DOI:10.3969/j.issn.1005-202X.2024.12.017

医学人工智能

基于人工智能的多模态影像组学特征挖掘及分析软件设计

陈业，李翰威，胡德斌，齐宏亮，陈宏文
南方医科大学南方医院医学工程科，广东 广州 510515

【摘要】针对影像组学研究需要使用多款软件，常存在数据不兼容、算法参数无法调节等问题，开发一款基于人工智能的影像组学分析建模软件，为医生和科研工作者提供支持图像预处理、特征提取、特征筛选、建模分析和数据可视化的影像组学集中解决方案。使用一个公开数据集对软件功能进行测试，创建8组分析模型，一一完成对测试集数据的分类预测并输出关键性能指标，通过参数调优，使得模型性能进一步优化，验证软件的可用性。该软件的使用可使研究人员更多聚焦课题本身，减少不必要的开发负担，对于影像组学研究朝着更加方便、高效的方向发展具有积极的推动作用。

【关键词】影像组学；人工智能；软件设计；二分类
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2024)12-1578-07

Design of a software for multimodal radiomics features mining and analysis based on artificial intelligence

CHEN Ye, LI Hanwei, HU Debin, QI Hongliang, CHEN Hongwen
Department of Medical Engineering, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China

Abstract: Various types of software needed in radiomics studies come with the problems such as data incompatibility and hyperparameter tuning. Therefore, an artificial intelligence-based software is developed for radiomics studies, providing doctors and researchers a solution with image preprocessing, feature extraction, feature selection, modeling analysis and data visualization. The usability of the software is demonstrated using a public data set. Eight sets of feature selectors and classifiers are established for classification predication on test data set and key performance indicator output. Through hyperparameter tuning, the model is further optimized. Researchers will focus more on the research itself rather than unnecessary development efforts, and radiomics studies will become more convenient and efficient with the software addressed.

Keywords: radiomics; artificial intelligence; software design; binary classification

前 言

医学成像能够在治疗前、治疗中、治疗后无创地可视化肿瘤的放射学表型[1-2]。目前，临床上的影像分析局限于医师对图像的主观判断，仅对CT密度、PET/CT的标准摄取值(Standard Uptake Value, SUV)以及图像灰度值进行简单统计，可能会导致疾病的漏诊和误诊，且简单的视觉分析无法捕捉到病灶更深层次的信息，无法满足精准医疗和个体化治疗的要求[3-4]。影像组学是人工智能与医学影像大数据结合的新技术，通过从医学图像的感兴趣区域(Region of Interest, ROI)中挖掘定量特征，利用机器学习将其与临床信息(分型、疗效和预后等)进行关联，从而揭示与病灶关联的深层信息，是目前影像学领域最受关注的研究前沿之一[5-7]。影像组学的研究流程包括以下几个步骤：(1)医学影像数据的获取；(2)ROI勾画；(3)高通量特征的提取与归一化；(4)特征筛选，目的是保留与疾病诊断相关性最高的特征，避免过拟合[8]；(5)对提取的特征进行建模；(6)利用ROC曲线下面积AUC评估模型的性能[9]。针对上述步骤，目前已经有研究人员开发出相应的软件，比如3DSlicer、ITK-SNAP[10]等用于ROI勾画。CGITA(Chang Gung Image Texture Analysis)可用于分子影像、CT、MRI纹理数据分析[11]。IBEX(Imaging Biomarker Explorer)能够构建特征集，方便用户及逆

【收稿日期】2024-07-23
【基金项目】国家重点研发计划(2023YFC2414601)；南方医科大学南方医院院长基金(2022B030, 2022B016)
【作者简介】陈业，硕士，研究方向：图像处理及软件开发，E-mail: 15915724843@163.com

行特征提取，但不能显示三维结果[12]。PyRadiomics[13]是一个开源Python包，是一个经过测试和维护的特征提取平台。目前使用较多的影像组学建模分析工具有SPSS Modeler、FAE（FeAture Explorer）[14]等。这些软件通常需要手动为提取的特征列表添加标签列，需要根据标签数值分开提取特征，然后通过合并矩阵得到完整的特征列表，而且对特征提取以及建模算法也没有提供灵活的参数设置方法，是影响影像组学特征提取和建模结果可重复性和通用性的不可忽视的因素[15]。另外，科研人员在影像组学研究的不同环节往往需要使用不同的软件，数据兼容性问题也会降低研究效率。

