ICS 11.040.99
CCS C30

中华人民共和国国家标准

GB/T XXXXX—XXXX

人工智能医疗器械 质量要求和评价 第5部分：预训练模型

Artificial intelligence medical device—Quality requirements and evaluation—Part 5: Pre-trained models

立项草案稿

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前　言

本文件按照GB/T 1.1-2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件是国家标准GB/T XXXX《人工智能医疗器械 质量要求和评价》的第5部分。GB/T XXXX已经发布了以下部分：

——第1部分：术语；

——第2部分：数据集通用要求；

——第3部分：数据标注通用要求；

——第4部分：可追溯性。

——第5部分：预训练模型。

请注意本文件的某些内容可能涉及专利。本部分的发布机构不承担识别这些专利的责任。

本文件由国家药品监督管理局提出并归口。

本文件起草单位：

本文件主要起草人：

引　言

近年来，人工智能医疗器械不断发展，成为医疗器械标准化领域的一个新兴方向。我国已初步建立人工智能医疗器械标准体系。在该标准体系中，GB/T XXXX《人工智能医疗器械 质量要求和评价》是基础通用标准，为开展细分领域的标准化活动提供指导，拟由八个部分组成。

——第1部分：术语。目的在于为人工智能医疗器械的质量评价活动提供术语。

——第2部分：数据集通用要求。目的在于提出数据集的通用质量要求与评价方法。

——第3部分：数据标注通用要求。目的在于提出数据标注环节的质量要求与评价方法。

——第4部分：可追溯性。目的在于明确人工智能医疗器械的可追溯性通用要求与评价方法。

——第5部分：预训练模型。目的在于规范人工智能医疗器械采用的预训练模型质量。

——第6部分：合成数据。目的在于规范人工智能医疗器械采用的合成数据质量要求与评价方法。

——第7部分：安装验证。目的在于加强人工智能医疗器械安装验证环节的质量控制。

——第8部分：伦理要求。目的在于从技术层面实现人工智能伦理的要求，保护人的权益。

本部分为人工智能医疗器械使用的预训练模型质量评价的相关工作提供了思路，也为后续制定细分算法模型专用质量要求提供了依据。

人工智能医疗器械使用的预训练模型来源广泛，包括医疗器械厂家、第三方供应商、第三方服务平台、网络开源等。受技术、商业和政策等因素限制，预训练模型的技术细节、研发过程和质量控制全套记录一般较难获取。为了有效地控制和追溯人工智能医疗器械终产品的质量，本文件对人工智能医疗器械使用的预训练模型本身和相应说明文档提出质量要求和评价方法，以引导人工智能医疗器械厂家从内部加强质量控制。

由于预训练模型本身不属于医疗器械且技术路线处于快速发展阶段，本文件不对预训练模型本身的研发过程进行约束，避免限制创新。对于预训练模型的版本变更、采用动态更新的第三方服务、具备自学习能力等情形，本文件基于具体时间和具体版本进行质量评价，版本变更和更新参照医疗器械变更规定和标准进行管理。

对于研发企业，本文件为预训练模型的遴选、质量控制提供了依据。对于第三方检测机构，本文件为预训练模型的测试活动提供了依据。

人工智能医疗器械 质量要求和评价 第5部分：预训练模型

1 范围

本文件规定了人工智能医疗器械使用的预训练模型的通用质量要求和评价方法。
本文件适用于人工智能医疗器械使用的预训练模型。
本文件不适用于预训练模型的研发过程评价。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

YY/T 1833.1 人工智能医疗器械 质量要求和评价 第1部分：术语
YY/T 1833.2-2022 人工智能医疗器械 质量要求和评价 第2部分：数据集通用要求
YY/T 1833.3-2022 人工智能医疗器械 质量要求和评价 第3部分：数据标注通用要求

3 术语和定义

YY/T 1833.1界定的以及下列术语和定义适用于本文件。

3.1

预训练模型 pre-trained model

已在数据集上经过训练的计算模型，可用作新任务的基础。

注：预训练模型狭义上通常用于迁移学习，其中模型在一个源任务上进行预训练，然后在下游任务上进行微调；在广义上，预训练模型可能作为机器学习模型的初始值，供下游任务进行微调。附录A.1对预训练模型的来源和类别进行扩展性说明。

3.2

预训练模型说明 pre-trained model description

陈述预训练模型各种性质的文档。

3.3

任务域 task domain

机器学习模型需要解决的特定问题或任务的领域。

注：在人工智能医疗器械的应用中，任务域的示例有：影像病灶检测、影像学征象分类、影像ROI区域分割、超声视频分割、影像报告生成、心电信号检测、图像生成、流程优化等。

3.4

源任务 source task

用于训练机器学习模型的初始任务。

注： 附录A.2对源任务、下游任务和预训练模型的关系进行说明。

3.5

下游任务 downstream task

预训练模型利用预训练中获得的泛化能力所要解决的现实任务，通常根据对应任务的数据集和任务类型通过对预训练模型进行微调来实现。

3.6

数据模态 data modality

数据集中收集或表示数据的类型或形式。

注：人工智能医疗器械中的常见数据模态包括医学图像、文本、信号数据和结构化数据。

3.7

超参数 hyperparameter

事先给定的，用来控制学习过程的参数。

注：超参数包含模型层数，模型节点数，以及学习率等。

3.8

微调 fine-tuning

为提升人工智能模型的预测精确度,一种先以大型广泛领域数据集训练，再以小型专门领域数据集继续训练的附加训练技术。

[来源：GB/T 41867-2022，3.4.29]

