农业大数据学报 2023, 5(4) : 13-23
Journal of Agricultural Big Data
DOI: 10.19788/j.issn.2096-6369.230402

研究论文

从水稻病害识别出发探索农业数据共享新模式

张濛濛[1,2]，王秀娟[1,3]，康孟珍[1,2*]，华净[1,3]，王浩宇[1,3]，王飞跃[4,5]

1. 中国科学院自动化研究所多模态人工智能系统全国重点实验室，北京 100190；2. 中国科学院大学人工智能学院，北京 1000049；3. 中国科学院自动化研究所北京市智能化技术与系统工程技术研究中心，北京 100190；4. 澳门科技大学创新工程学院，澳门 999078；5. 中国科学院自动化研究所复杂系统管理与控制全国重点实验室，北京 100190

摘要：准确高效地识别作物病害类型，有助于农户及时采取有效的针对性预防措施，从而降低因病虫害导致的减产风险和经济损失。然而，在其他领域能达到 SOTA 效果的识别模型，在农业领域特别是水稻病害识别的应用中，却面临目前已有的水稻病害数据量不足、种类不丰富以及数据质量不高等问题。本研究采用多种经典卷积神经网络，并利用迁移学习的方法在两个不同的数据集上进行训练。验证了除模型结构带来的优化外，训练数据集本身对于训练结果也具有重要影响。但目前农业领域开源数据较少，几乎没有综合性的数据开源平台可供利用。这一现象与高质量农业数据获取难度大且成本高、大多数从业人员教育水平相对较低、分布式训练系统不成熟、数据安全问题得不到保障等因素密切相关。针对农业领域训练中高质量数据缺乏的问题，在本文中提出了基于联邦学习框架构建农业数据共享平台的新思路。

关键词：水稻病虫害识别；卷积神经网络；分布式训练；联邦学习；开源数据共享平台

1 引言

水稻是我国的主要农作物之一，具有重要的经济战略地位。目前，病虫害的识别主要依靠人工观测方法，不仅耗时耗力，识别率较低，还不能及时发现病虫害，延误最佳诊断时间，从而影响病虫害识别的准确度和时效性[1]。为了提高识别的准确度和效率，一些学者用传统机器学习方法识别水稻病虫害。传统机器学习方法的思路如下：首先，通过特征工程，从数据集中提取不同水稻病虫害的特征。常用的特征提取方法包括 SIFT(Scale Invariant Feature Transform)[2]、HOG(Histograms of Oriented Gradients)[3]、SURF (Speeded-Up Robust Features)[4]等。然后，再使用 K 近邻聚类[5]或支持向量机（SVM）[6]等传统机器学习的代表性方法，对提取的特征进行分类，从而实现对不同水稻病虫害类型的准确识别。

近年来，卷积神经网络模型在众多图像识别任务中表现出色。自从 2012 年 AlexNet 模型在 ILSVRC（ImageNet Large-Scale Visual Recognition Challenge）中脱颖而出后，卷积神经网络在全世界掀起了研究热潮。为了不断提升卷积神经网络的性能，学者们相继提出了 VGGNet、ResNet、InceptionNet 等一系列卷积神经网络。随着人工智能在农业领域的应用逐渐普及，部分学者逐渐将卷积神经网络应用到农作物病虫害识别领域。Mohanty 等[7]在 PlantVillage 数据集上使用 AlexNet 和 GoogleNet 模型，对 14 种作物的病虫害进行了识别；Lu 等[8]在 500 张图片的数据集上使用卷积神经网络模型，对 10 种水稻病虫害进行分类；Liu 等[9]使用卷积神经网络，对 5136 张水稻病虫害图片进行识别。虽然水稻病虫害识别领域已经取得了一定进展，但是目前的研究中仍存在以下不足：（1）实验数据集规模较小，难以有效优化深度卷积神经网络中的大量模型参数；涵盖病害种类不齐全，不能满足实际的应用需求；（2）公开数据集质量参差不齐，不同数据集使用同一模型训练时效果差别显著；（3）农业领域气候多变、农产品品种多样，因此农业病害数据获

收稿日期：2023-10-30；**录用日期：**2023-11-27
基金项目：国家重点研发计划资助项目（No. 2021ZD0113701）
联系方式：第一作者张濛濛，E-mail：zhangmengmeng2022@ia.ac.cn；通信作者康孟珍，E-mail：mengzhen.kang@ia.ac.cn。

取困难、数据采集成本高，而且开源数据相对较少，缺乏专业的大型农业数据共享平台。

卷积神经网络模型结构在发展过程中逐渐向更深的网络方向改进，导致训练过程中有大量的模型参数需要拟合。为了更好地拟合这些模型参数，对于数据有如下要求：（1）用于深度模型训练和测试的样本都需要满足独立同分布的条件；（2）必须有足够多的训练样本才能学习到一个优秀的深度分类模型[10]。然而在农作物病虫害识别的实际应用中，这两个条件往往难以满足。能否采集到农作物病虫害的数据，往往取决于农作物的生长周期内是否发生了病虫害；不同的病虫害往往发生在不同的地区，需要跨地区采集数据；同一种病虫害的不同阶段，农作物的表型也有区别。所以农作物病虫害数据集的构建过程，不仅需要耗费大量的人力物力成本，同时也难以采集到足够多的数据供模型学习使用。

迁移学习能够很好地解决模型训练数据稀缺的问题。它的主要思想是将预训练模型在其他数据集上已经学到的知识，迁移运用到更多的相关领域中，从而减少对于相关领域训练数据量的需求[11-13]。预训练模型一般是在大型数据集上进行的训练，例如图像领域的 ImageNet 数据集。该数据集一共有超过 1400 万张的标注图片，大约有 2.1 万类。丰富的类别和充足的图片数量，让该数据集能够为预训练模型提供充足有效的通用先验知识，进而让模型能够通过少量数据微调后，被应用到丰富多样的下游场景中。

