**1. Title & Author Information**

* 标题：
* Predicting Malignancy of Pulmonary Nodules Across Hospitals Using Transferable Machine Learning on Small and Imbalanced Datasets
* 作者及机构
* 通讯作者邮箱
* 贡献说明（CRediT）

**2. Abstract（摘要）**

* **Background**：小样本、类别不平衡、域偏移在医学AI中的挑战
* **Methods**：提出预训练Transformer+TCA跨域适应框架
* **Results**：在两个真实医院数据集（A: 295条，B: 190条）上验证AUC/SEN/SPE
* **Interpretation**：模型具备可推广性和临床部署潜力

**3. Introduction（引言）**

**3.1 医学场景问题**

* 医疗AI部署的主要瓶颈：数据异质性、小样本、高偏移
* 多中心部署难以泛化

**3.2 方法学背景**

* Transformer 在表格数据的成功应用（预训练）
* TCA 作为经典、稳健的无监督域对齐方法

**3.3 技术挑战与动机**

* 表格基础模型在域外泛化不足
* 小样本条件下UDA方法稳定性差
* 医疗数据分布偏移 + 不平衡加剧模型崩溃

**3.4 研究目标与贡献**

* 集成预训练表格Transformer与TCA
* 提出TCA-enhanced Foundation Model
* 实证其在跨医院、小样本任务中的有效性
* 对比多个UDA方法，验证其稳定性与优势

**4. Methods（方法）**

**4.1 数据集与样本**

本研究使用来自中国两家权威肿瘤中心的临床结构化数据，构建了跨机构的训练-测试数据集框架，以评估所提方法在真实医学场景中的泛化能力。

**4.1.1 数据来源**

* **训练集（数据集A）**来源于中山大学肿瘤防治中心（Sun Yat-sen University Cancer Center, Guangzhou, China），共纳入 295 例孤立性肺结节（solitary pulmonary nodules, SPNs）患者，诊断时间为 **2011年1月至2016年12月**。
* **测试集（数据集B）**采自河南省肿瘤医院（Henan Tumor Hospital, Zhengzhou, China），共包含 **190 例SPNs患者**，时间范围为 **2013年1月至20xx年x月**，用于外部独立验证。

所有纳入患者均满足以下标准：① 胸部CT发现孤立性肺结节，且病灶直径 ≤ 3 cm；② 无肺外恶性肿瘤；③ 诊断前 7 日内具备完整的临床信息、CT影像和实验室检查结果；④ 最终诊断均基于经病理证实的组织学结果，包括经CT引导下经皮穿刺、支气管镜、胸腔镜或手术切除获得的病理报告。

**4.1.2 标签定义与任务目标**

每个样本均包含一个二分类标签（Label），表示肺结节的良恶性质：

* 1 表示**恶性结节（malignant）**
* 0 表示**良性结节（benign）**

本研究的核心任务为：构建一种能够在跨医院环境中鲁棒预测肺结节良恶性的二分类模型。

**4.1.3 特征变量**

* 数据集 A（中山大学肿瘤中心）包含 **63 个结构化特征**，全部可用于建模。
* 数据集 B（河南省肿瘤医院）包含 **58 个结构化特征**，为数据集 A 特征的子集，主要由于部分医院采集指标缺失。

所有特征涵盖了人口学信息、基础生命体征、血液生化、肿瘤标志物、影像征象等常见肺部肿瘤风险因素，具有较高的临床价值。在建模过程中，所有特征使用编号格式（Feature1–Feature63）进行匿名化处理，以避免敏感信息泄露。在论文撰写与可解释性分析中，参考特征对照表进行还原：

表 2. 特征编号与原始医学名称映射示例（节选）

|  |  |
| --- | --- |
| **Feature** | **Original Feature Name** |
| Feature63 | upper lobe |
| Feature2 | age |
| Feature46 | Lobe location upper |
| Feature61 | DLCO1 |
| Feature40 | CYSC |
| Feature56 | VC |
| Feature42 | CEA |
| Feature39 | CRE |
| Feature43 | NSE |
| Feature48 | diameter |

**4.1.4 样本分布与类别不平衡**

两个数据集中，样本标签存在一定程度的不平衡。在训练集A中，恶性样本约为良性样本的 **1.78 倍**，测试集B亦存在类似比例失衡。这种不平衡结构对模型的泛化与稳定性构成挑战，需在建模策略中加以考虑。

| **数据集** | **样本总数** | **恶性（Label=1）** | **良性（Label=0）** | **恶性占比** |
| --- | --- | --- | --- | --- |
| A（训练） | 295 | 192 | 103 | 65.1% |
| B（测试） | 190 | 122 | 68 | 64.2% |

**4.1.5 伦理声明**

所有数据均为历史电子病历记录，且已完成全面脱敏处理，不包含可识别个人身份的信息。所有患者在入院时已签署书面知情同意书，授权其临床数据用于科学研究。本研究获得中山大学肿瘤防治中心和河南省肿瘤医院伦理委员会批准，研究过程遵循《赫尔辛基宣言》的伦理规范。

**4.2 数据预处理**

为提升模型对跨机构医疗数据的泛化能力，我们设计了一套精炼、高效的数据预处理流程。该流程依托于预训练表格基础模型（Pre-trained Tabular Foundation Model）所提供的内置机制，能够在无需大量特征工程的前提下自动完成标准化、编码与分布对齐等操作。该模型具备结构先验与任务无关的预训练能力，在本研究中进一步基于源域数据进行有监督优化。整体预处理流程由以下三个关键模块组成。