3.9

本体 ontology

对一个论域中存在的概念及其关系和性质的可共享的、形式化的、显现的描述；使用该本体所建立的模型可以被其他人员或系统共享。

注：本文件中的论域指一个问题或任务的上下文或领域，用于描述问题的范围或限定，以便更好地理解和解决问题。

4 预训练模型说明要求

4.1 概述

预训练模型提供方应描述与该预训练模型本体相关的过程、方法和所使用数据集的全部相关细节。包括预训练模型的版本标识（发布版本号）、模型框架、用于预训练的训练数据及数据集构建方法、标注方法、训练过程、模型适用性及其他相关细节。

注： 附录A.3给出预训练模型提供方的扩展性描述，附录B给出4.2~4.5的示例。

4.2 模型框架描述

4.2.1 模型结构

预训练模型说明应描述模型的整体结构及来源。如模型结构属于对已公开结构的修改，应详细描述二者的差异；如模型结构属于未公开的自研结构，则需详细描述该模型构成的详细数学或结构表述。

注： 附录A.4给出常见的深度学习和非深度学习的模型结构示例。

4.2.2 模型节点

预训练模型说明应描述模型节点信息，该节点信息宜包含神经元节点数量、分布方式、激活函数、连接方式等。

注：神经元节点为深度神经网络的最小组成节点；如模型为非神经网络的算法，应说明模型的算法类型以及其最小组成单位。

4.2.3 模型参数

预训练模型说明应描述模型中带有参数的各层相应的参数元数据。该元数据宜包括模型各层的节点数量、组合关系和类别。

注： 附录A.5给出模型参数的元数据示例。

4.2.4 模型数据表示

预训练模型说明应描述预训练模型中使用的数据表示。如数据的抽象程度、数据类型及规模、用于将原始数据映射到预训练模型中的编码方法、数据必要的预处理、后处理等。数据抽象程度应描述可被模型处理的数据被抽象化的程度，可包含原始数据（动态数据、图像数据、信号数据等）和特征数据（由特征提取器获得的抽象特征数据）。

注：抽象程度是对从数据到特征进行抽象过程复杂性和深度的一种度量描述，随着特征的传递逐渐增加。预训练模型的输入可能是原始数据或已被提取的特征数据。附录A.6给出特征的概念。

a） 数据类型

应描述模型输入接受的数据类型，该数据类型可包括整数、浮点数、字符串、列表、数组等。

b） 数据维度和大小

应描述模型中接受数据的维度以及各维度下的数据元素个数。

c） 编码方式

应描述原始数据如何在模型中被映射，包括使用各种编码方法，如独热编码（one-hot encoding）、标签编码（label encoding）、序列编码（ordinal encoding）、频数编码（count encoding）、目标编码（target encoding）等。

注：附录A.7给出常见的各种编码方式的描述。

d） 数据处理

应描述模型能够接受的输入数据的限定、输出数据的形式以及输入输出数据需进行的必要处理。如适用，应描述模型接受数据所需要的预处理，如归一化，缩放。如适用，应描述模型输出数据的后处理，如拼接、插值、平滑等。

4.2.5 模型可解释性

如适用，预训练模型说明宜描述模型的可解释性。

注： 附录A.8给出模型可解释性的含义。

4.3 源任务数据描述

4.3.1 数据模态

预训练模型说明应描述对预训练模型进行训练时，使用的源任务数据模态，如语音、自然图像、医学图像等。如适用，应对医学数据的采集设备进行说明。应描述是否使用仿真数据对预训练模型进行训练，如使用了仿真数据，应详细描述数据生成方法和步骤。如适用，应描述训练数据标注的相关信息，该信息宜包含标注的来源、数量以及标注质量的说明。

4.3.2 数据量

预训练模型说明应描述用于训练预训练模型的数据总量。如适用，应描述每种数据模态的数据量；应描述预训练过程中数据划分方式。该说明应包含训练集、验证集和测试集的数据量,以及训练集、验证集和测试集之间的分布差异。

4.3.3 数据质量

预训练模型说明应描述用于预训练模型训练数据质量。数据质量应包括数据的准确性、完整性和一致性。应说明用于选择训练数据的质量标准，如数据来源、可靠性和相关性，以及数据清洗的步骤。

对于模型训练过程中使用的数据集，如满足YY/T 1833.2-2022的要求，应在预训练模型说明文档中按照YY/T 1833.2-2022第4章数据集说明文档的要求进行说明。如数据集建立过程依从特定标准，应在预训练模型说明文档中进行说明。对带有标注结果的数据集，如满足YY/T 1833.3-2022的要求，应在预训练模型说明文档中按照YY/T 1833.3-2022第4章数据标注任务说明文档的要求进行说明。

注： 附录A.9给出部分人工智能医疗器械数据集建立过程的规范和标准。

如适用,预训练模型说明应描述用于保护受试者隐私的技术手段,如数据去标识化、数据匿名化等。如适用，训练数据说明应描述数据去标识化或数据匿名化的规则。

4.3.4 基准性能

预训练模型说明宜描述源任务中预训练模型在训练集、验证集和测试集的相关基准性能。

注：附录A.10对基准性能指标的选取给出说明。

4.4 训练过程描述

4.4.1 学习方法

预训练模型说明应描述预训练模型训练所设置的源任务，以及所采用的学习方法等。

4.4.2 训练设置

预训练模型说明宜描述预训练模型训练所采用的数据增强方式、模型权重初始化方式、优化器，以及主要超参数的设置情况。

4.4.3 任务域

预训练模型说明应描述预训练模型生成过程中源任务的任务域。如使用了多种源任务，应说明所有设置的源任务所对应的任务域。

4.5 模型适用性描述

4.5.1 适用数据模态

预训练模型说明应描述预训练模型进行前向推理所适用的数据模态。如适用，可对模型微调后所能处理的数据模态进行说明。

注1：预训练模型可能是适用于通用领域的数据模态，也可能被专门设计来处理医学图像，如MRI或CT扫描，或医疗文本，如电子健康档案（Electronic Health Record, EHR）数据。

