Computer Science and Application 计算机科学与应用, 2026, 16(11), 50-59
Published Online November 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.1611325

Hans汉斯

基于随机森林外观特征筛选的大米品种识别研究

郑馨怡，张莎莎

广州新华学院计算机与人工智能学院，广东 东莞

收稿日期：2026年10月1日；录用日期：2026年10月29日；发布日期：2026年11月9日

摘 要

为筛选具有明确外观含义的大米品种识别指标，本文基于公开大米数据集开展随机森林特征筛选与分类对比研究。采用85%/15%的分层随机划分构建训练集与测试集，并在训练集内进行5折交叉验证。根据随机森林特征重要性构建不同规模的原始特征子集，比较径向基函数核支持向量机、随机森林和多层感知机等的分类表现，并与主成分分析降维方案、ANOVA F检验和L1正则化逻辑回归筛选方案进行对比。结果表明，随机森林筛选的13个原始特征包含5个形态特征、2个形状特征和6个颜色特征，结合多层感知机时，测试集宏平均F1为0.9981，接近全特征多层感知机的0.9986，输入特征数量减少87.7%。以逻辑回归开展的特征组对比显示，颜色特征组在三类单独特征组中表现最好，全部特征联合输入的分类性能最高。11项候选特征在5个训练折中均进入Top-13，表现出较高的跨折选择一致性。研究表明，随机森林筛选能够减少输入变量并保持较高分类性能，同时保留原始外观指标的名称和含义。

关键词

大米品种识别，随机森林，特征重要性，特征选择，机器学习

Rice Grain Variety Identification Using Random Forest-Based Selection of Appearance Features

Xinyi Zheng, Shasha Zhang

School of Computing and Artificial Intelligence, Guangzhou Xinhua University, Dongguan Guangdong

Received: October 1, 2026; accepted: October 29, 2026; published: November 9, 2026

文章引用: 郑馨怡, 张莎莎. 基于随机森林外观特征筛选的大米品种识别研究[J]. 计算机科学与应用, 2026, 16(11): 50-59. DOI: 10.12677/csa.2026.1611325

Abstract

To identify rice-variety indicators with clear appearance interpretations, this study conducted random-forest (RF)-based feature selection and comparative classification analysis using a publicly available rice dataset. An 85%/15% stratified random split was used to construct the training and test sets, and five-fold cross-validation was performed within the training set. Based on RF feature importance, subsets of original features with different sizes were constructed, and the classification performance of a support vector machine with a radial basis function kernel (RBF-SVM), RF, and a multilayer perceptron (MLP) was compared. The RF-based feature selection approach was also compared with dimensionality reduction using principal component analysis (PCA) and feature selection using the ANOVA F-test and L1-regularized logistic regression. The results showed that the 13 original features selected by RF comprised five morphological features, two shape features, and six color features. When combined with MLP, this RF-selected subset achieved a test-set macro-averaged F1 score of 0.9981, close to 0.9986 obtained by the full-feature MLP, while reducing the number of input features by 87.7%. Logistic-regression comparisons among feature groups showed that the color feature group performed best among the three individual groups, whereas combining all feature groups yielded the highest classification performance. Eleven candidate features were selected in the Top-13 subset in all five training folds, indicating high selection consistency across folds. These results demonstrate that RF-based feature selection can reduce the number of input variables while maintaining high classification performance and preserving the names and meanings of the original appearance indicators.

Keywords

Rice Variety Identification, Random Forest, Feature Importance, Feature Selection, Machine Learning

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

水稻是重要的粮食作物，大米的外观特征是大米品质评价和类型识别的重要信息来源。传统大米质量评价通常采用目视检查和人工测量，已有研究利用商品大米的形态与颜色参数开展自动分类，为基于外观信息的大米识别提供了依据[1]。人工评价与分类过程可能存在耗时较长、成本较高等局限，而图像分析与机器学习方法已被用于根据大米的形态、形状和颜色特征识别不同品种[2]。因此，研究基于大米外观信息的自动分类方法，并分析具有判别作用的关键指标，对辅助大米品种识别具有应用意义。