**4.2.1 特征筛选：基于 Transformer 的递归特征消除**

我们采用基于模型输出的重要性评分，结合递归特征消除（Recursive Feature Elimination, RFE）方法，从原始 63 个特征中筛选出 8 个最具诊断价值的特征。所用模型为预训练表格基础模型，其在执行监督训练时会对每个特征的贡献度进行估计，并通过置换重要性迭代剔除低信息量特征。

最终选出的 **8 个特征子集（BEST\_8\_FEATURES）** 在源域和目标域中均表现出良好的稳健性与解释性，同时保留了关键的肿瘤标志物、生化指标及影像参数，为后续建模提供了简洁而有效的输入空间。

### **4.2.2 类别变量编码机制**

结构化医学数据中通常包含诸如性别、既往病史、病灶部位等非数值类别变量。在本研究中，所采用的预训练表格基础模型（Pre-trained Tabular Foundation Model）具备自动化、并行性与多样性处理能力，针对类别变量采用以下策略：

* **自动识别类别特征**：模型在预处理阶段会自动判断哪些特征为离散类别型变量，依据其唯一值数量与样本频率分布进行推断。
* **双路径编码策略（并行）**：在 32 个集成子模型中，不同子成员对类别特征采用不同的处理方式：
  + **序数编码（Ordinal Encoding）**：将每个类别映射为有序整数值（如 0, 1, 2...），保留类别间的可比性；
  + **数值视角转换（Numerical Interpretation）**：将类别特征直接视作连续型数值处理，跳过显式编码。
* **扰动机制（Diversity Injection）**：为提升集成鲁棒性，部分子模型在类别映射后引入顺序扰动（如随机打乱编码顺序），从而增加模型的表达多样性。

该机制在“生成 32 个集成推理子模型”时自动并行执行，每组子模型采用不同的类别处理路径与扰动策略，使得最终模型能更好适应多分布、异质性强的小样本医疗数据。

相较于传统独热编码（one-hot encoding）方法，该策略有效避免了维度膨胀问题，尤其适用于中小样本条件下的泛化建模任务。

### **4.2.3 模型内置数据预处理流程**

在模型训练与推理阶段，预训练表格基础模型会自动执行一套结构感知的数据预处理流程。根据图 4-3 所示，整体流程将针对每个子配置生成共 **32 个集成推理成员**，每个成员独立执行以下步骤：

#### （1）特征顺序重排（ShuffleFeaturesStep）

每个子模型首先对输入特征顺序进行**rotate式循环位移**，即通过位移重排方式构造新的特征排列结构。该策略可打破Transformer对特征位置信息的偏倚，有效提升模型结构的多样性与泛化能力。

#### （2）特征分布变换与降维（ReshapeFeatureDistributionsStep）

根据不同配置，子模型对输入特征执行如下两类变换策略：

* **配置 A（启用变换）**：对所有特征执行**分位数变换（Quantile Transformation）**，将原始数据分布映射为标准正态分布（𝒩(0,1)），以减弱偏态与异常值干扰。变换公式如下：

x′=Φ−1(F(x))

* 其中 F(x)为经验分布函数，Φ−1为正态分布反累积分布函数。此外，模型使用奇异值分解（SVD）对变换后的特征进行降维（保留20维），再与原始特征拼接形成最终输入。
* **配置 B（关闭变换）**：直接输入原始特征，跳过任何分布转换与降维操作。

该设计允许模型覆盖不同的特征分布假设，增强跨域稳定性。

#### （3）类别变量处理（EncodeCategoricalFeaturesStep）

每个子模型根据**类别变量编码机制**，选择使用序数编码或数值视角方式处理类别特征，并依据配置执行类别扰动。该步骤确保类别变量在不同子模型间的表示存在差异性，有助于集成结构充分捕捉特征表达的多样性。

#### （4）并行推理与集成融合

32 个模型成员将基于各自配置完成独立训练与预测，最终所有成员的输出通过以下方式融合为最终分类结果：

* **Softmax 激活 + 概率平均**：所有子模型预测的类别概率取平均值；
* **类别不平衡调整（逆频率加权）**：在融合阶段对预测概率进行类别权重修正，使用训练集类别频率的倒数作为加权系数，缓解类别不平衡对预测结果的影响。

最终输出为一组类别概率得分，经加权融合后生成模型预测结果。

**4.2.4 模型内置预处理流程总览**

图示

AI 生成的内容可能不正确。

**4.2.5 对比传统表格建模流程**

传统的表格数据建模通常依赖手动执行标准化、类别编码、缺失值填补等预处理操作，流程复杂且易引入偏差。而本研究所使用的预训练表格基础模型在设计上已集成这些处理步骤，实现了：

* 全流程自动执行，无需额外工程处理；
* 面向小样本、异质分布的结构感知建模能力；
* 与Transformer架构深度耦合的预处理逻辑，有效提升跨域稳定性。

该策略为临床AI系统在多机构部署与迁移学习中提供了更具现实价值的解决方案。

**4.3 模型架构**

## **4.3.1 表格预训练模型（**Pre-trained Tabular Foundation Model**）**

本研究采用一种预训练的表格Transformer模型作为基础分类器，用于从结构化医学数据中学习样本表示并进行跨域迁移预测。该模型基于**Per-Feature Transformer 架构**，具备结构感知、上下文自适应与集成推理能力，特别适用于小样本、高维度的医疗数据建模。