注2：前向推理是机器学习模型在训练完成后，对特定任务进行处理的过程，如在医疗领域进行辅助决策过程中分割、分类、增强等任务处理过程。

4.5.2 适用任务类型

预训练模型说明应描述对于预训练模型进行前向推理所适用的任务类型（如图像分割、图像分类、信号识别等）、源任务域和下游任务域之间的相似度、所能处理的场景、样本形式的具体情况以及预期的性能。如适用，应描述在其适用下游任务中可采用的健壮性，泛化性测试方法。

4.5.3 适用环境

预训练模型说明应描述对于模型推理和训练所适用的硬件资源以及软件环境。

注：适用的硬件资源是指运行预训练模型所需的计算能力和资源（如CPU、GPU和服务器节点等，通过每秒浮点运算次数、每秒操作次数等指标来度量），如模型微调所需的计算能力以及软件环境；以及人工智能医疗器械在实施或部署环境中运行模型所需的计算能力以及软件环境。

4.5.4 下游任务训练需求

预训练模型说明宜描述预训练模型迁移到新的下游任务域的数据需求，即输出空间映射发生改变时，重新进行训练需要的新下游任务对应的数据集（包括所需数据量、数据模态等）。如适用，应描述在微调中采用的训练设置需求，包括微调采用的数据增强方法、模型训练方法等。

5 预训练模型质量特性

5.1 概述

本部分内容涵盖预训练模型的质量特性、整体风险等要素，宜根据模型的预期用途、应用场景对预训练模型开展质量评价，形成技术报告，作为对预训练模型质量的验证依据。

5.2 质量特性

5.2.1 可训练性

预训练模型提供方应声称预训练模型的可训练性指标，并提供书面证据。如适用，应使用损失函数值、目标数据分布的拟合程度等指标。

注： 可训练性是指预训练模型能够在训练过程中迭代优化。

5.2.2 架构可扩展性

预训练模型提供方应声称预训练模型架构能否通过增加计算能力和资源提高推理和训练效率，以及对计算能力和资源的需求要求，并提供书面证据。如适用，应描述所需的最低和最高硬件配置等形式描述。

注： 附录A.11给出架构可扩展性的说明。

5.2.3 可迁移性

预训练模型提供方应声称预训练模型微调后，在人工智能医疗器械下游任务中的预期性能，并提供书面证据。

注：对于模型微调后的预期性能，附录A.10给出常用场景下的性能指标举例。

5.2.4 模型效率

预训练模型提供方应声称预训练模型效率，包括模型推理计算量、资源利用率、精度，并提供书面证据。可采用如下指标：

e） 预训练模型完成前向推理所需的计算量；

f） 预训练模型对算力和存储空间的占用率；

g） 如适用，相同架构的预训练模型的不同超参数（如参数量）配置下的精度和对应的推理时间。

5.2.5 输出重复性

预训练模型提供方应声称预训练模型输出的重复性，确保其在给定相同测试数据情况下，模型产生的输出是一致的。

注：输出一致是指输出内容的内在含义一致，若输出是量化指标，指标保持相同或处于同一值域范围；若输出为描述性文字，文字含义保持一致。

5.2.6 健壮性

预训练模型应声称预训练模型的健壮性，确保其在具有不同程度的多样性、与任务域存在偏差的数据集上产生正确输出的能力，该能力应满足如下要求：

h） 预训练模型应确保模型在输入数据有噪声时性能符合预期，产生正确输出。

i） 预训练模型应对训练集分布外数据点（Out-of-Distribution）具有拟合能力。

5.2.7 泛化性

预训练模型提供方应声称预训练模型的泛化性，并提供书面证据。如适用，应根据预训练模型预期用途和适用环境，对预训练模型研发使用的训练集与真实世界中陌生样本之间的差异进行分析。

5.2.8 对抗安全

预训练模型提供方宜声称预训练模型的对抗安全性。如适用，应提供案例说明模型处理的对抗攻击种类和面对该种对抗攻击的性能。

5.2.9 隐私保护

预训练模型提供方应声称预训练模型采用的隐私保护措施，满足如下要求：

a） 应采用合适的措施，如差分隐私处理，确保预训练模型不会因攻击产生训练数据的泄露，包括训练数据的分布以及对单个训练数据案例的推断。

b） 应确保对预训练模型代码所产生的数据上传和数据储存操作采用保护措施。

6 预训练模型质量符合性评价方法

6.1 通则

预训练模型质量评价包含对预训练模型说明、质量要求等的评价，由模型提供方提供待评价的预训练模型本体、说明文档等内容。

6.2 预训练模型说明评价

对预训练模型说明中的预训练模型描述内容的完整性、准确性进行检查，判断其结果是否符合第4章的要求；对于能证明已被广泛使用的预训练模型，预训练模型说明文档的形式根据实际情况确定。