现有水稻品种识别研究主要包括两类技术路线：一类从大米图像中提取人工设计的形态、形状、纹理和颜色特征，再利用机器学习分类器进行识别；另一类利用深度神经网络从图像中学习判别特征。已有研究沿上述路线开展水稻品种分类，并取得较高的识别精度[1]-[4]。Coronel-Reyes 等[5]比较了大米形态特征在 PCA 降维前后多种机器学习分类器的表现；Li 等[6]结合残差学习与多尺度特征提取，并利用 Grad-CAM 展示模型关注的大米图像区域。两类方法在特征获取方式上各有特点：人工设计的外观特征具有明确的变量定义，便于将模型输入与大米的几何或颜色属性联系起来；深度学习方法能够自动学习图像表示，但所学特征通常难以直接对应单个外观指标。相关综述指出，该领域仍面临图像采集条件变

化、模型计算开销和实际部署等挑战[7]。

近期特征工程综述将预测性能、特征可解释性及重采样稳定性列为重要评价维度[8]。主成分分析(Principal Component Analysis, PCA)通过线性组合压缩特征空间，而原始特征筛选保留变量的名称和定义。本文据此比较筛选后的原始特征、同维度 PCA 表示与全特征输入，考察精简外观指标后的分类性能。

已有研究将特征筛选用于水稻品种分类及相关识别任务。Naik 等[9]采用基于方差分析的特征筛选方法开展水稻品种识别；Phan 和 Nguyen [10]则针对水稻种子纯度识别研究特征选择方法。在通用机器学习任务中，Bhuiyan 等[11]比较了 RF 重要性、置换特征重要性与统计筛选方法，并通过交叉验证评价特征缩减后的分类表现。本文针对去壳米粒的品种分类，比较不同规模特征子集与分类器的组合。根据特征选择研究对预测性能和重复入选情况的联合评价思路[12]，本文在比较分类性能的同时统计跨折入选频率，以识别具有较高选择一致性的候选指标。

基于上述分析，本文以去壳大米为研究对象，采用随机森林(Random Forest, RF)特征重要性对 106 个原始外观特征进行排序，围绕特征精简、分类性能与选择一致性开展研究。首先，构建不同规模的特征子集，在固定分类器参数下比较径向基函数核支持向量机(Support Vector Machine with a Radial Basis Function Kernel, RBF-SVM)、RF 和多层感知机(Multilayer Perceptron, MLP)的交叉验证表现，分析特征数量与分类性能的关系；其次，进一步比较全特征、PCA-13 和 RF Top-13 的分类表现，并引入 ANOVA F 检验和 L1 正则化逻辑回归作为特征选择基线，在 RBF-SVM 和 MLP 下评价不同筛选方法的表现；最后，结合特征组对比、跨折入选频率及候选特征的外观含义，分析所选指标的组成与选择一致性。

2. 数据与方法

2.1. 数据集与特征描述

本文采用包含 Arborio、Basmati、Ipsala、Jasmine 和 Karacadag 五个品种的大米外观特征数据集[2]。数据集共包含 75,000 个样本，每个品种 15,000 个，各类别样本数量相同。研究对象为去壳后的大米，每个样本由 106 个外观特征及对应的品种标签组成，分类任务为根据外观特征预测样本所属的品种。本文以已提取的数值特征作为模型输入，不直接对原始图像进行分类。

106 个外观特征包括 12 个形态特征、4 个形状特征和 90 个颜色特征，分别描述大米的尺寸与轮廓、几何比例及颜色分布，具体组成见表 1。该数据集已被用于水稻品种分类及特征筛选研究[9]。

Table 1. Composition of rice grain appearance features
表 1. 大米外观特征组成

特征类别	数量	代表性特征	描述内容
形态特征	12	面积、周长、主轴长度、偏心率、圆度	大米尺寸与轮廓属性
形状特征	4	形状因子 1~4：由大米面积、主轴长度与次轴长度构建的几何比例指标	大米几何形状
颜色特征	90	不同颜色空间中的均值、标准差、熵及小波统计量	颜色分布与变化特征
合计	106	-	-

