第 XX 卷第 XX 期
XXXX 年 XX 月

红 外 与 毫 米 波 学 报
J. Infrared Millim. Waves

Vol. XX, No. XX
XX, XXXX

文章编号:1001-9014(XXXX)XX-0001-11

DOI:10. 11972/j. issn. 1001-9014. XXXX. XX. 001

太赫兹光谱技术结合卷积神经网络鉴别三七产地的研究

王俊涛[1], 王胜峰[1], 李秋叶[2], 彭 滟[1*]

(1. 上海理工大学 光电信息与计算机工程学院 太赫兹技术创新研究院,上海 200093;
2. 文山州检验检测认证院,云南 文山 663099)

摘要:三七作为珍贵中草药,其药效品质与皂苷含量密切相关,而皂苷含量呈现显著产地差异性。为准确鉴别三七产地并保障药材质量,该研究提出结合太赫兹精密光谱技术与卷积神经网络算法的新方法。实验收集中国云南省红河自治州、昆明市、曲靖市和文山自治州四个产区的40份三七样本,分别采用太赫兹光谱与高效液相色谱技术进行检测分析。基于获取的光谱和色谱数据,研究构建并训练了卷积神经网络模型以实现产地分类。实验结果显示,太赫兹光谱技术结合卷积神经网络模型的分类准确率达到92. 5%,较高效液相色谱数据结合同类型模型的分类准确率(82. 5%)提升显著。该发现证实太赫兹光谱技术在中草药成分解析与产地溯源方面具有应用潜力,为中药材的快速无损检测与精准识别提供了新型科学手段。

关 键 词:太赫兹光谱;卷积神经网络;三七;产地鉴别

中图分类号:O43 **文献标识码**: A

Research on the identification of *Panax notoginseng* origin using terahertz spectroscopy combined with convolutional neural networks

WANG Jun-Tao[1], WANG Sheng-Feng[1], LI Qiu-Ye[2], PENG Yan[1*]

(1. Terahertz Technology Innovation Research Institute, School of optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Wenshan Prefecture Inspection, Testing and Certification Institute, Wenshan 663099, China)

Abstract: As a valuable Chinese herbal medicine, *Panax notoginseng* exhibits therapeutic efficacy and quality closely associated with its saponin content, which demonstrates significant geographical variations. To accurately authenticate the geographical origin and ensure medicinal quality, a novel method integrating terahertz precision spectroscopy with a convolutional neural network (CNN) algorithm was proposed. 40 *Panax notoginseng* samples from 4 regions in Yunnan Province, China—Honghe Autonomous Prefecture, Kunming, Qujing, and Wenshan Autonomous Prefecture—were analyzed using terahertz spectroscopy and high-performance liquid chromatography (HPLC). A CNN model was constructed and trained based on the acquired spectral and chromatographic data to classify the geographical origins. Experimental results revealed that the terahertz spectroscopy combined with the CNN model achieved a classification accuracy of 92. 5%, significantly outperforming the 82. 5% accuracy attained by the HPLC-CNN model. This finding highlights the potential of terahertz spectroscopy in component analysis and geographical traceability of herbal medicines, providing a novel scientific approach for rapid, non-destructive, and precise identification of Chinese medicinal materials.

Key words: terahertz spectroscopy, convolutional neural network, *Panax notoginseng*, origin identification

收稿日期:2024-11-30,**修回日期**:2025-01-01 **Received date**:2024-11-30,**Revised date**:2025-01-01

基金项目:云南省院士(专家)工作站项目(202505AF350094);上海市"科技创新行动计划"技术标准项目(24DZ2200900)

Foundation items:Supported by the Yunnan Province Academician (Expert) Workstation Project (202505AF350094);Shanghai "Science and Technology Innovation Action Plan" Technical Standard Project (24DZ2200900).

作者简介(Biography):王俊涛(1999-),男,江苏苏州人,硕士研究生,主要研究领域为太赫兹技术在生物医学领域的应用,E-mail: 892945208@qq. com

[*] **通讯作者**(Corresponding author): E-mail: py@usst. edu. cn

引言

三七(*Panax notoginseng*),隶属于五加科,是一种多年生草本植物,是珍贵中草药之一,主要分布于中国云南省[1]。三七含有多种生物活性成分,如皂苷[2-5]、氨基酸[6]、多肽[6]、多糖[7]、黄酮类[8]等,其中皂苷为主要有效成分,对人体心血管和免疫系统有显著的积极影响。在三七的质量评价中,人参皂苷R_1、Rg_1和Rb_1的含量被认为是关键指标,而这些皂苷成分的含量受到地理气候条件的显著影响。不同地理环境下栽培的三七,其品质属性存在差异,进而对药效产生影响。地理环境对三七质量和药效的影响主要通过气候和土壤因素实现,如文山适宜的温度、降水及土壤条件利于总皂苷积累,而其他地区因条件差异,虽也有药效,但在某些方面不如文山。例如,文山地区的三七总皂苷含量较高,通常可达5.5%以上,而红河等其他地区相对较低;文山地区的三七根形态粗壮、颜色均匀淡黄,而其他地区的三七根相对较细,颜色因土壤和气候条件的不同而有所变化。这些差异影响药效的机制在于,皂苷类成分对止血、抗炎等作用关键,含量高的三七效果更显著;多糖类成分影响免疫调节,黄酮类成分关乎抗氧化效果[9-12]。文山是三七的原产地和主产地,被认为是道地药材,质量最好。但因受到连作障碍,三七种植地向外扩展至红河、昆明、曲靖等其他地区。因此,精确鉴别三七产地对于确保药材质量具有重要意义。

传统三七鉴别的主要方法有性状鉴别和显微鉴别。性状鉴别主要依赖于专业人员的视觉、触觉、嗅觉和味觉对药材的形状、大小、表面特征及气味等进行观察和感知。尽管这种方法直观且易于操作,但一旦药材被加工成片状或粉末,其鉴别难度便显著增加[13]。此外,感官评价方法容易受到人为因素和外界环境的干扰,导致其可靠性不足[14, 15]。显微鉴别则通过显微镜技术对药材的组织结构、细胞形态和内含物进行分析,这种方法适合于粉末状药材的鉴别。然而,由于同属药材在显微特征上的相似性,这使得它们之间的区分变得困难[16]。为了提高鉴别的准确性,现代研究中常采用高效液相色谱法(High performance liquid chromatography, HPLC)来进行中草药的定性识别和定量分析[17, 18]。尽管HPLC法具有较高的准确度,但其操作复杂、分析时间长、成本高且效率低[19-21]。

