

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 9 月 15 日 (15.09.2022)



(10) 国际公布号
WO 2022/188490 A1

- (51) 国际专利分类号:
G06T 7/00 (2017.01) *G16B 20/20* (2019.01)
G06T 7/11 (2017.01) *G06N 3/04* (2006.01)
G06K 9/62 (2022.01) *G06N 3/08* (2006.01)
G16B 40/00 (2019.01)
- (21) 国际申请号: PCT/CN2021/137310
- (22) 国际申请日: 2021 年 12 月 12 日 (12.12.2021)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202110264535.X 2021 年 3 月 11 日 (11.03.2021) CN
- (71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY

OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

- (72) 发明人: 张圣海 (ZHANG, Shenghai); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。李志成 (LI, Zhicheng); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。赵源深 (ZHAO, Yuanshen); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。孙秋畅 (SUN, Qiuchang); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。梁栋 (LIANG, Dong); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(54) Title: SURVIVAL TIME PREDICTION METHOD AND SYSTEM BASED ON IMAGING GENOMICS

(54) 发明名称: 一种基于影像基因组学的生存预测方法和系统

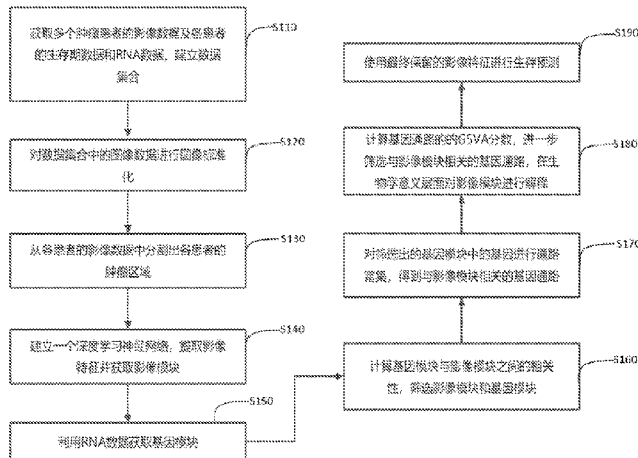


图 1

- S110 Obtain image data of a plurality of tumor patients and survival time data and RNA data of the patients, to establish a data set
- S120 Perform image normalization on the image data in the data set
- S130 Separate tumor regions of the patients from the image data of the patients
- S140 Establish a deep learning neural network to extract image features and obtain image modules
- S150 Use the RNA data to obtain gene modules
- S160 Calculate correlations between the gene modules and the image modules to screen the image modules and gene modules
- S170 Perform pathway enrichment on genes in the gene modules obtained after screening, to obtain gene pathways related to the image modules
- S180 Calculate GSVA scores of the gene pathways to further screen out a gene pathway related to the image modules, and interpret the image modules biologically
- S190 Use image features finally retained to perform survival time prediction

(57) Abstract: Disclosed in the present invention is a survival time prediction method and system based on imaging genomics. The method comprises: obtaining image data of tumor patients and survival time data and RNA data of the patients, to establish a data set; separating tumor regions of the patients from the image data; inputting the image data of the patients into a neural network to extract image features and cluster same to obtain a plurality of image modules; using the RNA data to obtain gene modules of the patients; performing screening according to correlations between the gene modules and the image modules to select a plurality of strongly correlated gene modules and image modules; performing pathway enrichment on genes in the selected gene modules to obtain gene pathways related to the image modules; calculating gene set variation analysis scores of the gene pathways, and retaining a gene pathway

(74) 代理人: 北京中巡通大知识产权代理有限公司
(BEIJING ZHONG XUN TONG DA INTELLECTUAL
PROPERTY AGENCY CO., LTD.); 中国北京市
西城区榆树馆胡同东口1层106, Beijing
100195 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家
保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,
BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

strongly correlated to the image modules; and using retained image features to perform survival time prediction. The present invention can improv the biological interpretability of survival time prediction, and also improves the generalization capability of deep learning.