本研究旨在开发一款影像组学自动分析软件，兼容主流的医学影像格式，能自动完成数据分集、数据预处理与归一化、特征提取、特征筛选与降维、建模及性能评估，并提供图形化界面用于算法调参。

1 影像组学自动分析软件设计

传统的医学影像分析方法通常需要大量的人力和时间成本，并且容易受到主观因素的影响。而基于人工智能的软件可以自动化地提取特征并进行分析，大大提高分析的效率和准确性[16-17]。多模态影像组学的方法可以帮助医学研究人员深入了解疾病的发病机制和进展过程，从而为新药研发、疾病预防和治疗提供科学依据[18-19]。本研究设计的多模态影像组学特征挖掘及分析软件主要包括数据准备、特征提取与筛选、自动建模与分类预测、数据可视化4个模块，软件的技术路线如图1所示。读取医学影像数据及其ROI文件，设置特征提取参数，包括图像滤波器、特征类型等，提取特征。导入提取的特征列表并进行归一化处理，选择合适的特征筛选算法去除冗余特征，然后选择合适的算法构建分类模型，最后给出模型的性能评价指标。

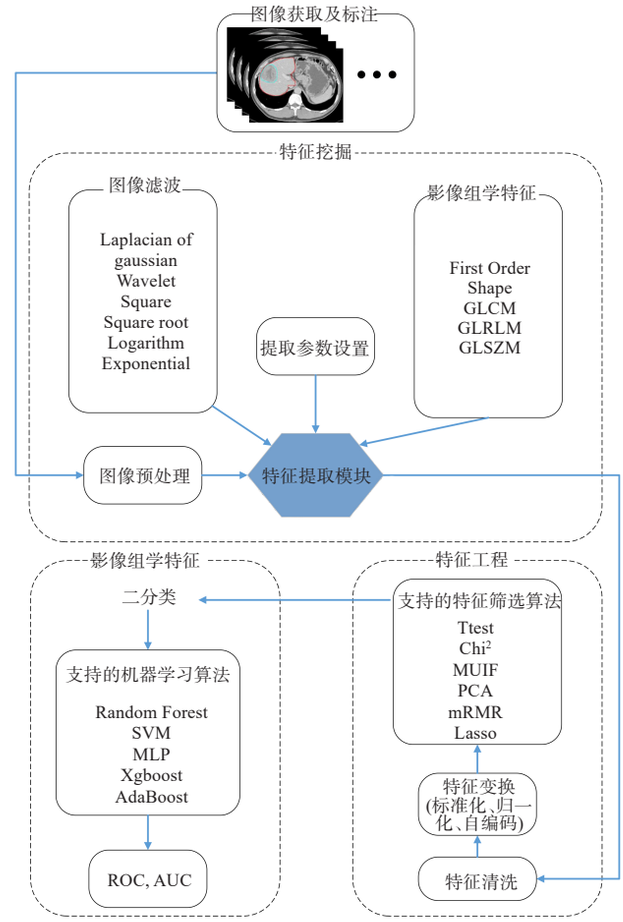


图1 软件技术路线图

Figure 1 Software roadmap

1.1 数据准备

本研究设计的影像组学分析软件能够处理不同类型的影像数据，包括但不限于MRI、CT、PET等图像数据。这些影像数据可能来自于不同的设备、不同的成像模态，具有不同的格式、分辨率和特征。支持目前主流的医学影像格式，如DICOM（.dcm）、NEITY（.nii或者.nii.gz）等。用户首先需要将医学影像文件和对应的mask文件（即勾画出的ROI文件，支持.nii或者.nrrd格式），按照规定的文件目录结构保存，目录结构如图2所示，其中P代表标签为1的病例的目录，N代表标签为0的病例的目录。软件可以根据设定的比例，将数据分为训练集和测试集。数据标准化是很多机器学习分类算法的共同需求：消除

