用于光谱数据分类的一维卷积神经网络和机器学习算法：新冠肺炎案例研究

arXiv:2301.10746v1 [cs.NE] 24 Jan 2023

Breno Aguiar Krohling
自然启发计算实验室
LABCIN- UFES 圣埃斯皮里图联邦大学 巴西维多利亚
brenokrohling@gmail.com

RenatoA. Krohling自然启发计算实验室 LABCIN - UFES
PPGI -UFES 圣埃斯皮里图联邦大学 巴西维多利亚
krohling.renato@gmail.com

摘要

机器学习和深度学习算法日益广泛地应用于解决各个知识领域中的问题。其中，化学计量学受益于这些算法在光谱数据分析中的应用。通常，支持向量机和偏最小二乘等算法会应用于光谱数据集，以执行分类和回归任务。本文介绍了一维卷积神经网络（1D-CNN），用于评估其在光谱学所获取光谱数据上的有效性。在大多数情况下，光谱信号存在噪声，且不同类别之间存在重叠。首先，我们开展了广泛的实验，将一维卷积神经网络与机器学习算法以及化学计量学中常用的标准算法进行比较，针对文献中最知名的可用数据集开展光谱数据分类。接下来，我们将近期通过光谱学采集的SARS-CoV-2病毒（该病毒会引发新冠肺炎）光谱样本用作案例研究。实验结果表明，一维卷积神经网络的表现优于机器学习算法和标准算法，平均准确率达到96.5%，特异度为98%，敏感度为94%。这些令人期待的结果表明，未来可将一维卷积神经网络用于自动化系统，以诊断新冠肺炎及其他病毒性疾病。

关键词 卷积　　卷积神经网络· 机器学习· 光谱学· 新冠肺炎　　.

1 引言

随着光谱测量技术的发展以及化学计量学领域数据分析研究的推进，光谱学的应用日益受到研究关注，并在实践中得到广泛应用。尤其是红外（IR）和近红外（NIR）光谱测量的应用，它们能够提供无法通过简单观察获得的信息。红外光谱被认为是化学计量学中最重要的工具之一（Stuart，2000），因为它成本低、速度快，且能够分析处于各种状态的样品，因此也是一种非侵入性方法。此外，机器学习和深度学习（如人工神经网络）已被引入以往由统计分析方法和／或经典人工智能算法主导的领域。随着神经网络应用在过去几年中的不断增长，这类算法开始应用于各种不同性质的问题，包括光谱学问题。

光谱学数据存在一些问题，例如维度高、数据集规模小。因此，通常需要采用噪声滤波和降维等预处理技术，才能进行适当的分类。然而，人工神经网络（NN）已知能够很好地处理高维性，并且在不同应用中取得了显著成果。不过，NN 需要大量数据才能正常运行。在这种情况下，有必要分析采用标准方法与转向新方法之间的权衡。

卷积神经网络可用于解决具有挑战性的现实问题，例如诊断 SARS-CoV-2病毒。随着 COVID-19大流行的发展，出于卫生和公共卫生方面的考虑，有必要大规模开展人群测试。SARS-CoV-2病毒测试面临两大主要障碍：1）测试费用；2）获取结果所需的时间。前者主要影响欠发达国家，后者则影响世界各地的防控计划。通过利用光谱仪采集信息进行测试，并借助稳健的人工智能模型处理数据，可以快速、低成本地获取患者的诊断信息，从而使公共和私营医疗服务机构都能开展此类检测。

随着机器学习技术的发展，光谱学中常用的标准算法（如支持向量机和偏最小二乘）正逐渐被近年来的方法所取代，例如卷积神经网络（Zhang等人，2019年）以及（Yuanyuan和Zhibin，2018年）。然而，卷积神经网络通常需要大量数据来训练模型。尽管已有公共数据库（Kaewseekhao等人，2020年）、（Chauvergne等人，2020年）和（Zyubin等人，2020年），但公共领域中仍然缺乏光谱数据集。因此，有必要研究此类算法在训练样本数量较少的问题上的表现。此外，光谱数据具有某些特征，例如样品在特定波长处相互重叠，这使分类任务变得困难；由Yin等人 （2021）提供的 COVID-19 数据集便是一个例子。

本研究的主要贡献有两点：

- 我们将一维卷积神经网络应用于 7 个可用的公开光谱数据集，以验证结果并与传统机器学习算法进行比较。

- 我们展示了一项涉及新冠肺炎光谱数据的案例研究，并说明了一维卷积神经网络的可行性。

本文其余部分安排如下：第2节介绍相关研究；第3节介绍方法，包括预处理技术、机器学习算法以及一维卷积神经网络；第4节展示结果并进行比较分析和讨论；第5节总结全文，并指出未来研究方向。

2 相关研究

光谱技术应用于多个知识领域，可用于解决多种现实问题，例如皮肤病变体内 分类（McIntosh等人，2001年），以及农产品品质判定（Hayati等人，2020年）。McIntosh等人（2001年）采用统计方法研究皮肤病变的表征，具体使用配对t检验。研究采用线性判别分析（LDA）对癌性病变进行分类，结果表明，该方法准确率较高，可行性良好，能够以非侵入性方式区分皮肤病变。

Gniadecka等人 (2004) 将光谱技术与多层感知机神经网络相结合，用于皮肤癌的早期诊断。该研究旨在将黑色素瘤（一种致死率较高的皮肤癌）与其他皮肤病变区分开来，例如色素痣和基底细胞癌。研究考察了癌组织中化学变化的影响及其对分类的重要性。结果显示，敏感度和特异度较高，这两项指标对于黑色素瘤检测至关重要。 Xiang等人 (2010) 使用红外光谱数据诊断子宫内膜癌。该研究包含77个样本，包括癌变组织和对照组。研究采用了降噪技术，如Savitsky-Golay滤波器，并使用主成分分析进行数据降维。处理后的数据用于训练人工神经网络，所得结果表明，该方法对于疾病的早期诊断有效。

Wu等人 (2012) 使用光谱数据，无创地测定云芝（一种药用真菌）粉末样品中的蛋白质和多糖含量。作者比较了应用于原始数据的人工神经网络，以及应用主成分分析降维后数据的人工神经网络。该神经网络通过反向传播算法进行训练，使用降维后的数据取得了更好的结果。 Liu等人(2017) 使用深度自编码器（DAE）从红外光谱数据中提取并降低特征维度。为评估其有效性，研究将该算法应用于Cigarettes数据集，并与文献中已有的方法进行比较，例如主成分分析；随后又应用了K近邻算法。结果表明，与同期方法相比，使用DAE的模型能够提取出区分能力更强的特征。

Peng 等人（2018年）提出了一种用于提取频域光谱信息的算法。通过各主成分的熵贡献结合遗传算法确定主成分。随后使用偏最小二乘法（PLS）将筛选出的主成分用于构建回归模型，与使用原始数据得到的结果相比，模型性能有所提升。

Cui 和 Fearn（2018年）对三个光谱数据集进行了测试，以验证卷积神经网络相较于偏最小二乘回归方法的有效性。结果表明，卷积神经网络的卷积层可充当光谱的预处理器，无需应用信号处理技术。此外，结果还表明，即使使用样本数量较少的数据集，卷积神经网络在用于验证模型的指标方面仍具有优势。

Yuanyuan和Zhibin (2018年) 将卷积神经网络集成应用于玉米、汽油和混合气体数据集，并将该方法与光谱学领域的经典方法（如偏最小二乘）以及使用反向传播训练的神经网络进行了比较。结果表明，所提出的模型在所有研究案例中均取得了更优的性能，显示出其在光谱分类方面的价值。

Lima 等 (2019年) 研究了非黑色素瘤皮肤癌（基底细胞癌和鳞状细胞癌）与光化性角化病及正常皮肤之间的鉴别。该研究使用了从活体组织和离体组织中采集的光谱数据。研究采用 t 检验找出光谱中样本间存在显著差异的区域，并利用选定区域通过欧氏距离和马氏距离鉴别病变。结果表明，使用离体组织数据时具有良好的稳健性。

