褚楚,罗雪路,王海童,等.基于机器学习和中红外光谱的牦牛奶掺假预测模型研究[J].华中农业大学学报,2025,44(2):116-124.
DOI:10.13300/j.cnki.hnlkxb.2025.02.012

基于机器学习和中红外光谱的牦牛奶掺假预测模型研究

褚楚[1],罗雪路[1],王海童[1],温佩佩[1],杜超[1],丁考仁青[2],拉毛草[3],张淑君[1]

1.华中农业大学动物科学技术学院、动物医学院,武汉 430070; 2.甘南藏族自治州畜牧工作站,合作 747000;
3.甘肃省碌曲县动物疫病预防控制中心,碌曲 747200

摘要 为监督和规范牦牛奶生产和销售,对牦牛奶中奶牛奶的掺假比例进行定量预测,研发新的快速检测技术,通过中红外光谱技术结合机器学习算法建立检测牦牛奶中掺加奶牛奶的预测模型,以76份纯牦牛奶、76份掺加10%奶牛奶的牦牛奶、76份掺加25%奶牛奶的牦牛奶、76份掺加50%奶牛奶的牦牛奶为研究对象,利用5种光谱预处理算法、6种定性和12种定量机器学习算法,分别建立鉴别纯牦牛奶和掺加奶牛奶的牦牛奶的二分类定性模型和预测掺加奶牛奶比例的定量回归模型。结果显示,基于支持向量机建模算法、无预处理光谱建立的鉴定纯牦牛奶和掺加奶牛奶的牦牛奶的预测模型效果最好,该模型验证集AUC为0.95,准确性0.84,灵敏度0.93,特异性0.87,可用于纯奶和掺假奶的鉴定。利用贝叶斯正则化神经网络建模算法和一阶导数光谱预处理算法建立了预测牦牛奶中奶牛奶掺加比例的最佳定量模型,该模型R_P^2=0.88,$RMSE_V$=6.57%,RPD=2.89%。结果表明,中红外光谱技术结合机器学习算法可有效地鉴定出掺加奶牛奶的牦牛奶,并可检测出掺假的比例。

关键词 中红外光谱(MIRS); 机器学习; 牦牛奶; 牛奶掺假; 预测模型

中图分类号 O657.33;TS252.7 **文献标识码** A **文章编号** 1000-2421(2025)02-0116-09

牦牛是高海拔高寒地区重要的生物,是一种非常珍贵的畜种。牦牛奶中蛋白质、乳糖、共轭亚油酸和钙含量较高,被认为是一种天然浓缩的牛奶[1]。牦牛奶的营养价值和价格均高于奶牛奶,导致不法生产商可能在牦牛奶生产中掺入奶牛奶以增加利润[2]。牦牛奶掺假行为对消费者的健康和财产安全有不利影响,因为这些掺假奶制品中的奶牛奶成分可能会使一部分人群诱发过敏等不良反应[3]。因此,从法律、消费者保护和消费者信心的角度,快速检测牦牛奶中掺加奶牛奶的含量以确保牦牛奶质量安全非常重要。

迄今为止,已经开发了几种分析方法用于检测不同动物奶中奶牛奶的掺加,如PCR[4]、酶联免疫吸附测定法、毛细管电泳法[5]、聚丙烯酰胺凝胶电泳、高效液相色谱法[6]等,但上述技术均存在耗时、成本较高和无法大批量检测等问题,无法对乳制品行业中原料奶掺假进行大规模筛查。中红外光谱(Mid-infrared spectroscopy,MIRS)技术是一种实时在线的生化指纹技术,与传统方法相比,具有快速、灵敏、低成本和高通量等优点[7]。MIRS技术基于电磁辐射与化学键之间的相互作用,目前已应用于预测牛奶的脂肪酸组成[8]、蛋白质组成[9]、矿物质含量[10]、奶牛健康状况(如是否发生乳腺炎、酮病等)[11-12]和繁殖状态(如奶牛是否妊娠)[13-14]。MIRS技术结合适当的机器学习算法,能够从光谱中提取定性和定量信息,从而快速对食品进行表征化和分类。因此,MIRS技术可能是检测和量化牦牛奶中掺加奶牛奶的理想解决方案。MIRS技术已成功检测出水牛奶[15-19]、山羊奶[17,20]和骆驼奶[21-22]中掺加的奶牛奶,但目前还没有关于MIRS技术检测牦牛奶中掺加奶牛奶的报道。

本研究利用MIRS不同预处理方法和机器学习建模方法,探究MIRS用于检测和量化牦牛奶中欺诈性的奶牛奶掺加的可能性,并且根据最优MIRS预处理和最优建模算法开发用于检测牦牛奶中掺加奶牛奶的定性鉴定模型及定量回归模型,旨在提高预测模型的准确性,为大规模检测牦牛奶中奶牛奶的掺加情况提供新思路。

收稿日期: 2024-04-07
基金项目:国家重点研发计划-政府间国际科技创新合作(2021YFE0115500);湖北省国际合作项目(2022EHB043)
褚楚,E-mail:chu1999@webmail.hzau.edu.cn
通信作者: 张淑君,E-mail:sjxiaozhang@mail.hzau.edu.cn

1　材料与方法

1.1　试验材料

从我国青海地区牦牛养殖场中采集健康状况良好的牦牛的奶样共76份，从我国华中和华北地区的荷斯坦奶牛场采集健康状况良好的奶牛的奶样154份，所有样本的采集时间为2021年。牦牛属于草原散养，荷斯坦奶牛属于集约化饲养。样品采集后倒入采样瓶中，依次编号，并向每个采样瓶里立即加入溴硝丙二醇防腐剂，缓慢摇晃使其充分溶解。样品采集完成后放于4 ℃环境进行保存，立刻运输至华中农业大学动物遗传育种实验室进行掺假样品的制备。