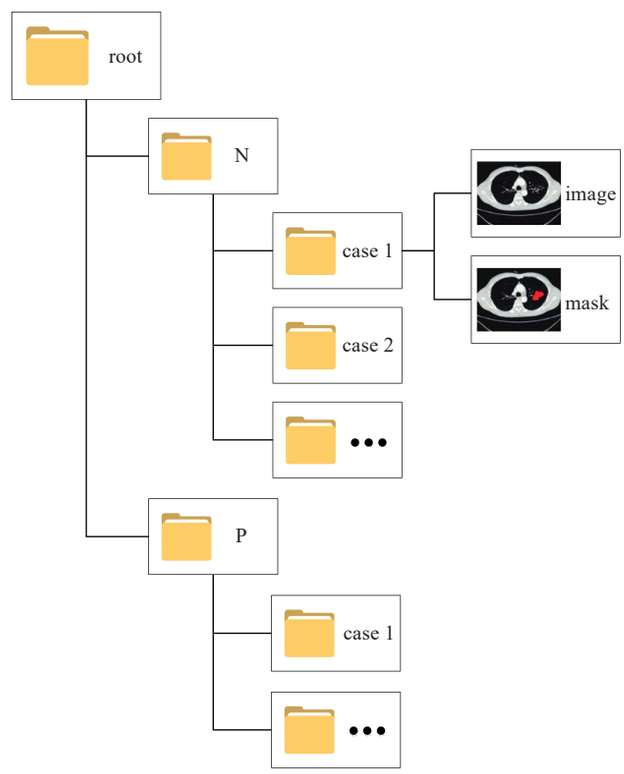


图2 文件目录结构

Figure 2 Data directory structure

特征之间不同量纲对结果造成的影响，同时使数据符合标准正态分布。本软件使用的是Z-Score标准化方法，用于评估样本点到总体均值的距离。

1.2 特征提取与筛选

本研究使用PyRadiomics实现影像组学特征提取，PyRadiomics是基于图像生物标志物标准化倡议的开源的医学影像特征提取工具，也可以应用于经过信号转换处理的图像，比如小波变换、对数（Log）变换等。用户需要设置好数据文件所在目录以及需要提取的特征类型，点击“开始提取”就可以进行特征提取。图3为提取的特征列表的一部分，其中每一行代表一个病例，每一列代表一个特征值，表头部分第一列是各个病例的序号，第二列标记病例的label，即阴性或阳性，表头其余部分包含特征的名字，本软件可以根据文件目录自动为特征列表插入label列。由于提取的特征原始表格包含很多字符串格式或者空字符的特征值，这些值是无法进行后续特征选择和建模分析的，软件会自动识别并剔除这些值。获取高通量的特征后，大量冗余的特征容易造成影像组学分析预测模型过拟合，从而达不到最真实的结果，因此需要对特征进行筛选以获得本质特征[20]。本软件内置多种特征选择和降维方法，如Ttest、卡方检测、互信息（Mutual Information, MUIF）、主成分分析（Principal Component Analysis, PCA）、最大相关最小冗余、Lasso。每种算法都提供了用户界面用于参数设置。

index	label	diagnostics_Image-original_Maximum	diagnostics_Mask-original_VoxelNum
BraTS19_2013_0_1	0	990	21279
BraTS19_2013_15_1	0	1701	4487
BraTS19_2013_16_1	0	1300	5522
BraTS19_2013_1_1	0	207	9329
BraTS19_2013_24_1	0	153	43406
BraTS19_2013_28_1	1	220	50041
BraTS19_2013_29_1	0	301	28736
BraTS19_2013_6_1	1	993	91102
BraTS19_2013_8_1	1	1549	16887
BraTS19_2013_9_1	0	260	20173
BraTS19_TCIA09_141_1	0	2448	93298
BraTS19_TCIA09_177_1	1	3462	10557
BraTS19_TCIA09_254_1	1	360	77985
BraTS19_TCIA09_255_1	0	391	110787
BraTS19_TCIA09_312_1	0	2393	1501

