# 内容

## 1 前言

### 1.1 背景

在医学研究中，组织学为肿瘤诊断和治疗的提供了重要信息。解剖病理学家能通过评估组织学特征，如核异型、有丝分裂活动、细胞密度和组织结构等，同时结合相应的细胞以及更高层面，如器官、系统等，提供的信息对病情发展进行分类和分级。

随着基因测序的进步，目前，直接观察基因组中生物标志物，如基因表达和表观遗传修饰变化，以分析肿瘤病症已成为可能。但组织分析仍是预测患者未来病程的重要且不可或缺的工具。组织特征表现出分子层面改变所带来的总体影响。同时，组织能提供直观的视觉信息以帮助医疗人员判断癌症侵入性。然而，组织分析有极大的缺陷——即其是高度主观、不能重复得到相同结果的。

与之相反的是，对计算机而言，其结果由特定的一系列计算得出。显然地，在计算过程不变的情况下，同样的数据输入，其结果是一定不变的。

此外，计算机能提取组织学成像的全部信息，而人仅能凭经验观察特定区域。如此，组织学成像中被人忽视的数据能被计算机利用以使诊断更全面。如本研究中所用的10倍镜下的每一例切片成像会划分为5e+05张左右的96\*96像素的小区域图像。设想若人工需要1秒钟分析一张，则每一例需要近14小时。研究中共使用了759例病理图像，则仅图像信息提取步骤上就需要10626小时，即442.75天。即使10人组成研究组，也需要不停止、不降速地工作44天左右方能完成信息提取。可见，人是不大可能能提取组织学成像的全部信息的。而本研究使用的计算机（见附录），在只执行一次计算且不考虑I/O延迟的情况下，处理全部图像仅需要15小时。故计算机有能力胜任这类图像的大数据分析，而由人去实现是高成本而不现实的。

因此，通过计算机对组织学成像进行分析，不但能克服人工进行组织分析的缺陷，而且能提取人工进行组织分析忽略的信息。可见，应用计算机分析组织学成像以辅助医疗诊断极具前景。

### 1.2 现状

目前，计算组织学成像分析受到了极大的关注。随着相关技术的进步和计算性能的提升，许多图像分析算法已经被开发用于分级，分类和鉴定多种癌症类型的淋巴结转移。

深度卷积神经网络（Deep Convolutional Neural Networks, DCNN）是一个重要的图像分析工具。它的出现打破了许多图像分析挑战的记录。DCNN从原始图像数据中学习预测特征的能力具有泛化性，即其不被图像类型或分析任务限制。这一特性极其利于医学图像分析。

医学图像分析往往依赖于特征工程提取方法。如使用分割算法，明确圈定感兴趣的结构以测量一些特定的结构特征。经验上，这些特征能包含能分析病情的信息。因此，研究人员会使用这些特征的数据来训练预测患者预后的模型。与之不同的是，DCNN的特征学习范式是自适应学习。它能根据具体的学习目标的不同，将图像信息转换为含有相应预测信息的特征。这种特征学习避免对特征提前定义带来的偏差。同时，由于分割算法的结果易受一些无关变量的影响，其很难有普遍的应用。具体而言，分割算法一般借助图像颜色或边缘等信息，而这类信息易受图像缺陷，如光斑，以及一些外界变化，如染色时间造成颜色明度的变化等的影响。在这一点上，DCNN极具优势，因为它不需要使用人工编写的分割算法。

2018年，Google的AI研究部门Deepmind开发了基于CNN技术的NASNet模型。在ImageNet图像分类中，NASNet在测试集上实现了82.7％的预测准确度。该结果超过了他们之前建立的所有模型。此外，NASNet的性能比之前发布的所有模型高1.2％，与arxiv.org未公布的最佳结果相当。此外，可以调整NASNet的大小以产生一系列模型，这些模型具有良好的性能，同时计算成本很低。大型NASNet具有最高的准确性，同时只需要一半SENet——arxiv.org上报告的最佳结果——的计算成本。可见，NASNet模型在图像分析应用上极具潜力。

虽然特征学习已逐渐成为DCNN的低难度任务，但在医学应用上它面临特殊的挑战。首先，训练DCNN需要大量标记数据。目前相应的高质量数据不足限制了DCNN的能力。此外，DCNN是一个“黑箱”体系，难以进行逐步分析。因此，很难解释DCNN的预测机制。然而，在医学应用上，解释是必不可少的。不过，尽管如此，DCNN仍成功应用于各类医学图像分析。

临床中的许多重要问题涉及事件时间预测，如预测癌症的总生存期和病情进展时间。虽然在其他应用中取得巨大成功，深度学习还没能普遍地解决这些问题。生存分析可以被认为是二元分类问题，即以一特定时间跨度为分界，预测其生存时间是长于还是短于该时间。然而，以分类的方式进行生存分析有严重的缺陷。首先，数据收集时间跨度少于这一特定时间的案例均不能被使用。同时，分类只能讨论该特定时间的生存情况，而不能模拟其他时间的生存情况。为了解决这个问题，统计学家设计了一系列“时间-事件”模型，如Cox回归模型。这类模型可以使用具有任意时间跨度的案例，并且能对任意时间点的生存率的进行估计。

在设计基于神经网络的Cox回归模型的早期探索中，有研究使用包含数十特征的训练集成功训练模型。然而，这种神经网络模型并不优于基本的线性Cox回归模型。近期，有研究使用了更先进的“深层”神经，包含多个为优化Cox比例风险可能性（Cox proportional hazard likelihood）而设计的网络层。接着，他们使用临床信息数据训练模型。这样产生的模型在预测方面具有比基本的线性Cox回归模型更好的表现。

从组织学成像直接训练癌症生存模型更为困难。这种“时间-事件”模型和使用DCNN分析组织学成像有相似的挑战，即数据量不足；也有不同的挑战，即肿瘤内异质性和特征隐蔽性。首先说明数据量不足这一相似挑战：对于图像分析而言，每个像素的特征是不仅由其独立的信息（位置，RGB或CMYK）决定，更由其所在的背景决定。整合这些背景信息需要使用卷积运算。但由此建立的模型是极其复杂、巨大的——往往包含千万个的参数。确定这些参数需要有庞大的训练集提供数据。然而，一般能获得的数据量不能满足这一条件，如本研究使用的TCGA-BRCA数据库，其中仅有1063份案例。借鉴训练一般图像识别模型的方法，多个对组织学成像的研究使用了“数据增强”（data augmentation）技术以应对这个问题。该技术通过随机对原始数据进行“微调”，如改变图像的亮度、对比度以及旋转图像等，以生成新的数据，或多或少地能缓解数据量不足带来的问题。至于肿瘤内异质性这一问题则尤为独特且不易解决。因为肿瘤组织通常包含一系列体现不同程度病情进展的区域。找到方法整合一个病人样本中不同区域的信息，以预测该病人的生存结果是使这类模型有效的关键。此外，与风险相关的肿瘤特征往往在组织外观的细微变化中。这些细微变化需要通过多种组织学标准检验才能发现。而人类病理学家需要多年的专业培训才能应用相应的检验去发现、解读这些细微变化。因此，至今，由组织学成像建立有效的生存模型仍极具挑战。

近期，有一神经胶质瘤相关的研究提出的基于DCNN的生存模型解决了上述问题。该研究通过专家分析划定筛选出神经胶质瘤图像中包含关键信息的区域。然后，用这些区域以建立训练集。同时，通过对这些区域使用“数据增强”以解决数据不足问题；通过以随机选择一个样本中多个区域、选择第二大风险的预测结果代表整个样本结果的方法解决“肿瘤内异质性”问题；通过以19层视觉几何组（Visual Geometry Group, VGG）卷积网络架构建立庞大模型解决“特征隐蔽”这一问题。研究实现了0.741 一致性指数（concordance index, c index），高于所有之前类似研究的结果。不仅如此，其准确性甚至超过了人类专家使用相应临床分类标准得到的预测。然而，训练该模型需要专家先对图像进行划分。同时，由于该模型庞大，其不易在当前的个人计算机上训练及使用。

