数据融合技术在中药研究领域的应用

顾志荣[1]，毛小文[2]，祁　梅[1]，郭　燕[2]，葛　斌[1*]

1. 甘肃省人民医院 药剂科，甘肃 兰州　730000
2. 甘肃中医药大学药学院，甘肃 兰州　730000

摘　要：数据融合技术是对多来源的数据信息进行统筹、集成的技术，以提高决策模型的灵敏度、特异性和准确率。数据融合技术结合多变量模型是研究中药复杂体系的有力工具，目前已应用于中药的基原鉴别、产地溯源与鉴别、质量控制与评价、加工炮制与制剂研究、资源形成研究等诸多领域。融合所纳入的源数据主要是中药的化学物质信息，包括各类色谱、光谱信息，无机元素及有机成分的含量信息，电子鼻、电子眼、电子舌等传感器信息以及代谢组学信息等；所采用的多变量模型如主成分分析、层次聚类分析、偏最小二乘-判别分析、正交偏最小二乘-判别分析、支持向量机、人工神经网络、随机森林、决策树、线性判别分析等。未来，数据融合有望与人工智能（artificial intelligence，AI）相结合，源数据中纳入生物医学数据与组学数据，拓展应用于中药活性物质筛选，预测患者对药物反应、药物相互作用、药物-靶点相互作用，中药新药开发以及栽培种植等更多领域，同时应积极开发集数据融合与多变量建模功能于一体的中药研究相关软件系统。

关键词：数据融合；中药研究；源数据；多变量模型；人工智能

中图分类号：R28；TP391　　　**文献标志码：**A　　　**文章编号：**0253 - 2670(2023)18 - 6159 - 13

DOI: 10.7501/j.issn.0253-2670.2023.18.033

Application of data fusion in field of traditional Chinese medicine research

GU Zhi-rong[1], MAO Xiao-wen[2], QI Mei[1], GUO Yan[2], GE Bin[1]

1. Department of Pharmacy, Gansu People's Hospital, Lanzhou 730000, China
2. School of Pharmacy, Gansu University of Chinese Medicine, Lanzhou 730000, China

Abstract: Data fusion is a technology that coordinates and integrates data information from multiple sources to improve the sensitivity, specificity and accuracy of decision models. Data fusion combined with multivariable model is a powerful tool to study the complex system of traditional Chinese medicine (TCM), which has been applied to many research fields such as species identification, origin tracing and identification, quality control and evaluation, processing and preparation, and resources formation research. The source data included in the fusion is mainly chemical substance information of TCM, including various types of chromatographic and spectral information, content information of inorganic elements and organic components, sensor information such as e-nose, e-eye and e-tongue, and metabolomics information, etc. The multivariate models used are principal component analysis (PCA), hierarchical cluster analysis (HCA), partial least squares-discriminant analysis (PLS-DA), orthogonal partial least squares-discriminant analysis (OPLS-DA), support vector machine (SVM), artificial neural network (ANN), random forests (RF), decision trees, linear discriminant analysis (LDA), etc. In the future, data fusion is expected to be combined with artificial intelligence (AI), incorporate biomedical and omics data into source data, and expand applications in more fields such as screening of active substances in TCM, prediction of patient's response to drug, drug-drug interactions, drug-target interactions, development of new TCM, and cultivation of TCM, etc. At the same time, software systems related to TCM research that integrate data fusion and multivariate modeling function should be actively developed.

Key words: data fusion; traditional Chinese medicine research; source data; multivariate model; artificial intelligence

中药是典型的复杂化学体系，同时也是复杂生物信息学体系。中药研究包括了资源学、鉴定学、成分分析、质量评价、加工炮制、物质基础、作用机制、新药研发等诸多方面，不同的研究内容、研

收稿日期：2023-05-10

基金项目：甘肃省青年科技基金计划项目（21JR7RA634）；甘肃省自然科学基金资助项目（20JR5RA154）

作者简介：顾志荣（1988—），男，硕士研究生，主管中药师，研究方向为中药药效物质基础及作用机制。E-mail: guzr8817@163.com

*通信作者：葛　斌（1965—），男，教授，主任药师，硕士生导师，从事中药药效物质基础及作用机制研究。E-mail: gjy0630@163.com

究方法与研究技术会产生不同的数据信息，同一研究内容的不同研究手段（如不同仪器设备）也会产生不同的数据信息，这些单一来源的数据信息往往只能反映中药内在信息的一个或少数几个方面，导致分类、鉴别、识别、评价、决策等的不准确性，而要得到全面、准确、可靠的研究结果则需要统筹尽可能多来源的数据进行综合分析以得到科学决策[1]。

数据融合（data fusion）技术是对不同来源的数据进行统筹、集成的技术，通过将互补的中药多源信息进行多级别、多角度、多层次的统筹融合，采用适宜算法剔除不相关和冗余的数据（“脏数据”），尽可能多地提取和保留特征性信息，得到“一谱多息”的一维数据，从而实现数据间的互补性与协同性，获得比单一来源数据分析更全面、准确、可靠的研究结果。因此，数据融合技术与复杂中药体系研究与具有较高的契合度，俨然已成为了这方面研究的利器之一。本文首先对数据融合技术进行概述，然后系统归纳和总结该技术在中药研究领域得到较好应用的几个方面，最后展望该技术在中药研究领域可能发挥重要作用的研究思路，以期对未来研究有所启发。

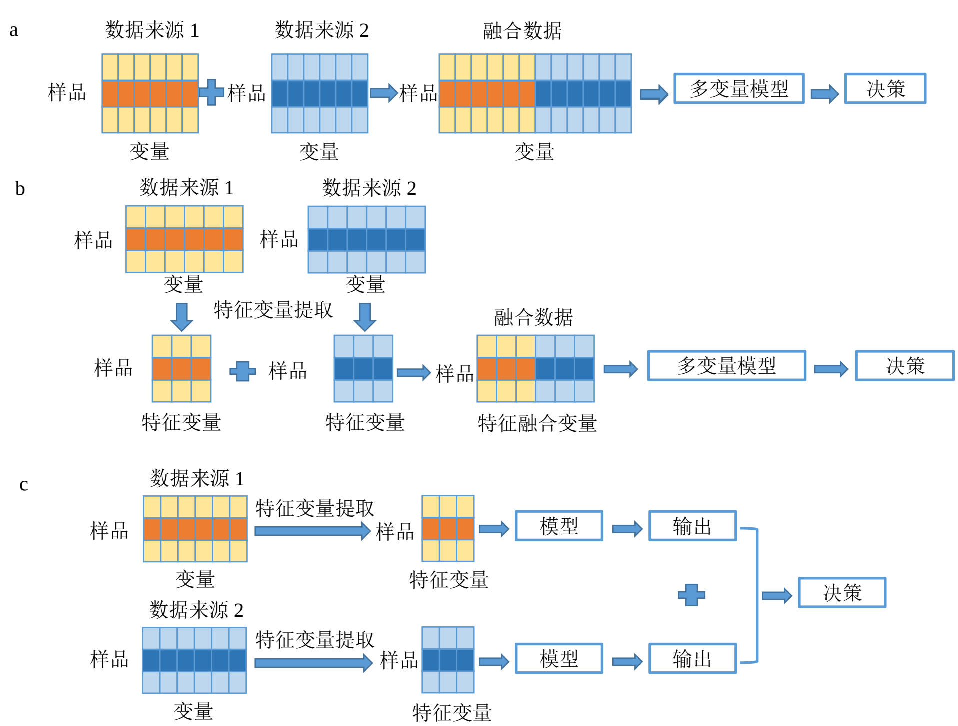
1　数据融合技术概述

1.1　定义

数据融合技术属于信息科学领域范畴，该技术通过数学算法将不同时间、不同空间、不同来源、不同表示的数据信息进行有机融合，提高针对基对象（underlying object）的决策模型的灵敏度（sensitivity，SEN）、特异性（specificity，SPE）和准确率（accuracy，ACC），从而产生单一来源数据无法获得的决策结果，降低决策风险，并得出更全面、准确、可靠的结论[2]。数据融合是新一代智能信息技术的重要基础，目前已在农业[3]、食品[4]、公共卫生（如新型冠状病毒肺炎疫情预测）[5]、医学影像分析[6]、疾病分期[7]、生理状态评估与预测[8]、分子生物学检测[9]、情绪识别[10]、交通运输[11]等领域得到了广泛应用。

