图像分类中的特征提取分析：综合研究

Ammara Khan[1], Muhammad Tahir Rasheed[2], Hufsa Khan[2,*]

[1] 旁遮普大学植物学研究所，巴基斯坦拉合尔 51000
[2] 深圳大学计算机科学与软件工程学院，广东深圳 518060

摘要： 已经有许多利用深度学习的实际应用，特别是在图像分类领域。一个常见的发现是，某些领域数据高度倾斜，这意味着大部分信息属于少数多数类别，而少数类别中的信息很少或没有。重要的是要承认，倾斜的类别分布对机器学习算法构成了重大挑战。因此，在数据分布不平衡的情况下，大多数机器和深度学习算法在高度不平衡时无效或可能失败。在本研究中，考虑基于深度学习的知名模型，对不平衡数据集进行全面分析。特别是，确定了最佳特征提取器模型，并检查了最新特征提取模型的当前趋势。此外，还进行了 1991 年至 2022 年的文献计量分析，以确定全球关于不平衡蘑菇数据集图像分类的科学研究。总之，我们的研究结果可以为研究人员提供快速基准测试参考和替代方法来评估图像分类研究中不平衡数据分布的趋势。

关键词： 蘑菇分类；文献计量分析；特征提取；分类准确率

DOI: 10.57237/j.cst.2023.04.001

An Analysis of Feature Extraction in Image Classification: A Comprehensive Study

Ammara Khan[1], Muhammad Tahir Rasheed[2], Hufsa Khan[2,*]

[1]Institute of Botany, University of the Punjab, Lahore 51000, Pakistan
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

Abstract: There have been many real-life applications utilizing deep learning, especially in the area of image classification. A common finding is that some domain data are highly skewed, which means that most of the information belongs to a small number of majority classes, and there is little or no information in the minority classes. Due to which in case of imbalanced data distribution, the majority of machine and deep learning algorithms are not effective or may fail when it is highly imbalanced. In this study, a comprehensive analysis of imbalanced dataset is conducted by considering deep learning-based well-known models. In particular, the best feature extractor model is identified and the current trend of latest feature extraction model is examined. Moreover, a bibliometric analysis is carried out from 1991 to 2022 in order to identify the global scientific research on the image classification of imbalanced mushroom dataset. In summary, our findings may offer researchers a quick benchmarking reference and alternative approach to assessing trends in imbalanced data distributions in image classification research.

Keywords: Mushroom Classification; Bibliometric Analysis; Features Extraction; Classification Accuracy

*通信作者: Hufsa Khan, hufsakhan@email.szu.edu.cn

收稿日期: 2023-08-12; 接受日期: 2023-10-07; 在线出版日期: 2023-10-09

1 引言

目前，不平衡数据的分类已成为数据挖掘领域最常见的问题之一。当数据不平衡时，标准分类模型表现不佳。由于最近信息技术发展到大数据时代，大量数据正在被处理。在许多领域，机器学习，模式识别和数据挖掘方法被广泛用于处理大量数据。研究表明，有偏差的分布会导致各种领域的问题，包括信用卡欺诈，软件缺陷预测，文本挖掘，模式识别和基因挖掘。在上述领域中，正确地对少数实例进行分类至关重要，因为少数实例的错误分类比多数实例的成本更高[1]。此外，许多传统分类器，如逻辑回归，决策树和支持向量机，都是为了最大限度地提高分类精度而设计的，因此少数类可能被归类为多数类[2]。

作为这项研究的一部分，我们特别关注蘑菇的数据。蘑菇是人类良好的营养来源。如今，蘑菇之所以如此受欢迎，是因为它们具有高营养价值，包括维生素 B，C 和 D，铜，β-葡聚糖，钾，硒，钠，钙，磷，镁等矿物质以及大量蛋白质[3]。然而，蘑菇的识别和分类一直是不同知识领域许多应用和研究的主题，尤其是在区分有毒蘑菇和可食用蘑菇时。通常，有两种方法，可以通过使用图像数据，也可以通过使用属性。第一个是计算机视觉研究，将各种算法和机器学习技术应用于蘑菇图像。在该算法中，从图像域中提取的特征被用于对没有背景的蘑菇图像进行分类。此外，在这个特定的知识领域中，已经研究了这种方法和算法的许多变体，如神经网络（NN），支持向量机（SVM），决策树或 k 近邻（kNN）[4-6]。其次，基于属性的研究的动机要么是毒性与可食性的生物学问题，要么是评估新的算法和分类器[7, 8]。

蘑菇数据的可食用或有毒分类已经使用了许多方法。然而，有一点需要注意，在现实生活中，现有的数据可能并不总是完美的。它们可能包括缺失数据[9, 10]，异常值，噪声数据或不平衡数据[11]。在数据集严重不平衡的情况下，传统的二进制或多类分类会导致偏向于实例数量高得多的类。因此，在这种情况下，建模和检测少数类的实例是极其困难的。这样，在预测的可食用或有毒蘑菇的标签中可能会出现歧义。为了解决这个问题，可以应用数据集的初步重新采样，例如向数据集添加新元素或删除旧元素。重新采样的方法多种多样，因此选择最好的方法可能很有挑战性。此外，特征提取也有助于提高分类精度。为了对图像进行分类，用于特征提取的模型能够有效地从图像中提取重要特征是至关重要的。

在这项研究中，受深度学习在各个领域的广泛应用的启发，如图像增强[12-14]，缺失数据[10, 15]，生物数据，面部识别数据，我们探索了使用基于深度学习的模型从蘑菇数据集中提取特征的可能性，以确认蘑菇数据不平衡时的分类准确性。此外，我们还提出了一些实用指南，并解释了不平衡蘑菇数据如何影响基于深度学习的特征提取模型的性能。本文探索并分析了几种基于深度学习的特征提取模型，以确定最佳的特征提取器模型。随后，对最佳特征提取器模型进行了实验测试。在通过使用上采样，下采样和混合采样应用重采样时，根据最佳特征提取器模型的性能对其进行了进一步研究。它概述了这些局势的关键方面及其带来的挑战。

此外，本研究还进行了文献计量分析，以确定 1991 年至 2022 年“蘑菇分类”类别的研究活动状态，观察过去十年的演变和增长。这项研究的意义在于，这项研究对从事蘑菇分类工作的科学家来说很重要，因为它涵盖了国际范围内的科学生产研究。通过文献计量分析，研究人员可以更好地了解蘑菇分类研究的趋势，这可以作为未来研究的基准。据我们所知，在文献中，没有基于类不平衡数据的蘑菇分类方法，也没有对蘑菇分类进行文献计量学研究。

综上所述，本文的主要贡献如下：

1. 确定了一个最佳特征提取模型，并通过实验研究了该模型对重采样的影响。此外，还分析了基于深度学习的算法性能与不平衡类分类率之间的关系。
2. 在蘑菇分类领域，进行了文献计量学分析。本研究旨在评估 1991 年至 2022 年蘑菇分类的研究活动状况及其演变和增长。
3. 实验研究结果表明，即使使用了公认的特征提取模型，不平衡分类也会影响模型的性能。

论文的其余部分组织如下：第 2 节简要回顾了现有文献中的相关工作。在第 3 节中，我们进行了文献计量学分析。在不平衡蘑菇分类数据集的情况下，最佳特征提取器模型如第 4 节所示。第 5 节介绍了实验装置，数据集，性能评估标准以及对结果的讨论。最后，我们提请注意今后的工作和第 6 节的结论性意见。

2 文献综述

本节的目的是概述与不平衡分类相关的现有工作，特别是在蘑菇数据集的情况下。最初，蘑菇分类是基于人工评估来识别蘑菇的特征[7]。后来，使用数据挖掘分类方法来区分食用蘑菇和毒蘑菇。然而，随着技术的出现，现代技术得到了发展，如数据挖掘，已应用于这一研究领域[16, 17]。使用称为 WEKA 的分析工具对三种监督学习算法（即朴素再见、支持向量机（SVM）和决策树（DT））进行比较[18]。北美蘑菇数据集是从实验结果的在线存储库中考虑的。通过评估训练数据并进行 10 倍交叉验证，他们发现决策树算法和支持向量机算法的准确性完全相同。另一方面，在处理速度方面，决策树被认为是比 SVM 更好的算法。基于这些发现，作者建议使用 DT 对蘑菇进行分类，而 Naive Bayes 在所有测试模型中的准确率最低。此外，无论蘑菇是可食用的还是有毒的，都会对其进行分类[19]。研究了许多分类算法，但决策树方法表现最好。

