一种基于梅尔频率倒谱系数和单层卷积神经网络的声音疾病检测系统

Xiaoping Xie、Hao Cai、Can Li,

摘要—嗓音疾病的发病率逐年上升。 使用软件进行远程诊断是技术发展的趋势，具有重要的实际价值。在嗓音疾病中，常见的会导致声音嘶哑的疾病包括痉挛性发声障碍、声带麻痹、声带小结和声带息肉。本文提出了一种可广泛应用于临床的嗓音疾病检测方法。我们与中南大学湘雅医院合作，收集了 61 位不同患者的语音样本。我们提取梅尔频率倒谱系数参数作为输入特征，以数据形式描述语音。我们提出了一种结合梅尔频率倒谱系数参数和单卷积层卷积神经网络的创新模型，以实现快速计算和分类。我们取得的最高准确率为 92%，全面超越了原有研究成果和国际先进水平。此外，我们使用高级嗓音功能评估数据库（AVFAD）评估所提出系统的泛化能力，准确率达到 98%。在临床数据集和标准数据集上的实验表明，对于嗓音疾病的病理检测，我们的方法在准确率和计算效率方面均有显著提升。

索引词—病理性嗓音障碍、梅尔频率倒谱系数、深度学习、疾病检测

一、引言

全球约有 3%–9% 的人患有不同程度的嗓音疾病或出现声音疾病症状。嗓音疾病会影响人们正常的交流、生活和工作质量 [1]。多种因素会导致嗓音病理，例如某些器官严重受损、空气污染、吸烟、饮食习惯以及压力 [2]。尤其是因职业需要而长期大声、频繁说话的人群，例如教师、歌手、拍卖师、律师和演员，患嗓音疾病的概率更高 [3]。临床上，医生通常通过喉镜检查确定喉部病变的位置，并通过喉肌电图诊断喉部肌肉和声带疾病。这两种临床嗓音疾病检测方法在很大程度上依赖医生的临床经验和主观判断，且成本较高 [4]。健康声音与各种病态声音的语音状况和声学特征之间存在直观的生物学对应关系，这也为发声器官产生并传递的语音波形提供了生理学解释。通过提取和处理患者的语音信号特征，可以对嗓音质量进行客观、科学的评估 [5]。作为一种计算机辅助病理评估，我们可以从多个方面着手，包括感知评估、声学评估、空气动力学评估和内窥镜成像，并结合机器学习方法，从数据角度进行预测和评估[6]。与此同时，这些方法还能避免喉镜检查给患者造成二次损伤，并帮助医生以相对低廉的成本诊断患者的嗓音疾病 [7]。

语音的产生过程可描述如下：首先，肺部呼吸，气流使声带规律振动并产生声波，最后声波经口腔和鼻腔传出。通常，患有嗓音疾病的患者发出的声音听起来紧张、生硬，或极其微弱并伴有喘息声。换言之，他们的音质较差。常见的嗓音疾病中，会导致发声障碍的疾病包括痉挛性发声障碍、声带麻痹、声带结节和声带息肉。

痉挛性发声障碍是一种神经性嗓音障碍，由声襞紧张所致。痉挛性发声障碍患者的声襞会不自主地紧张，从而影响语音 [8]。声带麻痹主要由迷走神经和喉返神经周围组织的病变、侵犯、压迫及其他损伤引起，进而导致声带运动障碍。其临床表现为声音嘶哑、无力，甚至失语症 [9]。声带结节是一种特殊类型的

第一条脚注中的这段文字将注明您提交论文供审阅的日期，该日期由 IEEE 填写。按照 IEEE 的格式，支持信息（包括赞助方及经费支持致谢）应显示在此处，而不是放在文章末尾的致谢部分。例如：“本研究部分由美国商务部通过资助编号 123456 提供支持。”通讯作者姓名列于经费信息之后，例如（通讯作者：M. Smith）。您也可以在此注明作者是否作出了同等贡献，或是否有共同第一作者。

接下来的几段应包含作者目前的任职信息，包括当前地址和电子邮箱。例如，第一作者 A. Author 任职于美国科罗拉多州博尔德市 80305 的美国国家标准与技术研究院（电子邮箱：author@ boulder.nist.gov）。

第二作者 B. 曾就职于莱斯大学，地址为美国得克萨斯州休斯敦，邮编 77005。现就职于科罗拉多州立大学物理系，地址为美国科罗拉多州柯林斯堡，邮编 80523（电子邮箱：author@lamar.colostate.edu）。第三作者 C. Jr. 就职于科罗拉多大学电气工程系，地址为美国科罗拉多州博尔德，邮编 80309；目前借调自日本筑波的国立金属材料研究所，邮编 305-0047（电子邮箱：author@nrim.go.jp）。

此处可提及补充材料以及动物/人类权益声明。本文一幅或多幅图的彩色版本可在线访问：http://ieeexplore.ieee.org

慢性喉炎，可定义为声带边缘出现的小型良性增厚。主要临床症状是声音粗糙，通常表现为高音调时声音嘶哑，并伴有发音延迟和音色变化 [10]。声带息肉是发生于固有声带浅表层的良性增生性病变，主要临床症状是声音嘶哑。

研究人员通常会针对一种或两种疾病的预测分类开展算法研究。单一分类和二分类的准确率通常已达到 90% 以上，但四分类研究的准确率不到 80% [11][12]。

在本文介绍的研究中，我们提出了一种基于深度学习的方法，可有效区分上述四种疾病。本文结构如下：第二节介绍相关工作；第三节描述数据集；第四节阐述方法，包括数据预处理、特征提取和算法模型；第五节展示实验测试及结果。在这一节中，我们采用多种评价指标，测试系统对临床音频的预测分类效果。最后，第六节使用一个更具说服力的数据集，讨论模型的效果。