样本名称/编号　标签　特征名称　特征值

图3 软件提取的特征列表格式

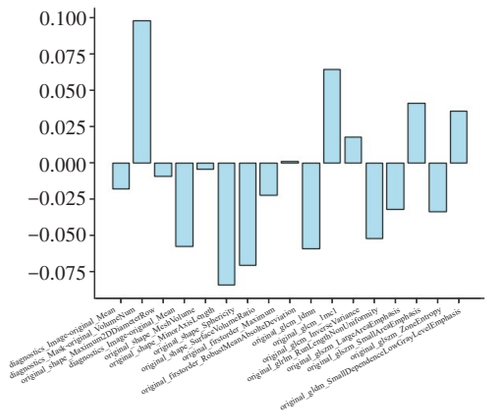
Figure 3 Format of feature matrix extracted by the software

1.3 自动建模与分类预测

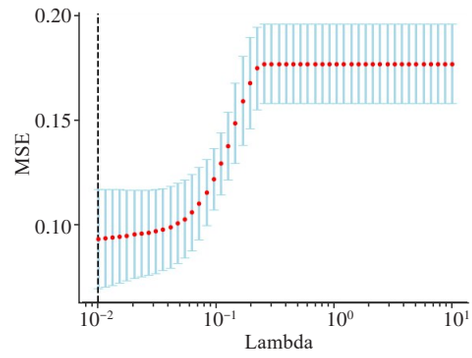
大多数影像组学研究属于二分类问题[21]。通过机器学习算法构建一个能够对新数据进行分类预测的模型，并且通过AUC来评估分类模型的性能，AUC越接近1说明结果越好[22]，使用训练好的模型对新的影像数据进行分类预测，不同模型算法的选择将对分类性能产生影响[23-24]。软件支持常用机器学习算法集成，包括支持向量机（Support Vector Machine, SVM）、XGboost（eXtreme Gradient Boosting）、随机森林、多层感知器（Multilayer Perceptron, MLP）等。由于机器学习算法包含的参数较多，不同参数组合适用的场景不同，导致分类结果也不同，与特征筛选一样，本软件为每种机器学习算法都设置对应的参数设置入口，方便研究人员根据实际情况调整参数，优化分类结果。通过开源方式支持自定义机器学习算法集成。

1.4 数据可视化

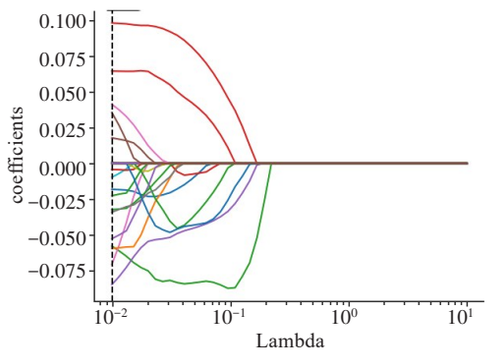
软件设置实时信息展示窗口，可以在特征提取和建模分析过程中实时打印输出过程以及报错信息，方便研究人员跟进流程进度。通过点击“显示特征”按钮可以显示当前提取出来的特征矩阵。点击“保存特征”可以将提取的特征，以csv文件格式进行保存，以便导入本软件的自动建模与分类预测模块进行建模分析，也可以导入其他分析软件进行特征分析。为了评估模型性能，软件在每次建模分析完成后在信息展示窗口打印输出模型的常用评估指标，包括准确率ACC（预测正确的样本占总样本的比例）、精确率P（预测正确的样本占正例样本的比例）、敏感性SEN（在实际为正例的样本中，被预测为正例的占比）、特异性SPE（在实际为负例的样本中，被预测为负例的占比）、阴性预测值NPV（检测为阴性的人群中确实为阴性的占比）等，软件提供多种常见的图表绘制，包括特征权重图、ROC图（带AUC值）、拟合轨迹图、特征热点图等，如图4所示。