另一项研究使用相同的方法，使用不同的分类器模型（即朴素贝叶斯，ZeroR 和贝叶斯）对蘑菇数据集进行分类。使用 WEKA 工具测试网络分类器，发现 ZeroR 是与其他分类器相比最差的分类器，具有最低的准确度和精度[20]。研究还发现，与其他分类器相比，朴素贝叶斯是 ZeroR 之后精度和精度最低的最差分类器。此外，在另一项研究中，研究并比较了三种不同算法的性能，包括朴素贝叶斯，ripple-down rule（RIDOR）和贝叶斯网络，以及提，乞求和堆叠集成分类器 [21]。结果表明，机器学习的分类比手工排序更快、更准确。因此，这些方法展示了基于机器学习的方法的技术能力，并突出了它们在几个领域的应用潜力。在另一项研究中，作者强调了使用逻辑回归，支持向量机和多粒度识别有毒蘑菇的级联森林分类器的重要性[22]。作为多粒度级联森林的结果，实现了最高的精度。P. Maurya [6]，使用基于机器学习方法的五种不同分类器来使用纹理特征对蘑菇进行分类。使用了不同的监督分类器算法，包括支持向量机（SVM），K 近邻（KNN），决策树，集成树和判别分析。在这些算法中，SVM 表现出最高的准确度性能，超过了其他提到的分类器。根据 A. Subramaniam [23]，可以使用主成分分析（PCA）算法将食用蘑菇与不可食用蘑菇区分开来。基于训练图像的数量，所提出的方法显示出 85%至 96%的成功率。随着训练图像数量的增加，该方法的性能有所提高。类似地，[24]使用主成分分析（PCA）算法从数据集中提取最佳特征，然后使用决策树算法根据蘑菇的行为特征对其进行分类。该分类是通过使用从 UCI 机器学习库收集的 8124 个具有 22 个行为特征的数据来完成的。在这些特征中，蘑菇的“气味”最高。此外，可以使用朴素贝叶斯、人工神经网络和自适应神经模糊推理系统对各种可食用和不可食用蘑菇进行分类[25]。他们的结果表明，自适应神经模糊推理系统在准确性方面优于其他分类器，而 Naive Bayes 算法仍然是最差的分类器。

在另一项研究中，从互联网上选择了一个数据集来对蘑菇进行分类[5]。实验采用不同的机器学习算法进行，KNN 分类器的准确率最高。在另一项研究中，多层传感器被用来将蘑菇分为有毒和可食用两类[26]。在他们的研究中，使用“JustNN”软件对 8124 个样本实现了显著的准确性。还有其他研究使用多层感知神经网络开发深度学习模型，并开发移动应用程序来识别蘑菇是否有毒[27, 28]。在一项研究中，作者提出了一种蘑菇诊断辅助系统（MDAS）[29]，它主要是一种智能手机设备，具有三种机制，即网络应用程序（服务器），统一数据库和移动应用程序。他们的工作主要集中在幼稚间隔与决策树的分类率上。作为第一步，他们提出的系统选择了最常见的蘑菇，然后根据它们的属性对它们进行分类。蘑菇诊断辅助系统是 Al mejibli [30]为蘑菇采集者开发的另一款移动应用程序，以确保蘑菇采集的安全。此外，他们将决策树分类器的性能与朴素贝叶斯分类器的性能进行了比较，发现决策树在错误测量，正确分类样本和错误分类样本方面具有更好的性能。在另一项研究中，提出了一种快速有效的傅里叶变换红外（FT-IR）光谱和模式识别算法[31]识别有用的蘑菇特征用于分类。

3 文献计量分析

本文献计量分析的目的是确定 1991 年至 2022 年在科学网数据库蘑菇分类研究领域内的研究活动的现状和演变。它已被用作分析出版物分布和特征的统计方法。这种方法使我们能够了解特定领域的演变，同时深入了解该领域的新兴领域[32]。换言之，这类研究可能对所有领域的科学家都有帮助，他们被阅读和参考的期刊选择淹没了，或者希望知道自己的文章可以在哪里发表。我们利用了 Web of Science 数据库中的可用信息，以便在选择过程中尽可能客观。该数据库是文献计量分析和文献综述中最常用和最广泛使用的数

据库之一[33]。该数据库显示了广泛的研究领域[34]。作为一个检索主题，本研究使用了“蘑菇”和“分类”这两个术语来实现文献计量分析的目的。通过这个主题搜索，我们可以识别那些在标题，摘要和/或关键词中包含关键词“蘑菇”和“分类”的出版物。为了确保搜索结果的稳健性，在 1991 年至 2022 年的搜索词中添加了引号。第二种搜索方法是使用搜索关键字“蘑菇”和“分类”和“可食用”来使搜索更加精确和简洁，我们只发现了 89 个结果，这是一个太小的数字，无法进行文献计量分析。当使用第三种方法搜索关键词“蘑菇”和“分类”和“可食用”和“有毒”时，我们只找到了 15 个结果，这不足以进行文献计量分析。因此，本文采用“蘑菇”和“分类”两个关键词来分析这一领域的研究进展。然而，从 1991 年到 2022 年，使用蘑菇分类的搜索参数发现了近 384 篇文章。搜索包括所有类型的出版物，如期刊文章，会议记录，笔记，评论，会议摘要，编辑材料，书籍章节等。因此，为了集中分析 Web of Sciences 中最经典的研究，我们只选择了“期刊文章”和“评论”，导致文献数量从 384 篇减少到 331 篇，如图 2 所示。我们纳入了评论，尽管它们不被认为是重要的科学贡献，因为它们提供了对通常影响未来研究的搜索主题的强烈观点。Web of Science 中的每篇出版物都包含许多详细信息，例如出版年份，作者姓名和地址，出版物的标题和摘要，来源期刊，主题类别和参考文献。一个名为 VOSviewer [35]的免费程序用于分析 384 份出版物的科学网导出数据。该程序分析并可视化出版物输出和引用次数、国家和组织合作、关键词分析以及期刊出版关系。该方法利用 VOSviewer（相似性可视化）映射方法来计算和定位二维地图中的主题，以便两个项目之间的距离反映它们彼此的相似性或相关性。作为 VOSviewer 映射的结果，每个主题被计算并放置在二维地图中，使得两个项目之间的距离准确地反映了它们之间的相似性或相关性。在 VOSviewer 聚类方法中，主题被分组到不同的组中，并且每个组被分配特定的颜色。每个小节都详细解释了如何解释可视化。一般来说，这种解释可以概括如下：圆圈的大小和标签的字体表示出现的次数，颜色表示聚类，两个圆圈之间的距离表示它们的相关性和相似性[32]。地图可以沿 x 轴和 y 轴自由旋转和翻转，x 轴和 y 轴没有特殊含义。

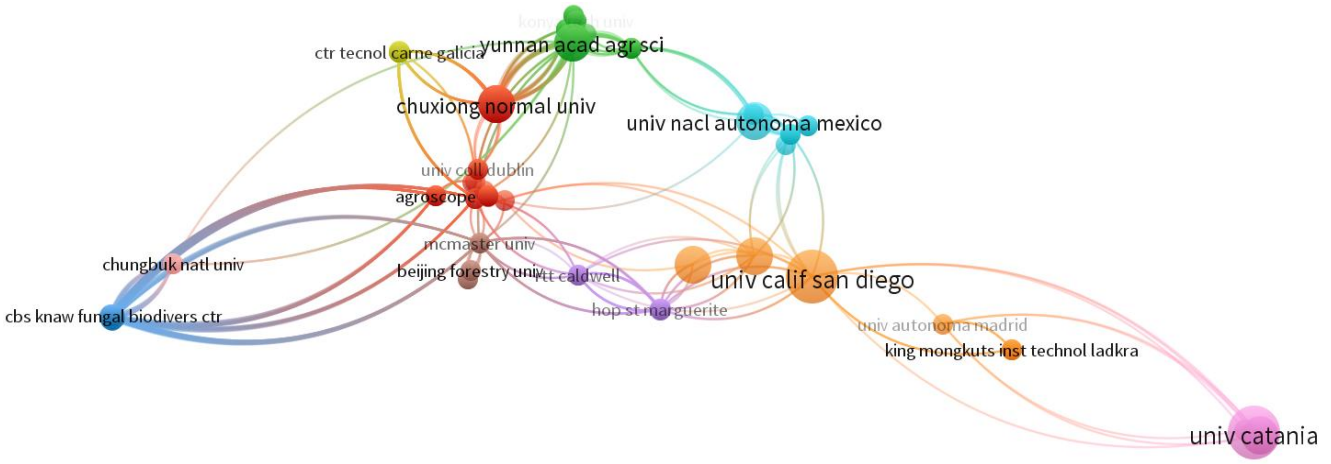


图 1 发布蘑菇分类材料的机构网络

3.1 出版物产出和引文得分

某一研究领域或学科发展趋势的一个重要指标是同行评审出版物的数量。自 1991 年以来，使用蘑菇分类的出版物数量有所增加，如图 2 所示。据统计，目前有 384 种期刊发表了有关蘑菇分类的文章。蘑菇分类的出版物总数达到 384 篇，大约花了 20 年时间（从 1991 年到 2022 年）。从图 2 中可以看出，自 1991 年以来，有关蘑菇分类的出版物数量有所增加。根据这个数字，目前有 384 种期刊发表了有关蘑菇分类问题的文章，而 1991 年，只有一篇出版物讨论了蘑菇分类问题。直到 2015 年，每年都会出版有关该主题的有限出版物。自 2016 年以来，出版物数量每年都在增加，但 2019 年和 2020 年除外，这两个年份出现了下降。此外，2021 年出版物数量达到峰值（n=38）。从统计数据来看，蘑菇研究的引用次数呈上升趋势，如图 2 所示。从图 2 可以看出，1991 年至 2003 年期间，每篇