1.2　分类

1.2.1　以融合方式分类　数据融合技术按照融合方式可分为低级融合（low-level fusion）、中级融合（mid-level fusion）与高级融合（high-level fusion）[12-15]，融合方式示意图[13]见图 1（以 2 类数据融合为例）。低级融合不需对数据进行任何处理，在多变量建模



a-低级数据融合　b-中级数据融合　c-高级数据融合

a-low-level data fusion　b-mid-level data fusion　c-high-level data fusion

图 1　按融合方式分类的数据融合技术

Fig. 1　Data fusion technology classified by fusion mode

分析前将不同来源的数据集直接连接成单一数据集即可，其局限性在于有用信息的增加同时会造成冗余信息的增加。中级数据融合首先从单个数据集中提取潜在特征变量，然后将特征变量连接成单个数据集以进行多变量建模，该策略能有效剔除冗余信息干扰，提取出关键信息，从而解决“维度诅咒”（curse of dimensionality）（也称“维数灾难”，即数据量随着维度呈指数增长）问题，但其缺点是需为每个数据集开发一个特征变量提取模型。高级数据融合最为复杂，需对不同的来源数据进行特征变量提取，并单独建立多变量模型进行分析，然后对多变量模型分析结果进行统筹分析以做出决策。

1.2.2 以应用方式分类 从应用方式来分类，数据融合可分为物理模型、特征技术推理模型及感知模型 3 类，分类及相应的算法[16-17]如图 2 所示。物理模型将目标对象的实际采样数据与各物理模型或预先存储的目标信息进行匹配，以此实现研究对象的识别和分类，多采用仿真、估计及句法等技术。特征技术推理模型是最常用的数据融合技术，该技术把目标对象的特征数据或统计信息映射到识别空间，然后对识别空间中的数据进行分析处理，据此可分为基于参数和基于信息论的 2 种方法。感知模型通过模拟人类处理问题的过程来实现自动制定策略，常见方法包括逻辑模板、基于知识的专家系统以及模糊集理论。

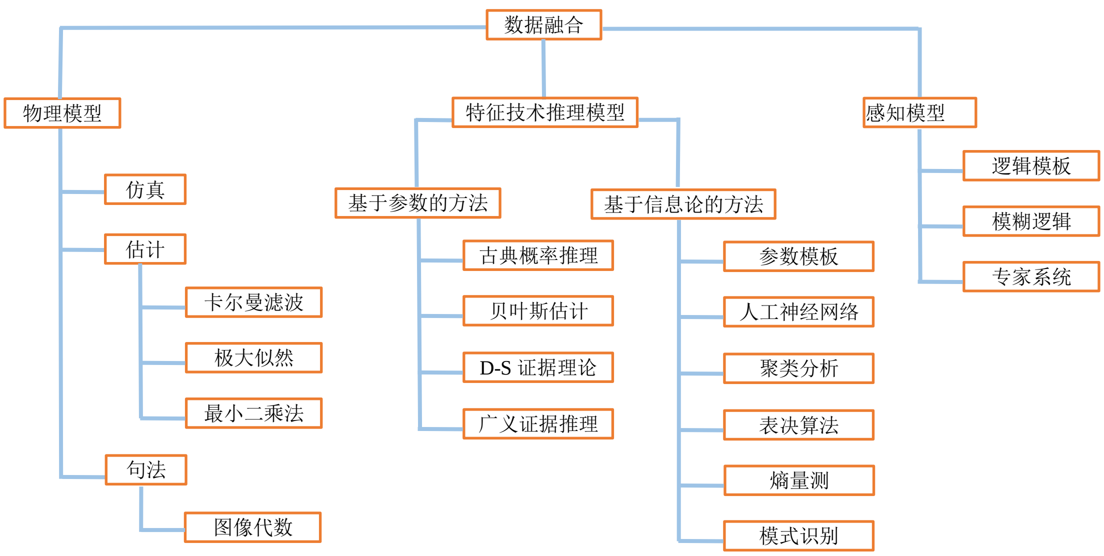


图 2 按应用方式分类的数据融合技术

Fig. 2 Data fusion technology classified by application mode

1.3 算法

算法是数据融合的关键核心，目前可选择的算法较多，常用的有贝叶斯估计、模糊逻辑、卡尔曼（Kalman）滤波、极大似然估计、最小二乘法、聚类分析、人工神经网络等。贝叶斯估计[18]是数据统计意义上的数据融合算法，其理论基础是贝叶斯定理，即在给出证据的条件下，计算假设事件为真的条件概率或后验概率；而古典概率推理是在给定的假设事件下，计算本次量测是来源于某一目标或事件的概率。模糊逻辑[19]基于模糊集合理论，在某一模糊集合中的某一元素的隶属度是对该元素属于该集合的一种度量，但一个元素可能属于多个不同的模糊集合，而模糊逻辑允许在控制语言中包含不精确的状态变量，因此优势在于可减少计算时间。卡尔曼滤波[20]是一种最优化自回归算法，对于解决大部分问题都可以得到最优解，应用方式包括简单卡尔曼滤波器、扩展卡尔曼滤波器、高斯混合滤波器及交互式多模型滤波器等。极大似然估计[21]是利用已知的样本结果信息，反推最大概率导致这些样本结果出现的模型参数值，即先假定其具有某种确定的概率分布，但是其参数未知，然后基于训练样本对概率分布的参数进行估计。最小二乘法[22]是一种通过最小化误差的平方和来寻找数据的最佳函数匹配，简便地求得未知的数据，并使求得的数据与实际数据之间误差的平方和为最小的数学优化方法。聚类分析[23]是一组将研究对象分为相对同质的群组的统计分析技术，包括系统聚类、动态聚类、有序样品聚类、模糊聚类等，常采用 K-均值、K-中心点等算法来实现。人工神经网络[24]是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法模型，该模型依靠系

统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到信息处理的目的。

数据融合方式及算法的选择通常高度依赖于采集的数据结构、数据属性和已开发的应用程序，因此没有一种方法可以通用，并不是高级数据融合一定会优于中级和初级融合，因此在数据融合之前，应针对不同的数据源筛选不同的数据预处理方法及融合算法[7]。

1.4 多变量模型

完成数据融合后，开发适宜的多变量模型用于最终的决策非常重要，常用的多变量模型包括主成分分析（principal component analysis，PCA）、层次聚类分析（hierarchical cluster analysis，HCA）、偏最小二乘-判别分析（partial least squares-discrimination analysis，PLS-DA）、正交偏最小二乘-判别分析（orthogonal partial least squares-discrimination analysis，OPLS-DA）、支持向量机（support vector machine，SVM）、人工神经网络（artificial neural network，ANN）、卷积神经网络（convolutional neural network，CNN）、随机森林（random forest，RF）、K 最近邻算法（K-nearest neighbor algorithm，KNN）、C5.0 决策树（C5.0 decision tree）、机器学习（machine learning，ML）、深度学习（deep learning，DL）、线性判别分析（linear discriminant analysis，LDA）、二次判别分析（quadratic discriminant analysis，QDA）、逻辑回归（logistic regression，LR）等，模型性能多通过 SEN、SPE 和 ACC 来评价[25]。其中，SEN 表示模型正确识别属于目标类别样本的能力；SPE 表示模型拒绝属于其他类别的样本的能力；ACC 也称为效率，是结合 SEN 和 SPE 的度量，表示模型正确识别能力的大小。

SEN＝真阳性/(真阳性＋假阴性)

SPE＝真阴性/(真阴性＋假阳性)

