

DOI: 10.13652/j.spjx.1003.5788.2024.80820

# 基于传统机器学习和 LSTM 的近红外光谱玉米中 ZEN 和 DON 定性分析方法

高 曼<sup>1</sup> 钱承敬<sup>2</sup> 丁子元<sup>2</sup> 罗云敬<sup>1</sup> 翟 晨<sup>3</sup>

(1. 北京工业大学化学与生命科学学院, 北京 100124; 2. 中粮营养健康研究院营养健康与食品安全北京市重点实验室, 北京 102209; 3. 中国农业科学院北京畜牧兽医研究所, 北京 100193)

**摘要:** [目的] 建立快速、准确的玉米中真菌毒素检测分析方法。[方法] 基于近红外光谱技术建立玉米中的玉米赤霉烯酮(ZEN)和脱氧雪腐镰刀菌烯醇(DON)含量的定性模型, 使用 144 个自然污染的玉米样品, 采集玉米粉末和毒素提取液两种不同样品制备的近红外光谱, 采用 5 种预处理方法对原始光谱进行处理, 选择两种特征波长筛选方法进一步提取光谱中有效信息, 通过最临近分类算法、最小二乘支持向量机、随机森林 3 种机器学习算法以及深度学习的长短期记忆网络(LSTM)算法建立 ZEN 和 DON 的污染分类模型。[结果] LSTM 算法在玉米粉末光谱数据上的表现优于其他算法, 其中 ZEN 的最佳定性模型测试集分类准确率达 97%, DON 的最佳定性模型测试集分类准确率达到 83%。[结论] LSTM 算法能有效缓解过拟合问题, 显著提升模型分类性能。

**关键词:** 近红外光谱; 玉米; 脱氧雪腐镰刀菌烯醇; 玉米赤霉烯酮; 定性模型; LSTM

## Qualitative analysis methods of ZEN and DON in maize based on near-infrared spectroscopy, traditional machine learning, and LSTM

GAO Man<sup>1</sup> QIAN Chengjing<sup>2</sup> DING Ziyuan<sup>2</sup> LUO Yunjing<sup>1</sup> ZHAI Chen<sup>3</sup>

(1. College of Chemistry and Life Sciences, Beijing University of Technology, Beijing 100124, China;  
2. Beijing Key Laboratory of Nutrition, Health and Food Safety, COFCO Institute of Nutrition and Health,  
Beijing 102209, China; 3. Institute of Animal Science of CAAS, Beijing 100193, China)

**Abstract:** [Objective] To establish a rapid and accurate detection and analysis method for mycotoxins. [Methods] Based on near-infrared spectroscopy, qualitative models for the contents of zearalenone (ZEN) and deoxynivalenol (DON) in maize were established. 144 naturally contaminated maize samples were used, and near-infrared spectra were collected from two sample preparations, namely maize powder and toxin extract. Five pretreatment methods were applied to process the original spectra, and two characteristic wavelength screening methods were selected to further extract effective information from the spectra. Pollution classification models for ZEN and DON were established using three machine learning algorithms, namely k-nearest neighbor classification algorithm, least squares support vector machine, and random forest, as well as the long short-term memory (LSTM) network algorithm of deep learning. [Results] The LSTM algorithm outperforms other algorithms on the spectral data of maize powder. For ZEN, the classification accuracy of the test set of the optimal qualitative model reaches as high as 97%, while that of the optimal qualitative model for DON is 83%. [Conclusion] The LSTM algorithm can effectively alleviate the overfitting problem and significantly improve the classification performance of the model.

**Keywords:** near-infrared spectroscopy; maize; deoxynivalenol; zearalenone; qualitative model; LSTM

基金项目: 国家重点研发计划项目(编号: 2019YFC1605301)

通信作者: 罗云敬(1971—), 女, 北京工业大学教授, 博士。E-mail: luoyj@bjut.edu.cn

翟晨(1988—), 女, 中国农业科学院高级工程师, 博士。E-mail: zhaichen@caas.cn

收稿日期: 2024-08-15 改回日期: 2025-11-01

引用格式: 高曼, 钱承敬, 丁子元, 等. 基于传统机器学习和 LSTM 的近红外光谱玉米中 ZEN 和 DON 定性分析方法[J]. 食品与机械, 2026, 42(4): 59-67.

**Citation:** GAO Man, QIAN Chengjing, DING Ziyuan, et al. Qualitative analysis methods of ZEN and DON in maize based on near-infrared spectroscopy, traditional machine learning, and LSTM[J]. Food & Machinery, 2026, 42(4): 59-67.

玉米作为中国的主要谷物粮食作物之一,在收获、储存、运输过程中易受到真菌毒素的污染,真菌毒素可能引发急性或慢性中毒,甚至具有致癌、致畸和致突变的危害<sup>[1]</sup>。玉米赤霉烯酮(ZEN)<sup>[2]</sup>,具有类雌激素活性,会对神经系统、心脏、生殖系统造成毒害。脱氧雪腐镰刀菌烯醇(DON)<sup>[3]</sup>则会刺激胃黏膜,引发中枢神经的呕吐和拒食反应。2022年的中国玉米副产物及全价料霉菌污染调查报告<sup>[4]</sup>指出,玉米中ZEN和DON的检出率分别为93.24%和97.15%。

ZEN和DON的传统检测方法主要有高效液相色谱法、高效液相色谱串联质谱法等<sup>[5-6]</sup>。Lijalem等<sup>[7]</sup>建立的同位素稀释—液相色谱—串联质谱方法,可以实现玉米中ZEN及其5种主要代谢产物的准确测定。但这些方法试验周期较长、设备昂贵、成本高。因此,开发高效的检测分类真菌毒素技术非常关键。

