更大的全连接神经网络能实现更好的泛化吗?

Mohammad Jalili Torkamani内布拉斯加大学林肯分校计算机学院美国内布拉斯加州林肯市
mjalilitorka
mani2@huskers.unl.edu

Suhib Hamid农学与园艺学内布拉斯加大学林肯分校美国内布拉斯加州林肯市
shamid2@huskers.unl.edu

Moaz N. Ishag生物系统工程 内布拉斯加大学林肯分校美国内布拉斯加州林肯市
mishag2@huskers.unl.edu

arXiv:2611.29039v1 [cs.LG] 21 Nov 2026

摘要——**更大的神经网络容量并不一定能提升模型在未见数据上的性能。本研究考察在 Fashion-MNIST 数据集上，较深的全连接神经网络架构是否比浅层架构具有更好的泛化能力。我们在两种学习率和两种 L2 正则化设置下比较两种架构，共形成八种配置。两种架构分别包含 95,410 和 561,470 个可训练参数，相差约 5.9 倍。我们根据验证损失选择检查点和最终配置，并在留出测试集上评估所有配置。选出的较大架构配置取得了 89.82% 的验证准确率和 88.77% 的测试准确率。在每种架构中选择验证损失最低的配置后，两者的测试准确率分别为 88.77% 和 88.52%，相差 0.25 个百分点。在训练设置相同的对比中，较大架构仅在四组比较中的一组取得了更高的测试准确率。类别级错误主要集中在外观相似的服装类别上，而定性像素移除分析表明，模型对形状和边界较为敏感。总体而言，研究结果表明，仅增加架构容量并不能保证更好的泛化能力，这凸显了选择神经网络架构时考虑优化和正则化的重要性。**

I. 引言

选择合适的神经网络架构，需要确定增加模型容量能否提升模型在未见数据上的性能。尽管更深、更宽的神经网络能够表示更复杂的函数，但其对泛化的影响取决于训练条件以及从数据中学到的模式。因此，对神经网络架构进行实证比较时，应考察提高架构复杂度是否能提升留出数据上的性能，以及观察到的优势能否在不同的优化和正则化设置下持续存在。

人工智能和机器学习中的模型性能受架构设计、数据表示、训练流程和评估策略的影响。研究人员比较了神经网络和线性回归方法在污水管道状况预测中的表现 [1],, 并探讨了基于树的分类器的超参数优化和交叉验证 [2]。Ensafi 等人 [3] 将特征选择、类别不平衡处理和分类器堆叠结合起来，用于门诊预约预测。深度学习方法还将时空处理和注意力机制用于脑电图分类 [4], ，以及噪声注入和多任务学习用于农产品图像分类 [5]。这些研究表明，预测性能取决于模型设计和训练的多个方面，因此有必要在比较时兼顾架构容量、优化和正则化。

输入表示、任务构式和部署约束带来了其他需要考虑的因素。Mehrnia 等人[6] 比较了 MedSAM1 和 MedSAM2 在心脏磁共振成像分割中的表现，结果显示，尽管提示要求和处理时间存在差异，两者的分割准确率相当。其他研究将工艺参数纳入铁电极化的基于Transformer的预测 [7]，并采用基于Transformer的强化学习，在不重新训练的情况下调度更大规模的问题实例 [8]。有研究通过在可穿戴设备与远程服务器之间分割计算来考察部署约束 [9],，也有研究采用生成式方法扩充规模较小且不平衡的图像数据集 [10]。除传统准确率之外，开放集入侵检测研究还考察了识别未见攻击与保持精确率之间的权衡 [11], ；而关于在旅游规划中采用生成式人工智能的研究，则探讨了感知有用性、态度以及对错误信息的担忧 [12]。总的来说，这些研究表明，人工智能系统的评估取决于其设计选择和预期应用。