本文旨在利用迁移学习的方法，在不同的数据集上使用多种神经网络模型进行水稻病害识别，以对比分析不同卷积神经网络模型在农业数据集上的训练效果，同时研究不同数据集在相同模型上的表现情况。旨在验证除了模型结构之外，农业数据集本身对神经网络训练效果的重要性，并探讨一种新的农业数据共享和农业模型训练模式——基于联邦学习框架[14]的农业数据共享平台。探索新的农业数据共享和农业模型训练模式，不仅是为了推动农业领域科研的发展，更是为了更好地服务于农业生产。目前，数字信息鸿沟是阻碍农业生产力发展，限制农业经济和社会一体化进程的关键障碍之一[15]。综合考虑社会信息和物理信息的分布式农业服务系统是弥合这一信息鸿沟的有效方式[16–18]。基于联邦学习框架的农业数据共享平台，不仅为高质量科研数据共享提供了解决方案，平台建设中关于分布式系统部署、链上农业数据安全保护、智能合约与通信激励方式的讨论，或将受益于分布式农业服务系统的启发，进一步推动农业智能服务系统的发展。

2 方法

2.1 研究数据与方法

2.1.1 研究数据

本文使用了两个数据集，包括 Prajapati 等人使用的 4 种水稻叶片病斑数据集 A[19]，以及农业科研院所的 5 种水稻叶片病斑数据集 B[20]。

(1)数据集 A

数据集 A 包括 train 和 validation 两部分的数据，共 3355 张图片。其中 train 集有 2685 张图片，validation 集有 670 张图片。该数据集包括褐斑病、稻瘟病、白叶枯病等 3 种水稻病害叶片以及健康叶片。不同病害及健康叶片所包含的图片数量如表 1 所示。

表 1 数据集 A

Table 1 Dataset A

数据集组成	Train	Validation
褐斑病	417	105
稻瘟病	623	156
白叶枯病	452	113
健康叶片	1191	297

(2)数据集 B

数据集 B 包括胡麻斑病、稻瘟病、白叶枯病、纹枯病、细菌性条纹斑病等 5 种水稻病害类型，共 6280 张图片。数据集中的图片来源主要是科研人员实地采集的稻田病虫害图片，图片质量较高。不同病害所包含的图片数量如表 2 所示。

表 2 数据集 B

Table 2 Dataset B

数据集组成	数量
胡麻斑病	1537
稻瘟病	1677
白叶枯病	914
纹枯病	740
细菌性条纹斑病	1412

2.1.2　研究方法

根据迁移学习的特性，在本文中使用了 AlexNet、Vgg16_bn、Resnet50、Inception_V3 等 4 种经典的卷积神经网络的预训练模型。其中模型的预训练参数使用的是 ImageNet 数据集上预训练得到的参数。ImageNet 数据集包含了丰富的图片类别和充足的图片数量，能够提供有效的通用先验知识，是很多利用迁移学习进行下游任务时会选择的预训练数据集。针对两个水稻数据集，通过冻结各模型除全连接层以外的所有参数，并使用适合于当前任务的全连接层替换掉原本的全连接层，然后仅对全连接层参数进行微调，本研究分别在两个数据集上进行训练和测试。尽管迁移学习已经大量减少了训练中所需的数据量，但是训练数据的质量还是会在很大程度上影响迁移学习网络在下游任务中的效果。因此，为了验证模型的优化与农业数据集本身的质量在农业病虫害识别任务中的重要性，本文采用两种不同的方式进行对比分析：（1）纵向比较各模型在同一数据集上的训练结果；（2）横向比较同一卷积神经网络在不同数据集上的表现情况。

2.2　实验设置

2.2.1　实验环境

使用深度学习框架 Pytorch 软件进行实验验证。电脑配置为 CPU：Intel i7-137000@ 2.1GHz;GPU 配置为 NVISDIA GeFore RTX4070ti，显存大小 12G。

2.2.2　实验数据

本实验使用的 4 种卷积神经网络模型最后一层均为全连接层，所以需要固定输入图片的大小。实验中 Inception_V3 使用的是 229×229 大小的图片，其他使用的是 224×224 大小的图片。其中数据集 A 将 train 数据按照 7:3 分割为训练集和验证集，将 validation 数据集用于测试，数据集 B 按照约 6:1:3 的比例划分为训练集、验证集与测试集。训练集与验证集用于模型训练，测试集只用于模型的精度评价。

2.2.3　数据增强

深度卷积神经网络(CNN)在图像处理任务中取得了显著的成绩。然而，它们的高表达能力有过拟合的风险。因此，在丰富数据集的同时，提出了数据增强技术来防止过拟合[21−23]。实验的 4 个神经网络模型均使用 ImageNet 数据集的预训练参数。ImageNet 数据集在 R、G、B 三个通道上的均值为（0.485，0.456，0.406），标准差为（0.229，0.224，0.225），因此对本实验使用的水稻病害数据集也采用 ImageNet 上的均值和标准差做归一化处理。

本实验对训练集图片进行了随机垂直、水平翻转、随机灰度化以及随机旋转 30°等数据增强操作，而对于验证集则不做额外的操作，只将其切割到适合模型的大小。实验中使用 4 种卷积神经网络的预训练模型，分别在两个数据集上迭代训练 40 次后测试。

3　实验结果与分析

3.1　不同模型对结果的影响

对比分析不同模型在同一数据集上的训练曲线（图 1），我们发现随着模型的深度和结构不断优化，训练的收敛迭代次数逐渐减少，训练精度不断上升。

AlexNet 是最基础的卷积神经网络模型，仅由 5 个卷积层以及 3 个全连接层组成，且初始的卷积核大小为 11×11，感受野较大[24]，在病兆较小且于图片中并不占主要位置的数据集 A 上，不能有效捕捉到有效特征，只有 64.52%的准确率，见表 3。

表 3　实验数据集划分

Table 3　Dataset division

Datasets	Train size	Validation size	Test size
Dataset A	1878	805	671
Dataset B	3846	550	1884