(57) 摘要: 本发明公开了一种基于影像基因组学的生存预测方法和系统。该方法包括: 获取肿瘤患者的影像数据及各患者的生存期数据和RNA数据, 建立数据集合; 从影像数据中分割出各患者的肿瘤区域; 将各患者的影像数据输入神经网络, 以提取影像特征并进行聚类, 得到多个影像模块; 利用RNA数据获取各患者的基因模块; 根据基因模块与影像模块的相关性进行筛选, 选出多个强相关的基因模块和影像模块; 对所选出的基因模块中的基因进行通路富集, 得到与影像模块相关的基因通路; 计算基因通路的基因集变异分析分数, 保留与影像模块具有强相关的基因通路; 使用保留的影像特征进行生存预测。本发明可提高生存预测方面的生物学可解释性、并同时提高深度学习的泛化能力。

一种基于影像基因组学的生存预测方法和系统

技术领域

[0001] 本发明涉及医学图像处理技术领域，更具体地，涉及一种基于影像基因组学的生存预测方法和系统。

背景技术

[0002] 肿瘤的发病率和死亡率很高，已经成为人类因疾病死亡的主要原因。对于肿瘤病人的生存预测分析可以给临床医生提供必要的预后信息，帮助临床医生建立明确的治疗方案，提高肿瘤的治愈率，同时有效降低患者的负担，改善患者预后的生活质量。美国癌症联合会推出的TNM分期系统在肿瘤临床实践中得到了广泛应用，是指导预后的重要工具。但是，许多研究发现TNM不能对不同肿瘤特征的患者的生存结局差异进行有效的区分。

[0003] 影像基因组学主要用于研究医学成像表型和肿瘤基因组之间的潜在联系。在预后预测方面，利用深度学习的方法，通过影像学标记结合肿瘤基因组，可以预测疾病状态和预后情况，进而无创地评估肿瘤的生物学行为，对肿瘤个体化治疗起到重要作用。目前已有一些研究探索将基因表达数据与分子信息和临床关联起来，但影像基因组学还很少被应用到肿瘤生存预测方面。

[0004] 现在的单纯基于深度学习影像的生存预测方法虽然已取得巨大进展，但在临床实践中仍然面临一些亟待解决的难题。例如，深度学习方法可以自动提取抽象的影像特征，但是其预测过程是端到端的，只有直接结果，无法提供诊断依据和病因病理，不能被医生完全信任和接受。以脑胶质瘤生存预测为例，医生可通过各种检测方式，结合患者的临床症状来进行生存预测；然而深度学习方法是通过对人工神经网络学习大量带有标记的训练数据进行提取特征，得到的模型在临床上很难解释输入和输出间的因果关系。而可解释性对于医学领域非常重要，可解释性有利于医生更好地理解生存预测方法的强项与不足，并明确生存预测方法的知识边界，了解其在何种情况下有效，从而恰当地信任和使用生存预测方法来进行预测。因此，在现有技术中，深度学习方法缺乏可解释性的问

题导致了医生对其得出结果的置信度较低，难以支持生存预测研究中的因果推理。而且，在医学领域，数据驱动的深度学习算法由于样本数据不足、训练样本分布与真实样本分布不一致等问题，均可能导致算法的性能下降严重，而现有的深度学习方法缺乏可解释性又会进一步导致泛化能力变差。因此，目前基于深度学习影像的生存预测方法泛化能力受到很大的质疑和挑战。

发明概述

技术问题

[0005] 本发明的目的是克服上述现有技术的缺陷，提供一种基于影像基因组学的生存预测方法和系统，解决了基于影像的深度学习生存预测方法缺乏生物学可解释性以及缺乏泛化能力的问题。

问题的解决方案

技术解决方案

[0006] 根据本发明的第一方面，提供了一种基于影像基因组学的生存预测方法。该方法包括以下步骤：

[0007] 步骤S1，获取多个肿瘤患者的影像数据及各患者的生存期数据和RNA数据，建立数据集合；

[0008] 步骤S2，从影像数据中分割出各患者的肿瘤区域，该肿瘤区域包括真实肿瘤区域和疑似肿瘤区域；

[0009] 步骤S3，将各患者的影像数据输入神经网络，以提取影像特征并进行聚类，得到多个影像模块；

[0010] 步骤S4，利用RNA数据获取各患者的基因模块；

[0011] 步骤S5，根据基因模块与影像模块之间的相关性进行筛选，选出满足第一相关性条件的多个强相关的基因模块和影像模块；

[0012] 步骤S6，对所筛选出的基因模块中的基因进行通路富集，得到与影像模块相关的基因通路；

[0013] 步骤S7，计算基因通路的基因集变异分析分数，保留与影像模块的相关性满足第二相关性条件的强相关的基因通路，该基因通路用于对影像模块在生物学层面上进行解释；