文章的平均被引次数不到 100 次。从图 2 中可以看出，随着出版物数量的不断增加，蘑菇分类研究变得越来越重要。从图 2 可以看出，蘑菇分类出版物随着时间的推移呈指数增长。

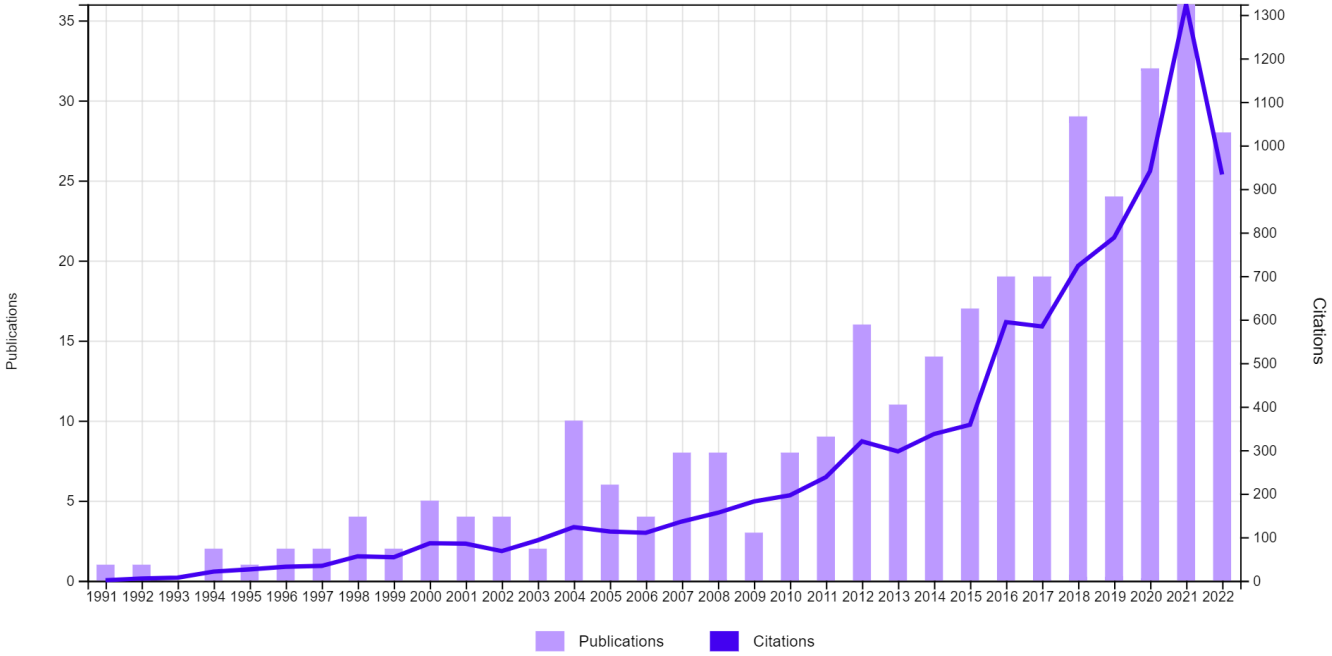


图 2 一段时间内的出版物和引用

3.2 国家和机构文献计量分析

根据 Web of Science 数据，每个出版物都根据出版物中列出的地址分配了一个国家/地区。一个国家在蘑菇分类方面发表的论文数量在一定程度上反映了该国的研究实力。共有 27 个不同的国家参与蘑菇的分类。本研究包括对发表蘑菇分类文章的国家之间的合作网络进行的分析，并使用 VOSviewer 进行网络连接，如图 4 所示。蘑菇领域的国家至少出版了一份有关该主题的出版物。该网络不包括未与其他国家连接的国家。在图 4 中，我们可以看到国家间合作网络的结果。圆圈的大小和出版物的数量之间存在显着的相关性，链接的强度表明了合作的强度。协作集群由颜色表示。可以区分两个主要集群：一个以中国为中心（绿色集群），另一个以美国为中心（蓝色集群，右。聚集在美国周围的集群中，墨西哥是主要贡献者，中国是韩国。德国被认为是第三个国家，较小的集群（棕色集群，右下。在其他科学研究领域发现，合作国家在地理上具有相关性，并且往往聚集在出版物产量最高的国家周围。在蘑菇分类领域，有多家机构从事研究。共有 118 个不同的研究机构参与了 384 篇带有机构信息的出版物（一位作者可能隶属于多个机构，或者一篇文章可能由来自不同机构的多个作者撰写。从图 1 中可以看出，蘑菇分类研究由大约 118 个不同的组织进行。分析结果显示，加州大学圣地亚哥分校和卡塔尼亚大学是全球领先的两所研究机构。总体而言，蘑菇分类的大部分研究发生在美国，意大利和中国（橙色簇在美国，粉色簇在意大利，红色簇在中国）。

3.3 关键词分析

为了更深入地了解蘑菇分类的主题领域和研究趋势，分析该领域出版物的标题和摘要中使用的术语是有用的。通过分析关键词，可以确定蘑菇分类研究的确切背景和主题。所有名词短语均摘自 331 篇有关蘑菇分类的出版物的标题和摘要。搜索包括任何出版物中出现的所有关键字，共识别出 409 个术语。为了可视化研究作者选择的关键词的联系和共现，使用了 VOSviewer 工具。在地图中，每个关键词都由一个圆圈表示，其直径代表其与其他关键词的链接数量。因此，圆圈越大表示关键词之间的链接数量越多。两个圆圈之间的线的粗细表示两个关键字一起出现的频率。下图 3 说明了蘑菇分类研究中最常用的关键词。图 3 说明了蘑菇分类研究中最常用的关键词。该数字代表提取的出版物中关键词共现率的至少五倍。据分析，

分类、蘑菇、真菌、鉴定是蘑菇分类研究中最常用的关键词。如图 3 所示，蘑菇分类中最常用的关键词是“分类”（21 次）。从图 3 中可以清楚地看出，重点放在理解和概念化分类与真菌，分类与蘑菇之间的关系上。本研究发现最常见的关键词是“分类”，这与蘑菇分类文献的计量分析结果一致。

另一方面，在 43 种不同期刊上总共发表了 384 篇出版物。这个领域仍然需要探索，从各个角度研究蘑菇分类的有效性。图 5 显示，用大圆圈表示的期刊即“journal of ethnobiology and ethno”和“food analysis methods” 在蘑菇分类领域的出版物较多。该领域的主要期刊是（Journal of fungi）和（fungal biology），但我们可以看到这些期刊上的出版物较少。在从“医学”到“食品科学”学科类别的期刊中找到有关蘑菇分类的出版物更为常见。在植物学领域，这方面的工作较少。有必要探索这一领域，并需要在相关期刊上发表更多文章，这对研究界有利。

4 方法论

所提出的方法的概述可以在下一节中找到。此外，在分类不平衡的情况下，数据级别被定义为最佳特征提取器模型。在数据级方法中，对训练样本进行操作以平衡类别分布，例如，对少数类别进行过采样、对多数类别进行欠采样或将两者组合。