用于建立模型的样本应该代表生产中的实际情况。根据成本测算，在原奶中添加10%的掺假溶液，可使奶农每年多增收6万元，是奶农普通年收入的2倍多[23]。此外，根据巴西警方的调查，掺假的液态奶常含有10%～15%的掺假物[24]。因此，本研究奶牛奶的掺加比例设置在10%～50%。将奶牛奶添加至纯牦牛奶中，并按照0%、10%、25%和50%（V/V）进行混合。共制备了76个牦牛奶-10%奶牛奶混合物（掺加10%奶牛奶的牦牛奶），76个牦牛奶-25%奶牛奶混合物，76个牦牛奶-50%奶牛奶混合物。

1.2　仪器、设备和试剂

MilkoScan^{TM}FT＋，傅里叶变换中红外光谱仪（FTIR），丹麦FOSS公司；涡漩振荡器；离心管。

1.3　中红外光谱的采集

将新鲜的纯牦牛奶样品及掺加奶牛奶的牦牛奶样品运输至奶牛生产性能测定（dairy herd improvement，DHI）中心，利用MilkoScan FT＋仪器进行分析，以获取样品的MIRS、乳脂率、乳蛋白率、乳糖率、尿素氮含量和总固形物含量。每个样品在机器上扫描2次，最终结果（MIRS光谱和乳成分）输出2次的平均值。

1.4　光谱预处理

牛奶和牦牛奶的MIRS由5 008～925 cm^{-1}范围内的1 060个波点组成。为了遵循比尔定律，在建模前将光谱从透射率转换为吸光度[11]。牛奶MIRS中存在大量的背景噪声和无用信息，为去除光谱采集过程中环境、仪器及操作引起的系统误差，建模前需先对光谱进行预处理。本研究采用的光谱预处理方法包括无预处理、一阶导数（first derivative，1D）、二阶导数（second derivative，2D）、标准正态变量变换（standard normal variate transformation，SNV）和Savitsky-Golay平滑（SG平滑）。SNV主要用于消除粒径和表面散射光对光谱的影响，导数处理和平滑处理可以有效消除基线和其他背景噪声的干扰。结果仅展示最佳光谱预处理。

1.5　建模波段选择

5 008～2 968 cm^{-1}的区域被认为是噪音区，因此本研究中将此波段去除。研究[25]表明，1 773～2 802 cm^{-1}的区域内不包含有价值的信息，1 692～1 604 cm^{-1}区域与水的吸收有关，因此这2个波段也不参与建模。最后，剩余244个波点用于建模（2 968～2 802、1 773～1 692和1 604～925 cm^{-1}）。

1.6　模型建立

将数据集随机划分为校准集（80%）和验证集（20%）2部分，校准集用于训练模型，验证集数据独立于校准集，用于验证模型的性能。本研究共涉及2种模型，即：（1）二分类模型：纯牦牛奶样本定义为阴性，掺加奶牛奶的样本定义为阳性，此类模型可用于鉴别纯牦牛奶和掺加奶牛奶的牦牛奶，即鉴定牦牛奶中是否掺加了奶牛奶；（2）定量回归模型：将掺加奶牛奶的比例看作连续型变量，此类模型可用于预测牦牛奶中掺加奶牛奶的体积比。

本研究选择了最常用的6种分类算法和12种回归算法进行建模：偏最小二乘判别分析（partial least squares discriminant analysis，PLSDA）、分类回归树（classification and regression tree，CART）、随机森林（random forest，RF）、梯度增强机（gradient boosting machine，GBM）、支持向量机（support vector machine，SVM）和朴素贝叶斯（naive bayes，NB）等6种机器学习算法构建二分类定性模型。偏最小二乘回归（partial least squares regression，PLSR）、SVM、贝叶斯正则化神经网络（bayesian regularized neural network，BRNN）、尖峰和平板回归（spike and slab regression，SSR）、投影寻踪回归（projection pursuit regression，PPR）、CART、岭回归（ridge regression，RR）、最小绝对收缩和选择算子（least absolute shrinkage and selection operator，LASSO）、弹性网回归（elastic net regression，EN）、RF、GBM和极致梯度提升（extreme gradient boosting，XGB）等12机器学习算法构建定量回归模型。除偏最小二乘（partial least squares，PLS）算法以外，其余算法归为现代统

计机器学习算法。本研究中使用的所有机器学习算法都使用了R语言中的caret包，所有分析均使用R统计软件4.2.2版本进行。

重复5次的十折交叉验证用于构建预测模型和选择各种算法的关键参数。PLS潜在变量的最大数量设定为20个。BRNN的隐层数范围为1到4个，RF的mtry数为3、10、20、50、100、300、700、1 000和2 000。SVM的计算基于带核或径向基函数核的支持向量机，在caret软件包中使用method="svmLinear"或"svmRadial"作为参数来实现。对于"svmLinear"，C值为0.01、0.05、0.1、0.25、0.5、0.75、1、1.25、1.5、1.75、2和5；对于"svmRadial"，C值为0.01、0.05、0.1、0.25、0.5、0.75、1、1.25、1.5、1.75、2和5，sigma值为0.01、0.02、0.03、0.04、0.05、0.06、0.07、0.08、0.09、0.1、0.25、0.5、0.75、0.90。其余算法使用默认的内置参数。

1.7 模型性能的评价指标

模型性能通过2种方式进行评估：模型建立过程（校准集）和外部验证过程（验证集）。

使用校准集和验证集的准确性、敏感性、特异性和接受者操作特征曲线（recipient operation characteristic curve，ROC）下的面积（area under curve，AUC）等4个指标评估二分类模型的预测性能。准确性是指被正确分类的比例；敏感性指阳性数据被正确预测为阳性的比例；特异性指阴性数据被正确预测为阴性的比例[26]。ROC曲线通常用于评估诊断工具的性能，表示在不同的分类阈值下模型的真阳性率和假阳性率之间的关系，AUC是最常见的ROC汇总度量，取值为0至1。当0.9<AUC<1时，表明模型性能极好；当0.8<AUC<0.9，表明模型性能良好；当0.7<AUC<0.8，表明模型性能中等；当0.6<AUC<0.7，表明模型性能较差；当0.5<AUC<0.6，表明模型性能极差；当AUC=0.5，表明模型没有预测价值，类似随机猜测[27]。