评估标准也不仅限于预测准确率。语言模型研究探讨了偏好不一致情况下的沟通
以及身份条件化交互的质量 [13] 。同样，关于人工智能辅助决策的研究强调应当适当地依赖模型建议 [15]。虽然这些研究涉及人工智能性能的不同方面，但本研究具体关注通过留出分类准确率衡量的泛化能力。

本研究旨在探讨，在 Fashion-MNIST（一个包含十类灰度服装图像的基准数据集）上，更深的全连接神经网络架构是否比更浅的架构具有更好的泛化能力 [16]。我们比较了仅含一个隐藏层的基本架构与含三个隐藏层的扩展架构。在学习率和 L2 正则化设置相匹配的条件下，使用 Adam优化对这些架构进行评估 [17]。验证损失用于模型选择，而单独的测试集则保留用于

评估泛化能力。由于这些架构在深度和宽度上均有差异，本比较考察的是它们的整体性能，而非单独剥离神经网络深度的影响。

II. 问题定义

我们考虑使用 28×28灰度输入进行监督式十分类，像素强度范围为 $[0,255]$。每个输入都归一化至 $[0,1]$ ，并展平为 784 维向量后再输入神经网络。神经网络通过 Softmax 输出层生成十个目标类别上的概率分布，并选择预测概率最高的类别作为输出。通过最小化预测概率分布与独热编码真实标签之间的分类交叉熵，学习模型参数。

III. 实验设计

A. 数据与预处理

我们使用 Fashion-MNIST [16], , 其中包含 60,000 个训练样本和 10,000 个测试样例。原始训练集使用 TensorFlow Datasets 的固定切片 train[:90%] 和 train[90%:], , 分别划分为 54,000 个训练样本和 6,000 个验证样本。这种划分依据的是数据集提供的顺序，而非另行随机打乱或按类别分层的划分。单独的 10,000 个测试样例仅用于最终评估，不会影响模型选择。

像素强度从 $[0,255]$ 到 $[0,1]$, 进行归一化，并且每个 28×28 输入在送入全连接网络之前都会展平为一个 784 维向量。

B. 架构与训练

我们比较了两种全连接神经网络架构，它们在深度、宽度和参数总数方面各不相同。表 I 总结了它们的结构特征。两种架构均接收一个经过归一化的 784 维输入向量，在隐藏层中使用 ReLU 激活函数，并通过一个包含十个单元的 Softmax 输出层生成类别概率。

基本架构包含一个有 120 个单元的隐藏层，共有 95,410 个可训练参数。扩展架构包含三个隐藏层，分别有 500、300 和 60 个单元，共有 561,470 个可训练参数。因此，扩展架构的可训练参数数量约为基本架构的 5.9 倍。通过这一比较，我们可以在训练设置相同的条件下，考察深度和容量显著不同的架构的泛化性能。

L2正则化仅应用于隐藏层全连接层的权重核；偏置和输出层核不进行正则化。优化目标为分类交叉熵均值加上 $\lambda\sum_\ell\|W_\ell\|_F^2$，其中 $\lambda\sum_\ell\|W_\ell\|_F^2$ 对隐藏层权重核求和。所有报告的训练、验证和测试损失均不包含此惩罚项，只衡量交叉熵。

表 I 基本型与扩展型全连接神经网络的架构比较。

属性	基本	扩展
输入维度	784	784
隐藏层	1	3
隐藏单元	120	500, 300, 60
隐藏层激活函数	ReLU	ReLU
输出单元	10	10
输出激活函数	Softmax	Softmax
可训练参数	95,410	561,470
相对参数量	1.0×	5.9×

训练样本在分批处理前会使用大小为 1,024 的缓冲区进行随机打乱。所提供的实现未设置显式随机种子，因此初始化和训练顺序无法固定，不能精确复现。每种配置代表一次训练轮次。

所有模型均使用 Adam [17] 和 $\beta_1=0.9$以及 $\beta_2=0.999$进行优化。我们采用 64 的批量大小，并对每种配置最多训练 40 个轮次。