### **输入结构与Transformer编码**

模型以结构化表格样本作为输入，其中每一列特征被视作一个 token，构成输入序列。经特征嵌入与位置编码后，序列被送入Transformer编码器：

* **嵌入维度**：128
* **注意力头数**：4
* **前馈网络维度**：512
* **Transformer 层数**：10

该架构通过注意力机制建模各特征间复杂交互，形成高表达性的样本表示。

### **上下文学习机制（In-Context Learning）**

模型在推理阶段采用**上下文学习机制**：将训练样本与测试样本拼接为一个上下文序列输入Transformer，训练标签作为监督信号参与前向传播，测试标签留空。整个预测过程**无需梯度更新**，实现高效、高泛化能力的即时推理。

* 输入上下文维度为 [ntrain+ntest,1,f][n\_{\text{train}} + n\_{\text{test}}, 1, f]
* 使用参数 single\_eval\_pos 明确训练与测试边界
* 所有推理操作在 inference mode 下执行，避免引入训练偏差

### **集成预测机制与温度缩放**

为增强模型鲁棒性与泛化能力，系统生成**32 个集成子模型**，每个成员采用不同的特征排列、样本子集、类别扰动与预处理配置，形成结构多样性。

整体流程将针对每个子配置生成共 **32 个集成推理成员**，每个成员独立执行以下步骤：

#### （1）特征顺序重排（ShuffleFeaturesStep）

每个子模型首先对输入特征顺序进行**rotate式循环位移**，即通过位移重排方式构造新的特征排列结构。该策略可打破Transformer对特征位置信息的偏倚，有效提升模型结构的多样性与泛化能力。

#### （2）特征分布变换与降维（ReshapeFeatureDistributionsStep）

根据不同配置，子模型对输入特征执行如下两类变换策略：

* **配置 A（启用变换）**：对所有特征执行**分位数变换（Quantile Transformation）**，将原始数据分布映射为标准正态分布（𝒩(0,1)），以减弱偏态与异常值干扰。变换公式如下：

x′=Φ−1(F(x))

* 其中 F(x)为经验分布函数，Φ−1为正态分布反累积分布函数。此外，模型使用奇异值分解（SVD）对变换后的特征进行降维（保留20维），再与原始特征拼接形成最终输入。
* **配置 B（关闭变换）**：直接输入原始特征，跳过任何分布转换与降维操作。

该设计允许模型覆盖不同的特征分布假设，增强跨域稳定性。

#### （3）类别变量处理（EncodeCategoricalFeaturesStep）

每个子模型根据**类别变量编码机制**，选择使用序数编码或数值视角方式处理类别特征，并依据配置执行类别扰动。该步骤确保类别变量在不同子模型间的表示存在差异性，有助于集成结构充分捕捉特征表达的多样性。

每个成员独立输出预测结果后，所有输出经如下策略融合为最终预测：

* **Softmax后平均（probability averaging）**：对每个成员的输出 logits 先执行 softmax 运算，转换为类别概率分布，随后在类别维度上求平均：

y^=1N∑i=1Nsoftmax(zi)\hat{y} = \frac{1}{N} \sum\_{i=1}^{N} \text{softmax}(z\_i)

文本

AI 生成的内容可能不正确。

其中z\_i 为第 i 个集成成员的输出 logits，N=32。

* **温度缩放（Temperature Scaling）**：在 softmax 前对 logits 进行缩放：

zi′=ziTz\_i' = \frac{z\_i}{T}

图片包含 文本

AI 生成的内容可能不正确。

其中温度参数T > 0 控制输出分布的平滑程度。较高温度 TT 会生成更均匀的概率分布，减弱过拟合风险。

* **类别不平衡修正**：在输出阶段引入类别逆频率加权，提升罕见类别的预测敏感性。

医学数据中普遍存在标签分布偏斜问题，例如异常样本远少于正常样本。为此，模型在集成输出阶段引入**内置类别平衡机制**，自动调整输出概率以缓解不平衡带来的偏倚。该机制默认启用，并通过以下公式完成：

**重加权原理与公式**

设：

文本, 信件

AI 生成的内容可能不正确。

* p=softmax(z)∈RC\mathbf{p} = \text{softmax}(z) \in \mathbb{R}^Cp=softmax(z)∈RC 为模型输出的原始类别概率；
* π=(π1,π2,...,πC)\boldsymbol{\pi} = (\pi\_1, \pi\_2, ..., \pi\_C)π=(π1​,π2​,...,πC​) 为训练集中的类别频率向量，满足 ∑iπi=1\sum\_i \pi\_i = 1∑i​πi​=1；
* p^\hat{\mathbf{p}}p^​ 为加权修正后的输出概率。

则类别平衡修正过程为：

p^i=pi/πi∑j=1Cpj/πjfor i=1,2,...,C\hat{p}\_i = \frac{p\_i / \pi\_i}{\sum\_{j=1}^C p\_j / \pi\_j} \quad \text{for } i = 1, 2, ..., Cp^​i​=∑j=1C​pj​/πj​pi​/πi​​for i=1,2,...,C

其中：

* pip\_ipi​ 是模型原始预测的第 iii 类概率；
* πi\pi\_iπi​ 是训练集中第 iii 类的先验概率；
* 分母为归一化项，确保 ∑ip^i=1\sum\_i \hat{p}\_i = 1∑i​p^​i​=1。