6.3 质量特性评价

6.3.1 可训练性

根据预训练模型提供方所提供的训练用例和训练设置（包括超参数和适用环境），对预训练模型进行训练，通过记录损失函数值的收敛曲线等，判断其结果是否符合5.2.1。

6.3.2 架构可扩展性

调整预训练模型训练和部署环境的硬件资源和运行软件环境，记录预训练模型推理和训练的效率变化，判断其结果是否符合5.2.2。

6.3.3 可迁移性

根据预训练模型提供方所提供的训练用例和训练设置（包括采取的训练方法和数据增强方法），在下游任务域上测试并记录预训练模型的性能指标，判断其结果是否符合5.2.3。对于适用公开数据集进行性能测试的预训练模型，测试预训练模型对下游任务在公开数据集的性能，判断其结果是否符合5.2.3。

注1：性能指标的选择范围包括但不限于附录A.10中的适用指标。

注2：迁移训练主要是通过微调方法实现，例如大语言模型的低阶自适应微调、监督微调等。

6.3.4 模型效率

使用预训练模型对相应数据进行前向推理，按照以下条件判断其结果是否符合5.2.4。

a) 按照前向推理所需要的计算量进行前向推理，记录并判断是否运行正常。

b) 记录使用的算力和存储空间，判断是否符合声称。

c) 如适用，记录在不同参数量配置下的性能和推理时间，判断是否符合声称。

6.3.5 输出重复性

使用相同测试用例对模型进行前向推理测试，检查模型是否能够产生一致的输出，判断其结果是否符合5.2.5。

6.3.6 健壮性

在预训练模型适用的任务域中，对预训练模型所能处理的场景、样本形式的具体情况以及预期的性能进行评估，判断其结果是否符合5.2.6。

注：附录A.12给出供选取的健壮性的参考测试方法。

6.3.7 泛化性

使用模型适用的任务域中未包含在训练数据中的测试集，对未进行拟合的数据进行测试，判断其结果是否符合5.2.7。

注：附录A.12给出供选取的泛化性的参考测试方法。

6.3.8 对抗安全

对模型编写测试用例，验证其推理结果不会因误导性样本产生错误。使用黑盒或白盒方式产生欺骗性扰动，使用模型对这些添加扰动后的数据进行测试，验证模型是否能够抵御欺骗攻击，判断其结果是否符合5.2.8。

6.3.9 隐私保护

对预训练模型隐私保护措施进行检查，判断其结果是否符合5.2.9。

附 录 A
（规范性）
预训练相关要素的扩展说明

A.1 预训练模型

本文件中的预训练模型范围是人工智能医疗器械开发中使用的预训练模型，不特指针对医疗器械而研发的预训练模型。因此预训练模型也未限定只能从医疗器械制造商处获取的模型。使用方获取预训练模型的来源广泛，包括医疗器械厂家、第三方供应商、第三方服务平台、网络开源等均能获取合适的模型。

本文件对预训练模型的类型不做限定，该适用性涵盖了模型模态、参数规模等。本文件不对大语言模型进行单独的区分；本文件所声称的预训练模型，包括但不限于视觉预训练模型、大型语言模型（ Large Language Model，LLM）、多模态预训练模型等。

A.2 预训练模型源任务、下游任务和预训练模型的关系

预训练模型在源任务上进行预训练，通常能获得一定泛化能力。根据具体的下游任务对预训练模型进行微调，能获得针对具体下游任务的拟合能力。该种微调包含对模型超参数的调整、对模型重新训练，以及通过添加提示词的方式改变大型生成式语言模型功能等方式。

A.3 预训练模型提供方

本文件中预训练模型提供方是指将预训练模型提供给人工智能医疗器械开发方的责任方，在进行第三方测试时，模型提供方是将模型提供给第三方检测机构的责任方。预训练模型的开发方作为预训练模型提供方具有天然的优势，对于来源不明、开源、非医疗领域、预训练模型开发方不能提供明确预训练模型描述和特性等情况，人工智能医疗器械开发方要承担预训练模型提供方的角色，收集和完善预训练模型描述和特性等信息，承担相应的责任和风险。人工智能医疗器械的开发方在使用预训练模型开发人工智能医疗器械的过程中，根据任务和预训练模型质量特性慎重选择适当的预训练模型。

A.4 模型结构

本文件所声称的模型是以深度神经网络作为主要的技术路线进行描述，常用架构包括卷积神经网络、递归神经网络、循环神经网络、生成式对抗网络等类型。对于非深度神经网络的各类人工智能算法模型，如线性分类模型、支持向量机、决策树等，其结构描述方法各自有其习惯性方法。

A.5 模型参数

模型参数的元数据是指关于某数据的名字、意义、描述、来源、职责、格式、用途以及与其他数据的联系等信息。对于常用的深度神经网络，其模型参数的元数据包括神经元之间的连接权重、偏移、卷积核大小等。

A.6 数据抽象

特征是人工智能和模式识别研究领域的基本概念，指能表达模式本质的功能或结构特点的可度量属性，如尺度、纹理、形状等。好的特征能使同类模式聚集、不同类模式分离。例如，对于图像，能直接使用矩阵来表示，但为了降低处理的空间复杂度，对图像进行变换得到最能反映分类本质的特征。相应地，在一个模型中，特征提取是将一组原始数据缩减到较易管理的较小组的降维过程。从数据到特征的抽象是从输入数据中学习概括特征、消除不相干因素。随着特征通过神经网络模型层层前向传递，这种抽象的复杂性和深度逐渐增加。数据抽象程度定性地描述了数据被降维的程度，抽象程度可以包含未抽象的原始数据和已抽象的特征数据等。