每种架构均使用学习率 $\eta\in\{0.001,0.0005\}$ 和 L2 正则化系数 $\lambda\in\{0,0.0001\}$进行评估。结合两种架构、两种学习率和两种正则化设置，共得到八种配置。

提前停止会监控验证交叉熵，耐心值设为三个周期数。对于每次运行，都会恢复与验证损失最小值对应的权重。最终配置通过比较全部八次运行中的验证交叉熵，选择其中最低者确定。对于架构层面的比较，我们还会在每个架构内确定验证交叉熵最低的配置。这两种选择仅使用验证结果；测试结果仅用于最终评估。

C. 泛化评估

我们使用留出测试集评估泛化，该测试集始终与模型训练和模型选择相隔离。设E 表示分类错误的测试样例数量， $N=10{,}000$表示测试样例总数。观测到的测试错误率为

$$\hat{e}=\frac{E}{N},$$

这等价于一减去测试准确率。

此外，我们报告测试错误比例的 95% 威尔逊评分置信区间。威尔逊区间无需依赖正态近似的对称界限，并能提供一个处于有效概率范围内的区间。对于 $z=1.96$, 该区间的计算方法如下：

$$\frac{\hat{e}+\frac{z^2}{2N}\ \pm\ z\sqrt{\frac{\hat{e}(1-\hat{e})}{N}+\frac{z^2}{4N^2}}}{1+\frac{z^2}{N}}. \tag{1}$$

表 II 八种评估配置的性能。损失为不含 L2 惩罚项的交叉熵，在恢复的检查点处测量。周期数表示训练周期总数；误差区间为测试误差的 95% Wilson 区间。

模型	周期数	正则化	学习率	验证集交叉熵损失	验证集准确率（%）	测试集交叉熵损失	测试集准确率（%）	误差置信区间（%）
Basic-1	14	0.0	0.001	0.2995	89.35	0.3409	88.22	[11.16, 12.43]
Basic-2	20	0.0001	0.001	0.2975	89.47	0.3323	88.06	[11.32, 12.59]
基础版-3	15	0.0	0.0005	0.3077	88.87	0.3408	87.68	[11.69, 12.98]
基本-4†	21	0.0001	0.0005	0.2923	89.42	0.3300	88.52	[10.87, 12.12]
扩展版-1	8	0.0	0.001	0.3006	88.65	0.3531	87.56	[11.81, 13.10]
扩展版-2	9	0.0001	0.001	0.2992	88.78	0.3435	87.63	[11.74, 13.03]
扩展-3*	13	0.0	0.0005	0.2817	89.82	0.3300	88.77	[10.62, 11.86]
扩展版-4	9	0.0001	0.0005	0.3000	89.02	0.3379	87.90	[11.48, 12.75]

* 整体及 Extended 中的最低验证损失（0.2817）。 † Basic 中的最低验证损失（0.2923）。CE 表示交叉熵。Basic-1–4 的恢复检查点轮次分别为 11、17、12 和 18；Extended-1–4 的恢复检查点轮次分别为 5、6、10 和 6。