2.2. 实验设计

本研究采用分层随机划分，将 85%的样本作为训练集，15%的样本作为测试集。在训练集内进行 5 折分层交叉验证，每折的标准化、PCA 变换和各特征选择方法均在当前训练折上拟合，并应用于对应验证折，全部实验划分随机种子为 42。最终评价使用完整训练集重新拟合特征处理流程和分类器，再计算测试集指标。

实验包括特征组对比、特征数量比较、特征表示对比和特征选择方法比较四个部分。特征组对比以逻辑回归(Logistic Regression, LR)为分类器，评价形态、形状、颜色及全部特征的分类表现，LR 采用固定参数，最大迭代次数为 2000，其余参数采用默认设置；特征数量比较在固定分类器参数下评价不同规模的 RF Top-K 子集；特征表示对比考察全特征、PCA-13 与 RF Top-13 在不同分类器下的分类表现；特征选择方法比较以 13 项原始特征为主要比较规模，评价 RF、ANOVA F 检验和 L1 正则化逻辑回归筛选方案，并参考文献[9]采用 ANOVA 筛选 40 项特征的设置，进一步比较 RF Top-40、ANOVA Top-40 及 L1-LR 非零特征集的分类表现。全特征与 PCA 方案通过训练集内网格搜索选择分类器参数，各原始特征筛选方案在同一分类器下采用一致的固定参数。

2.3. 特征处理与筛选

全特征方案保留全部 106 个原始变量。PCA 方案先对原始特征进行标准化，再提取压缩后的 13 个主成分，其累计解释方差比为 94.87%。其中，前三个主成分的解释方差比分别为 41.93%、16.26%和 9.17%，累计为 67.36%。随后进行白化处理，以统一各主成分的方差尺度。

RF 筛选采用基于节点不纯度下降的特征重要性对原始变量排序[13]。排序模型包含 400 棵决策树(Decision Tree, DT)，最大深度为 20，叶节点最小样本数为 1。在每个交叉验证训练折中独立拟合 RF 排序模型，并构建 Top-10、Top-13、Top-20、Top-30 和 Top-50 特征子集，以全部 106 个特征作为基线。其中，RF Top-13 用于与 PCA-13 进行同维度比较。筛选后，对 RBF-SVM 和 MLP 的输入进行标准化，RF 分类器直接使用所选原始特征。

ANOVA 筛选对各原始特征计算单因素方差分析 F 统计量，以品种间均值差异相对于品种内变异的程度评价单个特征的判别作用，按 F 值降序排列，分别保留前 13 项和前 40 项特征。

L1 正则化逻辑回归筛选首先对输入特征进行标准化，再采用一对其余策略建立分类模型，记为 L1-LR。模型采用 liblinear 求解器，正则化参数 C 固定为 0.1，收敛容差为 10^{-4}，最大迭代次数为 10,000。以各类别模型中特征系数绝对值之和作为排序分数，将分数大于 0 的全部特征组成 L1-LR 非零特征集，并按该分数降序选取前 13 项，构成 L1-LR Top-13 特征集。上述参数预先固定，不根据测试集结果调整。本次实验各训练折及完整训练集均具有不少于 13 项的非零系数特征。