### 1.3 亮点

本文提出一种基于Nasnet的乳腺癌生存神经网络模型（SNAS）。它能分析乳腺癌组织学成像提供较为准确的“时间-事件”，即生存预测。该模型与使用专家分析同样图像提取的特征建立的生存分析模型能达到基本相当的表现，同时，该模型能由文中搭建的全自动的由.svs病理切片图像到生存模型的“端到端”系统生成。该系统由分割，提取，训练三部分构成。每一个部分均无需人为介入。此外，该模型大小中等，7.24e+07个参数，其中4.26e+07个参数为Nasnet以迁移Kaggle数据训练，故实际需要计算的参数量仅为2.97e+07个。

## 2 数据来源

### 2.1 癌症基因组图谱

癌症基因组图谱（The Cancer Genome Atlas, TCGA）是一始于2005年的项目。其主要工作是使用基因组测序和生物信息学对癌症相关的基因突变进行编目。如此，其通过应用高通量基因组分析技术，能帮助医疗工作者了解癌症的遗传基础，以提高医疗工作者诊断，治疗和预防癌症的能力。

TCGA隶属于美国国立卫生研究院（National Institutes of Health, NIH）的国家癌症研究所（National Cancer Institute， NCI）的癌症基因组学中心。同时，TCGA与NCI的基因组数据共享（Genomic Data Commons, GDC）研究计划并行。GDC为癌症研究界提供了统一的数据存储库。凭此，GDC实现了癌症基因组研究中的数据共享，促进了精准医疗的发展。当前，大量癌症图像分析相关的研究应用其中的影像数据。

本文的影像数据均来自GDC中TCGA-BRCA项目的数据,可由此<https://portal.gdc.cancer.gov/projects/TCGA-BRCA>访问。

### 2.2 切片选择

TCGA-BRCA提供快速冷冻（Flash Frozen）和福尔马林固定石蜡包埋（Formalin-Fixed Paraffin-Embedded, FFPE）这两种类型的切片影像。

快速冷冻样品通常在手术期间在冷冻液中制作，主要用于帮助外科医生确定肿瘤的边界是否清洁（即，肿瘤是否已在手术中被完全切除）。快速冷冻是一种快速且简单的过程，但经常会使组织受损，使其外观呈多孔状。于是，相应的，快速冷冻图像中与肿瘤相关的关键信息很可能已经丢失；FFPE切片是诊断医学的金标准。FFPE的制作需要先将样本固定在甲醛中，然后将其嵌入石蜡块中进行切割。它呈现的组织更为完整，使其更适合应用于计算分析。因此，本研究仅选择使用FFPE类型的数字影像数据。

### 2.3 专家识别特征

本文使用的专家识别特征数据库来自于近年发表的某一对乳腺癌形态学及相应分子机制的研究。在该研究中，15名病理学家从TCGA-BRCA获得图像数据对850例侵入性乳腺癌案例进行了组织病理学分析。他们评估了11项形态学特征，包括组织学分级：上皮小管形成（Epithelial Tubule Formation），核多形性（Nuclear Pleomorphism），有丝分裂数量（Mitotic Count）；原位癌情况：导管原位癌（Ductal Carcinoma in Situ, DCIS），小叶原位癌（Lobular Carcinoma in Situ, LCIS）；其他特征：基质炎症（Stromal Inflammation），坏疽（Necrosis），上皮癌百分比（% Cancerous Epithelium），大汗腺特征（Apocrine Features），淋巴血管侵入（Lymphovascular Invasion），基质中央纤维化聚集（Stromal Central Fibrotic Focus）。所有评估结果均通过相互评估（Inter-rater Reliability）检验。

本研究使用的专家识别特征信息均可于<http://legacy.dx.ai/tcga_breast/annotations.html>获得。

## 3 实验内容

总体的实验流程如下图。我们先以专家识别特征建立Cox HP模型得到参照基本线（baseline）。之后，我们对初始切片图进行区域分割及预筛选。在训练得到NASNet分类器后，使用该分类器对区域再次进行筛选。接着，我们将筛选区域作为样本内容，采用包括迁移学习在内的一系列方法对生存模型进行训练。最后对生存模型进行相应的优化和评估。

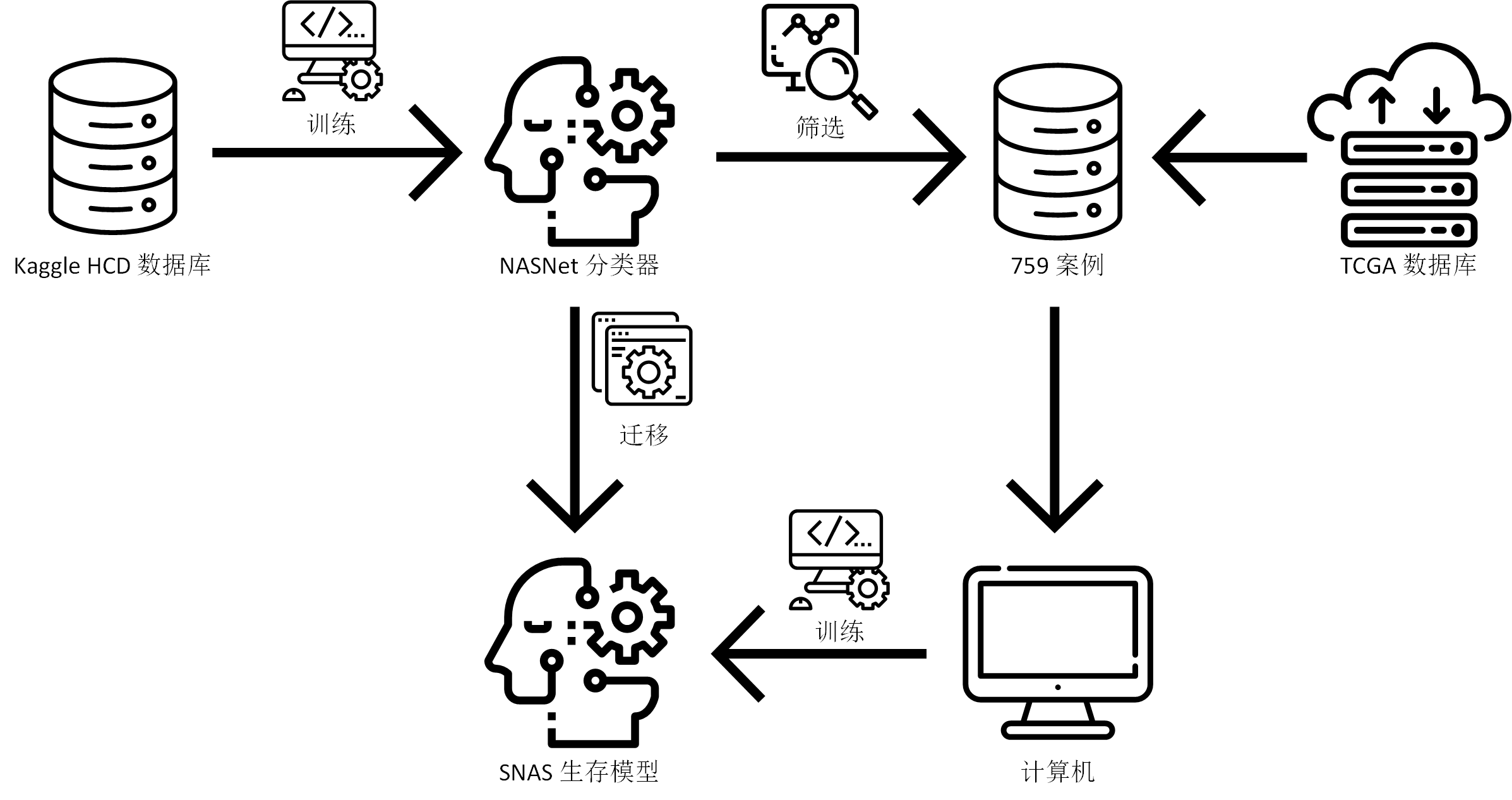


图 1

### 3.1 专家识别特征建立Cox HP模型

比例风险模型（Proportional hazards models, PH models）是统计学中生存模型的一类。通过分析在事件发生前经过的时间与和该时间量可能相关联的一个或多个协变量的关系，生存分析建立起“时间-事件”模型。在PH模型中，协变量中单位增加的效应与相应的危险率是乘积关系。例如，服用药物可能会使一个人发生中风的危险率降低一半。 （ Breslow, N. E. (1975). “Analysis of Survival Data under the Proportional Hazards Model”. International Statistical Review / Revue Internationale de Statistique. 43 (1): 45–57. doi:10.2307/1402659. JSTOR 1402659. ）