近年来近红外光谱结合化学计量法已应用于检测谷物中真菌毒素领域<sup>[8-9]</sup>,相对于传统化学方法,近红外检测快速且步骤简单。Kos等<sup>[10]</sup>以500  $\mu\text{g/kg}$ 为阈值建立玉米中DON的决策树分类模型,近红外分类模型准确率达到85%。Elem等<sup>[11]</sup>将近红外光谱数据结合偏最小二乘判别分析应用于大麦分类的探索性分析,根据欧盟阈值区分DON污染水平高低,训练集和检测集准确率高达91%和82%。De Girolamo等<sup>[12]</sup>利用近红外光谱和中红外光谱快速筛查小麦粉中脱氧雪腐镰刀菌烯醇,识别准确率为86%。

长短期记忆网络(LSTM)算法作为人工神经网络的一种特殊类型,结合了传统机器学习算法的稳健性与神经网络算法的灵活性。近年来,在中国LSTM算法常被用于建立食品品种、产地的分类模型,所建模型性能极佳<sup>[13-14]</sup>,然而对于微量物质如真菌毒素的检测与分类应用较少。研究拟将LSTM算法应用于真菌毒素的分类模型建立,利用其强大的序列建模能力来捕捉毒素数据中的复杂特征,以期进一步拓展LSTM算法在食品安全领域中的应用范围。

## 1 试验与方法

### 1.1 材料与仪器

试验样品:2023年收获的144份玉米样品,分别来自山东、河南、吉林、内蒙古、山西、新疆、辽宁、黑龙江、云南、河北、安徽、甘肃、江苏、陕西14个地区;

近红外光谱分析仪:MPA型,德国布鲁克光谱(Bruker Optics)有限公司;

超高效液相色谱仪:LC-30A型,岛津企业管理(中国)有限公司;

质谱仪:QTRAP 5500型,上海爱博才思分析仪器贸易有限公司;

高速冷冻离心机:5810 R型,艾本德(上海)国际贸易有限公司;

涡旋混匀器:VORTEX型,艾卡(广州)仪器设备有限公司;

超声波清洗机:KQ-2500E型,昆山市超声仪器有限公司。

### 1.2 样品前处理和近红外光谱采集

样品采集后密封保存,每份样品混匀后称取500 g进行粉碎,过40目筛,每份玉米粉末样品取出少部分待测。剩余部分再次混匀后称取2 g(精确至0.01 g)样品于50 mL离心管中,加入10 mL乙腈—水—甲酸溶液( $V_{\text{乙腈}}:V_{\text{水}}:V_{\text{甲酸}}=80:18:2$ ),强烈涡旋振荡1 min后,放入超声波清洗机中超声提取20 min,高速离心10 min(8 000 r/min, 4℃),取上清液1 mL至光程为2 mm的比色皿中待测。

玉米粉末使用近红外光谱仪和配备的漫反射附件进行测量,分辨率 $8\text{ cm}^{-1}$ ,样品扫描次数64,累加平均生成最终光谱。玉米液体毒素使用近红外光谱仪的透射模式进行测量,分辨率 $16\text{ cm}^{-1}$ ,样品扫描次数为64,累加平均生成最终光谱。

采用同位素内标—高效液相色谱—串联质谱法测定ZEN、DON含量<sup>[15]</sup>。

### 1.3 模型建立与评估

1.3.1 光谱数据预处理 获取的光谱数据除有效信息外还包含着冗余、干扰信息,需要对光谱进行处理以减少噪声、基线漂移和光源散射的影响<sup>[16]</sup>。采用5种预处理方法消除光谱干扰信息:一阶导数(1D)、二阶导数(2D)、SG平滑(SG)、多元散射校正(MSC)、标准正态变换(SNV)。平滑算法常用于消除光谱噪声,导数算法有效消除基线和背景干扰。SNV对每一条光谱进行标准化处理<sup>[17]</sup>,有效减小由于散射水平不同带来的光谱影响。MSC主要用来减小因颗粒大小不均造成的散射影响<sup>[18]</sup>。

1.3.2 光谱特征波长筛选 光谱经预处理后仍是高维、复杂的数据,需要进一步提取建模信息<sup>[19]</sup>。采用竞争自适应加权采样算法(CARS)和无信息变量消除(UVE)两种算法实现特征波段筛选。CARS通过竞争自组织神经网络对训练数据进行聚类 and 样本分类,根据每个样本所属的类别和重要性动态调整样本的权重,以实现样本平衡和有效性<sup>[20-21]</sup>。研究中CARS算法采用10折交叉验证,重复筛选特征波长100次选择最优筛选结果。UVE算法的建立是基于偏最小二乘算法建立的模型中添加一组白噪声变量,通过信息统计去除对建模贡献率最小的波长变量,最终筛选出特征波长变量<sup>[22]</sup>。

1.3.3 建模方法 采用最邻近分类算法(KNN)、最小二乘支持向量机算法(LS-SVM)及随机森林算法(RF)3种机器学习算法以及LSTM深度学习算法建立分类预测模型。在选择深度学习算法时,对比了LSTM和CNN在相

同条件下的模型性能。基于前期试验的结果,LSTM 表现出明显优势,而 CNN 在此类数据上的表现相对较差。因此,采用 LSTM 算法作为深度学习模型代表。

KNN 分类算法通过距离函数测量不同特征值之间的距离进行分类<sup>[23]</sup>,在决策判断毒素等级时,只参考待分类玉米样本周围的  $K$  个“邻居”样本的所属类别,选择多数样本类别作为该样品的类别,研究中 KNN 自动选择最优  $K$  值为 4。

LS-SVM 算法通过将传统支持向量机中的不等式约束改为等式约束,把解二次方程问题转化为解线性方程组问题<sup>[24]</sup>,提高了建模的速度。

RF 分类算法是一种集成学习算法,属于决策树算法的一种<sup>[25]</sup>。RF 模型基于多棵决策树的集成,其主要特点是每棵决策树都是由随机选择的不同玉米样本和不同特征光谱训练得到,通过取多颗决策树判断值的多数类别为待分类毒素类别,从而提高模型的泛化能力和准确性<sup>[26]</sup>。