利用校准集决定系数（coefficient of determination of calibration，R^2_C）、校准集均方根误差（root mean square error of calibration，$RMSE_C$）、验证集决定系数（coefficient of determination of validation，R^2_V）、验证集均方根误差（root mean squared error of validation，$RMSE_V$）、平均绝对误差（mean absolute error，MAE）及性能偏差比（ratio of performance to deviation，RPD）评估每种回归方法的性能。

Jabri等[28]对预测方程的R^2和RPD进行总结如下：根据R^2值可以将模型稳健性划分为4个等级：差（$R^2<0.66$）、中等（$0.66<R^2<0.81$）、好（$0.82<R^2<0.90$）和极好（$R^2\geqslant0.91$）。RPD越高越好，RPD>2的模型可以实现高准确性预测。

最佳模型的选择遵循以下规则：二分类模型，需要高的AUC、准确性、敏感性和特异性；定量回归模型，需要较高的R^2和RPD以及较低的RMSE和MAE。

1.8 统计分析

所有的分析和画图均使用R软件（版本4.3.1；https://www.r-project.org/）进行。使用t检验对平均值进行两两比较，所有检验的统计学显著性分析水平在$\alpha=0.05$。

2 结果与分析

2.1 牦牛奶和奶牛奶的营养物质含量对比

牦牛奶与奶牛奶的主要营养物质含量见表1。牦牛奶中乳脂率（8.18%）、乳蛋白率（5.26%）和总固形物含量（18.45%）显著高于奶牛奶（$P<0.05$），乳糖率（4.39%）和尿素氮含量（8.13 mg/100 g）显著低于奶牛奶（$P<0.05$）。差异最大的物质为乳脂率。

2.2 牦牛奶、奶牛奶和掺假牦牛奶的MIRS分析

牦牛奶、掺假牦牛奶（牦牛奶-奶牛奶混合物）和奶牛奶的原始MIRS图如图1所示。由图1可见，掺假牦牛奶中奶牛奶比例越高，则纯牦牛奶与掺假牦牛奶的光谱差异越大（图1A）。去除水区域后，可见最大的差异波段位于2 968～2 802 cm^{-1}（图1B）以及1 773～1 692 cm^{-1}（图1C），这些波段的吸光度主要与乳脂含量有关。纯牦牛奶MIRS与其掺假物MIRS之间的其他较明显差异位于牛奶指纹区（925～1 604 cm^{-1}）的乳蛋白（1 544 cm^{-1}）和乳糖（1 159 cm^{-1}和1 076 cm^{-1}）吸收区（图1D）。

纯牦牛奶与其含奶牛奶掺假物之间的MIRS存在差异，可使用简单的目视观察法进行粗略的区分，但当掺假量很低时，简单的目视观察法可能无法准确判断，更无法量化奶牛奶的掺加程度，因此，借助机器学习算法或许可以提取这些差异，进行鉴定或定量分析。