太赫兹波(Terahertz, THz),是指频率范围为0.1~20 THz的电磁波。它们占据了中红外和微波波段之间电磁频谱的很大一部分[22]。太赫兹波的独特优势在于其频率与大多数分子的振动和转动频率相吻合,能够通过共振吸收光谱提供分子的丰富特征信息[23-27]。太赫兹技术以其无损检测、高准确性、快速分析和强穿透力等特点,在多个领域内得到了广泛应用。近年来,太赫兹技术在中草药研究领域也取得了显著的进展。例如,2021年,Kou等人利用太赫兹光谱分析了西洋参的皂苷含量和产地[28];2022年,Shao等人应用太赫兹光谱检测艾绒的纯度和生长年限[29];2023年,Liu等人通过太赫兹光谱鉴别了陈皮的贮藏年份[30];同年,Pu等人也利用太赫兹光谱鉴别了陈皮的产地[31]。尽管太赫兹技术在中草药研究领域展现出巨大潜力,但其应用仍面临若干挑战。首先,中草药由多种复杂成分构成,这使得从丰富的光谱数据中准确提取关键信息并降低噪声和无效数据的影响成为一个难题。其次,太赫兹技术的有效应用依赖于一个庞大的光谱数据库,以便于进行准确的识别和分析。然而,当前该领域的数据库尚未达到理想的完善程度,这限制了太赫兹技术在中草药研究中的进一步发展。因此,构建和完善中草药的太赫兹光谱数据库,以及开发先进的数据处理技术以提高信息提取的准确性和可靠性,是当前研究中亟待解决的关键问题。

本研究提出了一种结合太赫兹精密光谱技术和神经网络算法的新方法,用于三七产地鉴别。首先,利用太赫兹光谱法和HPLC技术对皂苷Rb_1、Rg_1和R_1标准品进行检测,明确它们的太赫兹光谱特征和HPLC谱图特征。接着,从云南省4个主要三七产地(红河自治州、昆明市、曲靖市和文山自治州)收集了40份三七样本,并运用太赫兹光谱和HPLC技术,获取了这些样本的太赫兹光谱和HPLC数据,通过叠加10种噪声模式,生成400组样本数据,并分为训练集和验证集。最后,构建了卷积神经网络(Convolutional neural network, CNN)模型,并基于这些数据对模型进行训练,使其学习不同产地三七样本的数据特征。训练完成后,该CNN模型被应用于对三七样本进行产地分类,并通过与实际产地对比,以验证模型的分类准确性和可靠性。

1 材料与方法

1.1 实验材料

本研究所使用的人参皂苷标准品均购自Pure-

Chem Standard公司（中国成都），具体包括：人参皂苷 R_1（纯度>98%，CAS号：80，418-24-2）、人参皂苷 Rb_1（纯度>98%，CAS号：41，753-43-9）和人参皂苷 Rg_1（纯度>98%，CAS号：22，427-39-0）。

COC粉末购自Sigma-Aldrich公司（中国上海）。COC因其卓越的光学透明性和化学稳定性，在太赫兹光谱检测样品制备中被用作稀释剂。

三七样本采集自中国云南省的四个主要产地：红河自治州、昆明市、曲靖市和文山自治州。样本以块根形式收集，未经进一步纯化处理，以保持其天然状态和药用成分的完整性。

1.2　样本制备

太赫兹光谱分析样本制备：采用德国Retsch公司生产的MM400球磨机，以90 Hz的振动频率对三七样本研磨5 min，随后放置于红外烤灯下干燥，通过360目筛网获得平均粒径约为40 μm的粉末。筛选后的粉末与COC粉末在玛瑙研钵中混合均匀。最后，混合物经液压机在8吨压力下，压制成平均厚度为1 mm，直径为13 mm的三七压片样本。

HPLC分析样本制备：在10 mL圆底烧瓶中加入0.6 g预先干燥的三七粉末。向烧瓶中加入50 ml分析级甲醇（99.99%纯度，CAS号：67-56-1，购自美国Fisher公司），并将混合物浸泡过夜以确保充分提取。将浸泡后的混合物置于80 ℃的水浴中进行2 h的回流加热，以提高提取效率。加热完成后，将混合物冷却至室温，然后称重，如有溶剂蒸发导致重量减少，用甲醇补足。最后，振荡混合物以确保提取液的均匀性，并通过过滤得到清澈的滤液，用于后续的HPLC分析。

1.3　实验仪器与参数

太赫兹光谱分析设备与条件：使用布鲁克光学公司（Bruker Optics，Germany）生产的傅立叶变换红外光谱仪（vectex80v）进行太赫兹光谱分析。该仪器配备自冷汞灯作为远红外光源，以及氘代L-丙氨酸三甘氨酸硫酸盐检测器。有效波数覆盖范围30～680 cm^{-1}，信噪比优于10000:1。扫描参数设置为分辨率2 cm^{-1}，扫描次数128次，扫描速度5 kHz。为减少水蒸气干扰，所有测量在约22 ℃的真空环境中进行。实验所需采集的数据为1～20 THz，对应波数为33.3～666.7 cm^{-1}，数据均在仪器的有效波数范围内，其余超出所需范围的数据被剔除。每个样本经过3次测试，每次测试持续约3 min，单个样本的测试总耗时约为9 min。得到的结果为3次测试的平均值。

HPLC分析设备与试剂：使用安捷伦科技（Agilent Technologies，MA，USA）生产的Agilent 1200高效液相色谱仪进行HPLC分析。该仪器配备了双泵系统、自动进样器、柱温箱以及双波长紫外-可见光检测器。实验参数设定为：泵流速1.3 mL/min，进样量10 μL，柱温箱温度30 ℃，检测波长203 nm。每个样本经过3次测试，每次测试持续约70 min，单个样本的测试总耗时约为210 min。得到的结果也为三次测试的平均值。

1.4　卷积神经网络

1.4.1　CNN模型的选择与建立

CNN因其在数据分类、图像识别等领域的卓越性能而广泛应用于机器学习[32-37]。CNN的核心优势在于其能够自动从输入数据中学习并提取特征，这一特性使其在处理分类、回归等任务时表现出色[38-40]。面对三七样本这类涉及多种因素导致非线性变化的复杂系统，传统的分析方法已难以满足需求。因此，本研究采用一维CNN模型对三七的产地进行识别与分类。

本研究设计的CNN模型架构图如图1所示，包含两个卷积层、两个池化层、一个Dropout层和一个全连接层。卷积层（卷积核尺寸3，填充1，步幅1）用于提取光谱数据的局部特征，池化层（内核大小和步幅均为2）用于降低数据维度。Dropout层（丢失率0.5）减轻过拟合，全连接层通过sigmoid激活函数实现分类。模型训练采用Adam优化器，批量大小64，学习率0.0001。

1.4.2　数据处理

在本研究中，我们使用Python中的pandas库来导入Excel中的THz光谱数据（1～20 THz）和HPLC数据（0～70 min）。通过pandas.read_excel()方法，将Excel文件中的数据加载为DataFrame，并进一步转化为张量形式，以便后续的神经网络模型训练。

