域偏移和标签噪声条件下组织病理图像常见不确定性估计方法的基准评测

arXiv:2301.01054v2 [eess.IV] 6 Jul 2023

预印本

Hendrik Mehrtens,[1] Alexander Kurz,[1]Tabea-Clara Bucher,Titus J.Brinker[2]德国癌症研究中心肿瘤数字生物标志物部（DKFZ），德国海德堡 {hendrikalexander.mehrtens},{alexander.kurz},{tabea.bucher}@dkfz-heidelberg.de titus.brinker@nct-heidelberg.de

2022年9月

摘要

近年来，深度学习在组织病理学应用中的使用日益增多。然而，尽管这些方法展现出巨大潜力，在高风险环境中，深度学习模型需要能够判断自身的不确定性，并在错误分类风险较高时拒绝输入。在本研究中，我们对用于全视野切片图像分类的常用不确定性与鲁棒性方法进行了严格评估，重点关注选择性分类任务，即模型应在不确定时拒绝分类。我们在图块级别上针对域偏移和标签噪声开展实验，并在切片级别进行实验。实验中，我们比较了深度集成、蒙特卡洛丢弃、随机变分推断、测试时数据增强，以及由后几种方法组成的集成。我们发现，方法集成通常能带来更好的不确定性估计，并增强对域偏移和标签噪声的鲁棒性；然而，与经典计算机视觉基准的结果相反，其他方法并未表现出系统性的增益。在各种方法中，拒绝最不确定的样本，能够可靠地显著提高分布内和分布外数据上的分类准确率。此外，我们还开展了实验，在不同标签噪声条件下比较这些方法。最后，我们公开了代码框架，以促进组织病理学数据不确定性估计的进一步研究。

关键词 深度学习、不确定性估计、鲁棒性、组织病理学、域偏移、标签噪声

1 引言

研究表明，深度神经网络（DNN）在许多医学任务中表现相当，甚至优于人类专业人员 [Esteva 等，2017，Haenssle 等，2018，Hekler 等，2019], 。尽管如此，它们在临床实践中仍鲜少得到应用。一个常见的原因是 [Begoli 等，2019，Kompa 等，2021，van der Laak 等，2021] 它们无法提供经过良好校准的预测不确定性估计 [Guo 等，2017], ，因此专业人员无法判断系统决策的可靠性；而在医学决策等高不确定性领域，这是必要条件。

理想情况下，经过良好校准的深度学习系统应能够判断并传达其对每项预测的确定性的准确估计，不仅告知人类从业者系统当前的可靠性，还能使

[1]贡献相同[2]通讯作者

系统自动拒绝输入 [Band 等，2021，Jaeger 等， 2023] 并将其转交给人类检查，从而提高系统的可靠性。

实践中的另一个问题是，基于深度神经网络的系统极易受到“域偏移”的影响，即训练和部署环境中的数据分布存在差异。在机器学习领域的相关文献中，[Hendrycks 和 Dietterich，2019年，Ovadia 等人， 2019] 注意到，即使图像受到轻微的人工扰动，也会使深度神经网络极易受到影响，导致分类性能急剧下降。在数字病理学领域，这类偏移十分常见 [Stacke 等人， 2021]。它们源于不同诊所采用了不同的数据采集流程，例如染色流程或扫描仪不同；患者特征（性别、年龄等）的分布变化也可能导致这类偏移。模型在新的数据环境中表现不可预测，这限制了人们对系统决策的信心，尤其是在高风险环境中，例如影响患者治疗方案的决策。

提升机器学习系统泛化能力的传统方法是对输入数据进行增强，这一方法目前正在数字组织病理学领域受到广泛研究 [Tellez 等， 2019]。近年来，已有多种新技术被提出，用于在域偏移下进行不确定性量化和泛化。然而，这些方法大多主要在计算机视觉和目标检测的一般场景中进行评估，针对其在数字病理学领域的研究则十分有限。

全视野切片图像（WSI）是深度学习面临的一项挑战，因为图像尺寸很大，通常达到十亿像素级，而且标注数据有限。此外，由于 WSI 的评估容易出错，且观察者间变异较大 [Karimi 等， 2020], 标签可能存在程度很高的标签噪声，因此纳入预测确定性至关重要。

[Linmans 等， 2020] 探讨了多头集成网络检测全切片图像中分布外（OOD）输入的能力。 [Thagaard 等， 2020] 比较了深度集成 [Lakshminarayanan 等， 2017]和蒙特卡洛 Dropout（MCDO） [Gal 和 Ghahramani， 2016] 在图块级估计预测不确定性的能力。与我们的研究独立且同期开展的是， [Linmans 等， 2023] 比较了组织病理切片上主流的不确定性估计方法，重点关注在乳腺和前列腺组织中检测异源组织的分布外检测。在 Jaeger [Jaeger 等， 2023] 关于如何对不确定性估计方法进行有意义且可比较的评估的见解基础上，我们研究了不确定性估计方法拒绝不确定预测的能力。我们的评估在图块级以及切片级进行，后者是临床上更相关的任务。

我们仔细比较了深度集成 [Lakshminarayanan 等、 2017], 蒙特卡洛 Dropout [Gal 和 Ghahramani、2016], 随机变分推断（SVI） [Blundell 等、 2015], 测试时数据增强（TTA） [Ayhan 和 Berens，以及这些方法的集成在域偏移下的鲁棒性。我们研究了基于不确定性的拒绝选项对模型性能的影响，观察到高确定性预测的准确率显著提高，并比较了多种根据预测估计最终确定性的指标。我们模拟了肿瘤标注边缘区域的标签噪声影响，并测量了各方法对这种人为引入的标签噪声的鲁棒性。

在实验中，我们使用了多个肿瘤组织病理学数据集。对于图块级实验，我们采用了 Camelyon17 [Bándi 等 2019] 数据集。Camelyon17 乳腺癌组织数据集共包含 50 张带有标注肿瘤区域的切片。这些切片采集自五家不同的诊所，使用了三种不同的扫描仪，从而能够在真实的域偏移场景下评估不同方法。

在我们的切片级实验中，我们使用来自癌症基因组图谱（TCGA）的组织切片。Liu 等人 [、 2018], 评估了该数据集的分子特征，并提供了每个样本的突变状态信息，例如微卫星不稳定性（MSI）。考虑到直接从全组织切片进行预测这一具有临床意义的任务，我们将不确定性估计应用于基于癌症基因组图谱（TCGA）结直肠数据集中的全切片图像进行 MSI 预测。

我们的主要贡献如下：

- 详细分析拒绝不确定样本在多大程度上能够提升图块级和切片级的分类性能。

- 系统比较在域偏移条件下最具代表性的不确定性估计方法，并从选择性分类、网络校准和组织病理学数据上的分类器性能等方面进行评估。

- 研究标签噪声对全切片图像分类的影响，以及所纳入的不确定性估计方法抵御标签噪声的鲁棒性。

- 发布一个易于扩展的代码仓库[1]，以促进针对深度神经网络的不确定性估计的进一步研究。

2 方法

在本节 中，我们介绍了本研究所采用的方法、不确定性度量和评估设置。

2.1 不确定性估计方法

在不确定性估计中，给定输入 $\mathbf{x}$ 和训练数据 $\mathcal{D}$，我们希望计算输出$\mathbf{y}$的后验预测分布。该分布可通过对模型参数 $\mathbf{w}$ 进行贝叶斯模型平均来表示：

$$p(\mathbf{y}|\mathbf{x},\mathcal{D}) = \int p(\mathbf{y}|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathcal{D})\,\mathrm{d}\mathbf{w} \tag{1}$$

然而，神经网络的后验预测分布在解析上难以处理。因此，近年来已有多种近似方法被提出。其他方法并非基于贝叶斯统计，但仍常用于量化预测不确定性。下面简要介绍几种最具代表性的方法。

随机变分推断（SVI）： [Blundell 等人 2015] 和 [Graves 2011] 通过在神经网络的每个参数上施加高斯分布来近似后验分布。作为近似目标，我们最小化估计的证据下界（ELBO）[Blei 等人， 2017]。我们采用 SVI 的 Flipout 形式[Wen 等人， 2018]，以稳定训练过程。

蒙特卡洛丢弃（MCDO）： [Gal 和 Ghahramani， 2016] 指出，丢弃操作原本是一种用于稳定神经网络训练的正则化方法，但也可以用来近似神经网络的真实后验分布。为进行近似，在推理过程中，对同一输入进行多次前向传播并启用丢弃层，然后将结果汇总。通过该方法得到的预测分布可视为从后验分布近似中抽取的样本。

深度集成：深度集成 [Lakshminarayanan 等人， 2017] 由多个架构相同、但使用不同随机初始化进行训练的神经网络组成。推理时，以所有集成成员预测结果的均值作为预测。 [Ovadia 等，2019，Ashukha 等人， 2020]指出，在校准以及域偏移下的鲁棒性方面，深度集成优于许多其他方法。

测试时数据增强（TTA）：与在推理时使用多个模型的深度集成不同，TTA 在推理时通过以不同方式增强输入，多次使用同一个模型。通常，推理时会采用与训练时相同的增强方式。 [Ayhan 和 Berens，2018年，Ashukha 等人， 2020] 指出，TTA 在鲁棒性和校准方面表现良好；由于只需训练一个模型，其性能可接近深度集成，同时训练时间更短。

集成变体：此外，我们还实现并测试了随机变分推断、蒙特卡洛 Dropout 和测试时增强方法的集成变体。

2.2 不确定性指标

根据前文介绍的方法生成的预测，已有文献提出了多种指标来估计预测不确定性。

校准领域中最常用的方法 [Guo 等， 2017] 是所谓的预测“置信度”，即 Softmax 输出的最大值。其核心思想是，远离决策边界的决策可以被视为确定的，而最不确定的决策则靠近边界；对于 $1/N$，边界位于此处，其中 N 表示类别数。其他常用的不确定性指标包括预测的熵 [Mobiny 等，2019，Band 等， 2021] ，以及类集成输出中各预测之间的方差 [Nair 等，2020，Leibig 等， 2017]。在 B 中，我们比较了这些方法，并得出结论：常用的置信度度量非常适合此任务。

[1] https://github.com/DBO-DKFZ/uncertainty-benchmark

2.3 评估设置

本节介绍不同的评估设置，以及各设置中使用的性能指标。我们在具有病灶级标注的数据上进行图块级评估，同时也进行切片级评估；在切片级评估中，我们仅使用切片级标签。

拒绝选项： 在临床场景中，模型应能够将不确定性较高的预测转交给人类专业人员进行评估，这称为选择性分类 Geifman 和 El-Yaniv [2017]。适用于此任务的模型应对错误预测赋予比正确预测更高的不确定性，从而通过在某一确定性水平设置阈值，剔除大量错误预测。为了比较多个模型在这方面的能力，我们计算准确率-拒绝率曲线 [Nadeem 等， 2009], 绘制模型达到的准确率与数据集中被拒绝数据点所占百分比之间的关系。我们还比较准确率-拒绝曲线的曲线下面积（$AUARC$）。Jaeger 等人建议采用这种评估方法[Jaeger 等， 2023], 他们对故障检测领域的不同评估实践进行了比较和评估。

校准： 为了衡量校准情况，我们采用期望校准误差（ECE） [Guo 等，2017，Nixon 等， 2019]。对于数据集中的每个数据点，模型都会给出一个预测；每个样本的输出概率或"置信度"平均而言应与该预测的正确性相符。换言之，我们预期置信度为 60% 的预测在约 60% 的情况下是正确的。为验证这一点，我们将预测划分为预先确定数量 M 的区间 B，每个区间的置信度范围相等。然后，将每个区间内平均准确率与置信度之间的绝对差值求和：

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \Big| \tag{2}$$

H其中， $|B_m|$ 表示第 m个分箱中的样本数， n 表示样本总数 s.