Seoni等人(2019年)利用正常皮肤组织和患有光化性角化病的皮肤组织的红外光谱数据，研究了血红蛋白浓度。光化性角化病是一种良性皮肤病变，若不治疗，可能发展为癌性病变。研究采用多变量方差分析进行统计分析，以评估健康皮肤组与含有光化性角化病的皮肤组之间是否存在差异。结果表明，总体而言，通过统计方法可以区分这两组。

Zhang等人(2019年)提出了一种新方法，该方法基于具有一维卷积的卷积神经网络，并在两个卷积层中采用Inception结构。该方法在四个公开光谱数据集上进行了评估：玉米、Tablet、小麦和土壤。所得结果优于其他卷积神经网络架构和偏线性方法，例如偏最小二乘和支持向量回归（SVR）。此外，还对原始光谱信号应用了预处理方法，以研究其对预测结果的影响。作者得出结论，信号降噪是提升算法性能的必要步骤。

Chen 和 Wang（2019年）评估了在不对光谱进行降维的情况下，从卷积神经网络中提取特征的能力。研究使用玉米数据集开展实验，并将结果与传统化学计量学方法（如偏最小二乘）以及神经网络训练算法（如极限学习机（ELM）和反向传播）进行比较。实验结果表明：1）如果方法/算法以原始光谱作为输入，卷积神经网络的性能更好；2）如果将标准方法/算法与降维算法和变量选择相结合，所得结果与卷积神经网络相似，方法之间不存在统计学差异。

Chen 等人(2020年) 提出使用通过极限学习机算法训练的人工神经网络集成。所提出的算法通过随机初始化权重，使

隐藏层在模型之间形成差异。该方法在 3 个数据集 Tecator、Shootcut_2002和 Tablet 上进行了测试。该方法与偏最小二乘进行了比较，其表现优于传统化学计量学算法。Yan 等人 (2020年) 收集了中药原料水解过程中的红外光谱数据。研究使用卷积神经网络和偏最小二乘结合实时数据建立了回归模型。所提出的卷积神经网络架构表现优于偏最小二乘，但处理成本和计算时间更高。因此，对于实时应用，需要分析结果之间的最大时间间隔延迟，以决定应采用哪种方法。

3 方法

3.1 数据预处理

光谱学是一种基于分子中原子振动的技术。光谱可按波长进行分析和分类，例如紫外线、可见光和红外线。每种波长都能提供用于分析样品的独特信息，因此在采集数据时，有必要了解最适用的波长范围。例如，红外波长下获得的光谱广泛用于需要识别样品有机成分的应用中（Stuart，2000）。这种技术生成的数据称为光谱，它是通过向样品发射一定量的辐射，并确定样品在给定能级下吸收或反射了辐射的哪一部分而获得的（Stuart，2000）。

通过光谱仪采集的数据可能受到不同类型的干扰，从低质量设备造成的干扰，到不受控环境中的外部干扰，例如光散射和数据噪声等。光谱信号中的噪声成分是光谱分析中的一项难题，因为采集信号的不稳定可能导致错误结论。对样品进行多次读取并取光谱平均值可以降低这种不稳定性，但无法消除所有现存噪声。因此，信号处理技术成为光谱分析中的关键步骤，用于去除或减少光谱信号中不需要的噪声成分。

选择合适的技术进行数据预处理，会直接影响模型性能和数据可解释性（Lee 等人，2017）。最常用的预处理技术可分为两类：1）散射校正；2）光谱导数（Rinnan 等人，2009）。散射校正算法用于降低样本之间的差异。光谱导数算法使用卷积运算，通常通过对光谱中的点拟合多项式来实现。

在这项工作中，我们使用光谱分析中最常见的预处理滤波器之一——Savitzky-Golay（SG）滤波器（Savitzky 和 Golay，1964），并评估其对分类结果的影响。Savitzky 和 Golay（1964）提出了一种基于多项式插值平滑数据的算法。他们证明，对光谱中的一组点拟合多项式，然后按照点逼近间隔在后续点处求值，其效果等同于更复杂的卷积运算。此外，他们还表明，该算法能够去除数据中的噪声，同时保留

光谱波长中的形状和峰值，从而在平滑后的数据中保留原有特征。

3.2 机器学习

计算领域发展最快的领域之一是人工智能。人工智能的应用遍及医学、工程、金融等领域（Pannu，2015）。人工智能的一个分支是机器学习，它由能够从已知数据中获取知识的算法组成，并能在此基础上推断出针对新问题实例的解决方案，而无需为此进行明确编程（Alpaydin，2020年）。在现有问题中，监督学习问题可谓最为常见。算法会寻找模式，以学习如何将输入 $\mathbf{x}$ 与其对应的输出 y相关联。通常，监督学习算法通过示例学习，如何根据输入 y $\mathbf{x}$ **推断输出**，并估计 $p(y|\mathbf{x})$ (Goodfellow等人，2016）。本研究使用的机器学习算法包括偏最小二乘判别分析、支持向量机、k近邻（Alpaydin，2020年），以及基于爬山法和最小生成树的多目标聚类（Esgario和Krohling，2018年）。

3.3 一维卷积神经网络

人工神经网络（人工神经网络）由称为人工神经元的简单计算单元构成，这些单元彼此相连（Schalkoff，1997）。神经元之间的连接形成复杂的层，用于计算和传播信息。训练人工神经网络的算法称为反向传播。借助该算法，可以训练人工神经网络执行分类和回归／预测任务。标准人工神经网络由三个主要层组成：数据输入层、隐藏层和输出层。

卷积神经网络（卷积神经网络）最初由LeCun等人（1989）提出，是一种专为处理网格格式数据（如图像和时间序列）而设计的网络，因为它们能够捕捉空间和时间特征（Goodfellow等人，2016）。2012年，一种基于卷积神经网络的模型赢得了ImageNet数据库上的LSVRC-12竞赛，使这种架构声名鹊起（Krizhevsky等人，2012）。卷积神经网络与传统人工神经网络的主要区别在于，其架构中包含卷积层和池化层。卷积神经网络至少在一个层中使用卷积运算，而非矩阵运算（Goodfellow等人，2016）。下文将介绍卷积神经网络的主要组成部分及其在训练过程中的工作原理。

卷积层： 卷积是对两个函数执行的线性运算，并生成第三个函数。在卷积神经网络中，卷积用于提取特征，其中卷积的第一个函数由网络的输入数据表示，第二个函数则由一个卷积核表示，卷积核是一种用于提取特征的滑动滤波器。通过对这两个函数执行卷积运算，可以生成特征图，其中包含原始输入数据中的相关信息。根据问题领域和数据特征，卷积神经网络可以包含一维、二维或三维卷积层。例如，图像识别应用使用二维卷积，而数据来自时间序列的应用则使用一维卷积。

执行卷积运算后，将特征图应用于激活函数。激活函数旨在为模型引入非线性（Goodfellow等人，2016）。选择激活函数时，必须考虑一些特性，例如：由于该运算会频繁执行，因此计算成本必须较低。由于神经网络的训练过程基于梯度下降，激活函数还必须可微。线性整流单元（ReLU）激活函数在神经网络中被广泛使用，因为它具备这些特性。对于所有负值，该函数返回零；对于正值，则返回其自身。其定义为 $ReLU(x) = max(0, x)$。

池化层：池化层通常应用于卷积运算之后，其作用是降低特征图的维度，以消除冗余信息并保留主要特征（Goodfellow等人，2016）。与卷积类似，滑动窗口会在特征图上移动并生成新的输出。然而，池化运算是确定性的，通常会计算滑动窗口内元素的最大值或平均值（Zhang等人，2020年）。应用最大池化技术时，分析数据组中的最大值会传递到后续层；而平均池化技术则会传递数据的平均值。

全连接层：仅组合卷积层和池化层不足以完成分类/回归任务；它们在架构中的主要作用是从给定数据中提取特征。因此，提取出的特征需要通过分类器来完成任务，这种分类器称为全连接层。通常，卷积神经网络中的全连接层由一种称为多层感知机（MLP）的人工神经网络表示（Zhang等人，2020年）。多层感知机是一种前馈神经网络，由输入层、若干隐藏层和输出层组成。

