是一种最自然的信息载体，包含了丰富的声学信息，其获取也相对便捷且成本低廉。通过对前人的阿尔兹海默症识别研究工作可知，目前深度学习方法被广泛的应用在阿尔兹海默症识别的研究中，并取得了较大的研究进展。本章针对语音模态的数据，利用深度学习的方法展开了阿尔兹海默症识别的研究。

3.2 相关工作

介绍了语音去噪声的相关方法、语音中常见的几类特征以及本章所用到的三种特征及其提取流程。

3.2.1 语音端点检测与消音

通过对数据集中多条语音的试听发现有一个共同特性，即几乎每一条语音中都含有至少一处沉默片段，这类片段没有说话人的语音信息，但偶尔伴随着设备或外界产生的不可抗力噪声干扰，这种现象可能会降低语音特征的质量，在模型训练中形成噪声，从而影响模型的最终性能。

找到语音的有效起止位置并去除该区间内语音，得到不含有沉默片段的语音可以避免上述情况。端点检测技术(Voice Activity Detection，VAD)[55]可以准确定位有效声音信号和噪声信号的区间，其实现方法包含频域方法和时域方法，其中，频域方法包含熵谱法、倒谱法。时域方法包含基于短时能量和短时平均过零率的双门限算法，本质上都是根据语音和噪声相同参数所表现出的不同特征进行区分，确定真正的有效语音区间，然后再配合语音消音法将噪声信号内的语音帧去除，得到的是只含有说话人发声片段的语音。经这样处理后不仅减少了语音特征中的噪声干扰，还减少了无效数据量，使得模型训练速度在一定程度上有所提升，加快了模型的训练过程。

本文中所使用的语音端点检测步骤及参数设置如下：

(1) 在采样率为16000的条件下输入原始语音，将语音信号进行分帧、加窗处理并进行归一化便于计算；

(2) 为减少背景噪声的影响，增强算法的鲁棒性，在语音信号中引入高斯白噪声，也有助于消除全零值；

(3) 计算每帧语音的能量，若某一帧前后 5 帧都不存在受试者语音信号，则该帧被认为是无受试者语音帧，将该帧删去，反之则当作有语音帧进行保留。

3.2.2 语音特征提取

从语音中提取的特征一般可分为以下四种类别，音质特征[56]、韵律学特征[57]、频谱相关特征[58]和深度学习相关特征。音质特征是用来评价语音的质感和清晰度，主要呈现的是语音的质量，常见于声乐器和音响设备的研究中。韵律学特征是情绪的内在体现，富含多种信息，其中常见的有音调、响度和语速等，现今常被用于情感识别的研究中。频谱相关特征是指通过傅里叶变换（Fourier Transform，FT）[59]将时域信号转换为频域信号的过程中得到的特征，常见的频谱相关特征有：基于线性预测的倒谱系数(Linear Predictive Cepstral Coefficient，LPCC) [60]、感知线性预测倒谱系数(Perceptual Linear Predictive，PLP)[61] 、梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients，MFCC)[62]等。语音的深度学习特征主要是指一些已经被定义的特征集，如常见的语音特征集有GeMAPS、eGeMAPS[63]、IS10_paraling（Interspeech比赛中曾经使用的特征集，包含1582维相关语音特征）和ComParE[64]等。

众所周知语音特征的选择对深度学习任务中模型的效果至关重要，在本章的研究中，通过查阅相关文献和初步实验后最终确定利用频谱相关特征来进行阿尔兹海默症识别的研究，具体选取频谱相关特征中的声谱图(Spectrogram，Spec)[65]、梅尔声谱图(Melspectrogram，Melspec)[66]和梅尔频率倒谱系数来展开研究。三种特征具有很强的相关性，声谱图是由原始语音经过预加重、分帧、加窗、傅里叶变换后得到；梅尔声谱图是在声谱图的基础上经过梅尔频率滤波器，将频率转化为梅尔刻度所得；梅尔频率倒谱系数是在梅尔声谱图的基础上经过对数转换和反离散余弦变换(Discrete Cosine Transform，DCT)[67]后得到，三种特征提取流程见图3-1，下面对本章所选用的三种语音特征分别进行简要介绍。