在数据预处理阶段，我们首先对导入的数据进行了详细的检查和清洗。对于THz光谱数据和HPLC数据，我们检查了数据的完整性和一致性，确保每个样本在1～20 THz的频率范围内和0～70 min的时间范围内都有完整的数据信息。对于缺失或异常的数据点，我们采用了插值或删除的方法进行处理，以保证数据的质量。对于数据中的噪声和波动，我们采用了平滑滤波的方法进行处理，以降低数据的噪声水平，突出数据的主要特征。同时，

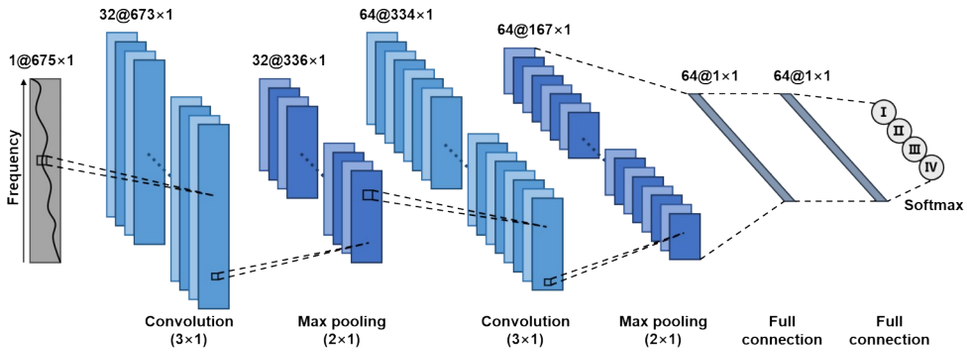

图1 1dCNN网络架构图

Fig. 1 1dCNN network architecture diagram

我们还对数据进行了归一化处理，将数据值缩放到[0,1]区间内，这有助于提高神经网络模型的收敛速度和训练效果，也便于后续的联合分析。

完成数据清洗和归一化处理后，我们基于40个样本，按照9:1的比例随机划分训练集和验证集，确保彼此相对独立。为了提升CNN模型在实际应用中的泛化能力，并确保其能够对现实环境中采集的含噪声光谱数据进行准确分类，我们采取了一种数据增强策略。具体而言，我们通过生成正态分布的噪声矩阵来对光谱进行扰动，这种噪声模拟现实数据采集中的随机误差。考虑到样本数量较少，为了增强数据的多样性，我们对每个数据样本添加十组随机噪声，进而提升模型的鲁棒性和泛化能力。基于40个样本，最终生成了400个样本数据，其中训练集360个，验证集40个，为CNN模型的训练和验证提供了充足且具有代表性的数据支持。

1.4.3 评价指标

在评估神经网络模型的性能时，准确率、精确率和召回率是三个核心指标，它们从不同角度描述了模型在分类任务中的表现。

准确率是用于衡量模型正确预测的样本数占总样本数的比例。准确率的计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

其中，TP(真正例)表示正确预测为正例的样本数，TN(真负例)表示正确预测为负例的样本数，FP(假正例)表示错误预测为正例的样本数，FN(假负例)表示错误预测为负例的样本数。准确率提供了模型整体分类性能的概览，但在类别分布不均衡的情况下，它可能无法全面反映模型的性能。

精确率反映了模型在预测正样本时的准确性，是衡量模型避免将负样本错误标记为正样本的能力。即在所有被预测为正例的样本中，真正为正例的比例。精确率的计算公式为：

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

召回率，也称为查全率或灵敏度。召回率衡量的是模型识别所有实际正例的能力，即在所有实际为正例的样本中，被模型正确预测为正例的比例。召回率的计算公式为：

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

2 结果

2.1 三七的3种主要皂苷标准品的太赫兹光谱分析

R_1、Rg_1和Rb_1是三七中的关键皂苷成分。图2展示了三种皂苷标准品的结构式及太赫兹吸收光谱。它们因为分子结构差异而具有的独特吸收特征。具体来说，R_1在7.3～10.9 THz频率范围内显示一个较宽且强度较高的尖峰，并在11.0～13.5 THz频率范围内有两个峰宽近似但峰高不一的小吸收峰。Rg_1则在8.9～10.7 THz和10.8～13.6 THz频率范围内各呈现两个半峰宽较窄、相对强度较低的吸收峰。Rb_1主要在6.8～10.7 THz频率范围内表现出一个平缓但强度较高的宽峰，在10.8～13.1 THz频率范围内有两个峰高近似的弱吸收峰。

这些吸收峰与人参皂苷分子的特定振动模式相对应。峰高能够指示相应振动模式的相对强度，高峰值强度通常与分子中较为显著的振动模式相关联。峰宽可能与分子振动模式集中程度有关，较大的峰宽表明存在较强的分子间相互作用或多重振动模式的叠加[41-47]。例如，R_1分子在7.3～10.9 THz频段内的强吸收峰，主要是由于其分子内部的C－O－C伸缩振动模式，这种振动模式在该频段内

能量集中,导致吸收峰强度较高。Rg_1分子在8.9~10.7 THz和10.8~13.6 THz频段内的多个吸收峰,是由于其分子中的多个羟基(-OH)的弯曲振动模式较为集中,且这些羟基之间的分子间氢键作用相对较弱,因此吸收峰呈现为多个相对独立的峰。Rb_1分子在6.8~10.7 THz处的宽峰,则是由于其分子内部存在多个不同类型的振动模式,如C-H伸缩振动、C-C伸缩振动等,这些振动模式相互叠加,同时分子间的范德华力作用较强,导致吸收峰宽度较大。结合这些皂苷标准品的特征峰与三七样品的光谱进行分析,可以实现对三七样本中人参皂苷成分的鉴定分析。