$ACC=(SEN\times SPE)^{1/2}$

2 数据融合技术在中药研究领域的应用

2.1 中药基原鉴别

基原鉴别是对中药的种属来源进行鉴别，以确定其正确的学名进而确定其真伪。非正品来源的药材可能影响临床用药安全，大黄就是典型的例子。Sun 等[26]基于近红外光谱（near infrared spectrum，NIR）和中红外光谱（midlle infrared spectrum，MIR）的数据融合鉴别了正品大黄和非正品大黄，将源数据直接连接完成初级融合，将源数据采用小波压缩（wavelet compression，WC）和区间偏最小二乘法（interval partial least squares，iPLS）来选取特征向量后再连接完成中级融合，建立 PLS-DA、软独立模式分类法（soft independent modeling of class analogy，SIMCA）、SVM、ANN 4 种模型进行鉴别，结果中级融合明显优于初级融合及 NIR、MIR 数据单独建模，4 种模型的 ACC 分别为 97.14%、94.12%、100%及 100%。厚朴为木兰科植物厚朴 *Magnolia officinalis* Rehd. et Wils.或凹叶厚朴 *M. officinalis* Rehd. et Wils. var. *biloba* Rehd.et Wils.的干燥干皮、根皮及枝皮，无法根据外观特征来区分，采用扩增片断长度多态性（amplified fragment length polymorphism，AFLP）分子标记技术鉴别 2 种基原也未成功[27]。为了解决这一难题，Jing 等[28]基于电子鼻（e-nose）、电子舌（e-tongue）和 8 种指标成分（木兰花碱、木兰苷 A、木兰苷 B、和厚朴酚、厚朴酚、辣薄荷基厚朴酚、厚朴碱、β-桉油醇）数据进行数据融合，建立 LDA 和 QDA 模型来鉴别 2 种厚朴基原，结果表明低级融合与 LDA 结合的鉴别能力优于融合前 3 种方法；中级融合与 LDA 结合的鉴别能力（ACC 达到 100%）显著优于与 QDA 结合（ACC 在 96.67%～98.33%），说明电子鼻、电子舌和多指标成分数据的中级融合结合 LDA 是鉴别 2 种厚朴基原的有力工具。Qi 等[29]采用紫外光谱（ultraviolet spectroscopy，UV）和傅里叶变换红外光谱（Fourier transform infrared spectrum，FTIR）的低级数据融合结合 PLS-DA、KNN、SVM 和 RF 4 种模型鉴别不同种类和来源的牛肝菌，结果表明 SVM 的鉴别性能最佳，SEN、SPE 及 ACC 均达到 100%。孙飞等[30]建立了 NIR 和 MIR 中级数据融合结合 PLS-DA 模型来鉴别姜半夏及其伪品姜虎掌南星，该策略明显优于 NIR 和 MIR 单独所建 PLS-DA 模型的鉴别能力，准确率达到 100%。

2.2 中药产地溯源与鉴别

不同产地的中药品质具有显著差异，建立特异、灵敏、准确的中药产地溯源与鉴别方法是目前中药研究领域的热点与难点之一。中药的产地溯源与鉴别方法主要包括基于传统经验、显微特征、指标成分或有效成分差异及色谱、光谱、生物（如 DNA）指纹图谱差异等的鉴别方法，但这些方法的 SEN、SPE 及 ACC 均难以令人满意，目前学术界仍在寻找更佳的中药产地溯源与鉴别方法。数据融合能够同时融合多维中药特征数据，整合各种地理及质量

特征来提高产地溯源的准确性。Qi 等[31]采用中级融合将余甘子的傅里叶变换近红外光谱（Fourier transform-near infrared spectrum，FT-NIR）与 FTIR 光谱进行特征提取，融合成特征数据矩阵，采用调整后的 RF 算法开发地理溯源模型，SEN、SPE 和 ACC 分别为 97.65%、99.85%和 97.63%，预测准确率达到 100%，实现了 17 个余甘子产地的地理来源鉴别及快速、无损、绿色质量检测。张娇等[32]鉴别云南、四川和广西的 9 个产地所产滇黄精，采集衰减全反射-傅里叶变换红外光谱（attenuated total reflection-Fourier transform infrared spectrum，ATR-FTIR）和紫外-可见吸收光谱（ultraviolet-visible absorption spectrum，UV-Vis）数据并分别进行低、中、高级数据融合，建立 RF 模型进行产地鉴别，2 种融合前光谱及低级融合的 RF 模型其 ACC 尚可，但 SEN、SPE 较低；中级融合 RF 模型的 SEN、SPE 均较好，ACC 达到 97.50%；高级融合的 RF 模型鉴别效果较差，并存在过拟合现象；因此中级融合 RF 模型的产地鉴别能力最好。Wu 等[33]对比研究了 FTIR 与 UV-Vis 的低、中、高级数据融合结合网格搜索支持向量机（grid search-support vector machine，GS-SVM）和 RF 模型对 6 个产地野生滇重楼的产地溯源，与中、低级融合结合 GS-SVM 策略相比，高级融合结合 RF 模型的溯源结果更令人满意，SEN、SPE 和 ACC 均超过 90%。可以看出，目前学术界研究较多的是融合相同样品的不同技术来源的数据，而另一个可借鉴的思路是融合不同药用部位的同一技术或多个技术来源的数据，以提高溯源与鉴别能力。如 Shen 等[34]基于滇龙胆根茎、茎、叶 3 个部位的高效液相色谱（high performance liquid chromatography，HPLC）指纹图谱的低、中级数据融合建立 RF 与 OPLS-DA 模型来实现 4 个不同纬度滇龙胆的产地鉴别，结果低级融合结合 OPLS-DA 鉴别效果最好，ACC 高于 99%。数据融合技术用于中药产地溯源与鉴别的其他典型研究见表 1。

总结可知，数据融合结合多变量模型用于中药产地溯源与鉴别具有明显优势，SEN、SPE 和 ACC 均得以大幅提高，相关分析策略是有效和值得推广的。还可以看出，数据融合依赖于源数据的属性，并没有一种融合方法可以通用，并不是高级数据融合一定优于中、初级融合；所建立的多变量模型也会影响决策结果，也没有一种模型可以通用；不同源数据对于融合结果的贡献是有区别的，在融合之前应针对不同数据源采用不同的数据预处理方法。

2.3 中药质量控制与评价

中药质量控制的核心难点在于确定能够科学、合理、全面反映其内在质量的质量标志物（quality marker，Q-Marker），目前常用的 Q-Marker 包括单一或有限的有效成分、指标性成分、有效部位、浸出物等。为了筛选科学、合理、全面的 Q-Marker，目前采用的常用方法有药效成分筛选法、指纹图谱法、谱效关系法、网络药理学法等[51-52]。Liao 等[53]为了确定中成药脑得生（片/胶囊/颗粒）的 Q-Marker，采用数据融合技术将 5 个不同波长（210、238、250、275、322 nm）下的特征多波长超高效液相色谱（ultra-high performance liquid chromatography，UPLC）指纹图谱生成单一的融合 UPLC 指纹图谱，并结合抗氧化活性筛选 Q-Marker，所建模型包括 HCA 和 PCA，最终确定了 11 种成分作为脑得生的 Q-Marker。该研究证明了多波长融合 UPLC 指纹图谱结合生物活性筛选 Q-Marker 的可行性，其优势在于多波长融合 UPLC 指纹图谱所反映的化学成分类型更为全面，不容易造成活性成分的遗漏，且结合生物活性较单纯基于化学成分的筛选方法更为科学及可靠。多年生中药普遍存在，不同生长年限中药的鉴别和质量控制是中药质量评价的重要内容。当前，由于生长激素、肥料等的过量使用，许多多年生中药被人为缩短生长年限，虽然产量提高，但质量与安全性备受全社会质疑。数据融合技术在不同生长年限中药的鉴别与质量控制中具有明显优势，其数据来源的多维性能够使鉴别与评价性能显著提高。Zhang 等[54]基于 ATR-FTIR 和 UV-Vis 的中级融合建立 PLS-DA 与 HCA 模型，准确鉴定了不同生长年限（2～5 年）黄精并进行了质量评价，SEN、SPE 和 ACC 均达到 100%，4 年生黄精的质量最优，该策略为不同生长年限的中药鉴别及质量评价提供了新的思路。数据融合技术在快速测定方面也具有显著优势，如 Wu 等[55]分别采用 ATR-FTIR、RS 及 ATR-IR 与 RS 融合光谱建立了联合区间偏最小二乘法（synergy interval partial least squares，siPLS）、SVM 和 PCA 模型用于快速测定黄酒总抗氧化能力和总酚含量，结果表明基于 RS 的模型性能优于 ATR-IR，融合光谱的 SVM 模型优于基于相同信息的 siPLS 模型和基于 ATR-IR 或 RS 光谱的 siPLS 模型，说明融合光谱可以明显提高测量精度。数据融合技术用于中药质量控制与评价的其他典型研究见表 1。