II. 相关工作

A. 数据集

马萨诸塞州眼耳医院数据库（MEEI数据库）是使用最广泛的数据库之一，包含多种器质性、神经性、创伤性和心因性嗓音障碍的录音。然而，MEEI数据库记录的语音样本来自不同的环境和频率，这可能会给研究人员深入研究带来缺陷 [13]。阿拉伯语嗓音病变数据库（AVPD）通过录制受试者的三个元音、连续语音和孤立词来收集样本。尤其是，该数据库提供了每个样本的感知严重程度 [14]。萨尔布吕肯语音数据库（SVD）包含患有多种病变的受试者的语音样本，这些病变的录音可分为功能性和器质性 [15]。VOICED 是一个专注于自动语音障碍检测与分类的数据库，但其样本量比其他数据库小，这是它的不足之处。AVFAD 是一个开放获取资源，包含临床确诊为声带病变的 346 名患者和 363 名无嗓音异常者的样本。它包含延长元音 /a, i, u/、句子朗读、平衡文本朗读和自发语音 [16]。

B. 特征

模式识别需要提取特征。一组能够有效表征病理性嗓音特征的声学特征参数，可以使预测结果更加准确。L. Salhi 等人利用基频、共振峰和小波系数等传统声学特征，实现了正常语音与病理语音的分类系数 [17]。A. Leonardo、M. Forero 等人提取时域和频域参数，并利用这些参数表征声门信号基频的变化 [18]。Ghulam Muhammad 等人将 MPEG-7 音频特征应用于哮喘疾病的语音分类，并取得了令人满意的结果 [19]。还有一些研究提取韵律、声道和激励特征 [12]。为丰富特征，部分学者构建并筛选出包含不同特征的显著特征集，例如扰动特征、频谱与倒谱特征以及复杂度特征 [20]。在 [21-25], 中，梅尔频率倒谱系数（梅尔频率倒谱系数）广泛用于语音信号特征提取，并取得了优异成果。

C. 算法数十年来

研究已证明，机器学习方法（支持向量机、高斯混合模型、极限学习机）应用于健康与疾病语音分类时，效果有效且准确。研究人员研究并应用了多种深度学习架构，包括五层网络、五层卷积神经网络和循环神经网络。在实验中，五层卷积神经网络取得了令人满意的结果，灵敏度为 96%，特异度为 18% [26]。此处提到的灵敏度和特异度与医学领域中的召回率和精确率相关。打个比方，灵敏度是指给定一系列样本时，能够判断出的患者所占比例；特异度则是确定某种疾病的能力。在 [12] 中，有研究提出使用深度置信网络提取特征，并结合支持向量机完成分类。此外，一些研究人员构建并改进了深度神经网络模型，该模型包含 5 个隐藏层，每层有 200 个神经元。通过调整参数来平衡灵敏度和特异度，最终分别达到 93.1% 和 46 % [21]。传统机器学习模型（如支持向量机、随机森林、K近邻和梯度提升）在分类任务中也取得了可靠的结果 [23][25]。针对四分类问题，研究人员提出了结合卷积架构与前馈神经网络的方法，对四种嗓音障碍进行分类，准确率达到 57% [27]。

III. 材料

A. 临床样本

在征得患者及其家属同意后，我们选取了2021年1月至2022年12月期间湘雅医院的相关患者作为研究对象。医生通过病史询问、体格检查和电子喉镜检查等现代医疗手段对患者进行诊断。我们计划首先设计并测试嗓音疾病监测系统，目标是区分这些患者的特定语音音频。

患者在医务人员的陪同和指导下，发出指定音素。由于研究对象均为成年中国人，我们提出发音音素为汉语元音 /a, o, e, i, u, ü/ 和 /wo, ei/。

在大多数数据库或研究中，研究人员关注的是单个元音（/a/）或三个元音组合（/a, i, u/）的连续发音[29][30]。客观而言，这些研究仅考察了语音的一个方面；我们将其扩展至 8 个音素，以避免大多数语音病理检测策略的局限。在语音采集过程中，我们使用了录音笔、手机等相对简单的采集设备进行录音。我们希望实验的音频质量尽可能差，以贴合实际临床环境和设备条件。

最后，我们创建的患者音频数据集包含四种疾病患者的 1464 段语音音频片段，每段约 3 秒。数据集包含以下文件：

1) 音频文件八个音素中的每个都朗读三次，每次约 3 秒，音频文件的详细分布如表 I。 2) 病例报告医院诊断记录包括医生的临床判断和病理分析、患者的生理信息和日常生活习惯等。

表 I。音频文件分布

类别	/a/	/o/	/e/	/i/	/u/	/ü/	/w/	/ei/
痉挛性发声障碍	36	36	36	36	36	36	36	36
声带麻痹	54	54	54	54	54	54	54	54
声带结节	33	33	33	33	33	33	33	33
声带息肉	60	60	60	60	60	60	60	60

IV. 基于梅尔频率倒谱系数和卷积神经网络的方法设计

几乎所有自动语音识别系统的初始阶段，都是识别有助于内容检测或特征提取的音频成分。本文这部分研究工作分为三个部分，如图 1所示。首先，对采集到的 WAV格式音频信号进行预处理；然后提取预处理后音频信号的特征；最后建立算法模型，用于训练和测试。

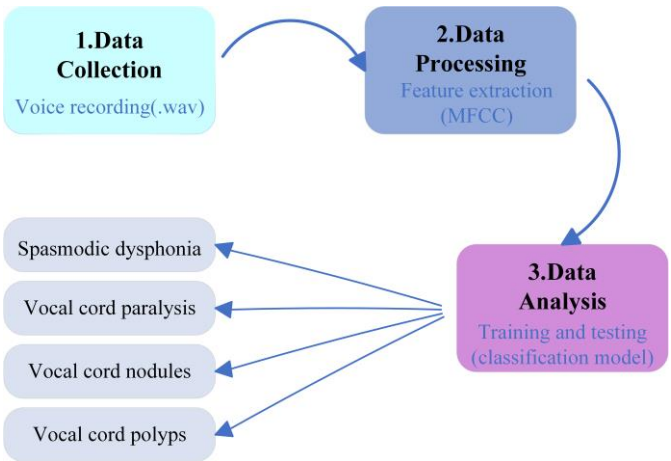

图 1. 病理语音分类框图

A. 音频预处理

采集到的元音音频，其能量分布主要集中在 1kHz 以下。高频带中的信号对本研究影响可忽略，因此我们跳过了预加重步骤。本文的预处理过程包括两个步骤，即语音活动检测和音频裁剪。