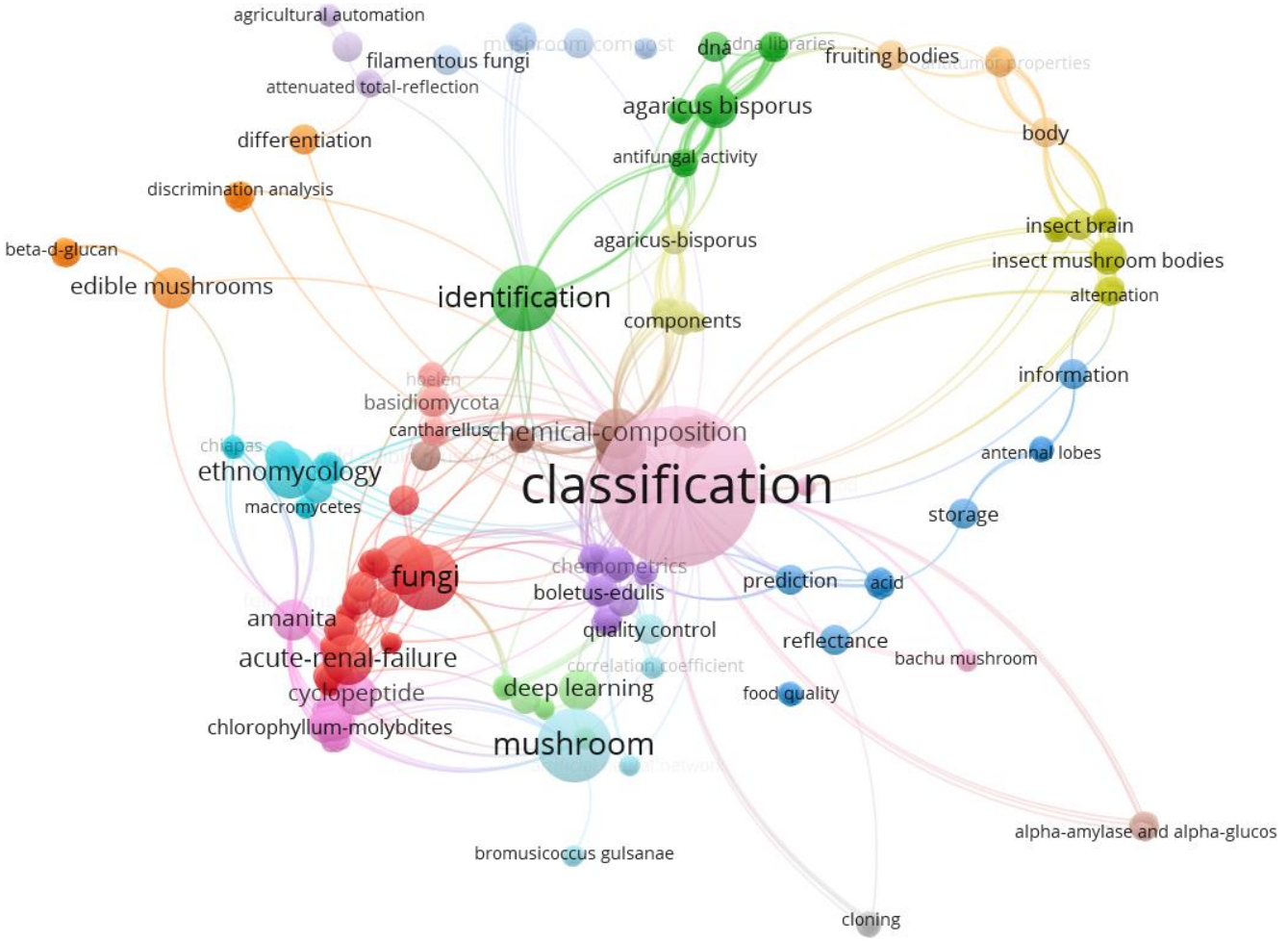


图 3 与蘑菇分类相关的关键词分析

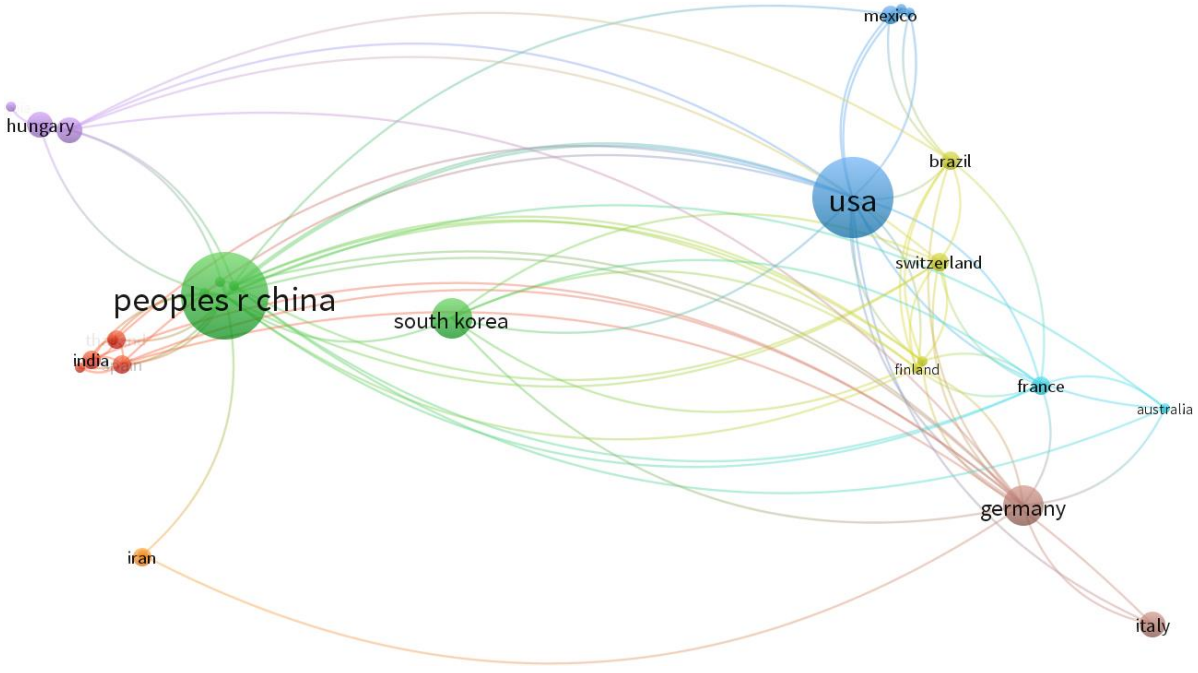


图 4 蘑菇分类合作国家网络

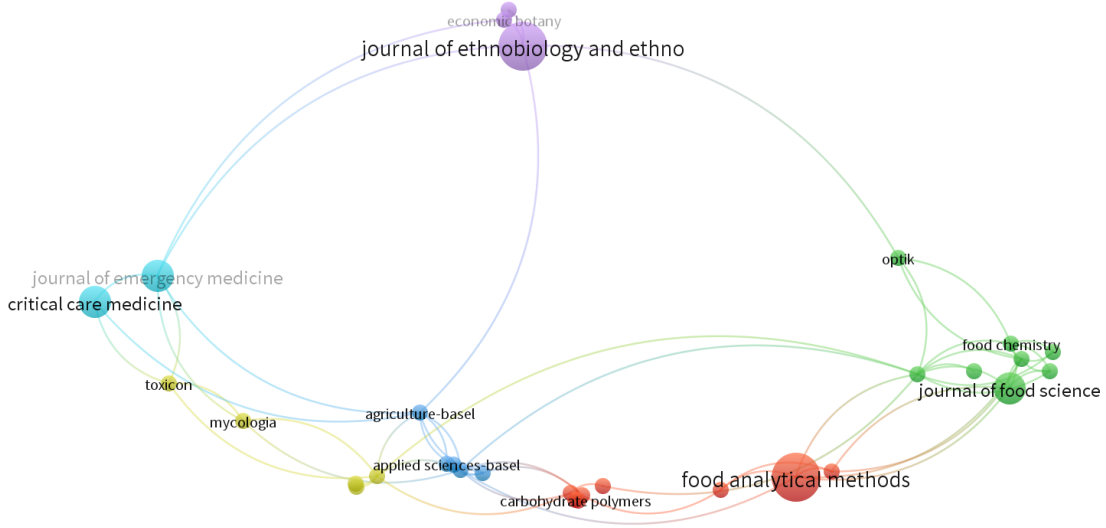


图 5 发表蘑菇分类研究的最活跃期刊概览

4.1 基于特征提取的网络架构

在分类和特征提取方面，机器学习和深度学习得到广泛应用。特征提取阶段之后是分类阶段，其中识别图像并为其分配类别或标签。特征提取阶段减少了数据中的冗余，这是图像处理和对象识别的重要组成部分。简而言之，特征提取是指从原始数据中选择最具代表性的信息的过程，它最大限度地减少了类内模式的可变性，并增加了类间模式的可变性。为了实现这一点，每个类都被分配了一组将其与其他类区分开的特征，同时保持类内特征差异的不变性。本研究的目的是分析不同的模型，以确定在蘑菇分类数据集不平衡的情况下哪个模型可以有效地提取信息最丰富的特征。因此，使用不同的模型（见表 1）来选择和提取特征。

一般来说，食用蘑菇和有毒蘑菇看起来相同，因此很难区分它们。多种有毒蘑菇因其大小和颜色而看似可食用[36]。一种称为霉菌毒素的成分的存在决定了蘑菇是否可以食用[24]。每种蘑菇类型的特征有助于确定蘑菇是否可食用或有毒。与上述类似，在深度学习

模型中，从蘑菇图像中提取的特征可用于区分可食用和有毒的。因此，为了使用深度学习来区分蘑菇，有必要确定能够区分它们的必要特征/特征。只有模型能够有效地提取特征才有可能。预计有效提取特征的模型将提供更好的分类结果。作为从原始输入数据中提取特征的结果，分类阶段利用它们来根据提取的特征的属性来识别特征类别。从所有选定的模型（参见表 1）中，我们试图识别可以从各种蘑菇分类图像中学习丰富特征表示的网络。换句话说，不一定所有众所周知的模型都是所有类型问题的良好特征提取器。