标签噪声： 医学标注通常含有未量化的标签噪声 [Joskowicz 等，2019，Jensen 等，2019，Karimi 等， 2020], ，这可能会降低监督式机器学习方法的性能。据我们所知，先前描述的方法尚未在医学领域针对标签噪声的鲁棒性进行比较。我们通过创建多个具有不同程度和类型标签噪声的数据集来评估标签噪声的影响，并在这些变化条件下评估各方法。

3 实验设置

3.1 数据集与数据处理

Camelyon17：我们在 Camelyon17 数据集的病灶级标注切片上开展实验 [Bándi 等，2019]。Camelyon17 数据集的这一部分包含 50 张乳腺淋巴结组织全视野图像，其中标注了转移性组织。这些切片取自荷兰的五家不同诊所，并使用三种不同的扫描仪获取，为评估诊所间域偏移对模型性能的影响提供了理想条件。

为了在分布内（ID）域和分布外（OOD）域之间引入分布偏移，我们将参与中心划分为两个不同的数据集。弱域偏移仅基于地理位置，因为分布外（OOD）数据集包含的扫描仪也出现在分布内（ID）数据集中。在这种划分下，中心 0、2 和 4 属于分布内（ID）数据集，分别包含 3D Histech、Hamamatsu 和 Philips 扫描仪；中心 1 和 3 则作为 OOD 数据集，这两个数据集均使用 3D Histech 扫描仪。

为了引入强域偏移，我们以一种方式划分各中心，使 OOD 数据集仅包含 ID 数据集中不存在的扫描仪，从而在图像采集过程中引入额外的技术偏移。为此，中心 0、1 和 3（使用 3D Histech 扫描仪）作为 ID 数据集，中心 2 和 4 则作为 OOD 数据集（分别使用 Philips 和 Hamamatsu 扫描仪）。

对于这两种划分，我们进一步将分布内数据划分为训练集、验证集和测试集。我们根据各中心切片所包含的标注肿瘤细胞面积对切片进行排序，并将中间的两张切片用作各中心的测试集。训练集和验证集通过随机划分图块生成，比例为 75%/25%。

图块本身按照 [Khened 等人， 2021] 的方法生成：对全切片图像的 HSV 饱和度分量进行中值滤波和大津阈值法处理，最后再应用开运算和闭运算膨胀。之后，提取尺寸为 256×256 的图块。切片上的肿瘤区域由多边形标注指示。利用这些标注计算每个图块的肿瘤覆盖率。肿瘤覆盖率超过 25% 的图块计为肿瘤图块，肿瘤覆盖率为 0% 的图块均计为非肿瘤图块。肿瘤覆盖率低于 25% 的肿瘤图块

则不纳入我们的标准训练，以尽可能降低因观察者间变异较大而产生标签噪声的风险 [Jensen 等，2019，Joskowicz 等，2019，Karimi 等，2020]。

癌症基因组图谱：在切片级分析中，我们使用来自癌症基因组图谱的结直肠癌（CRC）组织切片来预测微卫星不稳定状态，采用 [Kather 等，2019、Bilal 等， 2021]中所述的方法。遵循 [Kather 等， 2019],提出的流程，我们仅纳入标记为 MSI-H 和 MSS 的切片，并排除带有干扰标记或缺少分辨率信息的切片。因此，共有 322 张切片可用于训练和测试，其中 59 张切片标记为 MSI，263 张切片标记为 MSS。对于测试集，我们按地点定义域偏移，将提交中心纪念斯隆凯特琳癌症中心和罗斯威尔帕克提供的 60 张切片划入测试集。

3.2 训练设置与超参数调优

在图块级实验中，我们使用 ResNet-34 [He 等， 2016]，所有实验的批量大小均为 128，学习率均为 0.001。我们采用 Adam [Kingma 和 Ba，2017], 作为优化器；如果选定的交叉熵损失验证损失连续 3 个训练轮次未下降，则将学习率降低 10 倍。数据增强方面，我们遵循 [Tellez 等，2019] 的方法，进行随机裁剪至 224×224、随机旋转 90° 和颜色抖动（亮度：±20%，对比度：±30%，色相：±10%，饱和度：±10%）。输入数据使用训练数据的均值和方差进行归一化。根据验证集上的准确率选择最佳模型。训练期间，我们按类别样本数对训练集进行平衡；但不对验证集进行平衡，因为我们希望在所有可用数据上评估方法。

对于深度集成架构，我们按照近期文献 [Linmans 等，2020，Thagaard 等， 2020], 选取 $n = 5$ 个成员。对于蒙特卡洛 Dropout，我们在每个残差网络块之后放置一个丢弃层，其丢弃概率为 $p = 0.3$。这与 [Linmans 等，2020，Thagaard 等， 2020], 的做法形成对比：他们只在最后一层之前放置丢弃层，并观察到性能没有提升。在测试推理时，我们使用 10 个随机变分推断、蒙特卡洛 Dropout 和测试时增强样本。按照 [Wenzel 等， 2020], 的做法，我们增加一个超参数，用于权衡随机变分推断先验（与正态分布的 Kullback-Leibler 散度）对训练的影响。我们使用 Python 库 Optuna [Akiba 等，2019] 对该超参数、丢弃概率、丢弃层的位置以及学习率进行调优。

对于切片级实验，我们使用开源的 CLAM 方法 [Lu等人， 2021]。CLAM 流程包含两个核心部分：首先，使用在 ImageNet [Deng等人，2009], 上预训练的 ResNet-18 [He等人， 2016] 提取图块级特征，使每张切片都表示为一个图块级特征包。第二步，应用基于注意力的机制 [Ilse等人， 2018] 计算切片级预测。为训练该方法，我们采用 [Lu等人， 2021], 所述的相同设置，将切片级交叉熵损失与图块级 SVM 聚类损失相结合。在提取测试切片后，我们将剩余切片中的 80% 用于训练，20% 用于验证。我们确保 MSI 和 MSS 两种标签在所有数据划分之间均匀分布，并在训练时对两个类别使用加权采样。我们将初始学习率设置为 2^{-4}，通过提前停止监控验证损失，并将训练轮次的最小值设为 50、最大值设为 200。该方法的丢弃概率设置为 $p = 0.25$。在不确定性评估中，我们训练包含 $n = 5$个成员的深度集成；对于蒙特卡洛 Dropout，则在推理时运行 10 次。

我们的所有实验均使用 PyTorch 1.11，在搭载 CUDA 11.3 的 Nvidia GPU 上进行。

4 结果

我们训 练每种方法 5 次，每次使用不同的随机种子，并报告其平均性能或中位性能。

4.1 拒绝选项

我们根据不确定性方法检测错误预测的能力对其进行比较。为此，我们计算了（平衡）准确率-拒绝曲线（见第 2.3 节）。在图1中，对于弱域偏移和强域偏移下分布内中心和分布外中心的测试切片，我们比较了各方法在拒绝图块比例逐渐增加时的表现。为了确定被拒绝的最不确定图块，我们使用置信度度量。在弱偏移设置的分布内中心上，除测试时增强集成外，所有方法的表现都相近且良好；测试时增强集成虽然以更高的平衡准确率起步，但却是唯一一个随着拒绝率增加而未表现出平衡准确率单调上升的方法。在分布外中心上也可以观察到类似趋势：在拒绝超过 50% 的不确定图块后，测试时增强的表现趋于停滞。

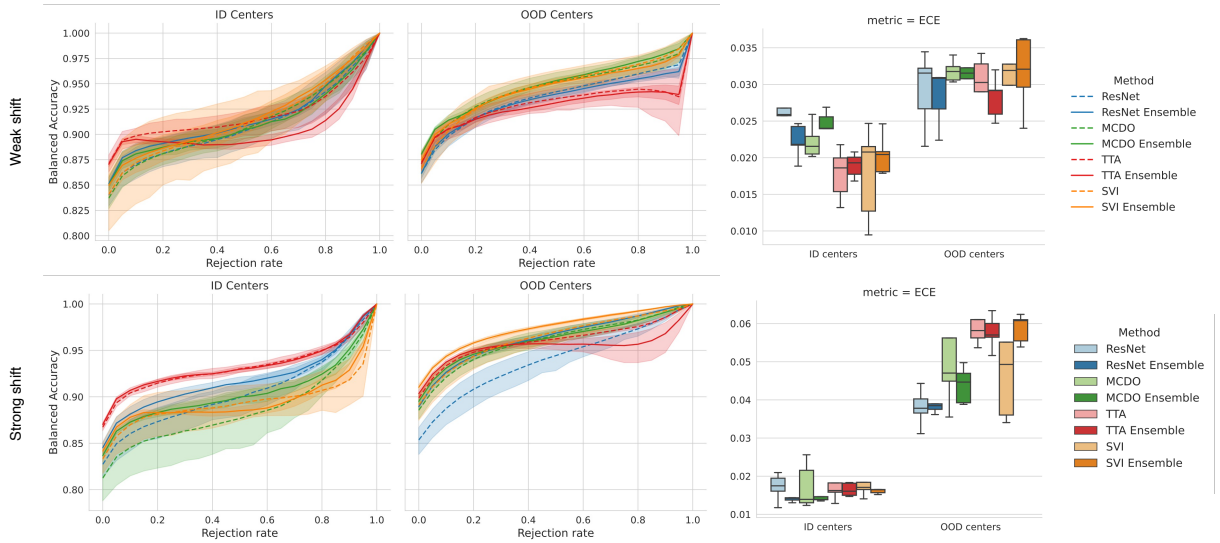


图1：弱域偏移和强域偏移设置下，分布内和分布外中心的（平衡）准确率-拒绝曲线，以及校准误差。横轴表示数据集中被拒绝输入所占的比例，并与剩余图像上达到的平衡准确率进行对比。被排除点的顺序由预测置信度决定（见 B）。我们绘制了全部 5 次试验的平均曲线，阴影区域表示 95% 置信区间。右侧图中，我们比较了不确定性估计方法对网络校准的影响，指标为分布内和分布外中心所有测试集切片的期望校准误差中位数。