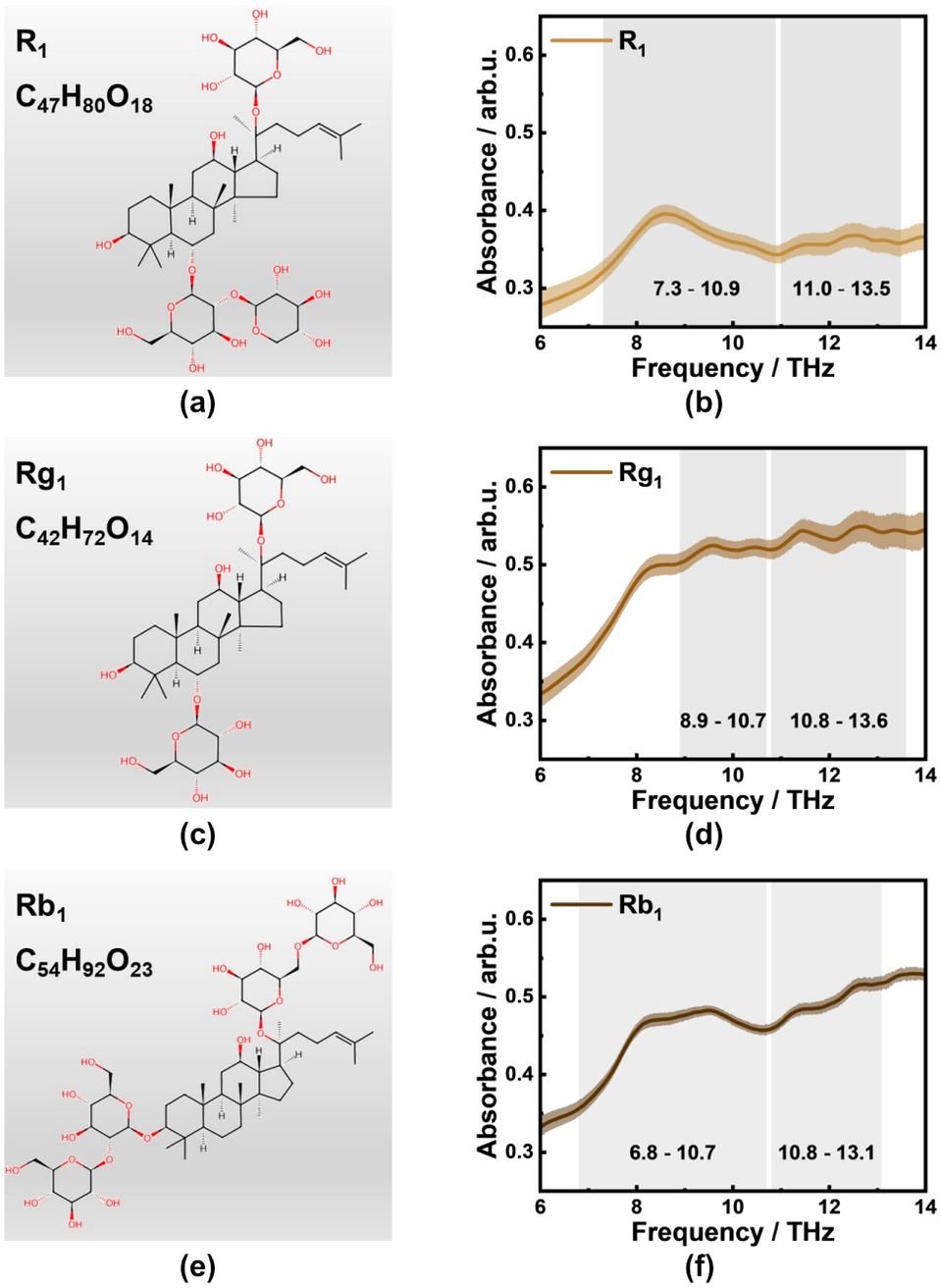

图2　三七主要皂苷的结构式及太赫兹吸收光谱:(a-b) R_1,(c-d) Rg_1,(e-f) Rb_1;误差棒为同一种皂苷样品重复测量偏差,吸收峰范围由寻峰算法结合手动校验得出

Fig. 2　Structural formula and terahertz absorption spectrum of main saponins of *Panax notoginseng*: (a-b) R_1, (c-d) Rg_1, (e-f) Rb_1; error bars represent intra-sample variability, The absorption peak range is obtained by the peak-seeking algorithm combined with manual calibration

2.2 不同产地的三七样本的太赫兹光谱分析

利用太赫兹技术，我们能够对三七中的皂苷成分进行鉴定分析。为此，我们对来自不同产地的三七样本进行了检测。具体而言，我们在云南省的红河自治州、昆明市、曲靖市和文山自治州四个主要三七产地各采集了10个样本，共计40个样本进行分析。所有样本均使用傅里叶变换红外光谱仪进行了测试。图3展示了各产地三七样本的太赫兹吸收光谱结果。来自4个不同产地的三七样本在8.2 ~ 9.6 THz、10.1 ~ 11.3 THz、11.7 ~ 13.7 THz三个频率范围内均展现出特征峰，这些峰与图2中所示的3种皂苷的特征峰区间相对应。8.2 ~ 9.6 THz区间主要反映了R_1、Rg_1和Rb_1的特征吸收，10.1 ~ 11.3 THz区间主要反映了R_1和Rb_1的特征吸收，而11.7 ~ 13.7 THz区间则主要与R_1和Rg_1的吸收相关联。不同产地样本间特征峰高度和宽度存在差异，这些差异可能归因于3个主要方面：(1)内在因素，例如三七中皂苷、多糖、氨基酸等成分含量的差异，这是影响特征峰差异的主要因素。(2)外在因素，三七生长环境引入的未知因素，例如气候变化和土壤中微量元素的差异等。(3)系统误差，包括仪器组件、测量环境和操作流程的差异，但在标准化操作下，这些因素的影响较小。

2.3 CNN模型结合太赫兹光谱鉴别三七产地

图4展示了CNN模型对太赫兹光谱数据集进行四分类的结果。该模型针对4个不同产地的三七样本进行了分类：Ⅰ红河(五角星)、Ⅱ昆明(圆形)、Ⅲ曲靖(三角形)、Ⅳ文山(正方形)。三维散点图(图4(a))直观地比较了模型的预测类别与样本的实际类别。每个点在三维空间中的位置由其预测类别(X轴)和实际类别(Y轴)决定，而样本编号则沿Z轴分布。图中的3D立体实心图形集中在对角线上，表明这些样本预测正确。相反，那些偏离对角线的2D空心图形则表示预测错误的样本。混淆矩阵(图4(b))提供了模型分类性能的精确统计信息。矩阵的行代表实际类别，列代表预测类别，每个单元格中的数字表示该类别的样本数量。对角线上的蓝色单元格表示正确分类的样本，而非对角线上的红