2.3 鉴别纯牦牛奶与掺加奶牛奶的牦牛奶（牦牛奶-奶牛奶）的二分类定性模型

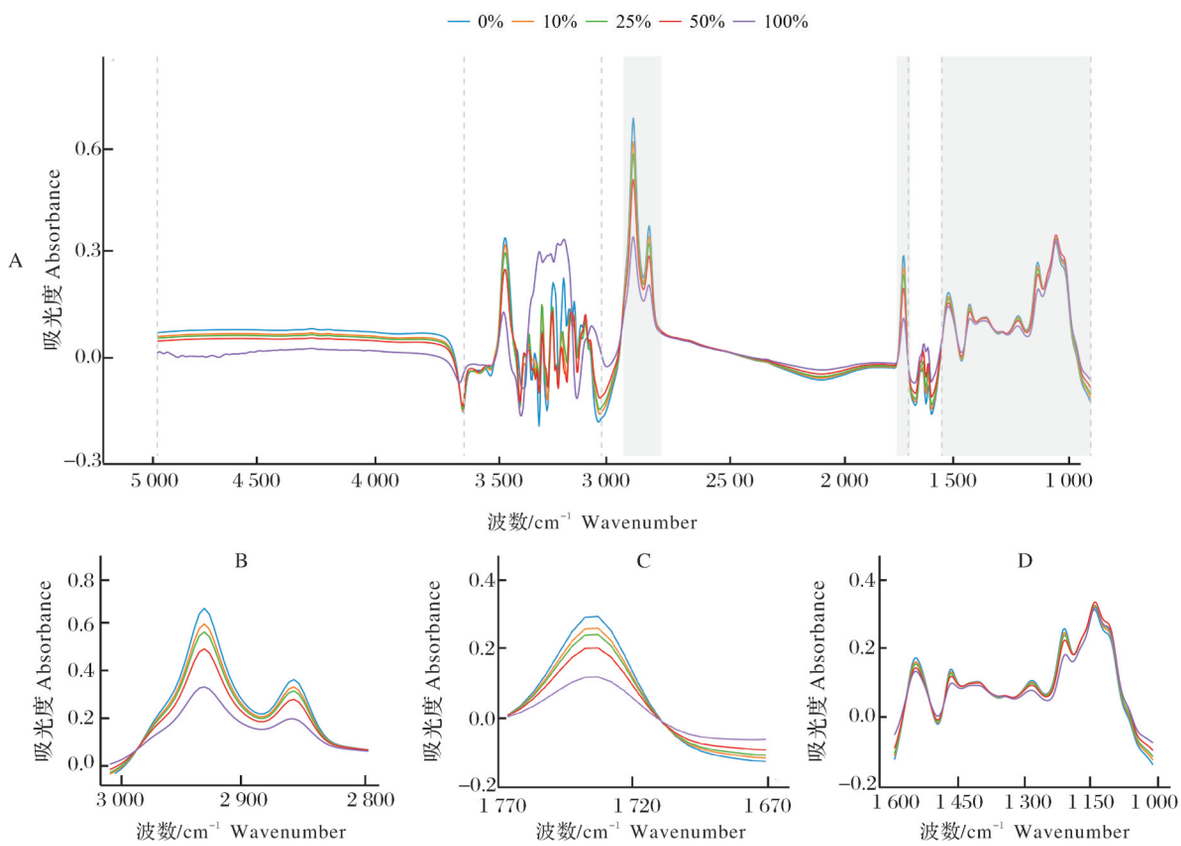
利用6种机器学习算法建立的鉴定纯牦牛奶与

表 1　牦牛奶和奶牛奶的营养物质组成

Table 1　The nutrient composition of yak milk and dairy cattle milk

牛奶分类 Milk classification	乳脂/% Fat	乳蛋白/% Protein	乳糖/% Lactose	尿素氮/(mg/100 g) Urea	总固形物/% Total solids
牦牛奶 Yak milk	8.18±1.55b	5.26±0.75b	4.39±0.32a	8.13±2.46a	18.45±1.94b
奶牛奶 Dairy cattle milk	3.54±1.00a	3.59±0.34a	4.89±0.40b	10.52±1.79b	12.68±1.22a

注：同一列中标不同字母者差异显著，$P<0.05$。Note：Means with different letters differ significantly，$P<0.05$.



A：全光谱图；B、C、D代表灰色区域的光谱图。0%、10%、25%、50%和100%代表牦牛奶-奶牛奶混合物中奶牛奶的体积百分比；灰色区域代表建模波段。A represents the full spectrum; B, C, D represents the spectrum in the grey region. 0%, 10%, 25%, 50% and 100% represent the volume proportion of dairy cattle milk in the yak milk-dairy cattle milk mixture; Grey area represents modeling band.

图 1　牦牛奶、奶牛奶和不同掺假含量（10%~50%）的牦牛奶-奶牛奶混合物的中红外光谱图

Fig.1　Mid-infrared spectroscopy of yak milk, dairy cattle milk and yak milk-dairy cattle milk mixture with different adulteration ratios (10%-50%)

掺加奶牛奶的牦牛奶的二分类模型如表 2 所示。SVM算法最优，校正集AUC为0.95、准确性0.87、敏感性0.98、特异性0.90，验证集AUC为0.95、准确性0.84、敏感性0.93、特异性0.87；PLSDA算法次之，NB算法表现最差。SVM算法与目前最常用的定性模型PLSDA（验证集AUC为0.93、准确性0.78、敏感性0.93、特异性0.82）相比，验证集AUC、准确性、敏感性和特异性分别提高了0.02、0.06、0.00和0.05；与表现最差的NB算法（验证集AUC为0.75、准确性0.60、敏感性0.80、特异性0.65）相比，验证集AUC、准确性、敏感性和特异性分别提高了0.20、0.24、0.13和0.22。

综合6种机器学习建模算法和5种光谱预处理算法建立的30种模型的性能评价指标（表2）发现，使用无预处理的MIRS和SVM建模算法建立的鉴定纯牦牛奶与掺加奶牛奶的牦牛奶的二分类定性模型产生了最高预测准确性，AUC为0.95、准确性0.84、敏感性0.93、特异性0.87，该模型可将牦牛奶样归为无

表2 二分类定性模型在校正集和验证集中的性能

Table 2 Performance of binary qualitative model in calibration set and validation set

建模算法 Modeling algorithm	预处理 Pretreatment	校正集 Calibration set				验证集 Validation set			
		AUC	准确性 Accuracy	敏感性 Sensitivity	特异性 Specificity	AUC	准确性 Accuracy	敏感性 Sensitivity	特异性 Specificity
PLSDA	None	0.95	0.84	0.97	0.87	0.93	0.78	0.93	0.82
SVM	None	0.95	0.87	0.98	0.90	0.95	0.84	0.93	0.87
CART	1D	0.82	0.72	0.90	0.76	0.79	0.64	0.80	0.68
RF	1D	1.00	1.00	1.00	1.00	0.85	0.82	0.80	0.82
GBM	None	0.88	0.75	0.85	0.77	0.87	0.76	0.87	0.78
NB	1D	0.78	0.60	0.84	0.66	0.75	0.60	0.80	0.65

注：结果只显示了最佳的光谱预处理。表4同。Note：The results demonstrate only the optimal spectral preprocessing. The same as below in Table 4.

掺假和有掺假2类。从表3可以看出，本研究建立的预测模型可以鉴别出纯牦牛奶的准确性是93%，鉴别出掺加了50%奶牛奶的牦牛奶的准确性为100%，鉴别出掺加了25%奶牛奶的牦牛奶的准确性约90%，然而，当奶牛奶掺加比例小于10%时，预测准确性不高，只有62%。

表3 二分类最优模型预测校准集和验证集的准确性结

Table 3 Accuracy of optimal binary classification model predicting calibration set and verification set data %

掺加奶牛奶的体积百分比 Proportion of adulteration of dairy cattle milk	预测准确性 Predictive accuracy	
	校正集 Calibration set	验证集 Validation set
0	98	93
10	68	62
25	93	89
50	100	100

