文章编号：1672-8785(2024)09-0044-09

基于太赫兹时域光谱和 PCA-SVM 算法的甜蜜素含量分析

王睿璇[1,2,3] 谭智勇[1,2,3*] 曹俊诚[1,2,3*]

(1. 中国科学院上海微系统与信息技术研究所，上海 200050；
2. 集成电路材料全国重点实验室，上海 200050；
3. 中国科学院大学材料与光电研究中心，北京 100049)

摘 要：光谱分析是研究太赫兹(THz)辐射与物质相互作用的重要手段。采用全光纤式 THz 时域光谱(THz Time-Domain Spectroscopy，THz-TDS)系统测试了不同含量甜蜜素样品的透过率光谱，发现甜蜜素的特征吸收峰位置在 1.4 THz 和 1.7 THz 附近；采用主成分分析结合支持向量机(PCA-SVM)的方法建立了甜蜜素含量回归模型，然后将其预测结果与遗传算法结合偏最小二乘(GA-PLS)模型进行分析比较，并引入决定系数(R^2)和预测均方根误差(*RMSE*)来评价建模效果，对以 10%含量梯度制作的样品集进行检测。研究结果表明，采用 PCA-SVM、SVM 和 GA-PLS 方法建立的预测模型的 *RMSE* 分别为 1.885%、1.926%和 2.432%。因此，PCA-SVM 方法的预测效果最优，且预测数据与实际数据均表现出良好的相关性，获得了效果良好的含量回归预测模型，为甜蜜素含量的检测与分析提供了一种有效手段。

关键词：太赫兹时域光谱；主成分分析；支持向量机；含量回归预测模型

中图分类号：O433 **文献标志码**：A **DOI**：10.3969/j.issn.1672-8785.2024.09.006

Analysis of Saccharin Content Based on Terahertz Time-Domain Spectroscopy and PCA-SVM Algorithm

WANG Rui-xuan[1,2,3], TAN Zhi-yong[1,2,3*], CAO Jun-cheng[1,2,3*]

(*1. Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China;*
2. State Key Laboratory of Materials for Integrated Circuits, Shanghai 200050, China;
3. Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)

收稿日期：2024-03-04

基金项目：国家自然科学基金项目(61927813；61991432)

作者简介：王睿璇(1999-)，女，江苏淮安人，硕士研究生，主要研究方向为太赫兹光谱测量与机器学习算法分析。

*通讯作者：E-mail：zytan@mail.sim.ac.cn(谭智勇)；jccao@mail.sim.ac.cn(曹俊诚)

Abstract：Spectral analysis is an important means of studying the interaction between THz radiation and matter. The transmittance spectra of samples with different levels of saccharin are tested using an all-fiber THz-TDS system, and it is found that the characteristic absorption peaks of saccharin are located around 1.4 THz and 1.7 THz. PCA-SVM method is used to establish the regression model of saccharin content, and the prediction results are analyzed and compared with the GA-PLS model. Correlation coefficients and *RMSE* are introduced to evaluate the modeling effect, and the sample set made with a 10% content gradient is tested. The research results show that the *RMSE* of the prediction models established using PCA-SVM, SVM and GA-PLS methods are 1.885%, 1.926% and 2.432%, respectively. Therefore, the PCA-SVM method has the best prediction performance, and the predicted data show a good correlation with the actual data. A content regression prediction model with good performance is obtained, which provides an effective means for the detection and analysis of saccharin content.

Key words：THz-TDS; principal component analysis; support vector machine; content regression prediction model

0 引言

环己烷氨基磺酸钠($C_6H_{12}NNaO_3S$)又称甜蜜素，是目前使用比较广泛的食品添加剂。在水果罐头、果冻、面包、糕点、果糕、坚果等食品的制作过程中，甜蜜素的添加是必不可少的。在面包、糕点和果糕等食品中，甜蜜素的含量相对较高。已有研究表明，当甜蜜素添加过量时，会损害人的肝脏和神经系统，尤其对于代谢能力相对较弱的孕妇以及老人儿童来说，危害更加明显，甚至可能会导致胎儿畸变或者癌症等病变[1]。在 2017 年 10 月 27 日世界卫生组织国际癌症研究机构所公布的清单中，甜蜜素被列为了第三类致癌物。而在我国《食品安全国家标准食品添加剂使用标准》(GB2760-2011)中，对甜蜜素的添加量有明确准定，但超范围、超限量的使用仍然屡见不鲜，尤其是在饮品、蜜饯、糕点等食品中[2]。