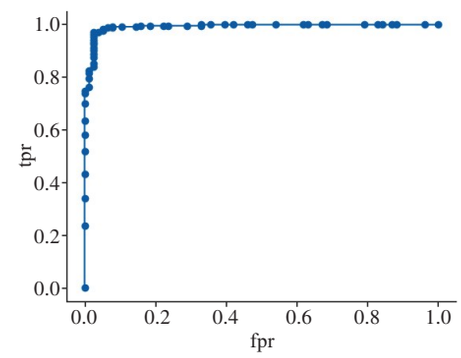
a：特征权重图



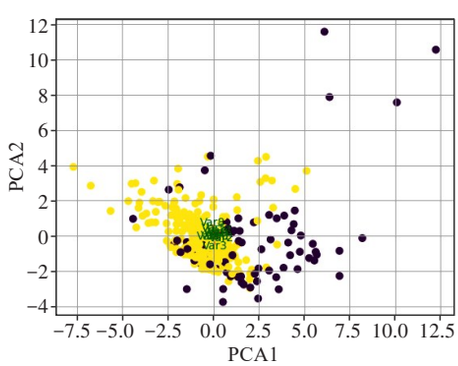
b：Lasso回归拟合均方误差图



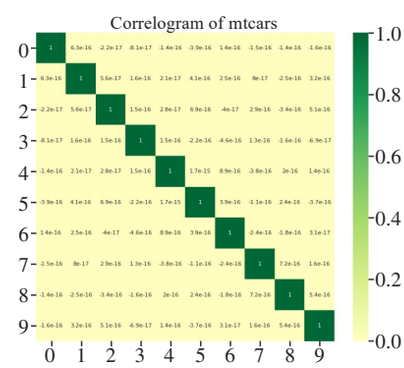
c：拟合系数轨迹图



d：ROC曲线图



e：PCA散点图



f：特征热点图

图4 可视化输出图像

Figure 4 Result visualization

软件采用Python 3.8.4和PySimpleGUI，基于Windows平台开发，对硬件要求较低，双核CPU，2 GB内存，256 GB硬盘的电脑即可运行该软件。在数据处理中，特征选择与降维以及建模分析中使用的算法大部分由Sklearn 1.3.0和Scipy 1.10.1库实现，并且使用XGboost 1.7.6库提供XGboost模型实现。通过开源的方式（https://github. com/DarthLeehom/RadiomicsTool/tree/master），研究者可以增加自定义的特征选择算法或者分类器模型，验证算法的有效性，软件界面如图5所示。

2 影像组学自动分析软件功能测试

为了得到更有价值的分析结果，应尽量选择并使用同种扫描设备去完成影像的采集[25]。本研究使用的测试数据来自多模态脑肿瘤挑战2019数据集（BraTS19）[26]，包含335例病人的MRI图像，高、低级别胶质瘤患者分别为259例和76例，每例病例包含4种模态的MRI序列（T_1、T_2、T_1CE、FLAIR）以及一个标签文件，图像和标签文件的格式均为nii.gz。测试电脑的配置为6核12线程CPU，频率3.00 GHz，8 GB内存，1 TB硬盘。提取的特征包括形态特征、一阶统计

图5 软件主界面

Figure 5 Software main interface

量、灰度共生矩阵、灰度行程矩阵、灰度区域大小矩阵、灰度依赖矩阵和邻域灰度差矩阵，使用T_1CE模态进行特征提取，最终从每例病例的原生图像中提取129个特征。特征筛选算法使用Ttest、PCA、MUIF和Lasso，其中MUIF设置保留特征20个。建模分析方面，设置训练集和测试集大小的比例为3:1，算法选择SVM或MLP。根据采用的特征筛选和建模分析算法不同，可以得到不同的分析模型，表1列举其中8组模型及其各自的评价指标。每个模型的ROC曲线如图6所示。MUIF+Lasso+SVM最终筛选得到的特征数量最少，但是模型评价指标并不高。Ttest+Lasso+SVM、Ttest+Lasso+MLP和MUIF+Lasso+SVM 3个模型的AUC值最高，其中Ttest+Lasso+SVM在模型评价指标方面表现最好。

