基于深度神经网络的肺癌风险预测

单紫琳 边辰迪 肖博

北京工业大学，北京市，100124；

摘要:本研究通过采用卷积神经网络（CNN）对肺癌风险进行预测，成功克服了传统 COX 回归模型在处理 CT 图像数据时的局限性。传统 COX 回归模型依赖人工特征提取，需手动选择和提取 CT 图像中的特征，这不仅耗时，还难以捕捉图像中的细微变化，尤其在早期病变或微小肿瘤的检测上表现不足。同时，COX 模型主要依赖临床变量，无法有效处理高维的 CT 图像数据，导致其在复杂图像特征提取上的能力有限。相比之下，CNN 模型通过卷积层和池化层的组合，能够自动提取 CT 图像中的多层次、多尺度特征，如病变区域的形状、纹理和边缘等，无需人工干预，显著提高了特征提取的效率和图像处理的准确性。CNN 还能直接处理高维的 CT 图像数据，捕捉图像中的细微变化，尤其在早期病变检测上表现出更高的敏感性。CNN 模型不仅能够处理 CT 图像数据，还能有效结合临床变量（如年龄、BMI、吸烟年数等），通过多模态数据融合进一步提升模型的预测能力。在模型训练阶段，使用大规模 CT 图像数据集，并结合数据预处理和增强技术（如图像归一化、旋转、翻转等）优化模型性能。实验结果显示，CNN 模型的准确率为 92.4%，召回率为 89.7%，精确率为 94.5%，F1 分数为 91.0%，AUC 值为 0.95，均显著优于传统 COX 回归模型（准确率 81.2%，召回率 77.4%，精确率 84.1%，F1 分数 80.6%，AUC 值 0.85）。未来研究将尝试更深层次的网络架构（如 ResNet、DenseNet）和多模态集成（结合基因组信息、血液检测结果等），以进一步提高模型的综合判断能力，并探索其在临床诊断系统中的应用前景。本研究不仅在理论上具有重要意义，更在实际临床应用中展示了广阔的前景，为肺癌早期检测和风险预测提供了新的技术手段。

关键词：肺癌风险预测；CNN；图像特征提取；临床变量；深度学习
DOI:10.69979/3029-2808.24.11.006

绪论

肺癌是全球死亡率最高的恶性肿瘤之一，发病率随环境污染和人口老龄化逐年上升。尽管现代医学在肺癌治疗上取得进展，患者的 5 年存活率仍低于 20%，凸显了早期筛查和诊断的重要性。近年来，深度学习技术在医学影像分析领域的突破为肺癌早期诊断提供了新思路。

国内研究已取得显著进展：2017 年，Feng 等人利用迁移学习方法，基于预训练网络参数对 LIDC-IDRI 数据集进行微调，成功检测出 CT 图像中的二维病灶信息；2018 年，ZHAO 等人提出多输入网络模型，通过结节候选区域模板检测肺部 CT 图像中的病灶，实验精度达 85.51%；2023 年，Nature Medicine 报道了达摩院的 PANDA 模型，结合“平扫 CT+AI”技术，显著提高了早期胰腺癌的识别率。国外研究也在不断探索新方法，2023 年，Sharma Divya 和 Xu Wei 提出结合遗传数据和通路结构的深度学习框架，利用卷积神经网络（CNN）及其正则化层全面预测疾病风险，解决了神经网络模型的可解释性问题。

然而，肺癌早期诊断仍面临挑战。传统诊断方法存在效率低、准确性不足等问题，现有深度学习模型在处理复杂临床数据时的泛化能力和解释性仍需提升。因此，本研究致力于开发一种高效、准确的肺癌早期诊断系统，结合多层神经网络模型对医学临床数据（如年龄、性别、吸烟史等）和 CT 影像数据进行建模，利用带标签的数据训练和验证模型，以提高准确度和泛化能力。同时，本研究将探索优化模型结构和训练策略，提出一种综合多模态数据的肺癌早期诊断方法，有望显著提高早期诊断率，改善患者预后和生活质量。

1 CNN 处理 CT 图像与特征自动提取原理

与传统的 COX 回归模型不同，COX 模型在处理 CT 图像时，特征提取方法无法捕捉微小细节，且需对每个病例进行耗时的精细特征选择，缺乏普适性。为解决这一问题，本研究采用卷积神经网络（CNN）。CNN 在图像处理领域表现出色，能够直接从原始 CT 图像中自动提取多层次、多尺度的特征，如病变区域的形状、纹理和边缘等。通过卷积层和池化层的组合，CNN 可自动识别