- [0014] 步骤S8，使用最终保留的影像特征进行生存预测。
- [0015] 根据本发明的第二方面，提供一种基于影像基因组学的生存预测系统。该系统包括：
- [0016] 数据采集单元：用于获取多个肿瘤患者的影像数据及各患者的生存期数据和RNA数据，建立数据集合；
- [0017] 肿瘤区域提取单元：用于从影像数据中分割出各患者的肿瘤区域，该肿瘤区域包括真实肿瘤区域和疑似肿瘤区域；
- [0018] 影像模块提取单元：用于将各患者的影像数据输入神经网络，以提取影像特征并进行聚类，得到多个影像模块；
- [0019] 基因模块提取单元：用于利用RNA数据获取各患者的基因模块；
- [0020] 第一筛选单元：用于根据基因模块与影像模块之间的相关性进行筛选，选出满足第一相关性条件的多个强相关的基因模块和影像模块；
- [0021] 基因通路分析单元：用于对所筛选出的基因模块中的基因进行通路富集，得到与影像模块相关的基因通路；
- [0022] 第二筛选单元：用于计算基因通路的基因集变异分析分数，保留与影像模块的相关性满足第二相关性条件的强相关的基因通路，该基因通路用于对影像模块在生物学层面上进行解释；
- [0023] 生存预测单元：用于使用最终保留的影像特征进行生存预测。

发明的有益效果

有益效果

- [0024] 与现有技术相比，本发明的优点在于：在现有技术中，基于深度学习影像的生存预测方法没有考虑到生物学上的可解释性，另外由于样本数据不足、训练样本分布与真实样本分布不一致等问题，极大地限制了医生对于基于深度学习影像的生存预测方法的置信度，本发明所提出的基于影像基因组学的生存期预测方法可以提高模型的生物学可解释性、并同时提高深度学习方法的泛化能力。
- [0025] 通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

对附图的简要说明

附图说明

[0026] 被结合在说明书中并构成说明书的一部分的附图示出了本发明的实施例，并且连同其说明一起用于解释本发明的原理。

[0027] 图1是根据本发明一个实施例的基于影像基因组学的生存预测方法的流程图。

发明实施例

本发明的实施方式

[0028] 现在将参照附图来详细描述本发明的各种示例性实施例。应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

[0029] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。

[0030] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

[0031] 在这里示出和讨论的所有例子中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它例子可以具有不同的值。

[0032] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

[0033] 参见图1所示，本发明提供的基于影像基因组学的生存预测方法包括以下步骤。

[0034] 步骤S110，获取多个肿瘤患者的影像数据及各患者的生存期数据和RNA数据，建立数据集合。

[0035] 具体地，多个患者的影像数据为患者的影像集合，例如表示为 $V = \{v_i, i=1, \dots, N\}$ ， N 为患者的数量，每个个体数据 v_i 表示为一个影像样本。

[0036] 在一个实施例中，所建立的数据集合中的患者患有同一种肿瘤疾病，如脑胶质瘤。影像数据为患者进行治疗前的影像，包括但不限于MRI、CI及PET等影像。生存期为影像数据产生时至患者死亡之间的时间段。患者的影像数据例如来源于TCIA（The Cancer Imaging Archive，癌症影像存档）和TCGA（The Cancer Genome Atlas，癌症基因图谱）中共同收录的患者影像数据集。本发明对患者

的影像数据个数不做限定，一般情况下，数据量越多，关联关系越准确。

[0037] RNA（核糖核酸）数据例如包括的核苷酸序列，单核苷酸多态性、结构、性质以及相关描述等。

[0038] 步骤S120，对数据集中的图像数据进行图像标准化。

[0039] 通过图像标准化，可以将不同量级的数据转化为统一量度，以便于后续分析。例如图像数据标准化的方法包括但不限于Z-Score标准化、最大最小标准化及小数定标法等方法。