考虑到采集过程中的客观因素，患者的录音通常是一段较长的音频，其中包含一些无用信息。语音活动检测可以在信号数据中找到有效语音片段的起点和终点，从而便于截取音频片段。最后，我们将有效音频截取为若干段长度为 0.5s 的短音频数据，每段仅包含一个音素。

B. 特征提取

在特征提取阶段，我们参考了 [27][31] 中的做法。梅尔频率倒谱系数和梅尔频谱图在表征语音信号信息方面效果良好。梅尔频率倒谱系数在一定程度上模拟了人耳对声音的感知特性。人耳的听觉特性与梅尔频率的增加相一致：实际频率低于 1000Hz 时呈线性分布，高于 1000Hz 时则呈对数增长。

梅尔频率倒谱系数的提取过程如图2所示。整个过程可描述如下：输入 WAV格式的语音信号后，对其进行预加重、分帧、加窗和 FFT 处理。此操作将连续的语音波形离散化，并将其转换为由较大数值构成的一维数组。随后通过一组梅尔尺度三角带通滤波器，再进行对数运算并缩放纵坐标。最后，通过离散余弦变换计算 n 阶梅尔频率倒谱系数。

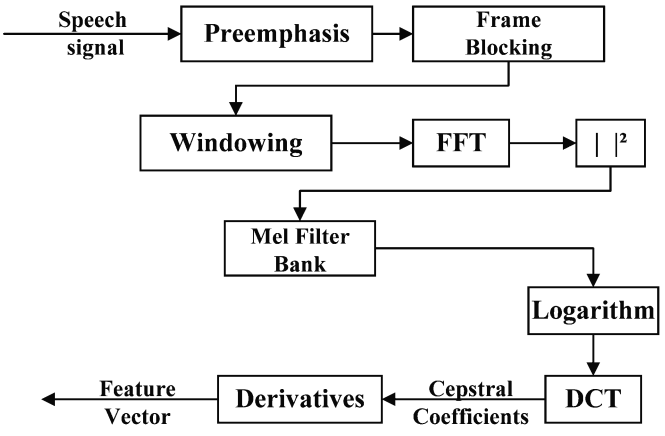

图2. 梅尔频率倒谱系数的整体流程

● 预加重（提高语音的高频分辨率）

$$signal[i] = signal[i+1] - a * signal[i] \quad (1)$$

a：预加重系数，取值为 0.97 .

● 分帧（将采样点收集成一个观测单元）

$$frame\ num = \frac{(N - win + inc)}{inc} \quad (2)$$

● 频带中的频谱 W(n)

$$x_1[i] = x_0[i] * w[i] \quad (3)$$

将每一帧的 f 与汉明窗中对应的元素相乘，使整体更加连续，并避免不连续边界处出现吉布斯现象。

● 快速傅里叶变换（FFT）

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\left(\frac{2\pi}{N}\right)kn} \quad (4)$$

获得 FFT 数组中的每个元素表示一帧语音信号的能量，也可以体现不同语音的特征。同时，对频谱的模取平方即可得到信号的能量谱。

● 梅尔滤波器组

对频谱进行平滑处理，以消除谐波的影响，并突出原始语音的共振峰。值得注意的是，这一操作表明，基于梅尔频率倒谱系数的语音识别模型不会受到输入语音基频的影响。

$$Y(m) = \sum_{k=1}^{N} H_m(k) |X[k]|^2 \quad (5)$$

$H_m(k)$: fre quency response

● 梅尔频谱图：通过

将能量谱通过梅尔滤波器，即可得到如图 3 所示的梅尔频谱图。梅尔频谱图以更直观的方式描述音频流在不同频率下的振幅。每个频率的振幅以热图形式呈现，颜色越亮，信号能量越高，表明声音样本集中在特定频率范围内。图 3 中的四幅图展示了选取的四位嗓音疾病患者发出 /a/ 音时的梅尔频谱图。理论上，只要分辨率足够高，我们就能通过这类图区分各个样本。但实际上，这只能视为一种可视化描述，仍需从数学角度进行客观判断。

● 离散余弦变换（去相关并剥离低频部分）

$$s(m) = log\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right) \quad (6)$$

$0 \le m \le M$　M: the number of filters

$$c(n) = \sum_{m=0}^{N-1} s(m) \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \quad (7)$$

n=1,2,3,4...L L：梅尔频率倒谱系数的阶数

按照标准做法，保留前 13 个梅尔频率倒谱系数，其余部分将被舍弃 [32]。我们尝试选取 13、40、128 个梅尔频率倒谱系数进行实验（包括静态梅尔频率倒谱系数以及一阶、二阶和高阶差分系数），以有效、全面地表征语音信号的静态和动态特征。最后，我们提取了二维输入特征。