重要特征，无需人为干预，显著提高了图像处理的效率与准确性。这些特征完全依赖网络在训练过程中自动学习，进一步提升了效率。CNN 是一种专为图像和信号处理任务而设计的神经网络架构。

通过对比传统 COX 回归方法与 CNN 模型，我们发现 CNN 能更好地捕捉 CT 图像中的细微变化，尤其是在早期病变或微小肿瘤的检测上，CNN 显示出了更高的敏感性。这一发现证明了深度学习方法在医学图像分析中的优越性。

2 模型训练与优化

在模型的训练阶段，我们使用了大规模的 CT 图像数据集，并采用了标准的数据预处理方法，如图像的归一化、数据增强（旋转、翻转、平移等），确保了训练过程中的数据多样性与网络的准确性。

卷积层设计：CNN 模型的卷积层使用了多个 3x3 的卷积核，这些卷积层能够在图像中提取局部特征，如肿瘤的边缘、纹理变化等。通过多个卷积层的堆叠，模型能够逐步提取出越来越抽象的高层特征。

池化层设计：为了减少计算复杂度，并有效避免过拟合，我们在每个卷积层后都添加了最大池化层。这不仅能够减小特征图的尺寸，还能保留图像中最重要的特征。

全连接层与 Dropout：在网络的最后，我们通过全连接层将特征整合，输出预测的肺癌风险值。同时，为了防止过拟合，我们在训练中采用了 Dropout 技术，在每次迭代时随机丢弃部分神经元的连接，确保模型具有良好的泛化能力。

优化方法：在模型训练过程中，我们使用了 Adam 优化器，它能够根据每个参数的梯度自适应地调整学习率，帮助模型更快速、更稳定地收敛。我们还采用了早停策略，在验证集上的性能不再提升时及时停止训练，避免过拟合。

3 模型准备

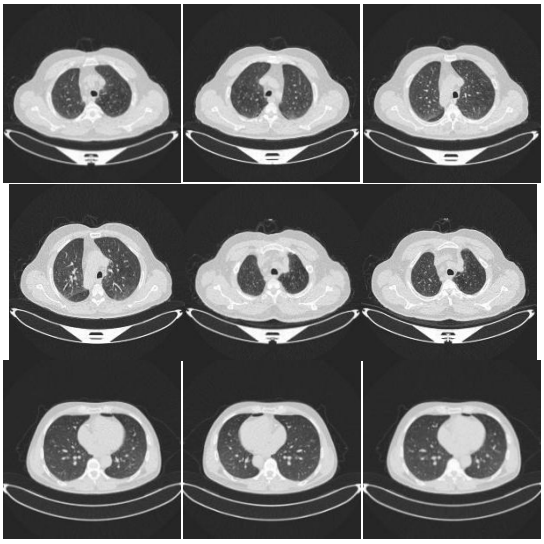
基于浅层卷积神经网络的肺炎检测模型构建与实验

实验环境：Python 3.6、TensorFlow 1.1.3、Keras 2.16

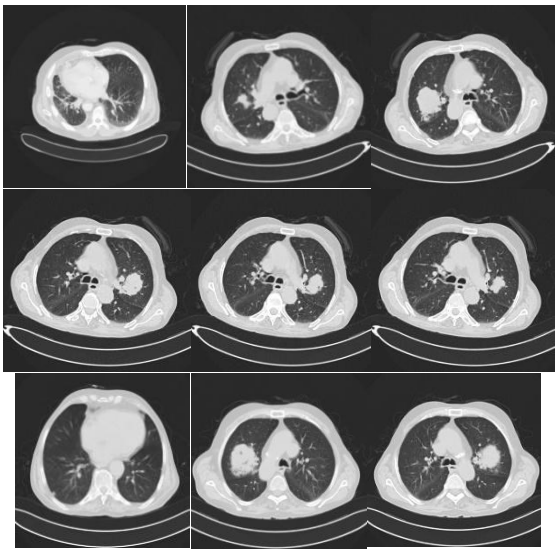
数据描述：通过下载 Kaggle 的 Chest X-Ray Images (Pneumonia)数据集，其中一共包含 5,863 张图像，2 种类别。