A.7 数据编码方式

独热编码（One-Hot Encoding）是一种原始数映射为能被模型接受数据格式的编码方法，将分类数据转换为二进制向量，其中每个类别用唯一的整数表示，并且只有一个元素为1，其余为0，用于机器学习中处理分类特征。

标签编码（Label Encoding）是一种原始数映射为能被模型接受数据格式的编码方法，将分类数据转换为整数标签，其中每个类别被分配一个唯一的整数值，用于将分类特征转化为机器学习模型可以处理的数值形式，但不会创建二进制向量，因此可能需要引入类别之间的顺序关系。

序列编码（Ordinal Encoding）是一种将原始数映射为能被模型接受数据格式的编码方法，将分类数据转换为整数序号，其中类别之间存在一定的顺序关系。

频数编码（Count Encoding）：是一种将原始数映射为能被模型接受数据格式的编码方法，将分类数据转换为该类别在数据集中的频数或出现次数,该方法提供关于类别的频率信息。

目标编码（Target Encoding）：是一种将原始数映射为能被模型接受数据格式的编码方法，将分类数据转为该类别在目标变量上的平均值或其他统计信息,该方法能够捕获类别与目标变量之间的关联性。

A.8 深度学习可解释性

深度神经网络的可解释性目前仍然是研究难点，尚未形成公认的理论解释其本质原理。目前对可解释性的定义，一般包含如下内涵：

a) 深度神经网络的中间件能被人类理解和分析的程度；

b) 深度神经网络的空间输出与网络内的特征分量之间的对应关系；

c) 从深度神经网络的特征中提取有意义的信息的能力。

可解释性度量方法处于持续研究过程中，已有方法如：局部可解释性测试、使用偏移模型的输入信息对被测模型输出进行预测、测量预测输出和实际输出的偏移程度。

可解释性使深度神经网络模型操作和决策过程能够被理解和解释给人类用户或监管机构，这对伦理具有重要影响。可解释性强的模型，其内部逻辑和决策可追溯，有助于建立透明度和问责制。可解释性可以帮助识别和纠正模型中的潜在偏见和不公平性，有助于确保模型的决策不会对不同人群产生不公平的影响。

A.9 数据集建立过程规范

现有的国家标准和行业标准尚未对人工智能医疗器械使用的数据集质量管理进行规范。国外先进标准IEEE 2801-2022（Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence）提出一套推荐的质量管理体系，涵盖了数据集管理的全生命周期和影响数据集质量的关键因素。如参照该标准建立医疗人工智能数据集的质量管理规则，将有利于提高整体数据质量。

A.10 医学领域中的预训练模型的性能指标

A.10.1 图像分类

对于适用于图像分类任务的预训练模型，通常采取以下性能指标评估预训练模型在下游任务域上的性能表现。

d) 准确率（Accuracy）：正确分类的图像占总数的百分比，表达式见公式(A.1)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \cdots\cdots (A.1)$$

式中：

TP——真阳性：正确分类的阳性样本；

FP——假阳性：被错误归类为阳性的阴性样本；

TN——真阴性：正确分类的阴性样本；

FN——假阴性：错误分类为阴性的阳性样本。

e) 精确度（Precision）：真阳性预测数除以真阳性和假阳性预测数的总和，表达式见公式(A.2)

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{(A.2)}$$

式中：

TP——真阳性：正确分类的阳性样本；

FP——假阳性：被错误归类为阳性的阴性样本。

f) 召回率（Recall）：真阳性预测的数量除以真阳性和假阴性预测的总和，表达式见公式(A.3)：

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \text{(A.3)}$$

式中：

Recall——召回率。

TP——真阳性：正确分类的阳性样本。

FP——假阳性：被错误归类为阳性的阴性样本。

g) F1 分数：用于评估二分类问题的性能，其平衡了精确度和召回率，可进行更全面的评估，表达式见公式(A.4)：

$$\text{F}_1 = \frac{2\cdot(\text{P}\cdot\text{R})}{\text{P+R}} \quad \text{(A.4)}$$

式中：

F_1——F1分数。

P——精确度：阳性预测数除以真阳性和假阳性预测数的总和。

R——召回率：正确预测的阳性实例（真阳性）占所有实际阳性实例的比例。

A.10.2 图像分割

对于适用于图像分类任务的预训练模型，通常采取以下性能指标评估预训练模型在下游任务域上的性能表现。

a) IoU 系数：模型预测掩膜与基准真实掩膜之间的重叠程度的指标，表达式见公式(A.5)：

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}} \quad \text{(A.5)}$$

式中：

IoU——交并比。

TP——真阳性：正确分割的像素；

FP——假阳性：错误包含在分割中的背景像素；

FN——假阴性：在分割过程中漏掉的像素。

b) Dice 系数：Dice 系数的计算与 IoU 相似，量化预测掩码与真实掩码之间的重叠程度，表达式见公式(A.6)：

$$\text{Dice} = \frac{2\cdot\text{TP}}{2\cdot\text{TP+FP+FN}} \quad \text{(A.6)}$$