LSTM 是一种循环神经网络的变体,克服了循环神经网络处理长序列时的梯度消失和梯度爆炸问题。玉米粉末样本的近红外光谱数据为 972 维向量,玉米液体毒素样本的近红外数据则为 1 167 维的向量,均为长序列数据,因此尝试采用 LSTM 网络模型对其进行分析。搭建的 LSTM 模型架构如图 1 所示。

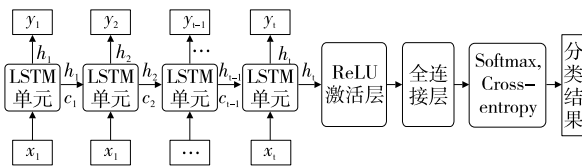


图 1 LSTM 网络模型

Figure 1 LSTM network model

根据时间步长将样本的近红外光谱数据分成若干个向量后按顺序输入 LSTM 单元中,保留最后一个时间步的输出作为这一层的输出  $y$ ,经过 ReLU 激活层,非线性变换使模型能够学习更复杂的数据模式,输入全连接层后,再通过 softmax 函数和 Cross-entropy 损失函数的配合使用,确保模型能够产生正确且可解释的分类概率输出。

1.3.4 模型评估参数 机器学习模型评估从敏感度、特异性、效率、准确度 4 个方面展开<sup>[27]</sup>,深度学习模型从召回率、精准率、调和平均数( $F_1$ -Score)、准确率 4 个方面展开。

召回率也为敏感度,表示模型能够正确识别正例的能力,召回率同样表明在实际为正的样本中被预测为正样本的概率,二者计算式为:

$$R = \frac{T_i}{T_i + F_i}, \quad (1)$$

式中:

$R$ ——召回率(敏感度),%;

$T_i$ ——正确归类为属于“ $i$ ”类样本数,个;

$F_i$ ——属于“安全”类但被模型归类为属于“ $i$ ”的样本数,个。

特异性也称为真负例率,指在实际为负例的样本中,模型正确预测为负例的比例,计算式为:

$$S = \frac{N_i}{N_i + P_i}, \quad (2)$$

式中:

$S$ ——特异性,%;

$N_i$ ——正确归类为不属于“ $i$ ”的样本数,个;

$P_i$ ——被模型归类为属于类“ $i$ ”但属于另一类的样本数,个。

效率为评估模型对每个类别的分类有效性,计算式为:

$$E = \sqrt{R_i + S_i}, \quad (3)$$

式中:

$E$ ——效率,%;

$R_i$ ——敏感度,%;

$S_i$ ——特异性,%。

准确率是正确分类数与总样本数的比值,它衡量整个模型的性能,呈现正确预测的比率,计算式为:

$$A = \frac{T_i + N_i}{T_i + N_i + P_i + F_i}, \quad (4)$$

式中:

$A$ ——准确率,%。

精准率表示预测结果为正类的数据中,有多少数据被正确预测,计算式为:

$$P = \frac{T_i}{T_i + P_i}, \quad (5)$$

式中:

$P$ ——精准率,%。

调和平均数综合评估精确率和召回率,值越大,模型表现越好,计算式为:

$$F_1 = \frac{2 \times R \times P}{R + P}, \quad (6)$$

式中:

$F_1$ ——调和平均数,%。

## 2 结果与讨论

### 2.1 分类阈值与毒素分析

《食品安全国家标准 食品中真菌毒素限量》(GB 2761—2017)对玉米中的 ZEN 最低限量为 60  $\mu\text{g}/\text{kg}$ ,故将 ZEN 分类阈值设置为 60  $\mu\text{g}/\text{kg}$ 。

《食品安全国家标准 食品中真菌毒素限量》(GB 2761—2017)对玉米中的 DON 最低限量为 1 000  $\mu\text{g}/\text{kg}$ 。部分有关玉米样品 DON 含量分析的研究<sup>[10,28]</sup>中,选择

500 和 1 000  $\mu\text{g}/\text{kg}$  作为分级统计阈值。Elem 等<sup>[11]</sup>研究对比了 500 和 1 750  $\mu\text{g}/\text{kg}$  两个分类阈值的建模效果,发现以 500  $\mu\text{g}/\text{kg}$  为分类阈值的分类模型表现更好,推测小阈值导致超标样品增加,模型更具代表性。故将玉米中的 DON 分类阈值设置为 500  $\mu\text{g}/\text{kg}$ ,以减弱样品分布不均匀性,避免模型出现过拟合<sup>[29]</sup>或者泛化能力不足的问题<sup>[30]</sup>。

样品中 ZEN、DON 含量测定值见表 1。样品毒素含量低于阈值限定值的归类为“安全”样品,反之则归类为“危险”样品。

2.2 ZEN 分类模型结果讨论

每份样品采集两条光谱,使用平均后光谱进行数据

建模。图 2 为 144 份样品在 800~2 780 nm 的吸收光谱图。

2.2.1 传统机器学习模型结果 除去异常样品,共 134 份玉米粉末光谱数据及玉米液体毒素光谱数据,按照 3:1 比例将数据划分为 101 个训练集样本和 33 个测试集样本。光谱数据经过 5 种预处理后,基于 CARS 算法和 UVE 算法筛选特征波长,结合 KNN、LS-SVM、RF 3 种方法建立玉米中 ZEN 含量的分类模型,模型结果如图 3 所示。

结果表明,CARS 算法筛选特征波长优于 UVE 算法。对比 CARS 筛选数据组合 5 种预处理算法与 KNN、LS-SVM、RF 算法构建的定性模型发现,1D 和 SG 预处理算法表现较好。CARS-KNN 模型在训练集上的准确率稳定在 72%~84%,但是在测试集上表现不稳定,准确率下降