比例风险模型（Proportional hazards models, PH models）是统计学中生存模型的一类。通过分析在事件发生前经过的时间与和该时间量可能相关联的一个或多个协变量的关系，生存分析建立起“时间-事件”模型。在PH模型中，协变量中单位增加的效应与相应的危险率是乘积关系。例如，服用药物可能会使一个人发生中风的危险率降低一半。

Cox比例风险模型（Cox Proportional hazards model, Cox PH model）是PH模型中的一种。

其常被医学统计用于研究患者存活时间和一个或多个预测变量之间的关联。它和Kaplan-Meier曲线、logrank测试不同——Kaplan-Meier曲线和logrank测试是单变量分析，它们会忽略其他因素的影响，只能探究一个因素与生存时间的关联；同时它们需要的变量类型是分类类型，如疗法A、疗法B或男性、女性，而不适用于定量类型的变量。Cox比例风险回归分析则不同，它能同时评估多个风险因素对生存时间的影响。并且，除了分析分类类型变量外，它也能分析定量类型变量。

本文应用基于Python的生存分析应用lifelines,<https://lifelines.readthedocs.io/en/latest/index.html>,进行Cox PH模型的建立。

### 3.2 区域分割及预筛选

影像数据分析的一大挑战是样本的图像非常大。因此，其难以全部直接读取至随机读取内存（random access memory, RAM），或者应用一般的流水线（pipeline）进行分析。以本研究所使用的TACG-BRCA数据为例。其以Aperio公司SVS文件为格式，759个样本的10倍镜成像平均包含5.38e+08个像素点，40倍成像图平均包含8.61e+09个像素点。如果在RGB模式，uint8数据格式下读取，40倍成像的图像将占用近26GB的RAM空间，此外，之后的运算还需要空间存储中间结果。而当前常规的计算机一般配备8至16GB的RAM。因此，直接读取文件进行分析是不可取的。同时，我们观察到10倍镜成像下，肿瘤细胞聚集分布在图中的某些区域，一个细胞仅占8\*8的区域。此外，图中有大量的空白区域或重复的大块区域，直接使用全图是不明智的。从经验上讲，对空白区域可用简单的算法进行预筛选。

#### 3.2.1 显微放大倍数选择

10倍镜成像虽相较于40倍成像镜有一定的损失，但10倍镜成像下细胞占8\*8像素，观察尚包含一定的信息。

此外，后续我们NASNet分类器及迁移学习使用的Kaggle数据库（见方法）数据为10倍镜下成像。于是，我们在研究中也选用了10倍镜下的成像，以在计算机资源使用和有效信息保留上取得良好的平衡。

#### 3.2.2 openslide进行96\*96区域切割

openslide是基于C语言的读取此类大型图像的开源工具。本文使用最新的3.4.1版。

显而易见地，我们需要将整个图像分割为多个小的区域。在对每个区域进行分析之后，再将分析结果进行整合。当前图像深度学习使用的图片一般会缩小为96\*96的小图，以便于分析。因此，我们使用openslide将原始的大图在调节至10倍成像。然后，openslide将10倍镜成像图分割为96\*96的小图。

#### 3.2.3 OpenCV预选

OpenCV，即开源计算机视觉库，是一个开源的应用于计算机视觉及机器学习的软件。OpenCV是计算机视觉应用常用的平台。

在分割过程中，我们利用简单的灰度分析对区域进行预筛选。具体而言，我们将小图由RGB模式转化为灰度模式。通过计算，我们能得到全图的平均灰度值以代表整个小图。该平均灰度值在0~255之间。一般地，包含大量空白的区域的平均灰度值会很低。于是，我们能通过观察设置一个阈值，低于该值的即为包含大量空白的区域。这样，我们使用OpenCV执行上述操作，以筛选除去了空白的区域。

### 3.3 NASNet区域分类器

2017年5月，Google Brain设计了能产生人工智能(artificial intelligence, AI)的人工智能，AutoML。NASNet是Google的研究人员使用强化学习，以AutoML作为控制神经网络自动训练出来的神经网络模型。在2018年，NASNet已是图像识别领域的最佳模型。

#### 3.3.1 NASNet分类器架构

NASNet模型分为NASNet large模型和NASNet mobile模型。NASNet mobile是其中较小的模型，其硬件需求更小，但仍能实现强大的图像识别功能。

NASNet分类器的第一层为96\*96\*3的数据输入层。之后，则是NASNet mobile（文中的NASNet均指NASNet mobile）层。NASNet层会产生3组数据，下一层的合并层（concatenate layer）将3组数据合并。接着，一个随机丢失层（dropout layer）用于防止过度拟合（overfitting），增强模型鲁棒性（robustness）。最后是sigmoid函数激活的1维全连接层（dense layer）。

该分类器使用分类任务常用的二元交叉熵函数（binary cross entropy function）

作为损失函数（loss function）。

#### 3.3.2 Kaggle数据库

Kaggle是Google旗下的用于数据科学研究和机器学习的平台。Kaggle组织病理肿瘤识别（Histopathologic Cancer Detection, HCD)数据库是其2018年开展组织病理肿瘤识别挑战时创建的数据库。

Kaggle HCD数据库的源头是Camelyon16挑战数据库。该数据库包含400张40倍镜下的H&E染色切片全图。PCam数据库产生于将Camelyon16挑战数据库的40倍镜成像缩小为10倍成像，且分割为96\*96的小区域。本文训练NASNet分类器使用的Kaggle数据库是Pcam数据库的子集。Kaggle数据库和Pcam数据库的差别在于Pcam数据库中的重复的图片全部被去除了。（[1] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling. “Rotation Equivariant CNNs for Digital Pathology”. arXiv:1806.03962[2] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199–2210. doi:jama.2017.14585）

Kaggle HCD数据库以在图像中心（32\*32）区域是否包含至少一个像素的肿瘤组织为评判标准对所有图像数据进行了分类。60%的图像为阴性，即图像中心没有肿瘤组织；40%的图像为阳性，即图像中心至少有一个像素的肿瘤组织。于是，能实现出色分类的模型一定有分析肿瘤组织的能力。因为我们所研究的生存模型是以分析肿瘤组织的图像为基础，所以该数据库训练出的模型很适用于之后生存模型的迁移学习。

可于<https://www.kaggle.com/c/histopathologic-cancer-detection>访问Kaggle HCD数据库。

#### 3.3.3 “数据增强”训练

由于高质量数据的获取需要大量资源，如人力、金钱、时间，数据量不足这一问题的解决不能仅仅依靠收集更多数据这一手段。于是，在机器学习领域，通过对原始数据添加随机的轻微“扰动”生成新数据，即“数据增强”，研究人员很好地以低成本方式解决了数据量不足这一问题。