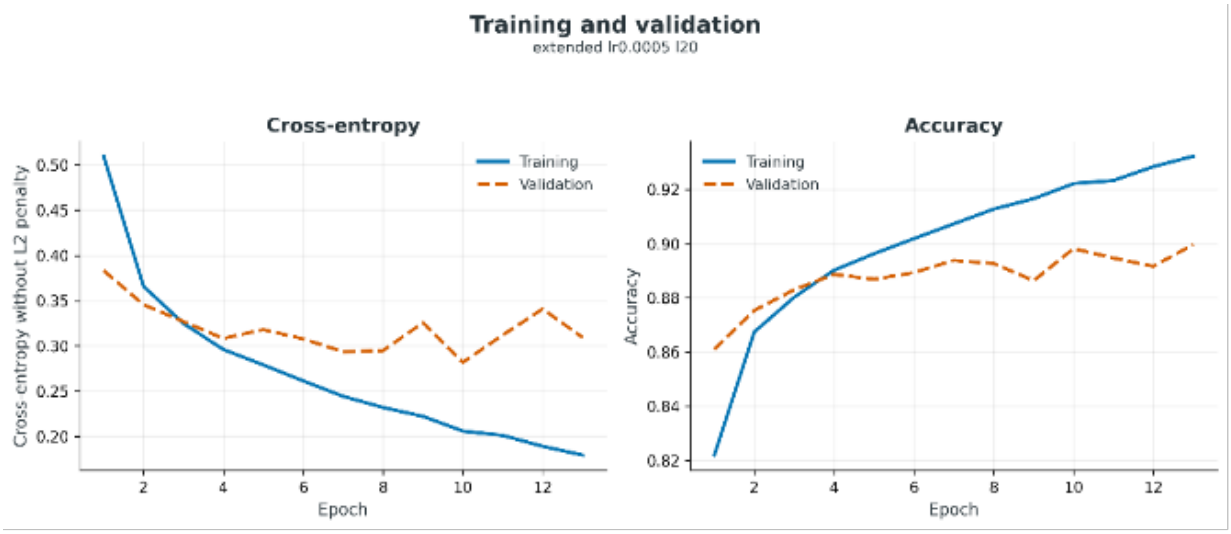

图1. 所选 Extended-3 配置的训练和验证交叉熵损失及分类准确率。

IV. 结果

A. 定量结果

表 II 总结了八种评估配置的性能。具有三个隐藏层的较大架构 Extended-3，在学习率为 0.0005 且不使用 L2 正则化的情况下，取得了最低的验证交叉熵 0.2817。因此，我们选择该配置进行最终分析。该配置的验证准确率为 89.82%，测试准确率为 88.77%，对应的测试错误率为 11.23%，95% 威尔逊置信区间为 10.62–11.86%。

在基础架构中，Basic-4 的验证交叉熵最低，为 0.2923；而在扩展架构内部以及全部八种配置中，Extended-3 的验证交叉熵最低。比较这两种依据验证集选出的配置，基础架构和扩展架构的测试准确率分别为 88.52% 和 88.77%。因此，尽管较大的架构拥有约 5.9 倍的可训练参数，其观测到的性能提升仍为 0.25 个百分点。然而，这一差异并不能证明具有统计显著性，也未考虑多次训练轮次之间的变异

轮次之间的变异。此外，在四种学习率和正则化设置相匹配的情况下，扩展架构仅有一种设置取得了更高的测试准确率。

两种架构的相对性能也会随训练配置而变化。在不进行 L2正则化的情况下，将学习率从 0.001 降至 0.0005，使扩展架构的测试准确率从 87.56% 提高到 88.77%，而基本架构的测试准确率则从 88.22% 降至 87.68%。正则化对不同架构也产生了不同影响。对于采用较低学习率的扩展架构，引入 L2正则化后，验证准确率从 89.82% 降至 89.02%。这些观察结果表明，架构的相对性能会随优化和正则化设置而变化，并非始终倾向于较大的模型。

图 1 展示了所选 Extended-3 配置的训练和验证性能。训练损失持续下降，而验证损失在初始下降后出现波动。训练与验证性能之间不断扩大的差距表明存在轻微过拟合，因为持续优化并不能始终改善验证性能。提前停止通过恢复