表1：图1中所示准确率-拒绝曲线的曲线下面积的均值和标准差（$AUARC$）。

偏移	方法	ID 中心	分布外中心
Weak	残差网络	$0.9105_{\pm 0.0190}$	$0.9375_{\pm 0.0094}$
	残差网络集成	$0.9169_{\pm 0.0071}$	$0.9354_{\pm 0.0044}$
	MCDO	$0.9124_{\pm 0.0134}$	$0.9470_{\pm 0.0112}$
	蒙特卡洛 Dropout 集成	$0.9126_{\pm 0.0104}$	$\mathbf{0.9494}_{\pm 0.0034}$
	TTA	$\mathbf{0.9200}_{\pm 0.0140}$	$0.9305_{\pm 0.0169}$
	测试时增强集成	$0.9051_{\pm 0.0053}$	$0.9283_{\pm 0.0056}$
	SVI	$0.9156_{\pm 0.0329}$	$0.9467_{\pm 0.0148}$
	随机变分推断集成	$0.9183_{\pm 0.0133}$	$0.9459_{\pm 0.0051}$
强	残差网络	$0.9057_{\pm 0.0169}$	$0.9402_{\pm 0.0150}$
	残差网络集成	$0.9184_{\pm 0.0070}$	$0.9639_{\pm 0.0043}$
	MCDO	$0.8864_{\pm 0.0303}$	$0.9609_{\pm 0.0093}$
	蒙特卡洛 Dropout 集成	$0.9048_{\pm 0.0075}$	$0.9625_{\pm 0.0070}$
	TTA	$0.9321_{\pm 0.0047}$	$0.9607_{\pm 0.0024}$
	测试时增强集成	$\mathbf{0.9332}_{\pm 0.0024}$	$0.9547_{\pm 0.0095}$
	SVI	$0.8948_{\pm 0.0269}$	$0.9643_{\pm 0.0110}$
	随机变分推断集成	$0.8958_{\pm 0.0058}$	$\mathbf{0.9739}_{\pm 0.0015}$

相比之下，在强偏移设置下，TTA 和测试时增强集成在 ID 中心的整条曲线上均取得最高的平衡准确率，而蒙特卡洛 Dropout 和随机变分推断方法的表现最差。在分布外中心，排名大多恰好相反：随机变分推断集成的准确率-拒绝曲线最高，而基线残差网络和测试时增强方法表现最差。令人惊讶的是，分布外中心的平衡准确率有一部分高于 ID 中心。在 A 中，我们将详细阐述这一分布外数据上的现象。此外，我们还使用测试数据集中所有切片的中位数平衡准确率来评估并比较这些方法的分类性能。

表1展示了弱域偏移和强域偏移场景下准确率-拒绝曲线的曲线下面积（$AUARC$）。对于弱域偏移的 ID 中心，除测试时增强集成外，所有方法相较于基线残差网络都提升了 $AUARC$，其中测试时增强、随机变分推断集成和残差网络集成表现最佳。然而，在强偏移的分布内数据上，所有随机变分推断和蒙特卡洛 Dropout 方法的表现都更差，或与

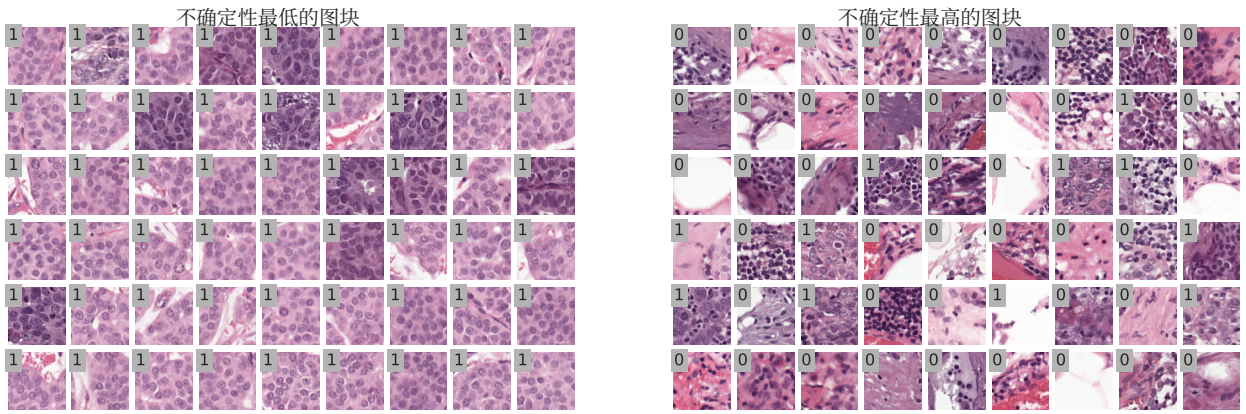


图 2：基于残差网络集成的置信度计算得出的分布内测试集中最不确定和最确定的图块。数字表示真实标签：“1”表示肿瘤，“0”表示非肿瘤图块。

ResNet 基线相当。在这里，只有两种测试时增强方法和残差网络集成显著优于 ResNet 基线。通过使用 $AUARC$ 指标，我们可以确定，在弱偏移设置的分布外中心上，蒙特卡洛 Dropout 集成表现最佳；而在强偏移设置的分布外中心上，随机变分推断集成表现最佳，同时测试时增强方法在弱偏移的分布外数据上表现不佳。在强偏移的分布外数据上，所有方法都优于基线。

根据我们的实验，没有哪一种方法能被认定为在所有数据划分中表现最佳，因为各方法的排名在分布内数据与分布外数据之间，以及在弱偏移设置与强偏移设置之间都有很大差异。不过，在域偏移场景下，集成方法的表现通常优于单一方法。附录 C 中提供了按中心进行留一设置时不确定性方法排名的进一步分析，得出了相同结论。

图 2 展示了在强偏移设置下，使用残差网络集成计算得到的分布内测试数据中置信度最高和不确定性最高的一组图块。我们观察到，神经网络似乎对肿瘤图块（标签 1）最有把握；这些图块被肿瘤组织覆盖，并具有相似的细胞结构。对于右侧不确定性最高的图块，则看不出它们之间存在可比的结构。这些图块既包含肿瘤组织，也包含非肿瘤组织，而且往往位于不同组织类型之间的边界区域。许多不确定性较高的图块似乎位于标注的肿瘤区域边界处；我们怀疑这些区域的标签噪声程度更高，而且由于健康组织和肿瘤组织中都存在某些特征，因此更难区分。

4.2 校准

在图 1 右侧的图表中，我们使用期望校准误差（ECE；见第 2.3 节）评估弱偏移和强偏移下的模型校准。按照 A 中的说明，ECE 值是对所有切片的校准误差取中位数得到的。在弱偏移设置的 ID 中心上，测试时增强和随机变分推断方法表现尤为出色。在强偏移设置的 ID 中心上，表现最好的方法是残差网络集成和蒙特卡洛 Dropout 集成，而测试时增强和随机变分推断方法的表现较差。

在分布偏移下，表现最佳的实际上是 ResNet 基线方法及其集成变体。其他所有方法在弱偏移和强偏移设置下的表现都更差；测试时增强集成在弱偏移下表现仍相当不错，但在强偏移设置下表现最差，与随机变分推断集成并列。

总之，与呈现的其他指标类似，我们无法确定表现最佳的方法。只有集成方法几乎始终优于对应的单一方法。表2列出了校准实验的详细数值结果。

4.3 标签噪声

为了将图块级观察结果转换为切片级结果，我们将图块级预测拼接回切片级肿瘤置信度图，并在图3中展示了一张 Camelyon17 切片示例。观察生成的置信度图时，我们可以看到，标注肿瘤区域边界处的肿瘤置信度较低。

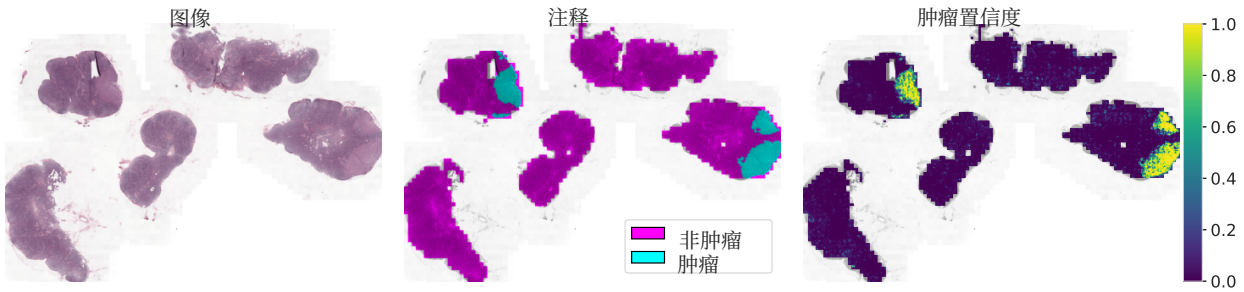


图3：带有真实标注和模型预测的全视野图像（患者_017_节点_2）。预测由残差网络集成生成。靠近标注边界的区域中，肿瘤置信度会降低，而不确定性则会增加。这一结果在整个数据集中都一致。

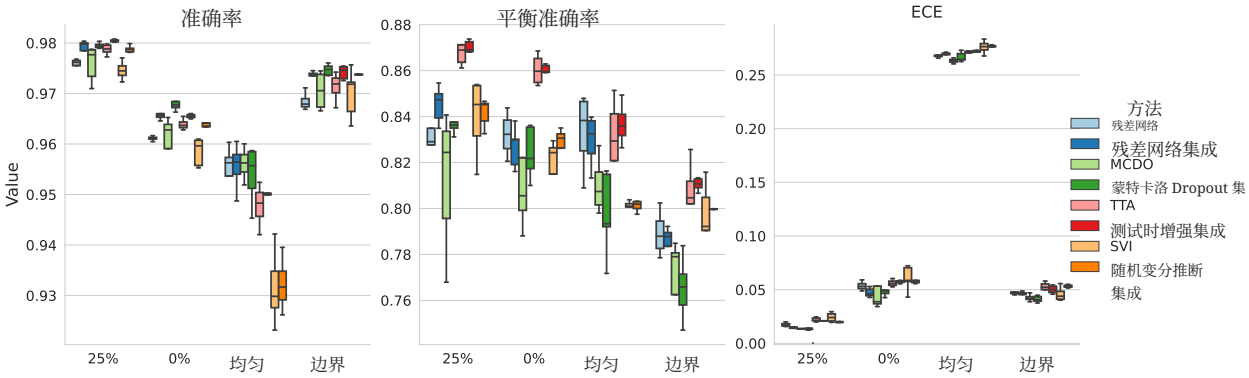


图4：所提出的方法在标签噪声条件下于不同数据集上的表现。我们比较了原始数据集（25%）、肿瘤图块的肿瘤覆盖率阈值为0%的数据集（0%）、对图块标签施加均匀噪声（均匀），以及随机翻转位于标注边界处的图块标签（边界）时的表现。

基于这些观察结果，在标签噪声实验中，我们研究了标注肿瘤区域边界区域内不精确肿瘤标注的影响。为此，我们以强偏移场景的分布内 Camelyon17 数据集为基础，通过向标注引入不同类型的标签噪声，构建了三个补充训练数据集。首先，我们创建一个数据集，将肿瘤图块的肿瘤覆盖率纳入阈值设为 0%。此前，为了降低标签噪声的可能性，我们排除了所有被肿瘤标注覆盖不足 25% 的图块（见第 3.1 节）。另外两个数据集则通过对 0% 阈值数据集的训练划分施加随机标签噪声来创建。首先，我们对整张切片施加均匀标签噪声，以 25% 的概率翻转图块类别。由于这类标注噪声不能反映现实世界中的观察者间变异，我们接着对标注的边界区域进行标签翻转。我们翻转位于标注多边形边界处、因而未被标注的肿瘤细胞完全覆盖的肿瘤图块的标签。我们将每个图块发生这种情况的概率设为 25%。