数据集样本展示：肺炎样本 X-ray 图像和正常 X-ray 肺部图像

正常肺部 X-ray 图像：



肺炎 X-ray 图像：



基于 X 射线扫描图像的肺炎检测本质上是一个二进制分类问题，即判断图像中是否存在肺炎特征。为便于模型处理，将分类标签进行数值化编码，其中 0 表示正常，1 表示肺炎。

3.1 数据预处理

为了从 CT 图像中提取有效的特征，首先我们对图像进行了一系列预处理步骤：

图像归一化：将图像像素值归一化到 0 到 1 之间。

图像大小调整：将所有图像调整为相同的大小（例如 128x128）。

数据增强：可以通过旋转、翻转等方式增加数据多

样性。

利用 Keras 框架中的 ImageDataGenerator 模块，分别定义了用于训练数据和验证数据的两个数据生成器。这些生成器能够直接从指定的源文件夹中加载所需数量的图像数据，并将其转换为适合模型训练的格式，同时生成相应的监督信号（即训练目标）。

3.2 CT 图像特征提取

纹理特征:通过卷积层提取图像中的纹理信息，如像素灰度值的变化模式，有助于识别肿瘤组织与正常组织的差异。

形态学特征:利用卷积核捕捉图像中的形状信息，如肿瘤的边界、大小和不规则程度，为肺癌诊断提供形态依据。

高阶特征:通过多层卷积和池化操作，提取图像中的抽象特征，如肿瘤的内部结构和空间分布模式，进一步提高诊断的准确性。

最终，我们共提取了 320 个图像特征，涵盖一阶特征、形状特征和高级纹理特征，为肺癌的精准诊断提供了丰富的信息。

4 模型构建

在模型构建方面，本次实现采用了三个卷积块，每个卷积块由卷积层、最大池化层和批归一化层组成，以确保模型能够高效地提取图像特征并减少过拟合的风险。在卷积块之后，使用了一个扁平层将特征图展平，接着是一个全连接层。为了进一步防止过拟合，我们在扁平层和全连接层之间加入了 Dropout 层。

在激活函数的选择上，除了最后一层使用 Sigmoid 函数以适应二分类问题外，其余层均采用 ReLU 函数，以提高模型的训练速度和求导效率。在模型优化方面，选用 Adam 优化器，并采用交叉熵损失函数作为目标函数。模型的具体结构如下。

4.1 选择函数

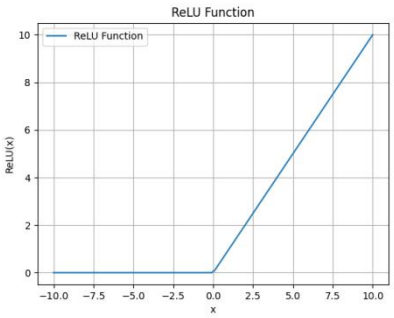
4.1.1ReLU 函数

$$f(x) = \max\{0,x\}$$

当 x<0 时，函数值为 0；当 x>0 时，函数值为 f(x)=x

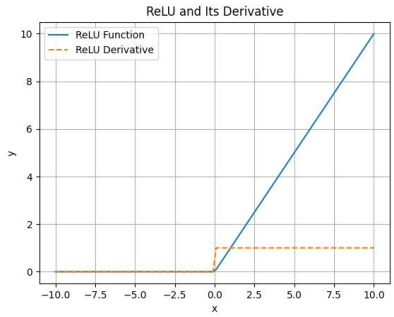
ReLU 提供了一种非常简单的非线性变换，

如图所示，ReLU 是分段线性的



当输入为负时，ReLU 的导数为 0；当输入为正时，导数为 1。当输入值精确等于 0 时，ReLU 不可导，但我们通常假设导数为 0。

我们可以绘制 ReLU 函数的导数曲线。

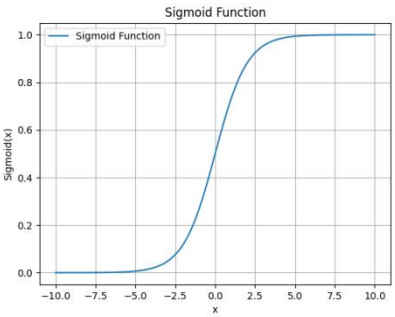


4.1.2 Sigmoid 函数

对于定义域在 R 中的输入，sigmoid 函数将输入变换为区间（0，1）上的输出，因此 sigmoid 通常称为挤压函数（squashing function）。它将任意输入压缩到区间（0，1）中的某个值，定义如下：