本研究进一步对Ttest+Lasso+SVM模型中的Lasso进行参数调优，采用10折交叉验证，具体参数设置如图7所示。采用设定参数的模型运行后最终

表1 模型测试集的性能评价指标

Table 1 Performance evaluation indexes of each model

模型名称	特征提取数量	筛选后特征数量	模型评价指标				
			准确率	精确率	敏感性	特异性	阴性预测值
Ttest+Lasso+SVM	129	18	0.91	0.93	0.96	0.73	0.84
Ttest+Lasso+MLP	129	23	0.89	0.91	0.96	0.58	0.76
PCA+Lasso+SVM	129	30	0.88	0.89	0.96	0.64	0.84
PCA+Lasso+MLP	129	20	0.90	0.92	0.96	0.63	0.80
MUIF+Lasso+SVM	129	11	0.83	0.84	0.96	0.46	0.80
MUIF+Lasso+MLP	129	12	0.88	0.92	0.92	0.73	0.73
MUIF+PCA+SVM	129	20	0.85	0.84	0.99	0.52	0.94
MUIF+PCA+MLP	129	20	0.90	0.89	0.99	0.67	0.95

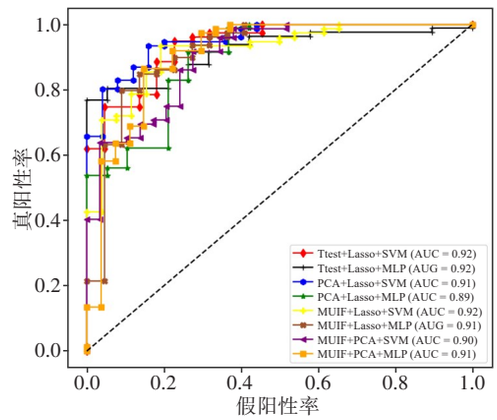


图6 全部模型测试集的ROC曲线

Figure 6 ROC curves of all models on the test set

筛选出6个特征，包括连通域mask个数、2个形态特征、2个灰度共生矩阵、1个灰度依赖矩阵，各个特征的权重如图8所示。特征数量随Lasso的参数Alpha的变化如图9所示。模型的AUC进一步提升，如图10所示。

3 讨 论

本研究开发一套基于人工智能的影像学分析软件，应用强大的计算机图像处理能力和多元的大数据挖掘方法，结合计算机和生物医学工程领域相关知识，为肿瘤辅助诊断、疗效评估和预后预测临床研究等提供实现平台。本研究开发的影像组学分析软