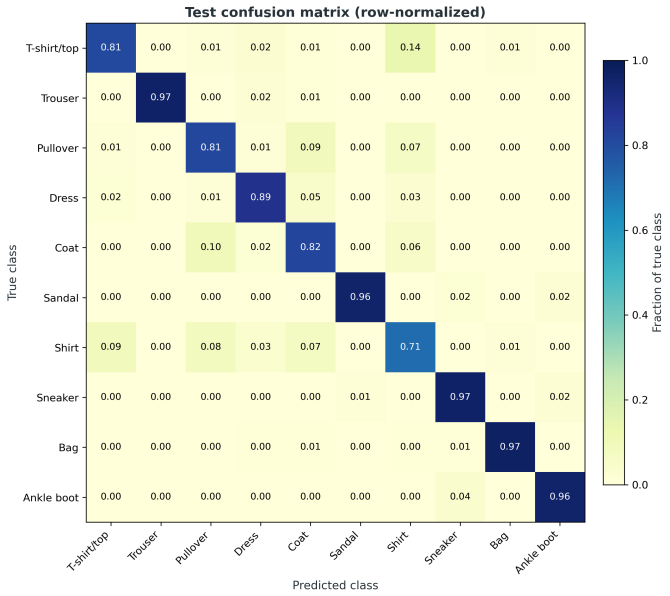

图 2. 所选 Extended-3 配置的行归一化测试混淆矩阵。行表示真实类别，列表示预测类别。

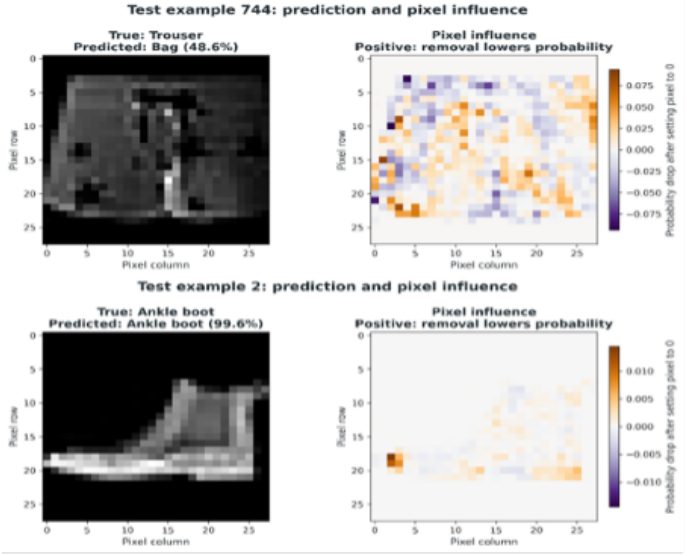

图 3. Extended-3 的代表性测试预测结果和逐像素留一影响图。上：被误分类为包的裤子。下：正确分类的短靴。正向影响表示移除某个像素会降低模型赋予原始预测类别的概率。

与验证损失最小时对应的模型参数来应对这一现象。

图 2 中的行归一化混淆矩阵进一步揭示了类别级性能。Extended-3 在裤子、运动鞋和包这几类具有较鲜明视觉特征的类别上，准确率均达到 0.97。相比之下，衬衫的准确率明显较低，仅为 0.71。被误分类的衬衫样例最常被归为 T恤/上衣（0.09）、套头衫（0.08）或外套（0.07）。这些错误主要出现在视觉上相似的上身服装类别之间，这表明全连接架构难以区分整体形状和结构相近的类别。

B. 像素级影响分析

为进一步研究所选 Extended-3 模型的预测，我们对具有代表性的测试样例进行了逐像素留出影响分析。我们逐一将每个像素设为零，并测量模型原始预测类别对应概率的变化。正的影响值表示移除相应像素会降低该类别的预测概率。

图 3 展示了一个错误预测和一个正确预测。在第一个示例中，裤子被误分类为包，预测概率为 48.6%。相应的影响图突显了图像轮廓和内部的像素，这与折叠衣物宽阔、类似包的外观相符。在第二个示例中，短靴被正确分类，预测概率为 99.6%；具有显著影响的像素主要集中在鞋头和鞋跟区域。

这些影响图为单个预测提供了定性见解，而非模型整体行为的证据。尽管如此，这些示例表明，与形状和边界相关的信息会影响所选模型的分类决策。

V. 讨论与局限性

较大的架构取得了观测到的最高测试准确率，但在匹配的训练设置下，其优势并不稳定。尽管经验证集选出的配置在测试准确率上相差 0.25 个百分点，但这一差异并不能证明其具有统计显著性，也不能说明其在不同训练轮次中具有可复现性。此外，两种架构对学习率和 L2正则化变化的响应不同，表明它们的相对性能取决于训练配置。这些发现表明，评估架构规模时应同时考虑优化和正则化的选择，而不应将其孤立看待。