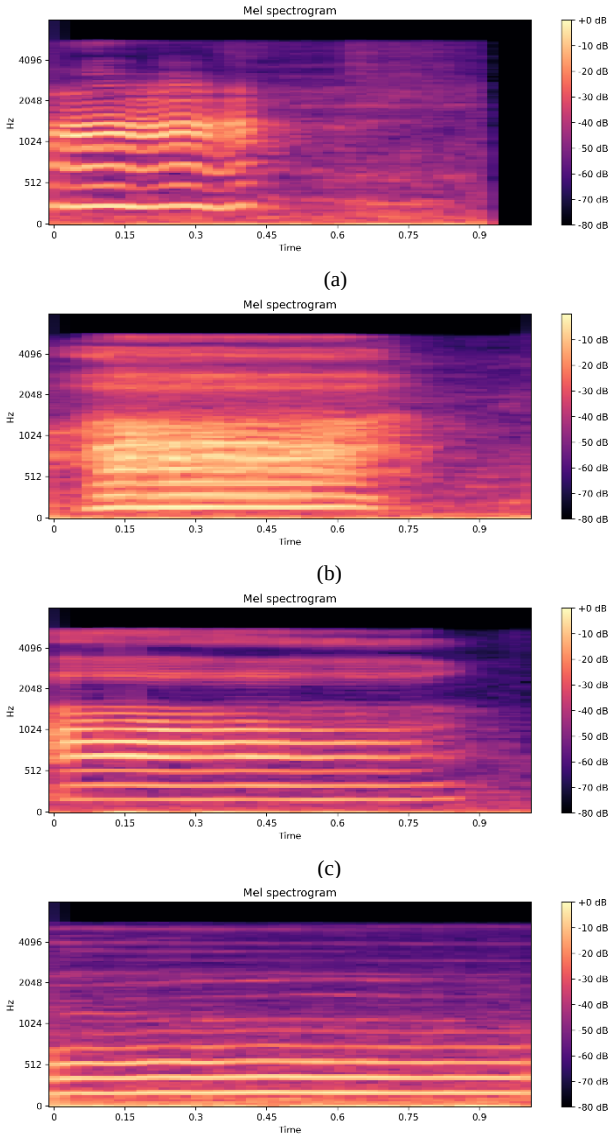

图 3。四类病理样本的梅尔频谱图。(a)痉挛性发声障碍患者的梅尔频谱图；(b)声带麻痹患者的梅尔频谱图；(c) 声带结节患者的梅尔频谱图；(d) 声带息肉患者的 梅尔频谱图。

C. 卷积神经网络模型

本文的研究目标是对嗓音疾病患者的语音进行分类，并识别其所患疾病类型。模型的输入是经过处理的梅尔频率倒谱系数特征张量，输出是疾病标签向量。我们希望模型具备结构简单、效率高、准确率高等特点。因此，我们研究了一种基于学习的模型，其结构如图 4所示。

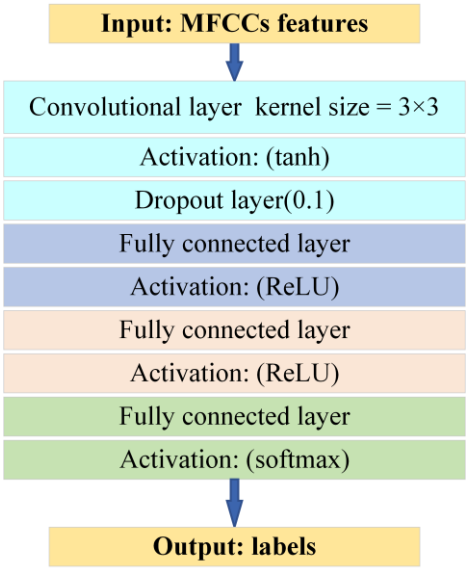

图 4. 卷积神经网络模型架构

在图像分类问题中，存在许多“冠军”模型，它们无一例外都具有非常复杂的深层网络结构，例如 GoogleNet、VGG、ResNet 等。本文也尝试将这些模型应用于嗓音疾病分类，但结果并不理想，准确率和召回率都不高。

卷积神经网络的核心思想是稀疏交互，而多层神经元对前一层数据的选择性感知不利于本文中的语音数据处理。特征数据会不断被压缩和映射，多层卷积运算更关注特征的局部相关性，却缺乏对特征的理解和全局表达。我们尝试了几种多层卷积网络结构，首先使用了 5 层二维卷积层，其卷积核大小为 3×3。同时，我们尝试改变卷积核函数的大小（例如 5 × 5和 7 × 7）以及层数。最终发现，使用 3×3 卷积核函数的一层卷积取得了令人满意的结果。同时，通用逼近定理指出，包含有限个神经元的单隐藏层前馈网络可以在 n[33]的紧子集上逼近连续函数。我们认为，具有一个隐藏层的网络可以逼近我们所需的分类函数。公式（8）显示，输入 X 与卷积核函数 k 进行互相关运算，并加上一个标量偏置，得到卷积层的输出 Y_0 。

$$Y_0 = X * k + b_0 \tag{8}$$

需要指出的是，双曲正切函数通常用于循环神经网络的网络结构中，而ReLU函数则常用于卷积神经网络。在多层卷积神经网络中，ReLU函数的输出既有>1 也有<1 值。经过多层叠加和相乘后，若干大于1和小于1的值会相互补偿。从某种意义上说，这抵消了梯度爆炸的可能性。然而，如果只有一层卷积，就不会出现上述补偿情况，使用ReLU函数会导致不收敛，如图5所示。同时，ReLU函数还会导致神经元死亡，这是由其数学特性导致的。我们还尝试了带泄漏的ReLU函数。结果有所改善，但仍存在波动。使用双曲正切函数作为激活函数可以有效解决波动问题。

$$Y_1 = tanh(Y_0) = \frac{e^{Y_0} - e^{-Y_0}}{e^{Y_0} + e^{-Y_0}} \tag{9}$$

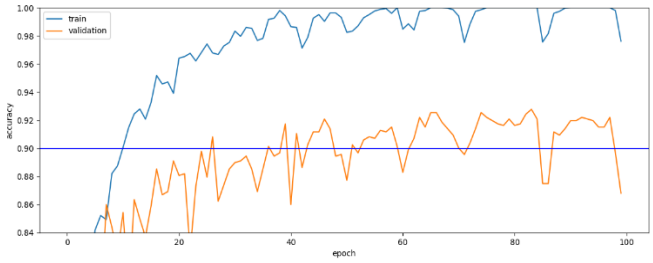

图5. ReLU激活函数的准确率变化

对于我们训练的样本，训练模型容易因依赖局部特征而发生过拟合。Dropout层以概率p暂时丢弃一半隐藏神经元，并更新未被丢弃神经元的参数，重复执行该操作，如图6所示。左图展示了一般的层结构，右图中带有 × 的圆圈表示暂时被丢弃的神经元。该操作相当于对不同神经网络的过拟合（包括正向拟合和反向拟合）取平均[34]。它还减少了具有固定关系的隐藏节点之间的相互作用，效果类似于L1和L2正则化，并增强了网络的鲁棒性。Dropout层的操作还增加了特征稀疏性，并提高了泛化能力。