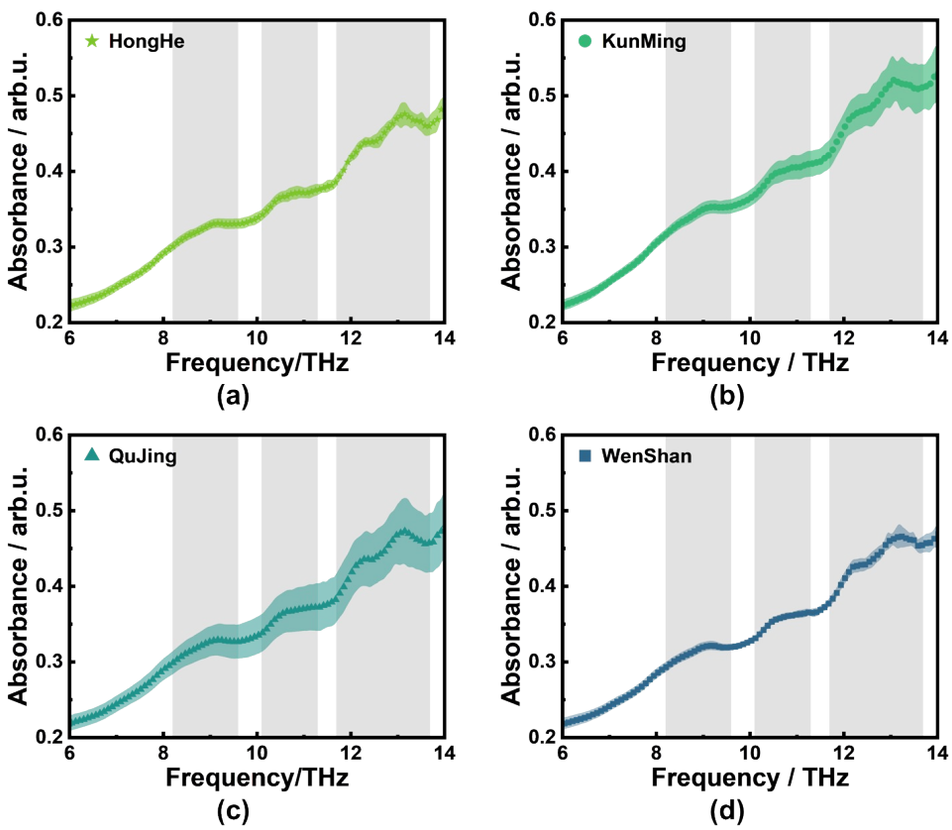

图3 云南省四个产地三七的太赫兹吸收光谱：(a)红河，(b)昆明，(c)曲靖，(d)文山；误差棒为同一产地10个样本间的偏差

Fig. 3 Terahertz absorption spectra of *Panax notoginseng* from four origins in Yunnan Province: (a) Honghe, (b) Kunming, (c) Qujing, and (d) Wenshan; the error bar is the deviation between 10 samples from each origin

色单元格则表示错误分类的样本。在40个样本中,有3个样本被错误分类,总体准确率达到了92.5%。具体误分类情况如下:1个红河样本被误判为昆明、1个昆明样本被误判为文山、1个文山样本被误判为红河。值得注意的是,该模型对曲靖产地的样本预测表现出色,精确率和召回率均达到了100%。而对于红河、昆明和文山产地的样本,模型的预测精确率和召回率也达到了90%。

2.4　HPLC分析三七中的皂苷成分

除了采用太赫兹技术对三七样本进行了检测外,我们还利用HPLC技术对皂苷标准品以及来自不同产地的三七样本进行了检测。在进行HPLC检测之前,所有样本均经过预处理步骤,以确保获得准确的皂苷提取液,具体处理方法详见第1.2节。随后,将这些提取液注入HPLC系统,以测定其中皂苷成分的保留时间。通过HPLC检测得到的皂苷标准品谱图以及各产地三七的代表性色谱图如图5所示。图5(a)展示了皂苷标准品的检测结果,可以看到,R_1、Rg_1和Rb_1的保留时间分别为26.43、29.88和53.37 min,三者可以很好地区分开。在4个产地三七的色谱图中,我们可以观察到三种主要皂苷成分——R_1、Rg_1和Rb_1的保留时间分别集中在26、29和61分钟左右。这些皂苷的保留时间在峰高和峰面积等方面表现出一定的差异,这种差异可能源自以下几个方面:(1)皂苷含量差异,不同产地的三七在皂苷含量及相对比例上存在差异,这就体现在HPLC色谱图中峰高和峰面积的差异;(2)色谱柱差异,色谱柱的材料、尺寸或填料径粒等方面的差异也可能影响皂苷的分离效果,进而导致保留时间的差异。

三种皂苷中,Rb_1的保留时间与标准品出现了较大的偏差,我们将三七色谱图中61分钟左右的特征峰与Rb_1标准品的峰进行形状和面积的对比,结合不同皂苷在色谱柱中的洗脱顺序、相对保留时间和相对强度,因此确定该特征峰为Rb_1。色谱柱的柱材会随水相使用过程中的冲击逐渐坍塌,需及时更换。而不同色谱柱之间存在材料、尺寸、填料粒径等方面的差异。因此,色谱柱更换后,各个皂苷的保留时间会发生改变。三七样品是一个复杂的体系,其中可能含有多种成分,这些成分可能会与Rb_1发生相互作用,影响其在色谱柱中的保留时间。例如,样品中的其他皂苷、多糖、蛋白质等成分可能会与Rb_1竞争色谱柱上的活性位点,或者改变流动相的性质,进而导致Rb_1的保留时间与标准品有所不同。在色谱分析过程中,温度、流动相组成、pH、离子强度等实验条件的微小变化都会引起保留时间的改变。尽管在实验过程中会尽量控制这些条件的稳定性,但仍然可能存在一些难以完全避免的波动,这些波动在不同样品的分析中可能会累积,导致三七样品与标准品的保留时间差异较大。

在对4个产地三七的HPLC分析中,我们发现R_1和Rb_1的峰值相对较小,而Rg_1的峰值则显著较高。这表明,在三七中,R_1和Rb_1的含量相对较低,而Rg_1的含量则相对丰富。特别是文山产地的三七,其Rg_1的峰值在所有样本中最高,这说明文山三七中Rg_1的浓度或纯度最高。Rg_1作为三七中的关键活性成分,其含量高低直接关系到三七的药效和品质。

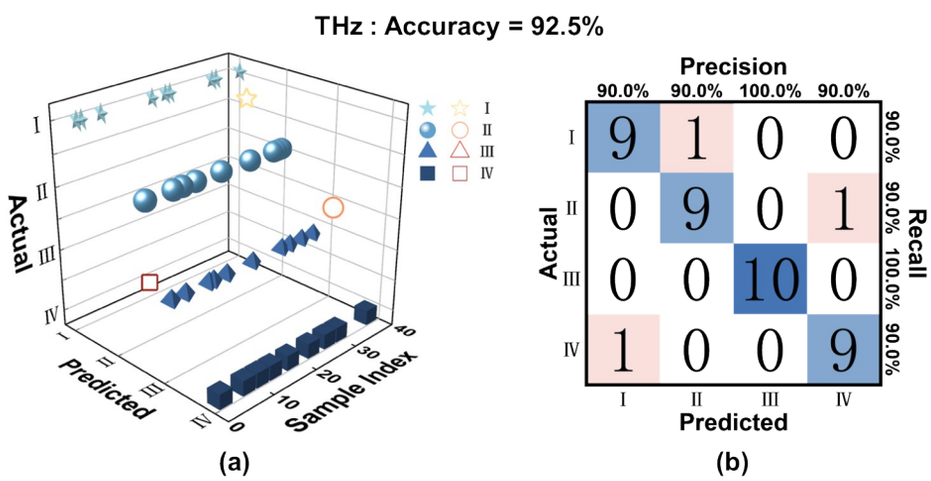

图4　CNN模型对太赫兹光谱数据的预测结果:(a)散点分类图,(b)混淆矩阵

Fig. 4　Prediction results of CNN model for terahertz spectral data: (a) scatter classification plot, (b) confusion matrix