式中：

Dice——Dice系数。

TP——真阳性：正确分割的像素。

FP——假阳性：错误包含在分割中的背景像素。

FN——假阴性：在分割过程中漏掉的像素。

c) 像素准确度（Pixel accuracy）：评估正确分类像素的百分比，表达式见公式(A.7)：

$$\text{Pixel Accuracy} = \frac{TP+TN}{N} \quad \cdots\cdots (A.7)$$

式中：

Pixel Accuracy——像素准确度。

TP——真阳性：正确分割的像素。

TN——真阴性：正确分类的背景像素。

N——总像素数

d) 平均精度（mAP）：用于实例分割任务，结合了不同对象类别的精度和召回率，表达式见公式(A.8)：

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \quad \cdots\cdots (A.8)$$

式中：

mAP——平均精度。

n——数量：被评估对象类别的数量。

AP_i——类别平均精度 i。

A.10.3 文本处理

对于适用于文本处理的预训练模型，通常采取以下性能指标评估预训练模型在下游任务域上的性能表现。

a) 精确度（Precision）:正确预测为阳性的实例（真阳性）在所有预测为阳性的实例中所占的比例，表达式见公式(A.9)：

$$\text{Precision} = \frac{TP}{TP+FP} \quad \cdots\cdots (A.9)$$

式中：

Precision ——精确度。

TP——真阳性：正确预测为阳性的实例数量。

FP——假阳性：预测为阳性但实际为阴性的实例数量。

b) 召回率（Recall）：正确预测为阳性的实例（真阳性）占所有实际阳性实例的比例，表达式见公式(A.10)：

$$\text{Recall} = \frac{TP}{TP+FN} \quad \cdots\cdots (A.10)$$

式中：

TP——真阳性：正确预测为阳性的实例数量。

FN——假阴性：预测为阴性但实际为阳性的实例数量。

c) F1 分数：F1 分数是精确度和召回率的调和平均值，表达式见公式(A.11)：

$$F1 = \frac{2\cdot(P\cdot R)}{P+R} \quad \cdots\cdots (A.11)$$

式中：

F_1——F1分数。

P——精确度：正确预测为阳性的实例（真阳性）在所有预测为阳性的实例中所占的比例。

R——召回率：正确预测为阳性的实例（真阳性）占所有实际阳性实例的比例。

A.11 架构可扩展性

可扩展性主要体现的是模型对计算资源的支持能力。对预训练模型已经验证过的典型部署环境进行描述的主要内容，包含：

——指定算力规格及架构的单个人工智能服务器，能支持的最大模型参数和结构，包括服务器算力、集成的人工智能加速卡个数、服务器内加速卡间的互联方式、人工智能加速处理器类型等；

——部署能够支持的最大 AI 加速卡个数，以及对模型分布式并行切分的策略设计。

A.12 泛化性，健壮性测试参考方法

A.12.1 泛化性测试方法

a) 子群体组合测试法：将具有相似特征的案例分组，形成案例子群体，并将不同子群体的样本组合起来，形成多样化的测试集。在测试过程中，要求被测模型能对测试样本进行预测，且模型性能不在子群体间产生显著的统计偏差（具体统计方法根据任务进行适当选择）。

b) 压力样本测试方法：使用目标数据库中的非典型或难以分类的样本来测试模型性能。要求模型性能在难测样本中到达模型提供方说明的性能基线。

c) 模型提供方说明的其他测试方法。

A.12.2 健壮性测试方法：

a) 自然噪声样本测试方法：从数据库中选择自然噪声较大的样本，在这些大的噪声样本中测试预训练模型的识别精度，要求模型能产生正确的结果。

b) 不合格样品测试方法：从数据库中抽取其他类型的样本，以及在正确的样本中混入其他有明显错误的样本，用被测预训练模型进行判别，要求模型能够避免被误导而产生错误的结果。

c) 模型提供方说明的其他测试方法。

注：自然噪声是一个相对的概念，根据任务的目标和性能要求的不同，采用适当的指标来度量噪声，以及确定噪声的水平和阈值。

附 录 B
（规范性）
模型说明描述示例

B.1 肿瘤分割预训练模型

B.1.1 概述

肿瘤和器官的精准分割是影像辅助诊断、放射治疗中靶区和危及器官勾画的关键步骤。但是不同应用目的，甚至不同医疗机构、专家对于肿瘤和器官在影像上的边界认知不同，因此，建立肿瘤分割的预训练模型，再根据不同任务的目的进行微调，将成为快速、低成本、系统化、系列化、高质量进行相关研究、开发临床系统、研发人工智能医疗器械的可行技术路线。

B.1.2 预训练模型描述

训练模型训练框架为Unet，采用全监督方式进行模型训练，训练数据为CT模态的浮点型二维扫描数据，标签采用one-hot编码标注的人体表（Body）区域。

B.1.3 模型框架描述

B.1.3.1 模型结构

采用Unet网络结构，使用4层结构，包含一个输入层（Input）、多个编码器（Encoder）、多个解码器（Decoder）和一个输出层（Output）。输入层是整个网络的输入，也是第一个编码器的输入。每个编码器包含两个卷积层和一个最大池化层（Maxpooling）以实现下采样。每次下采样后，编码器输出的特征图维度和大小会下降，通道数会增加。每个解码器包含一个反卷积层（Deconv）、一个跳跃连接（Skip connection）和两个卷积层，以实现上采样。每次上采样后，解码器输出的特征图维度和大小会增加，通道数减少。编码器和解码器中的卷积层均由卷积运算（Conv）、批量正则（Batchnorm）、线性整流函数（Rectified Linear Unit，RelU）激活函数构成。输出层由卷积运算（Conv）组成，每个通道表示一个器官分割概率图。

B.1.3.2 模型节点

预训练模型中的节点共9.04M个，分布方式采用前馈结构，即由Encoder和Decoder组成。Encoder通过多个卷积层将输入图像逐渐降采样为低分辨率的特征图，而Decoder则通过反卷积和跳跃连接的方式逐渐恢复高分辨率的特征图。节点间连接方式主要为两种，一种是通过局部连接方式在Encoder/Decoder部分的上下层间连接，另一种是通过跳跃连接将Encoder的特征图与Decoder的特征图进行拼接。