最终 Top-13 特征名单由完整训练集上的 RF 重要性排序确定，跨折入选频率用于稳定性分析。

2.4. 分类模型与参数设置

全特征与 PCA 对照实验采用八种分类器，包括 RBF-SVM、MLP、RF、LR、DT、自适应提升(Adaptive Boosting, AdaBoost)、K 近邻(K-Nearest Neighbors, KNN)和高斯朴素贝叶斯(Gaussian Naive Bayes, GNB)。除 GNB 采用默认参数外，其余分类器均通过训练集内的 5 折分层交叉验证进行网格搜索，以宏平均 F1 均值选择参数。RBF-SVM 固定采用径向基函数核，惩罚参数 C 取 0.1 或 1.0，核参数 γ 取 scale 或 0.01。RF 的决策树数量取 200 或 400，最大深度取 20 或不限制，叶节点最小样本数取 1 或 3。MLP 的隐藏层结构取(64, 32)或(128, 64, 32)，正则化参数取 0.0001 或 0.001，初始学习率固定为 0.001，最大迭代次数为 1000，并启用早停。AdaBoost 采用最大深度为 3 的决策树作为基学习器，算法采用 SAMME，弱学习器数量取 100 或 200，学习率取 0.5 或 1.0。KNN 的邻居数量取 3、5 或 9，权重采用均匀权重或距离权重。LR 的正则化强度倒数参数 C 取 0.1、1.0 和 10.0，最大迭代次数设为 2000，其余参数采用默认设置。DT 的最大深度取 10 或不限制，叶节点最小样本数取 1 或 3。各模型采用相同的 5 折分层交叉验证划分，启用随机打乱。

不同 Top-K 特征子集的比较采用固定分类器参数。RBF-SVM 设置 C = 1.0、γ = scale；RF 设置 400

棵决策树、最大深度 20 和叶节点最小样本数 1；MLP 采用(128, 64, 32)的隐藏层结构、ReLU 激活函数、0.001 的正则化参数和初始学习率，最大迭代次数为 1000，并启用早停。同一分类器在不同特征数量下保持上述参数设置一致。

ANOVA、L1-LR 与 RF 筛选方法的比较采用 RBF-SVM 和 MLP 作为下游分类器，其参数与上述 Top-K 实验保持一致。

2.5. 评价指标与稳定性分析

本文以准确率和宏平均 F1 值评价分类性能。准确率表示正确分类样本占总样本的比例；宏平均 F1 值为五个品种 F1 值的算术平均，使各类别在评价中具有相同权重。测试集结果用于评价最终模型的分类表现。

采用跨折入选频率评价特征选择的一致性，计算方法为某特征在 5 个训练折中进入 Top-13 的次数除以 5。频率越高，表示该特征在不同训练折中重复入选的程度越高。同时，按形态、形状和颜色类别统计候选特征的组成，分析精简特征子集包含的外观信息。

3. 结果与分析

3.1. 不同外观特征组的分类表现

表 2 给出了以 LR 为分类器时，不同特征组的 5 折交叉验证结果。颜色特征的宏平均 F1 均值为 0.9889，高于形态特征的 0.9744 和形状特征的 0.9668，在三类单独特征组中表现最好。

联合使用形态、形状和颜色特征时，LR 的宏平均 F1 为 0.9988 ± 0.0002，高于各单独特征组，说明联合输入有助于提高 LR 的分类性能。采用 PCA-13 时，宏平均 F1 为 0.9974 ± 0.0004，表明 13 个主成分在压缩输入维度后仍能支持较高的分类性能。

Table 2. Five-fold cross-validation performance of LR with different feature groups and PCA-13
表 2. LR 在不同特征组及 PCA-13 表示下的 5 折交叉验证表现

特征组	特征数量	准确率(均值 ± 标准差)	宏平均 F1 (均值 ± 标准差)
全部特征	106	0.9988 ± 0.0002	0.9988 ± 0.0002
形态特征	12	0.9744 ± 0.0006	0.9744 ± 0.0006
形状特征	4	0.9668 ± 0.0023	0.9668 ± 0.0023
颜色特征	90	0.9889 ± 0.0014	0.9889 ± 0.0014
PCA 特征	13	0.9974 ± 0.0004	0.9974 ± 0.0004

3.2. PCA 对不同分类器的影响

表 3 显示，在各自参数选择方案下，采用 PCA-13 后，GNB 的测试集宏平均 F1 由 0.7693 提高至 0.9751；其余七种分类器的结果均低于全特征方案。

RBF-SVM 和 MLP 的宏平均 F1 分别由 0.9988 和 0.9986 降至 0.9981 和 0.9982，仍保持较高的分类性能。RF 由 0.9982 降至 0.9955，DT 由 0.9965 降至 0.9912，降幅较大。在 PCA-13 方案中，MLP 表现最好，RBF-SVM 次之。