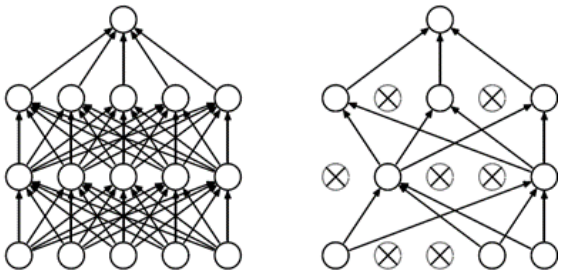

图6. Dropout层示意图

$$r_j \sim Bernoulli(p = 0.1)$$
$$Z_1 = W_1(r_j * Y_1) + b_1 \tag{10}$$

将 m 个一维数据展平（1×192），然后使用三个全连接层映射特征，并将最终输出维度缩减为 4（即最终分类标签的数量）。最后一个线性层采用‘softmax’函数激活，其余层采用 ‘ReLU’ 激活。映射过程如图 7所示，公式 11 至 13 展示了运算过程。

$$H_1 = ReLU(Z_1 W_2 + b_2) \tag{11}$$
$$H_2 = ReLU(H_1 W_3 + b_3) \tag{12}$$
$$O = softmax(H_2 W_4 + b_4) \tag{13}$$

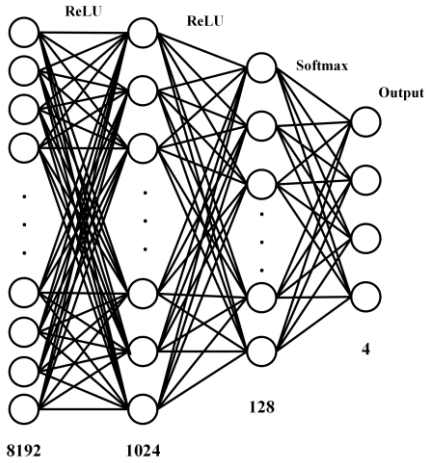

图 7。全连接层架构

V. 结果

在机器学习训练过程中，我们不仅要训练模型，还要利用数据调整其超参数并评估模型的性能。这里，我们将数据集随机划分为三部分：训练集、验证集和测试集。数据集的 80% 用于在训练集上训练模型，10% 用于验证，以确定最优网络结构，最后 10% 用于评估模型效果。

评估模型效果时，我们选取了三个常用的评价指标：准确率、召回率和 F1 分数[35]。同时结合 ROC 图，从“准确”和“完整”两个角度对模型进行客观评估。

首先，我们开始测试梅尔频率倒谱系数的数量，并从 13、40、50、128 中找出最佳数量。结果如表 II所示。

表 II。不同数量梅尔频率倒谱系数的准确率

梅尔频率倒谱系数	13	40	50	128
准确率	73%	83%	85%	92%

找到最合适的输出特征参数后，我们描述模型的整体准确率和损失，如图 8 和图 9所示。

在图 8中可以看到，随着训练轮次增加，模型的训练准确率趋于稳定在 1，而测试准确率在上升后逐渐稳定在约 0.92。在图 9中，训练集的模型交叉熵损失接近 0.01，测试集的损失也稳定在约 0.3。损失曲线符合正常的变化趋势，没有出现过拟合、学习不足等情况。

混淆矩阵通常用作一级指标。如图 10所示，以模型分类的样本数和实际类别的样本数分别作为横坐标和纵坐标，直接描述模型的分类性能。通过热力图的形式，可以非常直观地看出每个类别的预测值与实际值之间的差异。

混淆矩阵统计样本数量，即真实样本和预测样本的数量。在图 10中可以看到每个类别预测正确和错误的数值，混淆矩阵中有几个音频被误判。然而，在实际应用中，我们会对同一位患者的多个音频进行投票。四个类别中得票数最多的类别就是最终预测的疾病类别。实际诊断准确率应超过 92 %，达到 98% 至接近 100%；这也是实际应用与学术研究之间的差异。

然后，我们计算底层指标，得到次级指标结果，如 表 III所示。

表 III语音数据集上的性能

类别	精确率	召回率	F1 分数	支持
痉挛性发声障碍	0.9252	0.9519	0.9384	104
声带麻痹	0.8230	0.8158	0.8194	114
声带结节	0.7624	0.8280	0.7938	93
声带息肉	0.9018	0.8632	0.8821	234

对于需要进行多分类的四个标签，我们在对模型进行分类时将阈值设为 0.5。因此，我们绘制了 ROC 曲线，如图 11所示，以反映模型对各标签的灵敏度。同时，我们计算 ROC 曲线下面积（ROC-AUC 分数），以反映分类器的效果。曲线下面积的值越高，对分类器的评价就越好。