[0040] 步骤S130，从各患者的影像数据中分割出各患者的肿瘤区域。

[0041] 例如，为了提高生存预测的准确性，肿瘤区域除了包括真实肿瘤区域外，还包括疑似肿瘤区域。

[0042] 步骤S140，建立一个深度学习神经网络，提取影像特征并获取影像模块。

[0043] 可采用多种类型的神经网络，包括但不限于ResNet（残差网络）及其衍生网络、VGG（计算机视觉组）及其衍生网络等。将该深度学习神经网络的最后一个全连接层的输入视为网络从影像数据中提取到的影像特征，提取这些影像特征并进行聚类，得到若干个影像模块。聚类的方法包括但不限于K-Means聚类、均值漂移聚类等方法。

[0044] 步骤S150，利用RNA数据获取基因模块。

[0045] 获取基因模块的方法包括但不限于使用WGCNA分析（Weighted correlation network analysis，加权基因共表达网络分析）。加权基因共表达网络分析（WGCNA，Weighted correlation network analysis）是一种系统生物学方法，用来描述基因间的相关模式。

[0046] 步骤S160，计算基因模块与影像模块之间的相关性，筛选影像模块和基因模块。

[0047] 具体地，首先，对影像模块做降维处理，得到每个影像模块的特征值。

[0048] 使用的降维方法包括但不限于PCA（Principal Component Analysis，主成分分析）、LDA（Linear Discriminant Analysis，线性判别分析）等。

[0049] 然后，通过降维得到的特征值计算基因模块与影像模块之间的相关性，用于计

算相关性的函数包括但不限于皮尔逊相关系数、斯皮尔曼相关系数等。

[0050] 最后，根据基因模块和影像模块之间的相关性，可以筛选出与若干个强相关的基因模块和影像模块，对于相关性不大的基因模块和影像模块不保留。例如，使用相关性 $p=0.05$ 为阈值，仅保留正相关 $p<0.05$ 的或负相关 <0.05 的基因模块和影像模块。

[0051] 步骤S170，对筛选出的基因模块中的基因进行通路富集，得到与影像模块相关的基因通路。

[0052] 通路富集的方法包括但不限于使用Metascape网站、Cytoscape软件、ClusterProfiler R包等，通路富集得到的通路可以用来表征影像模块，即对影像模块在生物学层面上进行解释。例如某影像模块与某基因模块相关，且基因模块与代谢有关，那么可以认为影像模块在生物学上的意义与代谢有强相关的联系。

[0053] 步骤S180，计算基因通路的GSVA分数，进一步筛选与影像模块相关的基因通路，在生物学意义层面对影像模块进行解释。

[0054] GSVA (Gene Set Variation Analysis, 基因集变异分析) 将基因表达矩阵转换成通路富集分数矩阵，通路富集分数矩阵中的分数代表了某通路在某病例上的激活程度。挑选出步骤S170中通路富集中各通路的GSVA分数，使用这些GSVA分数再和步骤S160筛选后得到的影像模块做相关性分析，计算相关性的函数包括但不限于皮尔逊相关系数、斯皮尔曼相关系数等。

[0055] 根据基因通路和影像模块之间的相关性，可以筛选出与若干个强相关的基因通路和影像模块，对于相关性不大的基因通路和影像模块不保留。例如使用相关性 $p=0.05$ 为阈值，仅保留正相关 $p<0.05$ 的或负相关 <0.05 的基因通路和影像模块。使用最终保留的通路对影像模块进行解释，这样影像模块就拥有了生物学上的可解释性。

[0056] 需说明的是，在实际应用中，可根据数据量、执行速度和预测精确度等方面的要求设置适当的相关性阈值，并且与步骤S160中的相关性阈值不必须相同。

[0057] 步骤S190，使用最终保留的影像特征进行生存预测。

[0058] 在删去不相关的影像模块之后，相当于在生物学层面对影像特征进行了特征选择，最终影像模块中的影像特征就是所需要的具有生物学意义的影像特征，这

些具有生物学意义的影像特征能够使神经网络的可解释性和泛化能力都变得更
强。在找到与基因强相关的影像模块之后，对神经网络进行重新训练，在实际
应用中，可以采用多种训练方案。