THz 波介于微波和红外光之间，是指频率在 0.1～10 THz、波长在 0.03～3 mm 范围内的电磁辐射。许多物质分子的振动和转动能级对应于这一波段[3]。当 THz 辐射与这些物质相互作用时，相应的物质特征信息会在 THz 光谱中显现出来，尤其在表征碳水化合物、氨基酸、脂肪酸和维生素等方面，THz 光谱尤为有效[4]。除此之外，由于 THz 辐射的能量比大部分生物组织的化学键键能低，所以在对物质进行检测时，相比于其他检测方法，THz 辐射一般不会对所测样品造成电离损伤，可以实现无损检测[5]。THz 检测技术与化学计量法相结合，可以减少大量无关变量，促进数学分析过程，已经被广泛应用于食品工业检测中，比如有害化合物检测、抗生素和微生物检测、水分检测、异物检测、检验和质量控制等[6]。

从化学式来看，甜蜜素属于一种大分子有机物，和葡萄糖等糖类物质类似，在远红外及 THz 频段有显著的特征吸收峰[7]。Wang Y T 等人[8]采用傅里叶光谱仪测量了包括甜蜜素在内的五种人造甜味剂在 1000～1500 cm^{-1} 波数(30～45 THz)范围内的远红外光谱。其中，甜蜜素样品在 1258～1119 cm^{-1} (33.9～38.1 THz)有一个比较宽的吸收，在 1030 cm^{-1} (31.2 THz)左右有一个窄一些的吸收峰。郭祥帅等人[9]采用 THz-TDS 测量了甜蜜素在 0.1～2.5 THz 频段的吸收谱，并通过理论计算模拟了对应的特征吸收峰位置。实验结果显示，甜蜜素在 1.4 THz、1.68 THz、2.1 THz 和 2.45 THz 处均有吸收。

在食品添加剂的实际检测应用中，由于添加剂的含量大小和种类比较复杂，因此仅仅得到添加剂的吸收特征峰是远远不够的，还需要对添加剂的含量进行定量分析。在 THz 光谱分析领域，主成分分析(Principal Component Analysis，PCA)、多元线性回归(Multiple Linear Regression，MLR)、偏最小二乘(Partial

Least Squares，PLS)等都是被广泛应用的机器学习算法[10]。其中，PCA 方法被广泛应用于 THz 光谱的特征提取和降维，可以用较少的自变量来完成对原始变量的最大表示[11]。采用 MLR 可以成功鉴别牛奶中是否含有三聚氰胺[12]。PLS 可以线性化潜在变量。利用区间 PLS、向后间隔 PLS、移动窗口 PLS 等方法都可以避免光谱区域选择的主观性，并已经成功应用于基于 THz 光谱的噻苯咪唑研究[13]。而支持向量机(Support Vector Machine，SVM)则是一种基于结构风险最小化理论的监督学习方法，泛化能力比较强，对于非线性数据有较好的识别能力，在识别抗生素[14]、转基因棉花种子[15]以及区分小麦不同霉变阶段[16]等方面都有很好的应用。

本文首先采用小麦粉作为填充材料，制备了不同质量含量的甜蜜素样品；然后采用 THz-TDS 系统测量不同含量甜蜜素样品的 THz 透过率光谱，接着采用 PCA-SVM 方法建立相应的回归预测模型，并将其与 GA-PLS 模型的预测结果进行分析比较；最后对以 10% 含量梯度制作的样品集进行检测，获得了用不同方法建立模型后的 *RMSE* 与 R^2，为甜蜜素含量的检测提供了一种有效手段。