B.1.3.3 模型参数

预训练模型由Input、Encoder、Decoder、Output模块组成，这四个模块的最小构成单元包含2维卷积（Conv2d）、ReLU激活函数、2维批量正则（BatchNorm2d）、2维最大池化（MaxPool2d）、2维随机失活（Dropout2d）五部分，其中Conv2d的卷积大小为3×3、Dropout2d的置零率为0.5、MaxPool2d的池化窗口为2。具体每层输出的数据维度和大小如下表所示。

注：“输出的大小”中的-1代表该维度随着预训练时的batch而变动。

表B.1模型结构及参数

主体结构	层名称	卷积大小	输出的大小
Input	Conv2d-1	3×3	[-1, 1, 256, 256]
	ReLU-2	-	
	BatchNorm2d-3	-	
	Dropout2d-4	-	
	Conv2d-5	3×3	
	ReLU-6	-	
	BatchNorm2d-7	-	
	Dropout2d-8	-	

上表（续）

主体结构	层名称	卷积大小	输出的大小
Encoder1	Conv2d-9	3×3	[-1, 32, 256, 256]
	ReLU-10	-	
	BatchNorm2d-11	-	
	Dropout2d-12	-	
	Conv2d-13	3×3	
	ReLU-14	-	
	BatchNorm2d-15	-	
	Dropout2d1-16	-	
	double_conv2D-17	3×3	
	MaxPool2d-18	-	
Encoder2	Conv2d-19	3×3	[-1, 64, 128, 128]
	ReLU-20	-	
	BatchNorm2d-21	-	
	Dropout2d-22	-	
	Conv2d-23	3×3	
	ReLU-24	-	
	BatchNorm2d-25	-	
	Dropout2d-26	-	
	double_conv2D-27	3×3	
	down2D-28	3×3	
	MaxPool2d-29	-	
Encoder3	Conv2d-30	3×3	[-1, 128, 64, 64]
	ReLU-31	-	
	BatchNorm2d-32	-	
	Dropout2d-33	-	
	Conv2d-34	3×3	
	ReLU-35	-	
	BatchNorm2d-36	-	
	Dropout2d-37	-	
	double_conv2D-38	3×3	
	down2D-39	3×3	
	MaxPool2d-40	-	
Encoder4	Conv2d-41	3×3	[-1, 256, 32, 32]
	ReLU-42	-	
	BatchNorm2d-43	-	
	Dropout2d-44	-	
	Conv2d-45	3×3	
	ReLU-46	-	
	BatchNorm2d-47	-	
	Dropout2d-48	-	
	double_conv2D-49	3×3	
	down2D-50	3×3	
	MaxPool2d-51	-	

上表（续）

主体结构	层名称	卷积大小	输出的大小
Encoder5	Conv2d-52	3×3	[-1,512,16,16]
	ReLU-53	-	
	BatchNorm2d-54	-	
	Dropout2d-55	-	
	Conv2d-56	3×3	
	ReLU-57	-	
	BatchNorm2d-58	-	
	Dropout2d-59	-	
	double_conv2D-60	3×3	
	down2D-61	3×3	
Decoder1	ConvTranspose2d-62	3×3	[-1,256,16,16]
Decoder2	Conv2d-63	3×3	[-1, 256, 32, 32]
	ReLU-64	-	
	BatchNorm2d-65	-	
	Dropout2d-66	-	
	Conv2d-67	3×3	
	ReLU-68	-	
	BatchNorm2d-69	-	
	Dropout2d-70	-	
	double_conv2D-71	3×3	
	up2D-72	3×3	
	ConvTranspose2d-73	3×3	
Decoder3	Conv2d-74	3×3	[-1, 128, 64, 64]
	ReLU-75	-	
	BatchNorm2d-76	-	
	Dropout2d-77	-	
	Conv2d-78	3×3	
	ReLU-79	-	
	BatchNorm2d-80	-	
	Dropout2d-81	-	
	double_conv2D-82	3×3	
	up2D-83	3×3	
	ConvTranspose2d-84	3×3	
Decoder4	Conv2d-85	3×3	[-1, 64, 128, 128]
	ReLU-86	-	
	BatchNorm2d-87	-	
	Dropout2d-88	-	
	Conv2d-89	3×3	
	ReLU-90	-	
	BatchNorm2d-91	-	
	Dropout2d-92	-	
	double_conv2D-93	3×3	
	up2D-94	3×3	
	ConvTranspose2d-95	3×3	

上表（续）

主体结构	层名称	卷积大小	输出的大小
Decoder5	Conv2d-96	3×3	[-1,32,256,256]
	ReLU-97	-	
	BatchNorm2d-98	-	
	Dropout2d-99	-	
	Conv2d-100	3×3	
	ReLU-101	-	
	BatchNorm2d-102	-	
	Dropout2d-103	-	
	double_conv2D-104	3×3	
	up2D-105	3×3	
Output	Conv2d-106	3×3	[-1, 16, 256, 256]
	ReLU-107	-	
	Conv2d-108	3×3	[-1, 1, 256, 256]
	outconv2D-109	3×3	