表 1 样品毒素含量

Table 1 Toxin contents in samples

| 毒素  | 样品数目 | 最低含量/<br>( $\mu\text{g}\cdot\text{kg}^{-1}$ ) | 最高含量/<br>( $\mu\text{g}\cdot\text{kg}^{-1}$ ) | 平均含量/<br>( $\mu\text{g}\cdot\text{kg}^{-1}$ ) | 阈值/<br>( $\mu\text{g}\cdot\text{kg}^{-1}$ ) | 安全样品<br>数量 | 危险样品<br>数量 |
|-----|------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|---------------------------------------------|------------|------------|
| ZEN | 144  | 0                                             | 1 955                                         | 77                                            | 60                                          | 95         | 39         |
| DON | 144  | 0                                             | 5 350                                         | 539                                           | 500                                         | 87         | 57         |

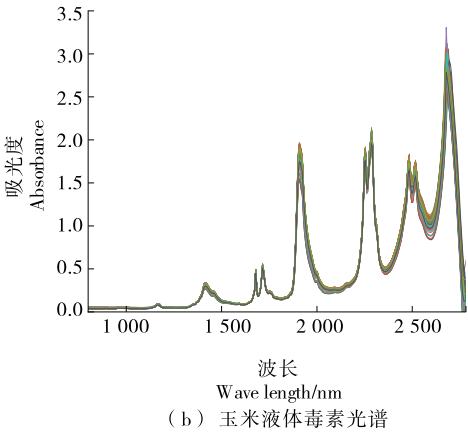
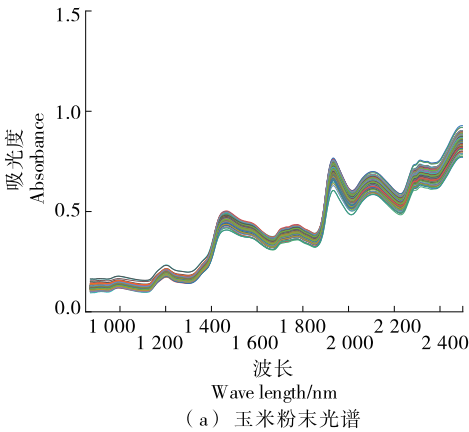


图 2 样品原始光谱图

Figure 2 Original spectrogram of samples

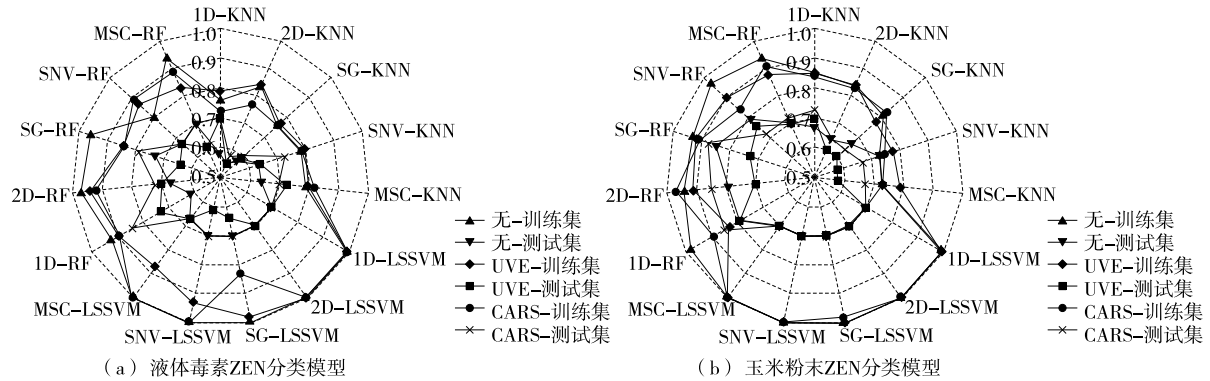


图 3 ZEN 分类模型结果

Figure 3 Results of ZEN classification models

至58%~73%。CARS-LS-SVM模型在训练集上的准确率高达100%,而在测试集上准确率较低且结果固化,显示出显著过拟合现象。CARS-RF模型的表现较好,训练集准确率范围稳定在80%左右,测试集准确率范围也稳定在72%~88%,显示出RF模型结合CARS筛选方法的强大预测能力。最优ZEN分类模型为基于玉米粉末数据的

SG-CARS-RF模型,模型测试集灵敏度、特异性、效率和准确率分别为96%,60%,76%,88%。

最优模型对应的原始平均光谱和特征波段位置如图4所示。玉米粉末特征波段较为分散,分别位于1 250, 1 500, 1 750, 2 100, 2 250 nm附近,液体毒素特征波段则集中于2 000~2 750 nm。

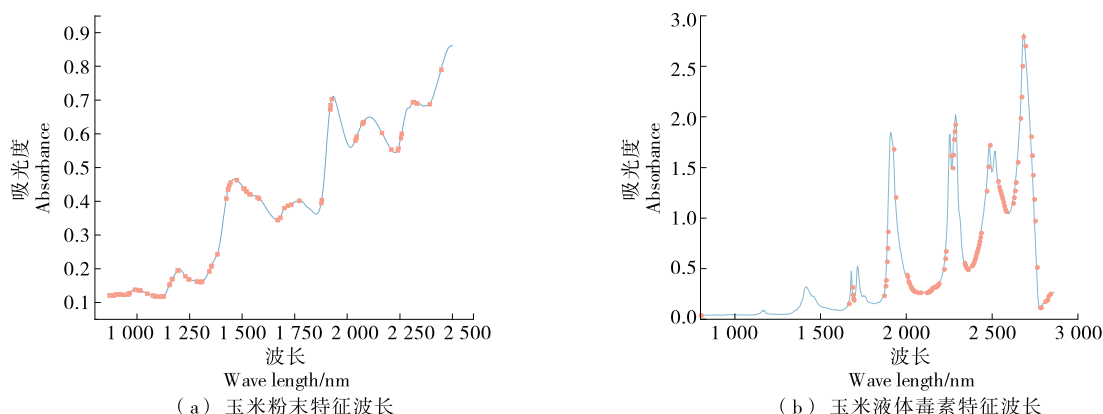


图4 特征波长筛选图

Figure 4 Screening diagrams of characteristic wavelength

**2.2.2 LSTM模型结果** LSTM模型通过其内置的正则化机制展现了在处理序列数据时的优越性。LSTM的训练轮次设定为1 000,学习率为0.01,每经过500轮,模型学习率递减为原来的0.1倍,模型正则化参数设定为0.000 1。按照3:1比例划分为训练集和测试集。ZEN分类模型的测试集准确率结果如图5所示。

