（四）数据介绍

1.数据来源

（1）预训练数据集

预训练数据集的训练集和验证集来自Kaggle网站的“Dataset of Tomato Leaves”数据集的训练集和验证集，地址为<https://www.kaggle.com/datasets/sunilgautam/dataset-of-tomato-leaves>。该数据集包括9种染病番茄叶片图像和1组健康番茄叶片图像，共11617张图片，按3:1的比例划分为训练集和验证集，该数据集主要用于预训练过程，以便于寻找到模型训练的最佳超参数，以提高后续训练的效果和效率。

（2）番茄叶片数据集

为分析诊断番茄叶片相关疾病并建立相应模型，本文以Kaggle网站上的“Tomato Leaves Dataset”数据集为研究对象，地址为<https://www.kaggle.com/datasets/ashishmotwani/tomato/data?select=valid>。研究分类了Bacterial\_spot、Early\_blight等十种番茄叶片疾病。该数据集图片从实验室与野外环境中收集，共收录了32534张图片数据，并将其按照疾病类型分为11组（其中10组为疾病样本，一组为健康样本），每组样本容量均在1000张以上。

**表一 “Tomato Leaves Dataset”数据集各标签数据数量一览表**

|  |  |  |  |
| --- | --- | --- | --- |
| 图像数据标签 | 训练集数据数量 | 验证集数据数量 | 总计 |
| Bacterial\_spot | 2826 | 732 | 3558 |
| Early\_blight | 2455 | 643 | 3098 |
| Healthy | 3113 | 792 | 3905 |
| Late\_blight | 2754 | 739 | 3493 |
| Leaf\_Mold | 2882 | 746 | 3628 |
| Septoria\_leaf\_spot | 1747 | 435 | 2182 |
| Spider\_mites Two-spotted\_spider\_mite | 1827 | 457 | 2284 |
| Target\_spot | 2039 | 498 | 2537 |
| Tomato\_mosaic\_virus | 2153 | 584 | 2737 |
| Tomato\_yellow\_leaf\_curl\_virus | 3051 | 805 | 3856 |
| powdery\_mildew | 1004 | 252 | 1256 |

模型的外部测试集同样来自Kaggle网站的“Dataset of Tomato Leaves”数据集的测试集（已确认该测试集与预训练模型的训练集、验证集及无交集），测试集共包含2914张番茄叶片图像。该测试机用于训练后模型的功能测试，以确保模型能够在没见过的数据上进行准确的诊断，

2.数据特征与划分

数据集中的32534张图片数据分别以.jpg、.png、.jpeg文件的格式存储，其中.jpg格式32025张，.png格式505张，.jpeg格式4张。

在训练模型时，我们将整个数据集的约20.5%划分为验证集数据，共6683张图片，用于评估模型性能，并判断训练是否有效。数据集中余下的25851张图片作为训练集，参与模型的训练，模型通过对训练集中数据的分析和学习，不断优化参数，从而提升模型自身的准确性和判断能力。

**表2 番茄叶片数据集划分**

|  |  |  |
| --- | --- | --- |
| 数据集类型 | 图像数量 | 用途 |
| 预训练训练集 | 8715 | 寻找模型训练的最佳超参数 |
| 预训练验证集 | 2902 | 寻找模型训练的最佳超参数 |
| 训练集 | 32025 | 模型训练 |
| 验证集 | 6683 | 评估模型性能 |
| 外部测试集 | 2914 | 模型功能测试 |

3.数据预处理

（1）训练集数据预处理

训练集数据的预处理主要包括统一尺寸、随机变化、图像归一化和标准化四个步骤。研究中中，由于数据集中图像数据的尺寸不一，首先要将图像数据大小统一为同一尺寸，便于模型训练。接着，利用旋转、平移、缩放、翻转以及图像调节等操作，提高数据多样性。随后，将图像归一化为张量格式并进行标准化处理，使图像数据更适合模型训练。通过以上操作，增强数据特征，提高数据多样性，防止模型过拟合。

（2）验证集数据预处理

验证集数据的预处理主要包括裁剪、统一尺寸、图像归一化和标准化操作。目的是保持验证集数据的一致性与标准性，以便准确评估模型性能。