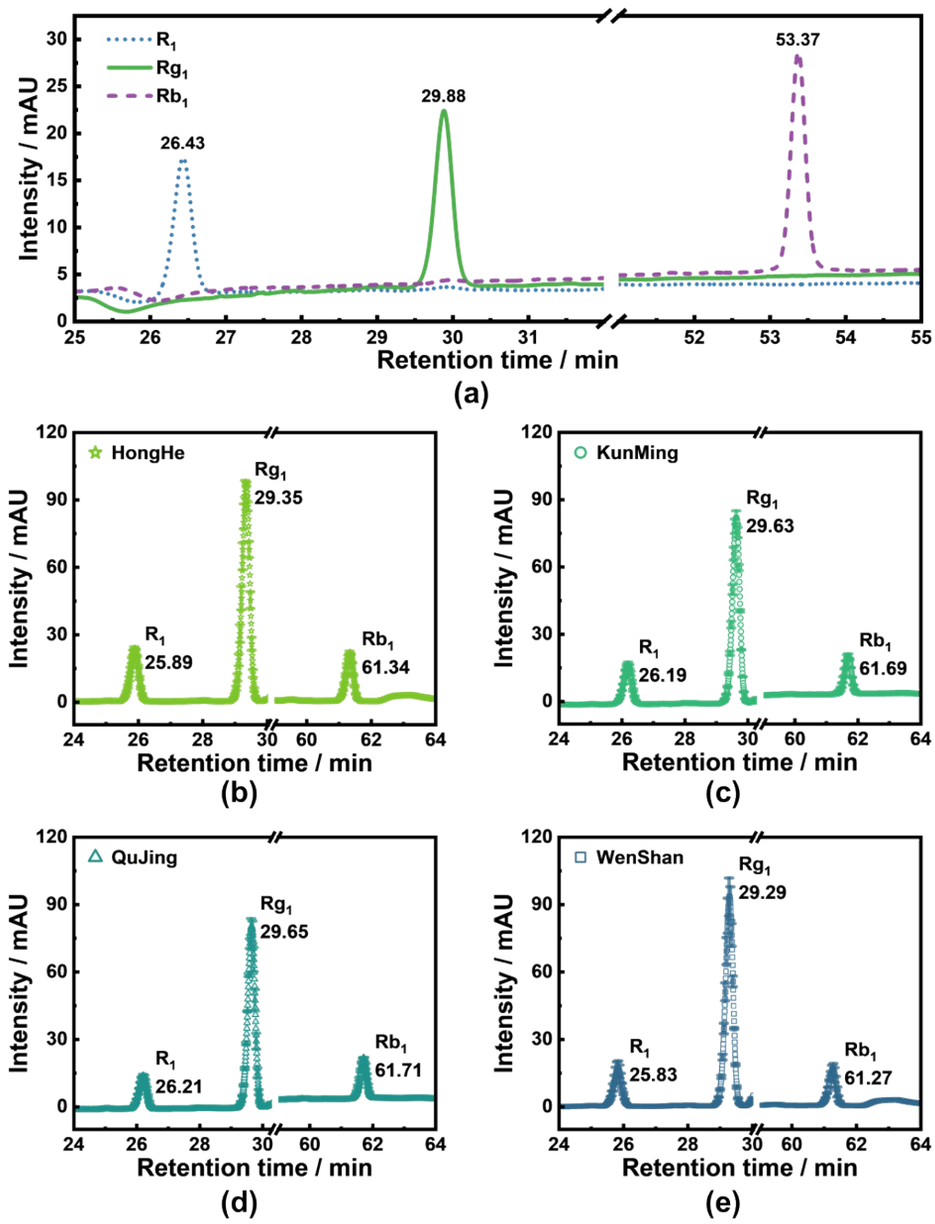

图5 皂苷标准品及云南省四个产地三七的HPLC检测结果:(a)皂苷标准品,(b)红河,(c)昆明,(d)曲靖,(e)文山;误差棒为各产地10个样本之间的误差

Fig. 5 HPLC test results of saponin standard product and *Panax notoginseng* from four origins in Yunnan Province: (a) saponin standard product, (b) Honghe, (c) Kunming, (d) Qu Jing, (e) Wenshan; the error bar is the error between 10 samples from each origin

2.5 CNN模型结合HPLC鉴别三七产地

同样地,我们将HPLC数据分为四个类别,分别对应于三七的四个主要产地:红河、昆明、曲靖和文山。利用这些数据,我们训练了CNN模型,并进行了预测。预测结果如图6所示。三维散点图(图6(a))直观地展示了CNN模型预测结果与实际产地之间的对比。图中使用不同的形状来代表不同的产地:五角星-红河、圆形-昆明、三角形-曲靖、正方形-文山。每个点的位置由其预测类别(X轴)和实际类别(Y轴)决定。图中的3D立体实心图形集中在对角线上,表示模型正确预测的样本。而2D空心图形偏离对角线,则表示模型预测错误的样本。混淆矩阵(图6(b))提供了模型分类性能的详细统计信息。矩阵的行代表实际类别,列代表预测类别,每个单元格中的数字表示该类别的样本数量。对角线上的蓝色单元格表示正确分类的样本,而非对

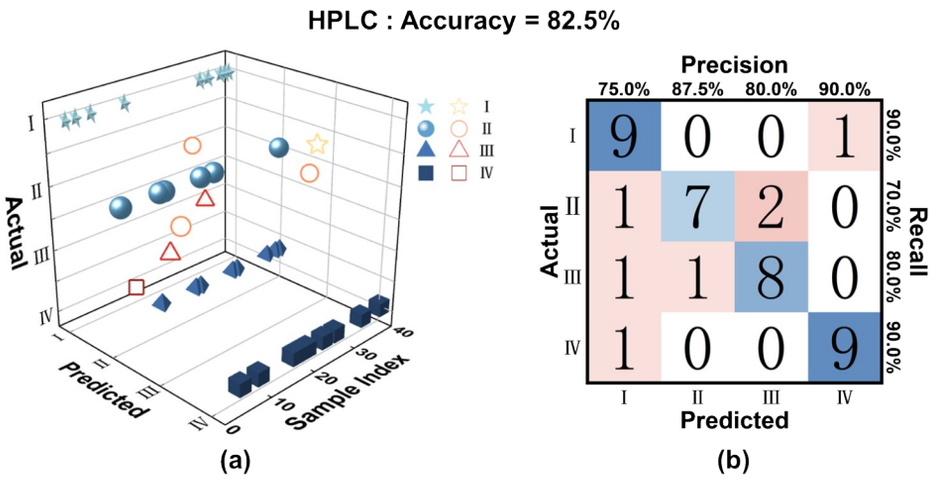

图6　CNN模型对HPLC数据的预测结果:(a)散点分类图,(b)混淆矩阵

Fig. 6　Prediction results of CNN model on HPLC data: (a) scatter classification plot, (b) confusion matrix