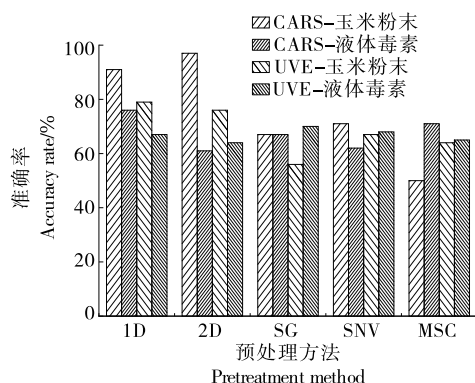


图5 ZEN-LSTM模型结果

Figure 5 Results of ZEN-LSTM model

CARS-LSTM和UVE-LSTM整体模型表现范围一致(50%~80%),说明特征波长筛选算法的选择对LSTM模型基本性能影响较小。而CARS-LSTM和UVE-LSTM结合不同的预处理算法时,模型结果的差异较为显著。具体而言,在结合1D、2D预处理算法时,CARS-LSTM模型的准确率就显著优于UVE-LSTM模型,CARS结合2D预

处理方法在玉米粉末中表现最佳,而UVE结合SG预处理方法在液体毒素中表现最优。结果表明,选择合适的特征波长选择方法和预处理方法对于提高LSTM模型在真菌毒素检测中的准确性至关重要。

**2.2.3 模型对比分析** 基于玉米粉末建立的ZEN分类模型整体性能优于液体毒素的ZEN定性模型。传统机器学习算法中,最优玉米粉末的ZEN分类模型为SG-CARS-RF模型(正确率为89%),而最优液体毒素模型为1D-CARS-RF模型(正确率为84%)。LSTM算法中,最优玉米粉末的ZEN分类模型为2D-CARS-LSTM模型(正确率为97%),显示出显著的分类准确度的提升,最优液体毒素模型为1D-CARS-LSTM模型(正确率为76%),整体模型稳定性较好。最优模型性能参数见表2,混淆矩阵如图6所示。

表2 ZEN分类模型性能参数

Table 2 Performance parameters of ZEN classification models

| 模型类型    | 分组  | 召回率/% | 精准率/% | F <sub>1</sub> -Score/% | 准确率/% |
|---------|-----|-------|-------|-------------------------|-------|
| 2D-CARS | 训练集 | 100   | 100   | 100                     | 100   |
| LSTM    | 测试集 | 100   | 96    | 98                      | 97    |

## 2.3 DON分类模型结果讨论

**2.3.1 传统机器学习模型结果** 将144个样品按照3:1比例随机划分为训练集108个和测试集36个。经过5种预处理后的光谱数据,分别基于全光谱范围、CARS算法和UVE算法筛选特征波长,并结合KNN、LS-SVM和RF 3种

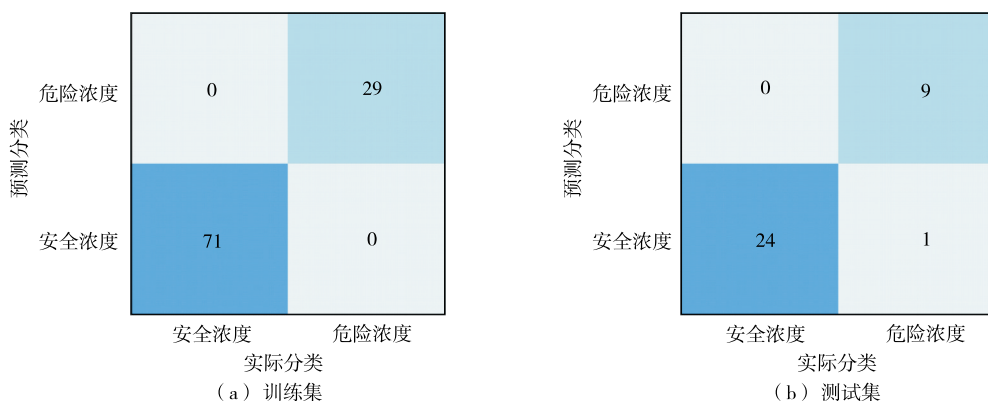


图 6 2D-CARS-LSTM 模型混淆矩阵

Figure 6 Confusion matrices of 2D-CARS-LSTM models

方法建立玉米中 DON 的分类模型。模型结果如图 7 所示。

CARS-KNN 定性模型在训练集上准确率稳定,但未达到 90%,在测试集上性能波动较大,准确率低于 70%。CARS-LS-SVM 模型除 SG 预处理外,训练集准确率均达

到 100%,然而测试集性能不佳,同样展现显著过拟合的现象。CARS-RF 模型的表现较好,在测试集准确率为 81%。总体来看,CARS 和 UVE 在一定程度上能够改善模型的泛化性能,但效果因建模算法的区别而异。