1 实验测量

1.1 全光纤 THz-TDS 系统

本文用于测量样品 THz 透过率谱的设备为 Menlosystem 公司生产的 Tera K15 型全光纤 THz-TDS 系统。该系统的泵浦激光中心波长为 1560 nm，信噪比高于 70 dB，测量谱宽覆盖 0.1～3.5 THz。整个系统的光路部分放置于密闭空间(见图 1)。其中，用于测量透过率谱的样品架的通光孔径为 4.5 mm。为避免空气中水汽吸收的影响，密闭空间在整个测量过程中维持室温和干燥环境。干燥环境的获得通过干燥氮气吹扫来实现。干燥环境的相对湿度小于 2%，以保证实验数据的准确性。

1.2 样品制备

本文所用甜蜜素纯样由河南蜜丹儿商贸有限公司生产，符合 GB1886.37 生产标准。为了实现不同含量甜蜜素样品的制备，本文采用与食品形态最为接近的小麦粉作为填充剂。小麦粉为益海(泰州)粮油工业有限公司生产的“香满园美味富强小麦粉”。将甜蜜素混合于小麦粉中，通过控制甜蜜素在混合物中的重量比

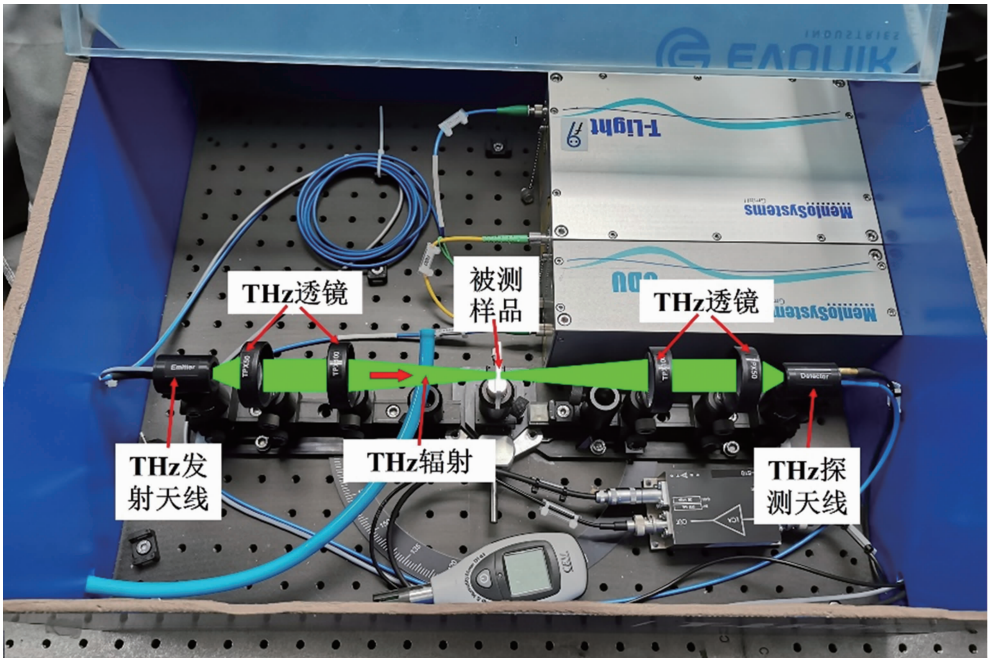

图 1 Tera K15 型 THz-TDS 系统的照片及光路示意图

图 2 制备的样品照片

来实现对甜蜜素含量的控制。具体制备过程如下：首先对甜蜜素进行初步研磨，然后按照不同含量称量甜蜜素与小麦粉，并将其混合均匀（每个样品质量约为 150 mg）。通过压片机用 15 MPa 的压力压制 1.5 min，形成片状样品。该片状样品的厚度约为 500 μm，直径为 13 mm。以 10%的含量差分组制备了 11 组样品，每组共制得至少 8 个样品。

2 预测模型

2.1 数据预处理与模型评价指标

对实验测量取得的 THz 时域光谱信号进行傅里叶变换，得到频域信号。将样品的频域信号除以背景的参考频域信号，得到样品的透过率 T：

$$T(\omega)=\frac{E_{sam}(\omega)}{E_{ref}(\omega)} \tag{1}$$

式中，$E_{sam}(\omega)$是样品的频域信号，$E_{ref}(\omega)$是背景的频域信号。

为了对预测模型进行更好的比较评价，所采用的评价指标主要是 *RMSE*、R^2 和相关系数 r，计算方法如下：

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{2}$$

$$R^2=1-\frac{SSR}{SST}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{3}$$