相较于 AlexNet，Vgg16_bn 网络深度有了提升，共有 16 层，其中有 13 个卷积层以及 3 个全连接层，并且采用较小的感受野，让网络能够捕捉到输入图片更加细节的特征[25]，加入了 Batch Norm 层提升模型的稳定性。从上文的 Vgg16_bn 的训练结果上可以看到，训练集和验证集的准确率在迭代次数较多的情况下波动减小，并且在验证集上的准确度相较于之前有了 0.2 左右的提高，稳定性也更好了。

ResNet50 在 Vgg16_bn 的基础上，网络层次进一步加深了，学习能力进一步提升，并且加入了残差机制，避免网络因为加深而导致的梯度消失或爆炸问题[26]。ResNet50 在数据集 A 上的准确率相较于 Vgg16_bn 提升了约 17.5%,在数据集 B 上也提升了约 9%。

Inception 模型相较于之前的模型，通过多个卷积核提取图像不同尺度的信息，最后进行融合，可以得到图像更好的表征，并且提出了 BN 算法对于每一层的输出进行正则化，使得模型各层之间的输入输出满足同

一正态分布，进而提升模型的效果[27]。Inception_V3 继续增加了原模型中的分支数量，并且将原本对称形式的卷积核变成了非对称形式，在处理更多、更丰富的空间特征以及增加特征多样性方面的效果更好了。在减少了计算量的同时，模型的学习能力也获得了更进一步的提升[28]。Inception_V3 模型在数据集 A 上的准确率相较于 Vgg16_bn 提升了 4%，在数据集 B 上提升效果不明显，仅约 0.16%。

由此可见优化模型对于提升同一数据集上的训练效果是有效的，拥有更强学习能力的网络，能够在相同数据集上学到更多的特征知识用于下游任务[29]。这与目前众多学者们的结论是一致的。所以随着计算视觉领域的发展，大量的优化模型被提出，近期更是有针对多应用场景都体现出强大智能推理能力的“大模型”出世。

3.2 不同数据集对结果的影响

模型的能力确实一定程度上会影响训练结果，但数据对于训练结果的影响可能更大[30]。通过对比同一模型在两组数据上的表现（图 1，表 4），我们发现相同模型在数据集 B 上效果显著更好，尤其是在使用 AlexNet 这一浅层神经网络的时候，数据集 A 的训练及测试结果准确率只有 64.52%，而数据集 B 的准确率高达 86.15%。在训练其他学习能力更强的模型时，数据集 B 也收敛得更快、准确率更高。例如在 ResNet50 模型训练中，第 5 次迭代时数据集 B 基本已经收敛到模型的最佳效果，而数据集 A 却仍然还有剧烈波动，且精度较低。可见好的训练数据可以很大程度上影响模型的训练效果。这一点在近期大模型的训练中也得到验证——大模型初始训练时需要大量的语料库输入[31−33]，训练完成后还需要高质量的 prompt engineering，使用恰到好处的 prompt 对模型进行微调，从而让大模型能够更准确地理解用户输入的意图。可见，数据集本身对于模型的影响是非常大的。使用更加符合应用场景的训练数据得到的模型，也会在下游任务中取得更好的表现。

表 4 两个数据集在四种模型上的测试准确度

Table 4 Accuracy of two datasets by 4 different models

Model	Dataset A	Dataset B
AlexNet	64.52%	86.15%
Vgg16_bn	67.99%	90.29%
ResNet50	85.47%	99.20%
Inception_V3	89.32%	99.36%