表 1 数据融合技术在中药研究中的应用

Table 1 Application of data fusion technology in TCM research

类型	药物	研究目的	源数据	融合方式	多变量模型	模型结果	研究结果	文献
中药产地溯源与鉴别	黄芪	5 个省级产区黄芪的产地溯源	可见/短波近红外光谱（380～1030 nm）、NIR（874～1734 nm）成像	中级融合	SVM、LR、CNN	中级融合结合 3 种模型的判别效果均较理想	可准确鉴别，3 种模型的 ACC 均高于 98%	35
	滇龙胆	4 个省级产区 17 个产地的滇龙胆鉴别	FTIR、HPLC	低、中级融合	PLS-DA	低级融合结合 PLS-DA 判别效果最好	可准确鉴别，ACC 为 100%	36
	茯苓	野生和栽培茯苓的产地溯源	FTIR、HPLC	低、中、高级融合	PLS-DA、RF	中级融合结合 PLS-DA 判别效果最好	可准确溯源，ACC 为 95.24%	37
	茯苓	8 个茯苓产地的溯源	FTIR 及 2 种 HPLC（242、210 nm）	低、中级融合	PLS-DA	低级融合结合 PLS-DA 判别效果最好	可准确溯源，ACC 为 100%	25
	牛肝菌	9 个牛肝菌产地的溯源	FTIR、ICP-AES	中级融合	SVM	中级融合结合 SVM 判别效果最好	可准确溯源，ACC 为 90.91%	38
	牛肝菌	11 个牛肝菌产地的溯源	ICP-AES、UV-Vis、FTIR	中级融合	PLS-DA、SVM	PLS-DA、SVM 效果均较好，SVM 略优于 PLS-DA	可准确溯源，ACC 在 90.63%～100%	39
	牛肝菌	7 个牛肝菌产地的溯源	FTIR、UV	中级融合	SVM	中级融合结合 SVM 判别效果最好	可准确溯源，ACC 为 100%	40
	三七	云南 5 个三七产地的溯源	FTIR、NIR	低、中、高级融合	RF	高级融合结合 RF 溯源效果最好	可准确溯源，ACC 在 98%～100%	41
	重楼	云南 5 个野生重楼产地的溯源	FTIR、NIR	低、中、高级融合	PLS-DA、RF	FTIR 比 NIR 贡献更多；高级融合结合 RF 溯源效果最好	可准确溯源，ACC 为 100%	13
	东方草莓（藏药直打萨曾）	5 个东方草莓产地的溯源	HPLC、ICP-MS	中级融合	HCA、PCA、PCA-LDA、C5.0	中级融合结合 C5.0 模型溯源效果最好	可准确溯源，ACC 为 98.9%	42
	蜂蜜	8 个植物来源的蜂蜜鉴别	FTIR、NIR、RS、PTR-TOF-MS、电子鼻	低、中、高级融合	PLS-DA	高级融合结合 PLS-DA 鉴别效果最好	可准确鉴别，ACC 为 99%	43
中药质量控制与评价	黄连	4 种黄连基原的鉴别及 Q-Marker 筛选	FT-NIR、FTIR	低、中级融合	PCA、HCA、PLS	基于变量重要性投影提取的中级融合结合 PLS 效果最佳	可准确鉴别 4 种黄连基原；木兰花碱、格陵兰黄连碱、表小檗碱、小檗碱可作为 Q-Marker	44
	橄榄油	鉴别优质（特级初榨）和不可食用的劣质橄榄油，特定识别主要异味	HS-MS、FTIR、UV-Vis	低、中、高级融合	PLS-DA	中级融合结合 PLS-DA 是鉴别优质和劣质的最佳策略，低级融合结合 PLS-DA 是识别主要异味的最佳策略	可鉴别质量优劣及识别异味（如霉味、酒味、酸败味等）	45
	百合	重金属 Pb、Cu、Zn 的快速检测	LIBS、NIR	低、中级融合	PLS	中级融合结合 PLS 效果最佳	可准确、快速同时测定重金属 Pb、Cu、Zn 含量	46
	含有舒缓类草药的复方制剂	复方制剂（包括纯草药、输液袋、片剂、胶囊和滴剂）中洋甘菊、西番莲、缬草、薰衣草的鉴别	HPLC、UV-Vis	中级融合	PLS-DA	中级融合的 PLS-DA 模型优于单一源数据建模	可准确鉴别复方制剂中的洋甘菊、西番莲、缬草、薰衣草	47
	小柴胡颗粒	外观性状的定性识别及指标性成分的定量预测	电子鼻、电子眼、电子舌传感器信号	中级融合	HCA、PLS-DA	中级融合结合 HCA、PLS-DA 均效果较好	可定性识别气味、颜色、味道，准确预测柴胡皂苷 B_2 和甘草皂苷 G_2 含量	48
	丹参滴丸	评价丹参滴丸的批间稳定性	缩酚酸和皂苷的 HPLC 指纹图谱	中级融合	相似度分析、PCA	融合的二元 HPLC 指纹图谱结合相似度分析和 PCA 效果较好	可实现丹参滴丸质量控制与评价	49
	参麦注射液	6 种皂苷和 4 种糖类的同时快速测定	NIR、UV	低、中级融合	PLS	中级融合优于初级融合，皂苷的分析结果优于糖类；NIR 与 UV 具有信息互补的协同效应	可准确、快速同时测定 6 种人参皂苷（Rg_1、Re、Rf、Rb_1、Rb_2、Rd）和 4 种糖类（*D*-果糖、*D*-葡萄糖、蔗糖、麦芽糖）含量	50

RS-拉曼光谱 PTR-TOF-MS-质子转移反应-飞行时间-质谱 HS-MS-顶空质谱 LIBS-激光诱导击穿光谱

RS-Raman spectrum PTR-TOF-MS-proton transfer reaction-time of flight mass spectrometry HS-MS-headspace mass spectrometry LIBS-laser-induced breakdown spectroscopy

2.4 中药加工炮制研究

加工炮制对中药的有效性、安全性均有重要影响，尤其对于含有毒性、烈性成分或通过炮制改变作用的中药。中药加工炮制的核心在于严格的过程质量控制（in process quality control，IPQC），科学、合理的 IPQC 是产品质量的重要保证。中药是多部位、多成分、多机制的复杂体系，加工炮制中 IPQC 的难点在于同时要兼顾色泽、气味、味道、质地、含量等多维度的质量要求，因此确定科学、合理同时又操作简便的质量控制指标尤为重要。但是现有的分析检测技术往往只能反映 IPQC 的单一维度指标，导致决策的科学性与可靠性欠佳，而数据融合技术可通过将多维度数据融合为一维数据而解决这一问题。Lan 等[56]建立了一种可靠、快速的方法来识别姜黄的颜色、气味等外观特征，筛选加工过程的关键质量指标，采用中级融合将 FT-NIR、电子鼻和色度计数据融合，融合矩阵通过传感器来精确模拟人的识别方式，对于外观特征的识别准确率达 100%；筛选出去甲氧基姜黄素和姜黄素是 IPQC 的关键指标；加工过程中产生的芳香衍生物与样品的颜色、气味等外观特征的变化密切相关，而蛋白质的变化则无显著差异。Dai 等[57]将 NIR 和高分辨质谱的中级融合与代谢组学及 PCA、PLS-DA 模型相结合，阐明了 4 个不同地理来源的麦冬在硫磺熏蒸前后的产地鉴别及质量变化特征，该研究是较早将代谢组学与数据融合相结合用于中药产地加工及鉴别的报道。Sun 等[58]将色度计数据与 NIR 进行中级融合，建立 PLS-DA 模型用于炒栀子的 IPQC，SEN、SPE 及 ACC 均达到 100%，所开发的方法能够准确判别出合格与不合格的的炒栀子，显著优于单独使用色度参数或 NIR 的鉴别模型。Dai 等[59]采用电子鼻和电子舌的中级融合建立 SVM 模型，以同时反映金耳在深层发酵中的气体（电子鼻）和液体（电子舌）特性，以麦角甾醇、还原糖和多糖作为深层发酵的质量控制指标，该策略能够实现对 3 个指标进行同时快速测定，为金耳深层发酵的工艺控制提供了技术支持。