3.3. 不同特征数量下的分类性能比较

表 4 列出了不同特征数量下三种分类器的五折交叉验证宏平均 F1，变化趋势如图 1 所示。随着特征

数量增加，三种分类器的性能总体提高。其中，RF 在 Top-10 条件下的宏平均 F1 最低，为 0.9947，增加至全特征后达到 0.9987，提升幅度最大。

Table 3. Test macro-F1 of classifiers with all features and PCA-13
表 3. 全特征与 PCA-13 条件下各分类器的测试集宏平均 F1

分类器	全特征	PCA-13	变化(百分点)
RBF-SVM	0.9988	0.9981	−0.07
MLP	0.9986	0.9982	−0.04
RF	0.9982	0.9955	−0.27
AdaBoost	0.9987	0.9964	−0.23
LR	0.9982	0.9974	−0.08
KNN	0.9970	0.9956	−0.14
DT	0.9965	0.9912	−0.53
GNB	0.7693	0.9751	+20.58

Table 4. Five-fold cross-validation macro-F1 scores for different numbers of RF-selected features
表 4. 不同 RF 筛选特征数量下的 5 折交叉验证宏平均 F1

特征数	RBF-SVM	MLP	RF
10	0.9960	0.9973	0.9947
13	0.9974	0.9978	0.9966
20	0.9980	0.9980	0.9975
30	0.9984	0.9981	0.9981
50	0.9989	0.9988	0.9986
106	0.9991	0.9988	0.9987

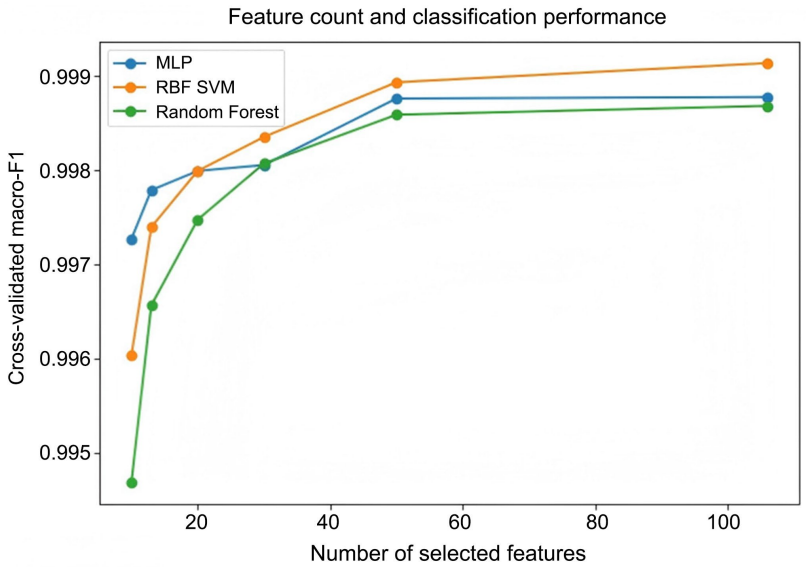


Figure 1. Cross-validation performance of three classifiers with different numbers of RF-selected features
图 1. 不同 RF 筛选特征数量下三种分类器的交叉验证性能

在 Top-10 和 Top-13 条件下，MLP 的宏平均 F1 分别为 0.9973 和 0.9978，均高于另外两种分类器；在 Top-50 和全特征条件下，RBF-SVM 分别达到 0.9989 和 0.9991，高于 MLP 和 RF。从 Top-50 增加至全特征时，三种分类器的性能提升均较小，MLP 的宏平均 F1 在四位小数精度下保持为 0.9988。

3.4. 不同特征处理方法的分类性能比较

表 5 和表 6 分别从特征表示与特征选择方法两个层面比较分类性能。表 5 比较全特征、PCA-13 与 RF Top-13 的分类表现；表 6 以 13 项原始特征为主要比较规模，比较 RF、ANOVA 和 L1-LR 三种筛选方法，同时报告 RF Top-40、ANOVA Top-40 及 L1-LR 非零特征集的结果。全特征与 PCA 方案分别通过训练集内网格搜索选择参数，各原始特征筛选方案在同一分类器下采用一致的固定参数。因此，表 5 反映各方案在所采用参数设置下的表现，表 6 用于比较统一分类器设置下不同特征选择方法的表现。