多层感知机将 f 映射为：$x \to y$ 函数，其中 x 是通过卷积层和池化层提取的特征，而 y 是期望的网络输出。这一过程称为人工神经网络训练，是一个旨在最小化损失函数的优化问题（Goodfellow等人，2016）。损失函数用于计算期望输出与神经网络实际输出之间的误差。所得误差用于重新计算网络的权重（Zhang等人，2020年）。卷积神经网络在分类问题中常用的交叉熵损失函数如下：

$$CrossEntropy(pa, pb) = -\sum_{i}^{k} pa_i \log pb_i \tag{1}$$

其中，pa_i 和 pb_i 分别表示类别 i的真实概率和预测概率，k 表示类别数。对于不平衡数据集，通常使用在计算中考虑各类别频率的损失函数，例如加权交叉熵。

除了参数调整算法之外，还可以采用权重正则化技术来避免过拟合等问题。过拟合是指模型在训练过程中表现良好，却无法有效处理未知数据。本研究采用了随机失活正则化技术。该技术以一定的概率移除部分神经元连接，以提高网络对未知数据的泛化能力（Zhang等人，2020年）。

将多个卷积层和池化层与全连接层相结合，即可构建类CNN架构。图1 展示了本研究中用于光谱数据分类的一维卷积神经网络架构。

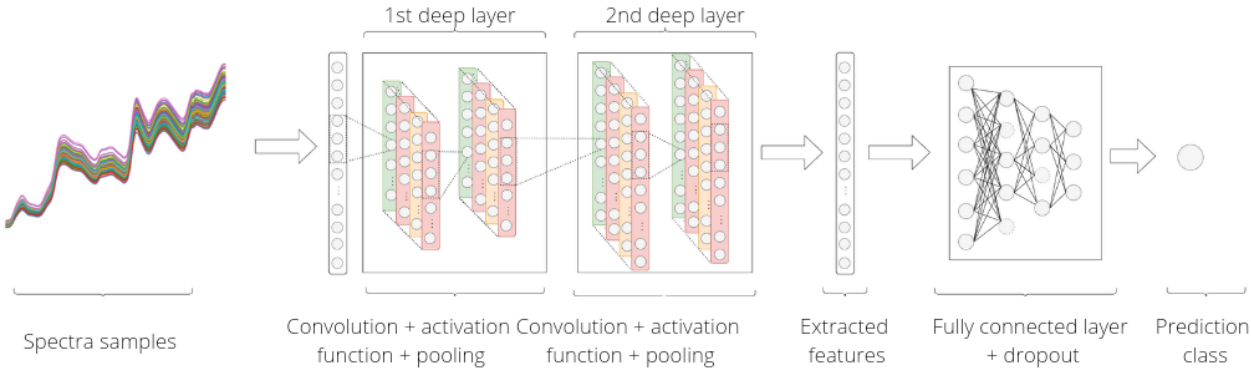


图 1： 一维卷积神经网络 由卷积层、池化层和全连接层组成的卷积神经网络架构 .

4 实验与结果

本节介绍结果并进行讨论。首先介绍实验中使用的数据集。其次，将算法应用于这些数据集，以便与机器学习算法和化学计量学标准算法进行比较。最后，介绍一个使用 SARS-CoV-2 病毒样品光谱数据的案例研究。

4.1 数据集

接下来，我们简要介绍这些数据集。Kosmowski和Worku (2018年) 收集了大麦、鹰嘴豆和高粱作物的红外光谱数据。每种谷物各构成一个数据集。大麦数据集由 1200 个大麦样本组成，分属 24 个谷物品种类别。鹰嘴豆数据集包含 950 个鹰嘴豆样本，分属 19 个类别；最后，高粱数据集包含 500 个高粱样本，分属 10 个类别。尽管数据集中的样本数量很多，但显然，这些样本分布在大量类别中。因此，每个类别中的样本频数较低。这三个数据集都具有良好的平衡性。Zheng等人 (2019年) 提出了一种使用极限学习机进行光谱分类的方法，并公开了相关数据。本研究使用了四个数据集：咖啡、肉类、油和水果。咖啡数据集中的样本划分为

数据集	样本	类别数
大麦	1200	24
鹰嘴豆	950	19
高粱	500	10
咖啡	56	2
Meat	120	3
Oil	120	4
水果	983	2

表 1： 每个数据集中包含的样本数量和类别数。

阿拉比卡和科尼隆咖啡豆，均匀分布在 56 个样本中。肉类数据集由 60 个鸡肉、猪肉和火鸡肉样本组成，分布同样均匀。油类数据集包含来自 4 个不同国家（希腊、意大利、葡萄牙和西班牙）的 60 个特级初榨橄榄油光谱样本。水果数据集由 983 个水果泥样本组成，分为草莓和非草莓两个类别。表 1 列出了各数据集中样本和类别的数量。

4.2 实验设置

本文研究的一维卷积神经网络和机器学习算法的训练均采用交叉验证进行。该方法用于减少偏差并避免模型过度估计。学习过程包括将数据集划分为 k 折，并将每一折分别用作模型的测试集。最后，计算该算法在数据集上的平均准确率和标准差。使用交叉验证 训练和评估算法的过程如算法 1所述。所有实验均采用 $k = 5$ 折，换言之，将数据集划分为 80% 用于训练和验证，另有 20% 专门用于测试。

算法1： 交叉验证评估

1 - 打乱样本。
2 - 将其分为 k 折。
对于折中的 f **执行**
    1 - 将 f 保留为测试集。
    2 - 使用剩余的折训练一个新模型
    3 - 使用折 f评估模型。
    4 - 保存模型准确率。
end
3 - 获取交叉验证平均准确率和标准差。
4 - 返回准确率最接近交叉验证平均准确率的模型。

PLS-DA、K近邻、支持向量机和卷积神经网络算法需要进行超参数调整。因此，可以为每个数据集选择最佳配置，从而最大限度地提升其性能。
偏最小二乘判别分析算法基于成分构建，旨在最大化变量的方差。每个成分都说明其所代表的方差程度。对于

每个数据集， c 选择了若干成分，这些成分共同代表数据总方差的 95%。

对于 K近邻和支持向量机算法，采用了该算法的一种变体 1 ，以便为每一轮交叉验证找到最佳超参数组合。用于训练模型的数据会重新划分为训练集和验证集，以找到最佳超参数组合。因此，对于每种训练折和测试折的组合，都可以为模型找到一组最优超参数。该过程结束时，计算算法得到的均值和标准差，并选取生成模型准确率最接近平均准确率的那组超参数，从而避免模型过拟合。寻找最佳超参数和平均准确率的过程见算法 2。

表 2 展示了 K近邻和支持向量机算法所使用的搜索空间。表 3 展示了偏最小二乘判别分析算法中，为解释 c 所需的成分数95% 的数据方差，以及 k个近邻的数量和核函数；它们在执行算法 2时所得结果最接近平均结果。

算法	超参数	搜索空间
KNN	k近邻	[2, 3, 5, 10, 15, 20, 24]
SVM	核函数	[线性、多项式、Sigmoid、径向基函数]

表 2： K近邻和支持向量机超参数的搜索空间。

Alg算法 2： 交叉验证评估——通过变化寻找最佳的 h 集合 yperp参数

1 - 打乱样本。 2 - 将其拆分为 k 折。**对于***f* 属于折 **执行**1 - 将 f 保留为测试集。**对于**每个集合 λ 的超参数 **执行** *v* 属于（折 - *f*）**执行**1 - 将 v 保留为验证集。 2 - 使用剩余的折训练一个新模型。 3 - 使用第 v折评估模型。 4 - 保存模型准确率。**结束**1 - 保存 λ 集的平均准确率和标准差。**结束**2 - 使用最佳 λ 集和剩余的折训练一个新模型。 3 - 使用第 f折评估模型。 4 - 保存准确率和模型

end

3 - 获取交叉验证平均准确率和标准差。4 - 返回模型和 λ 集，其准确率最接近交叉验证平均准确率。

数据集	组件数量	N. k-邻居	核函数
大麦	2	15	Sigmoid
鹰嘴豆	2	10	Sigmoid
高粱	2	10	Sigmoid
咖啡	2	3	RBF
Meat	3	3	RBF
Oil	6	5	RBF
水果	4	3	RBF