在图像中，“数据增强”的手段一般有翻转、旋转、亮度变化、切割等。我们在训练模型使用的“数据增强”方式如下表。

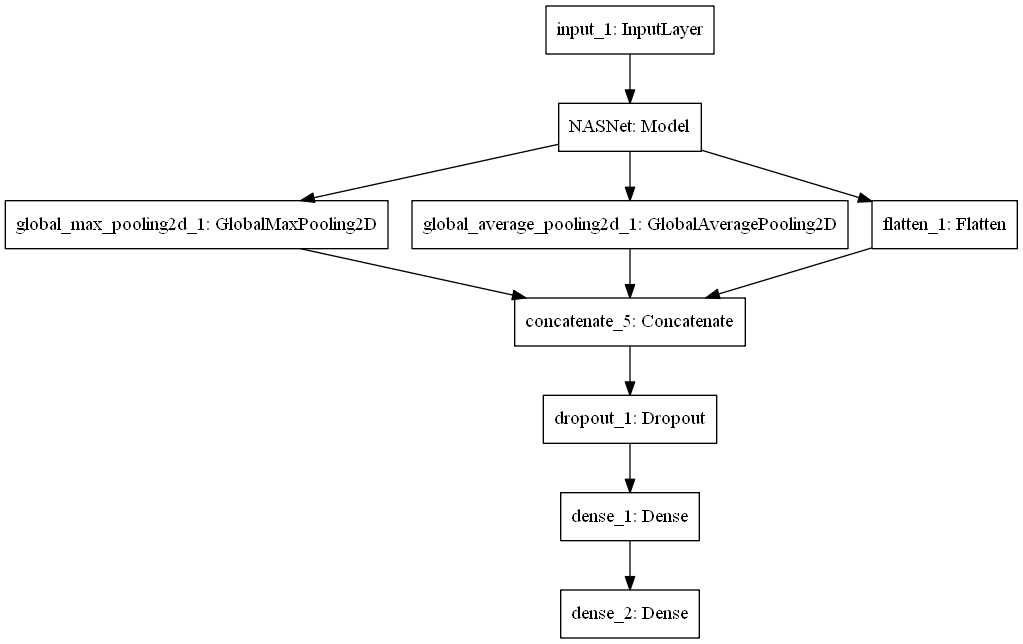
#### 3.3.4 区域采集量

事件发生的时间能影响生存模型的损失函数。因此，训练生存模型时，每次训练需要将数据按时间顺序输入。于是，在此处，一般图像识别为避免数据次序影响时所用的乱序（shuffle）方法一定不能使用。同时，分批（batch）的数据输入策略也不应使用。我们采用的方法是每次将全部样本一起输入进行训练。这样，每次训练时，一个样本需要从待选区域里面选择一张作为代表。此时，有效区域数量一定，如果待选区域数量过多，那么选择到有效区域的几率将大大减小。故我们需要设置NASNet分类器的筛选阈值，以最大程度地提高训练时训练数据选择到有效区域的可能。

### 3.4 训练

在训练时，我们运用了迁移学习（Transfer learning）策略。迁移学习是一种机器学习研究方法。简单而言，该方法将解决一个问题的“知识”存储，然后应用于不同但类似的问题。比如，可尝试使用识别轿车的模型对卡车进行识别。 （West, Jeremy; Ventura, Dan; Warnick, Sean (2007). “Spring Research Presentation: A Theoretical Foundation for Inductive Transfer”. Brigham Young University, College of Physical and Mathematical Sciences. Archived from the original on 2007-08-01. Retrieved 2007-08-05.）

#### 3.4.1 SNAS生存模型架构



chartflow

SNAS生存模型的第一层由96\*96\*3的数据输入层，之后则是NASNet层。NASNet层会产生3组数据，合并层将3组数据合并；合并的数据经过一个随机丢失层后输入一个全连接层；最后是基于全连接层的生存预测层。两个全连接层均使用L2函数

作为核规范器（kernel regularizer）、运算规范器（activity regularizer），以Glorot均衡起始器（Glorot uniform initializer,或称Xavier uniform initializer）以促进损失函数收敛，防止过度拟合，增强其普遍化的能力。（Understanding the difficulty of training deep feedforward neural networks）

从架构可见，我们将NASNet同样应用于生存模型中，设计得到我们的SNAS。NASNet含有近4.27e+07个参数。这些参数由学习Kaggle数据库得到。我们将NASNet层的参数冻结，以存储训练于Kaggle数据库能识别肿瘤区域的能力，实现迁移学习。

生存模型在机器学习中使用了一个特殊的损失函数，即负对数似然损失函数（negative log likelihood loss function）

#### 3.4.2 随机区域选择

如我们在区域采集量中提到的生存模型训练的需求——数据需要按照时间顺序排序，同时损失函数会因数据选取（不同时间点，数量大小）而变化。于是，我们每次选取一个区域代表样本，然后将所有样本输入进行训练。由于我们没有相应的专业人员对我们的区域进行人工筛选或鉴定，此处我们只能采用随机选取的方法，即从NASNet分类器筛选的某样本区域中随机挑选一个区域代表该样本。

### 3.5 批量预测

应对肿瘤内异质性，我们参考了PNAs文章中使用的方法，即随机选取多个区域进行预测，然后选择风险预测值最大或者第二大的结果代表此样本的生存风险。在临床中，医务工作者对肿瘤发展的评估是根据肿瘤患者以各病患处表现出的最严重的特征。比如，若患者某处的组织已表现出转移瘤特征，而其他地方的肿瘤尚处于良性阶段，该患者的生存情况将会依据转移瘤的特征进行评估。在TNM（Tumor，Node，Metastasis；肿瘤，结节，转移）系统中会被评估为M阶段。（Brierley, J.D.; Gospodarowicz, M.K.; Wittekind, Ch., eds. (2017). TNM classification of malignant tumors (8th ed.). Chichester, West Sussex, UK: Wiley-Blackwell. ISBN 978-1-4443-3241-4.）同时，从肿瘤学上讲，在很大程度上，癌症的治疗和存活是通过其是否保持局部化或已扩散到身体的其他位置来决定的。如果癌症已经能转移到其他组织或器官，患者的死亡可能性通常会显着增加。（ Klein CA (September 2008). “Cancer. The metastasis cascade”. Science. 321 (5897): 1785–7. doi:10.1126/science.1164853. PMID 18818347.）

#### 3.5.1 随机多区域选择

经验上讲，分析参考的区域越多，得到准确预测的可能性越大。然而，增加区域参考量的同时，需要的计算资源也会同时增加。于是，我们需要在计算资源的限制下，尽可能多地选择区域的情况。同时，由于用NASNet分类器选择出的有效区域数量有限，如果选择的区域数量太大，只会重复分析同一区域，平白消耗计算资源。结合计算机科学中对42的经验认同，在模型评估时，我们选择的42个区域供模型分析。

### 3.6 超参数（hyperparameter）优化

在机器学习中，虽然参数的值可以通过训练得出，但在学习过程开始之前，我们需要对模型本身设置参数。这一模型训练前需要设置的参数即为超参数。需要通过实验测试以实现合理超参数的设定。

超参数的选择能决定训练和测试模型所需的时间。同时，超参数能显著影响模型的表现。然而，超参数通常是连续或整数类型，其模型相应的表现结果变化不一定是有规律的；比较模型的表现又需要用新的超参数重头训练。因此，虽然找到良好的超参数组合是极其重要的，但在有限的计算资源下，找到最优的超参数组合又是不大可能实现的事。（“Claesen, Marc, and Bart De Moor.”Hyperparameter Search in Machine Learning." arXiv preprint arXiv:1502.02127 (2015)". arXiv:1502.02127. Bibcode:2015arXiv150202127C.）

目前我们采取的策略是：选择最有影响力的超参数进行组合，然后在这些搭配中，找到当前资源限制下，较优的一个超参数的组合。

#### 3.6.1 模型大小

从我们SNAS的模型架构可见，合并层后全连接层的大小直接影响着该模型的大小。我们需要从具有256节点（node）的全连接层和具有512节点的全连接层中选择一个适合该生存预测任务的。

#### 3.6.2 “数据增强”选择

我们在NASNet分类器训练中提到了“数据增强”方法。该方法能以低成本解决数据不足问题，以防止过度拟合，保证模型的普遍适用性。然而，有时，对图像的随机改变可能影响其中的关键信息。比如，肿瘤细胞的形态是判断其类型的依据；“数据增强”可能影响这些细胞的形态使模型不能从图像中得到准确的真实信息。于是，输入错误的数据，模型也只会输出不准确的结果。因此，在SNAS模型训练及测试中是否使用“数据增强”技术也是我们的超参数之一。