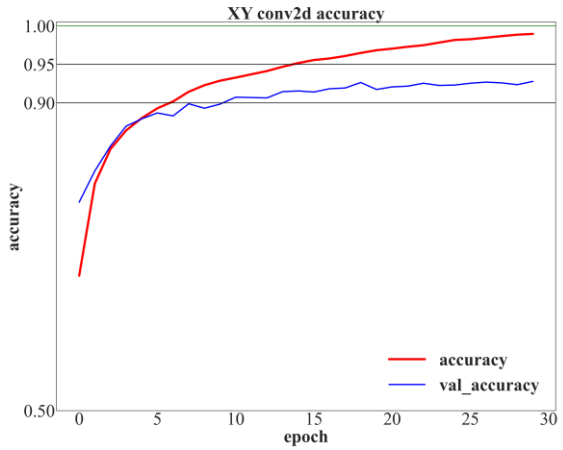

图 8. XY 数据集上模型的准确率

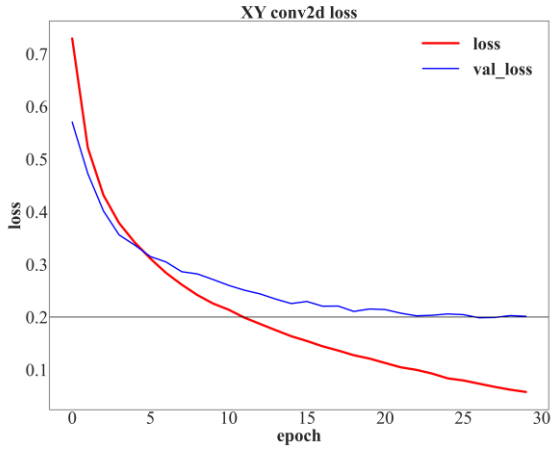

图 9. 模型在 XY 数据集上的损失

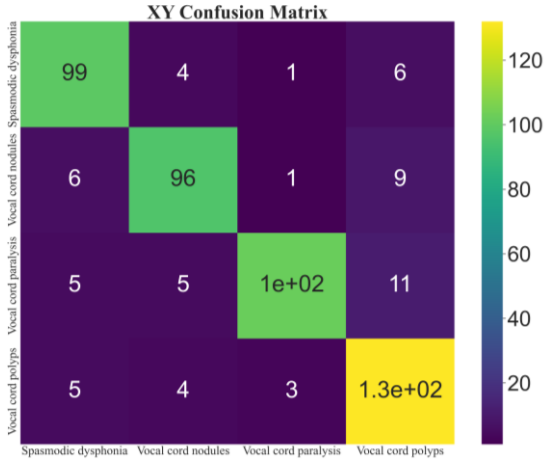

图 10. 模型在 XY 数据集上的混淆矩阵

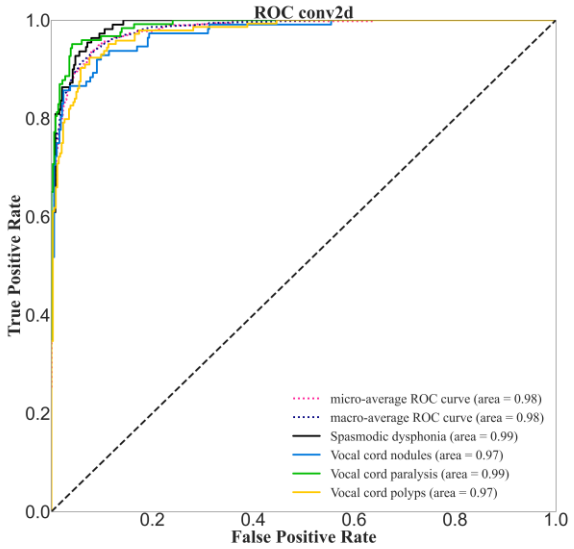

图 11. 模型在 XY 数据集上的 ROC 曲线

VI. 讨论

基于数据的自动化模型能够为临床医生提供客观的判断辅助。从上一节的模型测试结果可以看出，准确率达到了 92%。值得一提的是，我们的模型处理语音和音频的速度很快。大多数与嗓音疾病相关的检测与分类系统都使用准确率、特异度和灵敏度作为评价指标。然而，基于执行时间的评价同样重要，尤其是在临床应用语音病理检测与分类系统时。等待时间过长可能会影响患者的情绪和心理状态，也会降低患者对计算机辅助诊断工具的信任。

此外，我们认为该方法也具有很强的普适性，即泛化能力可靠。我们使用高级语音功能评估数据库（AVFAD）作为评估模型泛化能力的对象。在各种条件下，我们进行了若干相关实验。实验结果如表 IV-VI所示。

1) 从 AVFAD 中提取 13、49、50、128 个梅尔频率倒谱系数；2) 将单音或多音组合用作输入，我们尝试了 /i/、/a, i/、/a, i, u/ 三种条件，并各重复 3 次。

表 IV.使用三元元音时不同梅尔频率倒谱系数下的准确率

梅尔频率倒谱系数	13	40	50	128
准确率	92%	98%	98%	98%

表 V使用二元元音时不同梅尔频率倒谱系数下的准确率

梅尔频率倒谱系数	13	40	50	128
准确率	91%	97%	97.5%	98%

表 VI使用单个元音时不同梅尔频率倒谱系数下的准确率

梅尔频率倒谱系数	13	40	50	128
准确率	90%	97%	97%	98%

以三元元音作为分析对象的输入音频具有更好的分类效果，准确率约为98%。以下是随着梅尔频率倒谱系数增加而变化的准确率曲线图像。

梅尔频率倒谱系数值越高，检测准确率越高。对于大型数据集（AVFVD），准确率在约为50时达到峰值，随后趋于保持不变。然而，曲线振荡的情况仍然十分明显，而较大值（128）的梅尔频率倒谱系数有所改善，也就是说，稳定性更好，拟合效率也有所提高，正如图12-15所示。