2.4 预测牦牛奶中掺加奶牛奶比例的定量回归模型

最佳MIRS预处理及12种机器学习算法分别建立的预测牦牛奶中掺加奶牛奶比例的回归模型性能见表4。PLSR被认为是传统的基准方法，因为PLSR在化学计量分析中始终具有强大的预测性能，然而，在本研究中PLSR并没有表现出最佳效果。SSR、PPR、CART、RR、EN和LASSO算法均提供了较差的预测结果，RPD_V小于2。BRNN、GBM和XGB算法优于PLSR，其余算法差于PLSR。BRNN算法表现最优，其次是GBM和XGB，CART算法表现最差。BRNN算法与最常用的PLSR算法相比，$RMSE_V$降低了2.37%，R_V^2和RPD_V分别提高了0.10和0.76。

综合以上12种机器学习建模算法和5种MIRS预处理算法建立的60种模型的性能评价指标结果发现，利用BRNN建模算法和1D光谱预处理算法建立的量化牦牛奶中奶牛奶掺加比例的回归模型性能最优，其中，$RMSE_V$=6.57%，MAE_V=5.22%，R_V^2=0.88，RPD_V=2.89。

3 讨 论

3.1 牦牛奶、掺假牦牛奶（牦牛奶-奶牛奶混合物）和奶牛奶的中红外光谱分析

牛奶的光谱由5 008～925 cm^{-1}范围内的1 060个波点组成，划分为短波红外区（short-wavelength infrared，SWIR）、中波红外区（mid-wavelength infrared，MWIR）和长波红外区（long-wavelength infrared，LWIR）[29]。5 010～3 673 cm^{-1}被称为SWIR区域；3 669～3 052 cm^{-1}被称为SWIR-MWIR区域；3 048～1 701 cm^{-1}被称为MWIR-1区域；1 698～1 585 cm^{-1}被称为MWIR-2区域；1 582～925 cm^{-1}被称为MWIR-LWIR区域。本研究发现，牦牛奶、掺假牦牛奶（牦牛奶-奶牛奶混合物）和奶牛奶的光谱差异主要存在于MWIR-1、MWIR-2、SWIR-MWIR和MWIR-LWIR区域。MWIR-2及SWIR-MWIR区域与水吸收有关，这些光谱特征增加了不同牛奶样品之间吸光度的变异性，在预测牛奶物质成分及奶牛生理状态时这个区域通常被排除在外。MWIR-1区域主要的吸收峰是C—H、C=O，C—N和N—H键[30]，所有这些键都与乳脂含量有关。在这一区域，检测到一些吸光度差异较大的峰。第1个重要光谱区域位于2 968～2 802 cm^{-1}，该区域与Fat-B的C—H键振动有关[31]。第2个重要光谱区域位于1 773～1 692 cm^{-1}，此区域与Fat-A的羰基振动有关[31]，本研究建模过程用到了这2个光谱吸收区域。本研究

表 4　基于中红外光谱的牦牛奶中掺加奶牛奶含量的预测模型性能

Table 4　Performance of the predictive model for cow milk content adulteration in yak milk based on mid-infrared spectroscopy

建模算法	预处理	校正集 Calibration set				验证集 Validation set			
Modeling algorithm	Pretreatment	$RMSE_C$	MAE_C	R_C^2	RPD_C	$RMSE_V$	MAE_V	R_V^2	RPD_V
PLSR	None	8.83	6.87	0.78	2.14	8.94	7.02	0.78	2.13
SVM	None	9.44	6.50	0.75	2.00	9.15	6.39	0.77	2.08
SSR	None	11.24	8.85	0.65	1.68	10.46	8.41	0.70	1.82
PPR	SG	3.98	2.44	0.96	4.74	10.21	6.65	0.72	1.86
CART	1D	7.97	6.02	0.82	2.37	11.07	7.92	0.66	1.72
BRNN	1D	2.68	2.12	0.98	7.05	6.57	5.22	0.88	2.89
RR	None	11.31	8.92	0.64	1.67	10.60	8.49	0.69	1.79
EN	None	11.03	8.69	0.66	1.71	10.34	8.17	0.70	1.84
LASSO	None	10.93	8.58	0.67	1.73	10.25	8.07	0.71	1.85
RF	1D	3.59	2.72	0.97	5.26	9.00	6.60	0.79	2.11
GBM	1D	0.59	0.43	1.00	32.19	7.29	5.61	0.85	2.61
XGB	2D	2.92	2.30	0.98	6.46	8.14	6.53	0.82	2.33

建模用到的另外 1 个光谱区域是 MWIR-LWIR 区域，这是牛奶的指纹区，与乳蛋白[32]、尿素氮[33]和乳糖[34]吸收有关。

3.2　鉴定和预测牦牛奶中掺加奶牛奶模型的准确性

目前用于检测牦牛奶中奶牛奶掺假的方法有聚丙烯酰胺凝胶电泳[35]、酶联免疫吸附测定技术[1]和质谱法[36]等，这些技术的检测误差与准确性与本研究相似，但本研究使用的方法具有快速、环境友好以及大批量测定的优点。目前还没有关于 MIRS 技术预测牦牛奶中掺加奶牛奶的报道。本研究利用 6 种机器学习分类建模算法、12 种机器学习回归建模算法和 5 种光谱预处理方法建立了基于 MIRS 的牦牛奶（原料奶）中掺加奶牛奶的定性鉴定模型和定量预测模型，并筛选出 2 个最优模型：对于定性鉴别模型（二分类），SVM 建模算法和无预处理光谱建立的模型表现出最优预测性能，验证集 AUC 为 0.95、准确性 0.84、敏感性 0.93、特异性 0.87，对于定量回归模型，BRNN 算法和 1D 光谱预处理算法建立的模型表现出最优预测性能，$RMSE_V$=6.57%，MAE_V=5.22%，R_V^2=0.88，RPD_V=2.89。这些性能统计结果表明，本研究建立的 2 个模型性能良好，可以对牦牛奶中奶牛奶的掺加情况进行初步的预测[27-28]。但由于只有 10%、25%、50% 这 3 个掺加梯度，定量模型的性能结果仅供参考。