[0059] 例如，在重训练时，保持全连接层之前的网络的权重不变，在全连接层使用筛
选后的影像模块中的影像特征代替原有的影像特征，然后对全连接层的权重进
行重新训练。

[0060] 又如，使用影像模块的特征值代替原本的影像特征进行重训练。具体地，保持
全连接层之前的网络的权重不变，对筛选后的影像模块做降维处理，得到每个
影像模块的特征值，然后使用影像模块的特征值代替原有的影像特征，对全连
接层的权重进行重新训练。这里使用的降维方法包括但不限于PCA（Principal
Component Analysis，主成分分析）、LDA（Linear Discriminant Analysis，
线性判别分析）等。这种方法使用到的特征值数量比原有的影像特征数量显著
降低，从而增强了泛化能力。

[0061] 经过上述步骤训练和重训练获得的神经网络模型可用于针对个体病例进行预测
。例如，对于输入的病例图像数据，网络最终输出得到一个风险值，这个风险
值可以当作是一个预测因子。进一步地，结合每个病例的生存数据，构建回归
模型（如运用Lasso Cox回归建立预测模型）来对患者进行生存预测。

[0062] 相应地，本发明还提供一种基于影像基因组学的生存预测系统，用于实现上述
方法的一个方面或多个方面。例如，该系统包括：数据采集单元，其用于获取
多个肿瘤患者的影像数据及各患者的生存期数据和RNA数据，建立数据集合；肿
瘤区域提取单元，其用于从影像数据中分割出各患者的肿瘤区域，该肿瘤区域
包括真实肿瘤区域和疑似肿瘤区域；影像模块提取单元，其用于将各患者的影
像数据输入神经网络，以提取影像特征并进行聚类，得到多个影像模块；基因
模块提取单元，其用于利用RNA数据获取各患者的基因模块；第一筛选单元，其
用于根据基因模块与影像模块之间的相关性进行筛选，选出满足第一相关性条
件的多个强相关的基因模块和影像模块；基因通路分析单元，其用于对所筛选
出的基因模块中的基因进行通路富集，得到与影像模块相关的基因通路；第二
筛选单元，其用于计算基因通路的基因集变异分析分数，保留与影像模块的相

关性满足第二相关性条件的强相关的基因通路，该基因通路用于对影像模块在生物学层面上进行解释；生存预测单元，其用于使用最终保留的影像特征进行生存预测。

[0063] 综上所述，本发明所提出的基于影像基因组学的技术方案，首次结合了深度学习技术和生物信息技术，解决了目前基于影像的深度学习生存预测方法缺乏生物学可解释性以及缺乏泛化能力的问题。可应用于肿瘤的生存预测，有利于医生更好地理解生存预测方面的强项与不足，并明确生存预测方法的知识边界，了解其在何种情况下有效，从而恰当地信任和使用生存预测方法，用于临床指导。

[0064] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0065] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是一——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

[0066] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计

计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0067] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++、Python等，以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网（LAN）或广域网（WAN）—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列（FPGA）或可编程逻辑阵列（PLA），该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

[0068] 这里参照根据本发明实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

[0069] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其它设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0070] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列

操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0071] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。对于本领域技术人员来说公知的是，通过硬件方式实现、通过软件方式实现以及通过软件和硬件结合的方式实现都是等价的。