这种 **逆频率加权（inverse-frequency reweighting）** 策略能有效提升少数类的模型敏感性，同时压制多数类的过拟合倾向。

### 4.3.2 Transfer Component Analysis（TCA）

Transfer Component Analysis（TCA）是一种基于核映射的无监督域自适应方法，旨在学习一个投影子空间，在该子空间中源域和目标域的分布尽可能接近，同时保持样本结构信息。该方法特别适用于医学领域中的小样本、特征稳态但分布偏移的问题，如跨医院诊断数据迁移。

#### 核映射与优化目标

设源域样本为 $X\_s \in \mathbb{R}^{n \times d}$，目标域样本为 $X\_t \in \mathbb{R}^{m \times d}$，总样本数为 $n + m$，TCA通过在核空间中学习一个变换矩阵 $W \in \mathbb{R}^{(n+m) \times k}$ 来进行域对齐。

其优化目标为：

图片包含 图示

AI 生成的内容可能不正确。

min​ tr(W⊤KLK⊤W)+μ⋅tr(W⊤KHK⊤W)

其中：

* $K$ 为联合核矩阵；
* $L$ 为拉普拉斯矩阵，衡量源域与目标域之间的分布差异；
* $H$ 为中心化矩阵；
* $\mu$ 为正则化系数。

#### 联合核矩阵构建（线性核）

在本研究中，我们采用 **线性核函数**：

K(xi​,xj​)=xi⊤​xj​  
文本, 信件

AI 生成的内容可能不正确。

联合核矩阵 $K$ 构建如下：

K=[KssKstKst⊤Ktt]K = \begin{bmatrix} K\_{ss} & K\_{st} \\ K\_{st}^\top & K\_{tt} \end{bmatrix}K=[Kss​Kst⊤​​Kst​Ktt​​]

文本

AI 生成的内容可能不正确。

* $K\_{ss} = X\_s X\_s^\top$
* $K\_{tt} = X\_t X\_t^\top$
* $K\_{st} = X\_s X\_t^\top$

联合后 $$ 是对称的。

#### 拉普拉斯矩阵 $L$

拉普拉斯矩阵 $L$ 用于衡量域间差异，定义为：

L=[1n2⋅1n×n−1nm⋅1n×m−1nm⋅1m×n1m2⋅1m×m]L = \begin{bmatrix} \frac{1}{n^2} \cdot \mathbf{1}\_{n \times n} & -\frac{1}{nm} \cdot \mathbf{1}\_{n \times m} \\ -\frac{1}{nm} \cdot \mathbf{1}\_{m \times n} & \frac{1}{m^2} \cdot \mathbf{1}\_{m \times m} \end{bmatrix}L=[n21​⋅1n×n​−nm1​⋅1m×n​​−nm1​⋅1n×m​m21​⋅1m×m​​]

文本

AI 生成的内容可能不正确。

其含义为：

* 正值部分维持域内结构；
* 负值部分鼓励跨域样本接近；
* 权重比例可自动调节源/目标/跨域结构保持强度。

#### 中心化矩阵 $H$

用于消除样本均值偏移：

H=I−1n+m⋅11⊤H = I - \frac{1}{n+m} \cdot \mathbf{1} \mathbf{1}^\topH=I−n+m1​⋅11⊤

手机屏幕的截图

AI 生成的内容可能不正确。

应用于核矩阵后：

Kcentered=HKHK\_{\text{centered}} = H K HKcentered​=HKH

文本

AI 生成的内容可能不正确。

使得每一行与每一列的均值为 0，聚焦于相对结构而非绝对值。

#### 解法：最小二乘 + 特征值分解

TCA的优化目标最终转化为线性系统求解：

(I+μKLK)S=KHK\left( I + \mu K L K \right) S = K H K(I+μKLK)S=KHK

文本

AI 生成的内容可能不正确。

对 $S$ 进行对称特征值分解：

S=UΛU⊤S = U \Lambda U^\topS=UΛU⊤

文本

AI 生成的内容可能不正确。

选取前 $k$ 个特征向量构成变换矩阵：

W=U[:,1:k]W = U\_{[:, 1:k]}W=U[:,1:k]​

文本

AI 生成的内容可能不正确。

#### 特征变换与预测

得到 $W$ 后，源域与目标域样本映射为：

Zs=KsW,Zt=KtWZ\_s = K\_s W, \quad Z\_t = K\_t WZs​=Ks​W,Zt​=Kt​W

图片包含 图形用户界面

AI 生成的内容可能不正确。

其中 $K\_s$ 与 $K\_t$ 分别为源/目标样本与联合样本之间的核矩阵（前 $n$ 行和后 $m$ 行）。随后将 $Z\_s$ 与标签 $y\_s$ 输入分类器训练（如 Logistic Regression），并将该模型用于 $Z\_t$ 上的预测。

#### 医疗数据中TCA的优势

* **解释性强**：线性核保持了特征间的线性结构，适用于血压、BMI 等连续医学指标；
* **稳定性好**：使用最小二乘避免了特征值不稳定性；
* **对小样本友好**：联合核空间维度为 $n+m$，在本项目中为 $485$，在不扩展样本数量的条件下学习了完整的跨域结构；
* **可扩展性强**：TCA作为预处理模块，可轻松集成至表格预训练模型框架中，实现“结构对齐 + 表格建模”的二阶段组合。