所提出的策略如图 6 所示，其中绿色框表示蘑菇数据被传递到基于深度学习的模型，同时对样本数据集执行特征提取。因此，从该模型中提取的特征已被扩展并在红色框中突出显示，以实现更直观的可视化（见图 6。可以根据提取的特征来确定哪个模型对于提取最佳特征更有效。一旦提取出特征，它们就会被传递到神经网络模型进行分类（见图 6）。模型提取的特征被传递到扁平化层，扁平化层又被传递到全连接层。最后一层是 soft-max 层，如图 6 所示，它将图像分类为适当的类别。为了从每个模型中提取特征并根据其适用类别进行分类，每个模型都遵循相同的过程。根据此分析的结果，确定最佳特征提取器模型。根据图 6，绿色框代表不同的模型，每次在该框中使用不同的模型，而其余设置被认为对于所有模型都是相同的。使用从模型中提取的特征，我们决定当数据不平衡时哪个模型更有效，以提取最佳特征。

4.2 重采样数据的模型分析

首先，我们分析了不平衡数据集的最佳特征提取模型，然后在对数据集重新采样后进一步确认了其有效性。现实世界的二元分类任务（例如从图像中检测对象）通常涉及不平衡的数据集，这意味着某个类（次要类）的代表数量少于另一个类。对次要类别做出准确的预测通常很重要，但这可能非常具有挑战性，因为有关次要类别的可用信息非常少。可以通过首先对数据集重新采样（即添加或删除现有元素）来处理此问题。有许多方法可用于重采样，这就提出了哪种方法最合适的问题。重采样方法有多种类型，包括过采样、欠采样和混合采样。在过采样期间，会复制少数类的实例，以匹配多数类的数据大小。相反，欠采样涉及随机删除多数类的一些实例，以适应少数类的大小。在混合采样方法中，同时考虑过采样和欠采样。混合方法由于平衡了少数类和多数类之间的实例数量，从而显着提高了分类性能。我们的论文通过实验研究了重采样对分类精度的影响。

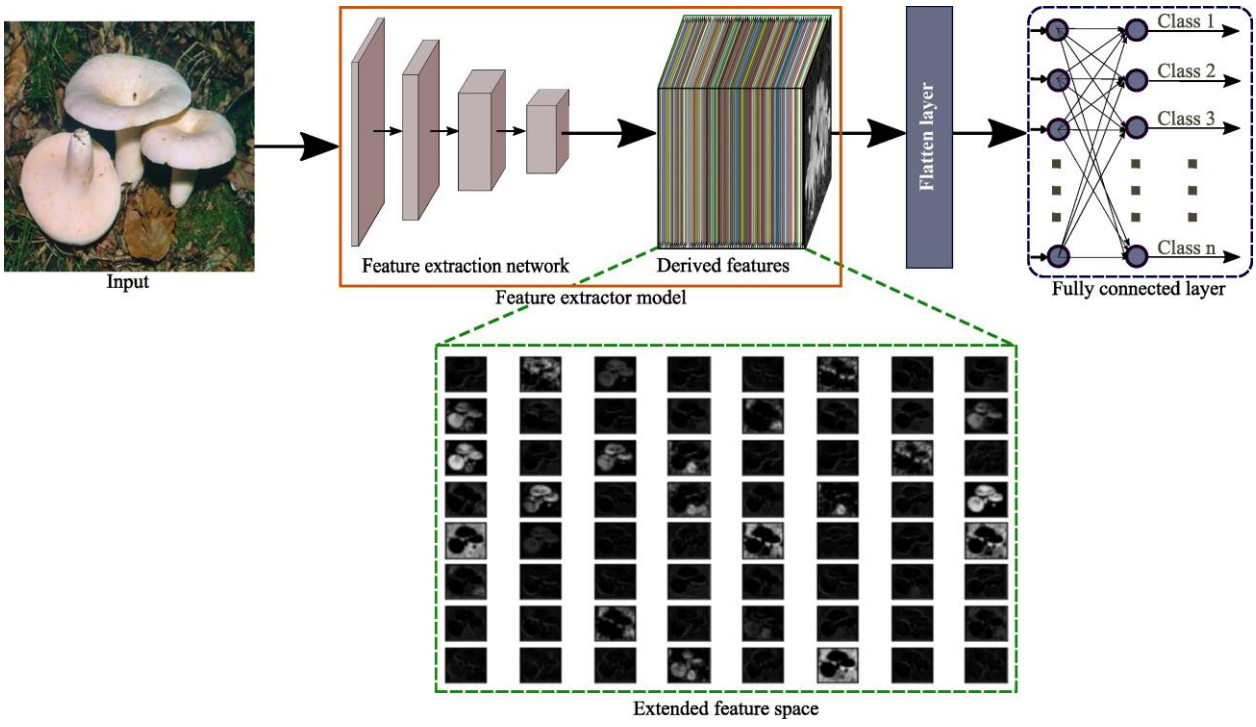


图 6 特征提取过程概述

5 实验结果与讨论

5.1 实验框架

本节介绍对从 Kaggle [37]机器学习存储库中选择的蘑菇数据集进行的实验结果。本节包含对所选数据集、实验设置、性能评估和结果讨论的详细描述。本研究的目的是分析数据不平衡情况下的最佳特征提取器模型，并验证其重采样的有效性。我们比较了十种最先进的基于深度学习的分类方法，即 CNN Decoder，ResNet50V2 [38]，InceptionV3 [39]，DenseNet201 [40]，DenseNet121 [41]，EfficientNetB0 [42]，Xception net [43]，EfficientNetB6 net [44]，MobileNetV2net [45]，和 VGG16 net [43]（在数据集不平衡的情况下）。此外，所有这些使用蘑菇数据训练的网络的性能评估都是使用准确度，精确度，召回率，F 分数和 Cohen 的 kappa 统计数据进行的。这些指标的详细描述可以在第 5.4 节中找到，并且在第 5.3 节中提供了实验和训练设置的描述。最后，我们通过比较不同算法获得的分类精度并对结果进行讨论来结束我们的研究。

5.2 数据集

在本研究中，为了进行实验，蘑菇数据集是从 Kaggle 机器学习存储库获得的[37]。该数据集由蘑菇图像组成，这些图像被认为是最常见的北欧蘑菇天才。该数据集包含 6,714 张图像，其中包括来自伞菌属，鹅膏属、牛肝菌属，Cortinarius，Entoloma，Hygrocybe，Lactarius，Russula 和 Suillus 属的样本。每个类别图像的样本如图 8 所示。此外，图 7 显示了 6714 个蘑菇样本中属于每个类别的蘑菇总数。图 7 表明大多数蘑菇属于乳杆菌属和牛肝菌属。这是因为这两种蘑菇的采集范围很广，并且具有多种口味和质地。因此，大多数可食用，不可食用和有毒蘑菇因其多样性而分为这两类。由于在各个领域中，大量数据正在逐渐产生，因此确保收集到的数据可靠且有价值非常重要，因为质量差的数据无法用于生成可靠的模型，并且会对分类准确性产生负面影响。文献中最近提出了几种方法来提高数据质量差的情况下的分类精度[9, 10, 15]。为了保证数据处理结果无差错，预处理是提高数据质量的重要步骤。在挖掘过程中，数据可能无法以最佳状态提供。因此，需要对数据进行处理才能获得显着的性能。由于所选数据集中图像的大小与系统的格式不兼容，因此必须对其进行预处理才能使用。为此，在预处理阶段，将每个样本的大小设置为 256×256×3。此外，为了确保特征域的一致性，我们应用归一化来避免大尺度值的破坏性影响。

