Optimization of Future Multifilter Surveys Toward Asteroid Characterization

Hanna Klimczak[1], Emil Wilawer[1], Tomasz Kwiatkowski[1] , Agnieszka Kryszczyńska[1], Dagmara Oszkiewicz[1],
Wojciech Kotłowski[2], and Francesca DeMeo[3]

[1] Astronomical Observatory Institute Faculty of Physics, Adam Mickiewicz University Słoneczna 36, 60-286 Poznań, Poland
[2] Institute of Computing Science Poznań, University of Technology ul. Piotrowo, 60-965 Poznań, Poland
[3] Department of Earth, Atmospheric and Planetary Sciences Massachusetts Institute of Technology 77 Massachusetts Avenue 54-427, Cambridge, MA 02139, USA

Abstract

The aim of this paper is to find a set of photometric passbands that will give optimal results for spectrophotometric classification of asteroids into taxonomic types and classes. For this purpose various machine-learning methods are used, namely multinomial logistic regression, naive Bayes, support vector machines, gradient boosting, and multilayer perceptrons. Sequential feature selection is performed to assess the contribution of each reflectance difference. We find that to determine the taxonomic complexes with a balanced accuracy of 85%, a set of five spectrophotometric bands is required. For taxonomy type determination with the balanced accuracy of 80% a set of eight bands is necessary. Furthermore, only a three-band system is enough for distinguishing the C-complex asteroids with 92% balanced accuracy. These results can be used for designing future asteroid multifilter sky surveys.

Unified Astronomy Thesaurus concepts: Asteroid surfaces (2209); Spectroscopy (1558); Multi-color photometry (1077)

1. Introduction

Taxonomic classifications based on color indices from sky surveys have been commonly used in past studies (Sykes et al. 2000; Ivezić et al. 2002; Popescu et al. 2016, 2018; Peña et al. 2020; Morate et al. 2021; Sergeyev & Carry 2021; Sergeyev et al. 2022). Carvano et al. (2010) created even a separate taxonomic system based on the SDSS color data. A multitude of studies were performed in the color space. Jedicke et al. (2004) studied the color—age relation of asteroid families from the S complex and studied the space weathering mechanism. Szabó et al. (2007) explored the colors of Jupiter Trojans and found that their color distribution is significantly different than that of the asteroids in the main belt. They also showed the leading swarm contains twice as much objects as the trailing one, which is consistent with stability studies. Parker et al. (2008) studied the size distribution of asteroid families using the color information. Rivkin (2012) explored the number of hydrated C-complex asteroids in the SDSS data. Several studies searched for V-type asteroids in the solar system in the relation to the missing mantle problem based on the color indices (Moskovitz et al. 2008; Oszkiewicz et al. 2014; Licandro et al. 2017). DeMeo et al. (2014) looked at the D type and DeMeo et al. (2019) at the A-type distribution across the main belt and explored their origins and implications to the solar system evolution. Moskovitz (2012) researched the colors of asteroid pairs in relation to their formation scenarios. Colors are also often used to classify near-Earth objects as they are easier to obtain than full spectra (Vaduvescu et al. 2017; Hromakina et al. 2021). Clearly, colors are one of the most commonly used parameters in solar system science exploration.

Furthermore, future sky surveys are expected to produce a large number of asteroid multifilter photometry, thus colors.

The Legacy Survey of Space and Time (LSST) located in Vera C. Rubin Observatory in Chile will be the largest wide-field ground telescope ever built. With a mirror's diameter of 8.4 m, this telescope will have an extremely wide field of view and will be capable of covering the entire sky in just 3 days. During the first 10 yr of operation, observing around 30,000 deg^2 of the southern sky in wavelength range 320–1080 μm with a set of six filters (u, g, r, i, z, y) it will discover around 6 millions solar system small bodies and provide over a billion photometric and astrometric measurements (LSST Science Collaboration 2009).

The Panoramic Survey Telescope and Rapid Response System, located at the Haleakala Observatory in Hawaii, began observations in 2010 and continues to collect data today. The observations are obtained using a set of 5 broadband filters g_{P1}, r_{P1}, i_{P1}, z_{P1}, y_{P1} and occasionally with sixth wide filter w_{P1} which spans over g_{P1}, r_{P1}, i_{P1} bands. Thanks to this mission, 129 new comets, 2900 new near-Earth asteroids (NEA), and over 100,000 main-belt asteroids have been discovered. In addition, for a quarter of a million main-belt asteroids, photometric properties have been obtained (Chambers et al. 2016).

Euclid is an ESA mission devoted to cosmology that is supposed to be launched in the second half of 2023 and to last 6 yr. Although focused mainly on dark matter, dark energy, and gravitational lensing, it will observe about 150,000 solar system objects, located mainly in the main belt. Euclid will be equipped with visual imaging channel operating in a large visible photometry band and near-infrared spectrometer and photometer operating in Y, J, H bands covering 1.1–1.85 μm wavelengths (Carry 2018).

In this work we search for an optimal set of photometric bands to be used for determination of the asteroid taxonomies and replace the more time-consuming spectroscopic characterizations. From many existing taxonomic schemes, we choose the most widely used BusDeMeo taxonomy (DeMeo et al. 2009), which is based on the asteroid reflectance spectra, $R(\lambda)$.