[0072] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。本发明的范围由所附权利要求来限定。

权利要求书

- [权利要求 1] 一种基于影像基因组学的生存预测方法，包括以下步骤：
- 步骤S1，获取多个肿瘤患者的影像数据及各患者的生存期数据和RNA数据，建立数据集合；
- 步骤S2，从影像数据中分割出各患者的肿瘤区域，该肿瘤区域包括真实肿瘤区域和疑似肿瘤区域；
- 步骤S3，将各患者的影像数据输入神经网络，以提取影像特征并进行聚类，得到多个影像模块；
- 步骤S4，利用RNA数据获取各患者的基因模块；
- 步骤S5，根据基因模块与影像模块之间的相关性进行筛选，选出满足第一相关性条件的多个强相关的基因模块和影像模块；
- 步骤S6，对所筛选出的基因模块中的基因进行通路富集，得到与影像模块相关的基因通路；
- 步骤S7，计算基因通路的基因集变异分析分数，保留与影像模块的相关性满足第二相关性条件的强相关的基因通路，该基因通路用于对影像模块在生物学层面上进行解释；
- 步骤S8，使用最终保留的影像特征进行生存预测。
- [权利要求 2] 根据权利要求1所述的方法，其中，在步骤S5中，所述根据基因模块与影像模块之间的相关性进行筛选，选出满足相关性阈值的多个强相关的基因模块和影像模块包括：
- 对影像模块作降维处理，得到每个影像模块的特征值；
- 通过降维得到的特征值计算基因模块与影像模块之间的相关性；
- 根据基因模块和影像模块之间的相关性，筛选出多个强相关的基因模块和影像模块。
- [权利要求 3] 根据权利要求2所述的方法，其中，采用主成分分析或线性判别分析对影像模块作降维处理，根据皮尔逊相关系数或斯皮尔曼相关系数计算基因模块与影像模块之间的相关性。
- [权利要求 4] 根据权利要求1所述的方法，其中，在步骤S7中，所述计算基因通路

的基因集变异分析分数，保留与影像模块相关的基因通路包括：

通过基因集变异分析将基因表达矩阵转换成通路富集分数矩阵，通路富集分数矩阵中的分数代表了相关通路在对应病例上的激活程度；

挑选出骤S6中得到的通路富集中各通路的基因集变异分析分数，使用这些基因集变异分析分数与步骤S5筛选后得到的影像模块做相关性分析；

根据基因通路和影像模块之间的相关性，筛选出满足设定的相关性阈值的多个强相关的基因通路和影像模块。

[权利要求 5] 根据权利要求1所述的方法，其中，在步骤S8中，所述使用最终保留的影像特征进行生存预测包括：

在筛选出与基因通路具有强相关的影像模块之后，对所述神经网络进行重训练，该神经网络的最后一个全连接层的输入作为从影像数据中提取到的影像特征，在重训练时，保持全连接层之前的网络权重不变，在全连接层使用所筛选出的影像模块中的影像特征代替原有的影像特征，然后对全连接层的权重进行重新训练。

[权利要求 6] 根据权利要求1所述的方法，其中，在步骤S8中，所述使用最终保留的影像特征进行生存预测包括：

在筛选出与基因通路具有强相关的影像模块之后，对所述神经网络进行重训练，该神经网络的最后一个全连接层的输入作为从影像数据中提取到的影像特征，在重训练时，保持全连接层之前的神经网络的权重不变，对所筛选后的影像模块做降维处理，得到每个影像模块的特征值，然后使用影像模块的特征值代替原有的影像特征，对全连接层的权重进行重新训练。

[权利要求 7] 根据权利要求1所述的方法，其中，步骤S1还包括：对数据集中的图像数据进行图像标准化，以将不同量级的数据转化为统一量度。

[权利要求 8] 根据权利要求5或6所述的方法，还包括：

对于待预测的病例的影像数据，将其输入经重训练的所述神经网络，输出得到一个风险因子；

根据所述风险因子结合该病例的生存数据，构建回归模型进行生存预测。

[权利要求 9]

一种基于影像基因组学的生存预测系统，包括：

数据采集单元：用于获取多个肿瘤患者的影像数据及各患者的生存期数据和RNA数据，建立数据集合；

肿瘤区域提取单元：用于从影像数据中分割出各患者的肿瘤区域，该肿瘤区域包括真实肿瘤区域和疑似肿瘤区域；

影像模块提取单元：用于将各患者的影像数据输入神经网络，以提取影像特征并进行聚类，得到多个影像模块；

基因模块提取单元：用于利用RNA数据获取各患者的基因模块；

第一筛选单元：用于根据基因模块与影像模块之间的相关性进行筛选，选出满足第一相关性条件的多个强相关的基因模块和影像模块；

基因通路分析单元：用于对所筛选出的基因模块中的基因进行通路富集，得到与影像模块相关的基因通路；

第二筛选单元：用于计算基因通路的基因集变异分析分数，保留与影像模块的相关性满足第二相关性条件的强相关的基因通路，该基因通路用于对影像模块在生物学层面上进行解释；

生存预测单元：用于使用最终保留的影像特征进行生存预测。

[权利要求 10]