表3： 偏最小二乘判别分析、K近邻和支持向量机取得最佳结果时的超参数。

最后，作为化学计量学领域的一种新近方法，一维卷积神经网络并没有针对涉及光谱数据分类的应用预先定义并预训练的架构。为了选择本研究中使用的架构和超参数，使用了软件 Optuna[1]。Optuna 是一种为机器学习问题开发的超参数优化器，用于搜索能够最大限度提升模型性能的超参数和架构组合。与算法2类似，它会在训练和验证数据集中搜索最佳模型；找到一组次优超参数后，再用测试集测试该模型。

表4 列出了各数据集在训练和测试阶段所选用的超参数。所有实验和算法均使用 scikit-learn[2]和 PyTorch[3] 库以 Python[4]编写。实验所用源代码可向作者索取。

	数据集						
超参数	大麦	鹰嘴豆	高粱	咖啡	Meat	Oil	Fuit
优化器	Adam	Adam	Adam	Adam	Adam	Adam	Adam
学习率	0.001	0.001	0.001	0.0001	0.001	0.001	0.001
训练轮数	5000	5000	5000	100	700	1000	1000
丢弃率	0.4	0.4	0.3	0.1	0.1	0.2	0.1
批次大小	50	50	50	6	10	10	50
Loss 函数	交叉熵	交叉熵	交叉熵	交叉熵	交叉熵	交叉熵	交叉熵

表4： 为一维卷积神经网络取得最佳结果的超参数。

[1]https://optuna.readthedocs.io/[2]
https://scikit-learn.org/[3]
https://pytorch.org/[4]https: //
www.python.org/

4.3 实验结果

本小节介绍各算法应用于本文数据集时所获得的结果。实验分为两部分：1）使用原始光谱数据；2）将 Savitzky–Golay（SG）滤波器作为预处理步骤。将实验分为两个阶段，旨在分析数据预处理对模型结果的影响。所有实验均在一台配备 Intel Core I5-8265U 处理器和 8GB 随机存取存储器的 Windows 10 计算机上进行。

原始数据： 第一种实验配置（未经预处理的数据）所得结果见表 5。对每个数据集均应用了所研究的算法（MOCHM、偏最小二乘判别分析、支持向量机、K近邻和一维卷积神经网络）。最佳结果以粗体显示。

数据集	MOCHM	偏最小二乘判别分析	SVM	KNN	一维卷积神经网络
大麦	-	10.66 ± 0.62	7.75 ± 0.54	16.08 ± 1.38	$\mathbf{47.4 \pm 4.7}$
鹰嘴豆	-	14.73 ± 1.12	23.58 ± 1.07	27.16 ± 4.48	$\mathbf{54.2 \pm 3.3}$
高粱	-	18.79 ± 3.31	15.79 ± 1.94	34.79 ± 2.99	$\mathbf{55.8 \pm 3.3}$
咖啡	51 ± 2.1	85.7 ± 3.4	42.7 ± 3.4	46.3 ± 1.37	$\mathbf{100 \pm 0}$
Meat	36 ± 0.2	70.83 ± 6.56	41.6 ± 3.1	87.5 ± 7	$\mathbf{95 \pm 3.3}$
Oil	44 ± 0.7	75 ± 2.04	41.6 ± 1.02	73.3 ± 3.11	$\mathbf{88.8 \pm 5.3}$
水果	95 ± 1.9	91.04 ± 1.02	88.3 ± 2.96	88.9 ± 1.75	$\mathbf{96 \pm 0.6}$

表 5： 未进行预处理的数据应用各算法后得到的平均准确率。

对于所有数据集，使用未经预处理的数据时，一维卷积神经网络取得的结果都更优。不过可以看到，大麦、鹰嘴豆和高粱数据集的平均准确率低于其他数据集。这是因为数据中类别数量较多，而训练阶段每个类别可用的样本数量较少，导致模型无法有效泛化。尽管如此，将一维卷积神经网络的结果与化学计量学中使用的标准算法（即偏最小二乘判别分析和支持向量机）进行比较，仍可发现差异。K近邻算法的结果也优于偏最小二乘判别分析和支持向量机算法：在高粱数据集上，其平均准确率达到 20%；相比之下，支持向量机在该数据集上的表现最差。在其他数据集中，偏最小二乘判别分析算法总体上取得了仅次于一维卷积神经网络的最佳结果。MOCHM 算法在大麦、高粱和鹰嘴豆数据集上的运行时间超过 60 分钟，因此在执行完成前便停止了。

预处理数据：第二组实验采用了相同的算法和数据集，但在预处理步骤中应用了 SG 滤波器。SG 滤波器采用标准超参数配置：窗口大小设为 11，使用 3 阶多项式和二阶导数。在表格 6 中展示了实验所得结果。

可以观察到，一维卷积神经网络并未像前一个案例中所报告的那样取得结果上的统治地位。对于大麦、鹰嘴豆 和高粱 数据集，K近邻优于

数据集	MOCHM	偏最小二乘判别分析	SVM	KNN	一维卷积神经网络
大麦	-	11.83 ± 2.77	14.25 ± 2.3	**64.91 ± 1.3**	46.16 ± 0.97
鹰嘴豆	-	15.78 ± 1.07	20.4 ± 4.28	**71.36 ± 2.11**	55.2 ± 1.1
高粱	-	25.99 ± 2.63	28.38 ± 5.87	**71.39 ± 3.3**	59.6 ± 2.1
咖啡	83 ± 1.07	87.4 ± 2.1	92.69 ± 6.1	80.11 ± 4.9	**100 ± 0**
Meat	81 ± 0.8	71.66 ± 4.24	58.33 ± 3.1	92.49 ± 4.08	**96.6 ± 2.07**
Oil	**91 ± 1.1**	75.8 ± 3.11	41.6 ± 1.02	79.1 ± 1.16	89.4 ± 1.12
水果	92 ± 2.2	79.2 ± 1.02	93.49 ± 0.86	91.86 ± 0.38	**94.2 ± 0.3**

表 6： 对经过预处理的数据应用各算法后得到的平均准确率。

卷积神经网络。对于其他数据集，卷积神经网络保持了优势，但油 数据集除外，在该数据集上，MOCHM取得了最高的准确率。
总体而言，使用预处理数据时，算法的性能提升是显著的。我们可以从两个方面评估这种提升：1）数据集上算法平均准确率的提高，重点关注 MOCHM 和 K近邻算法；2）标准差的降低，这一点在一维卷积神经网络中有所体现。

MOCHM 在 油 数据集上的准确率提升了约 50%，从而取得了更好的整体性能。K近邻算法在 大麦、鹰嘴豆 和 高粱 数据集上的准确率也取得了显著提升，超过了卷积神经网络。关于标准差的降低，卷积神经网络在所有数据集上都呈现出这一特征。尽管其准确率提升不如其他算法显著，但标准差的降低表明稳定性更高，从而生成了可靠性更高的模型。这一点在一维卷积神经网络于 大麦 数据集上取得的结果中表现得尤为明显：在未经预处理的实验中，其准确率为47.4 ± 4.7，而在经过预处理的实验中则为 46.16 ± 0.97。

尽管在所进行的实验中，一维卷积神经网络在准确率方面取得了出色的结果，但仍有必要分析结果之间的权衡。由于与其他算法相比，一维卷积神经网络的计算成本较高，因此有必要分析在哪些情况下更适合使用该算法。表 7 展示了每种算法在各数据集上训练模型所需的时间。显然，数据集越大，一维卷积神经网络所需的训练时间越长。此外，在大多数情况下，只有该算法需要数分钟来训练模型。对于样本数量众多且维度较高的问题，高效地使用预处理算法并结合 K近邻等简单算法，或许是比使用一维卷积神经网络更可行的替代方案。