角线上的红色单元格则表示错误分类的样本。总的来说,40例样本中有7例被错误分类,准确率为82.5%。其中昆明、曲靖、文山均有1例被误判为红河。另外还有1例红河被误判为文山,2例昆明被误判为曲靖,1例曲靖为误判为昆明。对于红河、昆明、曲靖和文山4个产地的分类精确率分别为75.0%、87.5%、80.0%和90.0%,召回率分别为90.0%、70.0%、80.0%和90.0%。综合来看,该模型结合HPLC数据对文山三七的分类效果最佳,精确率和召回率均达到90.0%。

与太赫兹光谱的结果相比,HPLC数据的整体准确率以及各个分类的预测精确率和召回率均较低。这可能归因于HPLC数据在特征表达上存在局限性,或者CNN模型对HPLC数据的适应性不如太赫兹光谱数据。在HPLC色谱图中,各个皂苷成分仅表现为一个特征峰,且峰高变化不大、保留时间相对稳定,导致HPLC提供的特征信息较为单一。相比之下,太赫兹光谱中各个皂苷成分在多个频率点上显示出特征吸收峰,且这些峰的高度和宽度差异显著。太赫兹光谱数据因此更为复杂,包含了更丰富的多维度特征信息。CNN模型从太赫兹这种更复杂的数据中学习并提取有效特征,这可能增加了模型的学习难度,进而影响了分类结果。并且,在本实验中,太赫兹检测所需时间比HPLC的耗时更短(单次测试太赫兹需3 min,HPLC需70 min)。这表明,在此应用场景下,太赫兹光谱技术可能因其无损、快速、准确的优势而更适合作为分类工具。

3　结论

在本研究中,我们创新地提出了一种结合太赫兹光谱技术与CNN算法的三七产地鉴别方法。通过利用太赫兹光谱技术和HPLC技术,对来自云南红河、昆明、曲靖和文山4个主要产地的40个三七样本进行了太赫兹光谱和液相色谱分析。在获取数据的基础上,我们构建并训练了一个CNN模型,旨在实现对三七样本产地的精确分类。研究结果表明,太赫兹光谱技术在三七产地鉴别中展现出了卓越的性能,其准确率高达92.5%,明显优于HPLC技术的准确率(82.5%)。这一显著的差异凸显了太赫兹光谱技术在捕捉与产地相关特征方面的潜在优势。太赫兹光谱能够提供关于分子振动和旋转的丰富信息,这些信息对于识别三七中特定化学成分的细微差异至关重要。本研究实现了三七药材产地的快速、无损、准确检测,为中草药的分类鉴定提供了一种新的科学工具,有望在中草药的品质控制、真伪鉴别以及产地追溯中发挥重要作用。

References

[1] Park H J, Kim D H, Park S J, et al. Ginseng in Traditional Herbal Prescriptions [J]. Journal of Ginseng Research, 2012, 36(3): 225-241.

[2] Liao P Y, Wang D, Zhang Y J, Yang C R. Dammarane-type glycosides from steamed notoginseng [J]. Journal of Agricultural and Food Chemistry, 2008, 56(5): 1751-1756.

[3] Qiu L, Jiao Y, Huang G K, et al. New Dammarane-Type Saponins from the Roots of *Panax notoginseng* [J]. Helvetica Chimica Acta, 2014, 97(1): 102-111.

[4] Wan J B, Zhang Q W, Hong S J, et al. 5,6-Didehydroginsenosides from the Roots of *Panax notoginseng* [J]. Mole-

cules, 2010, 15(11): 8169-8176.

[5] Yoshikawa M, Morikawa T, Kashima Y, et al. Structures of new dammarane-type triterpene saponins from the flower buds of *Panax notoginseng* and hepatoprotective effects of principal ginseng saponins [J]. Journal of Natural Products, 2003, 66(7): 922-927.

[6] Wang T, Guo R, Zhou G H, et al. Traditional uses, botany, phytochemistry, pharmacology and toxicology of *Panax notoginseng* (Burk.) FH Chen: A review [J]. Journal of Ethnopharmacology, 2016, 188: 234-258.

[7] Wang P P, Zhang L, Yao J, et al. An arabinogalactan from flowers of *Panax notoginseng* inhibits angiogenesis by BMP2/Smad/Id1 signaling [J]. Carbohydrate Polymers, 2015, 121: 328-335.

[8] Zhu J T T, Leung W K W, Cheung J K H, et al. A flavonol glycoside, isolated from roots of *Panax notoginseng*, protects the β-amyloid-induced neurotoxicity in cultured PC12 cells [J]. Neurosignals, 2006, 15(3): 150-150.

[9] Sun B, Xiao J, Sun X B, Wu Y. Notoginsenoside R_1 attenuates cardiac dysfunction in endotoxemic mice: an insight into oestrogen receptor activation and PI3K/Akt signalling [J]. British Journal of Pharmacology, 2013, 168(7): 1758-1770.

[10] Yang C Y, Wang J, Zhao Y, et al. Anti-diabetic effects of *Panax notoginseng* saponins and its major anti-hyperglycemic components [J]. Journal of Ethnopharmacology, 2010, 130(2): 231-236.

[11] Zhang W, Wojta J, Binder B R. Effect of notoginsenoside R_1 on the synthesis of tissue-type plasminogen activator and plasminogen activator inhibitor-1 in cultured human umbilical vein endothelial cells [J]. Arteriosclerosis and thrombosis : a journal of vascular biology, 1994, 14(7): 1040-1046.

[12] Zhao Y, Wang W, Han L, et al. Isolation, structural determination, and evaluation of the biological activity of 20(S)-25-methoxyl-dammarane-3β, 12β, 20-triol 20(S)-25-OCH_3-PPD, a novel natural product from *Panax notoginseng* [J]. Medicinal Chemistry, 2007, 3(1): 51-60.

[13] Zhao Z Z, Liang Z T, Guo P. Macroscopic identification of Chinese medicinal materials: Traditional experiences and modern understanding [J]. Journal of Ethnopharmacology, 2011, 134(3): 556-564.

[14] Chen H, Tan C, Lin Z. Identification of ginseng according to geographical origin by near-infrared spectroscopy and pattern recognition [J]. Vibrational Spectroscopy, 2020, 110: 174-180.

[15] Pisano P L, Silva M F, Olivieri A C. Anthocyanins as markers for the classification of Argentinean wines according to botanical and geographical origin. Chemometric modeling of liquid chromatography-mass spectrometry data [J]. Food Chemistry, 2015, 175: 174-180.

[16] Ji C, Zhang Q, Shi R, et al. Determination of the Authenticity and Origin of Panax Notoginseng: A Review [J]. Journal of Aoac International, 2022, 105(6): 1708-1718.

[17] Sun Y Z, Liu X Y, Fu X J, et al. Discrepancy Study of the Chemical Constituents of Panax Ginseng from Different Growth Environments with UPLC-MS-Based Metabolomics Strategy [J]. Molecules, 2023, 28(7): 2928.