图4展示了标签噪声实验的结果。详细结果见表3。所有方法均在带噪声的数据集上训练，但在原始测试数据集上评估。就准确率而言，集成方法显著优于相应的非集成方法。只有在均匀噪声数据集上，ResNet 和蒙特卡洛 Dropout 的表现与各自的集成方法相近（每种方法的准确率均为 $\sim 95.6\%$）。从平衡准确率指标来看，测试时增强和测试时增强集成大幅领先于其他所有方法。集成方法再次始终优于相应的单一方法。面对标签噪声时，蒙特卡洛 Dropout 和随机变分推断是最不稳健的方法，其表现往往不如单一的 ResNet 基线。

我们可以得出结论：集成方法不仅对域偏移和图像损坏更加稳健 [Ovadia 等人， 2019] 同样也能更好地应对标签噪声，在本例中即组织病理学图像中的标签噪声。然而，根据我们的实验，SVI 和蒙特卡洛 Dropout 并不适合处理标签噪声，通常只能带来略微改善，甚至会导致更差的结果。

然而，就校准误差而言，测试时增强表现不佳。与近期文献的结论相反，蒙特卡洛 Dropout 优于测试时增强和随机变分推断；后两者产生的校准预测整体最差 [Ashukha 等，2020，Ayhan 等人，

2020]。可以看到，与基线数据集相比，均匀标签噪声数据集上的期望校准误差大幅增加，另外两个带有标签噪声的数据集上的校准失准则略有增加。除原始数据集外，看不出集成方法能够降低校准误差的趋势。面对更大量的标签噪声时，集成似乎并不能改善校准；这与域偏移的情形（图1）相反，在该情形下，集成降低了校准误差。

4.4 切片级分析

对于切片级分析，我们旨在按照第 3.1 节所述，利用癌症基因组图谱（TCGA）数据集中的结直肠癌组织全切片图像（WSI）预测微卫星不稳定（MSI）状态。此前，我们已在图块级评估了常见不确定性方法的性能，因此首先尝试利用图块级不确定性推断切片级预测。 [Linmans 等人 2023]在同期开展的研究中提出了类似方法，用于前列腺癌的切片级预测。为训练图块级方法，我们首先使用亚型识别网络识别每张全切片图像中的肿瘤图块。随后，将对应全切片图像的二元切片级标签（微卫星稳定（MSS）或微卫星不稳定（MSI））赋予每个肿瘤图块，并训练由五个 ResNet-34 组成的集成模型，与此前的图块级评估相当。为计算切片级预测，我们对置信度最高的 1% 图块级分数对应的预测取平均。通过采用所述方法进行五次不同的运行，我们得到了测试集上的 AUROC 值及其均值和标准差 0.6243 ± 0.0610 。

除了聚合图块级预测之外，在全视野切片分类中还有更复杂的方法，其中之一是 CLAM 方法 [Lu等人， 2021]。我们在相同的数据划分上训练并评估了 CLAM 方法，5 次运行的 AUROC 值为 0.7168 ± 0.0328。性能显著提升，因此可以得出结论：与通过不确定性聚合图块级预测相比，基于注意力的 CLAM 方法更适合切片级预测任务。

继图块级实验之后，我们评估了域偏移下切片级选择性分类的能力。为此，我们在性能更优的 CLAM 方法之上实现了深度集成和蒙特卡洛丢弃。图 5 展示了切片级预测任务的准确率-拒绝曲线。从这些曲线中，我们可以

可以看到，基于不确定性拒绝切片能够提升性能，其效果与图块级实验相当。我们还观察到与图块级评估类似的趋势：方法集成的表现优于单一模型。

总的来说，我们可以得出结论：基于不确定性拒绝切片级预测是一种可行的选择性分类方法，并且能够显著提高预测准确率。

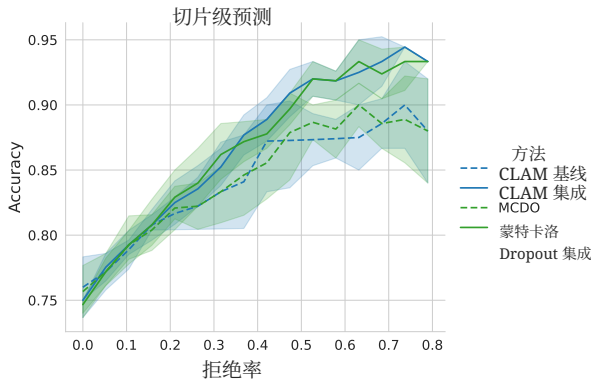


图 5: 准确率-拒绝率 使用 t 绘制的切片级微卫星不稳定性预测曲线 he CLAM 方法 [Lu等人，202 1]

5 讨论

在上一节中，我们全面比较了在组织病理学全视野切片上应对领域偏移的主要不确定性估计方法。本节将讨论我们的观察结果，并为其他尝试将不确定性估计应用于数字病理学的研究者提出建议。

我们的结果表明，可以可靠地检测错误预测，而恰当的方法能够提高模型对域偏移和标签噪声的鲁棒性，同时提供校准效果更好的预测。

在不确定性估计方法中，集成能够提供最可靠的不确定性估计，还能提升分类性能并改善网络校准。正如预期，将蒙特卡洛 Dropout、测试时增强和随机变分推断与集成结合起来，可以进一步提升分类性能，但也会大幅增加计算需求，这在某些医疗环境中可能难以实现。对于蒙特卡洛 Dropout 和随机变分推断，需要对神经网络进行架构调整，而且在我们的实验中，这两种方法的表现最不稳定。由于所有受考察方法的排名在不同设置之间差异很大（另见 C），因此无法确定一种适用于所有任务的最佳方法。

在分布内数据的错误预测检测方面，所有方法都能可靠地提升分类性能；随着拒绝率升高，错误预测检测在临床场景中也变得切实可行。测试时增强方法在选择性分类方面表现最不稳定，因为只有这些方法的准确率-拒绝曲线呈现非单调上升。对于每种方法，最大幅度的提升都出现在准确率-拒绝曲线的起始阶段，且斜率通常在拒绝率达到约 20% 的阈值时明显下降。根据我们的结果，在许多情况下，只需拒绝约 20% 的最不确定预测，就可能显著提高准确率。

我们观察到，从分布内中心到分布外中心，校准误差显著增大。然而，根据我们的实验，在所有域偏移和数据划分下，没有任何一种非集成方法在校准误差方面胜过其他方法。集成方法通常能改善测得的校准效果。

按不确定性对图块预测进行排序后，我们观察到确定性最高与不确定性最高的图块之间存在视觉上可识别的差异，这一现象在所有方法中都一致。在全切片图像中，对于位于肿瘤转移灶内部的图块，各网络给出的预测尤其具有确定性；而在肿瘤区域的边界处，不确定性最高，这一点在对切片级不确定性进行可视化时也可以观察到。在我们的标签噪声实验中，我们可以立即看到，将带有噪声标签的组织边界图块纳入训练会产生影响，并导致分类器性能下降。因此，应谨慎正确地准备训练数据，因为边界区域中由不准确的标注或切块过程引入的标签噪声可能会显著影响性能。在我们的实验中，集成方法和测试时增强表现出更强的标签噪声抵抗能力。

除了分析不确定性估计方法对图块级肿瘤预测任务的影响外，我们还将不确定性估计应用于结直肠癌组织切片上的切片级微卫星不稳定预测任务。由于临床诊断是在切片级进行的，因此，一个重要的发现是，借助不确定性估计，可以在切片级识别错误预测，并提高神经网络的性能。在这里，集成方法再次优于其单一对应方法，而蒙特卡洛 Dropout 并未优于基线 CLAM 方法。

处理组织病理学数据时，选择正确的指标是一项重要考量。在我们的实验中，我们注意到数据集的不同划分可能导致结果有很大差异，见 C。因此，评估需要采用鲁棒的指标。在图块级实验中，我们选择使用测试数据集中所有切片上该指标的中位性能，事实证明这种做法更加鲁棒，见 A。

6 结论

在组织病理学这一安全关键领域部署基于人工智能的诊断系统，需要采用具备不确定性感知能力的机器学习算法，以增强对模型预测结果的信任。

为此，我们比较了多种不确定性估计方法和不确定性指标在域偏移和标签噪声场景下的表现、校准能力，以及其在组织病理学场景中检测图块级和切片级错误预测的能力。结果表明，所有受研究的方法都能有效检测错误预测，并拒绝其不太可能正确分类的输入；这对于机器学习系统可靠、安全地部署具有重要的临床意义。在我们的实验中，随着被拒绝输入的比例增加，只有 TTA 方法未能使性能单调提升。此外，与单独使用的方法相比，集成方法能够改善组织病理学数据上的校准效果，无论是分布内数据还是分布外数据；另一方面，经典计算机视觉文献中纳入考察的鲁棒性和不确定性估计方法，在域偏移下都无法可靠地改善校准效果。

标签噪声会对预测质量产生重大影响。由于标签噪声在医疗数据中是一个常见问题，因此应特别注意识别可能标注错误的输入，并选择能够生成标注准确的数据点的数据准备技术。在应对标签噪声的鲁棒性方面，集成和测试时增强方法在我们的实验中表现最佳，但没有任何方法能够抵御标签噪声导致的校准能力大幅下降。

总体而言，集成方法在我们的所有实验中表现得最为稳定。与经典计算机视觉文献中的结果相反，蒙特卡洛 Dropout 和随机变分推断在大多数指标上的表现通常仅与基线残差网络相当或略好，但在推理时需要耗费更多计算资源，并且需要对神经网络架构进行修改。测试时增强在分布内数据上取得了最佳分类结果，但这一优势在我们的域偏移场景中并未延续。此外，在选择性分类任务中，它产生的准确率-拒绝率曲线表现最差。总体而言，各方法的排名并不一致，因此我们无法针对所有场景推荐一种表现最佳的方法。根据我们的经验，集成方法相较于各自的单一对应方法，能够带来相对稳定的提升，因此值得推荐；同时，它们在实践中也相对容易实现和使用。

虽然没有任何单一基准测试能够提供面面俱到的结果和见解，但我们希望本次评估能为组织病理学领域中不确定性方法的应用提供指导。我们发布的代码旨在便于复现，并可扩展用于后续研究。

致谢

这项研究由德国巴登-符腾堡州的社会事务与融合部资助。

参考文献

Andre Esteva、Brett Kuprel、Roberto A. Novoa、Justin Ko、Susan M. Swetter、Helen M. Blau 和 Sebastian Thrun。《利用深度神经网络进行皮肤癌皮肤科医生级分类》。*Nature*，542(7639):115–118，2017年2月。ISSN 1476-4687。doi:10.1038/nature21056。