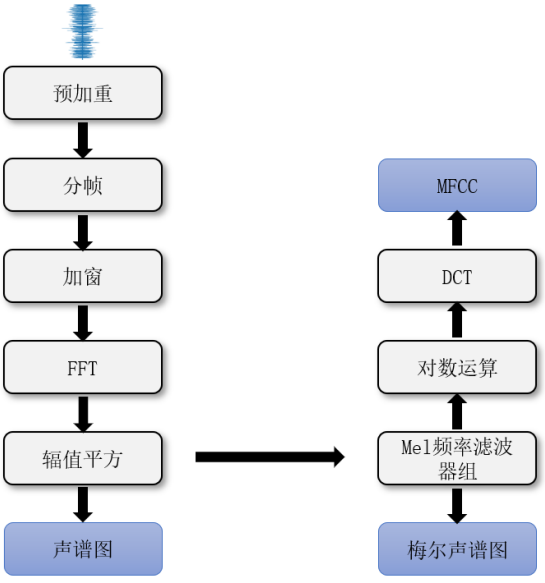

图3-1 声谱图、Mel声谱图和MFCC特征提取流程

声谱图由原始语音经过预加重、分帧、加窗、短时傅里叶变换(Short-Time Fourier Transform，STFT)[68]后等操作得到。在实际应用中，语音信号经过分帧和加窗处理，分割成一帧一帧的离散序列，该操作可视为采用短时傅里叶变换的结果，其计算方法见公式（3-1）。其中K是离散傅里叶变换后的频率点个数，k是频率索引，$0 \le k < K$。$X\ [k, l]$建立起索引为lL的时域信号和索引为k的频域信号的关系，对于采样率F_S，相应的索引对应为时间lL/F_S和频率kF_S/K。

$$X[k,l]=\sum_{n=0}^{N-1}x_l[n]e^{-\frac{j2\pi nk}{K}}=\sum_{n=0}^{N-1}w[n]x[n+lL]e^{-\frac{j2\pi nk}{K}} \quad (3\text{-}1)$$

声谱图经过一系列梅尔滤波器组后得到梅尔声谱图，梅尔声谱图也被称作为FBank。由于相邻滤波器存在重叠，因此梅尔声谱图的特征相关性较高，其计算方法见公式（3-2），在所提取到声谱图的基础上，求其平方得到能量谱，将每个滤波频带内的能量进行叠加，第k个滤波器输出功率谱为$X[k]$；将每个滤波器的输出取对数，得到相应频带的对数功率谱。

$$Y_{FBank}[k]=\log X[k] \quad (3\text{-}2)$$

梅尔频率倒谱系数在梅尔声谱图的基础上经过反离散余弦变换后即可得到，该特征是一种在说话人识别和自动语音识别中广泛被运用的特征，其计算方法见公式（3-3），其中M是三角滤波器个数，一共产生L个MFCC系数。

$$C_n=\sum_{k=1}^{M}\log X[k]\cos\left(\frac{\pi(k-0.5)n}{M}\right),n=1,2,3,\ldots,L \quad (3\text{-}3)$$

本章中所使用的语音特征提取工具是 librosa 工具库[69]，librosa 是一个强大的 python 语音信号处理的第三方库，因其操作文档清晰、功能齐全、安装方便等优点，被广泛应用在各类涉及数字信号处理的研究中。在使用 librosa 工具包

提取过程中，为确保提取到的特征较为优良，抗噪声水平高，涉及到多种参数的设置和调优，本章中所提取特征的参数设置见表 3-1：

表 3-1 语音特征提取中的相关参数设置

相关参数	本实验所设值
n_fft	1024
hop_length	512
segment_duration	3
segment_overlap	0.5
num_segments	None
sample_rate	22050
duration	6
sr	22050
n_mels	128

3.3 模型框架

本文的主要研究内容是使用深度学习的方法来探究阿尔兹海默症识别问题，基于此背景下，本章针对语音模态数据下的阿尔兹海默症识别展开了研究，提出了我们的阿尔兹海默症识别模型如图3-2。该模型由两部分组成，其中从左至右分别是特征提取模块和分类模块，特征提取模块是通过不同的语音特征分别输入模型进行训练或不同语音特征融合后再输入模型进行训练，分类模型就是将训练好的模型去做测试，经过softmax层后得到最大概率类别即为预测标签。

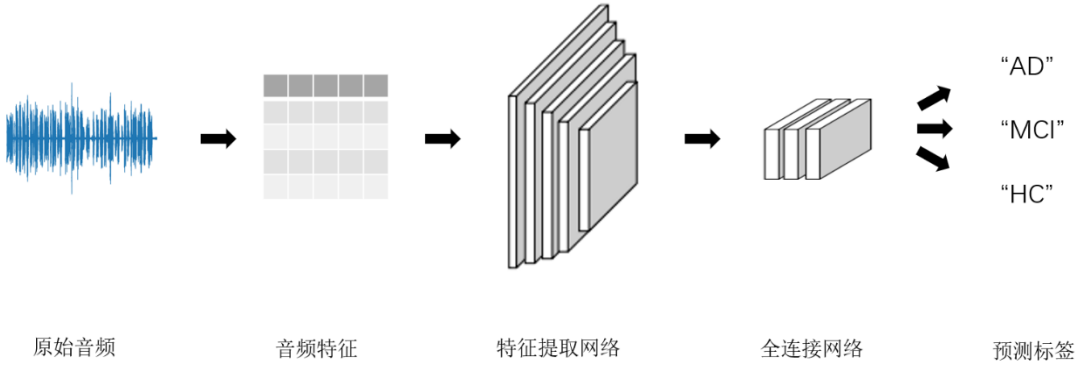

图 3-2 基于语音的阿尔兹海默症识别模型

3.3.1 基于预训练的特征提取网络

针对特定任务的深度神经网络模型的设计是一项极具挑战性和开创性的工作，一个好的模型不仅仅需要扎实的综合理论基础知识，还需要丰富的工程实践经验，花费很大气力去设计一个模型性能却鲜有提升的情况屡见不鲜。预训练模