[18] Zhu J Q, Fan X H, Cheng Y Y, et al. Chemometric Analysis for Identification of Botanical Raw Materials for Pharmaceutical Use: A Case Study Using *Panax notoginseng* [J]. Plos One, 2014, 9(1): e87462.

[19] Chen G, Zhang H, Jiang J M, et al. Metabolomics approach to growth-age discrimination in mountain-cultivated ginseng (*Panax ginseng* C. A. Meyer) using ultra-high-performance liquid chromatography coupled with quadrupole-time-of-flight mass spectrometry [J]. Journal of Separation Science, 2023, 46(22): 2300445.

[20] Kim J, Phung H M, Lee S, et al. Anti-skin-aging effects of tissue-cultured mountain-grown ginseng and quantitative HPLC/ELSD analysis of major ginsenosides [J]. Journal of Natural Medicines, 2022, 76(4): 811-820.

[21] Zhang G M, Hu S Y, Chen G, et al. Age identification of the root of Huanren mountain cultivated ginseng and differentiation with cultivated ginseng using terahertz spectroscopy [J]. Journal of Food Composition and Analysis, 2024, 125: 105790.

[22] Liu H B, Zhang X C. Terahertz spectroscopy for explosive, pharmaceutical, and biological sensing applications; proceedings of the *NATO Advanced Research Workshop on Terahretz Frequency Detection and Identification of Materials and Objects*, 2006 [C]. Berlin: Springer, 2007: 251-323.

[23] Liu B W, Peng Y, Hao Y F, et al. Ultra-wideband terahertz fingerprint enhancement sensing and inversion model supported by single-pixel reconfigurable graphene metasurface [J]. Photonix, 2024, 5(1): 10.

[24] Peng Y, Huang J L, Luo J, et al. Three-step one-way model in terahertz biomedical detection [J]. Photonix, 2021, 2(1): 12.

[25] Peng Y, Shi C J, Zhu Y M, et al. Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement [J]. Photonix, 2020, 1(1): 12.

[26] Shao Y N, Wang Y T, Zhu D, et al. Measuring heavy metal ions in water using nature existed microalgae as medium based on terahertz technology [J]. Journal of Hazardous Materials, 2022, 435: 129028.

[27] Shen Y, Yin Y X, Li B, et al. Detection of impurities in wheat using terahertz spectral imaging and convolutional neural networks [J]. Computers and Electronics in Agriculture, 2021, 181: 105931.

[28] Kou T Y, Ye J, Wang J, et al. Terahertz Spectroscopy for Accurate Identification of *Panax quinquefolium* Basing on Nonconjugated 24(R)-Pseudoginsenoside F_{11} [J]. Plant Phenomics, 2021, 2021: 6793457.

[29] Shao Y N, Zhu D, Wang Y T, et al. Moxa Wool in Different Purities and Different Growing Years Measured by Terahertz Spectroscopy [J]. Plant Phenomics, 2022, 2022: 9815143.

[30] Liu Y, Pu H B, Li Q, Sun D W. Discrimination of Pericarpium Citri Reticulatae in different years using Terahertz Time-Domain spectroscopy combined with convolutional neural network [J]. Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy, 2023, 286: 122035.

[31] Pu H B, Yu J X, Sun D W, et al. Distinguishing pericarpium citri reticulatae of different origins using terahertz

time-domain spectroscopy combined with convolutional neural networks [J]. Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy, 2023, 299: 122771.

[32] Abdeljaber O, Avci O, Kiranyaz S, et al. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks [J]. Journal of Sound and Vibration, 2017, 388: 154-170.

[33] Abdeljaber O, Sassi S, Avci O, et al. Fault Detection and Severity Identification of Ball Bearings by Online Condition Monitoring [J]. Ieee Transactions on Industrial Electronics, 2019, 66(10): 8136-8147.

[34] Erdenebayar U, Kim H, Park J-U, et al. Automatic Prediction of Atrial Fibrillation Based on Convolutional Neural Network Using a Short-term Normal Electrocardiogram Signal [J]. Journal of Korean Medical Science, 2019, 34(7): e64.

[35] Han D X, Chen J, Sun J. A parallel spatiotemporal deep learning network for highway traffic flow forecasting [J]. International Journal of Distributed Sensor Networks, 2019, 15(2): 1550147719832792.

[36] Harbola S, Coors V. One dimensional convolutional neural network architectures for wind prediction [J]. Energy Conversion and Management, 2019, 195: 70-75.

[37] Zhang Q X, Zhou D, Zeng X. HeartID: A Multiresolution Convolutional Neural Network for ECG-Based Biometric Human Identification in Smart Health Applications [J]. Ieee Access, 2017, 5: 11805-11816.

[38] Yu Y, Rashidi M, Samali B, et al. Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm [J]. Structural Health Monitoring-an International Journal, 2022, 21(5): 2244-2263.

[39] Zhao X L, Yao J Y, Deng W X, et al. Intelligent Fault Diagnosis of Gearbox Under Variable Working Conditions With Adaptive Intraclass and Interclass Convolutional Neural Network [J]. Ieee Transactions on Neural Networks and Learning Systems, 2023, 34(9): 6339-6353.

[40] Din N M U, Assad A, Dar R A, et al. RiceNet: A deep convolutional neural network approach for classification of rice varieties [J]. Expert Systems with Applications, 2024, 235: 121214.

[41] Dragoman D, Dragoman M. Terahertz fields and applications [J]. Progress in Quantum Electronics, 2004, 28(1): 1-66.

[42] Ariyoshi S, Ohnishi S, Mikami H, et al. Temperature dependent poly (l-lactide) crystallization investigated by Fourier transform terahertz spectroscopy [J]. Materials Advances, 2021, 2(14): 4630-4633.

[43] Banks P A, Kleist E M, Ruggiero M T. Investigating the function and design of molecular materials through terahertz vibrational spectroscopy [J]. Nature Reviews Chemistry, 2023, 7(7): 480-495.

[44] Banks P A, Song Z H, Ruggiero M T. Assessing the Performance of Density Functional Theory Methods on the Prediction of Low-Frequency Vibrational Spectra [J]. Journal of Infrared Millimeter and Terahertz Waves, 2020, 41(11): 1411-1429.

[45] Juliano T R, King M D, Korter T M. Evaluating London Dispersion Force Corrections in Crystalline Nitroguanidine by Terahertz Spectroscopy [J]. Ieee Transactions on Terahertz Science and Technology, 2013, 3(3): 281-287.

[46] King M D, Buchanan W D, Korter T M. Understanding the Terahertz Spectra of Crystalline Pharmaceuticals: Terahertz Spectroscopy and Solid-State Density Functional Theory Study of (*S*)-(+)-Ibuprofen and (*RS*)-Ibuprofen [J]. Journal of Pharmaceutical Sciences, 2011, 100(3): 1116-1129.

[47] Korter T M, Balu R, Campbell M B, et al. Terahertz spectroscopy of solid serine and cysteine [J]. Chemical Physics Letters, 2006, 418(1-3): 65-70.