H. A. Haenssle、C. Fink、R. Schneiderbauer、F. Toberer、T. Buhl、A. Blum、A. Kalloo、A. Ben Hadj Hassen、L. Thomas、A. Enk、L. Uhlmann、读者研究I级和II级组、Christina Alt、Monika Arenbergerova、Renato Bakos、Anne Baltzer、Ines Bertlich、Andreas Blum、Therezia Bokor-Billmann、Jonathan Bowling、Naira Braghiroli、Ralph Braun、Kristina Buder-Bakhaya、Timo Buhl、Horacio Cabo、Leo Cabrijan、Naciye Cevic、Anna Classen、David Deltgen、Christine Fink、Ivelina Georgieva、Lara-Elena Hakim-Meibodi、Susanne Hanner、Franziska Hartmann、Julia Hartmann、Georg Haus、Elti Hoxha、Raimonds Karls、Hiroshi Koga、Jürgen Kreusch、Aimilios Lallas、Pawel Majenka、Ash Marghoob、Cesare Massone、Lali Mekokishvili、Dominik Mestel、Volker Meyer、Anna Neuberger、Kari Nielsen、Margaret Oliviero、Riccardo Pampena、John Paoli、Erika Pawlik、Barbar Rao、Adriana Rendon、Teresa Russo、Ahmed Sadek、Kinga Samhaber、Roland Schneiderbauer、Anissa Schweizer、Ferdinand Toberer、Lukas Trennheuser、Lyobomira Vlahova、Alexander Wald、Julia Winkler、Priscila Wölbing 和 Iris Zalaudek。《人与机器对比：深度学习卷积神经网络在皮肤镜黑色素瘤识别中的诊断性能，与58名皮肤科医生进行比较》。*Annals of Oncology*：欧洲肿瘤内科学会官方期刊，29(8):1836–1842，2018年8月。ISSN 1569-8041。doi:10.1093/annonc/mdy166。

Achim Hekler, Jochen S. Utikal, Alexander H. Enk, Axel Hauschild, Michael Weichenthal, Roman C. Maron, Carola Berking, Sebastian Haferkamp, Joachim Klode, Dirk Schadendorf, Bastian Schilling, Tim Holland-Letz, Benjamin Izar, Christof von Kalle, Stefan Fröhling, Titus J. Brinker, Laurenz Schmitt, Wiebke K. Peitsch, Friederike Hoff- mann, Jürgen C. Becker, Christina Drusio, Philipp Jansen, Joachim Klode, Georg Lodde, Stefanie Sammet, Dirk Schadendorf, Wiebke Sondermann, Selma Ugurel, Jeannine Zader, Alexander Enk, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Julia Winkler, Priscilla Wölbing, Hiba Asper, Ann-Sophie Bohne, Victoria Brown, Bianca Burba, Sophia Deffaa, Cecilia Dietrich, Matthias Dietrich, Katharina Antonia Drerup, Friederike Egberts, Anna-Sophie Erkens, Salim Greven, Viola Harde, Marion Jost, Merit Kaeding, Katharina Kosova, Stephan Lischner, Maria Maagk, Anna Laetitia Messinger, Malte Metzner, Rogina Motamedi, Ann-Christine Rosenthal, Ulrich Seidl, Jana Stemmermann, Kaspar Torz, Juliana Giraldo Velez, Jennifer Haiduk, Mareike Alter, Claudia Bär, Paul Bergenthal, Anne Gerlach, Christian Holtorf, Ante Karoglan, Sophie Kindermann, Luise Kraas, Moritz Felcht, Maria R. Gaiser, Claus-Detlev Klemke, Hjalmar Kurzen, Thomas Leibing, Verena Müller, Raphael R. Reinhard, Jochen Utikal, Franziska Winter, Carola Berking, Laurie Eicher, Daniela Hartmann, Markus Heppt, Katharina Kilian, Sebastian Krammer, Diana Lill, Anne-Charlotte Niesert, Eva Oppel, Elke Sattler, Sonja Senner, Jens Wallmichrath, Hans Wolff, Anja Gesierich, Tina Giner, Valerie Glutsch, Andreas Kerstan, Dagmar Presser, Philipp Schrüfer, Patrick Schummer, Ina Stolze, Judith Weber, Konstantin Drexler, Sebastian Haferkamp, Marion Mickler, Camila Toledo Stauner, 和 Alexander Thiem。《人类与人工智能相结合，实现更优异的皮肤癌分类》。*EuropeanJournalof Cancer*，120：114–121，2019年10月。I SSN 0959-8049。doi:10.1016/j.ejca.2019.07.019。

Edmon Begoli、Tanmoy Bhattacharya 和 Dimitri Kusnezov。机器辅助医疗决策中对不确定性量化的需求。*Nature Machine Intelligence*，1(1):20–23，2019年1月。ISSN 2522-5839。doi:10.1038/s42256-018-0004-1。

Benjamin Kompa、Jasper Snoek 和 Andrew L. Beam。需要第二意见：在医疗机器学习中传达不确定性。*npj Digital Medicine*，4(1):1–6，2021年1月。ISSN2398-6352。doi:10.1038/s41746-020-00367-3。

Jeroen van der Laak、Geert Litjens 和 Francesco Ciompi。组织病理学中的深度学习：通往临床应用之路。*Nature Medicine*，27(5):775–784，2021年5月。ISSN 1546-170X。doi:10.1038/s41591-021-01343-4。

Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q. Weinberger。《现代神经网络校准》。载于第*34*届机器学习国际会议论文集，第1321–1330页。PMLR，2017年7月。

Neil Band、Tim G J Rudner、Qixuan Feng、Angelos Filos、Zachary Nado、Michael W Dusenberry、Ghassen Jerfel、Dustin Tran 和 Yarin Gal。《基于糖尿病视网膜病变检测任务对贝叶斯深度学习进行基准测试》。载于*NeurIPS2021*，第15页，2021年。

Paul F. Jaeger、Carsten Tim Lüth、Lukas Klein 和 Till J. Bungert。《呼吁反思图像分类中故障检测的评估实践》。载于*ICLR 2023*，2023年2月。

Dan Hendrycks 和 Thomas Dietterich。《对常见损坏和扰动下的神经网络鲁棒性进行基准测试》。载于*ICLR 2019*，第16页，2019年。

Yaniv Ovadia、Emily Fertig、Jie Ren、Zachary Nado、D. Sculley、Sebastian Nowozin、Joshua Dillon、Balaji Lakshminarayanan 和 Jasper Snoek。你能信任模型的不确定性吗？在数据集偏移下评估预测不确定性。载于神经信息处理系统进展，第 32 卷。Curran Associates, Inc.，2019年。

Karin Stacke、Gabriel Eilertsen、Jonas Unger 和 Claes Lundström。测量组织病理学深度学习中的域偏移。*IEEE* 生物医学与健康信息学杂志，25(2):325–336，2021年2月。ISSN 2168-2208。doi:10.1109/JBHI.2020.3032060。David Tellez、Geert Litjens、Péter Bándi、Wouter Bulten、John-Melle Bokhorst、Francesco Ciompi 和 Jeroen van der Laak。量化数据增强和染色颜色归一化对用于计算病理学的卷积神经网络的影响。《医学图像分析》，58:101544，2019年12月。ISSN 1361-8415。doi:10.1016/j.media.2019.101544。Davood Karimi、Haoran Dou、Simon K. Warfield 和 Ali Gholipour。使用噪声标签进行深度学习：探索医学图像分析中的技术与补救方法。《医学图像分析》，65:101759，2020年10月。ISSN 1361-8415。doi:10.1016/j.media.2020.101759。Jasper Linmans、Jeroen van der Laak 和 Geert Litjens。利用多头卷积神经网络高效检测数字病理学中的分布外数据。载于第三届医学影像与深度学习会议论文集，第 465–478 页。PMLR，2020年9月。Jeppe Thagaard、Søren Hauberg、Bert van der Vegt、Thomas Ebstrup、Johan D. Hansen 和 Anders B. Dahl。在数字病理学中，面对真实数据集偏移时，能否信任预测不确定性？载于 Anne L. Martel、Purang Abolmaesumi、Danail Stoyanov、Diana Mateus、Maria A. Zuluaga、S. Kevin Zhou、Daniel Racoceanu 和 Leo Joskowicz（编），医学图像计算与计算机辅助干预 – *MICCAI 2020*，《计算机科学讲义》，第 824–833 页，Cham，2020年。施普林格国际出版。ISBN 978-3-030-59710-8。doi:10.1007/978-3-030-59710-8 80。_Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。使用深度集成实现简单且可扩展的预测不确定性估计。载于神经信息处理系统进展，第 30 卷。Curran Associates, Inc.，2017年。Yarin Gal 和 Zoubin Ghahramani。将 Dropout 作为贝叶斯近似：表示深度学习中的模型不确定性。载于第 *33* 届国际机器学习会议论文集，第 1050–1059 页。PMLR，2016年6月。Jasper Linmans、Stefan Elfwing、Jeroen van der Laak 和 Geert Litjens。用于数字病理学分布外检测的预测不确定性估计。《医学图像分析》，83:102655，2023年1月。ISSN 1361-8415。doi:10.1016/j.media.2022.102655。Charles Blundell、Julien Cornebise、Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。载于第 *32* 届国际机器学习会议论文集，第 1613–1622 页。PMLR，2015年6月。Murat Seçkin Ayhan 和 Philipp Berens。测试时数据增强：估计深度神经网络中的异方差偶然不确定性。载于首届医学影像与深度学习会议论文集，第 9 页，2018年。Péter Bándi、Oscar Geessink、Quirine Manson、Marcory Van Dijk、Maschenka Balkenhol、Meyke Hermsen、Babak Ehteshami Bejnordi、Byungjae Lee、Kyunghyun Paeng、Aoxiao Zhong、Quanzheng Li、Farhad Ghazvinian Zanjani、Svitlana Zinger、Keisuke Fukuta、Daisuke Komura、Vlado Ovtcharov、Shenghua Cheng、Shaoqun Zeng、Jeppe Thagaard、Anders B. Dahl、Huangjing Lin、Hao Chen、Ludwig Jacobsson、Martin Hedlund、Melih Çetin、Eren Halıcı、Hunter Jackson、Richard Chen、Fabian Both、Jörg Franke、Heidi Küsters-Vandevelde、Willem Vreuls、Peter Bult、Bram van Ginneken、Jeroen van der Laak 和 Geert Litjens。从单个转移灶检测到患者层面的淋巴结状态分类：CAMELYON17 挑战赛。*IEEE* 医学影像汇刊，38(2):550–560，2019年2月。ISSN 1558-254X。doi:10.1109/TMI.2018.2867350。Yang Liu、Nilay S. Sethi、Toshinori Hinoue、Barbara G. Schneider、Andrew D. Cherniack、Francisco Sanchez-Vega、Jose A. Seoane、Farshad Farshidfar、Reanne Bowlby、Mirazul Islam、Jaegil Kim、Walid Chatila、Rehan Akbani、Rupa S. Kanchi、Charles S. Rabkin、Joseph E. Willis、Kenneth K. Wang、Shannon J. McCall、Lopa Mishra、Akinyemi I. Ojesina、Susan Bullman、Chandra Sekhar Pedamallu、Alexander J. Lazar、Ryo Sakai、Samantha J. Caesar-Johnson、John A. Demchok、Ina Felau、Melpomeni Kasapi、Martin L. Ferguson、Carolyn M. Hutter、Heidi J. Sofia、Roy Tarnuzzer、Zhining Wang、Liming Yang、Jean C. Zenklusen、Jiashan (Julia) Zhang、Sudha Chudamani、Jia Liu、Laxmi Lolla、Rashi Naresh、Todd Pihl、Qiang Sun、Yunhu Wan、Ye Wu、Juok Cho、Timothy DeFreitas、Scott Frazer、Nils Gehlenborg、Gad Getz、David I. Heiman、Jaegil Kim、Michael S. Lawrence、Pei Lin、Sam Meier、Michael S. Noble、Gordon Saksena、Doug Voet、Hailei Zhang、Brady Bernard、Nyasha Chambwe、Varsha Dhankani、Theo Knijnenburg、Roger Kramer、Kalle Leinonen、Yuexin Liu、Michael Miller、Sheila Reynolds、Ilya Shmulevich、Vesteinn Thorsson、Wei Zhang、Rehan Akbani、Bradley M. Broom、Apurva M. Hegde、Zhenlin Ju、Rupa S. Kanchi、Anil Korkut、Jun Li、Han Liang、Shiyun Ling、Wenbin Liu、Yiling Lu、Gordon B. Mills、Kwok-Shing Ng、Arvind Rao、Michael Ryan、Jing Wang、John N. Weinstein、Jiexin Zhang、Adam Abeshouse、Joshua Armenia、Debyani Chakravarty、Walid K. Chatila、Inode Bruijn、Jianjiong Gao、Benjamin E. Gross、Zachary J. Heins、Ritika Kundra、Konnor La、Marc Ladanyi,