同时，对于湘雅数据集，我们发现，小型数据集需要使用较大的梅尔频率倒谱系数才能取得更好的结果，而大型数据集取适中的值即可达到较高的准确率。

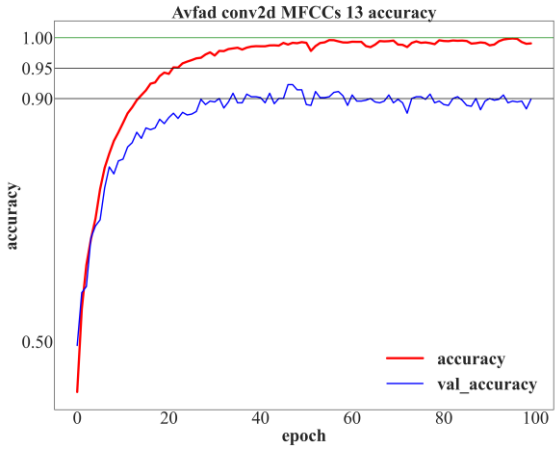

图 12. 使用 13 个梅尔频率倒谱系数时模型在 AVFAD 上的准确率

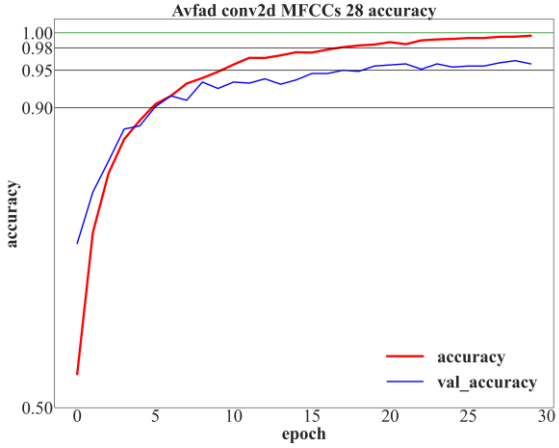

图 13. 使用 28 个梅尔频率倒谱系数时模型在 AVFAD 上的准确率

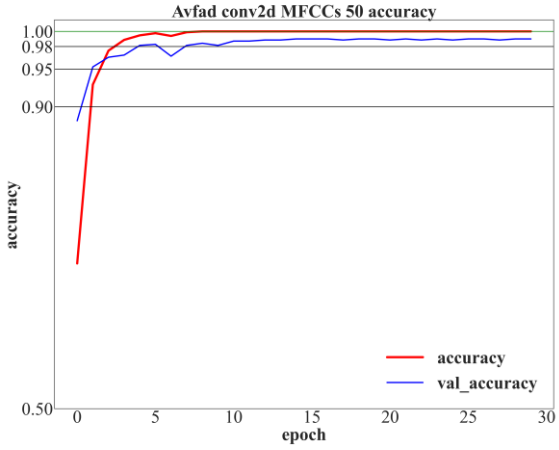

图 14. 使用 50 个梅尔频率倒谱系数时模型在 AVFAD 上的准确率

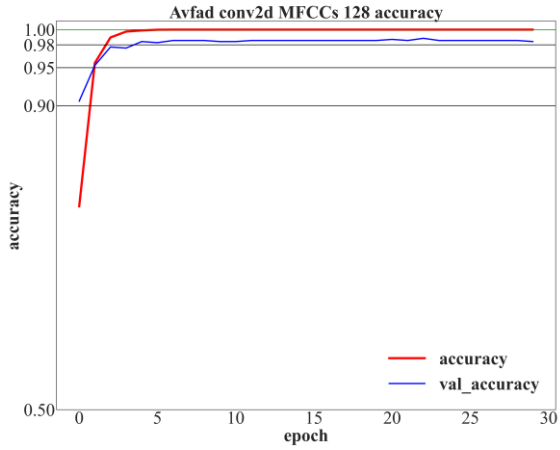

图 15. 使用 128 个梅尔频率倒谱系数时模型在 AVFAD 上的准确率

VII. 结论

为了满足湘雅医院的临床需求并提高嗓音疾病的诊断效率，我们提出了一种更加准确、快速的方法，用于对四种常见的嗓音疾病进行分类。首先，我们使用在临床环境中采集的正常音质患者音频（8 个音素）训练模型，调整参数，并测试分类效果，准确率达到 92%，召回率达到 86%。为证明该方法的普适性，我们在 AVFAD 上进行了验证，准确率超过 98%。

本文的贡献主要体现在以下几个方面：

1) 大多数关于嗓音疾病的病理研究都聚焦于健康语音和病理语音这两类，研究所用的音频也都是单元音 /a/。本文采集了 8 个音素，用于对疾病本身的四个类别进行分类。2) 与大多数基于卷积神经网络的语音分类模型不同，我们提出了一种单卷积层结构，并证明其分类效果更准确、速度更快。3) 在实验中，我们总结出了语音特征梅尔频率倒谱系数的一般规律：大型数据集采用适中的梅尔频率倒谱系数值，小型数据集采用较大的梅尔频率倒谱系数值，这能够弥补输入特征的缺陷，并取得良好的分类结果。

我们认为，湘雅数据集中的患者信息同样具有重要的研究价值，例如工作环境、饮食习惯等，这些信息可作为辅助语音疾病研究的重点。将这些因素纳入研究，可以使预测更具前瞻性。此外，通过图3中梅尔图的二维描述，我们认为每位患者的图像也具有特异性，我们还可以从中学习特征，并在此基础上应用先进的图像识别方法，这也是我们的研究方向。