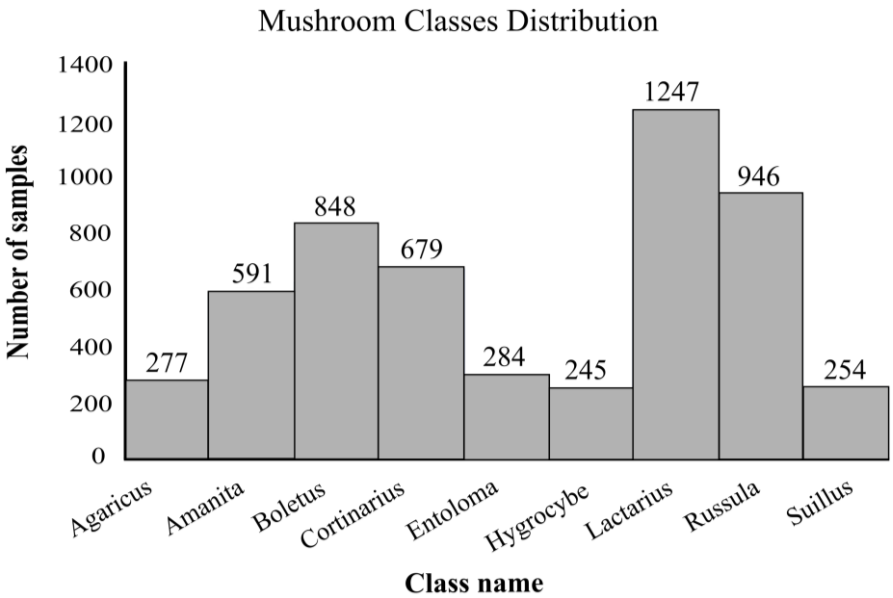


图 7 插图显示了所选蘑菇分类数据集中属于每个类别的蘑菇总数

5.3 实施和培训细节

本节描述用于训练/微调最先进模型的实验设置。最初，整个数据集分为两部分，即训练数据和测试数据，分别占 80%和 20%。这 80%的训练数据用于模型的训练/微调，而测试数据集被保留用于分析算法的效率。在本研究中，比较了十个模型，即 ResNet50V2 [38]，InceptionV3 [39]，DenseNet201 [40]，DenseNet121 [40]，EfficientNetB0 [41]，Xception net [42]，EfficientNetB6 net [43]，MobileNetV2 net [44]和 VGG16 网络[45]已经使用预训练的权重对选定的蘑菇数据集进行了微调。而 CNN Decoder 和 Vision Transformer（VIT）则经过训练。在微调过程中，每 25 个 epoch 后保存上述模型并在测试数据集上进行测试。这些模型使用 256×256 的 patch 大小进行训练和微调，并使用 Adam 优化器优化交叉熵损失函数。我们使用 Tensorflow 2.4 实现了这些架构，并在 NVIDIA Titan Xp GPU 上进行了训练/微调。所有模型都使用 10−4 的学习率训练/微调 1000 个时期。图 10 说明了不同模型所达到的精度。通过对上采样、下采样和混合采样训练数据分别进行微调，进一步分析最佳模型（即 VGG16 网络）。

5.4 绩效评估标准

我们的研究使用几个众所周知的指标来评估模型的性能，包括准确度，精确度，召回率，F-score，Cohen's Kappa。这项研究的目的是确定模型如何有效地区分不同类型的蘑菇。值得注意的是，较高的分类精度代表实验中正确分类图像的比例较高。这表明该模型足以区分不同类型的蘑菇。换句话说，不建议使用 MAE 和 RMSE 进行性能评估，因为较低的 MAE 或 RMSE 并不一定意味着分类精度较高。因此，我们选择了不同的评价指标来评估食用蘑菇，不可食用蘑菇和有毒蘑菇的性能。本研究选取的评价指标如下：

准确率：实验中，分类准确率是正确分类蘑菇图像类型占整个实例集的百分比，如式(1)所示

$$\text{Accuracy} = \frac{\text{Total number of correct prediction}}{\text{Total number of predictions}} \tag{1}$$

Precision：该指标用于衡量预测的正确性，如式(2)所示

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

召回率：该指标通过测量检测到正确参考对象的概率来表示真阳性率。它可以通过使用等式(3)来计算。

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

F-score：为了比较不同分类器的分类精度，使用 F-score。它将分类器的精度和召回率通过调和平均值结合到单个度量中，如方程式所示 (4)。

$$\text{F-score} = \frac{2*(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

方程式中的哪里(2)，(3)和(4)*TP* 是预测为正类的真阳性类的数量，*TN* 预测为负类的真负类数量，*FP* 预测为假负类的数量正类，*FN* 分别预测正类为负类和漏掉的正类数量。

同样，Cohen's Kappa 是一种基于统计的方法，用于衡量两个评估者之间的一致性程度或将项目分类为相互排斥的类别的法官血淋淋的。通常认为 Cohen's Kappa 系数(k)是比简单的百分比一致性更稳健的衡量标准计算，因为它考虑了以下可能性该协议可能是一个巧合。

$$\text{Kappa}(\kappa) = \frac{P_o - P_e}{1 - P_e} \tag{5}$$

其中 P_o 代表评分者之间观察到的相对一致性，P_e 代表机会一致性的假设概率。

5.5 结果与讨论

本节的目的是详细描述各种深度学习模型的比较，以便确定最合适的模型。表 1 总结了分类准确度，精确度，召回率，F-score，Cohen's Kappa 统计量。从表 1 中可以看出，数据集上获得的最高分数为红色，第二高分数为蓝色。结果表明，并非所有众所周知的模型在应用于所选的不平衡数据集时都同样有效。因此，这些模型的使用将取决于它们所应用的具体问题。这些模型根据问题展示了它们的性能，我们可以相应地使用它们。相反，如果我们不使用模型来解决特定问题，那么它的效率有多高并不重要。例如，最初，本研究总共考虑了 13 个模型，即 CNN Decoder，ResNet50V2 [40]，InceptionV3 [39]，DenseNet201 [40]，DenseNet121 [40]， EfficientNetB0 [41]，Xception net [42]，EfficientNetB6 net [43]，MobileNetV2 net [44]，VGG16 net [45]，VGG19 [45]，transformer net 和 VIT [44]，其中 VGG19 [45]，transformer net 和 VIT [46]被认为是最先进的和高效的模型。由于这三个模型对我们的问题表现出不到 23%的性能，因此尽管它们先进且高效，但我们并未将它们纳入本研究中。结果表明，

这些模型在处理高度不平衡的数据集时表现不佳。

实验结果表明VGG16模型比其他模型表现出更好的性能。这种差异的说明可以在图 9 中找到，它是表 1 的图形表示。该模型在数据集不平衡的情况下提出了更好的特征提取器，表明该网络提取的特征比其他模型提取的特征更具建设性。VGG16 模型的准确度很高，这表明该网络提取的特征对于分类是有用且有效的。VGG16 [45]是一种对象检测和分类算法，在 ImageNet 数据库中超过一百万张图像上进行训练。这导致网络学习许多不同类型图像的丰富特征表示。

此外，对数据进行重新采样，以确认 VGG16 [45]模型在下采样、上采样和混合采样条件下的稳健性。在图 11 中，我们显示了每种情况下的重采样数据。表 1 说明了当还考虑重采样数据时 VGG16 [43]模型的性能。根据表 1 中的实验结果，VGG16 [45]分类模型被证明在选定蘑菇数据集的多个分类模型中表现令人满意。表 1 显示，VGG16 [45]下采样情况下的结果明显比上采样和混合样本的结果差。为了克服类别不平衡的问题，欠采样是一种流行的数据预处理技术，它可以平衡数据集，以提高分类率并避免对大多数类别的示例产生偏差。在下采样训练数据集中总是使用完整的少数数据，如果数据集中存在一些嘈杂的少数样本，这可能会降低分类器的性能。下采样减少了多数类别可用的信息量。当少数实例的数量较小时，分类系统的性能可能会受到下采样的限制。因此，当使用下采样时，性能可能并不显着。相反，过采样的性能优于下采样，但不如混合采样。过采样可能会导致过拟合，因为它复制了少数类的实例，从而导致其性能下降。表 1 表明，混合采样获得的结果通常优于下采样和混合采样获得的结果，并且与 VGG16 [45]模型获得的结果相当。

此外，从表 1 可以明显看出，混合采样获得的结果与 VGG16 [45]模型获得的结果相当。这是因为混合采样平衡了前面提到的两种采样类型之间的样本，因此与其他两种采样方法（即下采样和上采样）相比，它可以显着提高分类性能。但是，平衡数据可能效率不高，这可能会导致性能下降。

另一方面，EfficientNetB0 [41]方法被发现是第二好的模型，表现出比其他方法更高的性能。然而，对于在选定的蘑菇数据集上训练的其他模型，结果表明它们的表现稍差。鉴于此，很明显这些算法无法在高度不平衡的数据集上表现良好。同时，CNN Decoder 和 InceptionV3 [39]在表 1 中被证明是较差的特征提取器，导致其精度明显低于其他模型。CNN 解码的性能较差是因为传统的分类算法在数据分布不平衡的情况下无效，并且当数据分布严重不平衡时可能无法提取有用的特征。我们有一个不平衡的蘑菇数据集，因此在我们的例子中性能并不显着。另外，如图 8 所示，所选蘑菇图像数据集包含不同的背景，因此提取的特征值包括背景和蘑菇图像特征。如果通过适当的分割方法去除所用蘑菇图像的背景，则可以增强该网络的性能，因为提取的特征值将仅包括蘑菇图像。