Augustin Luna、Moriah G. Nissan、Angelica Ochoa、Sarah M. Phillips、Ed Reznik、Francisco Sanchez-Vega、Chris Sander、Nikolaus Schultz、Robert Sheridan、S. Onur Sumer、Yichao Sun、Barry S. Taylor、Jioajiao Wang、Hongxin Zhang、Pavana Anur、Myron Peto、Paul Spellman、Christopher Benz、Joshua M. Stuart、Christopher K. Wong、Christina Yau、D. Neil Hayes、Joel S. Parker、Matthew D. Wilkerson、Adrian Ally、Miruna Balasundaram、Reanne Bowlby、Denise Brooks、Rebecca Carlsen、Eric Chuah、Noreen Dhalla、Robert Holt、Steven J. M. Jones、Katayoon Kasaian、Darlene Lee、Yussanne Ma、Marco A. Marra、Michael Mayo、Richard A. Moore、Andrew J. Mungall、Karen Mungall、A. Gordon Robertson、Sara Sadeghi、Jacqueline E. Schein、Payal Sipahimalani、Angela Tam、Nina Thiessen、Kane Tse、Tina Wong、Ashton C. Berger、Rameen Beroukhim、Andrew D. Cherniack、Carrie Cibulskis、Stacey B. Gabriel、Galen F. Gao、Gavin Ha、Matthew Meyerson、Steven E. Schumacher、Juliann Shih、Melanie H. Kucherlapati、Raju S. Kucherlapati、Stephen Baylin、Leslie Cope、Ludmila Danilova、Moiz S. Bootwalla、Phillip H. Lai、Dennis T. Maglinte、David J. Van Den Berg、Daniel J. Weisenberger、J. Todd Auman、Saianand Balu、Tom Bodenheimer、Cheng Fan、Katherine A. Hoadley、Alan P. Hoyle、Stuart R. Jefferys、Corbin D. Jones、Shaowu Meng、Piotr A. Mieczkowski、Lisle E. Mose、Amy H. Perou、Charles M. Perou、Jeffrey Roach、Yan Shi、Janae V. Simons、Tara Skelly、Matthew G. Soloway、Donghui Tan、Umadevi Veluvolu、Huihui Fan、Toshinori Hinoue、Peter W. Laird、Hui Shen、Wanding Zhou、Michelle Bellair、Kyle Chang、Kyle Covington、Chad J. Creighton、Huyen Dinh、HarshaVardhan Doddapaneni、Lawrence A. Donehower、Jennifer Drummond、Richard A. Gibbs、Robert Glenn、Walker Hale、Yi Han、Jianhong Hu、Viktoriya Korchina、Sandra Lee、Lora Lewis、Wei Li、Xiuping Liu、Margaret Morgan、Donna Morton、Donna Muzny、Jireh Santibanez、Margi Sheth、Eve Shinbrot、Linghua Wang、Min Wang、David A. Wheeler、Liu Xi、Fengmei Zhao、Julian Hess、Elizabeth L. Appelbaum、Matthew Bailey、Matthew G. Cordes、Li Ding、Catrina C. Fronick、Lucinda A. Fulton、Robert S. Fulton、Cyriac Kandoth、Elaine R. Mardis、Michael D. McLellan、Christopher A. Miller、Heather K. Schmidt、Richard K. Wilson、Daniel Crain、Erin Curley、Johanna Gardner、Kevin Lau、David Mallery、Scott Morris、Joseph Paulauskis、Robert Penny、Candace Shelton、Troy Shelton、Mark Sherman、Eric Thompson、Peggy Yena、Jay Bowen、Julie M. Gastier-Foster、Mark Gerken、Kristen M. Leraas、Tara M. Lichtenberg、Nilsa C. Ramirez、Lisa Wise、Erik Zmuda、Niall Corcoran、Tony Costello、Christopher Hovens、Andre L. Carvalho、Ana C. de Carvalho、José H. Fregnani、Adhemar Longatto-Filho、Rui M. Reis、Cristovam Scapulatempo-Neto、Henrique C. S. Silveira、Daniel O. Vidal、Andrew Burnette、Jennifer Eschbacher、Beth Hermes、Ardene Noss、Rosy Singh、Matthew L. Anderson、Patricia D. Castro、Michael Ittmann、David Huntsman、Bernard Kohl、Xuan Le、Richard Thorp、Chris Andry、Elizabeth R. Duffy、Vladimir Lyadov、Oxana Paklina、Galiya Setdikova、Alexey Shabunin、Mikhail Tavobilov、Christopher McPherson、Ronald Warnick、Ross Berkowitz、Daniel Cramer、Colleen Feltmate、Neil Horowitz、Adam Kibel、Michael Muto、Chandrajit P. Raut、Andrei Malykh、Jill S. Barnholtz-Sloan、Wendi Barrett、Karen Devine、Jordonna Fulop、Quinn T. Ostrom、Kristen Shimmel、Yingli Wolinsky、Andrew E. Sloan、Agostino De Rose、Felice Giuliante、Marc Goodman、Beth Y. Karlan、Curt H. Hagedorn、John Eckman、Jodi Harr、Jerome Myers、Kelinda Tucker、Leigh Anne Zach、Brenda Deyarmin、Hai Hu、Leonid Kvecher、Caroline Larson、Richard J. Mural、Stella Somiari、Ales Vicha、Tomas Zelinka、Joseph Bennett、Mary Iacocca、Brenda Rabeno、Patricia Swanson、Mathieu Latour、Louis Lacombe、Bernard Têtu、Alain Bergeron、Mary McGraw、Susan M. Staugaitis、John Chabot、Hanina Hibshoosh、Antonia Sepulveda、Tao Su、Timothy Wang、Olga Potapova、Olga Voronina、Laurence Desjardins、Odette Mariani、Sergio Roman-Roman、Xavier Sastre、Marc-Henri Stern、Feixiong Cheng、Sabina Signoretti、Andrew Berchuck、Darell Bigner、Eric Lipp、Jeffrey Marks、Shannon McCall、Roger McLendon、Angeles Secord、Alexis Sharp、Madhusmita Behera、Daniel J. Brat、Amy Chen、Keith Delman、Seth Force、Fadlo Khuri、Kelly Magliocca、Shishir Maithel、Jeffrey J. Olson、Taofeek Owonikoko、Alan Pickens、Suresh Ramalingam、Dong M. Shin、Gabriel Sica、Erwin G. Van Meir、Hongzheng Zhang、Wil Eijckenboom、Ad Gillis、Esther Korpershoek、Leendert Looijenga、Wolter Oosterhuis、Hans Stoop、Kim E. van Kessel、Ellen C. Zwarthoff、Chiara Calatozzolo、Lucia Cuppini、Stefania Cuzzubbo、Francesco DiMeco、Gaetano Finocchiaro、Luca Mattei、Alessandro Perin、Bianca Pollo、Chu Chen、John Houck、Pawadee Lohavanichbutr、Arndt Hartmann、Christine Stoehr、Robert Stoehr、Helge Taubert、Sven Wach、Bernd Wullich、Witold Kycler、Dawid Murawa、Maciej Wiznerowicz、Ki Chung、W. Jeffrey Edenfield、Julie Martin、Eric Baudin、Glenn Bubley、Raphael Bueno、Assunta De Rienzo、William G. Richards、Steven Kalkanis、Tom Mikkelsen、Houtan Noushmehr、Lisa Scarpace、Nicolas Girard、Marta Aymerich、Elias Campo、Eva Giné、Armando López Guillermo、Nguyen Van Bang、Phan Thi Hanh、Bui Duc Phu、Yufang Tang、Howard Colman、Kimberley Evason、Peter R. Dottino、John A. Martignetti、Hani Gabra、Hartmut Juhl、Teniola Akeredolu、Serghei Stepa、Dave Hoon、Keunsoo Ahn、Koo Jeong Kang、Felix Beuschlein、Anne Breggia、Michael Birrer、Debra Bell、Mitesh Borad、Alan H. Bryce、Erik Castle、Vishal Chandan、John Cheville、John A. Copland、MichaelFarnell、Thomas Flotte、Nasra Giama、Thai Ho、Michael Kendrick、Jean-Pierre Kocher、Karla Kopp、Catherine Moser、David Nagorney、Daniel O'Brien、Brian Patrick O'Neill、Tushar Patel、Gloria Petersen、Florencia Que、Michael Rivera、Lewis Roberts、Robert Smallridge、Thomas Smyrk、Melissa Stanton、R. Houston Thompson、Michael Torbenson、Ju Dong Yang、Lizhi Zhang、Fadi Brimo、Jaffer A. Ajani、Ana Maria Angulo Gonzalez、Carmen Behrens、Jolanta Bondaruk、Russell Broaddus、Bogdan Czerniak、Bita Esmaeli、Junya Fujimoto、Jeffrey Gershenwald、Charles Guo、Alexander J. Lazar、Christopher Logothetis、Funda Meric-Bernstam、Cesar Moran、Lois Ramondetta、David Rice、Anil Sood、Pheroze Tamboli、Timothy Thompson、Patricia Troncoso、Anne Tsao、Ignacio Wistuba、Candace Carter、Lauren Haydu、Peter Hersey、Valerie Jakrot、Hojabr Kakavand、Richard Kefford、Kenneth Lee、Georgina Long、Graham Mann、Michael

Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, Vésteinn Thorsson, Adam J. Bass, and Peter W. Laird. 胃肠道腺癌的比较分子分析。*Cancer Cell*，33(4):721–735.e8，2018年4月。ISSN 1535-6108。doi:10.1016/j.ccell.2018.03.010。

Alex Graves。《神经网络的实用变分推断》。载于神经信息处理系统进展，第 24 卷。Curran Associates, Inc.，2011年。

David M. Blei、Alp Kucukelbir 和 Jon D. McAuliffe。《变分推断：统计学家综述》。美国统计学会会刊，112(518):859–877，2017年4月。ISSN 0162-1459。doi:10.1080/01621459.2017.1285773。

Yeming Wen、Paul Vicol、Jimmy Ba、Dustin Tran 和 Roger Grosse。《Flipout：小批次上的高效伪独立权重扰动》。载于*ICLR 2018*，2018年。

Arsenii Ashukha、Dmitry Molchanov、Alexander Lyzhov 和 Dmitry Vetrov。《深度学习中的域内不确定性估计与集成的陷阱》。载于*ICLR 2020*，第 30 页，2020年。