B.1.3.4 模型数据表示

a) 数据抽象程度
模型输入数据为从 CT 设备获取到的含有人体组织、器官的 DICOM 原始图像数据，并且 CT 拍摄范围需要包含进行模型训练的人体组织、器官，数据输入到模型前未进行特征变换和特征选择。

b) 数据类型
模型输入数据的可接受类型为整数、浮点数。

c) 数据维度和大小
模型可以接受的数据维度为 1，大小为 65536（256×256）。

d) 编码方式
输入的训练数据采用将真实世界中的亮度值转化为计算机中的数值强度进行线性编码，真实世界中亮度值越高则计算机中的数值越大。进行预训练时的标签数据采用独热编码（one-hot encoding）。

e) 数据处理
模型的输入数据大小会先变形到大小为 65536 的 1 维向量，然后将数据归一化到（0，1）区间。模型的输出数据需要从大小为 65536 的 1 维向量恢复到输入数据维度和大小。

B.1.3.5 模型可解释性

本模型属于典型的基于编码器-解码器结构的深度卷积网络。通过模型激活可视化，可以观察到本模型在编码器前端主要提取细粒度边缘和角点信息，随着编码器逐步对特征进行卷积，模型编码的语意特征为更加抽象的高维语意特征。

B.1.4 源任务数据描述

B.1.4.1 数据模态

数据模态为由CT设备扫描的CT图像。

B.1.4.2 数据量

预训练模型的数据总量为100例CT模态的数据。数据集采用7：2：1的比例进行构建，即7份数据做训练集、2份数据做验证集、1份数据做测试集。其中，训练集将参与预训练模型的训练，且数据会影响各神经元权重的调整。验证集参与预训练模型的训练，但不影响各神经元权重的调整，仅作为一个调整训练过程的精度指标。独立测试集不参与训练过程，它将用于测试预训练完成后的模型精度。

B.1.4.3 数据质量

预训练模型使用的CT图像，来自多个临床机构，数据质量控制如下：

a) 医院级别及数量
   1) 医院级别：二级以上；
   2) 医院数量：3 家以上。
b) 入选标准为：
   1) 18 岁以上接受放射治疗患者；
   2) CT 影像层厚≤5mm；术式：（左右侧）保乳、根治；
   3) 从分布上考虑患者年龄、术式，以确保纳入对象的年龄、术式分布贴近临床实况。
c) 排除标准：
   1) 扫描部位先天畸形或解剖结构异常；
   2) 患者体位不符合常见临床扫描标准；
   3) 伪影、假体、植入物等导致影像不清晰难以辨别；
   4) 影像不符合 DICOM 标准；
   5) 研究者认为不合适。
d) 数据采集规范
   1) 采集流程：签署合作协议、采集和脱敏、数据质控、数据转移；
   2) 采集设备：16 排以上螺旋 CT，层厚≥2mm；
   3) 采集参数：常规成像、符合 DICOM 3.0 标准、常规剂量和低剂量扫描、分辨率 512×512 及以上；
   4) 数据脱敏：保留性别、年龄、检查时间等信息，其他信息进行脱敏处理；
   5) 采集人员：影像专业医师或技师；
   6) 脱敏人员：采集医疗机构指定的医师、技师或信息专业工程师；
   7) 审核人员：住院医师及以上；
   8) 数据接收人员：企业指定人员，具有医学影像处理基础知识和影像解剖培训。
e) 数据清洗规范：
   1) 清洗对象：CT 图像；
   2) 清洗规则：检查脱敏、可读性、数据完整性、伪影等
   3) 人员要求：企业指定人员，具有医学影像处理基础知识和影像解剖培训。
f) 数据标注规范：
   1) 标注内容：人体器官和肿瘤病灶外部轮廓；
   2) 标注流程：采用背靠背双人标注和专家仲裁方式；
   3) 标注人员：影像或临床专业医师及以上；
   4) 仲裁人员：影像或临床副主任医师及以上。

B.1.4.4 基准性能

本模型在其原任务，即标准乳腺癌临床靶区（CTV）勾画任务中的测试集中，其勾画结果相较于基准勾画结果，Dice系数为0.871。

B.1.5 训练过程描述

B.1.5.1 学习方法

预训练模型采用全监督学习的方式进行训练。

B.1.5.2 训练设置

本模型预训练过程中采用水平、垂直翻转训练数据以及增加随机的高斯噪声方式进行训练数据的增强。使用随机权重作为模型的初始化权重值，训练过程中的优化器为随机梯度下降(Stochastic Gradient Descent,SGD)，数据的迭代量（batch）为60，初始学习率为0.001，学习率将会随着训练的轮次（epoch）动态变化，变化的规律为每20个epoch，学习率下降10%。

B.1.5.3 任务域

预训练模型的任务域为CT图像中的人体表（Body）分割任务。

B.1.6 模型适用性描述

B.1.6.1 适用数据模态

预训练模型仅支持CT扫描数据。

B.1.6.2 适用任务类型

预训练模型仅支持图像分割任务。

B.1.6.3 适用环境

运行预训练模型所需的推荐计算资源如下表。

表B.2推荐计算资源

部件	规格
中央处理器	Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz
内存	32GB
数据硬盘	5TB
显卡	GTX 1080-Ti-11GB
操作系统	Ubuntu 16.04.2LTS

B.1.6.4 下游任务训练数据需求

预训练模型迁移到新任务域时的数据集规模将会发生改变。举例说明，人体表（Body）分割任务的模型训练采用100例CT数据完成。对于新的任务域，如乳腺癌靶区分割任务的模型训练仅需要采用50例CT数据完成。

本模型在微调中无需使用特殊的训练方法和训练设置，本模型共计在新下游任务对应的数据集上完成了50个轮次的训练。

参 考 文 献

[1] IEEE Std 2801-2022 Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence.

[2] 计算机科学技术名词审定委员会.计算机科学技术名词（第三版）[M].科学出版社,2002.