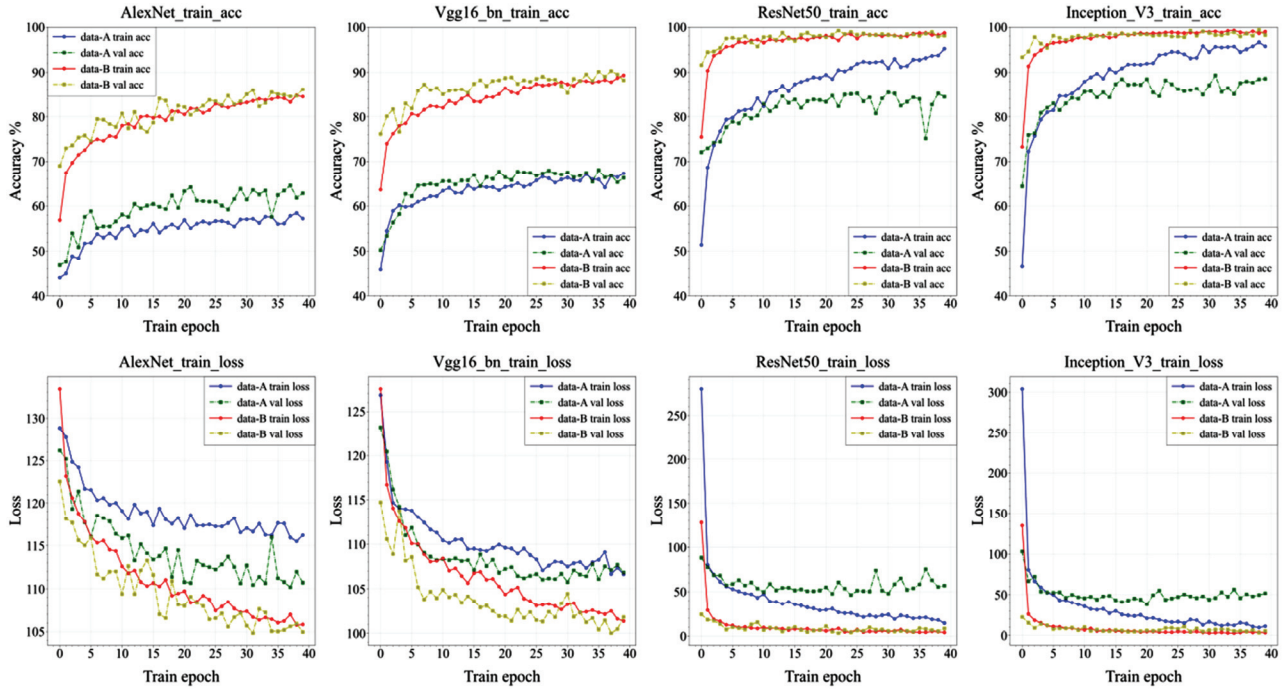

图 1 四种模型在两个数据集上的训练曲线图

Fig. 1 Training curves for 4 models on 2 datasets

4　分布式学习模型

4.1　联邦学习框架与分布式模型训练

近年来，为了在模型训练时保护数据的隐私安全，一种新的 AI 训练范式“联邦学习”（federated learning，简称 FL）应运而生。“联邦学习”是基于区块链提出的分布式学习模型。其在数据预处理、训练、评估和部署阶段的整个数据生命周期内都强调了数据的安全问题。联邦学习(也称为协作学习)是一种机器学习技术，它可以在不跨多个分散的边缘设备或服务器传输数据样本的情况下训练算法。这种策略不同于标准的集中式机器学习技术，后者将所有本地数据集上传到单个服务器；也不同于更传统的去中心化替代方案，后者通常假设本地数据样本是均匀分布的。联邦学习允许多个参与者在不共享数据的情况下协作开健壮的机器学习模型，从而解决数据隐私、数据安全、数据访问权限和对异构数据的访问等关键问题。“联邦学习”被广泛应用于国防、电信、物联网和制药行业等领域。FL 的核心理念是让模型通过分布式训练，直接在多种数据源中利用本地数据进行训练，再通过交换模型参数或中间结果来完成数据需求方的训练需求[34–36]。在构建分布式学习系统，以及利用多个数据源训练模型时，联邦学习采用包括安全多方计算、差分隐私、硬件加密在内的隐私保护技术，来防止数据信息的泄漏[37-38]。

分布式模型训练系统，能够让散布在边缘节点“孤岛”数据，在网络中被再次利用。但分布式建模环境下的数据安全隐患较大，没有能够掌握所有节点行为的信息中心，让任意参与数据价值交换的节点都存在利用欺诈行为恶意窃取其他节点信息的可能性[39]。为了更好地共享农业数据，建立一个可信、透明且可溯源的数据交换体系[40]，研究人员提出了基于区块链的 FL 框架。区块链是一种去中心化的分布式账本数据库[41]，具有不可编造、篡改以及撤销的特性，是一种全新地去中心化基础架构与分布式计算范式[42]，为构建安全可信和可便捷编程的智能建模新生态[43]提供了技术基础。基于区块链的 FL 框架，可以在数据建模过程中保护信息的同时，交换和验证移动设备上本地学习模型的更新[44]。在将区块链与区块链存储相结合的情况下，还能够为用户构建一个专注于隐私的分布式个人数据管理平台[45]。

基于区块链的 FL 框架，可以使用智能合约（smart contract）[46–48]作为协同建模引擎，利用区块链共识机制进行数据定价，从而实现去中心化的农业数据共享。智能合约是一种存放在区块链中的可编程合约，当合约成立的条件达到时，代码合约就会自动执行。例如双方用户通过电子签名达成数据使用的智能合约，其中合约包括数据的使用范围，双方约定的数据交易方式等。该交易数据无法被解密且只能用于智能合约签订的任务。数据用完后就会被销毁，并且合约会自动按照约定的交易内容完成交易。

4.2　农业数据共享平台原型

基于已有的去中心化区块链技术在其他领域的应用实例[49]，针对构建农业数据共享平台，本文提出具有 5 层结构的联邦学习框架原型。5 层结构分别为：物理层、基础技术层、中心管理层、激励与合约层、应用层。原型结构图如图 2 所示。关于各结构层的说明如下：

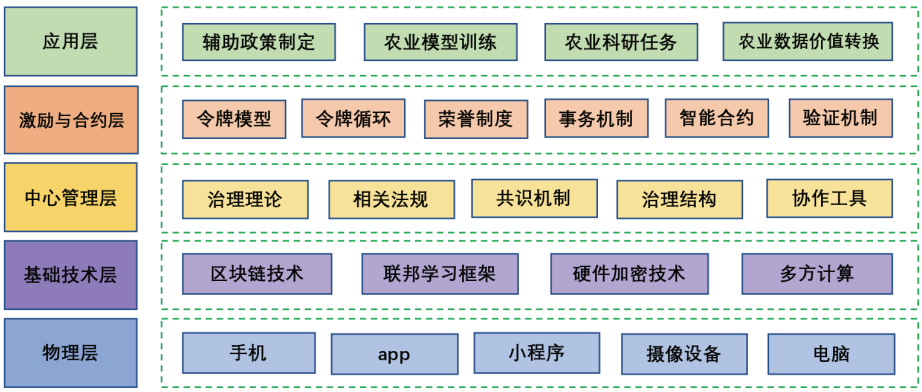

图 2　农业数据共享平台原型

Fig. 2 Prototype of the agricultural data sharing platform

（1）物理层：物理层封装了参与到农业数据共享平台中的各物理实体，包括如手机、APP、小程序、摄像设备、电脑等终端。这一层的关键是数据采集和应用。这些参与到农业数据共享平台的物理实体，在生产活动中产生的农业数据，就是能够在农业数据共享平台上流通并服务于应用层的数据。