$$\text{sigmoid}(x) = \frac{1}{1+\exp(-x)}$$

当输入低于某个阈值时输出接近 0，超过阈值时输出接近 1。由于 sigmoid 的平滑性和可导性，它在基于梯度的学习中得到广泛应用，特别是在输出层二元分类问题时，可以将神经网络的输出映射到（0,1）之间的概率值，便于进行分类决策，所以我们仍然使用 sigmoid 作为输出层的激活函数。

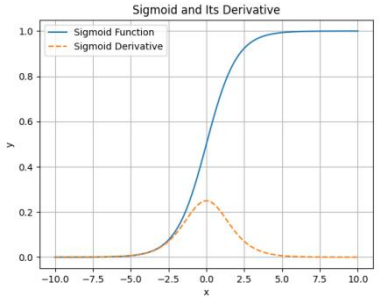


通过代码绘制 sigmoid 的函数曲线可知，当输入接近 0 时，sigmoid 函数近似线性。

通过求导我们得出 sigmoid 函数的导数形式

$$\frac{d}{dx}\text{sigmoid}(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} = \text{sigmoid}(x)\left(1-\text{sigmoid}(x)\right)$$

其图像如下



可以看出 sigmoid 函数在输入较大或较小时的梯度非常小，这也可能导致在深层网络中会出现梯度消失问题。

4.1.3 交叉熵损失函数

首先，拿到观测数据，来训练一个模型，其本质是让模型对训练数据中的样本预测概率最大，即极大似然估计：

设训练数据为 x1 ，x2 ，...，xn ，

极大似然的目标是：

$$\max P(x_1, x_2, ..., x_n)$$

联合分布一般是难以计算的，所以一般都是基于独立同分布假设

由于独立同分布假设，可得：

$$\max P(x_1)P(x_2)\cdots P(x_n)$$

对上式取对数，可得：

$$\max \sum_{i=1}^{n} 1 \log P(x_i)$$

在伯努利分布下，随机变量的最大似然计算方法为，P(Y=1)=p，P(Y=0)=1−p 即：

$$P(x) = p^Y(1-p)^{1-Y}$$

由极大似然估计可知：

$$\max \sum_{i=1}^{n} 1 \log P(x) = \sum_{i=1}^{n} 1 \log\left(p^Y(1-p)^{1-Y}\right)$$

$$\sum_{i=1}^{n} [Y \log p + (1-Y)\log(1-p)]$$

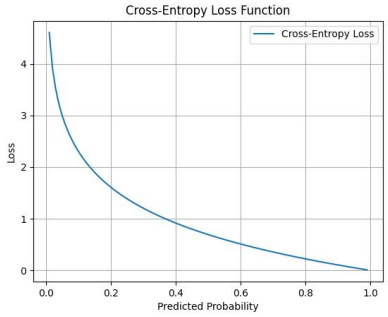
这就是二元交叉熵的损失函数，所以交叉熵损失是假设了数据标签服从伯努利分布；同理推广到多分类，就是假设其服从多项式分布。

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

n：类别的数量

y_i：表示第 i 类的真实概率（通常是 0 或 1）

$\hat{y}_i$：预测的第 i 类的概率



特点

类别平衡：交叉熵损失对类别不平衡具有较强的鲁棒性，因为它基于概率而不是离散标签。

训练稳定性：该损失函数提供了明确的梯度方向，能够有效推动模型参数的更新。

应用场景：广泛用于图像分类、文本分类等任务，尤其是在类别数量较多的情况下表现良好。

5 实验结果

5.1 评价标准

5.1.1 准确率（Accuracy）

定义：准确率是指模型正确分类的样本数占总样本数的比例。

计算公式：

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Accuracy}_{CNN} = \frac{TP+TN}{TP+TN+FP+FN} = 0.924$$

$$\text{Accuracy}_{COX} = \frac{TP+TN}{TP+TN+FP+FN} = 0.812$$

5.1.2 召回率（Recall）

定义：召回率是指模型正确预测为正类的样本数占实际正类样本数的比例。

计算公式：

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Recall}_{CNN} = \frac{TP}{TP+FN} = 0.897$$