### 4.3.3 集成架构：TCA增强的表格预训练模型

本研究构建了一种支持端到端训练与推理的集成式表格建模框架。该框架由源域特征选择模块、TCA域对齐模块、预训练表格模型和后处理模块组成，分别处理训练与测试流程，如图所示。

图示

AI 生成的内容可能不正确。图示

AI 生成的内容可能不正确。

#### 训练阶段（Train Pipeline）

训练阶段输入为源域数据 $$，首先通过特征选择模块筛选出稳定且共享的输入变量，随后送入预训练表格基础模型（tabular foundation model）进行训练。模型输出类别概率后，进一步结合类别分布进行权重调整，生成最终预测标签 $$。整个过程支持 32 路并行建模与融合，具有较强的泛化能力与鲁棒性。

#### 预测阶段（Test Pipeline）

测试阶段输入为目标域数据 $$ 和源域数据 $$，首先经过相同的特征选择处理，随后通过 Transfer Component Analysis（TCA）进行子空间映射与域间对齐，得到投影后的目标域特征 $$。该特征随后送入与训练阶段结构一致的表格预训练模型中进行推理，最终输出预测标签 $$。

#### 端到端支持与跨域适应性

整个框架结构一致、模块解耦，支持端到端训练与部署。特征选择模块与 TCA 模块均可泛化到任意目标域，使得模型具备较强的跨医院适应能力。

**4.5 验证设计**

为评估所提出的Pre-trained Tabular Foundation Model+TCA跨域医疗分类方法的有效性，本研究设计了包含内部验证、外部验证和方法对比的多层次验证框架。

内部验证（数据A的10折交叉验证）

在源域数据集A上采用10折交叉验证策略评估模型的内在性能。具体地，将数据集A随机划分为10个等大小的子集，保持各子集的类别分布平衡。依次选择9个子集作为训练集，余下1个子集作为测试集，重复10次获得稳定的性能评估。该验证旨在评估Pre-trained Tabular Foundation Model在源域数据上的分类能力，为后续跨域迁移提供性能基准。

外部验证（train on A，test on B）

采用"训练于A，测试于B"的外部验证策略模拟真实的跨医院部署场景。在源域医院A的数据上训练Pre-trained Tabular Foundation Model+TCA模型，随后在目标域医院B的数据上进行测试评估。该验证策略能够直接反映模型在面临域分布差异时的泛化能力，是评估跨域医疗AI系统实用性的关键指标。

方法对比

设计多组对照实验验证域适应方法的有效性：（1）Pre-trained Tabular Foundation Model无域适应基线：直接使用Pre-trained Tabular Foundation Model在源域训练、目标域测试，不进行域对齐；（2）传统临床基线方法：包括PKUPH、Mayo临床评分和Paper\_LR等经典医疗评估模型；（3）机器学习基线方法：SVM、决策树、随机森林、GBDT和XGBoost等传统机器学习算法；（4）TCA域适应方法：所提出的Pre-trained Tabular Foundation Model+TCA集成方法。通过系统性的方法对比，量化分析TCA域适应在Pre-trained Tabular Foundation Model跨域医疗分类中的有效性和优势。

该验证设计通过内外部验证结合、多基线对比的策略，为所提出方法提供全面而严格的性能评估。

**4.6 评估指标**

为全面评估Pre-trained Tabular Foundation Model+TCA域适应方法的性能表现，本研究构建了包含分类性能指标、统计置信度评估、可视化分析和域距离度量的多维度评估体系。

分类性能指标

采用医疗AI领域常用的分类性能指标进行评估：（1）AUC（Area Under Curve）作为主要评估指标，反映模型的整体分类能力和临床应用价值；（2）准确率（Accuracy）评估模型的总体分类正确率；（3）F1分数平衡精确率和召回率，适用于医疗数据的不平衡特性；（4）敏感性（Sensitivity/Recall）评估模型识别阳性病例的能力，在医疗筛查中至关重要；（5）特异性（Specificity）评估模型正确识别阴性病例的能力。各指标通过10折交叉验证获得均值和标准差，确保结果的稳健性。

置信区间评估

采用95%置信区间（95% CI）量化性能指标的不确定性，通过Bootstrap重采样方法计算AUC指标的置信区间。该方法进行1000次Bootstrap采样，对每次采样计算ROC曲线和AUC值，通过计算2.5%和97.5%分位数获得置信区间边界：

$$

$$

其中$$和$$分别为1000次Bootstrap采样所得AUC值的第2.5和97.5百分位数，为临床决策提供统计学支持和风险评估依据。

可视化分析

构建三类核心可视化图表评估模型性能：（1）ROC曲线展示不同阈值下敏感性与特异性的权衡关系，直观比较各方法的分类性能；（2）校准曲线（Calibration Curve）评估预测概率与实际阳性率的一致性，验证模型输出概率的可靠性；（3）决策曲线分析（Decision Curve Analysis）从临床净收益角度评估模型的实用价值，结合医疗成本效益进行综合评估。