Table 5. Independent-test macro-F1 under full features, PCA-13, and random-forest Top-13
表 5. 全特征、PCA-13 与 RF Top-13 下的独立测试集宏平均 F1

分类器	全特征	PCA-13	RF Top-13
RBF-SVM	0.9988	0.9981	0.9977
MLP	0.9986	0.9982	0.9981
RF	0.9982	0.9955	0.9969

由表 5 可见，在两种 13 维输入方案中，RBF-SVM 采用 PCA-13 时的宏平均 F1 为 0.9981，高于 RF Top-13 的 0.9977；MLP 采用两种方案时分别为 0.9982 和 0.9981，结果接近；RF 分类器采用 RF Top-13 时达到 0.9969，高于 PCA-13 的 0.9955。不同分类器对特征表示的响应存在差异，PCA 表示与原始特征筛选各有适用情形。

RF Top-13 结合 MLP 的测试集宏平均 F1 为 0.9981，接近全特征 MLP 的 0.9986，同时将输入特征数量由 106 项减少至 13 项，降幅为 87.7%。与 PCA 生成的线性组合相比，RF 筛选保留了原始指标的名称，便于结合其计算定义讨论米粒的几何属性和颜色统计特征。

Table 6. Test-set Macro-F1 scores of different feature selection methods
表 6. 不同特征选择方法的测试集宏平均 F1

特征选择方法	最终特征数	RBF-SVM	MLP
RF Top-13	13	0.9977	0.9981
ANOVA Top-13	13	0.9925	0.9959
L1-LR Top-13	13	0.9980	0.9977
RF Top-40	40	0.9986	0.9983
ANOVA Top-40	40	0.9983	0.9985
L1-LR 非零特征集	77	0.9988	0.9987

表 6 显示，在保留 13 项原始特征时，采用 RBF-SVM 的 L1-LR、RF 和 ANOVA 方案的测试集宏平均 F1 分别为 0.9980、0.9977 和 0.9925，L1-LR 方案的结果最高；采用 MLP 时，RF、L1-LR 和 ANOVA 方案分别为 0.9981、0.9977 和 0.9959，RF 方案的结果最高。因此，在相同特征数量下，筛选方法的性能排序随下游分类器而变化。

交叉验证结果显示，L1-LR Top-13 结合 RBF-SVM 的宏平均 F1 为 0.9983 ± 0.0005，高于 RF Top-13 的 0.9974 ± 0.0006；结合 MLP 时，两者分别为 0.9979 ± 0.0005 和 0.9978 ± 0.0006。L1-LR 与 RF 在 MLP

下的交叉验证均值较为接近，其排序与测试集结果有所不同，说明不能仅依据单次测试中的微小差异认定某一种筛选方法具有稳定优势。

当保留 40 项特征时，RF 和 ANOVA 方案的测试集表现较为接近。采用 RBF-SVM 时，RF 方案的宏平均 F1 为 0.9986，高于 ANOVA 的 0.9983；采用 MLP 时，ANOVA 方案为 0.9985，高于 RF 的 0.9983。L1-LR 非零特征集结合 RBF-SVM 和 MLP 分别获得 0.9988 和 0.9987，但其最终保留 77 项特征，输入规模明显大于 13 项和 40 项方案。因此，对不同方案的评价需要同时考虑分类性能与特征数量。

文献[9]采用 ANOVA 筛选 40 项特征并结合 SVM，报告测试准确率为 99.81%；本文 ANOVA Top-40 结合 RBF-SVM 在当前实验设置下的测试准确率为 99.83%。本文进一步考察了 13 项原始特征条件下不同筛选方法的分类表现，以及 RF 筛选指标的跨折入选一致性。

综上所述，RF Top-13 结合 MLP 能够以较少的原始特征保持较高分类性能，是本实验条件下兼顾输入精简与原始指标解释的可行方案。

3.5. 特征选择稳定性与外观含义

Table 7. Definitions, appearance interpretations, and cross-fold selection frequencies of RF Top-13 features
表 7. RF Top-13 特征的定义、外观含义及跨折入选频率