一种计算机可读存储介质，其上存储有计算机程序，其中，该程序被处理器执行时实现根据权利要求1至8中任一项所述方法的步骤。

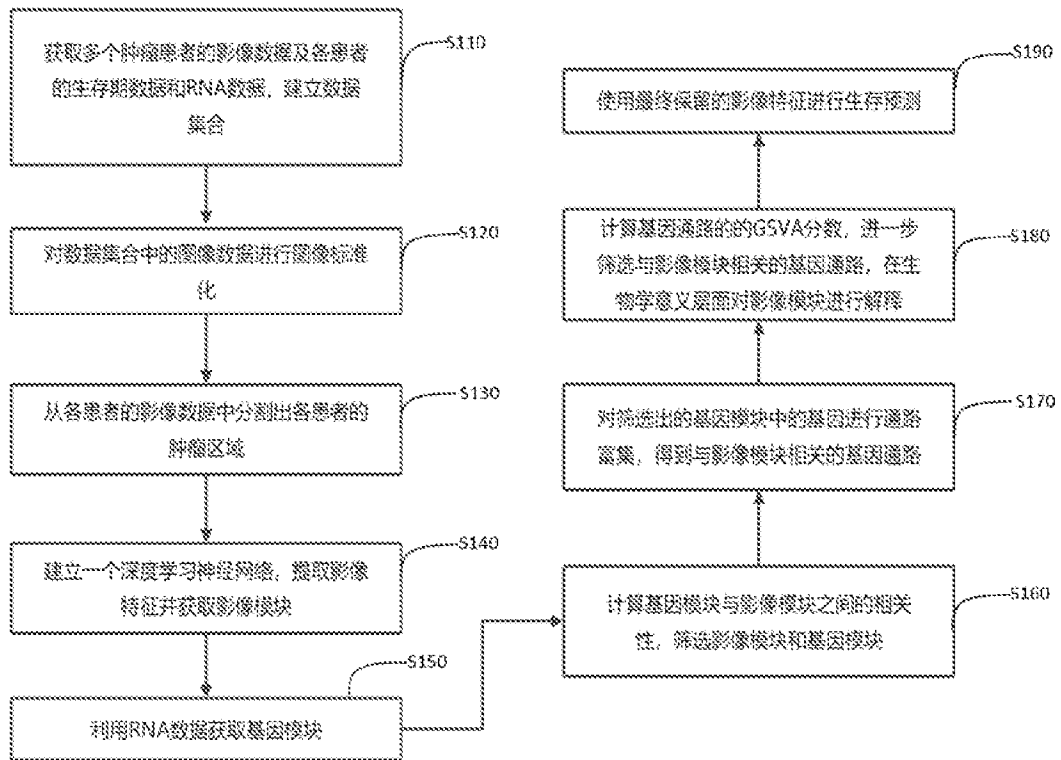


图 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/137310

A. CLASSIFICATION OF SUBJECT MATTER

G06T 7/00(2017.01)i; G06T 7/11(2017.01)i; G06K 9/62(2022.01)i; G16B 40/00(2019.01)i; G16B 20/20(2019.01)i;
G06N 3/04(2006.01)i; G06N 3/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T G06K G16B G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE: 影像, 基因, 生存, 预测, 肿瘤, 神经网络, 深度, 学习, 训练, 权重, 筛选, 强, 相关, 通路, 变异, 生物学, 特征值, 主成分, RNA, image?, gene, survival, prediction, tumor, neural network, depth, learning, training, weight, screening, strong, correlation, pathway, variation, biology, eigenvalue, PCA

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112907555 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 04 June 2021 (2021-06-04) claims 1-10	1-10
Y	CN 106815481 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 09 June 2017 (2017-06-09) description, paragraphs [0026], [0095] and [0109]	1-10
Y	CN 107169497 A (SUZHOU INSTITUTE OF BIOMEDICAL ENGINEERING AND TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 15 September 2017 (2017-09-15) description, paragraphs [0032]-[0041], and figure 1	1-10
A	CN 110097921 A (FUDAN UNIVERSITY) 06 August 2019 (2019-08-06) entire document	1-10
A	US 2019370969 A1 (SIEMENS HEALTHCARE GMBH) 05 December 2019 (2019-12-05) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