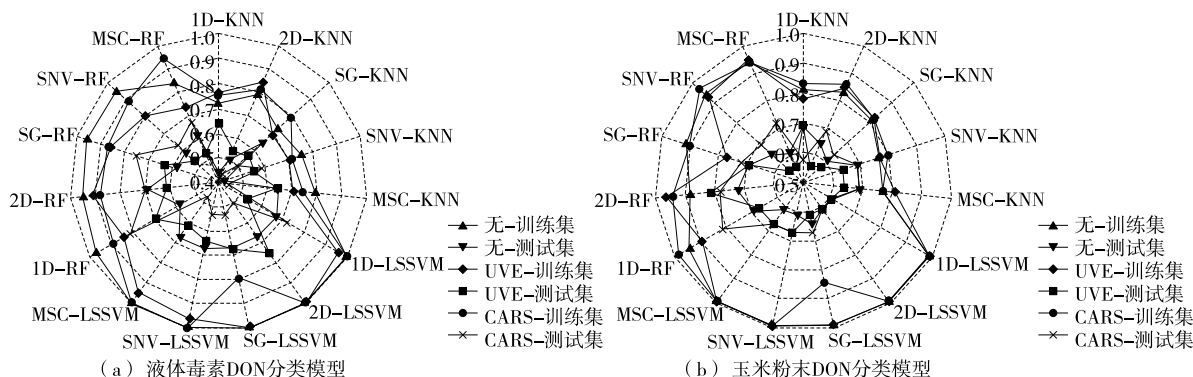


图 7 DON 分类模型结果

Figure 7 Results of DON classification models

最优模型对应的原始平均光谱和特征波段位置分别如图 8 所示。玉米粉末特征波段较为分散,主要位于 1 200,1 500,2 100,2 400 nm 附近,液体毒素特征波段则集中于 2 000~2 250 nm,波段密集。

**2.3.2 LSTM 模型结果** LSTM 参数同 2.1.2。按照 3:1 比例划分为训练集和测试集。DON 分类模型的测试集准确率结果如图 9 所示。

CARS-LSTM 和 UVE-LSTM 整体模型表现范围一致(50%~80%),结果表明,CARS-LSTM 和 UVE-LSTM 模型的整体性能存在一定的重叠。然而,通过选择合适的预处理方法,可以显著提高模型的准确性。在玉米粉末样品中,2D 预处理方法显著提高了 CARS-LSTM 模型的性能,均取得了最高准确率。而在液体毒素中,其优势并不突出。综合来看,强调了特征波长选择方法和预处理方法的协同优化的重要性。

**2.3.3 模型对比分析** 传统机器学习算法中,毒素提取液和玉米粉末的最优模型准确率均达到了 81%,然而,基于玉米粉末的 1D-CARS-RF 模型具有更高的灵敏度(90%)、特异性(64%)和效率(76%)。LSTM 模型中,基于玉米粉末数据的最优模型为 2D-UVE-LSTM(准确率为 83%),基于液体毒素的最优模型为 2D-UVE-LSTM(准确率为 69%)。最优 DON 分类模型性能参数见表 3,混淆矩阵如图 10 所示。

## 2.4 模型结果综合分析

对比玉米粉末和液体毒素两种不同样本类型的模型性能,玉米粉末的模型表现普遍更优。这可能与样本的光学特性和采集方式有关。玉米粉末的漫反射光谱包含了 972 个波数,而液体毒素的透射光谱则有 1 167 个波数,不同的光谱特性可能导致特征波长的分布和信息含量的差异。此外,结合前文的特征波长筛选结果分析,玉米粉