现有的关于 MIRS 技术预测水牛奶[15-19]、山羊奶[17,20]和骆驼奶[21-22]中掺加奶牛奶比例的研究得到的预测误差（$RMSE_V$）范围分别为 2.84%～7.42%、2.84%～8.03% 和 0.87%～0.99%，本研究建立的牦牛奶中掺加奶牛奶的定量预测模型的 $RMSE_V$ 为 6.57%，在水牛奶、山羊奶的预测误差范围内。与其他研究相比，本研究的优势在于数据量大和尝试利用多种建模算法及光谱预处理方法，从而能够充分挖掘 MIRS 中包含的有用信息，建立稳健的预测模型。由于本研究建模过程中使用的数据量不足以满足生产应用的需要，而且牦牛奶的数据仅来自我国的一个省份，没有利用其他省份的数据进行外部验证，因此这些方法在不同地理区域和不同规模的乳品业中的实用性还有待考虑。然而，从本研究获得的初步模型的结果参数来看，所开发的模型有望应用于不同地理区域和不同生产规模的乳品业。

3.3　现代统计机器学习算法与 PLS 算法之间的比较

预测结果的可靠性和准确性在很大程度上取决于模型的质量，模型的质量与建模数据集、光谱质量以及用于开发预测模型的算法（包括变量选择、光谱预处理和模型）有关[37]。由于涵盖共线、高维数据集，PLS 是将牛奶的 MIRS 数据与牛奶和动物性状相关联的首选及最传统的方法，但是对于变量之间的复杂关系（如非线性和互作），可能并不是理想的处理方法[38]。关于 MIRS 对原料奶掺加预测的研究，主要利用的建模算法为 PLS，一些牛奶 MIRS 的研究证明其他机器学习算法如随机森林、神经网络、决策树、神经网络等也能够有效处理牛奶 MIRS 数据[39]，均能较好地对复杂关系进行建模，但迄今为止，这些现代统计机器学习算法在 MIRS 分析中的应用仍然

较少，很少有学者探究利用MIRS信息预测动物原料奶中掺假的潜力以及与PLS算法的比较。

本研究利用多种现代统计机器学习算法建立了预测牦牛奶中掺加奶牛奶的模型，并与PLS算法进行比较。研究发现，现代统计机器学习算法对牦牛奶掺假的检测表现出优于PLS的性能，尤其是SVM算法和BRNN算法。SVM能够通过核函数将样本映射至较高维空间，有效增强模型的学习能力，适合处理非线性问题，且对样本数据分布无要求，对噪声、随机的容限度较大[40]。BRNN是1种将贝叶斯方法应用于神经网络的正则化技术。研究表明，与线性模型（如PLS）相比，神经网络能够提供较好的预测值[41]，但很容易受到过拟合的影响，在预测新数据时可能表现出较低的稳健性，具有贝叶斯正则化训练算法的神经网络可以避免这种过拟合。本研究使用了3种收缩方法（LASSO、EN和RR），总的来说，这些算法的预测性能相似（均差于PLSR），但LASSO和EN总是略优于RR。Sen等[34]也发现了类似的规律，这是在预期内的，因为LASSO和EN可以直接进行变量选择，而RR保留了所有变量。SSR性能同样差于PLSR，也是一种变量选择方法，但与LASSO和EN不同的是，SSR并不是基于缩放方法进行变量选择，而是采用贝叶斯方法。

本研究利用MIRS建立了检测和定量牦牛奶中掺加奶牛奶的预测模型，即基于SVM建模算法、无预处理光谱建立的鉴定纯牦牛奶和掺加奶牛奶的牦牛奶的预测模型和基于贝叶斯正则化神经网络建模算法、一阶导数光谱预处理建立的预测牦牛奶中奶牛奶掺加比例的定量回归模型。结果表明，MIRS具有预测牦牛奶中掺加奶牛奶的潜力，二分类预测模型整体准确性为84%，该模型鉴别纯牦牛奶的准确性为93%，鉴别出牦牛奶中掺加了50%奶牛奶的准确性为100%，鉴别出牦牛奶中掺加了25%奶牛奶的准确性约90%，鉴别出牦牛奶中掺加了10%奶牛奶的准确性约62%，定量预测模型的预测误差为6.57%。SVM算法在分类模型中表现较优，BRNN算法在定量模型中表现较优，在其他相关研究中也可考虑这2种算法的应用。为了更准确地对牦牛奶不同比例的掺假情况进行检测，所建立的模型还需要进行更多训练和优化。

参考文献 References

[1] REN Q R, ZHANG H, GUO H Y, et al. Detection of cow milk adulteration in yak milk by ELISA[J]. Journal of dairy science, 2014, 97(10): 6000-6006.

[2] 付尚辰，李玲，郑卫民，等. 掺假羊乳及其制品中牛乳的检测技术研究进展[J]. 食品安全质量检测学报，2021，12(8)：3000-3007. FU S C, LI L, ZHENG W M, et al. Research progress on adulteration detection technology of cow milk in goat milk and its products[J]. Journal of food safety & quality, 2021, 12(8): 3000-3007 (in Chinese with English abstract).

[3] CHAFEN J J S, NEWBERRY S J, RIEDL M A, et al. Diagnosing and managing common food allergies[J/OL]. Clinical governance, 2010, 15(4): 7[2024-04-07]. https://doi.org/10.1108/cgij.2010.24815dae.007 .

[4] 尹艳. 奶及奶制品鉴别方法的研究[D]. 北京：北京化工大学，2013. YIN Y. Study on identification methods of milk and dairy products[D]. Beijing: Beijing University of Chemical Technology, 2013 (in Chinese with English abstract).