2.5 中药制剂工艺研究

与中药加工炮制的 IPQC 相似，数据融合技术在中药制剂工艺研究中也具有很大优势。吴思俊等[60]引入数据融合技术，提出了一种中药制备工艺变更（酸水渗漉-氯仿萃取替代苯提取）前后质量一致性评价的新策略，将反映工艺整体质量的 NIR 与 RS 光谱分别进行低、中、高级数据融合，利用相似度匹配值和 PCA 中的 Hotelling T^2 和 DModX 统计量对工艺进行表征，结果表明中级融合效果最佳，酸水渗漉-氯仿萃取后所得盐酸青藤碱样品质量和苯提取无显著差异，而新工艺对降低生产成本和环境保护更有优势。湿法制粒是制备中药口服固体制剂的关键工序之一，常用的制粒方式有摇摆式制粒、高剪切湿法制粒（high shear wet granulation，HSWG）和流化床制粒。HSWG 具有黏合剂用量少、全密闭、产出颗粒圆整、流动性好等优点，但其颗粒增长机制及影响因素多而复杂，至今尚未完全阐明，控制其过程是该研究领域的热点之一[61]。Wang 等[62]采用数据融合技术建立了 HSWG 过程的“配方-过程-质量（formulation-process-quality）”模型，将 HSWG 实验数据通过数据融合技术融合成单一数据，生成了 1 个 HSWG 多样性数据配方库，基于融合后数据建立的 PLS 模型能够显著降低 HSWG 过程的预测误差，该研究策略为 HSWG 的科学阐释提供了新思路，并能节省实验时间和成本。中药制剂过程是一个受多因素、多变量影响的 IPQC 系统，具有多个鲜明的关键质量属性（critical quality attributes，CQAs），实时动态监测多 CQAs 对于 IPQC 具有重要意义。光谱技术是目前最常用的过程分析工具，也是中药加工炮制[63]及制剂[64]IPQC 中的常用技术与研究热点，而在这方面光谱融合技术的分析能力明显优于单一光谱技术。Zhang 等[65]通过 NIR 和 FTIR 的光谱融合，实现了实时、准确、快速监测小儿消积止咳口服液提取过程的 7 个 CQAs（辛弗林、槟榔碱、绿原酸、连翘酯苷、柚皮苷、橙皮苷和新橙皮苷），中级融合 PLS 模型优于单光谱 PLS 模型及低级融合 PLS 模型，该方法对同类剂型生产过程的快速监测具有较好参考价值。

2.6 中药资源形成研究

中药活性成分（次生代谢产物）的生物合成及中药资源特性受产地地理环境因子的影响，由此造成不同产地的中药品质及资源特性具有显著差异，即中药具有“道地性”特征。中药资源形成与道地性研究是中药研究领域的持续热点，当前主要从地理环境因子对中药有效成分、指标性成分、化学或生物指纹图谱等的影响展开研究，但所能纳入的化学物质信息有限，而数据融合技术可以解决这一问题。Zhang 等[66]采用 HPLC、GC 和 UPLC-QTOF/MS

中级数据融合建立 PCA、PLS-DA、OPLS-DA 模型研究宿主和环境因素对槲寄生代谢物的影响，发现 HPLC、GC 和 UPLC-QTOF/MS 3 种代谢物谱同时受宿主和环境因素的双重影响，而环境因素起关键作用，环境因素引起的主要差异代谢物为高圣草素、高圣草素-7-*O*-β-*D*-芹菜糖基(1→2)-*O*-β-*D*-葡萄糖苷、高圣草素-7-*O*-β-*D*-葡萄糖苷 3 种成分，研究证实中级融合结合多变量模型是确定宿主和环境因素对代谢物影响的有力策略，为同类研究提供了技术示范。未来，还可将 HPLC、GC、MS 等化学数据与电子鼻、电子眼、电子舌等性状的传感器数据相融合，纳入更多维度以科学、全面地揭示中药品质及资源特性的形成机制。

2.7　小结

近年来，中药研究领域越来越细化，研究的技术手段也得到了长足发展，数据融合技术使得对不同研究领域、不同技术手段、不同分析仪器等所产生的中药研究数据进行集成分析、稳健分析成为了可能，其应用目的是为了提高鉴别、计算或决策的 SEN、SPE 及 ACC。数据融合技术在中药研究领域的分析流程如图 3 所示。目前的应用领域包括中药的基原鉴别、产地溯源与鉴别、质量控制与评价、加工炮制研究、制剂工艺研究、资源形成研究等；融合所纳入的源数据主要是中药的化学物质信息，包括 UV、UV-Vis、MIR、FTIR、NIR、FT-NIR、RS、HPLC、UPLC、MS、PTR-TOF-MS、ICP-AES 等色谱、光谱、波谱信息，无机元素及有机成分的含量信息，电子鼻、电子眼、电子舌等色泽、气味的传感器信息以及代谢组学信息等。但是，数据融合仅改善了数据性能和建模简便性，融合数据仍需要适宜的多变量模型才能完成后续研究，即数据融合与多变量模型是提高 SEN、SPE 及 ACC 缺一不可的 2 个有机组成部分。在其他研究领域广泛使用的多变量模型仍然适用于中药研究领域，如 PCA、HCA、PLS、OPLS、SVM、RF、C5.0、ANN、KNN、CNN、ML、DL、LDA、QDA 等。

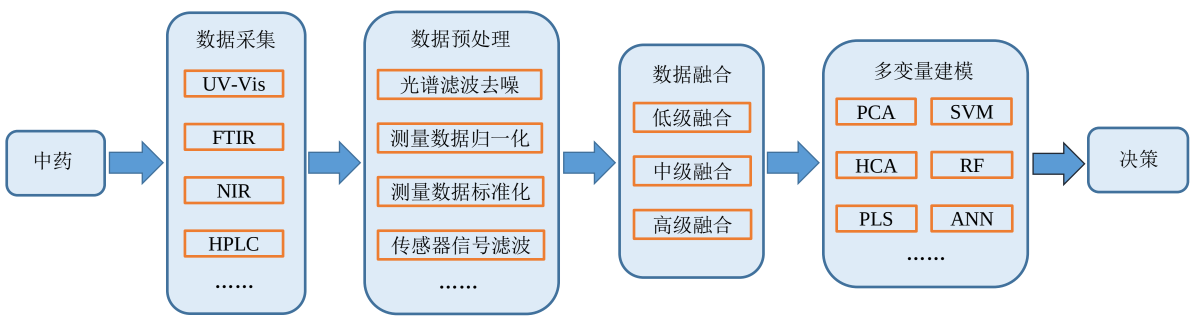


图 3　中药研究中的数据融合分析流程

Fig. 3　Analysis process of data fusion in TCM research

3　前景展望

3.1　数据融合与人工智能（artificial intelligence，AI）相结合用于中药研究

在中药分析中，除了提高 SEN、SPE 及 ACC，智能程度也是所极力追求的。未来，中药相关研究应与 ML、DL 等 AI 相结合，在大幅提高 SEN、SPE、ACC 的同时大幅提高分析的深度、智能程度及应用范围，并实现研究结果的实用化、产品化与产业化，真正满足中药栽培种植及产品研发、生产、质控、市场流通管理等实际需求。极限学习机（extreme learning machine，ELM）是一类基于前馈神经网络（feedforward neural network，FNN）的 ML 方法，目前在计算机视觉、生物信息学、环境科学等方面得到了广泛应用[67]。有报道将 UV-Vis、NIR 数据融合与 ELM 相结合用于农药杀虫剂的快速质量控制，取得了较好的效果[68]。也有报道可通过多特征数据融合结合 DL 技术，能够从微生物组数据中预测人类疾病[69]。

3.2　数据融合纳入中药生物医学数据

目前中药数据融合研究主要纳入的是中药化学物质信息及少量的电子舌、电子眼、电子鼻等外观性状方面的传感器信息，尚未有纳入中药生物医学信息的相关报道。目前，可纳入的中药生物医学信息包括细胞显微图像信息、性状图像信息、生物效应信息、药理药效信息、毒性反应信息等。当然，在数据融合中纳入生物医学信息，融合的难度将随之提升，因此可能需要引入或开发更先进的融合算法。