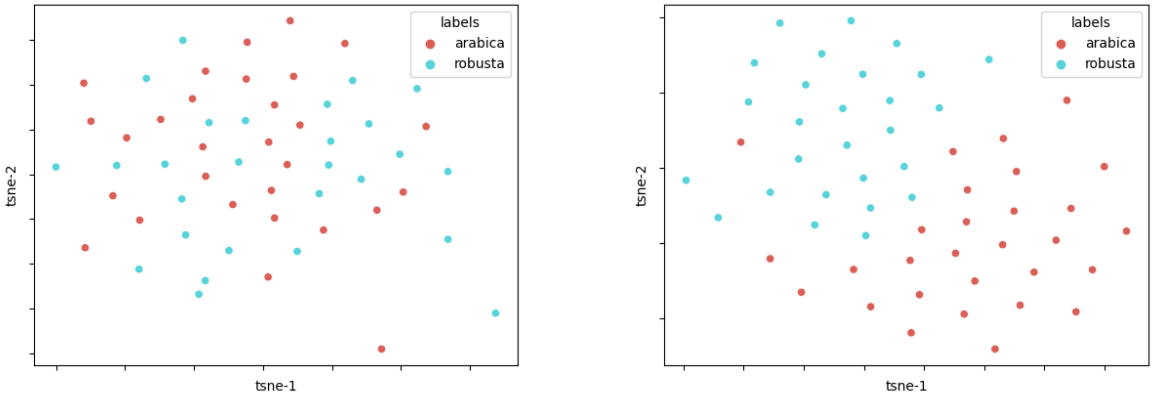
数据可视化： 为了以二维方式直观展示数据预处理的影响，我们对未进行和进行预处理的情况应用了 t-分布随机邻居嵌入（t-SNE）（Van der Maaten 和 Hinton，2008年）。

图2 展示了咖啡数据集的原始数据及预处理结果。无法从原始数据中观察到能够区分数据集中两个类别的清晰分组，因此简单的分类器难以通过

数据集	MOCHM	偏最小二乘判别分析	支持向量机	K近邻	一维卷积神经网络
大麦	-	0.1328	1.7173	0.3353	603.12
鹰嘴豆	-	0.0855	0.8775	0.2436	485.30
高粱	-	0.0255	0.2826	0.1020	443.05
咖啡	22.10	0.1196	0.0083	0.0795	4.22
Meat	18.30	0.0377	0.0136	0.0277	45.27
Oil	31.55	0.0152	0.0194	0.0231	106.07
水果	32.26	0.0377	0.3920	0.2325	81.13

表7：各算法的模型训练时间（秒）。

线性函数正确地分离样本。然而，观察预处理数据后，我们看到数据以紧凑的方式分组，表明SG 滤波器有助于模型区分数据。



(a) (b)

图2：使用 t-SNE 可视化咖啡数据集 (a) 原始数据和 (b) 预处理后的数据的空间排列。

结果表明，对光谱数据应用预处理算法以降低噪声十分重要，而且无论是提高准确率还是增强模型稳定性，这些算法都能因此取得更好的结果。卷积神经网络在光谱数据分析中展现出良好前景，因为结果表明其性能优于传统算法。值得一提的是，一维卷积神经网络在小型数据集上具有良好的泛化能力。此外，K近邻等算法证明是多数情况下可行且简单的选择；与数据预处理算法结合使用时，它们能够取得良好结果，且计算成本不高。基于所得结果，我们将在下一节进一步开展案例研究，使用一维卷积神经网络、偏最小二乘判别分析和K近邻，因为这些算法在此前的实验中取得了最佳平均性能。

4.4 新冠肺炎案例研究

SARS-CoV-2病毒（新冠肺炎）大流行已成为近几十年来全球面临的最严重且最具挑战性的健康问题。2019年12月，中国武汉区确诊首例病例（Zhu et al.，2020年），此后病毒迅速在中国及其他国家传播。2020年1月，世界卫生组织（WHO）宣布进入全球紧急状态（OMS，2020年）。

由于病毒传播迅速且传播率高，自首例确诊病例出现后的短短几个月内，全球已有144个国家报告发现该病毒（OMS，2020年）。为实施卫生限制措施、评估并报告病毒的传播情况，在出现病例的多个地区启动了大规模检测。然而，现有检测方法存在一些不理想的特性，可能影响为遏制病毒而采取的决策。首先，现有检测成本高昂，导致欠发达国家无法开展检测。其次，获得患者是否感染病毒的确切结果需要耗费时间。第三，假阴性检测率高。在上述问题中，假阴性率高是一个关键因素，会直接影响病毒遏制工作。假阴性检测者可能放松卫生机构实施的卫生措施，从而加速病毒传播（West 等人，2020）。近期，利用人工智能自动诊断新冠肺炎的研究兴趣日益浓厚（Khana 等人，2021）。

在此背景下，利用光谱仪数据进行自动化诊断的应用有望成为传统检测的替代方案。光谱仪便于携带且易于使用，结合计算分类模型，这类设备采集的数据或许能在几分钟内提供快速诊断。这些特性有助于改善当前检测中两项不利因素的影响：高昂的成本和获取结果所需的时间。最后，为了减少假阴性结果——这是当前检测中发现的第三个问题——分析结果时除准确率（ACC）外，还将纳入另外两项指标，即特异度（ESPEC）和敏感度（SE）；这两项指标在生物医学应用中十分常见。

特异度的计算公式为：

$$ESPEC = \frac{TN}{TN + FP} \tag{2}$$

其中，TN 表示真阴性值，即属于对照组且被分类器正确分类的样本；FP 表示假阳性值，即被判定为新冠肺炎阳性的对照组样本。

敏感度指标（也称为召回率）的计算公式为：

$$SE = \frac{TP}{TP + FN} \tag{3}$$

其中，TP 表示真正例，即被正确分类的病毒样本；FN 表示假负例，即属于新冠肺炎组但被分类为对照组的样本。

通过使用这些指标，我们可以回答两个重要问题：

- 在标记为对照组的样本中（此处指未感染新冠肺炎的样本），模型正确分类了多少个?
- 对于所有新冠肺炎阳性样本，模型将其中多少个分类为携带病毒?

尤其是对于新冠肺炎，由于其具有较强的传播性和较高的致死率，我们尤其关注回答第二个问题，因为这一信息与灵敏度指标直接相关。模型获得较高的灵敏度率，意味着我们能够确保样本中存在新冠肺炎时，将其分类为病毒阳性；也就是说，这能减少假阴性诊断的数量，而这正是当前用于检测该病毒的检测方案所面临的问题。

Yin等人(2021)收集了对照组和新冠肺炎感染患者的光谱样本，并将其公开供公众使用。所提供的数据集包含309个样本，其中150个属于对照组，159个属于病毒感染患者组。采集的光谱如图3所示。

表8 展示了各算法应用于新冠肺炎数据集（Yin等人，2021）时，在原始数据和经SG滤波器预处理的数据上得到的准确率、特异度和敏感度结果。与上一节所报告的其他数据集的结果类似，一维卷积神经网络在预处理步骤中使用SG滤波器时取得了最佳结果，平均准确率为 96.5 ± 1.1 。此外，值得注意的是，其特异度和敏感度数值都很高，平均分别为 98.06 ± 1.86 和 94.06 ± 2.43。如前所述，敏感度值低意味着假阴性数量较多，而这被认为是新冠肺炎诊断中最糟糕的情况。K近邻的结果便体现了这种情况：对于未经预处理的数据，其假阴性率平均超过25% ；对于预处理后的数据，则超过12% 。这一数值表明，该算法在新冠肺炎相关实际应用中的可行性不足。总体而言，偏最小二乘判别分析算法在所有分析指标上都取得了较有前景的结果，但仍逊于一维卷积神经网络。

新冠肺炎	ACC	特异度	SE
一维卷积神经网络	94.19 ± 1.6	96.82 ± 2.3	94.72 ± 3.39
Savitzky-Golay平滑 + 1D-卷积神经网络	**96.5 ± 1.1**	**98.06 ± 1.86**	**94.06 ± 2.43**
偏最小二乘判别分析	93.7 ± 1.81	92.85 ± 1.8	90.3 ± 2.1
Savitzky-Golay平滑 + 偏最小二乘判别分析	94.17 ± 0.7	95.12 ± 1.15	92.1 ± 1.9
KNN	83.17 ± 4.84	92.78 ± 7.1	73.59 ± 10.1
Savitzky-Golay平滑 + K近邻	93.2 ± 1.58	97.23 ± 0.9	88.71 ± 1.1