## 4 结果

一致性指数（concordance index, c index），通过量化排名（ranking）质量评估模型能力，是生存分析中模型表现评估的标准指标。 (On Ranking in Survival Analysis: Bounds on the Concordance Index）本研究中对模型优劣的评估均以c index为准。

如果数据存在访问缺失（censoring），即在数据收集时间内目标事件没有发生，使用如均方误差或平均绝对损失之类的损失函数是不合理的。在本研究中，许多（？）乳腺癌患者在数据收集时间内未死亡。因此，不应使用如均方误差或平均绝对损失之类的损失函数。此时，应当使用一致性指数（concordance index, c index）。该指数评估预测时间排序的准确性，以量化排名（ranking）质量评估模型能力，是生存分析中评估模型表现的标准指标。它实际上是另一种常见的损失函数AUC（）的泛化。和AUC类似，c index的数值有如下解释：0.5，随机预测；1.0，预测与事实完全一致；0.0，预测与事实完全不一致（乘以-1即得到1.0的预测）。质量一般的生存模型的c index通常介于0.55到0.7之间。因为生存分析相关的数据中往往会包含很多噪音，所以，即使模型非常优秀，其c index也不会特别高。

### 4.1 Cox HP基础模型表现

五次Cross Validation（CV）的平均CI为0.71， 标准差为0.03。可见，使用人工分析图像得出数据建模的预测能力中等。之后的深度学习模型若超过该预测水平，则说明该研究下使用的方法优于人工。

### 4.2 NASNet分类器表现

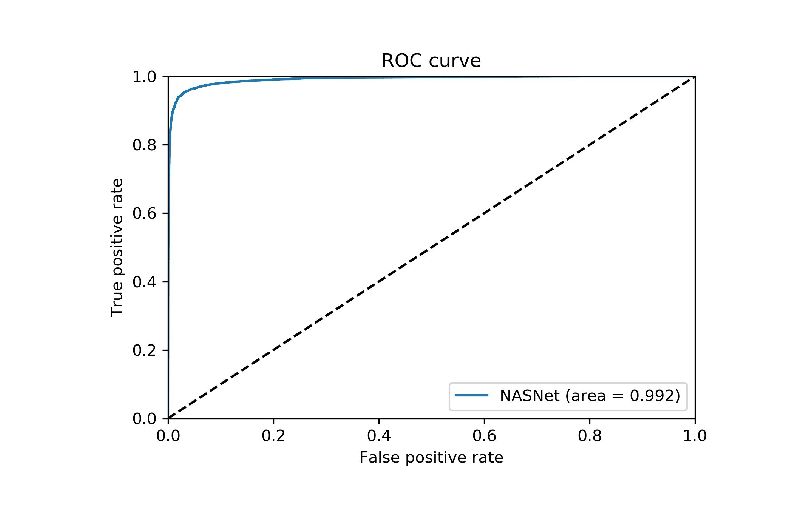


图1

### 4.3 切片区域可采集量

分类器按照区域内细胞是否与其在Kaggle HCD数据库中学习的肿瘤细胞特征相似来进行评估，若相似，则预测值大，反之，则预测值小。预测值在0~1之间，因此我们需要选择一个阈值在其中，以区分肿瘤组织区域及非肿瘤组织区域。

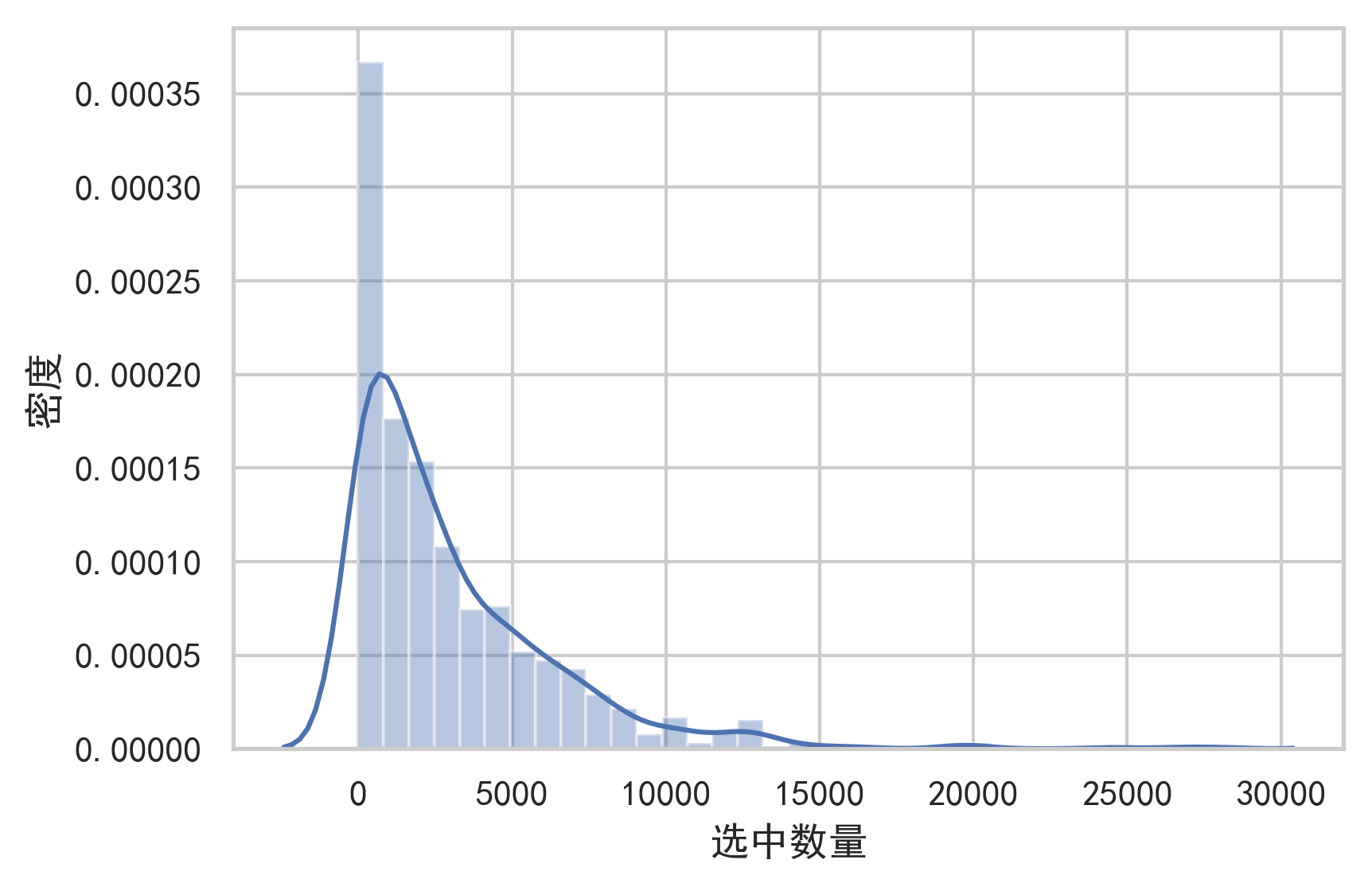
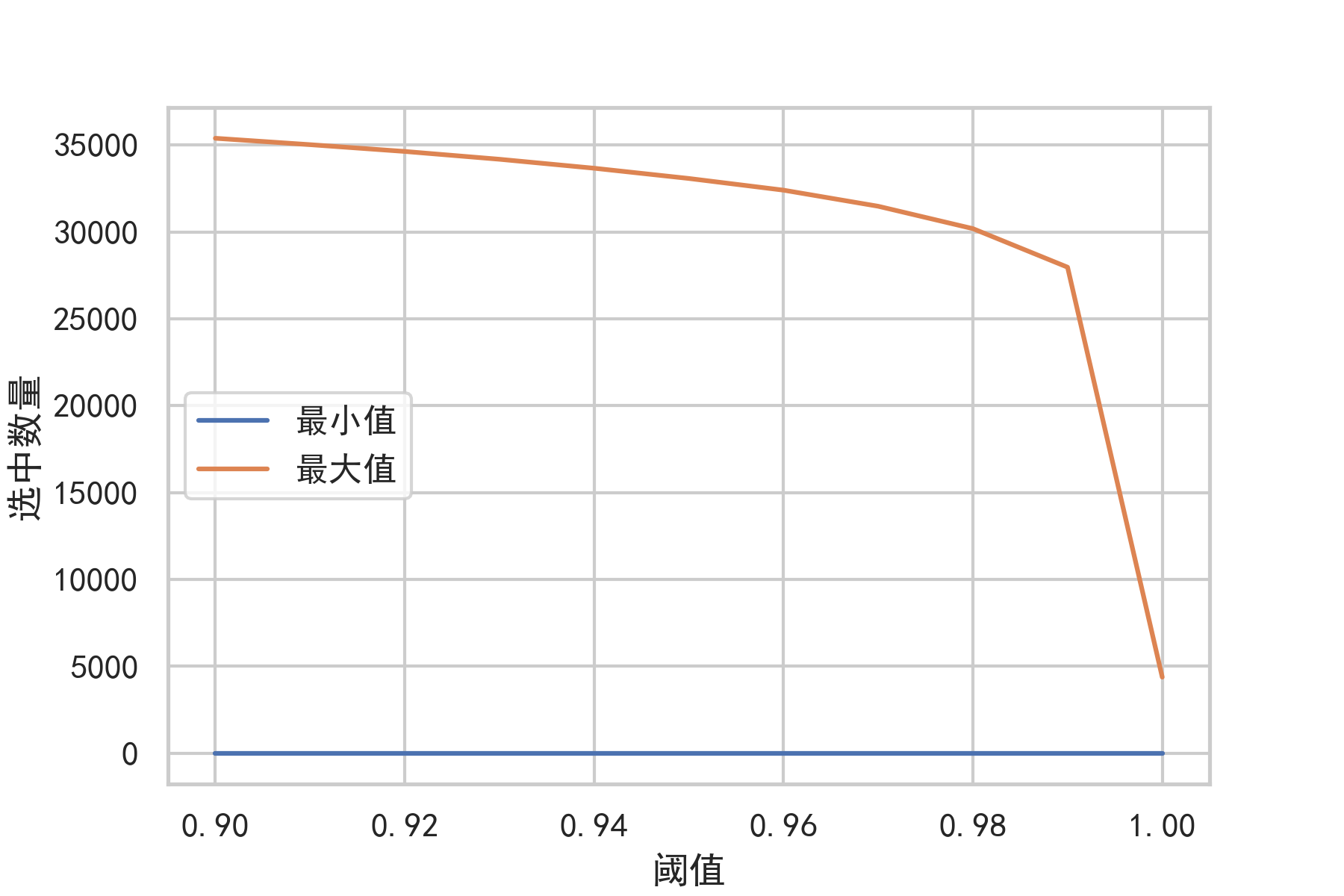
如图所示，NASNet分类器的分类结果数量分布呈幂分布（power distribution），即大量的样本的被选区域数量在？？以下的。同时，分类器阈值设置的高低（从0.95到趋于于1）对其中的最小值没有影响，只对最大值影响。因此，对阈值的设定只能调节样本被选区域量的上限，而不能调节被选区域量的下限。