域适应效果评估

通过降维可视化和标准化域距离度量评估域适应的有效性：

（1）PCA/t-SNE可视化：展示域适应前后源域和目标域数据分布的变化，直观反映域对齐效果。

（2）标准化域距离度量：采用多种标准化距离指标量化域间差异。对于任意两个域的数据$$和$$，首先进行标准化预处理：

$$

$$

$$

$$

标准化Wasserstein距离：

$$

$$

标准化KL散度：

$$

$$

MMD距离：考虑到TCA算法采用线性核进行域对齐，MMD度量作为辅助评估指标采用RBF核计算，以提供不同视角的域差异评估：

$$

$$

其中$$为特征维数，$$为一维Wasserstein距离，$$和$$为标准化后第$$个特征的概率分布，$$为RBF核函数。

需注意MMD度量采用RBF核而TCA使用线性核，因此MMD主要用于从非线性角度评估域间复杂分布差异，与TCA的线性域对齐形成互补评估。通过比较域适应前后距离指标的变化，从多个角度定量评估TCA算法的域对齐能力。

该多维度评估体系从分类性能、统计可靠性、临床实用性和域适应效果四个层面全面评估所提出方法，为Pre-trained Tabular Foundation Model + TCA在跨医院医疗分类任务中的有效性提供充分的实证依据。

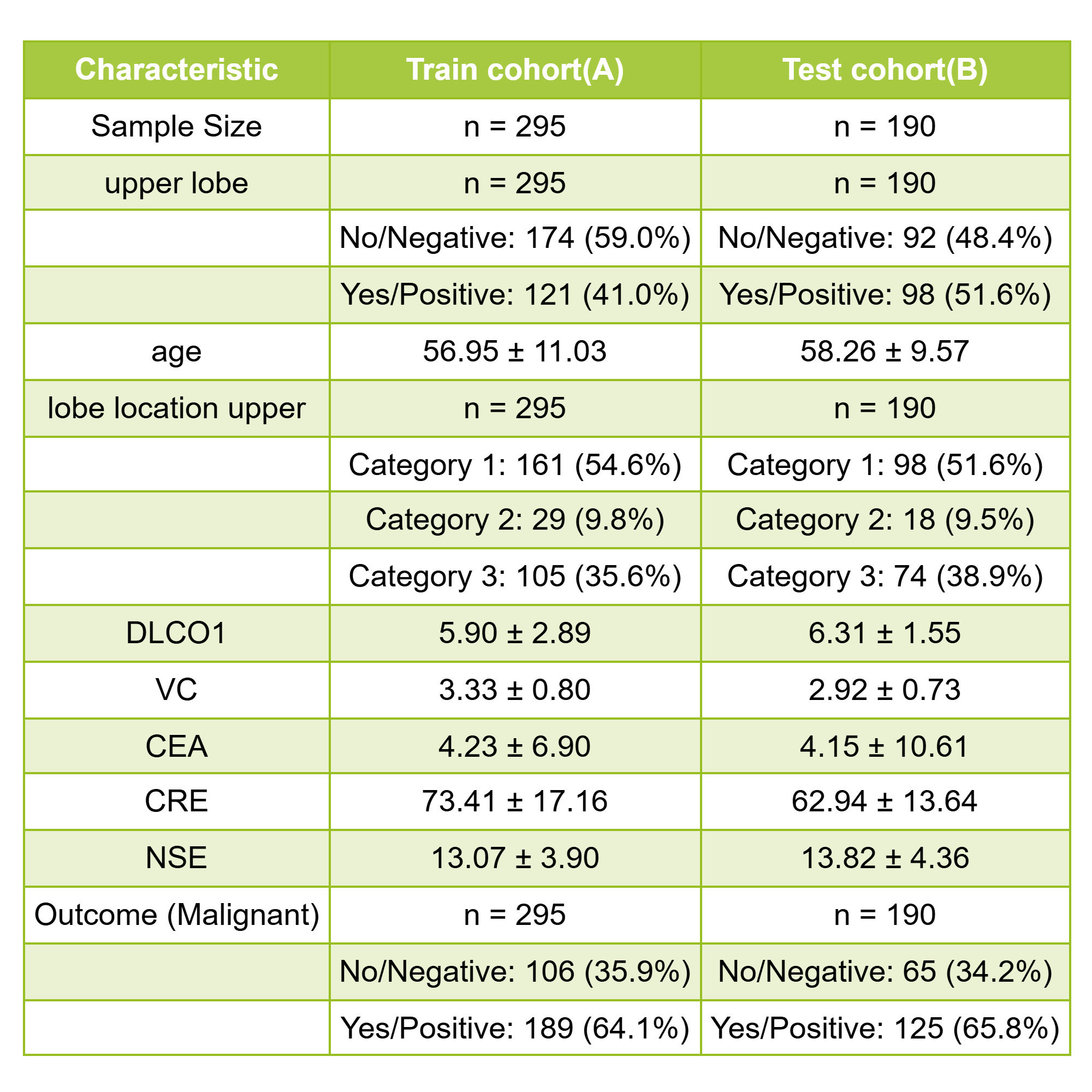
**5. Results（结果）**

**5.1 数据描述**

本研究使用两家医院的临床表格数据作为源域（训练集，cohort A）与目标域（测试集，cohort B），其统计特征如表 1 所示。为提升模型的泛化性能与稳定性，我们首先采用递归特征消除（Recursive Feature Elimination, RFE）算法对全部 63 个原始特征进行筛选。

最终，保留了最具判别力的 **8 个特征**，该特征集最初由 9 个特征构成，但由于目标域（数据集 B）中缺失 Feature40，因此在最终建模中将其排除。这些特征被统一用于训练与测试阶段，是本研究跨域迁移建模的基础。