（2）基础技术层：基础技术层概括了农业数据共享平台依赖的技术，例如区块链技术，联邦学习框架，硬件加密技术、多方计算等。其中关键技术是区块链技术以及联邦学习框架。区块链技术在平台部署时提供全局不可替代的身份登记，并可以分散存储不可篡改数据，让数据具有高可信度和低验证成本的特征，利于将农业数据平台推广到更大的使用群体。联邦学习框架为参与平台的各实体节点提供去中心化数据流通的范式，服务于应用层的各种应用需求。以农业模型训练为例，联邦学习框架的概念图如图 3 所示。其中模型提供方可以在平台上发布模型性能提升需要的数据，各物理实体作为平台网络中的节点可以提供他们想要在平台上交流的农业数据。模型提供方需要训练的局部模型，与数据提供方想要提交换取激励的数据，在平台中通过智能合约完成相应的训练任务，模型与数据本身的信息对于双方相互并不可见，可以有效保护双方的数据安全与知识产权。

（3）中心管理层：中心管理层封装了所有可能的治理政策和相关规定，包括治理理论和相关法规、共识机制、治理结构和协作工具。很多相关的去中心化组织是通过民主决策进行管理，但是大多数这样的去中心化组织的参与实体较少，在几十上百人左右，而我们的农业数据共享平台是从丰富农业数据的角度出发构建的，纳入平台的物理实体远超过一般的去中心化组织。民主决策方式在考虑治理效率时，不太适用于农业数据共享平台的管理。本文认为可以构建阶段式的管理策略，在平台构建初期采用民主决策的方式进行管理，当参与平台的物理实体不断增加后，开启社区管理模式，各节点可以在社区中提出话题，当话题热度达到一定阈值时开启社区投票，制定新的或者修订旧的治理政策和规定。

（4）激励与合约层：激励与合约层封装了通用和异构状态激励与合约的设计和实现。其中激励包括令牌模型，令牌循环和荣誉制度。合约相关内容包括合约的基础框架和模型的所有协议、代码功能、事务机制、验证机制。其中关于可信交易的智能合约部分，具有自动化、可编程功能。

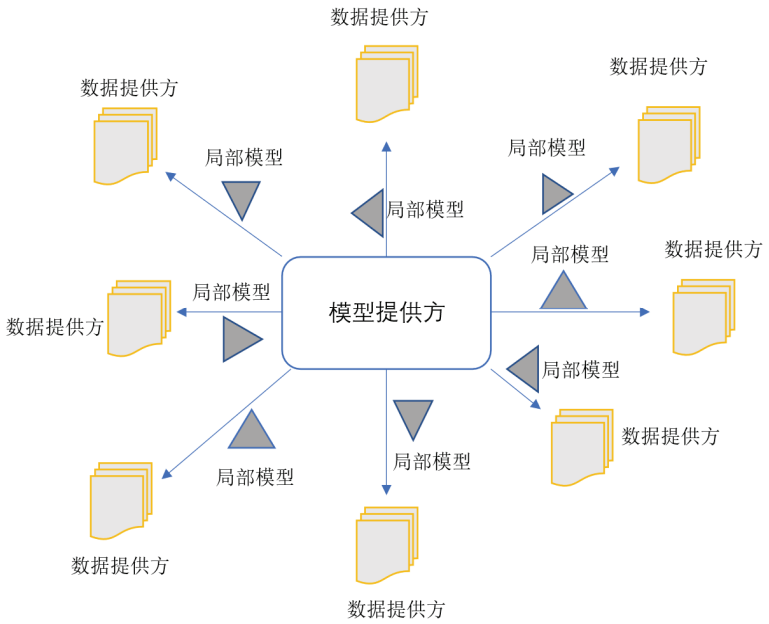

图 3　FL 框架下的农业模型训练

Fig. 3　Agricultural model training in FL framework

（5）应用层：应用层封装了基于区块链技术和联邦学习框架的农业数据共享平台潜在的应用场景和功能方法。根据农业数据可能被应用到的场景，以及智能合约与激励能被参与实体认可的程度，平台可以为政府相关部门的政策制定提供数据，为农业病虫害识别等农业模型的开发提供数据，为各农业科研任务提供数据，还可以为参与的各数据提供实体提供数据价值转换的渠道等。

5　讨论

本研究通过使用 4 种经典卷积神经网络在两个水稻叶片数据集上做病害识别，比较分析后发现模型与数据集本身对于训练结果都有重要影响。其中数据集的影响可能更大，因为所有的深度学习模型的训练，都离不开数据本身。优秀的模型会有更强大的学习能力、良好的训练数据会得到更好的训练结果，这些都是被学界和业界公认的事实，也与本实验所得的结果吻合。

网络规模数据训练出来的生成式大模型，在多任务场景都有较好的表现，引发了 AIGC[50] (AI-Generated Content)在交叉应用领域的火热发展。优秀的开源模型、高质量的数据集源源不断地被提出，又被进一步在生产生活和科研中使用，促进更新的模型和数据集产生，两者形成良性循环。但针对农业领域，目前智能化应用普及率仍然较低，这与农业数据的难获取、从业人员的受教育水平参差不齐、社会关注度较低等因素都有关系。虽然目前很多的通用的优秀模型是开源的，但是针对某个具体的应用领域时，这些模型往往需要使用该应用领域的数据，进行一定程度的再训练和微调，才能真正投入使用。而最终模型在使用时能够达到的效果，与再训练时使用的数据质量有着直接关联。少数被迁移运用到农业领域的小模型，往往因为数据原因，只针对一些很局限的应用场景，且并不开源。保护模型的知识产权是合理的，但这对于普及应用和科学研究并不利。

目前农业领域有一些知名开源数据集，例如 PlantVillage、Plantdoc 等，但这些数据集大多为几种常见的经济作物——苹果、蓝莓、葡萄、桃子、番茄、土豆、辣椒等，我国较多的小麦、水稻等粮食作物则较为少见[51]，并且关于每种作物的病害资料也并不齐全，使用者不知道数据集中是否已经包含了常见的主要病症，但大多数的农业病虫害检测研究又基本是基于这些开源数据实现的[8,52-53]。农业数据的不全面、不专业、不充足是影响农业大模型构建、阻碍农业领域科研和应用发展创新的重要原因之一。