从图中容易发现，在0.95~0.99之间，被选区域数量最大值的下降比较平缓，0.99至1（趋近于1）被选区域数量最大值很快下降。可见，0.99是最大值变化于0.95到1区间的拐点所在。因此，我们选择0.99作为基础的阈值。分类器预测值在0.99以上的区域将被纳入被选区域，以进行之后的训练和测试，在0.99以下的区域则暂时不选取。

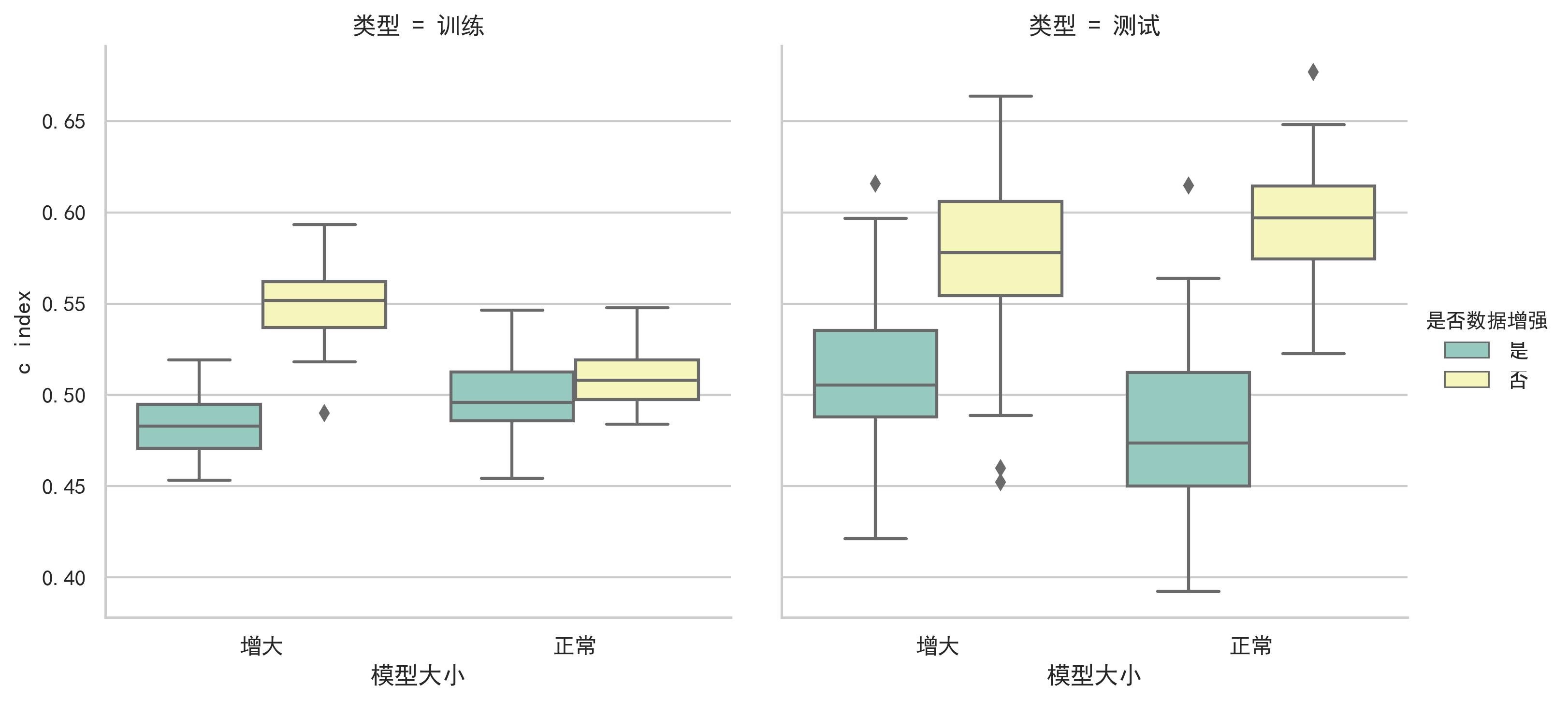
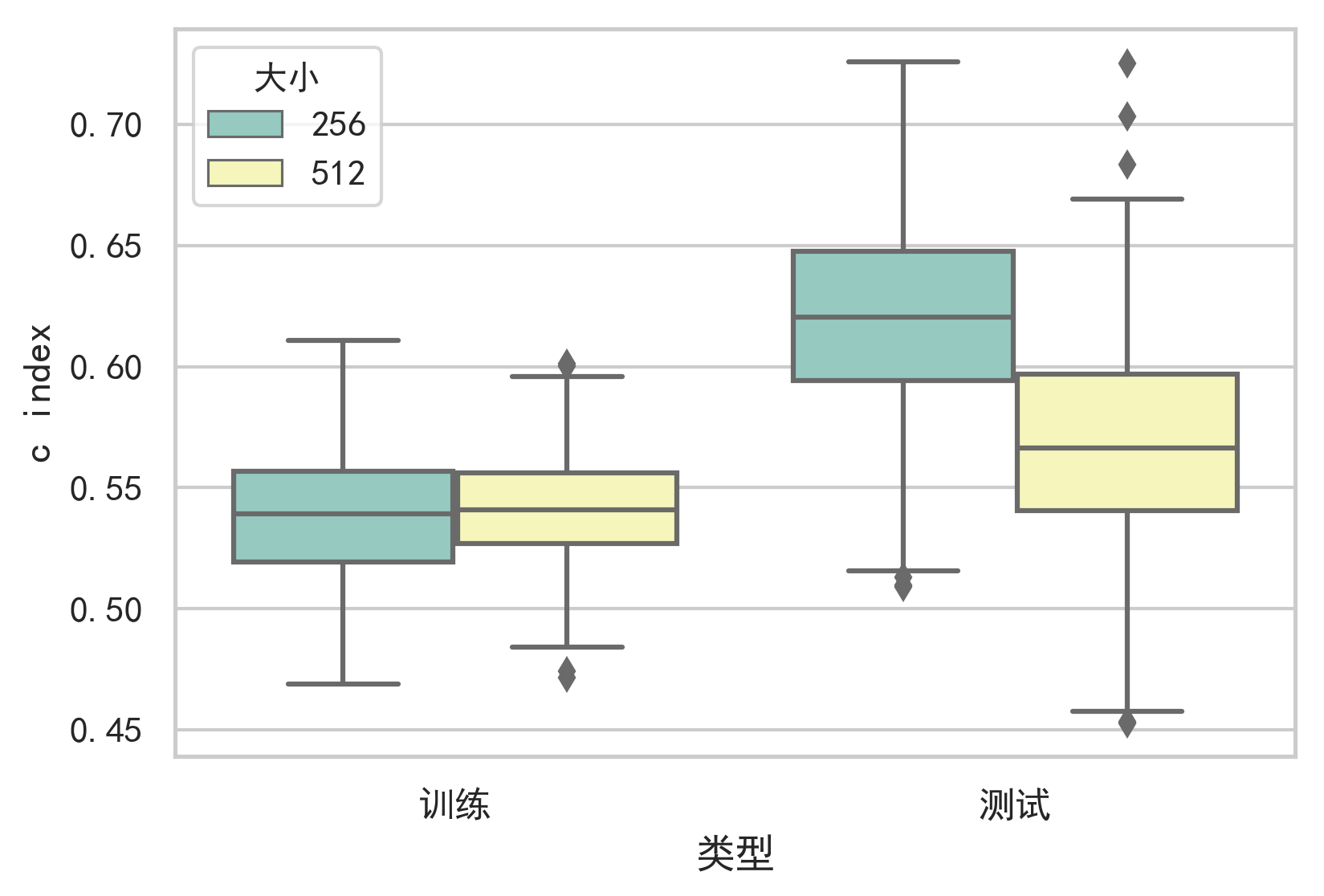
然而，我们认为“肿瘤细胞特征不明显的区域多”这一导致样本被选区域量很少的特点很有可能和样本的生存预后有关。某区域内的肿瘤细胞特征不明显说明该区域细胞分化良好。在细胞层面，病理学家会按细胞的分化程度来评估癌症的预后。因此，如果我们对那些选择区域少的样本只按研究设定的0.99以上来选取区域，它们将只有一个区域被选取，其很多区域细胞分化良好这一特征将丢失。我们希望能把“多区域分化良好”这一特征保留。于是，我们对那些选择区域少的样本相应地降低阈值，使它们各自均能有50个区域被选中。