类别	特征	定义或计算依据	外观含义	5 折入选频率
形态	ASPECT_RATIO	L/l	长短轴比例，描述米粒细长程度	100%
形态	COMPACTNESS	$\sqrt{4A/\pi}/L$	等效圆直径与长轴之比，描述形状紧致程度	100%
形态	ECCENTRICITY	$\sqrt{1-(l/L)^2}$	等效椭圆偏心率，数值越大越细长	100%
形态	MAJOR_AXIS	米粒区域等效椭圆的长轴长度 L	图像尺度下的纵向尺寸	100%
形态	ROUNDNESS	$4\pi A/P^2$	基于面积与周长的轮廓圆整程度	100%
形状	SHAPEFACTOR_2	l/A	短轴与面积的比例，受形状和尺度共同影响	80%
形状	SHAPEFACTOR_3	$4A/(\pi L^2)$	面积相对于以长轴为直径的圆面积的比例	100%
颜色	StdDevB	数据集 Lab-B 通道的像素值标准差 s_B	对应颜色通道的离散程度	100%
颜色	StdDevCb	YCbCr-Cb 通道像素值的标准差 s_{Cb}	蓝色色差分量的离散程度	100%
颜色	entropyB	数据集 Lab-B 通道的熵特征	对应颜色通道的统计描述量	100%
颜色	entropyCr	数据集 YCbCr-Cr 通道的熵特征	对应红色色差通道的统计描述	100%
颜色	entropyRR	数据集 RGB 红色通道的熵特征	对应红色通道的统计描述	100%
颜色	ALLdaub4b	数据集 Lab-B 通道的 db4 小波分解描述量	对应颜色通道经小波处理后的描述量	40%

注：A 为米粒区域面积，P 为周长，L 和 l 分别为等效椭圆的长轴与短轴长度，均按图像像素尺度计算。s_c 为通道 c 的像素值标准差，按公式 $s_c=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{i,c}-\bar{x}_c)^2}$ 计算，其中 N 为米粒区域像素数，$x_{i,c}$ 为第 i 个像素在通道 c 上的取值，$\bar{x}_c$ 为该通道的像素均值。几何特征及颜色标准差的计算定义、颜色特征的名称与通道归属参考文献[3]，特征字段名称沿用数据集原始标记。B 表示数据集标记的 Lab-B 通道，Cb 和 Cr 分别表示 YCbCr 的蓝色色差与红色色差通道，RR 表示 RGB 红色通道；ALLdaub4b 为文献所述的 db4 小波分解描述量。跨折入选频率为该特征进入 5 个训练折 RF Top-13 子集的次数除以 5。

表 7 列出了完整训练集上 RF 最终筛选出的 13 项特征，包括 5 项形态特征、2 项形状特征和 6 项颜色特征。其中 11 项在 5 个训练折中均进入 Top-13；SHAPEFACTOR_2 和 ALLdaub4b 分别入选 4 次和 2 次，入选频率为 80%和 40%。这表明多数最终入选特征具有较高的跨折选择一致性，但部分特征的入选仍随训练样本划分变化。

4. 结论

针对大米品种外观识别中原始指标较多、降维后特征含义不明确以及特征表示与分类器适配关系不清等问题，本文采用随机森林对原始外观指标进行重要性排序和子集筛选，并将筛选结果与 PCA 降维表示、ANOVA F 检验及 L1 正则化逻辑回归筛选结果进行比较，同时分析不同特征组的判别作用及候选指标的跨折稳定性。综合研究结果，基于原始外观指标的筛选策略能够在精简输入的同时保持较高分类性能，更适合构建可解释的大米品种识别方案。稳定入选的指标主要反映米粒的几何结构与颜色分布，说明品种差异并非由单一外观维度决定，而是由尺寸、轮廓和颜色统计信息共同表征。颜色信息具有较强的区分能力，联合形态与形状信息能够进一步提高分类性能，体现了三类外观信息的联合利用价值。不同分类器对特征表示的响应存在差异，因而特征选择应与分类器协同确定，不能仅依据降维后的累计方差判断方案优劣。综合性能、变量规模和可解释性，随机森林筛选后的原始指标与 MLP 的组合更适合作为后续识别模型的基础，并可为特征采集、指标体系构建和模型部署提供依据。