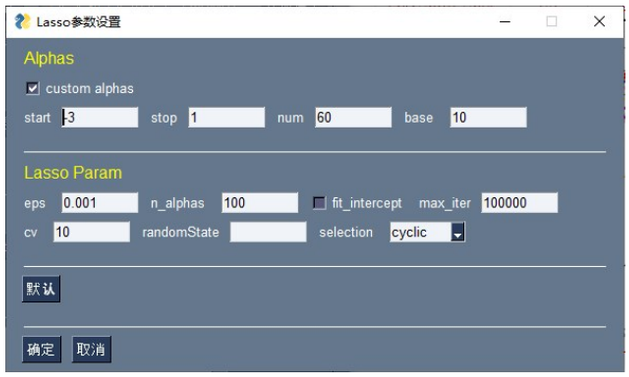


图7 Lasso参数设置

Figure 7 Lasso parameter setting

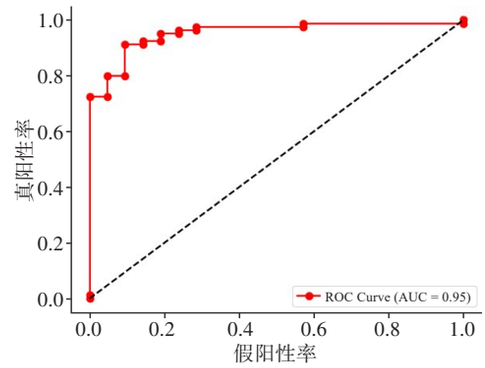


图10 Ttest + Lasso + SVM测试集的ROC曲线

Figure 10 ROC of Ttest + Lasso + SVM on the test set

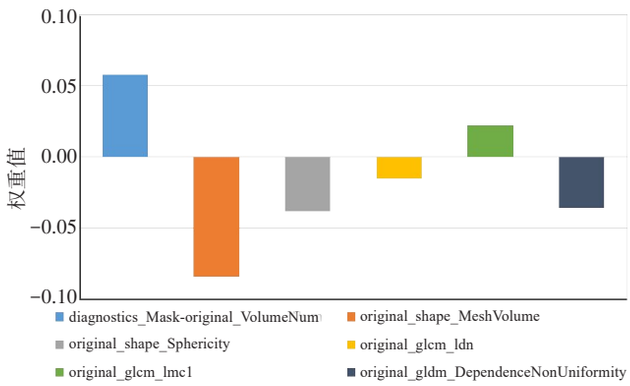


图8 Ttest + Lasso筛选后的特征权重

Figure 8 Feature weight after feature selection using Ttest and Lasso

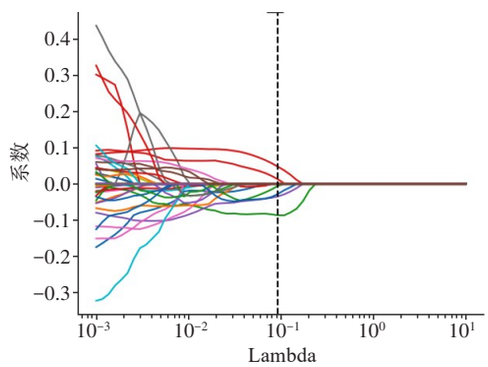


图9 特征数量随Alpha变化

Figure 9 Number of features varies with Alpha

件支持多种格式的医学影像特征提取，自动插入label，内置经典的特征选择和降维算法以及机器学习算法模型，方便研究者根据实际需求选择不同的分析模型组合。该软件提供丰富的模型评价指标用于衡量模型性能，并且支持输出常用图表，方便研究者进行数据可视化。此外，针对特征筛选算法和机器学习算法的超参数调优问题，该软件为每个算法提供用户界面用于参数设置。为了验证软件的可用性，本研究使用BraTs19数据集进行软件功能测试，验证影像组学挖掘及建模分析全流程的功能，降低用户的学习成本，提高科研效率。基于该软件开源平台，研究者可以实际使用场景创建自定义算法进行研究。

该软件还存在以下局限性：(1)对于影像组学分析全流程而言，该软件没有集成ROI勾画功能，ROI勾画需要实时交互，当涉及到半自动或自动分割算法时，计算密集度往往较高，对软件执行效率要求更高，更适合C或者C++开发[14]；(2)该软件目前可以方便地进行二分类问题研究，但是没有支持多标签的分类和回归问题研究。这些将在未来的软件版本迭代逐步完善。

4 结 论

针对医学影像数据分析的需求，本研究设计一款基于人工智能的多模态影像组学特征提取及分析软件，研究者可以使用该软件完成影像数据的特征提取、筛选以及分析建模全流程，支持算法调参、模型评价和数据可视化等功能。该软件对于医学影像领域的诊断、治疗和研究从传统的人工分析模式转向智能化的数据驱动模式，具有重要的意义。

【参考文献】

[1] 王旭. 功能医学成像技术在肾癌诊断中的应用现状[J]. 中国医疗器械信息, 2017, 23(21): 35-36.
Wang X. Application of functional medical imaging in the diagnosis of renal cell[J]. China Medical Device Information, 2017, 23(21): 35-36.

[2] Xu MZ, Chen ZY, Zheng JX, et al. Artificial intelligence-aided optical imaging for cancer theranostics[J]. Semin Cancer Biol, 2023, 94: 62-80.

[3] 陈武凡. 数字化医学成像研究进展与未来趋势[J]. 中国基础科学, 2014, 16(5): 21-28.
Chen WF. Progress and trends on digital medical imaging[J]. China Basic Science, 2014, 16(5): 21-28.