开源的高质量农业数据之所以少，与其获得成本还有获取难度有关[11]。在各种与 AI 方向交叉的领域中，农业是较少获得社会关注的应用方向，所以相关科研投入的经费、人力等都有限。为农业 AI 研究构造一个完整的开源数据平台，目前对于任何一个组织或个人来说，都是难以独立完成的。因此，需要探索一种合理的模式，能够举众人之力实现农业数据共享，构建全面专业的农业数据平台。

虽然高质量的农业数据难以获得，但也还是不断有学者们和商业组织从事农业领域的研究，并根据自己的研究方向构造需要的数据集，或者是相关的农业机构因生产生活需要，保存有一些农作物数据资料。但这些私人组织或机构的数据集往往并不开源。这些私有数据被使用完后，因为并没有开源，成为一次性消耗品，是极大的浪费。想要从零开始收集数据打造一个农业 AI 开源数据平台是不现实的，但是可以考虑将那些没有开源的“一次性数据”利用起来。这些“一次性数据”往往具有良好的标签和图片质量，是极具有价值的，而它们不开源的原因大多是因为获取成本较高、需要保护数据的安全性等。如果能在不公开数据的情况下，可以让其用于数据需求方的训练，并让数据提供方获得收益，那么“一次性数据”就有了被再次利用的可能性，从而可以为农业数据共享平台的建立提供支持。

联邦学习框架为构建农业数据共享平台提供了新思路，但也存在一些不同于开源数据平台的新挑战。挑战主要存在于以下三个方面：首先是安全问题。联邦学习框架虽然采用了多种隐私保护技术来克服分布式模型训练时的数据和模型的隐私安全问题，但是安全问题层出不穷，联邦学习框架的安全防护是与时俱进、不可间断的过程。其次是智能合约中数据定价机制的合理性问题。设想中的基于联邦学习框架的数据共享平台，供需双方根据数据为模型中间参数的提升效果来完成交易。但是对于模型效果的提升如何进行量化评价目前没有定论，这也许与模型本身的参数量、数据质量等多方面因素有关，需要进一步探索。并且如何将量化过程通过智能合约实现，以及如何以合理的数据定价机制进行交易、交易过程中应该遵循怎样的经济规律也需要进一步探讨。最后是农业数据共享

平台生态链的建设问题。农业数据共享平台的提出，是为了更好地促进农业领域的应用和科研发展，进而提供惠农、便农服务，在减轻人力负担的条件下实现农产品增产增收。人工智能已经融入各个领域，为各行各业带来了巨大变革，未来这种融合趋势和影响势必还将继续扩大。农业领域由于从业人员的受教育水平低，目前与人工智能的结合程度尚浅。结合分布式自主组织和数字智能[18,54]，为农业数据共享平台挖掘开辟一条生态链，为平台找到现实价值，是保证数据共享平台生机不断的必要条件。

联邦学习框架可以很好地解决农业数据共享平台构建中，分布式模型训练时面临的数据隐私安全问题。但是基于农业领域的科研应用生态环境，联邦学习框架的实现与部署仍存在以下挑战尚待进一步研究：（1）如何针对农业数据共享的应用场景，实现联邦学习框架部署时的数据安全；（2）如何利用智能合约和定价机制，完成非开源农业数据共享平台中的交易；（3）如何为农业数据共享平台构建相应的生态链环境。如上问题的解决是保障农业元宇宙形成和运营的重要支撑[17]。

6 结论

本研究使用四种经典卷积神经网络，在两个数据集上，通过迁移学习进行了水稻病虫害识别。对结果进行对比分析后，验证了除模型优化结构以外，训练数据本身对于模型训练结果的影响和重要性。与已有的相关研究工作相比，本研究主要针对农业模型训练中训练数据量不够以及训练数据质量不高的问题，结合农业领域的实际情况，提出了构建基于联邦学习框架的农业数据共享平台这一解决方案。

参考文献

[1] ARNAL BARBEDO J G. Digital image processing techniques for detecting, quantifying and classifying plant diseases[J/OL]. SpringerPlus, 2013, 2(1): 660[2023-08-29]. https://doi.org/10.1186/2193-1801-2-660.

[2] LOWE D G. Distinctive Image Features from Scale-Invariant Keypoints[J/OL]. International Journal of Computer Vision, 2004, 60(2): 91-110[2023-08-28]. http://link.springer.com/10.1023/B:VISI.0000029664.99615.94.

[3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C/OL]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05): 2005: 886-893. https://doi.org/10.1109/CVPR.2005.177.

[4] BAY H, ESS A, TUYTELAARS T, et al. Speeded-Up Robust Features (SURF)[J/OL]. Computer Vision and Image Understanding, 2008, 110(3): 346-359[2023-08-28]. https://linkinghub.elsevier.com/retrieve/pii/S1077314207001555.

[5] COVER T, HART P. Nearest neighbor pattern classification[J/OL]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27. https://doi.org/10.1109/TIT.1967.1053964.

[6] CORTES C, VAPNIK V. Support-vector networks[J/OL]. Machine Learning, 1995, 20(3): 273-297[2023-08-29]. https://doi.org/10.1007/BF00994018.

[7] MOHANTY S P, HUGHES D P, SALATHÉ M. Using Deep Learning for Image-Based Plant Disease Detection[J/OL]. Frontiers in Plant Science, 2016, 7[2023-08-28]. https://www.readcube.com/articles/10.3389%2Ffpls.2016.01419.

[8] LU Y, YI S, ZENG N, et al. Identification of rice diseases using deep convolutional neural networks[J/OL]. Neurocomputing, 2017, 267: 378-384[2023-08-28]. https://linkinghub.elsevier.com/retrieve/pii/S0925231217311384.

[9] LIU Z, GAO J, YANG G, et al. Localization and Classification of Paddy Field Pests using a Saliency Map and Deep Convolutional Neural Network[J/OL]. Scientific Reports, 2016, 6(1): 20410[2023-08-28]. https://www.nature.com/articles/srep20410.

[10] 庄福振,罗平,何清,等.迁移学习研究进展[J]. 软件学报. 2015,26(1): 26−39.