3.3 数据融合与组学技术相结合用于中药研究

组学技术属于系统生物学范畴，主要包括代谢组学、蛋白组学、基因组学、转录组学等，目前在中药研究领域以代谢组学研究最多，但将数据融合与代谢组学等组学技术相结合的报道较少。中药组学研究具有显著的整体性、系统性特征，这与中药的整体性、系统性不谋而合。以代谢组学为例，中药的代谢轮廓（谱）、代谢指纹体现为数千条可分辨的代谢谱线，代表了数以千计的代谢物，融合这些数据无疑会使中药的内在信息得到更为全面、准确、可靠的展现，从而增强决策的 SEN、SPE、ACC 及智能程度。

3.4 数据融合用于中药活性物质筛选

谱效关系法是重要的中药活性物质筛选方法之一，数据融合技术在中药谱效关系研究中的应用目前暂无相关报道。采用谱效关系筛选活性物质的可靠性与筛选效率在很大程度依赖于所建“谱”（化学物质谱）的质量，如“谱”所包含的成分数量、成分可识别性、成分代表性等。数据融合在中药谱效关系研究中的优势在于，相较融合前的单一技术来源的化学物质谱，融合后的“谱”代表了多层次、多角度、多部位的化学物质，具有更明显的“一谱多息”的功能，可通过一次建模达到从多个维度筛选中药活性物质的目的。另外，受相关研究启发，数据融合技术还可用于从数以万计的中药中筛选具有明确治疗作用的中药及其活性成分，但此类研究的难点在于基于大数据建立中药的“性-效-成分-靶点”多源信息数据库。Liu 等[70]基于多源数据融合和 CNN 模型识别治疗帕金森病的潜在药物，结果令人满意，对中药相关研究具有一定借鉴意义。

3.5 加强数据融合在中药研究领域的产品开发及产业化应用

目前在中药研究领域，数据融合需借助于 MatLab 等进口商业软件，普遍存在版权、编程等问题，限制了研究的便捷性与普及性。目前尚无中药研究专用的数据融合软件系统及相关的仪器设备，已有研究也极少涉及软件系统、仪器设备等的开发，研究成果也未见产业化应用。此外，目前所用的算法也较为传统，距离 AI 等新兴技术尚有较大距离。未来应大量采用多学科交叉技术，加快开发集数据融合与多变量建模功能于一体的中药研究相关软件系统和仪器设备，使得数据融合在中药研究领域实现工具化与智能化。

3.6 数据融合在中药其他研究中的应用

预测患者对中药的反应（patient’s response to drug，PRD）是个性化医疗中的关键问题，其中基于多组学数据融合的 PRD 预测是相关研究的一个热点方向。Yu 等[71]采用多组学与多层网络模型上的低维特征向量融合数据来建立 ML 模型用于预测药物反应，并将这种新方法命名为“基于多组学数据融合的药物反应预测”（drug response prEdiction based on multiOmics data fusion，DREMO）。DREMO 融合了细胞系之间的相似性和药物之间的相似性，首先基于基因表达谱、细胞突变、拷贝数变异（copy number variation，CNV）、药物化学结构和药物靶点等数据构建了与细胞系和药物相关的多层相似性网络，再使用低维特征向量融合多层网络的生物信息，然后应用 ML 模型来预测新的药物-细胞系关联，结果表明 DREMO 显著提高了预测能力。

药物/中药联合治疗中的药物-药物相互作用（drug-drug interaction，DDI）问题也是中药研究中一个具有挑战性的问题。近年来，使用 DL 技术来预测 DDI 的报道较多，但这些方法只考虑药物的单一信息，在稳健性和可扩展性方面存在不足。He 等[72]提出了一种基于图神经网络模型（graph neural network model，GNN）的多类型特征数据融合方法用于 DDI 预测，可有效融合分子图中的拓扑信息、药物之间的相互作用信息以及 SMILES（simplified molecular input line entry system，即简化分子线性输入规范，是一种用 ASCII 字符串明确描述分子结构的规范）序列中药物与局部化学环境之间的相互作用信息，结果表明 GNN 结合数据融合可以准确预测 DDI，稳健性和可扩展性优于其他方法。

数据融合技术结合 ML、DL 等还可用于预测药物/中药-靶点相互作用（drug-target interaction，DTI），对于发现新的中药活性成分及新用法具有借鉴价值。如 Song 等[73]提出了一种基于 DL 的多尺度特征融合方法——深度融合，用于预测 DTI，该方法基于相似性理论与 CNN 生成全局结构相似性特征，并使用变压器网络（transformer network）分别生成药物和蛋白质的局部化学的子结构特征，取得了令人满意的预测效果。

目前，中药的 PRD、DDI、DTI 研究尚处于探索阶段，网络药理学方法作为目前的研究热点，在中药“药效-成分-靶点”预测方面具有一定优势，但对中药 PRD、DDI、DTI 研究则涉及较少且无能为

力，这与其基础数据库内容涉及的多组学信息、中药之间的相互作用信息、中药的人用信息等较少，以及软件算法不涉及数据融合等原因有关，因此积极开发一种集相关基础数据库与数据融合功能于一体的软件系统是极为必要的。

3.7 数据融合在中药新药开发方面的应用

数据融合在中药新药开发中的应用涉及 2 个方面，即工艺开发和质量控制，数据融合在这 2 个方面均具有较好的应用前景，尤其在开发中药新剂型方面，如固体分散体、缓控释制剂、靶向制剂等。中药制剂是多组分复杂化学体系，制剂工艺及成分的物理、化学表征是一大难题，目前多借助各类色谱、光谱技术尽可能做到完整、科学的表征，而数据融合可将多种表征方法所得的数据进行融合，通过融合后的单一数据即可完成表征。目前，数据融合在质量控制中的报道较多，而在工艺研究中的报道较少。Fan 等[49]基于数据融合提出的多有效部位中药制剂质量控制与评价的新思路与新方法，可推广应用于其他制剂。Casian 等[74]开发了一个由 4 种互补仪器组成的过程分析技术（process analytical technology，PAT）工具，用于表征电纺制备的美洛昔康无定形固体分散体，4 种仪器分别提供了 NIR、RS、比色法和图像分析数据，通过中级融合，PAT 工具的互补性显著提高了模型性能，该研究是数据融合在 PAT 领域的首次应用，可有效处理高通量仪器提供的大数据分析，这对中药制剂研究是一个非常好的启示。目前，已有研究者将 PAT 理念引入中药制药研究中，但在实际应用中仍存在许多技术和仪器设备方面的难题。

3.8 数据融合在中药栽培种植研究方面的应用

中药栽培种植中的产量预测对于药农作出肥料、水、储存、现金流等方面的合理决策具有重要的现实意义。传统的产量预测多采用抽样调查法，需进行破坏性抽样，时间、人力成本较大。受相关研究的启发，可采用基于卫星、无人机遥感数据的数据融合并结合多变量建模、ML、DL 等进行中药种植面积与产量的预测。如 Xu 等[3]建立了基于时间序列无人机遥感数据的棉花产量估算模型，结合多光谱图像和棉铃像素覆盖，使用贝叶斯正则化反向传播神经网络来预测棉花产量，结果令人满意。此类研究能够为中药种植面积与产量的预测提供有价值的借鉴。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] Zhang H, Chen R Y, Xu C, *et al*. An integrated approach to discriminate the quality markers of traditional Chinese medicine preparation based on multi-dimensional characteristic network: Shenqi Jiangtang Granule as a case [J]. *J Ethnopharmacol*, 2021, 278: 114277.

[2] Borràs E, Ferré J, Boqué R, *et al*. Data fusion methodologies for food and beverage authentication and quality assessment-A review [J]. *Anal Chim Acta*, 2015, 891: 1-14.

[3] Xu W C, Chen P C, Zhan Y L, *et al*. Cotton yield estimation model based on machine learning using time series UAV remote sensing data [J]. *Int J Appl Earth Obs Geoinf*, 2021, 104: 102511.