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

式中，y_i 是模型观测值，$\hat{y}i$ 是模型预测值，$\bar{y}$ 是模型观测值的平均值。*SSR* 代表残差平方和（即模型预测值与观测值的差异平方和），*SST* 代表总平方和（即观测值与平均值之间的差异平方和）。而 x_i、$\bar{x}$ 分别代表对应的输入变量及其均值。*RMSE* 值越小时，说明模型精度越高；R^2 值越接近 1，说明模型的预测能力越强；r 越接近 1，则代表两个变量越相关，即模型的拟合能力越好。

2.2 PCA-SVM

作为一种分类器，SVM 的基本原理是寻找样本空间中使样本分开间隔最大的超平面。该算法可以处理非线性分类问题，计算开销较小，结果易解释。尤其在小样本情况下，SVM 可以解决高维问题，避免陷入局部极小点的问题。在数据获取的时候，制样容易产生误差，且在不同时间段测量，测试环境的波动也会对测得数据结果产生影响，因此选择泛化能力比较强的分类器 SVM 来建立模型。

对于非线性数据集，会加入核函数。一般采用线性核或者高斯核，将其映射到高维空间进行线性回归，从而达到分类的目的。针对多

分类的情况，采用一对多的方法(One-vs-Rest，OVR)，即训练 N 个分类器，将其中一个作为正例(其他作为反例)，选择预测置信度最大的类别标记为最终结果。

考虑到一条光谱数据特征量比较多，可以结合 PCA 方法对数据的协方差矩阵进行特征分解，选取较大特征值所对应的特征向量来表征原始数据，以压缩数据空间，从而将多元数据的特征在低维空间直观表现出来[17]。

2.3 GA-PLS

PLS 是一种统计建模方法，主要应用方面是数据分析和回归分析。与传统的最小二乘回归(Ordinary Least Squares，OLS)相比，该方法可以更好地解决数据多重共线性、维度高和样本量偏少等问题。其基本原理是分解自变量 X 和因变量 Y，通过构建潜在变量来描述两者之间的最大协方差。在保持样本间协方差的同时减少相关性，从而提高模型的效果和解释能力。

GA 用于寻找最优的特征子集，即选择最相关的、最有代表性的特征变量集合。将 PLS 模型的 $RMSE$ 作为适应度函数，通过遗传算法的迭代优化过程对特征子集进行筛选，从而获得使 $RMSE$ 最小的特征变量组合。然后对筛选得到的特征子集应用 PLS 回归分析，建立预测模型，并对模型进行分析评价。

采用 GA-PLS 方法建立模型，结合了 GA 的优化能力和 PLS 的特征提取与预测能力，可以处理高维度和多重共线性数据，并在特征选择和预测中具有良好的性能[18]。

3 实验结果及分析

3.1 THz 光谱分析

采用 THz-TDS 系统测量样品的 THz 光谱图时，由于 THz 辐射对水分比较敏感，为了减少空气中水分对光谱的影响，向样品仓中通氮气，将相对湿度控制在 2%以内。为了减小测量仪器噪声以及信号波动带来的误差，每条光谱数据为 50 次测量的平均值。考虑到样品制备压片过程中存在混合不均匀的情况，随机选择样品并变换位置再次测量。对每种含量组分测得 12 条光谱数据，共得到 11 种样品的 132 组数据。图 3 所示为对每一种甜蜜素含量测得的 12 条光谱数据进行平均后得到的 11 条光谱数据曲线。

可以观察到，作为天然产物的小麦粉是复杂的混合有机物，主要包含碳水化合物、脂肪、纤维和水分等。如图 3 所示，在 THz 波段没有明显的吸收峰，说明小麦粉在 THz 波段没有占主导的吸收物质，是食品类样品光谱研究过程中一种非常好的填充剂。对比甜蜜素的 THz 频谱信号和纯小麦粉的样品信号后发现，在 1.4 THz、1.7 THz 处有明显吸收，信号幅差约为 10 dB。观察吸收谱同样可以发现，在 1.4 THz、1.7 THz 左右有明显的吸收峰，而且随着甜蜜素含量的增加，吸收峰更加明显，吸收系数在 80～130 cm^{-1} 范围内。