最后的样本区域选取情况如下表。

|  |  |  |
| --- | --- | --- |
|  | 数量 | 比例 |
| 均值 | 2.98e+03 | 5.73e-02 |
| 标准差 | 3.38e+03 | 6.17e-02 |
| 最小值 | 5.00e+01 | 2.58e-04 |
| 最大值 | 2.80e+04 | 5.60e-01 |

### 4.4 SNAS超参数优化

如图2所示，在模型对测试集、训练集中案例的预测中，不使用数据增强的表现都显著地优于使用数据增强的表现，无论其是大模型还是正常模型。原因之一可能是对图像的随机改变影响了图中的关键信息。如“数据增强”中的“扭曲”、“颜色改变”等可能影响这些细胞的外观形态使模型从图像中得到的信息并不是真实准确的。于是，基于不准确的信息，模型产生了不准确的预测。因此，在本模型的训练及测试中，与一般的深度学习不同，应当不使用“数据增强”技术。

如图B所示，对测试集而言，使用的模型大小对表现没有显著性影响，即无论随机丢弃层后的全连接层有256个节点还是512个节点，其预测相似。就训练集而言，大模型的表现显著性优于小模型的。显而易见地，大模型比小模型多了2.97e+07个可训练参数，因此有更强的能力拟合训练集。然而，当我们将训练次数增加，此处为400，可见大模型对训练集的预测表现是显著优于对测试集的，说明大模型的预测已经处于过度拟合。在预测模型能力评估上，普遍性，即能用于未知的情况，是比是否拟合当前的数据重要的。因此，我们在256节点和512节点上，我们选择了256节点。

### 4.5 SNAS生存模型表现

我们以不使用数据增强的方式共计训练了600个256节点为中间全连接层的模型，每一次模型的参数均递归于上次的模型，这样的训练方法和推荐模型的on-fly相似。因为每一次训练均是从607样本中随机选取一个区域，有时该区域包含有效的信息，有时该区域不包含有效的信息，所以，虽然每一次模型的参数训练均基于上次的模型，后一个训练出的模型却不一定会比前一个好。为选取到其中最好的模型，我们将600次的模型均进行了存档，以方便后续使用最好的图像训练模型结合基因信息训练出能同时分析图像和基因组数据预测生存的模型。