这些发现的泛化能力受到若干局限因素的制约。首先，评估仅包含两种全连接架构，因此无法对模型规模与泛化之间的关系得出更广泛的结论。其次，报告的威尔逊置信区间量化了单个测试错误比例的不确定性，但无法确定在相同测试样例上评估的模型之间的差异是否具有统计显著性。要通过配对比较评估这种差异，需要模型的逐样本预测。第三，每种超参数配置只进行了一轮训练。鉴于神经网络优化具有随机性，在多个随机种子下重复实验将能更可靠地评估性能变异性。第四，这两种架构

在深度、宽度和参数数量上同时存在差异，因此无法单独确定其中任何一个因素的作用。未来采用容量大致匹配的架构进行实验，或许有助于区分深度与宽度及参数总量的影响。最后，这两个模型处理的都是展平的图像输入，且未明确保留空间关系，因此这些发现的适用范围仅限于专为结构化视觉数据设计的架构。

在更广泛的人工智能和机器学习文献中，本研究关注模型评估的一个维度：在留出基准测试集上的预测性能。其他评估框架还会考量分类准确率以外的方面。例如，语言模型研究探讨了偏好不一致情况下的沟通 [13]以及人口统计学差异对交互质量的影响 [14]。尽管这些研究涉及不同的任务和评估目标，但它们说明，应依据人工智能系统的预期应用，采用相应的标准进行评估。在我们的实验中，混淆矩阵和像素移除影响图通过呈现类别级错误模式以及对单个预测的定性见解，补充了总体准确率。不过，评估鲁棒性和公平性等其他维度，则需要单独的实验设计。

除了预测性能之外，实际的模型选择还必须考虑计算要求和部署约束。例如，Sazonov 等人的可穿戴干预框架通过分割计算处理带宽、功耗和延迟问题 [9]，，而 NoisyViT 研究则指出边缘设备延迟是部署方面的限制 [5]。这些考量促使未来的架构比较除了预测准确率之外，还应纳入推理延迟、内存消耗和能源需求。由于参数数量本身并不能决定部署效率，因此应主要从预测泛化的角度解读本研究结果。评估部署效率和用户采用情况不在本研究范围之内，可作为未来评估的一个方向。

VI. 结论

我们在四组匹配的学习率和 L2 正则化设置下，对比了两种全连接 Fashion-MNIST 架构。尽管较大的架构包含的参数约为小型架构的 5.9 倍，但仅在其中一种设置下取得了更高的测试准确率。分别在每种架构中选取验证损失最低的配置后，较大模型的测试准确率为 88.77%，较小模型为 88.52%。观察到的 0.25 个百分点差异仍受测试样本不确定性和训练变异性的影响。

对于本文考察的架构和单次运行实验，大幅增大模型规模并未带来稳定的泛化收益。未来的研究应区分深度、宽度和参数数量的影响，并使用不同随机种子重复训练，同时将评估扩展到更多数据集和计算成本指标。