为了尽量减少背景信号的干扰，选取有特征峰的一段数据。将甜蜜素样品的频谱数据除以参考的背景信号，可以得到样品的透过率谱(见图 4)。由于小麦粉的透过率较低，随着甜蜜素含量的减少，整体样品的透过率下降，呈一定的线性关系，在 1.4 THz、1.7 THz 处的吸收峰也趋于平稳。甜蜜素含量较高时，变化比较明显；但是含量比较低的时候，各条谱线变化较小。

3.2 模型建立与评价

由图 4 可知，甜蜜素在 1.1～1.8 THz 范围内有明显的特征峰，因此我们把这一段的光谱数据作为训练样本，按照含量进行分类标记。使用 PCA 进行降维，降维后特征数为 55。使用 SVM 进行分类模型建立时，采用 5 折交叉验证和网格法寻优。结果表明，惩罚系数为 390 时，效果最优。其中，核函数采用线性核，gamma 参数选用“scale”。建立回归预测模型时，按照 8∶2 划分训练集和测试集，同样采用网格法寻优，惩罚参数 C 选择为 400，系数 epsilon 选择为 1，效果较优。除此之外，采用 GA-PLS 方法建立模型作为对比。使用 GA 筛选特征子集，其中种群大小设置为 100，迭

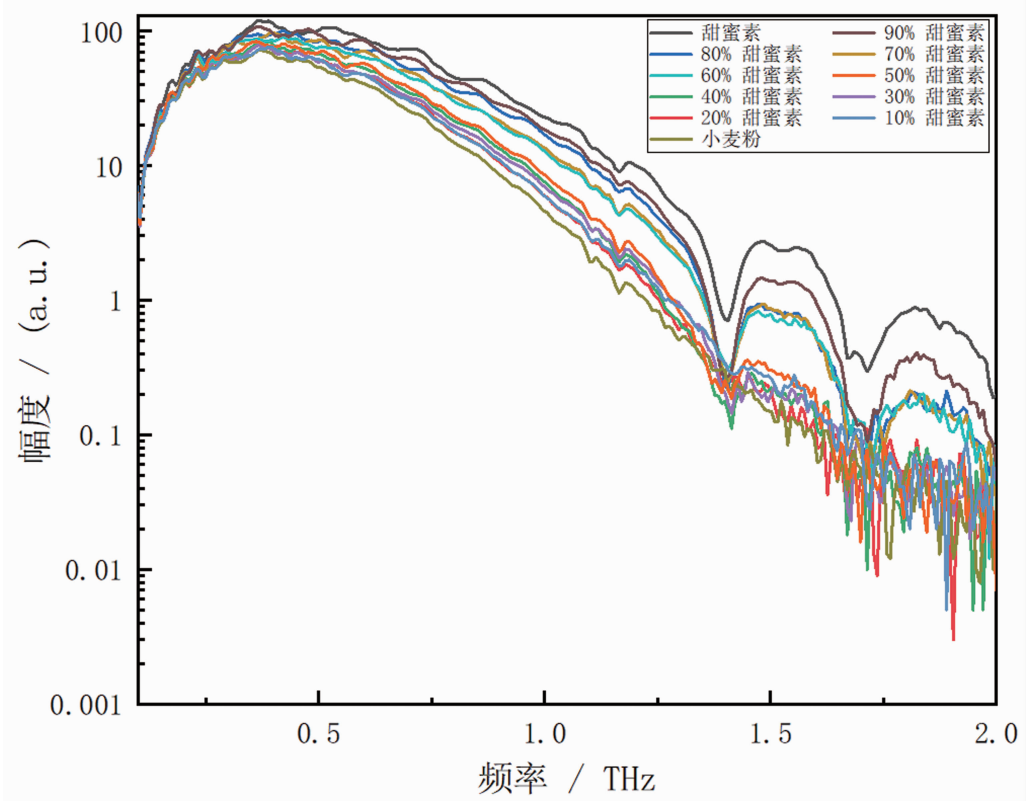

图 3 不同含量甜蜜素样品的 THz 频谱图

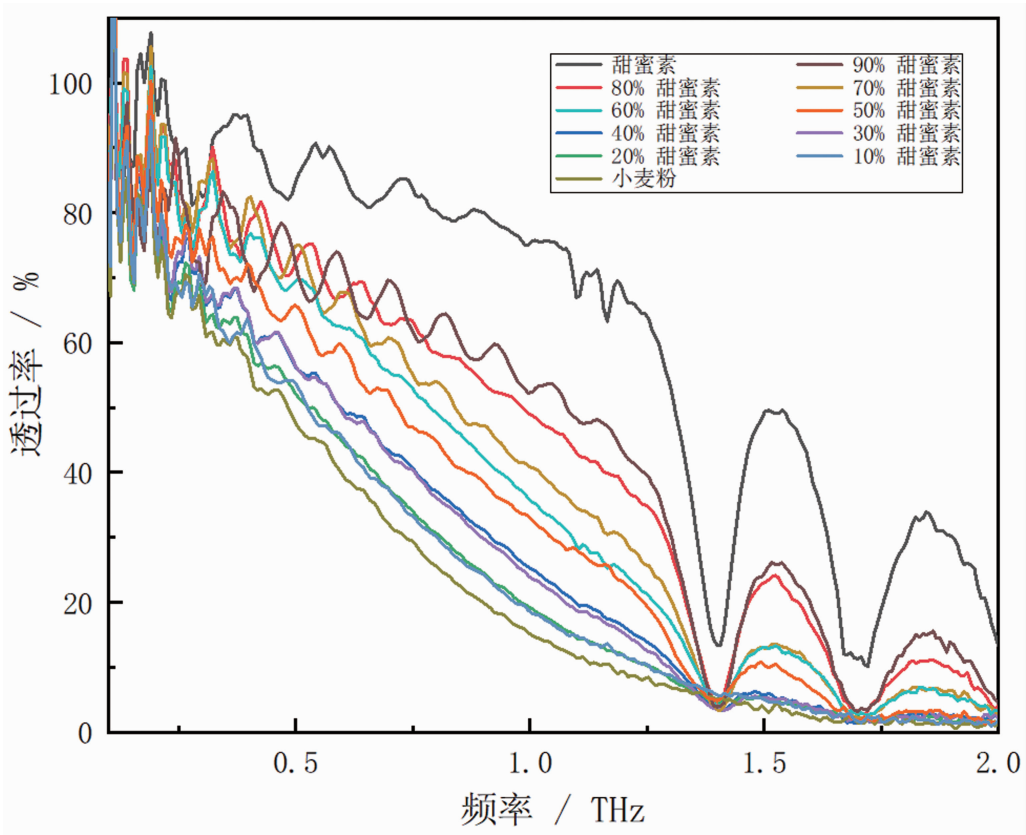

图 4 不同含量甜蜜素样品的透射谱

表 1 预测回归模型评价结果对比

模型方法	*RMSE*(%)	R^2	相关系数
SVM	1.926	0.984	0.992
PCA-SVM	1.885	0.985	0.993
PLS	2.573	0.868	—
GA-PLS	2.432	0.908	—