[4] Zhang ST, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis[J]. Med Image

Anal, 2024, 91: 102996.
[5] Kumar V, Gu YH, Basu S, et al. Radiomics: the process and the challenges[J]. Magn Reson Imaging, 2012, 30(9): 1234-1248.
[6] Limkin EJ, Sun R, Dercle L, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology[J]. Ann Oncol, 2017, 28(6): 1191-1206.
[7] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data[J]. Radiology, 2016, 278(2): 563-577.
[8] Kira K, Rendell LA. A practical approach to feature selection[M]// Sleeman D, Edwards P. Machine Learning Proceedings 1992. San Francisco, CA, USA: Morgan Kaufmann, 1992: 249-256.
[9] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.
[10] Kickingereder P, Burth S, Wick A, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models[J]. Radiology, 2016, 280(3): 880-889.
[11] Fang YH, Lin CY, Shih MJ, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images[J]. Biomed Res Int, 2014, 2014: 248505.
[12] Zhang LF, Fried DV, Fave XJ, et al. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics[J]. Med Phys, 2015, 42(3): 1341-1353.
[13] van Griethuysen JJ, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype[J]. Cancer Res, 2017, 77(21): e104-e107.
[14] Song Y, Zhang J, Zhang YD, et al. FeAture explorer (FAE): a tool for developing and comparing radiomics models[J]. PLoS One, 2020, 15 (8): e0237587.
[15] Bettinelli A, Marturano F, Avanzo M, et al. A novel benchmarking approach to assess the agreement among radiomic tools[J]. Radiology, 2022, 303(2): E30.
[16] 蒋西然, 蒋韬, 孙嘉瑶, 等. 深度学习人工智能技术在医学影像辅助分析中的应用[J]. 中国医疗设备, 2021, 36(6): 164-171.
Jiang XR, Jiang T, Sun JY, et al. Deep learning in computer aided analyses of medical images[J]. China Medical Devices, 2021, 36(6): 164-171.
[17] Zhao D, Wang W, Tang T, et al. Current progress in artificial intelligence-assisted medical image analysis for chronic kidney disease: a literature review[J]. Comput Struct Biotechnol J, 2023, 21: 3315-3326.
[18] 姚丽红, 朱丽红. 多模态影像组学检查在子宫颈癌放化疗中的临床应用及研究进展[J]. 中国妇产科临床杂志, 2023, 24(4): 431-433.
Yao LH, Zhu LH. Clinical application and research progress of multimodal radiomics examination in radiotherapy and chemotherapy for cervical cancer[J]. Chinese Journal of Clinical Obstetrics and Gynecology, 2023, 24(4): 431-433.
[19] Tang W, Zhang Y, Yu XL, et al. Diagnostic value of the dual-modal imaging radiomics model for subpleural pulmonary lesions[J]. Eur J Radiol, 2023, 166: 111000.
[20] 杨凤, 杨明. 影像组学运用中的常用工具与方法[J]. 医学影像学杂志, 2023, 33(5): 882-885.
Yang F, Yang M. Common tools and methods in the application of radiomics[J]. Journal of Medical Imaging, 2023, 33(5): 882-885.
[21] James G, Witten D, Hastie T, et al. An introduction to statistical learning: with applications in R[M]. New York, NY: Springer US, 2021.
[22] 刘鹏, 王丽嘉, 马超. 影像组学及分析工具浅谈[J]. 生物医学工程与临床, 2022, 26(4): 511-518.
Liu P, Wang LJ, Ma C. Introduction of radiomics and analytical tools [J]. Biomedical Engineering and Clinical Medicine, 2022, 26(4): 511-518.
[23] Lambin P, Leijenaar RT, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine[J]. Nat Rev Clin Oncol, 2017, 14(12): 749-762.
[24] Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis [J]. Eur J Cancer, 2012, 48(4): 441-446.
[25] 刘静宇, 刘颖, 张帆. 影像组学的临床应用研究及挑战[J]. 山西大同大学学报(自然科学版), 2022, 38(4): 67-72.
Liu JY, Liu Y, Zhang F. Clinical application research and challenge of radiomics[J]. Journal of Shanxi Datong University (Natural Science Edition), 2022, 38(4): 67-72.
[26] Center for Biomedical Image Computing & Analytics. Multimodal brain tumor segmentation challenge 2019: data[EB/OL]. https://www.med.upenn.edu/cbica/brats2019/data.html.

（编辑：陈丽霞）