表8：新冠数据集的结果。

图 4 展示了采用预处理的各算法的混淆矩阵。每个矩阵对应的模型，其准确率最接近通过交叉验证技术获得的平均准确率。与一维卷积神经网络相关的混淆矩阵显示

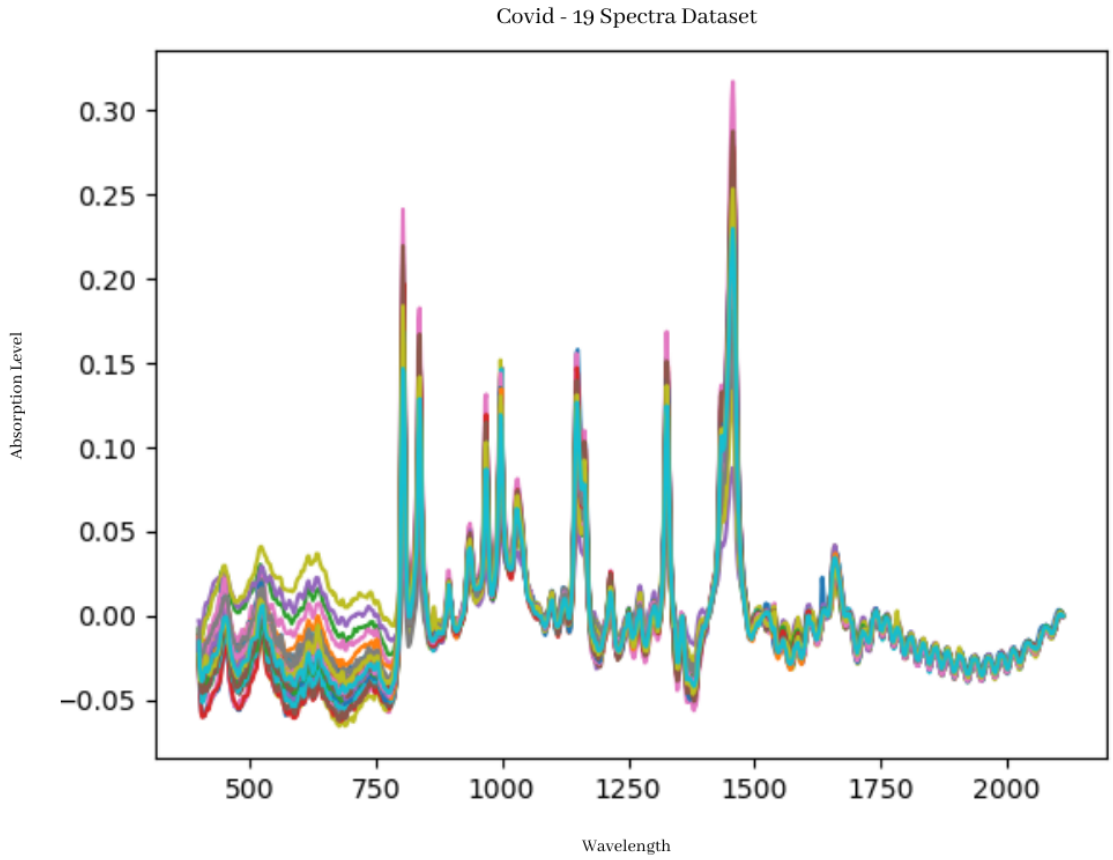
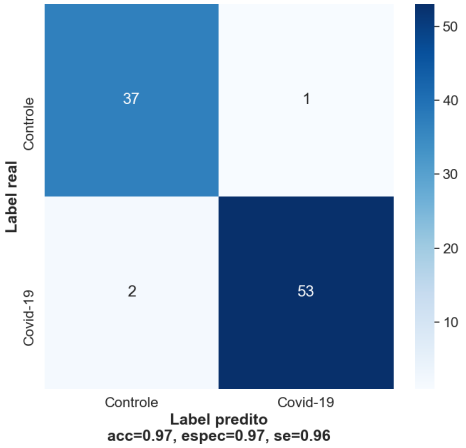


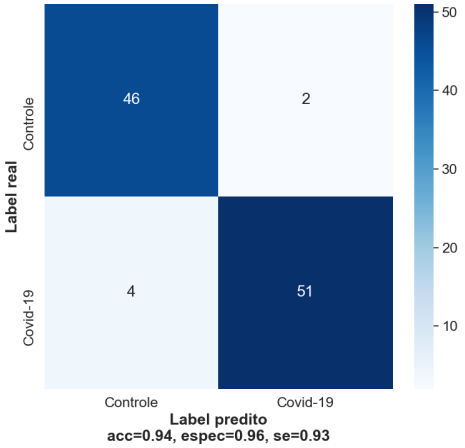
图 3： 从新冠肺炎数据集中采集的光谱中各波长（纳米）的吸收水平（Yin等人，2021）。

命中率较高，且只有两个假阴性，这证实了此前展示的结果。K近邻混淆矩阵显示假阴性数量较多，与表 8中的结果一致。

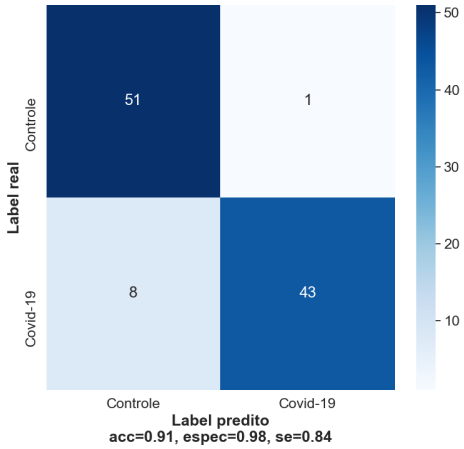
数据的空间排列如图5所示。图5a 展示了原始数据的可视化结果，图5b 展示了预处理后的数据可视化结果，而图5c 和5d 分别展示了一维卷积神经网络从原始数据和预处理数据中提取的特征。图5a 显示，原始数据存在重叠，在两个类别之间未能很好地分组。图5b 中，尽管应用SG滤波器



(a) SG + 1D-CNN - 混淆矩阵



(b) SG + 偏最小二乘判别分析 - - 混淆矩阵

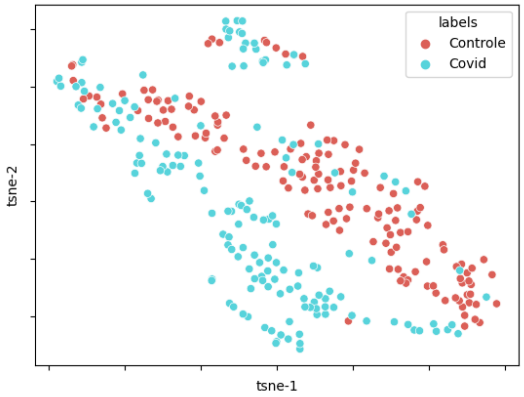


(c) SG + K近邻 - - 混淆矩阵

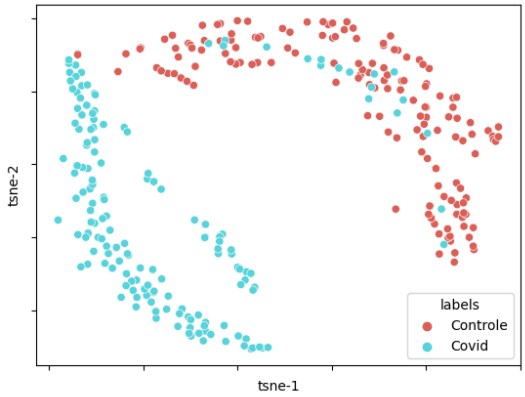
图 4： 模型得到的混淆矩阵。

改善了数据的分离效果，使其形成定义更清晰的组，但其中一组的数据仍存在重叠，这使得缺乏特征提取机制的算法（如K近邻）难以完成分类任务。这个问题可以通过一维卷积神经网络中的特征提取器解决。图5c 和5d展示了一维卷积神经网络提取的特征，可以看出，与图5a 和5b相比，组的排列更加合理。尤其是图5c中可以看到，即使不使用预处理技术，一维卷积神经网络也能够对数据进行分组。