[4] Jiang S, Ni C D, Chen G L, *et al*. A novel data fusion strategy based on multiple intelligent sensory technologies and its application in the quality evaluation of Jinhua dry-cured hams [J]. *Sens Actuat B*, 2021, 344: 130324.

[5] Aljaaf A J, Mohsin T M, Al-Jumeily D, *et al*. A fusion of data science and feed-forward neural network-based modelling of COVID-19 outbreak forecasting in IRAQ [J]. *J Biomed Inform*, 2021, 118: 103766.

[6] DeRamus T P, Wu L, Qi S, *et al*. Multimodal data fusion of cortical-subcortical morphology and functional network connectivity in psychotic spectrum disorder [J]. *Neuroimage Clin*, 2022, 35: 103056.

[7] Zheng H, Cai A M, Zhou Q, *et al*. Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment [J]. *Anal Chim Acta*, 2017, 991: 68-75.

[8] Li D Z, Gao W. Physiological state assessment and prediction based on multi-sensor fusion in body area network [J]. *Biomed Signal Process Contr*, 2021, 65: 102340.

[9] Liu Q, He D, Xie L. Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas System using attention boosted deep learning and network-based gene feature [J]. *PLoS Comput Biol*, 2019, 15(10): e1007480.

[10] Maithri M, Raghavendra U, Gudigar A, *et al*. Automated emotion recognition: Current trends and future perspectives [J]. *Comput Methods Programs Biomed*, 2022, 215: 106646.

[11] Ounoughi C, Ben Yahia S. Data fusion for ITS: A systematic literature review [J]. *Inf Fusion*, 2023, 89: 267-291.

[12] Robert C, Jessep W, Sutton J J, *et al*. Evaluating low- mid- and high-level fusion strategies for combining Raman and infrared spectroscopy for quality assessment of red meat

[J]. *Food Chem*, 2021, 361: 130154.

[13] Pei Y F, Zuo Z T, Zhang Q Z, *et al*. Data fusion of Fourier transform mid-infrared (MIR) and near-infrared (NIR) spectroscopies to identify geographical origin of wild *Paris polyphylla* var. *yunnanensis* [J]. *Molecules*, 2019, 24(14): 2559.

[14] Stavropoulos G, van Vorstenbosch R, Jonkers D M A E, *et al*. Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation [J]. *Anal Chim Acta*, 2021, 1183: 339001.

[15] Li Y, Xiong Y M, Min S G. Data fusion strategy in quantitative analysis of spectroscopy relevant to olive oil adulteration [J]. *Vib Spectrosc*, 2019, 101: 20-27.

[16] Lawrence A. Klein 著. 戴亚平译. 多传感器数据融合理论及应用 [M]. 北京: 北京理工大学出版社, 2004: 115-130.

[17] 黄小红. 传感器网络数据融合技术研究及在温室控制中的应用 [D]. 成都: 电子科技大学, 2009.

[18] Ge L Z, Zou K L, Zhou H, *et al*. Three dimensional apple tree organs classification and yield estimation algorithm based on multi-features fusion and support vector machine [J]. *Inf Process Agric*, 2022, 9(3): 431-442.

[19] Zhao G Z, Chen A G, Lu G X, *et al*. Data fusion algorithm based on fuzzy sets and D-S theory of evidence [J]. *Tsinghua Sci Technol*, 2019, 25(1): 12-19.

[20] Zhang J, Liu H B, Sun X J, *et al*. Processing of building subsidence monitoring data based on fusion Kalman filtering algorithm [J]. *Alex Eng J*, 2021, 60(3): 3353-3360.

[21] Song R, Chen X Y, Fang Y C, *et al*. Integrated navigation of GPS/INS based on fusion of recursive maximum likelihood IMM and Square-root Cubature Kalman filter [J]. *ISA Trans*, 2020, 105: 387-395.

[22] Ren M J, Sun L J, Liu M Y, *et al*. A weighted least square based data fusion method for precision measurement of freeform surfaces [J]. *Precis Eng*, 2017, 48: 144-151.

[23] 冯建英, 石岩, 王博, 等. 基于聚类分析的数据挖掘技术及其农业应用研究进展 [J]. 农业机械学报, 2022, 53(S1): 201-212.

[24] Qiu J J, Shi M L, Li S Q, *et al*. Artificial neural network model- and response surface methodology-based optimization of *Atractylodis Macrocephalae Rhizoma* polysaccharide extraction, kinetic modelling and structural characterization [J]. *Ultrason Sonochem*, 2023, 95: 106408.

[25] Wang Q Q, Huang H Y, Wang Y Z. Geographical authentication of *Macrohyporia cocos* by a data fusion method combining ultra-fast liquid chromatography and Fourier transform infrared spectroscopy [J]. *Molecules*, 2019, 24(7): 1320.

[26] Sun W J, Zhang X, Zhang Z Y, *et al*. Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb [J]. *Spectrochim Acta A*, 2017, 171: 72-79.

[27] He J S, Chen L, Si Y, *et al*. Population structure and genetic diversity distribution in wild and cultivated populations of the traditional Chinese medicinal plant *Magnolia officinalis* subsp. *biloba* (Magnoliaceae) [J]. *Genetica*, 2009, 135(2): 233-243.

[28] Jing W G, Zhao X L, Li M H, *et al*. Application of multiple-source data fusion for the discrimination of two botanical origins of *Magnolia officinalis* cortex based on E-nose measurements, E-tongue measurements, and chemical analysis [J]. *Molecules*, 2022, 27(12): 3892.

[29] Qi L M, Li J Q, Liu H G, *et al*. An additional data fusion strategy for the discrimination of porcini mushrooms from different species and origins in combination with four mathematical algorithms [J]. *Food Funct*, 2018, 9(11): 5903-5911.

[30] 孙飞, 陈雨, 王凯洋, 等. 基于红外光谱数据融合的姜半夏鉴别方法研究 [J]. 北京中医药大学学报, 2019, 42(10): 862-868.

[31] Qi L M, Zhong F R, Chen Y, *et al*. An integrated spectroscopic strategy to trace the geographical origins of emblic medicines: Application for the quality assessment of natural medicines [J]. *J Pharm Anal*, 2020, 10(4): 356-364.

[32] 张娇, 王元忠, 杨维泽, 等. ATR-FTIR 和 UV-Vis 结合数据融合策略鉴别滇黄精产地 [J]. 光谱学与光谱分析, 2021, 41(5): 1410-1416.

[33] Wu X M, Zhang Q Z, Wang Y Z. Traceability of wild *Paris polyphylla* Smith var. *yunnanensis* based on data fusion strategy of FT-MIR and UV-Vis combined with SVM and random forest [J]. *Spectrochim Acta A*, 2018, 205: 479-488.

[34] Shen T, Yu H, Wang Y Z. Assessing geographical origin of *Gentiana rigescens* using untargeted chromatographic fingerprint, data fusion and chemometrics [J]. *Molecules*, 2019, 24(14): 2562.

[35] Xiao Q L, Bai X L, Gao P, *et al*. Application of convolutional neural network-based feature extraction and data fusion for geographical origin identification of *Radix Astragali* by visible/short-wave near-infrared and near infrared hyperspectral imaging [J]. *Sensors*, 2020, 20(17): 4940.

[36] 王琴琴, 沈涛, 左智天, 等. 基于数据融合和多指标定量对滇龙胆产地鉴别和质量评价 [J]. 中国中药杂志,

2018, 43(6): 1162-1168.

[37] Wang L, Wang Q Q, Wang Y Z, *et al*. Comparison of geographical traceability of wild and cultivated *Macrohyporia cocos* with different data fusion approaches [J]. *J Anal Meth Chem*, 2021, 2021: 1-13.

[38] Li Y, Zhang J, Li T, *et al*. Geographical traceability of wild *Boletus edulis* based on data fusion of FT-MIR and ICP-AES coupled with data mining methods (SVM) [J]. *Spectrochim Acta A Mol Biomol Spectrosc*, 2017, 177: 20-27.

[39] Qi L M, Liu H G, Li J Q, *et al*. Feature fusion of ICP-AES, UV-vis and FT-MIR for origin traceability of *Boletus edulis* mushrooms in combination with chemometrics [J]. *Sensors*, 2018, 18(1): 241.