Aryan Mobiny、Aditi Singh 和 Hien Van Nguyen。用于皮肤病变诊断的风险感知机器学习分类器。《临床医学杂志》，8(8):1241，2019年8月。ISSN 2077-0383。doi:10.3390/jcm8081241。

Tanya Nair、Doina Precup、Douglas L. Arnold 和 Tal Arbel。探索用于多发性硬化症病变检测与分割的深度网络不确定性度量。《医学图像分析》，59:101557，2020年1月。ISSN 1361-8415。doi:10.1016/j.media.2019.101557。

Christian Leibig、Vaneeda Allken、Murat Seçkin Ayhan、Philipp Berens 和 Siegfried Wahl。利用深度神经网络的不确定性信息进行疾病检测。《科学报告》，7(1):17816，2017年12月。ISSN 2045-2322。doi:10.1038/s41598-017-17876-z。Yonatan Geifman 和 Ran El-Yaniv。深度神经网络的选择性分类。载于神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。Malik Sajjad Ahmed Nadeem、Jean-Daniel Zucker 和 Blaise Hanczar。用于比较带拒绝选项的分类方法的准确率-拒绝率曲线（ARCs）。载于第三届系统生物学机器学习国际研讨会论文集，第65–81页。PMLR，2009年3月。Jeremy Nixon、Michael W Dusenberry、Linchuan Zhang、Ghassen Jerfel 和 Dustin Tran。衡量深度学习中的校准。载于计算机视觉与模式识别会议研讨会，第4页，2019年。Leo Joskowicz、D. Cohen、N. Caplan 和 J. Sosna。计算机断层扫描中结构手动轮廓勾画的观察者间变异。《欧洲放射学》，29(3):1391–1399，2019年3月。ISSN 1432-1084。doi:10.1007/s00330-018-5695-5。Martin Holm Jensen、Dan Richter Jørgensen、Raluca Jalaboi、Mads Eiler Hansen 和 Martin Aastrup Olsen。利用评定者间一致性改进卷积神经网络中的不确定性估计。载于 Dinggang Shen、Tianming Liu、Terry M. Peters、Lawrence H. Staib、Caroline Essert、Sean Zhou、Pew-Thian Yap 和 Ali Khan（编），医学图像计算与计算机辅助干预 – *MICCAI 2019*，《计算机科学讲义》，第540–548页，Cham，2019年。施普林格国际出版。ISBN 978-3-030-32251-9。doi:10.1007/978-3-030-32251- 9 59。_Mahendra Khened、Avinash Kori、Haran Rajkumar、Ganapathy Krishnamurthi 和 Balaji Srinivasan。用于全切片图像分割与分析的通用深度学习框架。《科学报告》，11(1):11579，2021年6月。ISSN 2045-2322。doi:10.1038/s41598-021-90444-8。Jakob Nikolas Kather、Alexander T. Pearson、Niels Halama、Dirk Jäger、Jeremias Krause、Sven H. Loosen、Alexander Marx、Peter Boor、Frank Tacke、Ulf Peter Neumann、Heike I. Grabsch、Takaki Yoshikawa、Hermann Brenner、Jenny Chang-Claude、Michael Hoffmeister、Christian Trautwein 和 Tom Luedde。深度学习可直接根据胃肠道癌症组织学预测微卫星不稳定性。*NatureMedicine*，25(7):1054–1056，2019年7月。ISSN 1546-170X。doi:10.1038/s41591-019-0462-y。Mohsin Bilal、Shan E Ahmed Raza、Ayesha Azam、Simon Graham、Mohammad Ilyas、Ian A Cree、David Snead、Fayyaz Minhas 和 Nasir M Rajpoot。开发并验证一种弱监督深度学习框架，从常规组织学图像预测结直肠癌的分子通路状态及关键突变：一项回顾性研究。《柳叶刀·数字健康》，3(12):e763–e772，2021年12月。ISSN 2589-7500。doi:10.1016/S2589- 7500(21)00180-1。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*2016*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第770–778页，美国内华达州拉斯维加斯，2016年6月。电气电子工程师学会。ISBN 978-1-4673-8851-1。doi:10.1109/CVPR.2016.90。Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于*ICLR2015*。arXiv，2017年1月。Florian Wenzel、Kevin Roth、Bastiaan S. Veeling、Jakub Swi.atkowski、Linh Tran、Stephan Mandt、Jasper Snoek、Tim Salimans、Rodolphe Jenatton 和 Sebastian Nowozin。深度神经网络中的贝叶斯后验究竟有多好？载于*ICML 2020*，2020年7月。Takuya Akiba、Shotaro Sano、Toshihiko Yanase、Takeru Ohta 和 Masanori Koyama。Optuna：新一代超参数优化框架。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，KDD '19，第2623–2631页，美国纽约州纽约市，2019年7月。计算机协会。ISBN 978-1-4503-6201-6。doi:10.1145/3292500.3330701。Ming Y. Lu、Drew F. K. Williamson、Tiffany Y. Chen、Richard J. Chen、Matteo Barbieri 和 Faisal Mahmood。全切片图像上的高效数据与弱监督计算病理学。《自然·生物医学工程》，5 (6):555–570，2021年6月。ISSN 2157-846X。doi:10.1038/s41551-020-00682-w。Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型层级图像数据库。载于*CVPR 2009*，第248–255页，2009年6月。doi:10.1109/CVPR.2009.5206848。Maximilian Ilse、Jakub Tomczak 和 Max Welling。基于注意力的深度多实例学习。载于第*35*届国际机器学习会议论文集，第2127–2136页。PMLR，2018年7月。Murat Seçkin Ayhan、Laura Kühlewein、Gulnar Aliyeva、Werner Inhoffen、Focke Ziemssen 和 Philipp Berens。对糖尿病视网膜病变检测深度神经网络诊断不确定性的经专家验证的估计。《医学图像分析》，64:101724，2020年8月。ISSN 1361-8415。doi:10.1016/j.media.2020.101724。

A 指标选择与详细结果

在本 节中，我们阐述采用这些指标的理由，并给出所有实验的详细结果。

组织病理学数据集通常存在严重的类别不平衡，例如，与肿瘤组织类别相比，健康组织等类别所占比例过高。因此，AUROC 和准确率等常用指标可能会产生误导性结果，因为它们对类别不平衡非常敏感。此外，不同切片上的类别分布可能存在显著差异：一张切片中几乎没有肿瘤细胞，而下一张切片则可能主要由大面积转移灶构成。

在我们的实验中，这就是为什么我们能在图1中观察到分布外中心上较高的平衡准确率值。弱偏移下的分布外中心3和强偏移下的分布外中心4都包含肿瘤组织含量异常高的切片，因此产生了许多肿瘤组织覆盖整个图块的图块；如图 2 所示，这些图块上的神经网络对其预测最有信心。肿瘤图块的分布因此向肿瘤覆盖率为100%的图块倾斜，而边界样本较少，进而使分布外中心上的平衡准确率值更高。在表4的强偏移场景下，中心4的 $AUARC$ 值高于中心2，这也体现了这一点。

这类问题使得可靠地估计分类器的性能变得困难，因为对数据集进行不同的划分，可能会导致测试数据集包含差异很大的切片；在我们的实验中，这对方法的测得分数和排名产生了很大影响。

因此，我们报告测试数据集中所有切片上所测指标的中位数。在我们的实验中，指标表现高于或低于平均水平的异常切片始终存在，但结果证明，我们的方法对数据集划分的选择更加稳定。

然而，我们没有在图1和表1的准确率-拒绝曲线中采用这种方法，因为解读多条准确率-拒绝曲线的中位数曲线颇具挑战性。这项任务的主要重点是考察图块级拒绝情况下所测指标曲线的变化，以及各方法的相对排名。下文列出了测试数据集中所有切片的中位数平衡准确率。

表2展示了 Camelyon17 实验中分类器性能的详细结果，并报告了测试数据集中所有切片的中位数平衡准确率和期望校准误差。

在两种域偏移下的分布内数据上，我们可以看到，每种集成方法的表现都优于其单独对应的方法，只有弱偏移下的测试时增强集成表现与之相当。然而，单独来看，蒙特卡洛 Dropout 和随机变分推断的表现均不如 ResNet 基线，而测试时增强的表现略优于基线。

在弱域偏移和强域偏移下，所有替代性的非集成方法都未能优于 ResNet 基线，这与组织病理学领域之外的其他报告相反 [Ovadia 等人， 2019]。然而，大多数集成方法都优于 ResNet 基线，并且表现始终优于相应的单一方法。

与我们的讨论类似 在第4节中，我们无法根据实验结果推断出最佳方法。不过，集成方法的表现优于 基线方法，因此始终值得推荐。

表3展示了标签噪声实验的详细结果。

表2：Camelyon17 上图块分类器性能的详细结果。我们报告了每种域偏移、数据划分和方法在 5 次运行中的中位数及四分位距。指标为测试数据集中所有切片的中位性能。

偏移	划分	方法	平衡准确率 ↑	期望校准误差 ↓
Weak	ID 中心	残差网络	0.8551 0.0264	0.0259 0.0010
		残差网络集成	0。8712 0。0014	0.0218 0.0025
		MCDO	0.8542 0.0356	0。0215 0。0024
		蒙特卡洛 Dropout 集成	0.8691 0.0054	0.0240 0.0016
		TTA	0.8689 0.0058	0.0186 0.0046
		测试时增强集成	0.8669 0.0022	0.0193 0.0023
		SVI	0.8321 0.0368	0。0208 0。0088
		随机变分推断集成	0.8731 0.0194	0.0204 0.0027
	分布外中心	残差网络	0.6961 0.0245	0.0315 0.0055
		残差网络集成	0.6863 0.0130	0.0308 0.0043
		MCDO	0.6842 0.0447	0。0318 0。0018
		蒙特卡洛 Dropout 集成	0.6944 0.0109	0.0315 0.0015
		TTA	0.6871 0.0324	0.0303 0.0038
		测试时增强集成	0.7046 0.0100	0.0292 0.0033
		SVI	0.6732 0.0346	0.0319 0.0029
		随机变分推断集成	0.6860 0.0113	0.0321 0.0065
强	ID 中心	残差网络	0.8019 0.0031	0.0175 0.0034
		残差网络集成	0.8131 0.0014	0.0142 0.0006
		MCDO	0.7952 0.0214	0.0139 0.0085
		蒙特卡洛 Dropout 集成	0。8140 0。0086	0.0139 0.0008
		TTA	0.8105 0.0098	0.0162 0.0024
		测试时增强集成	0.8222 0.0080	0.0161 0.0032
		SVI	0.7907 0.0104	0.0171 0.0019
		随机变分推断集成	0.8126 0.0074	0.0165 0.0008
	分布外中心	残差网络	0.7402 0.0735	0.0378 0.0037
		残差网络集成	0.7940 0.0005	0.0385 0.0014
		MCDO	0.7351 0.0250	0.0470 0.0113
		蒙特卡洛 Dropout 集成	0.7662 0.0095	0.0447 0.0077
		TTA	0.7317 0.0232	0。0582 0。0048
		测试时增强集成	0.7348 0.0295	0.0570 0.0036
		SVI	0.6962 0.0554	0.0493 0.0191
		随机变分推断集成	0.7486 0.0119	0.0608 0.0056