[11] PAN S J, YANG Q. A Survey on Transfer Learning[J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359. https://doi.org/10.1109/TKDE.2009.191.

[12] SAMANTA S, TIRUMARAI SELVAN A, DAS S. Cross-domain clustering performed by transfer of knowledge across domains [C/OL]//2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). 2013: 1-4. https://doi.org/10.1109/NCVPRIPG.2013.6776213.

[13] WANG J, KE L. Feature subspace transfer for collaborative filtering [J/OL]. Neurocomputing, 2014, 136: 1-6[2023-08-29]. https://linkinghub.elsevier.com/retrieve/pii/S0925231214002446.

[14] HUA J, WANG H, KANG M, et al. The design and implementation of a distributed agricultural service system for smallholder farmers in China[J/OL]. International Journal of Agricultural Sustainability, 2023, 21(1): 2221108[2023-09-07]. https://www.tandfonline.com/doi/full/10.1080/14735903.2023.2221108.

[15] HUA J, WANG H, KANG M, et al. The design and implementation of a distributed agricultural service system for smallholder farmers in China[J/OL]. International Journal of Agricultural Sustainability, 2023, 21(1): 2221108[2023-09-07]. https://www.tandfonline.com/doi/full/10.1080/14735903.2023.2221108.

[16] KANG M, WANG X, WANG H, et al. The Development of AgriVerse: Past, Present, and Future[J/OL]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(6): 3718-3727. https://doi.org/10.1109/TSMC.2022.3230830.

[17] WANG X, KANG M, SUN H, et al. DeCASA in AgriVerse: Parallel Agriculture for Smart Villages in Metaverses[J/OL]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(12): 2055-2062. https://doi.org/10.1109/JAS.2022.106103.

[18] 康孟珍,王秀娟,李冬,等. 基于联邦学习的分布式农业组织[J]. 智能科学与技术学报, 2022,4(2): 288-297.

[19] PRAJAPATI H B, SHAH J P, DABHI V K. Detection and classification of rice plant diseases[J/OL]. Intelligent Decision Technologies, 2017, 11(3): 357-373[2023-09-21]. https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/IDT-170301.

[20] 王忠培, 张萌, 董伟, 等. 基于迁移学习的多模型水稻病害识别方法研究[J].安徽农业科学, 2021, 49(20):236-242.

[21] TAKAHASHI R, MATSUBARA T, UEHARA K. Data Augmentation using Random Image Cropping and Patching for Deep CNNs[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(9): 2917-2931[2023-08-29]. http://arxiv.org/abs/1811.09030.

[22] BJERRUM E J, GLAHDER M, SKOV T. Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics[M/OL]. arXiv, 2017[2023-08-29]. http://arxiv.org/abs/1710.01927.

[23] HAN D, LIU Q, FAN W. A new image classification method using CNN transfer learning and web data augmentation[J/OL]. Expert Systems with Applications, 2018, 95: 43-56[2023-08-29]. https://linkinghub.elsevier.com/retrieve/pii/S0957417417307844.

[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[C/OL]// Advances in Neural Information Processing Systems: 卷 25. Curran Associates, Inc., 2012[2023-08-29]. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[25] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[M/OL]. arXiv, 2015[2023-08-29]. http://arxiv.org/abs/1409.1556.

[26] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[M/OL]. arXiv, 2015[2023-08-29]. http://arxiv.org/abs/1512.03385.

[27] SZEGEDY C, LIU W, JIA Y, et al. Going Deeper with Convolutions [M/OL]. arXiv, 2014[2023-08-29]. http://arxiv.org/abs/1409.4842.

[28] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[M/OL]. arXiv, 2015 [2023-08-29]. http://arxiv.org/abs/1512.00567.

[29] ESI NYARKO B N, BIN W, ZHOU J, et al. Comparative Analysis of AlexNet, Resnet-50, and Inception-V3 Models on Masked Face Recognition[C/OL]//2022 IEEE World AI IoT Congress (AIIoT). 2022: 337-343. https://doi.org/10.1109/AIIoT54504.2022.9817327.

[30] BARBEDO J G A. Factors influencing the use of deep learning for plant disease recognition[J/OL]. Biosystems Engineering, 2018, 172: 84-91[2023-08-29]. https://linkinghub.elsevier.com/retrieve/pii/S1537511018303027.

[31] LIU Q, JIANG Y. Dive into Big Model Training[M/OL]. arXiv, 2022[2023-08-29]. http://arxiv.org/abs/2207.11912.

[32] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling Laws for Neural Language Models[M/OL]. arXiv, 2020[2023-08-29]. http://arxiv.org/abs/2001.08361.

[33] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM[M/OL]. arXiv, 2021[2023-08-29]. http://arxiv.org/abs/2104.04473.

[34] SHELLER M J, EDWARDS B, REINA G A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data[J/OL]. Scientific Reports, 2020, 10(1): 12598 [2023-08-29]. https://www.nature.com/articles/s41598-020-69250-1.

[35] MAMMEN P M. Federated Learning: Opportunities and Challenges [M/OL]. arXiv, 2021[2023-08-29]. http://arxiv.org/abs/2101. 05428.

[36] NGUYEN T V, DAKKA M A, DIAKIW S M, et al. A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data[J/OL]. Scientific Reports, 2022, 12(1): 8888[2023-08-29]. https://www.nature.com/articles/s41598-022-12833-x.

[37] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical Secure Aggregation for Federated Learning on User-Held Data[M/OL]. arXiv, 2016[2023-08-29]. http://arxiv.org/abs/1611.04482.

[38] MOTHUKURI V, PARIZI R M, POURIYEH S, et al. A survey on security and privacy of federated learning[J/OL]. Future Generation

Computer Systems, 2021, 115: 619-640[2023-08-29]. https://www.sciencedirect.com/science/article/pii/S0167739X20329848.