[40] Yao S, Li T, Li J Q, *et al*. Geographic identification of *Boletus* mushrooms by data fusion of FT-IR and UV spectroscopies combined with multivariate statistical analysis [J]. *Spectrochim Acta A*, 2018, 198: 257-263.

[41] Li Y, Zhang J Y, Wang Y Z. FT-MIR and NIR spectral data fusion: A synergetic strategy for the geographical traceability of *Panax notoginseng* [J]. *Anal Bioanal Chem*, 2018, 410(1): 91-103.

[42] 张旭超, 党艺航, 付艺萱, 等. 数据融合法结合决策树优化东方草莓全草的地理溯源性研究 [J]. 药物分析杂志, 2022, 42(5): 845-855.

[43] Ballabio D, Robotti E, Grisoni F, *et al*. Chemical profiling and multivariate data fusion methods for the identification of the botanical origin of honey [J]. *Food Chem*, 2018, 266: 79-89.

[44] Qi L M, Ma Y T, Zhong F R, *et al*. Comprehensive quality assessment for *Rhizoma Coptidis* based on quantitative and qualitative metabolic profiles using high performance liquid chromatography, Fourier transform near-infrared and Fourier transform mid-infrared combined with multivariate statistical analysis [J]. *J Pharm Biomed Anal*, 2018, 161: 436-443.

[45] Borràs E, Ferré J, Boqué R, *et al*. Olive oil sensory defects classification with data fusion of instrumental techniques and multivariate analysis (PLS-DA) [J]. *Food Chem*, 2016, 203: 314-322.

[46] Zhao Q, Yu Y, Hao N, *et al*. Data fusion of laser-induced breakdown spectroscopy and near-infrared spectroscopy to quantitatively detect heavy metals in lily [J]. *Microchem J*, 2023, 190: 108670.

[47] Pérez-Ràfols C, Serrano N, Díaz-Cruz J M. Authentication of soothing herbs by UV-vis spectroscopic and chromatographic data fusion strategy [J]. *Chemom Intell Lab Syst*, 2023, 235: 104783.

[48] Zhang X, Wu H W, Lin L N, *et al*. The qualitative and quantitative assessment of xiaochaihu granules based on e-eye, e-nose, e-tongue and chemometrics [J]. *J Pharm Biomed Anal*, 2021, 205: 114298.

[49] Fan X H, Cheng Y Y, Ye Z L, *et al*. Multiple chromatographic fingerprinting and its application to the quality control of herbal medicines [J]. *Anal Chim Acta*, 2006, 555(2): 217-224.

[50] Jiang C, Liu Y, Qu H B. Data fusion strategy based on near infrared spectra and ultraviolet spectra for simultaneous determination of ginsenosides and saccharides in Chinese herbal injection [J]. *Anal Methods*, 2013, 5(17): 4467-4475.

[51] Zhang H B, Zhang Y, Zhang T J, *et al*. Research progress on quality markers of traditional Chinese medicine [J]. *J Pharm Biomed Anal*, 2022, 211: 114588.

[52] 阳长明, 杨平, 刘乐环, 等. 中药质量标志物 (Q-Marker) 研究进展及对中药质量研究的思考 [J]. 中草药, 2021, 52(9): 2519-2526.

[53] Liao J C, Wu Y S, Xu F F, *et al*. Comprehensive evaluation of NAODESHENG by combining UPLC quantitative fingerprint and antioxidant activity [J]. *J Pharm Biomed Anal*, 2021, 193: 113636.

[54] Zhang J, Wang Y Z, Yang M Q, *et al*. Identification and evaluation of *Polygonatum kingianum* with different growth ages based on data fusion strategy [J]. *Microchem J*, 2021, 160: 105662.

[55] Wu Z Z, Xu E B, Long J, *et al*. Comparison between ATR-IR, Raman, concatenated ATR-IR and Raman spectroscopy for the determination of total antioxidant capacity and total phenolic content of Chinese rice wine [J]. *Food Chem*, 2016, 194: 671-679.

[56] Lan Z W, Zhang Y, Sun Y, *et al*. A mid-level data fusion approach for evaluating the internal and external changes determined by FT-NIR, electronic nose and colorimeter in *Curcumae Rhizoma* processing [J]. *J Pharm Biomed Anal*, 2020, 188: 113387.

[57] Dai S Y, Lin Z Z, Xu B, *et al*. Metabolomics data fusion between near infrared spectroscopy and high-resolution mass spectrometry: A synergetic approach to boost performance or induce confusion [J]. *Talanta*, 2018, 189: 641-648.

[58] Sun F, Zhong Y C, Meng J, *et al*. Establishment of an integrated data fusion method between the colorimeter and near-infrared spectroscopy to discriminate the stir-baked *Gardenia jasminoides* Ellis [J]. *Spectrosc Lett*, 2018, 51(10): 547-553.

[59] Dai C X, Huang X Y, Huang D M, *et al*. Detection of

submerged fermentation of *Tremella aurantialba* using data fusion of electronic nose and tongue [J]. *J Food Process Eng*, 2019, 42(3): e13002.

[60] 吴思俊, 王龙, 吴红根, 等. 基于多源数据融合技术的盐酸青藤碱制备工艺一致性评价方法研究 [J]. 分析测试学报, 2022, 41(1): 121-127.

[61] 王政, 王雅雯, 曹君杰, 等. 中药高剪切湿法制粒过程物料可制造性分类研究 [J]. 中国中药杂志, 2021, 46(19): 4969-4977.

[62] Wang Z, Cao J J, Li W T, *et al*. Using a material database and data fusion method to accelerate the process model development of high shear wet granulation [J]. *Sci Rep*, 2021, 11: 16514.

[63] Gao M L, Zhang Y, Cheng F F, *et al*. A gradient-based discriminant analysis method for process quality control of carbonized TCM via Fourier transform near infrared spectroscopy: A case study on carbonized Typhae Pollen [J]. *Spectrochim Acta A*, 2022, 265: 120363.

[64] Zhao J E, Zhou Z M, Zhao F, *et al*. Development and validation of global prediction models for monitoring the manufacturing process of herbal medicine by ultraviolet spectroscopy [J]. *Acupunct Herb Med*, 2022, 2(2): 118-129.

[65] Zhang J, Xu X H, Li L, *et al*. Multi critical quality attributes monitoring of Chinese oral liquid extraction process with a spectral sensor fusion strategy [J]. *Spectrochim Acta A*, 2022, 278: 121317.

[66] Zhang R Z, Zhao J T, Wang W Q, *et al*. Metabolomics-based comparative analysis of the effects of host and environment on *Viscum coloratum* metabolites and antioxidative activities [J]. *J Pharm Anal*, 2022, 12(2): 243-252.

[67] Huang G, Huang G B, Song S J, *et al*. Trends in extreme learning machines: A review [J]. *Neural Netw*, 2015, 61: 32-48.

[68] Li Q Q, Huang Y, Zhang J X, *et al*. A fast determination of insecticide deltamethrin by spectral data fusion of UV-vis and NIR based on extreme learning machine [J]. *Spectrochim Acta A*, 2021, 247: 119119.

[69] Chen X J, Zhu Z F, Zhang W T, *et al*. Human disease prediction from microbiome data by multiple feature fusion and deep learning [J]. *iScience*, 2022, 25(4): 104081.

[70] Liu J E, Peng D D, Li J L, *et al*. Identification of potential Parkinson's disease drugs based on multi-source data fusion and convolutional neural network [J]. *Molecules*, 2022, 27(15): 4780.

[71] Yu L, Zhou D D, Gao L, *et al*. Prediction of drug response in multilayer networks based on fusion of multiomics data [J]. *Methods*, 2021, 192: 85-92.

[72] He C X, Liu Y R, Li H, *et al*. Multi-type feature fusion based on graph neural network for drug-drug interaction prediction [J]. *BMC Bioinformatics*, 2022, 23(1): 224.

[73] Song T, Zhang X D, Ding M, *et al*. DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions [J]. *Methods*, 2022, 204: 269-277.

[74] Casian T, Farkas A, Ilyés K, *et al*. Data fusion strategies for performance improvement of a Process Analytical Technology platform consisting of four instruments: An electrospinning case study [J]. *Int J Pharm*, 2019, 567: 118473.

[责任编辑 潘明佳]