参考文献

[1] Aznan, A., Gonzalez Viejo, C., Pang, A. and Fuentes, S. (2021) Computer Vision and Machine Learning Analysis of Commercial Rice Grains: A Potential Digital Approach for Consumer Perception Studies. *Sensors*, **21**, Article 6354. https://doi.org/10.3390/s21196354

[2] Koklu, M., Cinar, I. and Taspinar, Y.S. (2021) Classification of Rice Varieties with Deep Learning Methods. *Computers and Electronics in Agriculture*, **187**, Article ID: 106285. https://doi.org/10.1016/j.compag.2021.106285

[3] Cinar, I. and Koklu, M. (2021) Determination of Effective and Specific Physical Features of Rice Varieties by Computer Vision in Exterior Quality Inspection. *Selcuk Journal of Agricultural and Food Sciences*, **35**, 229-243. https://doi.org/10.15316/sjafs.2021.252

[4] Çinar, İ. and Koklu, M. (2021) Identification of Rice Varieties Using Machine Learning Algorithms. *Tarım Bilimleri Dergisi*, **28**, 307-325. https://doi.org/10.15832/ankutbd.862482

[5] Coronel-Reyes, J., Haro-Sarango, A., Delgado-Vera, C. and Triviño-Sánchez, J. (2026) Comparative Evaluation of Machine Learning Algorithms for the Identification and Morphological Classification of Rice Grains. *AgriEngineering*, **8**, Article 100. https://doi.org/10.3390/agriengineering8030100

[6] Li, X., Wang, Y., Monday, H.N. and Nneji, G.U. (2025) A Novel Residual Learning of Multi-Scale Feature Extraction Model for the Classification of Rice Grain Varieties. *Computers and Electronics in Agriculture*, **237**, Article ID: 110491. https://doi.org/10.1016/j.compag.2025.110491

[7] Islam, M.M., Himel, G.M.S., Moazzam, M.G. and Uddin, M.S. (2025) Artificial Intelligence-Based Rice Variety Classification: A State-of-the-Art Review and Future Directions. *Smart Agricultural Technology*, **10**, Article ID: 100788. https://doi.org/10.1016/j.atech.2025.100788

[8] Sonnadara, D., Khuat, T.T., Musial-Gabrys, K. and Gabrys, B. (2026) Interpretable Automated Feature Engineering: A Comprehensive Review with a Focus on Dynamic and Stationary Environments. *Artificial Intelligence Review*, **59**, Article No. 165. https://doi.org/10.1007/s10462-026-11573-y

[9] Naik, N.K., Subbarao, M.V., Sethy, P.K., Behera, S.K. and Panigrahi, G.R. (2024) Machine Learning with Analysis-of-Variance-Based Method for Identifying Rice Varieties. *Journal of Agriculture and Food Research*, **18**, Article ID: 101397. https://doi.org/10.1016/j.jafr.2024.101397

[10] Phan, T. and Nguyen, L.H.B. (2025) Enhancing Rice Seed Purity Recognition Accuracy Based on Optimal Feature Selection. *Ecological Informatics*, **86**, Article ID: 103044. https://doi.org/10.1016/j.ecoinf.2025.103044

[11] Bhuiyan, A.H., Tinny, T.K. and Uddin, M.F. (2026) Assessing the Effectiveness of Feature Selection Methods in Machine Learning Using Statistical Metrics. *Egyptian Informatics Journal*, **34**, Article ID: 100949. https://doi.org/10.1016/j.eij.2026.100949

[12] Elmakias, I., Kolsky, D. and Vilenchik, D. (2026) Feature Stability as a Trust Layer for Feature Selection: Resampling-Based Recurrence Profiles beyond Predictive Performance. *Mathematics*, **14**, Article 2372. https://doi.org/10.3390/math14132372

[13] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324