训练集包含 295 个样本，测试集包含 190 个样本。两组患者在多个关键变量上的分布情况如下：

****

这些特征在两个域中具有结构一致性和适度的分布差异，为后续的 TCA 对齐和分类建模提供了良好的基础。

### 5.2 模型性能评估

* **5.2.1 内部验证性能（源域 10 折交叉验证）**

为评估各方法在源域（Train Cohort A）上的泛化能力，我们在训练集上执行了标准的 10 折交叉验证，并在多个指标上进行了性能统计，包括 AUC、Accuracy、F1-score、Precision 和 Recall。图中展示了各方法在五项指标上的平均表现，并按 AUC 值降序排列以突出全局性能最优的模型。

* **Our Model（Pre-trained Tabular Foundation Model）在五项指标中均表现出领先性能**，在 AUC 上达到 **0.826**，在准确率（Accuracy）为 **0.743**，F1 值为 **0.807**，精确率（Precision）为 **0.786**，召回率（Recall）为 **0.841**。综合来看，该模型在阳性识别能力和整体分类精度上均优于其他模型。
* **对比传统机器学习方法**：
  + **RF（随机森林）和XGBoost**分列第三、第四，在 AUC 上分别为 0.752 和 0.742，略低于 Our Model，但依然具有良好的稳定性。
  + **GBDT**、**SVM** 和 DT（决策树）在准确率和 F1 分数上略有下降，显示出在高维医疗数据上建模能力有限。
* **临床评分系统**（如 **MAYO** 和 **PKUPH**）的性能明显低于基于数据驱动的方法，尤其是 MAYO，在多个指标上表现不佳，F1、Precision 和 Recall 为 **0.000**，显示其在当前数据场景下缺乏有效预测能力。
* Paper\_LR（文献对比模型）虽然在 AUC（0.763）和 F1（0.810）上有良好表现，但在准确率（0.722）和 Precision（0.723）上略逊于 Our Model，说明其可能存在过拟合或类别分布偏移问题。

图表

AI 生成的内容可能不正确。

* **5.2.2 外部验证性能（跨域部署 A→B）**

为模拟实际医疗系统的跨机构部署场景，本文采用“训练于医院 A，测试于医院 B”的验证策略，系统性评估模型的跨域泛化能力。图中展示了包含 **Our Model**、其未进行域适应的版本（w/o TCA）、传统机器学习方法、以及临床评分系统在内的多种方法在目标域数据集 B 上的表现，并按 AUC 值降序排序。

结果表明，**TCA增强的预训练表格基础模型（Pre-trained Tabular Foundation Model + TCA）** 在所有评估指标上均取得领先成绩，其中 AUC 达到 **0.709**，F1 得分为 **0.811**，Recall 高达 **0.944**，显著优于未使用域适应的版本（AUC = 0.698）。该结果充分说明 TCA 显著增强了模型的跨域迁移能力，提升了对目标域正例的识别效果。

相较之下，**未进行域适应的预训练模型（w/o TCA）** 虽在源域上表现优异，但在目标域中性能明显下降，表现出典型的域偏移效应，验证了跨域场景下引入域适应机制的必要性。

在传统方法中，**Paper\_LR、RF 及 PKUPH** 模型相对稳定，AUC 范围在 **0.63–0.67** 之间。其中 Paper\_LR 在 Recall 上达到 **0.943**，接近 TCA 模型，但其 Precision（0.682）与 Accuracy（0.674）表现逊色，说明其倾向过度识别阳性，存在一定的性能偏移。

**SVM、GBDT、XGBoost 及决策树（DT）** 等机器学习模型整体跨域泛化能力有限，AUC 低于 0.63，尤其是 XGBoost（0.567）与 DT（0.509），对目标域分布变化敏感，难以适应分布漂移。

此外，**Mayo 临床评分系统** 几乎完全失效，多个关键指标（F1、Precision、Recall）为 **0**，说明其在目标域上完全丧失判别力，反映出传统评分机制在泛化性和适应性方面的显著不足。

综上，本文提出的 **TCA增强型预训练表格模型** 在跨域环境中展现出卓越的泛化性能，优于多种经典方法与经验规则，验证了其在实际多机构部署中的可行性与实用价值。

图表

AI 生成的内容可能不正确。

* **5.2.3 ROC 曲线分析与可视化对比**

图表, 散点图

AI 生成的内容可能不正确。

图中左侧为源域（医院 A）上 10 折交叉验证的 ROC 曲线，展示不同模型在内部验证中的敏感性与特异性权衡。可以观察到，**预训练表格基础模型（Our Model）** 在源域上 AUC 达到 **0.820**，显著优于其他传统模型（如 SVM：0.717，XGBoost：0.734，RF：0.752），说明其具备较强的拟合与判别能力。

右侧图为目标域（医院 B）上的跨域部署 ROC 曲线，涵盖了本方法、无UDA版本、传统机器学习模型及临床评分方法的对比结果。**加入 TCA 域适应的预训练模型（Our Model）** 在目标域中依然保持最高的 AUC（**0.709**），曲线整体明显位于其他方法之上，显示其在处理分布漂移下具有更强的稳健性和识别能力。

相比之下，**未使用 TCA 的模型（w/o UDA）** 虽在源域中性能较强，但在目标域中曲线明显下沉（AUC = 0.698），反映出典型的跨域退化现象。而传统临床评分系统 Mayo（AUC = 0.500）曲线几乎与随机分类器重合，完全丧失预测能力。