600次训练中，模型中拟合训练集最好的达到c index 0.61，拟合验证集中最好的达到c index 0.75。

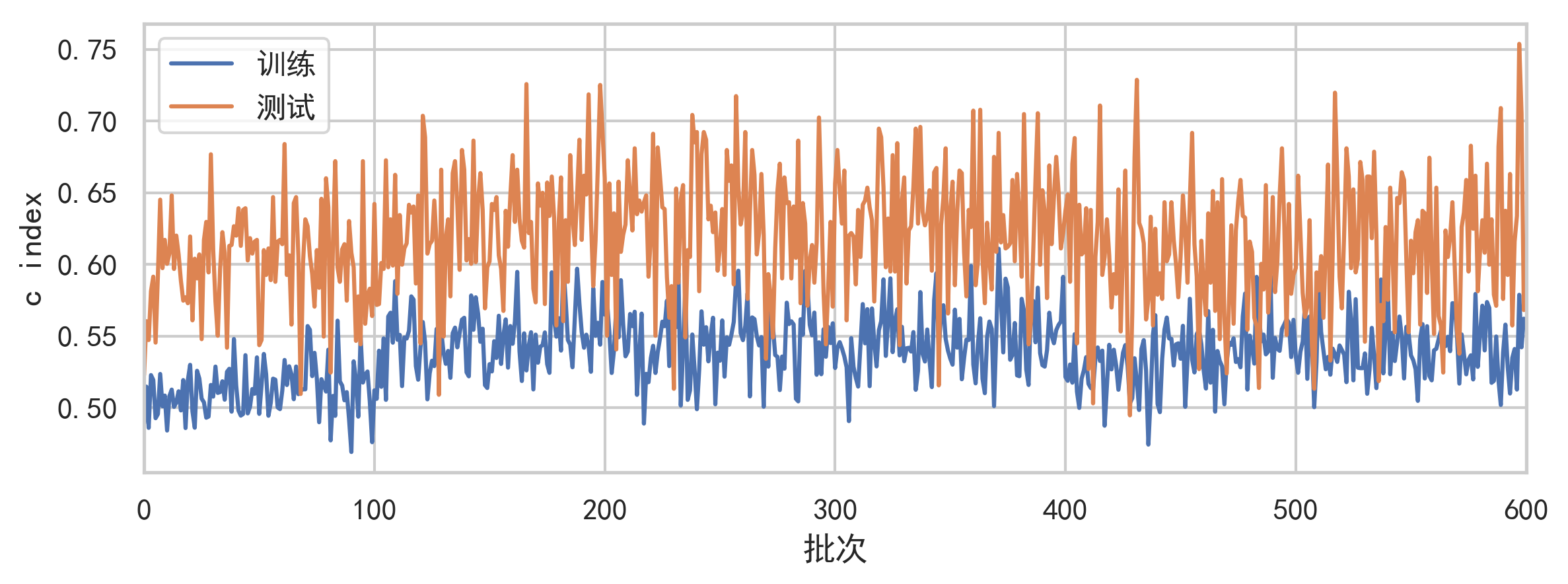


图2

### 4.6 数据数据量影响

然而，即使其中最好的模型，c index 0.61训练集表现，或0.75 c index的测试集表现也只属于中庸水平。其中很重要的原因是当前使用的训练数据量较少，如下图？？可见，使用同样的架构使用不同的数据量，训练的最终表现显著不同。同时，训练时的数据量越多，其表现越好。

### 4.7 端到端系统

值得注意的一点是，本文的模型训练需要任何的人力接入。因此，它是一个全自动的生存模型训练系统。该系统较好地解决了模型训练需要大量人力标注这一问题。以此，本文实现了从.svs文件到生存模型的流水线（end-to-end pipeline）。该流水线可于GitHub<https://github.com/Moo-YewTsing/Survival-Analysis-by-Breast-Cancer-Slides>参考。

## 5 讨论

### 5.1 乳腺癌特征

乳腺癌的生存情况取决于癌症类型，发展程度和年龄。（“Breast Cancer Treatment (PDQ®)”. NCI. 26 June 2014. Archived from the original on 5 July 2014. Retrieved 29 June 2014.）在发达国家，乳腺癌的生存率很高，（Breast cancer screening in developing countries - NCBI - NIH）比如，在美国，乳腺癌的五年生存率在89.4％左右。（“SEER Stat Fact Sheets: Breast Cancer”. NCI. Archived from the original on 3 July 2014. Retrieved 18 June 2014.） 在发展中国家，存活率却不容乐观，仅为57%。（Breast cancer screening in developing countries - NCBI - NIH）就世界范围而言，乳腺癌是女性的主要癌症类型，占所有病例的25.2％，占癌症死亡成因的14.7%。（World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 1.1. ISBN 978-92-832-0429-9.） 2018年，它导致了超过200万新病例，同时它在发达国家更常见（Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, in press.）。

多种分级系统能对乳腺癌进行分类。不同类型的乳腺癌会影响预后并影响治疗反应。

乳腺癌通常主要通过其组织学外观进行分类。大多数乳腺癌来源于导管或小叶内衬的上皮细胞，这些癌症被归类为导管或小叶癌。此外，原位癌由特定组织隔室（例如乳腺导管）内的低级癌或癌前细胞发展而来，它不会侵入周围组织。相比之下，浸润性癌并不局限于最初的组织隔室。（Merck Manual, Professional Edition Archived 10 November 2011 at the Wayback Machine, Ch. 253, Breast Cancer.）在细胞层面，分级将比较乳腺癌细胞的外观与正常乳房细胞的外观。像乳房这样的器官中，正常细胞会进行分化，即它们会形成反映其作为该器官一部分的功能的特定形态。癌细胞失去了这种分化能力。在癌症中，通常以有序的方式排列以构成乳管的细胞排列无序；细胞分裂不受控制；细胞核不均匀。病理学家按细胞的分化程度将其分为良好（低级），中度分化（中级）和低分化（高级），分别对应细胞的正常乳腺细胞中所见特征的丧失程度。分化差的癌症（其组织与正常乳腺组织细胞最不相似）的预后较差。在分期上，乳腺癌使用TNM系统分期。该系统基于肿瘤的大小（T），肿瘤是否已经扩散到腋窝的淋巴结（N），以及肿瘤是否已经转移（M）（即扩散到肿瘤发生源头远处的身体部分）。体积较大，淋巴结扩散和转移具有较大的分期数和较差的预后。

### 5.2 模型情况

该模型的主体为Google Brain的NASNet模型，该模型为Google Brain使用其开发的Auto ML自动探索生成，在普通的图像识别中，其能力已经表现优异。在本文研究完成时，我们尚未发现有其他研究将NASNet应用于生存分析模型。我们的结果表明NASNet是适用于生存分析的。毕竟从概念层面来讲，分析图像的生存模型需要先适用于图像识别的神经网络，以往的研究使用DCNN，我们使用NASNet，在概念层面均是应用于图像识别的神经网络。因此，既然NASNet在图像识别上优于其他的神经网络，我们认为使用NASNet替代DCNN是一个很好的策略。

事实上，可以通过模型预测热图可视化（heat map visualization）技术观察SNAS识别的部分。目前由于缺乏相关专业的人员支持，暂时无法对模型关注的区域进行分析。这些被关注区域的分析应当能为解释模型提供帮助，同时为如何优化模型提供依据。因此，此处的关注区域分析是补足该研究的重点之一。同时，或许模型的关注能力本身也能帮助医疗人员搜索重点区域，协助其进行诊断。 该模型大小适中，当前主流的个人计算机（3G显示卡内存，8G RAM内存）即可正常使用。目前的相关生存分析模型则相当巨大，如。对硬件要求较低能使更广泛的人群使用。与此同时，该模型的相对简单也更方便他人进行增强和修改。

### 5.3 数据情况

本研究使用的数据在深度学习领域相对而言是比较少的，因此不得不使用了迁移学习。在研究初期，我们将生存模型中的NASNet部分也作为参数学习对象，然而由于生存分析本是一件相对于普通图像识别困难的任务，同时训练样本量仅有607，模型无法实现数据拟合，Negative Log 损失函数的结果甚至出现overflow，即超过系统所能处理的范围，出现“Nan”(not a number)的错误数字。

我们的迁移学习使用了Kaggle HCD数据库。该数据库的数据标记的乳腺癌病理图像。我们所研究的生存分析问题正是关于乳腺癌。正因如此，由该模型训练出来的NASNet分类器不仅仅完成了样本有效区域筛选这一任务，同时其中的NASNet部分迁移接入我们之后训练的生存模型。我们研究的结果表明Kaggle HCD数据库具有相当高的质量，类似的研究如果遭遇高质量数据不足的问题时或许可以试着应用该数据库。

不过随着医疗信息电子化，以及医疗信息学的更广泛应用，数据不足这一问题将被逐渐解决。正如我们在结果数据量部分说明的，训练样本的数量越多，其训练出来的模型表现也就越出色。

### 5.4 应用

随着发达国家老龄化的加剧，以及全球人口人均寿命的普遍提高，癌症的发病率逐年提高。然而，此时人口结构变化，发达国家新生代人数逐年减少，发展中国家的人口出生率随着经济的发展也在下降中。这意味这在不久的将来，更重的医疗问题需要更少的人去解决。

文中提及的专家分析特征的工作从2011年11月开展，一直持续到2014年3月才结束。分析的850份样本历时近28个月。可见，直接使用人力对医疗图像进行分析的成本很高。使用这些专家分析的特征，我们建立的基本Cox HP模型对测试集的预测表现为0.71 c index，对训练集的拟合表现为0. c index。模型中拟合训练集最好的达到c index 0.61，预测验证集中最好的达到c index 0.75。因此，虽然该系统的模型总体表现并未超过基本Cox HP模型。但这是一个真正意义的全自动模型训练系统。因此，这类低人力成本模型，尽管没有超过人工模型，其仍然具有良好的应用发展前景。