型（Pre-Training Models，PTMs）[70]的横空出世很大程度解决了这种困扰，为众多深度学习科研工作者开辟了新的道路。从字面意思来看，预训练模型是先通过大量相关数据进行训练，得到训练好的基础模型后，再让基础模型针对具体下游任务的数据集继续训练，其中初次训练称为预训练（Pre-Training）阶段，第二次训练称为微调（Fune-Tuning）阶段。计算机视觉和自然语音处理领域的研究者对预训练模型定然不陌生，其作为深度学习模型的一种范式已经出现很多年，且不断的在实际应用场景中改良，当下已经被使用在众多深度学习领域及其相关交叉领域的研究中。

预训练思想是迁移学习（Transfer Learning）[71]的延伸，迁移学习就是依据先前学习到的知识来解决新的问题，常常可以达到类似的效果甚至略高于先前，且还减少了模型训练的时间，图3-3是迁移学习思想引申而来的预训练模型按标签的分类示意。预训练模型避免了深度学习中模型开发困难的问题，又充分利用迁移学习和自监督学习的方法一定程度解决了数据量不足的问题，提升了模型对集外数据的泛化能力。

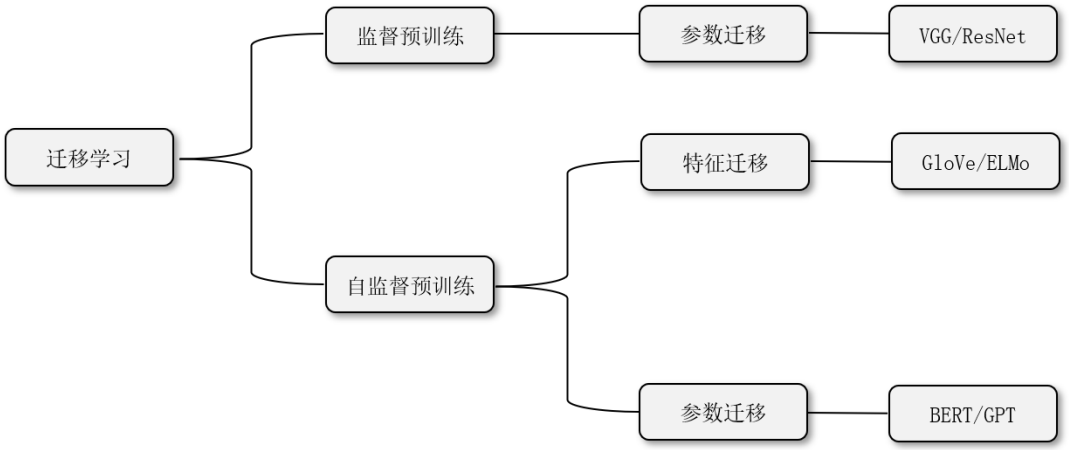

图3-3 预训练模型分类

经过搭建深度神经网络以及对前人在相关任务模型的研究对比，初步实验结果并不理想，均逊于预训练模型的性能，因此本文最终选择主流的预训练模型VGG19[72]、ResNet50[73]和EfficientNet-b8[74]作为阿尔兹海默症识别模型中的特征提取网络，下面分别对三种预训练模型进行简要介绍。

VGG极大的加深了卷积的层数，因而能提取到更多底层的特征，是模型性能增强的主要因素之一，其工作证明了增加网络的深度能够在一定程度上影响网络最终的性能。VGG有两种结构，分别是VGG16和VGG19，两者并没有本质上的区别，只是网络深度有所不同。从深度神经网络的发展来看，网络的深度对模型的性能至关重要，但不断地增加网络的深度，模型的性能并不会一直提升，甚至出现了下降，这就是深度网络的退化问题（Degradation problem）。

ResNet 的作者提出残差学习的思想来解决深度网络面临的退化问题，对于一个堆积层结构（几层堆积而成）当输入为 x 时其学习到的特征记为 $H(x)$，现

在希望其可以学习到残差 $F(x)=H(x)-x$,这样其实原始的学习特征是 $F\left(x\right)+x$，之所以这样是因为残差学习相比原始特征直接学习更容易。当残差为 0 时，此时堆积层仅仅做了恒等映射，理论上至少网络性能不会下降。实际中残差往往不会为零，这也会使得堆积层在输入特征基础上学习到新的特征，从而拥有更好的性能。残差学习思想的结构见下图 3-4。