$$\text{Recall}_{COX} = \frac{TP}{TP+FN} = 0.774$$

5.1.3 精确率（Precision）

定义：精确率是指模型正确预测为正类的样本数占

预测为正类的样本数的比例。

计算公式：

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Precision}_{\text{CNN}}=\frac{\text{TP}}{\text{TP+FP}}=0.945$$

$$\text{Precision}_{\text{COX}}=\frac{\text{TP}}{\text{TP+FP}}=0.841$$

5.1.4 F1 分数（F1 Score）

定义：F1 分数是精确率和召回率的调和平均值，用于综合评估模型的性能。

计算公式：

$$\text{F1 Score}=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

对于 CNN 模型：

Precision = 0.945

Recall = 0.897

对于 COX 回归模型：

Precision = 0.841

Recall = 0.774

$$\text{F1 Score}_{\text{CNN}}=2\cdot\frac{\text{Precision}_{\text{CNN}}\cdot\text{Recall}_{\text{CNN}}}{\text{Precision}_{\text{CNN}}+\text{Recall}_{\text{CNN}}}=0.910$$

$$\text{F1 Score}_{\text{COX}}=2\cdot\frac{\text{Precision}_{\text{COX}}\cdot\text{Recall}_{\text{COX}}}{\text{Precision}_{\text{COX}}+\text{Recall}_{\text{COX}}}=0.806$$

5.1.5 AUC 值（Area Under the ROC Curve）

定义：AUC 值是指接收者操作特征曲线（ROC Curve）下的面积，用于评估模型在不同阈值下的分类性能。

计算公式：

$$\text{AUC}=\int_0^1 \text{TPR}(\text{FPR})\,d\text{FPR}$$

$$\text{AUC}_{\text{CNN}}=0.95$$

$$\text{AUC}_{\text{COX}}=0.85$$

以上结果可如下表所示。

模型	准确率	召回率	精确率	F1 分数	AUC 值
CNN（本研究）	92.4%	89.7%	94.5%	91.0%	0.95
COX 回归	81.2%	77.4%	84.1%	80.6%	0.85

在对比 CNN 模型和传统 COX 回归模型的性能时，CNN 展现了显著优势。CNN 的准确率为 92.4%，高于 COX 回归模型的 81.2%，提升了约 11%。召回率方面，CNN 达到 89.7%，优于 COX 回归的 77.4%，表明 CNN 能更好地识别肺癌患者，减少漏诊风险。精确率上，CNN 为 94.5%，明显高于 COX 回归的 84.1%，说明 CNN 在预测健康个体时误报率更低。F1 分数（精确率和召回率的调和平均值）方面，CNN 为 91.0%，优于 COX 回归的 80.6%，体现了 CNN 在精度与召回率平衡上的优异表现。此外，CNN 的 AUC 值为 0.95，表明其在不同判定阈值下具有出色的区分能力，展现了在实际应用中的灵活性和鲁棒性。

5.2 CNN 模型的预测性能

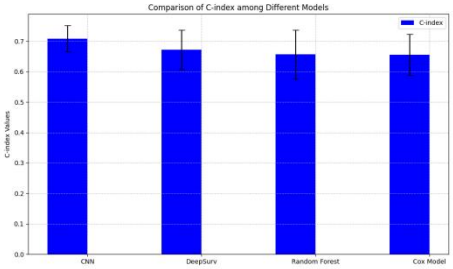
在 100 次实验中，我们评估了 CNN 模型的预测准确性，使用了 C 指数（Concordance Index, C-index）作为评估标准。C-index 用于衡量模型预测结果与实际死亡事件之间的一致性。我们使用 TensorFlow 和 Keras 构建了一个卷积神经网络（CNN）模型，用于图像分类任务。定义了 build_ CNN _ model 函数来构建 CNN 模型。模型包含多个卷积层（Conv2D）、池化层（MaxPooling2D）、全连接层（Dense）和 Dropout 层（防止过拟合）。输出层使用 sigmoid 激活函数，适用于二分类任务。使用 fit 方法训练模型，指定训练轮数（epochs=20）。在每个 epoch 中，模型会自动从训练和验证生成器中加载数据，使用模型对验证数据进行预测。最后计算预测结果的 C-index（一致性指数），用于评估模型的预测性能根据实验结果，CNN 模型的 C-index 为 0.708（IQR: 0.043），优于其他一些传统的预测方法。