表3：标签噪声条件下各实验的详细结果。我们报告中位数和四分位距，如箱线图所示。

实验	方法	准确率 ↑	平衡准确率 ↑	期望校准误差 ↓
25%	残差网络	0.9762 0.0010	0.8291 0.0073	0.0181 0.0020
	残差网络集成	0.9798 0.0015	0。8474 0。0105	0.0145 0.0011
	MCDO 蒙特卡洛 Dropout 集成	0.9777 0.0053	0.8244 0.0380	**0.0135** 0.0001
	TTA 测试时增强集成	0.9794 0.0006	0.8363 0.0026	0。0137 0。0010
	SVI 随机变分推断集成	0.9788 0.0013	0.8689 0.0076	0.0214 0.0026
		0.9805 0.0003	**0.8690** 0.0043	0.0207 0.0003
		0.9745 0.0019	0.8453 0.0219	0.0240 0.0066
		0.9787 0.0007	0.8452 0.0076	0.0196 0.0007
0%	残差网络	0.9612 0.0004	0.8323 0.0124	0.0531 0.0046
	残差网络集成	0.9658 0.0007	0.8299 0.0110	0.0460 0.0061
	MCDO 蒙特卡洛 Dropout 集成	0.9628 0.0048	0.8055 0.0229	**0.0387** 0.0165
	TTA 测试时增强集成	**0.9677** 0.0012	0.8217 0.0181	0.0491 0.0030
	SVI 随机变分推断集成	0.9637 0.0011	**0.8598** 0.0105	0.0558 0.0038
		0.9654 0.0005	0.8596 0.0029	0.0570 0.0014
		0.9596 0.0050	0.8243 0.0115	0.0586 0.0129
		0.9635 0.0008	0.8306 0.0062	0.0571 0.0020
均匀	残差网络	0.9563 0.0037	**0.8383** 0.0215	0.2678 0.0012
	残差网络集成	0.9564 0.0039	0.8325 0.0143	0.2698 0.0013
	MCDO 蒙特卡洛 Dropout 集成	0.9563 0.0034	0.8074 0.0143	**0。2633** 0。0024
	TTA 测试时增强集成	**0.9557** 0.0073	0.7934 0.0228	0.2644 0.0057
	SVI 随机变分推断集成	0.9483 0.0047	0.8294 0.0204	0.2712 0.0013
		0.9500 0.0003	0.8358 0.0092	0.2719 0.0012
		0.9299 0.0072	0.8010 0.0014	0.2765 0.0059
		0.9317 0.0057	0.8019 0.0030	0.2770 0.0011
边界	残差网络	0.9679 0.0016	0.7879 0.0119	0.0468 0.0015
	残差网络集成	0.9739 0.0005	0.7878 0.0060	0.0463 0.0015
	MCDO 蒙特卡洛 Dropout 集成	0.9706 0.0065	0.7790 0.0182	0.0420 0.0025
	TTA 测试时增强集成	**0.9747** 0.0018	0.7659 0.0134	**0.0405** 0.0042
	SVI 随机变分推断集成	0.9719 0.0029	0.8047 0.0098	0.0523 0.0053
		0.9746 0.0024	**0.8107** 0.0038	0.0500 0.0059
		0.9718 0.0058	0.7922 0.0144	0.0439 0.0074
		0.9738 0.0001	0.7997 0.0002	0.0532 0.0017

B 不确定性指标的比较

我们使用第 2.3 小节中定义的 $AUARC$ 指标，比较近期文献中最常用的不确定性指标，即置信度、熵和方差，以评估它们区分不确定预测与确定预测的能力。

置信度： 一种常用的方法，尤其是在校准领域 [Guo 等人， 2017], , 是取 softmax 的最大值，这也称为预测的“置信度”。其核心思想是，被认为确定的决策远离决策边界，而不确定的决策则接近边界 $1/N$，其中 N 是类别数。

熵： 另一种常用的指标 [Mobiny 等，2019，Band 等人， 2021] 是将预测的熵作为不确定性的度量。模型输出概率的熵计算如下：

$$H(\mathbf{y}|\mathbf{x},\mathcal{D}) = -\sum_{c\,\in\,\mathcal{C}} p(y=c|\mathbf{x},\mathcal{D}) \log p(y=c|\mathbf{x},\mathcal{D}) \tag{3}$$

由于最大可能熵会随类别数量而变化，我们将归一化熵计算为

$$\begin{aligned} H_{\text{norm}} &= \frac{H}{H_{\max}} \\ \text{with } H_{\max} &= -\sum_{i=1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right) = \log N \end{aligned} \tag{4}$$

使用归一化熵时，当 $H_{\text{norm}} \to 1$ 时，不确定性较高；当 $H_{\text{norm}} \to 0$时，不确定性较低。

在二分类场景下，置信度（$conf$）和熵的表现非常相似。给定两个预测 y_i 和 y_j，它们之间存在如下关系：

$$H(y_i) < H(y_j) \iff conf(y_i) > conf(y_j) \tag{5}$$

方差： 由于所有纳入的方法都会生成预测分布，因此可以使用该分布的方差作为不确定性的度量 [Nair 等，2020，Leibig 等， 2017]。如果所有预测器都对结果达成一致，则方差为零；而较高的方差表示不确定性较高。

在表4中，我们使用三种不同的不确定性指标比较了强偏移场景下的 $AUARC$ 值。

可以看出，使用置信度时，平均而言会产生略高的 $AUARC$ 值，而各方法的相对排序保持不变。由于置信度还可用于单个模型，与方差方法相比，我们建议使用置信度作为不确定性指标。

使用弱偏移设置，我们得出了相同的结论，但为简洁起见，此处不再赘述。置信度指标对应的数值与表1中的相同。

表4：强域偏移设置下，图1中准确率-拒绝曲线的曲线下面积均值和标准差（$AUARC$）。我们比较单次预测的置信度/熵、类似集成的方法中平均预测的置信度/熵，以及多次预测的方差。

不确定性度量	不确定性方法	ID 中心	分布外（中心2）	分布外（中心4）
置信度 / 熵	残差网络	$0.906_{\pm 0.017}$	$0.893_{\pm 0.031}$	$0.954_{\pm 0.010}$
置信度 / 熵	残差网络集成	$0.918_{\pm 0.007}$	$0.932_{\pm 0.010}$	$0.972_{\pm 0.004}$
置信度 / 熵	MCDO 蒙特卡洛 Dropout 集成	$0.886_{\pm 0.030}$	$0.914_{\pm 0.012}$	$0.971_{\pm 0.010}$
置信度 / 熵	MCDO 蒙特卡洛 Dropout 集成	$0.905_{\pm 0.008}$	$0.924_{\pm 0.012}$	$0.971_{\pm 0.007}$
置信度 / 熵	TTA 测试时增强集成	$0.932_{\pm 0.005}$	$0.897_{\pm 0.025}$	$0.970_{\pm 0.004}$
置信度 / 熵	TTA 测试时增强集成	$\mathbf{0.933}_{\pm 0.002}$	$0.884_{\pm 0.044}$	$0.963_{\pm 0.009}$
置信度 / 熵	SVI 随机变分推断集成	$0.895_{\pm 0.027}$	$0.931_{\pm 0.025}$	$0.972_{\pm 0.012}$
置信度 / 熵	SVI 随机变分推断集成	$0.896_{\pm 0.006}$	$\mathbf{0.949}_{\pm 0.003}$	$\mathbf{0.981}_{\pm 0.002}$
方差	残差网络集成	$0.914_{\pm 0.007}$	$0.928_{\pm 0.011}$	$0.972_{\pm 0.004}$
方差	MCDO 蒙特卡洛 Dropout 集成	$0.877_{\pm 0.033}$	$0.911_{\pm 0.011}$	$0.968_{\pm 0.012}$
方差	MCDO 蒙特卡洛 Dropout 集成	$0.898_{\pm 0.010}$	$0.915_{\pm 0.011}$	$0.969_{\pm 0.008}$
方差	TTA 测试时增强集成	$\mathbf{0.926}_{\pm 0.005}$	$0.877_{\pm 0.031}$	$0.965_{\pm 0.005}$
方差	TTA 测试时增强集成	$\mathbf{0.926}_{\pm 0.002}$	$0.861_{\pm 0.053}$	$0.959_{\pm 0.010}$
方差	SVI 随机变分推断集成	$0.890_{\pm 0.026}$	$0.929_{\pm 0.023}$	$0.970_{\pm 0.013}$
方差	SVI 随机变分推断集成	$0.888_{\pm 0.007}$	$\mathbf{0.942}_{\pm 0.004}$	$\mathbf{0.981}_{\pm 0.001}$

C 不同数据划分中各方法的排名

除了为 Camelyon17 数据集定义弱偏移和强偏移划分外，我们还定义了一种留一法场景：使用四个中心训练所有方法，并在剩余的分布外中心上评估其性能。由于域偏移的强度，以及因此导致的不同中心划分之间预测性能的下降幅度可能有显著差异，因此无法直接比较所得分数。相反，我们会在每次留一法实验中对所有方法进行排名。

表 5 展示了各方法在指标 $AUARC$、平衡准确率和期望校准误差方面的排名。从各列可以看出，没有哪种方法始终表现最佳；相反，许多方法的排名时高时低，这进一步印证了我们的结果：在组织病理学数据上，不存在一种在不确定性估计和鲁棒性方面表现最佳的方法。

表 5：在五种划分中（每次留出一个中心作为分布外数据），各方法在指标 $AUARC$、平衡准确率和期望校准误差上的排名。我们报告了各方法在分布内和分布外数据上的排名。

划分	指标	留出 out	MCDO	MCDO Ens.	残差网络	残差网络 Ens.	随机变分推断	随机变分推断 Ens.	TTA	TTA Ens.
ID 中心	AUARC	0	5	2	3	7	8	1	6	4
		1	3	7	5	6	2	8	4	1
		2	5	6	2	3	7	8	1	4
		3	5	6	2	3	8	7	1	4
		4	5	4	1	6	8	7	2	3
	平衡准确率	0	5	3	6	1	8	2	7	4
		1	2	6	7	8	1	3	5	4
		2	8	4	5	2	6	7	1	3
		3	8	6	7	4	3	2	1	5
		4	5	6	3	4	8	7	1	2
	ECE	0	5	7	1	3	2	8	4	6
		1	2	6	1	3	5	7	8	4
		2	2	4	5	7	6	8	1	3
		3	5	6	1	3	7	8	2	4
		4	2	4	1	3	8	7	6	5
分布外中心	AUARC	0	6	7	2	5	1	8	4	3
		1	4	5	3	1	2	6	8	7
		2	8	6	5	2	3	1	7	4
		3	8	6	1	4	7	3	2	5
		4	8	5	1	4	7	6	2	3
	平衡准确率	0	8	4	2	6	3	5	7	1
		1	3	2	7	5	1	6	8	4
		2	8	4	3	1	2	5	6	7
		3	7	4	3	5	8	6	1	2
		4	1	7	8	6	4	2	5	3
	ECE	0	3	6	1	4	2	7	5	8
		1	2	6	3	4	1	5	8	7
		2	8	4	7	2	5	1	3	6
		3	1	4	2	3	7	8	6	5
		4	2	4	1	3	6	8	5	7