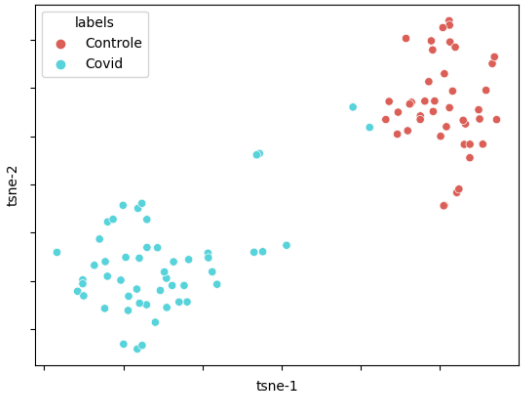
近期研究将光谱技术与分类算法相结合，分析从新冠肺炎样本中采集的数据。Yin等人（2021）将支持向量机算法应用于其数据集，以区分对照组和病毒感染组。Barauna等人（2020年）使用衰减全反射傅里叶变换红外光谱采集新冠肺炎样本的数据。在他们的研究中，对照组使用了111个样本，另有70个新冠肺炎阳性样本。作者将遗传算法与线性



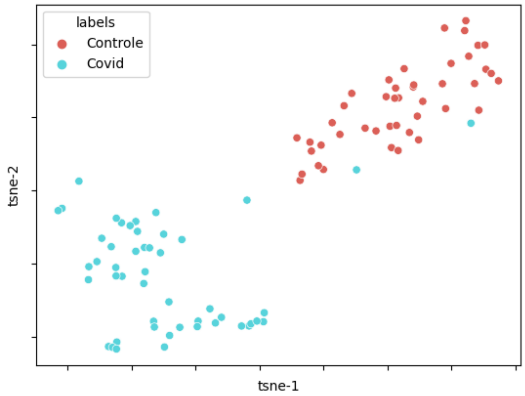
(a) 原始数据的空间排列。



(b) 预处理数据的空间排列。



(c) 特征的空间排列
由一维卷积神经网络从原始数据中提取。



(d) 特征的空间排列
由一维卷积神经网络从预处理数据中提取。

图 5： D ta 使用 t-SNE 对新冠肺炎数据集进行空间排列可视化 t.

判别分析（GA-LDA）应用于经SG滤波器预处理的数据，以执行分类任务。Carlomagno等人（2021）还提出使用卷积神经网络对新冠肺炎光谱样本进行分类，其数据集包含33个对照样本和30个病毒阳性样本。表9 展示了本研究及此前文献中提及的研究所获结果。

尽管各项研究所使用的数据集并不相同，但本研究采用的方法以及Carlomagno等人（2021）采用一维卷积神经网络后，在所用指标方面均取得了更优的结果。正如前文比较研究中所指出的，数据预处理等步骤会显著影响最终分类结果。值得一提的是，Yin等人（2021）使用相同的COVID-19数据集，结果显示一维卷积神经网络优于支持向量机。 Barauna等人（2020年）采用遗传算法线性判别分析得到的结果显示，其灵敏度值颇具潜力，比本研究所得结果高出1 %。然而，他们没有报告标准差值。由以下研究报告的结果

算法	ACC	特异度	SE
SG + 1D-CNN*	96.5 ± 1.1	98.06 ± 1.86	94.06 ± 2.43
Yin等人 (2021)*	91 ± 4	93 ± 6	89 ± 7
Barauna等人 (2020年)**	90 ± 0	89 ± 0	95 ± 0
Carlomagno等人(2021)**	97.8 ± 0	98 ± 0	97.5 ± 0

* 使用同一数据库进行的测试。

** 测试结果报告作者自己的数据集，而非公开可用。

表9： 我们的结果与文献中其他研究的比较，涉及通过 SARS-CoV-2 光谱学样本获得的不同光谱数据集。

Carlomagno等人（2021）和Barauna等人（2020年）可能会因采用交叉验证等方法而发生变化。总体而言，对于所研究的数据集，使用经SG滤波器预处理的数据和一维卷积神经网络所获得的结果，在准确率和敏感度指标方面表现出良好潜力，表明利用光谱数据构建自动化系统以辅助新冠肺炎诊断是可行的。

5 结论

在本研究中，我们提出了一种一维卷积神经网络，用于解决光谱数据分类问题。首先，我们对一维卷积神经网络与支持向量机、偏最小二乘判别分析、K近邻等机器学习算法，以及 MOCHM 进行了比较研究，旨在评估这些算法在化学计量学问题中的表现。我们研究了数据预处理的影响，并分析了这项技术如何影响各算法的性能。结果表明，在应用于光谱数据集时，一维卷积神经网络的平均表现优于所研究的其他算法。此外，数据预处理被证明是问题建模过程中的关键步骤，它会影响平均准确率的提升和标准差的降低，进而影响结果的可靠性。除此之外，对于所研究的数据集而言，数据集中样本数量较少似乎并未妨碍模型的泛化。鉴于光谱技术的特点有助于其应用于各种日常问题，我们使用一维卷积神经网络、偏最小二乘判别分析和 K近邻开展了一项案例研究，并采用来自新冠肺炎患者样本的数据进行比较。我们将这些算法应用于一个公开数据集，该数据集包含对照组患者和新冠肺炎感染者的样本。一维卷积神经网络结合 Savitzky–Golay 滤波器取得了最佳结果，并用于与其他涉及基于光谱数据诊断新冠肺炎的研究进行比较。结果表明，采用一维卷积神经网络的方法在准确率、特异度和敏感度方面表现最佳。尤其是在新冠肺炎检测等问题中，高灵敏度率非常理想且必不可少，因为假阴性代表最糟糕的情况。本研究获得的结果平均为：准确率 96%、特异度 98%、敏感度 94%，表明利用光谱技术获取的数据并以一维卷积神经网络作为分类器来诊断病毒是可行的。未来的研究可以进一步探讨数据预处理步骤中所用信号处理算法的影响。此外，许多问题还包含光谱之外的信息，这些信息也可以加以利用。

致谢

R.A. Krohling 感谢巴西研究机构 Conselho Nacionalde Desenvolvimento Científico e Tecnólogico（CNPq）提供的资助，项目编号：304688/2021-5。

参考文献

Ethem Alpaydin. 机器学习导论. 麻省理工学院出版社，2020年。

Valério G. Barauna、Maneesh N. Singh、Leonardo Leal Barbosa、Wena Dantas Marcarini、Paula Frizera Vassallo、Jose Geraldo Mill、Rodrigo Ribeiro-Rodrigues、Patrick H. Warnke 和 Francis L Martin。利用简便的衰减全反射傅里叶变换红外光谱和分析算法，在现场超快速检测 SARS-CoV-2 感染：高敏感度和高特异度。 *medRxiv*，2020年。doi：10.1101/2020.11.02.20223560。

C Carlomagno、D Bertazioli、A Gualerzi、S Picciolini、PI Banfi、A Lax、E Messina、J Navarro、L Bianchi、A Caronni 等。新冠肺炎唾液拉曼指纹：检测当前及既往 SARS-CoV-2 感染的创新方法。《科学报告》，11(1):1–13，2021。

Céline Chauvergne、Laurent Bonnal、Denis Bastianelli、Hélène Carrère、Yves Griveau、Marie- Pierre Jacquemot、Matthieu Reymond、Valérie Méchin、VirginieRossard 和 ÉricLatrille。不同光谱仪采集的有机样本近红外光谱数据集。*Datain Brief*，32：106264，2020年。

Hui Chen、Chao Tan 和 Zan Lin。用于近红外光谱多变量校准的极限学习机集成。《光谱化学学报*A*辑：分子与生物分子光谱学》，229：117982，2020年。

Yuan-Yuan Chen 和 Zhi-Bin Wang。基于卷积神经网络的近红外光谱端到端定量分析建模。《化学计量学杂志》，33（5）：e3122，2019年。

Chenhao Cui 和 Tom Fearn。用于多变量回归的现代实用卷积神经网络：近红外校准应用。*Chemometrics and Intelligent LaboratorySystems*，182：9–20，2018年。