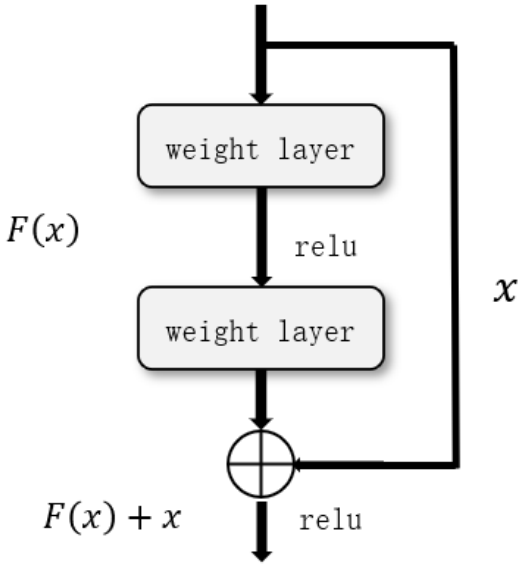

图3-4 残差学习的结构

增加网络的深度能够得到更加丰富、复杂的特征并且能够很好的应用到其它任务中，但网络的深度过深会面临梯度消失（Vanishing Gradient）[75]、训练困难的问题。增加网络的宽度能够获得更高细粒度的特征并且也更容易训练，但对于宽度很大而深度较浅的网络往往很难学习到更深层次的特征。增加输入网络的图像分辨率能够潜在获得更高细粒度的特征模板，但对于非常高的输入分辨率，准确率的增长也会放缓甚至倒退，并且大分辨率图像会增加计算量。EfficientNet 提出了NAS（Neural Architecture Search）的思想，该技术可以很好的来搜索网络的图像输入分辨率、网络的深度以及通道的宽度三个参数的合理化配置，以达到模型在训练中的参数自适应合理配置。

3.3.2 特征融合模块

特征融合是将不同分支或层次的特征进行组合，是提高深度神经网络表征能力的一种常用手段。现有的特征融合方法通常是对特征进行拼接或者求和操作，两种方法都是特征映射的固定线性聚合，实验前完全不知道这种组合是否适合特定的对象，所以可能并不是最佳的选择。

本章不仅将三种语音特征分别输入不同模型来进行研究，还进一步探究多特征融合对阿尔兹海默症识别模型的性能影响。从理论上来讲，多特征融合综合了

更多的有效信息，能实现多特征的优势互补并削弱劣势，可以让模型具有更好的鲁棒性和泛化能力。

不同于简单的拼接或者求和的方法，本章采用了多尺度通道注意力模块(Multi-Scale Channel Attention Module，MS-CAM)[76]来实现特征融合，其更好地融合了语义和尺度不一致的特征并缓解特征图的初始集成可能带来的问题，其融合过程见图3-5。

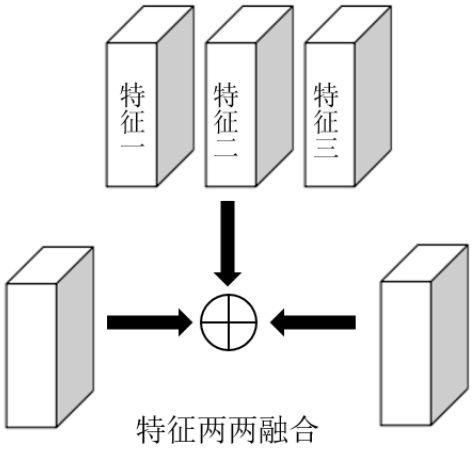

图3-5 多尺度通道注意力模块特征融合

3.4 实验结果分析

为验证本章所使用的语音特征、语音去噪方法、特征融合方法、预训练模型针对阿尔兹海默症识别的有效性，本节在 NCMMSC2021 数据集中展开大量实验进行探索。

3.4.1 实验数据

本章的实验均在语音模态的NCMMSC2021数据集中进行，该数据集来自于2021年全国人机语音通讯学术会议设立的阿尔兹海默症识别竞赛，全程由江苏师范大学、清华大学联合海天瑞声公司负责，旨在促进基于汉语口语数据对阿尔兹海默症的识别研究。

NCMMSC2021的训练集是来自123名受试者的280条录音音频，测试集选取的是长录音任务，录音内容包括自由谈话、看图说话、流畅性测试等。音频口音主要是以江苏方言为主，其中也包含少量普通话和各省方言，通过对数据集音频的抽样（280条中抽取210条）试听，其涵盖的语音任务如下：

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/096002051234010031