致谢

我们衷心感谢湘雅医院耳鼻咽喉科团队收集了如此宝贵的语音数据，并提供了专业的医学知识。此外，我们还要感谢阿威罗大学为我们提供 AVFAD，以验证我们的模型。

参考文献

[1] J. Morawska 和 E. N. Bogusz，《不同职业群体中嗓音障碍的风险因素与患病率——文献综述》，Otorynolaryngologia Przegla，D Kliniczny，第16卷，第3期，第94-102页，2017年。[2] Gómez、Pedro 等，《通过声带生物力学参数估计进行嗓音病理检测》，《Nonlinear Analyses and Algorithms for Speech Processing》：International Conference on Non-Linear Speech Processing，第242-256页，2005年。[3] A. Al-nasheri、G. Muhammad、M. Alsulaiman 和 Z. Ali，《使用相关函数研究不同频率区域中的嗓音病理检测与分类》，J. Voice，第31卷，第1期，第3-15页，2017年1月。[4] F. T. Al-Dhief、N. M. A. Latiff、N. N. N. A. Malik、N. S. Salim、M. M. Baki、M. A. A. Albadr 和 M. A. Mohammed，《基于物联网和机器学习算法的嗓音病变监测系统综述》，IEEE Access，第8卷，第64514–64533页，2020年。[5] O. Saz、J. Simón、W. R. Rodríguez、E. Lleida 和 C. V aquero，《认知障碍和言语障碍说话人的声学特征分析》，EURASIP J. Adv. Signal Process，第2009卷，第1期，第1-11页，2009年12月。[6] Mehta, Daryush D. 和 Robert E. Hillman，《嗓音评估：感知、声学、空气动力学和内窥镜成像方法的最新进展》，Current opinion in otolaryngology & head and neck surgery，第16卷，第3期，第211-215页，2008年。[7] W. Y uanbo、Z. Changwei、F. Ziqi、Z. Yihua、Z. Xiaojun 和 T. Zhi，《使用机器学习分类器进行嗓音病理检测与多分类》，载于 Proc. Int. Conf. Sens., Meas. Data Anal. Era Artif. Intell. (ICSMD)，第319-324页，2020年10月。[8] Murthy, G. L. N. 等，《使用随机森林框架检测痉挛性发声障碍嗓音病理的新算法》，2022 International Conference on Edge Computing and Applications (ICECAA)，第800-804页，2022年。[9] Maisel, Robert H. 和 Joseph H. Ogura，《声带麻痹的评估与治疗》，The Laryngoscope，第84卷，第2期，第302-316页，1974年。[10] Lancer, Jack M. 等，《声带结节：综述》，Clinical Otolaryngology & Allied Sciences，第13卷，第1期，第43-51页，1988年。[11] Wang, Syu-Siang 等，《用于改善病理性嗓音障碍学习的连续语音》，IEEE Open Journal of Engineering in Medicine and Biology，第25-33页，2022年。[12] Islam, Kazi Aminul、Daniel Perez 和 Jiang Li，《2018年FEMH嗓音数据挑战赛的迁移学习方法》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5252-5257页，2018年。[13] Cesari, Ugo 等，《健康嗓音和病理性嗓音的新数据库》，Computers & Electrical Engineering，第310-321页，2018年。[14] Mesallam, Tamer A. 等，《阿拉伯语嗓音病变数据库的开发及其使用语音特征和机器学习算法进行的评估》，Journal of Healthcare Engineering，2017年。[15] Pützer, Manfred 和 Jacques Koreman，《病理性声带振动模式的德语数据库》，Phonus 3，第143-153页，1997年。[16] Jesus, Luis MT 等，《高级语音功能评估数据库（AVFAD）：面向嗓音临床医生和语音研究的工具》，Advances in Speech-language Pathology。IntechOpen，2017年9月13日。[17] Salhi, Lotfi、Mourad Talbi 和 Adnane Cherif，《使用混合方法识别嗓音障碍：小波分析与多层神经网络》，International Journal of Electrical and Computer Engineering，第2卷，第9期，第3003-3012页，2008年。[18] Kohler, Manoela、Marley MBR Vellasco 和 Edson Cataldo，《使用声门信号参数分析和分类嗓音病变》，Journal of Voice，第30卷，第5期，第549-556页，2016年。[19] Muhammad, Ghulam 和 Moutasem Melhem，《使用 MPEG-7 音频特征进行病理性嗓音检测与二分类》，Biomedical Signal Processing and Control，第1-9页，2014年。[20] Arias-Londoño, Julián David 等，《面向2018年FEMH挑战赛的Byovoz自动语音状况分析系统》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5228-5232页，2018年。[21] Chuang, Zong-Ying 等，《用于检测和分类病理性嗓音的基于深度神经网络的方法》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5238-5241页，2018年。[22] Degila, Kevin、Rahhal Errattahi 和 Asmaa El Hannani，《面向2018年FEMH嗓音数据挑战赛的UCD系统》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5242-5246页，2018年。[23] Pham, Minh、Jing Lin 和 Yanjia Zhang，《使用机器学习诊断语音障碍》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5263-5266页，2018年。[24] Ju, Mingxuan 等，《用于嗓音疾病分类的多表示集成方法》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5258-5262页，2018年。[25] Pishgar, Maryam 等，《使用梅尔倒谱向量和支持向量机进行病理性嗓音分类》，arXiv预印本 arXiv:1812.07729，2018年。[26] Ramalingam, Archana、Sharat Kedari 和 Chandrasekar Vuppalapati，《2018年IEEE FEMH语音数据挑战赛》，载于 2018年IEEE国际大数据会议（Big Data）。IEEE，第5272-5276页，2018年。[27] Miliaresi, Ioanna、Kyriakos Poutos 和 Aggelos Pikrakis，《在深度学习网络中融合声学特征和医疗数据进行语音病理分类》，载于 2020年第28届欧洲信号处理会议（EUSIPCO）。IEEE，第1190-1194页，2021年。[28] Benba, Achraf、Abdelilah Jilbab 和 Ahmed Hammouch，《使用梅尔频率倒谱系数在倒谱域分析多种语音录音，以区分帕金森病患者与健康人群》，《国际语音技术杂志》，第449-456页，2021年。[29] D. Martínez、E. Lleida、A. Ortega、A. Miguel 和 J. Villalba，《通过校准和分数融合，使用multifocal工具包在萨尔布吕肯语音数据库上进行嗓音病理检测》，载于《伊比利亚语言语音与语言技术进展》，第99–109页，施普林格，柏林，海德堡，2012年。[30] N. Souissi 和 A. Cherif，《用于识别嗓音障碍的人工神经网络和支持向量机》，《国际高级计算机科学与应用杂志》，第7卷，第5期，第339–344页，2016年。[31] Ilapakurti, Anitha 等，《用于语音障碍检测的人工智能（AI）临床边缘计算》，载于 2019年IEEE第五届大数据计算服务与应用国际会议（BigDataService）。IEEE，第340-345页，2019年。[32] Tiwari, Vibha，《梅尔频率倒谱系数及其在说话人识别中的应用》，《新兴技术国际期刊》，第1卷，第1期，第19-22页，2010年。[33] Lu, Yulong 和 Jianfeng Lu，《用于表达概率分布的深度神经网络通用逼近定理》，《神经信息处理系统进展》，第3094-3105页，2020年。[34] N. Srivastava、G. Hinton、A. Krizhevsky、I. Sutskever 和 R. Salakhutdinov，《Dropout：一种防止神经网络过拟合的简单方法》，《机器学习研究杂志》，第15卷，第1期，第1929–1958页，2014年。[35] Zakariah, Mohammed 等，《基于梅尔频率倒谱系数和深度神经网络的语音病理检测分析研究》，《医学中的计算与数学方法》，2022年。