VII. 数据可用性

与本研究相关的源代码、实验配置和材料均已公开，网址为 https://github.com/MJaliliT/DLFCN2026。

参考文献

[1] M. Mohammadagha, M. Najafi, V. Kaushal, and A. Jibreen, “基于机器学习模型的钢筋混凝土管状况预测：威斯康星州案例研究中采用人工神经网络和多元线性回归的最新进展”，*Innovative Infrastructure Solutions*, vol. 11, no. 2, p. 65, 2026. [2] M. Mohammadagha, “基于树模型的机器学习预测的超参数优化策略：AdaBoost、决策树和随机森林的比较研究”，决策树和随机森林（*2025*年*4*月*11*日）, 2025. [3] M. Ensafi, A. Mozdgir, N. G. Tehrani, and O. Ahmadi, “开发一种基于机器学习的框架，利用堆叠算法提高门诊诊所预约出席情况的预测能力”，*Internationaljournalofindustrialengineeringandoperationalresearch*, vol. 8, no. 1, pp. 68–94, 2026. [4] A. Amini, M. Alijanpour, B. Latifi, and A. M. Nasrabadi, “Adhdeepnet：从原始 EEG 到诊断——通过时空处理、自适应注意力机制和原始 EEG 信号的可解释性改进多动症诊断”，*arXiv preprint arXiv:2509.08779*, 2025. [5] S. Esfandiari Fard, T. Ghosh, and E. Sazonov, “用于提升水果和蔬菜新鲜度检测及类型分类的多任务 NoisyViT”，*Sensors*, vol. 25, no. 19, p. 5955, 2025. [6] M. Mehrnia, M. Elbayumi, a nd M. S. Elbaz, “评估基础医学‘分割一切’（med-sam1、med-sam2）深度学习模型在三维延迟钆增强磁共振成像左心房分割中的表现”，*arXiv preprint arXiv:2411.05963*, 2024. [7] M. Y. Arzefouni, H. Yun, P. C. Doerschuk, A. Lal, and B. Davaji, “用于预测 HZO 铁电极化的过程感知深度学习”，发表于 *2026*年第*37*届*SEMI*年度先进半导体制造会议（*ASMC*）。IEEE，2026年，第1–5页。 [8] F. Ardali, M. A. Nyelele, and G. M. Knapp, “一种基于深度强化学习（DRL）的 Transformer 方法，用于求解开放车间调度问题”，*arXiv preprint arXiv:2606.13682*, 2026. [9] E. Sazonov, D. Hossain, S. E. Fard, M. B. Hossain, M. A. McCrory, J. Higgins, and J. G. Thomas, “针对饮食行为的及时自适应干预的通信与机器学习框架”，发表于 *2026*年第*6*届中东和北非通信会议（*MENACOMM*）。IEEE，2026年，第1–8页。 [10] I. Khazrak, S. Takhirova, M. M. Rezaee, M. Yadollahi, R. C. Green II, and S. Niu, “利用生成模型处理小规模且不平衡的医学图像数据集”，发表于人工智能与应用，2025年。[11] A. Safari and D. J. Kim, “提升物联网安全与信息系统韧性：一种极值机器方法”，2025年。 [12] A. Safari, “人工智能旅游采用的风险收益分析：旅行规划中的先前使用经验、熟悉度、有用性和人工智能错误信息担忧”，*Information Technology & Tourism*, vol. 28, no. 2, p. 48, 2026. [13] H. H. Balyani, S. P. M. Davoudi, A. Amiri-Margavi, A. G. Davodi, and A. Gharagozlou, “诚实的人工智能顾问：针对偏好不一致情境下大语言模型诚实性的预先指定基准”，*arXiv preprint arXiv:2606.01456*, 2026. [14] A. Amiri-Margavi, A. Gharagozlou, A. G. Davodi, S. P. M. Davoudi, and H. H. Balyani, “机会平等，互动不平等：对大语言模型公平性的反事实审计”，*arXiv preprint arXiv:2602.02932*, 2026. [15] A. Jabbari, S. Ashrafzadeh, S. Sarikhani, and S. Minayhashemi, “对人工智能决策支持系统的适当依赖：社会技术叙事综述”，*Information*, vol. 17, no. 10, 2026. [在线]。可访问：https://www.mdpi.com/

2078 2489 17 10 1005 - / / / H. Xiao, K. [16] Rasul, and R. Vollgraf, “Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集”，*arXiv preprint arXiv:1708.07747*, 2017. [17] D. P. Kingma and J. Ba, “Adam：一种随机优化方法”，*arXiv preprint arXiv:1412.6980*, 2014.