José GM Esgario 和 Renato A Krohling。利用逼近理想解排序法结合多准则信息的最小生成树聚类。载于*2018*年*IEEE*国际模糊系统会议 *(FUZZ-IEEE)*，第 1–7 页，2018年。

Monika Gniadecka、Peter Alshede Philipsen、Sonja Wessel、Robert Gniadecki、Hans Christian Wulf、Sigurdur Sigurdsson、Ole Faurskov Nielsen、Daniel Højgaard Christensen、Jana Hercogova、Kristian Rossen 等。通过拉曼光谱法和神经网络进行黑色素瘤诊断：完整癌组织中蛋白质和脂质的结构变化。《皮肤病学研究杂志》，122(2):443–449，2004年。

Ian Goodfellow、Yoshua Bengio、Aaron Courville 和 Yoshua Bengio。《深度学习》，第 1 卷。麻省理工学院出版社，剑桥，2016。

Rita Hayati、Agus Arip Munawar 和 F Fachruddin。增强近红外光谱数据，以提高完整芒果质量参数测定的预测准确度。《*Datain Brief*》，第 105571 页，2020年。

Benjawan Kaewseekhao、Noppadon Nuntawong、Pitak Eiamchai、Sittiruk Roytrakul、Wipa Reechaipichitkul 和 Kiatichai Faksri。通过拉曼光谱法和表面增强拉曼光谱检测的结核病患者血清蛋白质组学光谱数据集。*Data in Brief*，28:104891，2020年。

Muzammil Khana、Muhammad Taqi Mehrana、Zeeshan Ul Haqa、Zahid Ullaha、Salman Raza、Naqvia Mehreen 和 Ihsanb Haider Abbass。人工智能在 COVID-19 大流行中的应用：全面综述。*Expert Systems with Application*，185(1): 115695，2021。

Frédéric Kosmowski 和 Tigist Worku。用于品种鉴定的微型近红外光谱仪评估：埃塞俄比亚大麦、鹰嘴豆和高粱案例。*PloS one*，13 (3):e0193620，2018年。

Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。利用深度卷积神经网络进行 ImageNet 分类。神经信息处理系统进展，25：1097–1105，2012年。

Y. LeCun、B. Boser、J. S. Denker、D. Henderson、R. E. Howard、W. Hubbard 和 L. D. Jackel。将反向传播应用于手写邮政编码识别。*Neural Computation*，1(4)：541–551，1989年。

Loong Chuen Lee、Choong-Yeun Liong 和 Abdul Aziz Jemain。ATR-FTIR光谱数据预处理（dp）实践策略的当代综述。*Chemometrics andIntelligent Laboratory Systems*，163：64–75，2017年。

Ana Mara Ferreira Lima、Camila Ribeiro Daniel、Ricardo Scarparo Navarro、Benito Bodanese、Carlos Augusto Pasqualucci、Marcos Tadeu Tavares Pacheco、Renato Amaro Zângaro 和 Landulfo Silveira Jr.。通过拉曼光谱法对体内和离体的非黑色素瘤皮肤癌、角化病与正常皮肤组织进行鉴别。*Vibrational Spectroscopy*，100：131–141，2019年。

Ting Liu、Zhongren Li、Chunxia Yu 和 Yuhua Qin。基于深度自编码神经网络的近红外光谱特征提取。*Infrared Physics & Technology*，87：124–128，2017。

Laura M McIntosh、Randy Summers、Michael Jackson、Henry H Mantsch、James R Mansfield、Marilyn Howlett、A Neil Crowson 和 John WP Toole。利用近红外光谱无创筛查皮肤病变。《皮肤病学研究杂志》，116(1)：175–181，2001年。

OMS。存档：世界卫生组织时间线—新冠肺炎，2020年。网址 https://www.who.int/news/item/27-04-2020-who-timeline---covid-19。OMS。世界卫生组织《2019冠状病毒病（新冠肺炎）情况报告》，2020年。

Avneet Pannu。人工智能及其在不同领域的应用。人工智能，4(10)：79–84，2015年。

Dan Peng、Linqing Li、He Guo、Yanlan Bi 和 Guolong Yang。基于熵理论和遗传算法的近红外光谱频域信息提取。载于第五届国际信息科学与控制工程会议（*ICISCE*），第1143–1147页，2018年。

Åsmund Rinnan、Frans Van Den Berg 和 Søren Balling Engelsen。近红外光谱最常用预处理技术综述。《分析化学趋势》，28(10)：1201–1222，2009年。

Abraham Savitzky 和 Marcel JE Golay。通过简化的最小二乘法进行数据平滑和微分。《分析化学》，36(8)：1627–1639，1964年。

Robert J Schalkoff。人工神经网络。麦格劳-希尔高等教育，1997年。

Silvia Seoni、Federica Veronese、Vanessa Tarantino、Elisa Zavattaro、Massimo Salvi、Nicola Michielli、Bruno De Santi、Filippo Molinari、Paola Savoia 和 Kristen M Meiburger。采用冷刺激和近红外光谱对光化性角化病进行无创分析。载于第*41*届*IEEE*医学与生物学工程学会国际年会（*EMBS*），第467–470页，2019年。

Barbara Stuart。红外光谱。《柯克-奥思默化学技术百科全书》，2000年。

Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究期刊》，9(11)，2008年。

Colin P West、Victor M Montori 和 Priya Sampathkumar。新冠病毒检测：假阴性结果的威胁。载于《梅奥诊所学报》，第95卷，第1127页，2020年。

Li-yan Wu、Dong-sheng Yang、Ming-zhi Zhao 和 Fan-xin Meng。主成分分析-人工神经网络在近红外光谱法无损测定云芝中的应用。载于*2012*年第五届国际智能计算技术与自动化会议，第106–109页。IEEE，2012年。

Yuhong Xiang、Jing Tian、Zhuoyong Zhang 和 Yinmei Dai。基于近红外光谱和广义回归神经网络的子宫内膜癌诊断。载于*2010*年第六届国际自然计算会议，第3卷，第1228–1231页。IEEE，2010年。

Xu Yan、Sheng Zhang、Hao Fu 和 Haibin Qu。结合卷积神经网络与在线拉曼光谱监测山羊角水解过程。《光谱化学学报*A*辑：分子与生物分子光谱学》，226：117589，2020年。

Gang Yin、Lintao Li、Shun Lu、Yu Yin、Yuanzhang Su、Yilan Zeng、Mei Luo、Maohua Ma、Hongyan Zhou、Lucia Orlandini 等。基于血清拉曼光谱对新冠肺炎进行高效初筛。《拉曼光谱学杂志》，第 1–10 页，2021。

Chen Yuanyuan 和 Wang Zhibin。基于集成卷积神经网络的红外光谱定量分析建模。*Chemometrics and Intelligent LaboratorySystems*，181：1–10，2018年。

Aston Zhang、Zachary C. Lipton、Mu Li 和 Alexander J. Smola。动手学深度学习。2020年。 https://d2l.ai。

Xiaolei Zhang、Tao Lin、Jinfan Xu、Xuan Luo 和 Yibin Ying。Deepspectra：一种用于定量光谱分析的端到端深度学习方法。*Analytica Chimica Acta*，1058：48–57，2019年。

Wenbin Zheng、Hongping Shu、Hong Tang 和 Haiqing Zhang。基于核极限学习机的光谱数据分类。*Chemometrics and Intelligent Laboratory Systems*，192：103815，2019年。

Na Zhu、Dingyu Zhang、Wenling Wang、Xingwang Li、Bo Yang、Jingdong Song、Xiang Zhao、Baoying Huang、Weifeng Shi、Roujian Lu 等。2019年在中国肺炎患者中发现的一种新型冠状病毒。*New England Journal of Medicine*，382(8)：727–733，2020年。

Andrey Zyubin、Vladimir Rafalskiy、Anna Tcibulnikova、Karina Matveeva、Ekaterina Moiseeva、Alina Tsapkova、Ilia Samusev、Valery Bryukhanov 和 Maksim Demin。通过表面增强拉曼光谱法获得的健康人和心血管病理患者人血小板数据集。*Data in Brief*，29:105145，2020年。