[39] 吴超, 郁建兴. 面向公共管理的数据所有权保护, 定价和分布式应用机制探讨[J]. 电子政务, 2020 (1): 29-38.

[40] DENHARDT R B, DENHARDT J V. The New Public Service: Serving Rather than Steering[J/OL]. Public Administration Review, 2000, 60(6): 549-559[2023-09-15]. https://onlinelibrary.wiley.com/doi/10.1111/0033-3352.00117.

[41] bitcoin-zh-cn.pdf[EB/OL]. [2023-09-12]. https://nakamotoinstitute.org/static/docs/bitcoin-zh-cn.pdf.

[42] IUON-CHANG LIN, TZU-CHUN LIAO. A survey of blockchain security issues and challenges[J/OL]. International Journal of Network Security, 2017, 19(5). https://doi.org/10.6633/IJNS.201709.19(5).01.

[43] 袁勇, 王飞跃. 区块链技术发展现状与展望[J]. 自动化学报, 2016, 42(4): 481-494.

[44] KIM H, PARK J, BENNIS M, et al. Blockchained On-Device Federated Learning[M/OL]. arXiv, 2019[2023-09-15]. http://arxiv.org/abs/1808.03949.

[45] ZYSKIND G, NATHAN O, PENTLAND A “Sandy”. Decentralizing Privacy: Using Blockchain to Protect Personal Data[C/OL]//2015 IEEE Security and Privacy Workshops. 2015: 180-184. https://doi.org/10.1109/SPW.2015.27.

[46] WANG S, DING W, LI J, et al. Decentralized autonomous organizations: Concept, model, and applications[J/OL]. IEEE Transactions on Computational Social Systems, 2019, 6(5): 870-878 [2023-03-31]. https://ieeexplore.ieee.org/document/8836488/.

[47] DING W, HOU J, LI J, et al. DeSci Based on Web3 and DAO: A comprehensive overview and reference model[J/OL]. IEEE Transactions on Computational Social Systems, 2022, 9(5): 1563-1573[2023-03-31]. https://ieeexplore.ieee.org/document/9906878/.

[48] DING W W, LIANG X, HOU J, et al. Parallel Governance for Decentralized Autonomous Organizations enabled by Blockchain and Smart Contracts[C/OL]//2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI). Beijing, China: IEEE, 2021: 1-4[2023-03-31]. https://ieeexplore.ieee. org/document/9540069/.

[49] HOU J, DING W, LIANG X, et al. A Study on Decentralized Autonomous Organizations Based Intelligent Transportation System enabled by Blockchain and Smart Contract[C/OL]//2021 China Automation Congress (CAC). Beijing, China: IEEE, 2021: 967-971 [2023-03-31]. https://ieeexplore.ieee.org/document/9727429/.

[50] CAO Y, LI S, LIU Y, et al. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT[M/OL]. arXiv, 2023[2023-08-29]. http://arxiv.org/abs/2303.04226.

[51] WANG D, WANG J, LI W, et al. T-CNN: Trilinear convolutional neural networks model for visual detection of plant diseases[J]. Computers and Electronics in Agriculture, 2021, 190: 106468.

[52] Automated Identification of Northern Leaf Blight-Infected Maize Plants from Field Imagery Using Deep Learning[EB/OL]. [2023-08-29]. https://apsjournals.apsnet.org/doi/epdf/10.1094/PHYTO-11-16-0417-R.

[53] AGARWAL M, SINHA A, GUPTA S Kr, et al. Potato Crop Disease Classification Using Convolutional Neural Network[C/OL]//SOMANI A K, SHEKHAWAT R S, MUNDRA A, et al. Smart Systems and IoT: Innovations in Computing. Singapore: Springer, 2020: 391-400. https://doi.org/10.1007/978-981-13-8406-6_37.

[54] WANG Y F, KANG M Z, LIU Y L, et al. Can digital intelligence and cyber physical social systems achieve global food security and sustainability?[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(11): 2070-2080.

引用格式：张濛濛,王秀娟,康孟珍,华净,王浩宇,王飞跃. 从水稻病害识别出发探索农业数据共享新模式[J].农业大数据学报,2023,5(4):13-23.

CIATION: ZHANG MengMeng, WANG XiuJuan, KANG MengZhen, HUA Jing, WANG HaoYu, WANG FeiYue. A Novel Agricultural Data Sharing Mode Based on Rice Disease Identification[J]. Journal of Agricultural Big Data, 2023, 5(4):13-23.

A Novel Agricultural Data Sharing Mode Based on Rice Disease Identification

ZHANG MengMeng[1,2], WANG XiuJuan[1,3], KANG MengZhen[1,2*], HUA Jing[1,3], WANG HaoYu[1,3], WANG FeiYue[4,5]

1. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China; 2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China; 3. Beijing Engineering Research Center of Intelligent Systems and Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; 4. Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078, China; 5. National Key Laboratory of Complex System Management and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Accurate and efficient identification of crop diseases can enable farmers to take effective and targeted preventive measures in a timely manner, which is helpful to reduce the risk of yield reductions and economic losses caused by crop diseases. However, the recognition model that can achieve the effect of SOTA in other fields, especially in the application of rice disease identification, faces the challenge of insufficient available rice disease data, a limited range of disease varieties and low data quality. In this paper, a variety of classical convolutional neural networks are trained on two different datasets using transfer learning methods. We demonstrated that in addition to the optimization achieved through model structure, the training data set itself has an important impact on the training results. However, the scarcity of open-source agricultural data, coupled with the absence of a comprehensive open-source data sharing platform, remains a substantial obstacle. This issue is closely related to the difficulty and high cost of obtaining high-quality agricultural data, low level of education of most employees, underdeveloped distributed training systems and unsecured data security. To solve those challenges, this paper proposed a novel idea to construct an agricultural data sharing platform based on federated learning framework, aiming to address the deficiency of high-quality data in agricultural field training.

Keywords: rice disease identification; convolutional neural networks; distributed training; federated learning; open-source data sharing platform