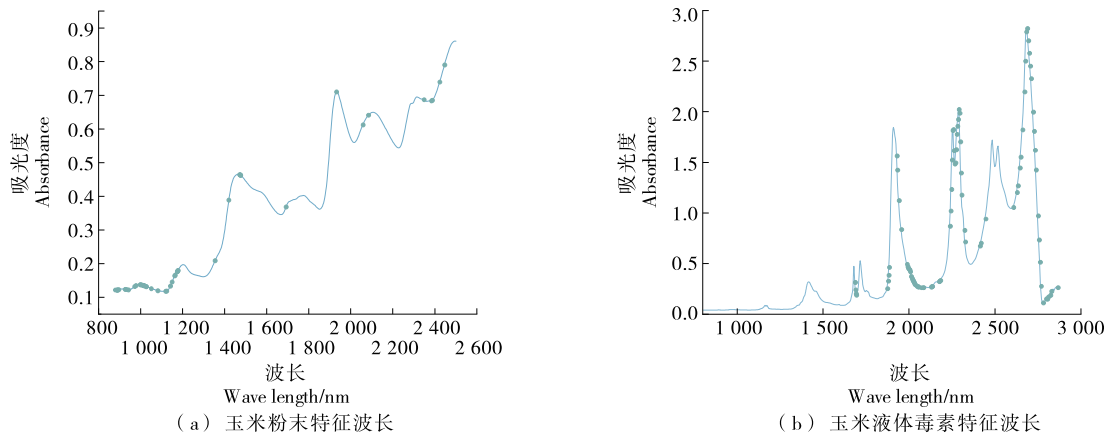


图8 特征波长筛选图

Figure 8 Screening diagram of characteristic wavelength

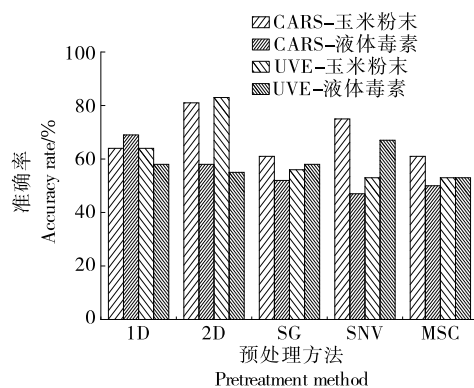


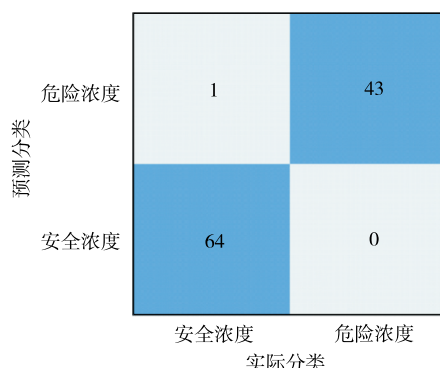
图9 DON-LSTM 模型结果

Figure 9 Results of DON-LSTM model

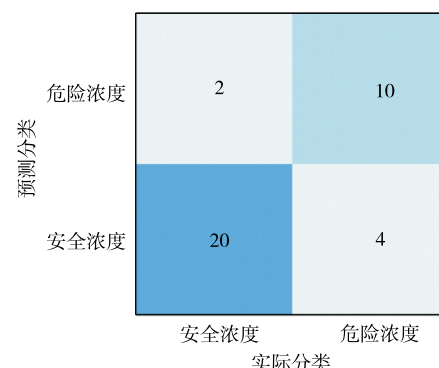
表3 DON 模型性能参数

Table 3 Performance parameters of DON model

| 模型类型   | 分组  | 召回率/% | 精准率/% | F <sub>1</sub> -Score/% | 准确率/% |
|--------|-----|-------|-------|-------------------------|-------|
| 2D-UVE | 训练集 | 98    | 100   | 99                      | 99    |
| LSTM   | 测试集 | 91    | 83    | 87                      | 83    |



(a) 训练集



(b) 测试集

图10 2D-UVE-LSTM 模型混淆矩阵

Figure 10 Confusion matrices of 2D-UVE-LSTM models

末由多种成分组成,其光谱吸收特性可能更为分散,而液体毒素由几种化合物组成,其光谱吸收可能更加集中。这些因素可能导致玉米粉末的数据包含了更丰富和全面的信息,有利于模型的学习和分类。因此,玉米粉末的近红外数据对于模型建立更为有效。

在机器学习算法模型训练中,玉米粉末及玉米液体毒素的模型过拟合现象都较为严重,推测其原因有二:①数据量有限,模型无法充分学习而过多依赖训练集中的特定数据点和噪声。②与算法特性有关,KNN算法处理高维数据时易受到噪声或不相关特征的影响;LS-SVM算法的正则化能力,易捕捉训练数据中的细节和噪声;RF模型复杂度与决策树数量、深度参数等相关,拟合表现较好。因此,优化正则化方法,对于提高模型的泛化能力和防止过拟合具有重要意义。

机器学习和深度学习算法建立在DON和ZEN的分类预测中的性能整体表现相当。但在建立的DON分类预测模型中,最优LSTM模型准确率(2D-UVE)较传统机器学习(1D-CARS)最优模型提高了3%。在建立的ZEN

分类预测模型中,最优 LSTM 模型准确率(2D-CARS)较传统机器学习(1D-CARS)最优模型接近 10%。说明 LSTM 结合适当预处理算法及特征波长筛选算法处理长序列数据,对于模型的性能提升有一定的保障。

### 3 结论

(1) 玉米赤霉烯酮的最优分类模型是基于玉米粉末的二阶导数—竞争自适应加权采样算法—长短期记忆网络模型,该模型在训练集上的分类准确率为 100%,在测试集上的分类准确率达 97%,实现了对玉米赤霉烯酮含量的精准分类。脱氧雪腐镰刀菌烯醇的最优分类模型是基于玉米粉末的二阶导数—无信息变量消除—长短期记忆网络模型,该模型在训练集的准确率为 97%,在测试集的准确率为 83%。

(2) 研究过程中,传统机器学习算法中往往存在过拟合现象。相较之下,长短期记忆网络模型表现更稳定,最优模型准确率提高了约 3%~10%。研究构建的长短期记忆网络分类模型性能显著提升,实现了对复杂毒素数据的精准分类。