[5] TRIMBOLI F, COSTANZO N, LOPREIATO V, et al. Detection of buffalo milk adulteration with cow milk by capillary electrophoresis analysis[J]. Journal of dairy science, 2019, 102(7): 5962-5970.

[6] BOSCO C D, PANERO S, NAVARRA M A, et al. Screening and assessment of low-molecular-weight biomarkers of milk from cow and water buffalo: an alternative approach for the rapid identification of adulterated water buffalo mozzarellas[J]. Journal of agricultural and food chemistry, 2018, 66(21): 5410-5417.

[7] NICOLAOU N, XU Y, GOODACRE R. Fourier transform infrared spectroscopy and multivariate analysis for the detection and quantification of different milk species[J]. Journal of dairy science, 2010, 93(12): 5651-5660.

[8] ZHAO X X, SONG Y T, ZHANG Y P, et al. Predictions of milk fatty acid contents by mid-infrared spectroscopy in Chinese Holstein cows[J/OL]. Molecules, 2023, 28(2): 666[2024-04-07]. https://doi.org/10.3390/molecules28020666.

[9] SOYEURT H, GRELET C, MCPARLAND S, et al. A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra[J]. Journal of dairy science, 2020, 103(12): 11585-11596.

[10] CHRISTOPHE O S, GRELET C, BERTOZZI C, et al. Multiple breeds and countries' predictions of mineral contents in milk from milk mid-infrared spectrometry[J/OL]. Foods, 2021, 10(9): 2235[2024-04-07]. https://doi.org/10.3390/foods10092235.

[11] MENSCHING A, ZSCHIESCHE M, HUMMEL J, et al. Development of a subacute ruminal acidosis risk score and its prediction using milk mid-infrared spectra in early-lactation cows[J]. Journal of dairy science, 2021, 104(4): 4615-4634.

[12] DENHOLM S J, BRAND W, MITCHELL A P, et al. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning[J]. Journal of dairy science, 2020, 103(10): 9355-9367.

[13] BRAND W, WELLS A T, SMITH S L, et al. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning[J].Journal of dairy science,2021,104(4):4980-4990.

[14] TIPLADY K M, TRINH M H, DAVIS S R, et al.Pregnancy status predicted using milk mid-infrared spectra from dairy cattle[J].Journal of dairy science,2022,105(4):3615-3632.

[15] SILVA L K R, UNIVERSITY S B S, GONÇALVES B R F, et al. Spectroscopic method (FTIR-ATR) and chemometric tools to detect cow's milk addition to buffalo's milk[J].Revista mexicana de ingeniería química,2019,19(1):11-20.

[16] CIRAK O, ICYER N C, DURAK M Z. Rapid detection of adulteration of milks from different species using Fourier transform infrared spectroscopy (FTIR) [J]. Journal of dairy research,2018,85(2):222-225.

[17] SEN S, DUNDAR Z, UNCU O, et al. Potential of Fourier-transform infrared spectroscopy in adulteration detection and quality assessment in buffalo and goat milks [J/OL]. Microchemical journal, 2021, 166: 106207 [2024-04-07]. https://doi.org/10.1016/j.microc.2021.106207.

[18] SPINA A A, CENITI C, PIRAS C, et al. Mid-infrared (MIR) spectroscopy for the detection of cow's milk in buffalo milk [J]. Journal of animal science and technology, 2022, 64(3):531-538.

[19] GONÇALVES B H, SILVA G, DE JESUS J, et al.Fast verification of buffalo's milk authenticity by mid-infrared spectroscopy, analytical measurements and multivariate calibration[J]. Journal of the Brazilian chemical society,2020,31:1453-1460.

[20] NICOLAOU N, XU Y, GOODACRE R.Fourier transform infrared spectroscopy and multivariate analysis for the detection and quantification of different milk species[J].Journal of dairy science,2010,93(12):5651-5660.

[21] BOUKRIA O, BOUDALIA S, BHAT Z F, et al. Evaluation of the adulteration of camel milk by non-camel milk using multispectral image, fluorescence and infrared spectroscopy [J/OL]. Spectrochimica acta part A: molecular and biomolecular spectroscopy, 2023, 300: 122932 [2024-04-07]. https://doi.org/10.1016/j.saa.2023.122932.

[22] SOUHASSOU S, BASSBASI, M, HIRRI A, et al. Detection of camel milk adulteration using Fourier transformed infrared spectroscopy FT-IR coupled with chemometrics methods [J]. International food research journal, 2018, 25(3):1213-1218.

[23] 内蒙古自治区统计局.内蒙古统计年鉴2012[M].北京:中国统计出版社,2012.Inner Mongolia Bureau of Statistics. Neimenggu statistical yearbook 2012 [M]. Beijing: Chinese Statistics Press,2012(in Chinese).

[24] SANTOS P M, WENTZELL P D, PEREIRA-FILHO E R. Scanner digital images combined with color parameters: a case study to detect adulterations in liquid cow's milk[J].Food analytical methods,2012,5(1):89-95.

[25] CHU C, WANG H T, LUO X L, et al.Possible alternatives: identifying and quantifying adulteration in buffalo, goat, and camel milk using mid-infrared spectroscopy combined with modern statistical machine learning methods [J/OL]. Foods, 2023, 12(20): 3856 [2024-04-07]. https://doi.org/10.3390/foods12203856.

[26] DELHEZ P, HO P N, GENGLER N, et al. Diagnosing the pregnancy status of dairy cows: how useful is milk mid-infrared spectroscopy?[J]. Journal of dairy science, 2020, 103(4): 3264-3274.

[27] FAWCETT T. An introduction to ROC analysis [J]. Pattern recognition letters,2006,27(8):861-874.

[28] JABRI M E, SANCHEZ M P, TROSSAT P, et al.Comparison of Bayesian and partial least squares regression methods for mid-infrared prediction of cheese-making properties in Montbéliarde cows[J].Journal of dairy science,2019,102(8):6943-6958.