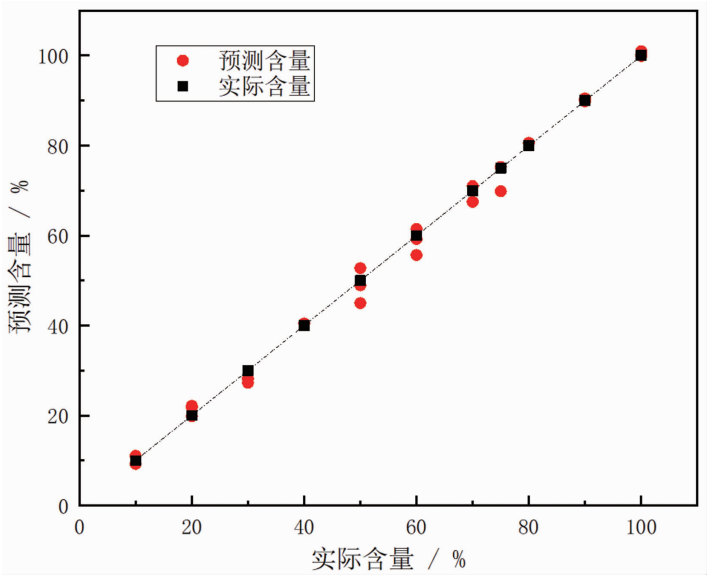

图 5 SVM 模型预测甜蜜素含量与实际含量的对比

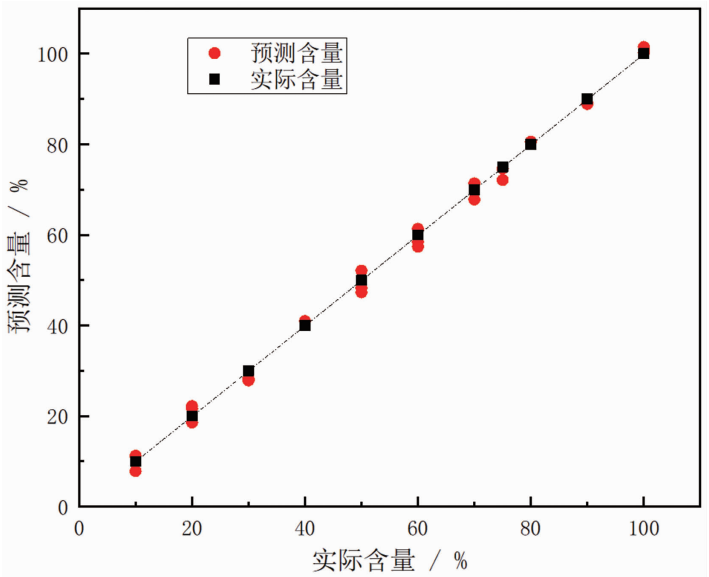

图 6 PCA-SVM 模型预测甜蜜素含量与实际含量的对比

代次数设置为 30；采用两点交叉、按位翻转的方法进行交叉变异操作，而选择操作则采用锦标赛方法。

使用 *RMSE*(%)、R^2 等指标对相关模型的效果进行对比评价(见表 1)。通过对比可知，在使用 PLS 算法时，如果不使用 GA 算法对光谱数据进行降维处理，则会导致数据中有些存在的噪声被捕捉，使得模型在测试集上表现效果不佳。R^2 值小于 0.9，说明模型无法有效地解释因变量的变化，拟合效果较差。由于

THz 光谱数据中存在非线性部分，采用 SVM 方法建立的回归模型比采用 PLS 算法时效果更好，模型的识别精度更高。预测的含量值与实际值相比误差较小，平均在 2%以内。R^2 也都大于 0.98，说明模型拟合效果较好，预测的准确度较高。

使用 SVM 建立模型的具体预测效果如图 5 所示。可以发现，在含量偏低或者偏高时，SVM 模型的预测效果较好，其预测的含量值与实际的甜蜜素含量值相比，误差较小。但是在 50%～75%含量范围内，预测值与实际值偏差较大，模型精度不足。

而采用 PCA 对光谱数据进行预处理，提取出主要的特征数据，然后再使用 SVM 进行预测，有效地解决了在 50%～75%含量处预测误差较大的问题。如图 6 所示，在各个含量梯度，预测的甜蜜素含量与实际的含量值相比，误差都比较小，可以有效地完成甜蜜素不同含量的识别。由于样品制备本身存在误差，精度不足无法制备含量更小的样品。因此，相比于实际应用，模型精度需要进一步提高。