整体而言，ROC 可视化直观展示了各模型在不同域下的表现差异，进一步验证了 TCA 在提升跨医院模型泛化能力方面的有效性。

* **5.2.4 域适应效果评估**

为全面验证 TCA（Transfer Component Analysis）在提升跨域分类性能中的有效性，本文从 **可视化分布**、**域距离度量** 与 **指标提升对比** 三个角度展开分析。

#### （1）可视化分析：PCA 与 t-SNE

图表, 散点图

AI 生成的内容可能不正确。

图中左侧展示了源域与目标域在未进行域适应前的分布差异，两个域的样本在特征空间中中心分离；而在 TCA 变换后，目标域样本（橙色）明显趋于源域样本（蓝色）聚集，分布重叠度提升，尤其在 t-SNE 图中呈现出良好的融合趋势，表明 TCA 能有效学习跨域共享子空间，实现不同医院数据的表示对齐。

#### （2）域距离量化对比：MMD / KL / Wasserstein

图表, 瀑布图

AI 生成的内容可能不正确。

TCA 显著缩小了四种典型的分布差异度量值：

* Normalized Linear Discrepancy 降低了 **0.070**，
* MMD、Wasserstein 与 KL 散度分别下降了 **0.018**、**0.006** 与 **0.022**。  
  这些量化指标变化进一步证明，TCA 能有效减弱源/目标域间的统计偏移。

#### （3）性能指标提升：与无UDA基线对比

图表, 条形图

AI 生成的内容可能不正确。

图中柱状图展示了 TCA 模型与未进行域适应的 baseline（TabPFN without TCA）在目标域上的五项关键指标对比：

* Accuracy 提升 **+5.3%**（从 0.658 → 0.711）；
* F1 分数提高至 **0.811**，优于 baseline 的 0.767；
* Recall 高达 **0.944**，显著优于 baseline 的 0.856，展现出强大的敏感性；
* Precision 与 AUC 指标亦有不同程度提升（分别为 +1.6% 和 +2.8%）。

综上结果，从可视化聚类趋势、分布差异收敛性，到具体分类性能指标的全面提升，均证明 TCA 能有效缓解医学数据在跨机构部署中普遍存在的域偏移问题，增强模型在目标域的泛化能力与鲁棒性。

### 5.3 可解释性分析

图表, 折线图, 直方图

AI 生成的内容可能不正确。

为了提升模型的可解释性并降低过拟合风险，本文对原始特征集进行了递归特征消除（Recursive Feature Elimination, RFE）分析，以筛选出最具判别力的特征子集。**与传统方法使用支持向量机（SVM）作为基础模型不同，本文使用Pre-trained Tabular Foundation Model作为RFE的评估器，以更好地反映下游任务的真实表现。**

（1）特征重要性排序与选择依据  
我们在源域数据集（共 63 个特征）上运行 RFE 算法，根据特征对模型性能（以 AUC、Accuracy、F1 等综合指标）影响程度进行排序，并初步获得最优的前 9 个特征。由于数据集 B中缺少 特征40，为保证跨域一致性，我们最终保留了前 8 个特征用于后续训练。

（2）为何选择 8 个特征（而非更多）  
如图所示（“Model Performance vs Number of Features”），随着所选特征数量增加，模型的 AUC、Accuracy 和 F1 指标在 8~10 维附近达到最优；当特征数量超过 15 后，各项性能指标出现不同程度的震荡与下降，泛化性能变差。这验证了如下两点：

* **奥卡姆剃刀原则**：在保证性能前提下，使用更少特征能提升模型稳定性与可解释性。使用前 8 个特征时已达到：  
  AUC = 0.837  
  Accuracy = 0.780  
  F1 = 0.833  
  表现与全特征模型相当甚至更优。
* **避免过拟合风险**：虽然在训练集上选用更多特征（如 Top-11）可能取得更高性能，但在目标域验证中会显著下降，说明过拟合于源域噪声特征，降低跨域迁移鲁棒性。（是否要加上对比实验结果呢？这样加上对比实验结果符合实验规范吗？）

（3）结论  
因此，选择最优子集（8个核心特征）不仅提升了模型的泛化能力，也符合实际部署中对模型紧凑性与可解释性的要求。同时，这一策略在后续 TCA 迁移学习中也有助于提高稳定性与可控性。

**6. Discussion（讨论）**

**6.1 主要发现**

* 强化了TCA在小样本场景的域泛化能力
* 表格基础模型的潜力首次在跨域医学数据中系统验证

**6.2 临床价值**

* 可部署性强：泛化+稳定+轻量级
* 辅助决策系统的可行性

**6.3 技术创新**

* 预训练表格Transformer + TCA首次集成
* 提供了一种新型高效的无监督跨域学习路径

**6.4 局限性**

* 数据量少，标签质量可能存在主观偏差
* 未测试大规模中心（>5医院）

**6.5 未来方向**

* 联邦学习 + TCA集成
* 多源域适应（MSDA）扩展
* 数据拓展

**7. Conclusion（结论）**

* 提出一种TCA增强的表格预训练模型，适用于小样本跨域医学分类任务
* 实证其在真实医疗数据上的可推广性
* 为未来多中心AI部署奠定了方法基础

**8. Contributors**

**9. Declaration of Interests**

**10. Code & Data Availability**

**11. References**

**12. Supplementary Materials**