28 February 2022

Date of mailing of the international search report

09 March 2022

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China**

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/137310

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112907555	A	04 June 2021	None			
CN	106815481	A	09 June 2017	WO	2018132997	A1	26 July 2018
CN	107169497	A	15 September 2017	None			
CN	110097921	A	06 August 2019	None			
US	2019370969	A1	05 December 2019	EP	3576020	A1	04 December 2019

国际检索报告

国际申请号

PCT/CN2021/137310

A. 主题的分类 G06T 7/00(2017.01)i; G06T 7/11(2017.01)i; G06K 9/62(2022.01)i; G16B 40/00(2019.01)i; G16B 20/20(2019.01)i; G06N 3/04(2006.01)i; G06N 3/08(2006.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																				
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06T G06K G16B G06N 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) WPI, EPDOC, CNPAT, CNKI, IEEE: 影像, 基因, 生存, 预测, 肿瘤, 神经网络, 深度, 学习, 训练, 权重, 筛选, 强, 相关, 通路, 变异, 生物学, 特征值, 主成分, RNA, image?, gene, survival, prediction, tumor, neural network, depth, learning, training, weight, screening, strong, correlation, pathway, variation, biology, eigenvalue, PCA																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 112907555 A (中国科学院深圳先进技术研究院) 2021年6月4日 (2021 - 06 - 04) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>Y</td> <td>CN 106815481 A (中国科学院深圳先进技术研究院) 2017年6月9日 (2017 - 06 - 09) 说明书[0026]-[0095], [0109]段</td> <td>1-10</td> </tr> <tr> <td>Y</td> <td>CN 107169497 A (中国科学院苏州生物医学工程技术研究所) 2017年9月15日 (2017 - 09 - 15) 说明书[0032]-[0041]段, 附图1</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 110097921 A (复旦大学) 2019年8月6日 (2019 - 08 - 06) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>US 2019370969 A1 (SIEMENS HEALTHCARE GMBH) 2019年12月5日 (2019 - 12 - 05) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 112907555 A (中国科学院深圳先进技术研究院) 2021年6月4日 (2021 - 06 - 04) 权利要求1-10	1-10	Y	CN 106815481 A (中国科学院深圳先进技术研究院) 2017年6月9日 (2017 - 06 - 09) 说明书[0026]-[0095], [0109]段	1-10	Y	CN 107169497 A (中国科学院苏州生物医学工程技术研究所) 2017年9月15日 (2017 - 09 - 15) 说明书[0032]-[0041]段, 附图1	1-10	A	CN 110097921 A (复旦大学) 2019年8月6日 (2019 - 08 - 06) 全文	1-10	A	US 2019370969 A1 (SIEMENS HEALTHCARE GMBH) 2019年12月5日 (2019 - 12 - 05) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 112907555 A (中国科学院深圳先进技术研究院) 2021年6月4日 (2021 - 06 - 04) 权利要求1-10	1-10																		
Y	CN 106815481 A (中国科学院深圳先进技术研究院) 2017年6月9日 (2017 - 06 - 09) 说明书[0026]-[0095], [0109]段	1-10																		
Y	CN 107169497 A (中国科学院苏州生物医学工程技术研究所) 2017年9月15日 (2017 - 09 - 15) 说明书[0032]-[0041]段, 附图1	1-10																		
A	CN 110097921 A (复旦大学) 2019年8月6日 (2019 - 08 - 06) 全文	1-10																		
A	US 2019370969 A1 (SIEMENS HEALTHCARE GMBH) 2019年12月5日 (2019 - 12 - 05) 全文	1-10																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期 2022年2月28日		国际检索报告邮寄日期 2022年3月9日																		
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 马鑫 电话号码 86-(10)-53961354																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/137310

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	112907555	A	2021年6月4日	无			
CN	106815481	A	2017年6月9日	W0	2018132997	A1	2018年7月26日
CN	107169497	A	2017年9月15日	无			
CN	110097921	A	2019年8月6日	无			
US	2019370969	A1	2019年12月5日	EP	3576020	A1	2019年12月4日