[29] 褚楚,张静静,丁磊,等.基于中红外光谱的牛奶中三种氨基酸含量预测模型的建立及应用[J].畜牧兽医学报,2023,54(8):3299-3312.CHU C, ZHANG J J, DING L, et al.Establishment and application of prediction model of three amino acids in milk based on mid-infrared spectroscopy[J].Acta veterinaria et zootechnica sinica, 2023, 54(8): 3299-3312 (in Chinese with English abstract).

[30] BITTANTE G, CECCHINATO A. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavelengths related to specific chemical bonds [J].Journal of dairy science,2013,96(9):5991-6006.

[31] KAYLEGIAN K E, LYNCH J M, FLEMING J R, et al.Influence of fatty acid chain length and unsaturation on mid-infrared milk analysis 1[J].Journal of dairy science,2009,92(6):2485-2501.

[32] KAYLEGIAN K E, HOUGHTON G E, LYNCH J M, et al. Calibration of infrared milk analyzers: modified milk versus producer milk 1 [J]. Journal of dairy science, 2006, 89(8): 2817-2832.

[33] HANSEN P W.Screening of dairy cows for ketosis by use of infrared spectroscopy and multivariate calibration [J]. Journal of dairy science,1999,82(9):2005-2010.

[34] SEN S, DUNDAR Z, UNCU O, et al. Potential of Fourier-transform infrared spectroscopy in adulteration detection and quality assessment in buffalo and goat milks [J/OL]. Microchemical journal, 2021, 166: 106207 [2024-04-07]. https://doi.org/10.1016/j.microc.2021.106207.

[35] 赵梦波.牦牛和犏牛乳的比较生物化学研究[D].成都:西南民族大学,2022.ZHAO M B. Comparative biochemical study on yak and yak milk[D].Chengdu: Southwest University for Nationalities,2022 (in Chinese with English abstract).

[36] 苗金梁,张九凯,周正火,等.不同乳源动物成分鉴别技术研究进展[J].食品安全质量检测学报,2021,12(18):7314-7323.MIAO J L, ZHANG J K, ZHOU Z H, et al.Research progress on identification technology of animal ingredients from different milk sources[J].Journal of food safety & quality,2021,12(18):7314-7323 (in Chinese with English abstract).

[37] MOTA L F M, PEGOLO S, BABA T, et al. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data[J]. Journal of dairy science, 2020, 104:8107-8121.

[38] SHADPOUR S, CHUD T C S, HAILEMARIAM D, et al. Predicting methane emission in Canadian Holstein dairy cattle using milk mid-infrared reflectance spectroscopy and other commonly available predictors *via* artificial neural networks [J].Journal of dairy science,2022,105(10):8272-8285.

[39] FRIZZARIN M, GORMLEY I C, BERRY D P, et al.Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods [J]. Journal of dairy science,2021,104(7):7438-7447.

[40] 方俊,王艺频.基于支持向量机技术的社会组织信用评估指标体系构建:以G省公益慈善类社会组织为例[J].广西大学学报(哲学社会科学版),2022,44(6):174-183.FANG J,WANG Y P. Construction of credit evaluation index system of social organizations based on support vector machine technology: taking charity social organizations in G province as an example [J]. Journal of Guangxi University (philosophy and social science), 2022,44(6):174-183 (in Chinese with English abstract).

[41] FERRAND-CALMELS M, PALHIÈRE I, BROCHARD M, et al.Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry[J].Journal of dairy science, 2014,97(1):17-35.

Prediction adulteration of yak milk based on machine learning and mid-infrared spectroscopy

CHU Chu[1], LUO Xuelu[1], WANG Haitong[1], WEN Peipei[1], DU Chao[1], Dingkaorenqing[2], Lamaocao[3], ZHANG Shujun[1]

1.*College of Animal Science and Technology, College of Veterinary Medicine, Huazhong Agricultural University, Wuhan 430070, China*;
2.*Animal Husbandry Station of Gannan Tibetan Autonomous Prefecture, Hezuo 747000, China*;
3.*Animal Disease Prevention and Control Center of Luqu County, Gannan Prefecture, Luqu 747200, China*

Abstract A predictive model for detecting the addition of milk to yak milk was established by combining mid-infrared spectroscopy (MIRS) with machine learning algorithms to supervise and regulate the production and sale of yak milk, further quantitatively predict the proportion of adulteration in yak milk and provide new technology of rapid detection. 76 samples of pure yak milk, 76 samples of yak milk adulterated with 10% milk, 76 samples of yak milk adulterated with 25% milk, and 76 samples of yak milk adulterated with 50% milk were used to establish binary qualitative models for distinguishing pure yak milk from yak milk adulterated with milk, and quantitative regression models for predicting the proportion of yak milk adulterated with milk with five spectral preprocessing algorithms, six qualitative and twelve quantitative machine learning algorithms. The results showed that the predictive model for identifying pure yak milk and yak milk adulterated with milk based on support vector machine modeling algorithm and the spectrum without preprocessing had the best performance. The validation set AUC, the accuracy, the sensitivity, and the specificity of the model was 0.95, 0.84, 0.93, and 0.87, which can be used for the identification of pure milk and adulterated milk. The optimal quantitative model for predicting the proportion of milk in yak milk was established using Bayesian regularized neural network modeling algorithm and first-order derivative spectral preprocessing algorithm. The model had R_P^2=0.88, RMSEV=6.57%, and RPD=2.89%. It is indicated that the combination of mid-infrared spectroscopy and machine learning algorithms can effectively identify yak milk adulterated with milk and detect the proportion of adulteration.

Keywords mid-infrared spectroscopy (MIRS); machine learning; yak milk; the adulteration of milk; predictive model

(责任编辑:赵琳琳)