图 8 为所选蘑菇分类数据集提供了每个类别的示例

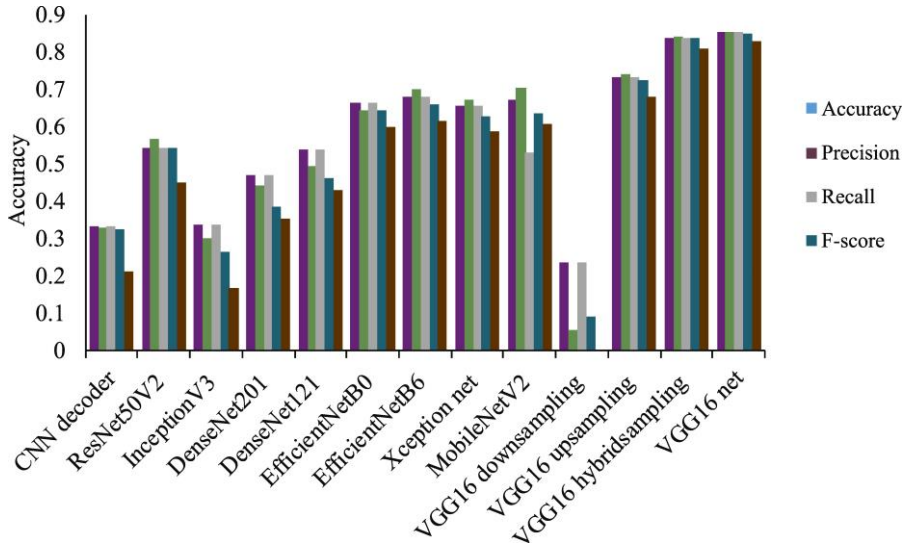


图 9 在不同时期数（范围从 0 到 1000）上进行的所有分类器的准确性。x 轴显示时期数，y 轴表示针对这些时期数实现的准确性

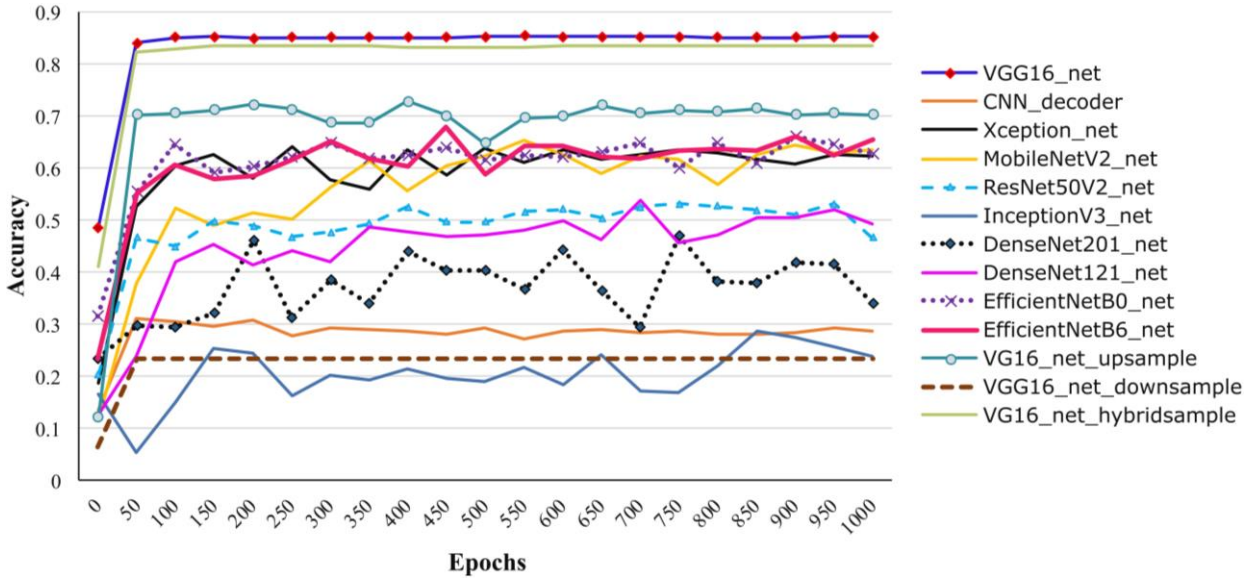
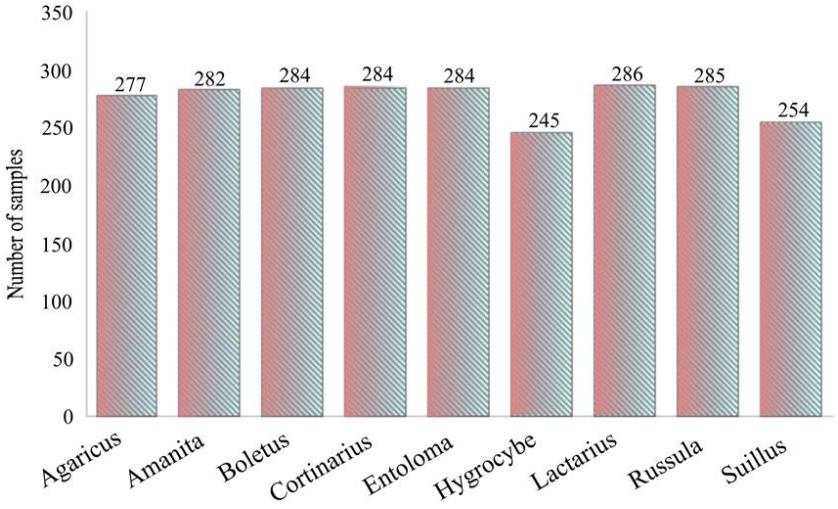


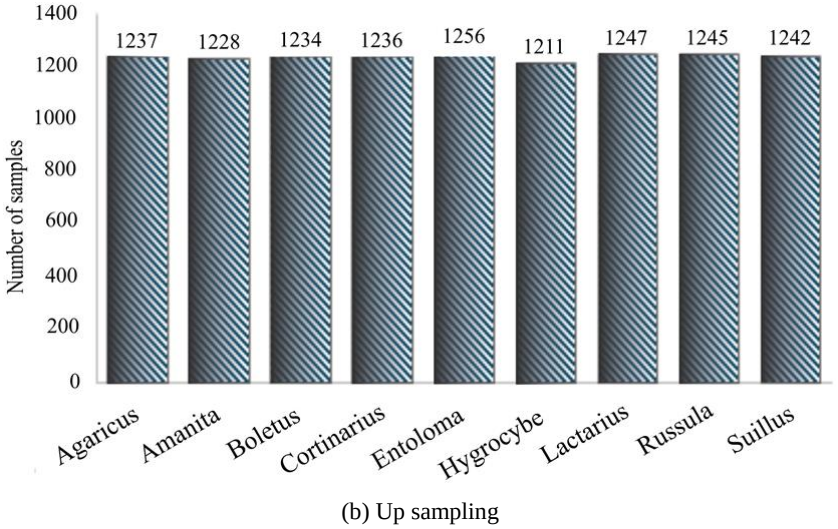
图 10 在 0 到 1000 范围内的不同历元数上进行的所有分类器的准确度比较。x 轴表示历元数，y 轴表示在这些历元上获得的准确度

表 1 使用各种众所周知的评估指标来评估基于深度学习的分类方法的分类准确性

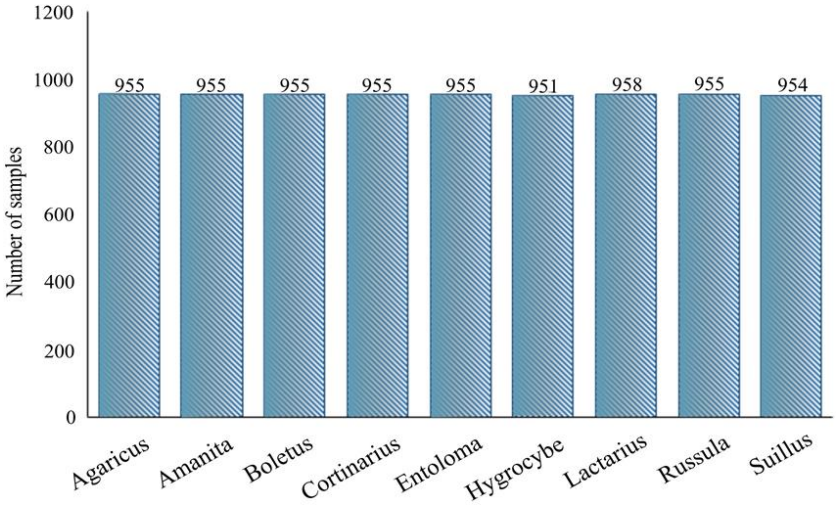
Methods	Accuracy	Precision	Recall	F-score	Kappa Statistic
CNN decoder	0.33239	0.32850	0.33231	0.32500	0.21139
ResNet50V2 [36]	0.54098	0.56488	0.54098	0.54368	0.44863
InceptionV3 [37]	0.33681	0.29995	0.33681	0.26342	0.16647
DenseNet201 [38]	0.47093	0.44068	0.47094	0.38611	0.35276
DenseNet121 [38]	0.53875	0.49414	0.53875	0.46322	0.43074
EfficientNetB0 [39]	0.66244	0.64334	0.66244	0.64111	0.59898
EfficientNetB6 [41]	0.67809	0.69727	0.67809	0.65776	0.61436
Xception net [40]	0.65350	0.67108	0.65350	0.62520	0.58484
MobileNetV2 [42]	0.66989	0.70196	0.52946	0.63343	0.60525
VGG16 net down-sampling	0.23547	0.05545	0.23547	0.08976	0.00000
VGG16 net up-sampling	0.73025	0.73804	0.73025	0.72497	0.68041
VGG16 net hybrid-sampling	0.83756	0.83888	0.83756	0.83571	0.80936
VGG16 net [43]	0.85395	0.85356	0.85395	0.85032	0.82802