(3) 研究建立的模型在内部数据集上表现良好,但其在更广泛样本来源和不同仪器、环境等条件下的泛化能力仍需进一步验证。

### 参考文献

- [1] HAO W, GUAN S, LI A P, et al. Mycotoxin occurrence in feeds and raw materials in China: a five-year investigation[J]. *Toxins (Basel)*, 2023, 15(1): 63.
- [2] JING S Y, LIU C M, ZHENG J, et al. Toxicity of zearalenone and its nutritional intervention by natural products[J]. *Food & Function*, 2022, 13(20): 10 374-10 400.
- [3] HOU S L, MA J J, CHENG Y Q, et al. The toxicity mechanisms of DON to humans and animals and potential biological treatment strategies[J]. *Critical Reviews in Food Science and Nutrition*, 2023, 63(6): 790-812.
- [4] 苏青云, 齐莎日娜, 雷元培, 等. 2022 年中国饲料和原料中霉菌毒素污染调查报告[J]. *饲料工业*, 2024, 45(9): 121-125.  
SU Q Y, QI S R N, LEI Y P, et al. A survey report on the mycotoxin contamination of Chinese raw materials and feed in 2022[J]. *Feed Industry*, 2024, 45(9): 121-125.
- [5] CAGLAYAN M O, ŞAHİN S, ÜSTÜNDAĞ Z. Detection strategies of zearalenone for food safety: a review[J]. *Critical Reviews in Analytical Chemistry*, 2022, 52(2): 294-313.
- [6] SZÉKÁCS A. Mycotoxins as emerging contaminants. Introduction to the special issue "rapid detection of mycotoxin contamination"[J]. *Toxins (Basel)*, 2021, 13(7): 475.
- [7] LIJALEM Y G, GAB-ALLAH M A, CHOI K, et al. Development of isotope dilution-liquid chromatography/tandem mass spectrometry for the accurate determination of zearalenone and its metabolites in corn[J]. *Food Chem*, 2022, 384: 132483.
- [8] TYSKA D, MALLMANN A, GRESSLER L T, et al. Near-infrared spectroscopy as a tool for rapid screening of deoxynivalenol in wheat flour and its applicability in the industry[J]. *Food Additives & Contaminants: Part A*, 2021, 38(11): 1 958-1 968.
- [9] FREITAG S, SULYOK M, LOGAN N, et al. The potential and applicability of infrared spectroscopic methods for the rapid screening and routine analysis of mycotoxins in food crops[J]. *Comprehensive Reviews in Food Science and Food Safety*, 2022, 21(6): 5 199-5 224.
- [10] KOS G, SIEGER M, MCMULLIN D, et al. A novel chemometric classification for FTIR spectra of mycotoxin-contaminated maize and peanuts at regulatory limits[J]. *Food Additives & Contaminants: Part A*, 2016, 33(10): 1 596-1 607.
- [11] ELEM T C, PIACENTINI K C, ALVES L T, et al. NIR spectroscopy and chemometric tools to identify high content of deoxynivalenol in barley[J]. *Food Additives & Contaminants: Part A*, 2020, 37(9): 1 542-1 552.
- [12] DE GIROLAMO A, CERVELLIERI S, CORTESE M, et al. Fourier transform near-infrared and mid-infrared spectroscopy as efficient tools for rapid screening of deoxynivalenol contamination in wheat bran[J]. *Journal of the Science of Food and Agriculture*, 2019, 99(4): 1 946-1 953.
- [13] 邓志扬, 廖强, 邵淑娟, 等. 基于自然语言处理的山楂果实品种近红外无损鉴别方法[J]. *食品工业科技*, 2023, 44(22): 249-256.  
DENG Z Y, LIAO Q, SHAO S J, et al. Nondestructive near-infrared identification of hawthorn fruit cultivars based on natural language processing[J]. *Science and Technology of Food Industry*, 2023, 44(22): 249-256.
- [14] 卢诗扬, 张雷蕾, 潘家荣, 等. 拉曼光谱结合 LSTM 长短期记忆网络的樱桃产地鉴别研究[J]. *光谱学与光谱分析*, 2021, 41(4): 1 177-1 181.  
LU S Y, ZHANG L L, PAN J R, et al. Study on the identification of the geographical origin of cherries using Raman spectroscopy and LSTM[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(4): 1 177-1 181.
- [15] 何卓霖, 李子琨, 穆蕾, 等. 同位素内标—高效液相色谱—串联质谱法同时测定小麦和玉米中 19 种真菌毒素[J]. *中国粮油学报*, 2022, 37(7): 150-157.  
HE Z L, LI Z K, MU L, et al. Simultaneous determination of isotope internal standard-high performance liquid chromatography-tandem mass spectrometry 19 mycotoxins in wheat and corn[J]. *Journal of the Chinese Cereals and Oils Association*, 2022, 37(7): 150-157.
- [16] MOKARI A, GUO S X, BOCKLITZ T. Exploring the steps of infrared (IR) spectral analysis: pre-processing, (classical) data modelling, and deep learning[J]. *Molecules*, 2023, 28(19): 1 177-1 181.

- 6 886.
- [17] BI Y M, YUAN K L, XIAO W Q, et al. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation[J]. *Analytica Chimica Acta*, 2016, 909: 30-40.
- [18] MARTENS H, STARK E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy[J]. *Journal of Pharmaceutical and Biomedical Analysis*, 1991, 9 (8): 625-635.
- [19] 蔡海辉, 周岭, 史舟, 等. 利用 CARS-BPNN 模型的南疆枣园土壤有机质高光谱反演[J]. *光谱学与光谱分析*, 2023, 43(8): 2 568-2 573.
- CAI H H, ZHOU L, SHI Z, et al. Hyperspectral inversion of soil organic matter in jujube orchard in southern Xinjiang using CARS-BPNN[J]. *Spectroscopy and Spectral Analysis*, 2023, 43(8): 2 568-2 573.
- [20] 李跑, 周骏, 蒋立文, 等. 窗口竞争性自适应重加权采样策略的近红外特征变量选择方法[J]. *光谱学与光谱分析*, 2019, 39(5): 1 428-1 432.
- LI P, ZHOU J, JIANG L W, et al. A variable selection approach of near infrared spectra based on window competitive adaptive reweighted sampling strategy[J]. *Spectroscopy and Spectral Analysis*, 2019, 39(5): 1 428-1 432.
- [21] LI H D, LIANG Y Z, XU Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. *Analytica Chimica Acta*, 2009, 648(1): 77-84.
- [22] ZOU X B, ZHAO J W, POVEY M J W, et al. Variables selection methods in near-infrared spectroscopy[J]. *Analytica Chimica Acta*, 2010, 667(1/2): 14-32.
- [23] ZHANG S C, LI X L, ZONG M, et al. Efficient KNN classification with different numbers of nearest neighbors[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(5): 1 774-1 785.
- [24] MA Y F, LIANG X, SHENG G, et al. Noniterative sparse LS-SVM based on globally representative point selection[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(2): 788-798.
- [25] RIGATTI S J. Random forest[J]. *Journal of Insurance Medicine*, 2017, 47(1): 31-39.
- [26] BECKER T, ROUSSEAU A J, GEUBBELMANS M, et al. Decision trees and random forests[J]. *American Journal of Orthodontics and Dentofacial Orthopedics*, 2023, 164(6): 894-897.
- [27] MORO M K, DE CASTRO E V R, ROMÃO W, et al. Data fusion applied in near and mid infrared spectroscopy for crude oil classification[J]. *Fuel*, 2023, 340: 127580.
- [28] FREITAG S, ANLANGER M, FOMINA P, et al. Attenuated total reflection mid-infrared spectroscopy to screen Austrian and French wheat from multiple years for deoxynivalenol[J]. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy*, 2025, 340: 126340.
- [29] ALI H, MUTHUOSS P, CHAUHAN C, et al. Machine learning-enabled NIR spectroscopy. Part 3: hyperparameter by design (hyd) based ANN-MLP optimization, model generalizability, and model transferability[J]. *AAPS Pharm Sci Tech*, 2023, 24(8): 254.
- [30] KERNBACH J M, STAARTJES V E. Foundations of machine learning-based clinical prediction modeling: part II-generalization and overfitting[J]. *Acta Neurochirurgica*, 2022, 134: 15-21.