4 结束语

本文针对当前食品加工过程中甜蜜素添加剂量超标的问题，将小麦粉作为填充剂，利用 THz 光谱可敏感检测有机物的特性并采用 THz-TDS 系统进行测量。结果表明，甜蜜素在 1.4 THz、1.7 THz 处均有明显吸收峰。基于 PCA-SVM 算法，选择 1.1～1.8 THz 范围内有明显特征吸收峰的光谱数据，构建出特征矩阵，建立了有关含量的预测回归模型，然后将其预测结果与 GA-PLS 模型进行分析比较，并引入 R^2 和 *RMSE* 对建模效果进行了评价。对 10%含量梯度甜蜜素样品的研究结果表明，采用 PCA-SVM 方法建立的预测模型的 *RMSE* 最小，为 1.885%（相应 R^2 为 0.985）。本文测量的甜蜜素样本量不是很大，有关甜蜜度含量的模型还可以根据实际含量范围作进一步优化。不过，通过本研究可以拓展 THz 光谱检测技术的应用范围，将其融入到 THz 时域光谱检测仪器后，可为食品添加剂的快速无损检测提供一种有效的手段。

参考文献

[1] 马麟莉，程亚萍，李树旺. 食品生产中甜蜜素的使用以及安全现状 [J]. **食品安全导刊**，2022，**29**(6)：1-3.

[2] 杨艳苹. 甜蜜素在食品中的应用及检测探究 [J]. **食品安全导刊**，2022，**29**(29)：112-114.

[3] Sun L，Zhao L，Peng R Y. Research progress in the effects of terahertz waves on biomacromolecules [J]. *Military Medical Research*，2021，**8**(1)：28-35.

[4] Chen T. Terahertz spectra identification of biomolecules based on PCA and fuzzy recognition [J]. *Chinese Journal of Quantum Electronics*，2016，**33**(4)：392-398.

[5] Tao Y H，Fitzgerald A J，Wallace V P. Non-contact，non-destructive testing in various industrial sectors with Terahertz technology [J]. *Sensors*，2020，**20**(3)：712-732.

[6] Chen Q，Jia S，Qin J，et al. A feasible approach to detect pesticides in food samples using THz-FDS and chemometrics [J]. *Journal of Spectroscopy*，2020，**8**(5)：1-10.

[7] Huang H，Shao S，Wang G，et al. Terahertz spectral properties of glucose and two disaccharides in solid and liquid states [J]. *iScience*，2022，**25**(4)：104102-104112.

[8] Wang Y T，Li B，Xu X J，et al. FTIR spectroscopy coupled with machine learning approaches as a rapid tool for identification and quantification of artificial sweeteners [J]. *Food Chemistry*，2020，**303**(1)：125404-125414.

[9] 郭祥帅. 基于太赫兹时域光谱技术的食品添加剂检测与分析 [D]. 青岛：山东科技大学，2020.

[10] Park H，Son J H. Machine learning techniques for THz imaging and time-domain spectroscopy [J]. *Sensors*，2021，**21**(4)：1186-1210.

[11] Yuan B，Wang W，Ye D，et al. Nondestructive evaluation of thermal barrier coatings thickness u-

sing Terahertz technique combined with PCA-GA-ELM algorithm [J]. *Coatings*, 2022, **12**(3): 390-401.

[12] Sun X, Zhu K, Hu J, et al. Nondestructive detection of melamine in milk powder by Terahertz spectroscopy and correlation analysis algorithm [J]. *Journal of Applied Spectroscopy*, 2019, **86**(4): 661-665.

[13] Ma Y, Wang Q, Li L. PLS model investigation of thiabendazole based on THz spectrum [J]. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 2013, **117**(3): 7-14.

[14] Guo J, Deng H, Liu Q, et al. A reliable method for identification of antibiotics by Terahertz spectroscopy and SVM [J]. *Journal of Spectroscopy*, 2020, **8**(10): 1-11.

[15] Qin B, Li Z, Chen T, et al. Identification of genetically modified cotton seeds by terahertz spectroscopy with MPGA-SVM [J]. *Optik*, 2017, **142**(8): 576-582.

[16] Ge H, Jiang Y, Xu Z, et al. Identification of wheat quality using THz spectrum [J]. *Optics Express*, 2014, **22**(10): 12533-12544.

[17] Li B, Zhang D, Shen Y. Study on terahertz spectrum analysis and recognition modeling of common agricultural diseases [J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2020, **243**(12): 1386-1425.

[18] Stefansson P, Liland K H, Thiis T, et al. Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations [J]. *Journal of Chemometrics*, 2020, **34**(3): 3195-3209.