

# Supplemental Information for Sorting Polyolefins with Near-Infrared Spectroscopy: Identification of optimal data analysis pipelines and machine learning classifiers

Bradley P. Sutliff<sup>1</sup>, Peter A. Beaucage<sup>2</sup>, Debra J. Audus<sup>1</sup>, Sara V. Orski<sup>1</sup>, and Tyler B. Martin<sup>1</sup>

<sup>1</sup>Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, United States

<sup>2</sup>NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, MD, 20899, United States

July 2024

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

Table S1: Abbreviations used in this SI

Abbreviation	Expansion
PE	Polyethylene
PP	Polypropylene
HDPE	High-Density Polyethylene
LDPE	Low-Density Polyethylene
LLDPE	Linear Low-Density Polyethylene
MDPE	Medium-Density Polyethylene
NIR	Near-Infrared
PLS-DA	Partial Least Squares Discriminant Analysis
SIMCA	Soft Independent Modelling by Class Analogy
SVM/SVC	Support Vector Machines/Support Vector Classifiers
KNN	k-Nearest Neighbors
RF/RFC	Random Forest (Classifier)
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
MLP(C)	Multilayer Perceptron (Classifier)
RBF	Radial Basis Function
GNB	Gaussian Naive Bayes

Table S2: Summary of the polymer state (pellet/powder) and presence of colorant for the commercial polymers. All other polymer samples were uncolored pellets.

	HDPE	MDPE	LDPE	LLDPE	PP	PP-co-PE
Uncolored Pellet	2	0	9	3	1	2
Colored Pellet	2	0	2	0	2	0
Uncolored Powder	8	0	0	0	0	0
Colored Powder	0	0	0	0	0	0

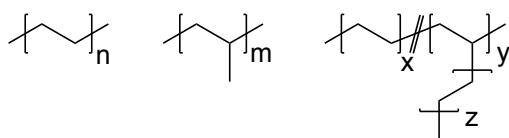


Figure S1: Reference skeletal structures of HDPE, PP, and LDPE, respectively. For LLDPE,  $z$  is less than 3.

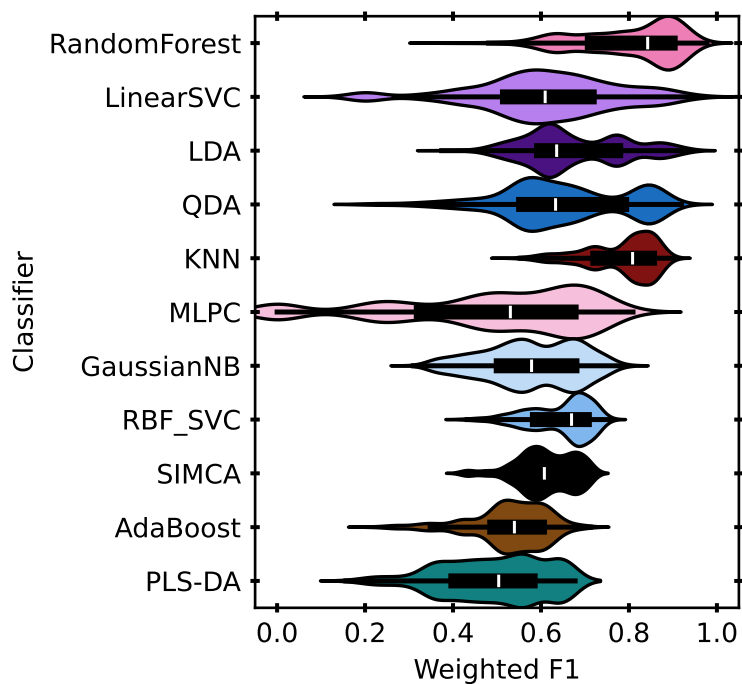


Figure S2: A violin plot showing the distribution of weighted F1 scores for each classifier. Center white dots indicate the median score, thick black boxes show the interquartile range, and black “whiskers” extend to 1.5 times the interquartile range. More details for the classifiers are provided in Tables S3 through S5, and Section 3.2 and Table 2 of the corresponding manuscript.

Table S3: Default hyperparameters and sources for the classifiers used in the initial classification study using all 1152 preprocessing pipelines. While listed values are the defaults, the arguments that were modified for the manuscript are identified in **bold**.

Common Name	Classifier	Source	Parameters
PLS-DA	PLSRegression	scikit-learn 1.2	<b>n_components=2</b> scale=True max_iter=500 tol=1e-06 copy=True
LDA	LinearDiscriminantAnalysis	scikit-learn 1.2	solver='svd' shrinkage=None priors=None n_components=None store_covariance=False tol=0.0001 covariance_estimator=None
QDA	QuadraticDiscriminantAnalysis	scikit-learn 1.2	priors=None reg_param=0.0 store_covariance=False tol=0.0001
LinearSVC	svm.LinearSVC	scikit-learn 1.2	penalty='l2' loss='squared_hinge' dual=True tol=0.0001 C=1.0 multi_class='ovr' fit_intercept=True intercept_scaling=1 class_weight=None verbose=0 random_state=None max_iter=1000
RBF_SVC	svm.SVC	scikit-learn 1.2	C=1.0 kernel='rbf' degree=3 gamma='scale' coef0=0.0 shrinking=True probability=False tol=0.001 cache_size=200 class_weight=None verbose=False max_iter=-1 decision_function_shape='ovr' break_ties=False random_state=None
SIMCA	SIMCA_classifier	custom code adapted from [1, 2]	n_components=3 alpha=0.05 simca_type='SIMCA'

Table S4: Continuation of the default hyperparameters and sources for the classifiers used in the initial classification study using all 1152 preprocessing pipelines. While listed values are the defaults, the arguments that were modified for the manuscript are identified in **bold**.

Common Name	Classifier	Source	Parameters
KNN	KNeighborsClassifier	scikit-learn 1.2	n_neighbors=5 weights='uniform' algorithm='auto' leaf_size=30 p=2 metric='minkowski' metric_params=None n_jobs=None
GaussianNB	GaussianNB	scikit-learn 1.2	priors=None var_smoothing=1e-09
MLPC	MLPClassifier	scikit-learn 1.2	<b>max_iter=200</b> hidden_layer_sizes=(100,) activation='relu' solver='adam' alpha=0.0001 batch_size='auto' learning_rate='constant' learning_rate_init=0.001 power_t=0.5 shuffle=True random_state=None tol=0.0001 verbose=False warm_start=False momentum=0.9 nesterovs_momentum=True early_stopping=False validation_fraction=0.1 beta_1=0.9 beta_2=0.999 epsilon=1e-08 n_iter_no_change=10 max_fun=15000
RandomForest	RandomForestClassifier	scikit-learn 1.2	n_estimators=100 criterion='gini' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features='sqrt' max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=True oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False class_weight=None ccp_alpha=0.0 max_samples=Non

Table S5: Continuation of the default hyperparameters and sources for the classifiers used in the initial classification study using all 1152 preprocessing pipelines. While listed values are the defaults, the arguments that were modified for the manuscript are identified in **bold**.

Common Name	Classifier	Source	Parameters
AdaBoost	AdaBoostClassifier	scikit-learn 1.2	estimator=None n_estimators=50 learning_rate=1.0 algorithm='SAMME.R' random_state=None base_estimator='deprecated'