(a) Down sampling



(b) Up sampling



(c) Hybrid sampling

图 11 不平衡蘑菇分类数据分别重新采样为下采样，上采样和混合采样

6 结论

现实生活中，食用菌和毒蘑菇的分配不均，从而出现阶层分配不平衡的问题。在本研究中，我们进行了实验分析，以确定哪种基于深度学习的模型是能够从数据集中提取信息特征的最佳特征提取器，这在数据集不平衡时可用于蘑菇分类。此外，进一步分析了最佳特征提取器模型在训练数据上采样，下采样和混合采样情况下的敏感性。实验结果表明，在重新采样数据的情况下，模型的稳健性会降低。结果还表明，不平衡数据的性能对模型选择更加敏感。此外，还进行了文献计量分析，展示了 1991 年至 2022 年期间全球蘑菇分类的科学研究趋势。该文献计量分析的结果表明蘑菇分类领域的出版物/合作不足。进一步探索这一领域将是有益的。在本研究中，文献计量分析包括蘑菇分类类别中已发表的所有研究方向。多种期刊发表有关该主题的文章表明该方向的出版物并不多，这表明该领域需要各种各样的研究主题。分析证实，分类精度对特征提取器更为敏感。未来一个可能的方向是设计更鲁棒的特征提取器，以提高不平衡数据分类的准确性。可以进一步观察到 VGG16 是相对更好的特征提取器。未来的另一个方向是设计使用 VGG16 作为特征提取器的高级分类器。

参考文献

[1] G. Douzas, F. Bacao, F. Last, Improving imbalanced learning through a heuristic oversampling method based on k-means and smote, Information Sciences 465(2018) 1-20.

[2] S. Ertekin, J. Huang, L. Bottou, L. Giles, Learning on the border: active learning in imbalanced data classification, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007, pp. 127-136.

[3] C. Krittanawong, A. Isath, J. Hahn, Z. Wang, S. E. Fogg, D. Bandyopadhyay, H. Jneid, S. S. Virani, W. W. Tang, Mushroom consumption and cardiovascular health: A systematic review, The American journal of medicine 134(5) (2021) 637–642.

[4] P. Heinemann, R. Hughes, C. Morrow, H. Sommer, R. Beelman, P. Wuest, et al., Grading of mushrooms using a machine vision system, Transactions of the ASAE 37(5) (1994) 1671–1677.

[5] M. A. Ottom, N. A. Alawad, K. Nahar, Classification of mushroom fungi using machine learning techniques, International Journal of Advanced Trends in Computer Science and Engineering 8(5) (2019) 2378–2385.

[6] P. Maurya, N. P. Singh, Mushroom classification using feature-based machine learning approach, in: Proceedings of 3rd International Conference on Computer Vision and Image Processing, Springer, 2020, pp. 197–206.

[7] J. Van De Vooren, G. Polder, G. Van der Heijden, Identification of mushroom cultivars using image analysis, Transactions of the ASAE 35(1) (1992) 347–350.

[8] M. N. Cáceres, M. A. González Arrieta, Automatic prediction of poisonous mushrooms by connectionist systems, in: Distributed Computing and Artificial Intelligence, Springer, 2013, pp. 341–349.

[9] H. Khan, X. Wang, H. Liu, Missing value imputation through shorter interval selection driven by fuzzy c-means clustering, Computers & Electrical Engineering 93(2021) 107230.

[10] H. Khan, X. Wang, H. Liu, Handling missing data through deep convolutional neural network, Information Sciences 595 (2022) 278–293.

[11] H. Khan, X. Wang, H. Liu, A study on relationship between prediction uncertainty and robustness to noisy data, International Journal of Systems Science 54(6) (2023) 1243–1258.

[12] M. T. Rasheed, D. Shi, LSR: Lightening super-resolution deep network for low-light image enhancement, Neurocomputing 505(2022) 263– 275.

[13] M. T. Rasheed, G. Guo, D. Shi, H. Khan, X. Cheng, An empirical study on retinex methods for low-light image enhancement, Remote Sensing 14(18) (2022) 4608.

[14] M. T. Rasheed, D. Shi, H. Khan, A comprehensive experiment-based review of low-light image enhancement methods and benchmarking low-light image quality assessment, Signal Processing (2022) 108821.

[15] H. Khan, H. Liu, C. Liu, Missing label imputation through inception-based semi-supervised ensemble learning, Advances in Computational Intelligence 2(1) (2022) 1–11.

[16] C. Salvador, M. R. Martins, H. Vicente, J. Neves, J. M. Arteiro, A. T. Caldeira, Modelling molecular and inorganic data of amanita ponderosa mushrooms using artificial neural networks, Agroforestry Systems 87(2) (2013) 295–302.

[17] A. T. Caldeira, J. M. Arteiro, J. C. Roseiro, J. Neves, H. Vicente, An artificial intelligence approach to bacillus amyloliquefaciens ccmi 1051 cultures: application to the production of anti-fungal compounds, Bioresource Technology 102(2) (2011) 1496–1502.

[18] A. Wibowo, Y. Rahayu, A. Riyanto, T. Hidayatulloh, Classification algorithm for edible mushroom identification, in: 2018 International Conference on Information and Communications Technology (ICOIACT), IEEE, 2018, pp. 250–253.

[19] J. H. J. C. Ortega, A. C. Lagman, L. R. Q. Natividad, E. T. Bantug, M. R. Resureccion, L. Manalo, Analysis of performance of classification algorithms in mushroom poisonous detection using confusion matrix analysis, International Journal 9(1.3) (2020).

[20] B. Sunita, D. Bishan, et al., Mushroom classification using data mining techniques., International Journal of Pharma and Bio Sciences 6(1) (2015).

[21] M. Husaini, A data mining based on ensemble classifier classification approach for edible mushroom identification, Int Res J Eng Technol (IRJET) 5 (2018).

[22] Y. Wang, J. Du, H. Zhang, X. Yang, Mushroom toxicity recognition based on multigrained cascade forest, Scientific Programming 2020 (2020).

[23] A. Subramaniam, B.-J. Oh, Mushroom recognition using pca algorithm, International Journal of Software Engineering and Its Applications 10(1) (2016) 43–50.

[24] S. Ismail, A. R. Zainal, A. Mustapha, Behavioural features for mushroom classification, in: 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE, 2018, pp. 412–415.

[25] S. Verma, M. Dutta, Mushroom classification using ann and anfis algorithm, IOSR Journal of Engineering (IOSRJEN) 8(01) (2018) 94– 100.

[26] A. Sameh, M. Moghayer, G. Mohanad, A. Mohammad, Classification of mushroom using artificial neural network., International Journal of Academic and Applied Research (IJAAR) 3(2) (2020).

[27] E. CENGİL, A. ÇINAR, Poisonous mushroom detection using yolov5, Turkish Journal of Science and Technology 16(1) (2021) 119–127.

[28] J. U. Lidasan, M. P. Tagacay, Mushroom recognition using neural network, International Journal of Computer Science Issues (IJCSI) 15(5) (2018) 52–57.

[29] E. S. Alkronz, K. A. Moghayer, M. Meimeh, M. Gazzaz, B. S. AbuNasser, S. S. Abu-Naser, Prediction of whether mushroom is edible or poisonous using back-propagation neural network., International Journal of Academic and Applied Research (IJAAR) 3(2) (2019).

[30] I. S. Al-Mejibli, D. H. Abd, Mushroom diagnosis assistance system based on machine learning by using mobile devices, Journal of AlQadisiyah for computer science and mathematics (2) (2017) Page– 103.

[31] L. Ma, R. Gao, H. Han, C. Chen, Z. Yan, J. Zhao, X. Lv, C. Chen, L. Xie, Efficient identification of bachu mushroom by flourier transform infrared (ft-ir) spectroscopy coupled with pls-gs-svm, Optik 224 (2020) 165712.

[32] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W. M. Lim, How to conduct a bibliometric analysis: An overview and guidelines, Journal of Business Research 133 (2021) 285–296.

[33] L. Leydesdorff, World shares of publications of the usa, eu-27, and china compared and predicted using the new web of science interface versus scopus, Profesional de la Informaci ón 21(1) (2012) 43–49.

[34] J. M. Merig ó, J.-B. Yang, A bibliometric analysis of operations research and management science, Omega 73 (2017) 37–48.

[35] N. Van Eck, L. Waltman, Software survey: Vosviewer, a computer program for bibliometric mapping, scientometrics 84(2) (2010) 523– 538.

[36] F. M. Cole, Amanita phalloides in victoria, Medical journal of Australia 158(12) (1993) 849–850.

[37] CatoDogo, Mushrooms classification common genus's images, https://www.kaggle.com/datasets/maysee/mushrooms-classificationcommon-genuss-images (2018).

[38] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European conference on computer vision, Springer, 2016, pp. 630–645.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[40] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700– 4708.

[41] L. Tao, V. Asari, An integrated neighborhood dependent approach for nonlinear enhancement of color images, in: International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004., Vol. 2, IEEE, 2004, pp. 138–139.

[42] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[43] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556 (2014).

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16×16 words: Transformers for image recognition at scale, arXiv preprint arXiv: 2010.11929 (2020).