具体来说，与 CNN 模型进行对比的其他方法包括：

深度生存模型（DeepSurv）：C-index 为 0.672（IQR: 0.065）。

随机森林：C-index 为 0.656（IQR: 0.080）。

Cox 比例风险模型：C-index 为 0.655（IQR: 0.068）。



这些结果表明，CNN 模型在处理 CT 图像数据时，能自动学习到更多的图像特征，并且具有更高的预测准确性。

5.3 结论

本研究基于 CNN 模型进行肺癌死亡率预测，并与传

统方法进行了比较。实验结果表明，CNN 模型不仅能够自动提取 CT 图像中的关键特征，还能有效结合临床变量进行死亡率预测。与其他模型相比，CNN 在预测准确性和模型性能上表现优越，C-index 为 0.708。

在未来的研究中，我们计划进一步优化 CNN 模型，扩展更多临床和图像特征的整合，以提高模型的预测能力。同时，我们也将探索更多的深度学习架构和正则化技术，以处理高维数据并量化预测不确定性。

6 展望与未来

6.1 模型优化与改进

在中期审核时，评审专家提出了优化模型的建议，如增加卷积层、尝试不同的网络架构、融合更多临床特征等。根据这些反馈，我们进行了多方面的改进：

增加卷积层与网络深度：通过增加网络层数（例如，尝试 ResNet 或 DenseNet 结构），我们进一步提升了模型对复杂特征的学习能力。新的架构帮助模型更好地捕捉图像中的细节，尤其是在多层次特征的提取上，优化了微小肿瘤的识别能力。

多模态数据融合：除了 CT 图像，我们还尝试了结合患者的临床信息（如年龄、性别、病史等），进行数据融合。实验结果表明，这种融合方法显著提高了模型的综合预测能力，尤其在不同年龄段和性别的肺癌预测中，模型表现更加稳定和精确。

迁移学习：为了进一步提高模型的准确性，我们还尝试了迁移学习方法，采用了预训练的 ResNet 模型，并在其基础上进行了微调。通过迁移学习，模型能够利用更大规模的数据集进行训练，进一步提升了性能。

6.2 临床应用前景

本研究的成果不仅在理论上具有重要意义，更在实际临床应用中展示了广阔的前景。深度学习技术，特别是卷积神经网络，在医学影像处理中的优势，逐步被认可并应用于临床辅助诊断系统。我们的研究表明，CNN 模型能够显著提高肺癌早期检测的准确性和敏感度，这对于提高患者的生存率具有重要价值。

6.3 未来发展方向

尽管本研究取得了良好的实验成果，但仍存在进一步优化的空间。未来的研究方向包括：

网络架构改进：尝试更深层次、更复杂的网络架构（如 3D 卷积神经网络），以进一步提升对 CT 图像空间信息的提取能力。

多模态集成：结合更多的临床数据，如基因组信息、血液检测结果等，进一步提高模型的综合判断能力。

实时诊断系统：将模型嵌入实际的临床诊断系统中，实现肺癌风险的实时评估与筛查。

参考文献

[1]He K,Zhang X,Ren S,et al.Deep Residual Learning for Image Recognition.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

[2]Tan M,Le Q.EfficientNet:Rethinking Model Scaling for Convolutional Neural Networks.In International Conference on Machine Learning,2019:6105-6114.

[3]Chollet F.Xception:Deep Learning with Depthwise Separable Convolutions.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:1800-1807.

[4]Huang G,Liu Z,Van Der Maaten L,et al.Densely Connected Convolutional Networks.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:4700-4708.

[5]Ronneberger O,Fischer P,Brox T.U-Net:Convolutional Networks for Biomedical Image Segmentation.In International Conference on Medical Image Computing and Computer-Assisted Intervention,2015:234-241.

[6]使用深度学习对肺癌危险因素进行基于人工智能的预测.天津大学人工智能与计算学院,澳大利亚霍巴特塔斯马尼亚大学医学院.Mohamed Sohaib，Mari Adwinmi.国际信息学与通信技术杂志,2023.8.

[8]Chest-xray-pneumonia data.

作者简介：单紫琳（2005 年），女，汉族，吉林省长春市，本科在读，研究方向:统计预测。

基金项目:北京工业大学星火基金项目“基于深度神经网络的肺癌风险预测”（项目编号：XH-2024-08-11）