Since our work is based on machine-learning methods, we use some terminology characteristic for data science. One of the most basic is a "feature": a measurable property of a data set to be analyzed. Features can be selected manually, based on prior knowledge of the data set, or automatically, with one of the machine-learning algorithms such as Sequential Feature Selection (Whitney 1971).

In Klimczak et al. (2021) we explored the performance of different machine-learning algorithms (multinomial logistic regression, naive Bayes, support vector machines, gradient boosting, multilayer perceptron) in the taxonomical classification of asteroids. The analysis was performed with two different parameter sets: one was the principal component directions (PCA), and the other was based on the reflectance values $R(\lambda)$ spaced every 0.05 μm in the 0.45–2.5 μm range. We also utilized the overal spectral slope. We found that multilayer perceptron and Stochastic Gradient Descent solver performed the best. Classification based on the reflectance values turned out more efficient than that based on PCA. Furthermore, we studied how many and which spectral features contribute the most to the classification task. It turned out that the top five features (spectral slope and reflectance at 1.05, 0.9, 0.65, and 1.1 μm) are enough to obtain high accuracy (93%) for the prediction of complexes and six features (spectral slope and reflectance at 1.4, 1.05, 0.9, 0.95, and 0.65 μm) to obtain (81%) accuracy for taxonomic types. Those results, however, assumed we are able to accurately determine the spectral slope, which requires several additional photometric bands.

For this reason, in Klimczak et al. (2022) we decided to accept more realistic approach and use the photometric bands already present in current and near-future sky surveys. We confirmed the superiority of multilayer perceptron in the classification of asteroids. Photometric systems of two surveys, Euclid and VISTA, reached the highest scores in classification of complexes (93%) and types (85%). Those results, however, were obtained with the assumption that an additional filter (with the central wavelength of $\lambda_\mathrm{n} = 0.55\,\mu$m) is added for normalization. A lack of such filter can result in lower accuracies. Before that Hietala (2020) studied the performance of neural networks in asteroid classification using LSST photometric filters. It was followed by a paper by Penttilä et al. (2022), in which a two-layer fully connected neural network was able to predict the Bus-DeMeo taxonomic types with up to 90.1% accuracy. The results of this work show that photometric filters are a promising data type for automatic classification.

In the present work, which is our third paper discussing the determination of asteroid taxonomies and classes from spectrophotometry, we decided to change the way of feature selection. Instead of scanning all $R(\lambda)$ and letting the machine-learning algorithms to select the most significant ones (as in (Klimczak et al. 2021), or limiting our choices to the bands already present in the sky surveys (as in Klimczak et al. 2022), we based our analysis on specific reflectance values that are connected with the features used to define asteroid taxonomy types and complexes, as described in Vilas & Gaffey (1989), Vilas (1994), Rivkin et al. (2002), DeMeo et al. (2009), and Binzel et al. (2019). In Section 2 we discuss the data, while Section 3 describes the parameterization and methodology used. In Section 4 we present our results. Conclusions are in Section 5.

2. Data

The data used for training the prediction models in this work consisted of reflectance spectra that in theory can be obtained by dividing the spectrum of the asteroid by the spectrum of the Sun. In practice this is almost impossible to be done so instead of the Sun, a spectrum of a carefully selected "solar analog" star is acquired on the same night, at similar airmass. As a result of the division we obtain a plot of reflectances R versus wavelength λ which show what percent of the incident solar radiation is reflected toward the observer at a given wavelength. The reflectance spectra are then normalized to unity at some selected reference wavelength λ_n. This step is necessary since the asteroid spectra are usually not flux-calibrated. Additionally, the flux reflected by the asteroid surface changes with its rotation while the flux of the solar analog is constant in time. The choice of the normalization wavelength, λ_n, is usually done in such spectral region were there are no absorption lines. Most authors (including DeMeo et al. 2009) choose the effective wavelength of the Johnson V band, $\lambda_\mathrm{eff} = 0.55\,\mu$m. Those who use the SDSS asteroid colors normalize at the r' band ($\lambda_\mathrm{eff} = 0.623\,\mu$m), because V-band magnitudes are not available in the SDSS survey. However, DeMeo et al. (2014) chose to normalize at the g' ($\lambda_\mathrm{eff} = 0.467\,\mu$m) band. In this work we are not limited by any particular filter set (we try to find the optimal one) so we will follow the DeMeo et al. (2009) convention and normalize at $\lambda_\mathrm{n} = 0.55\,\mu$m.

As we use the asteroid taxonomy based on the reflectance spectra we assume that the asteroid magnitudes observed by a sky survey are first converted to relative reflectances. This can be done with a formula:

$$R_\lambda = 10^{-0.4(C_\mathrm{ast} - C_\odot)}, \tag{1}$$

where C_ast is the color index of the asteroid ($C_\mathrm{ast} = m_\lambda - m_{\lambda_\mathrm{n}}$), normalized to unity at λ_n, and $C_\odot$ is the same color index, derived for the Sun.

For the analysis the same reflectance spectra of 504 objects as in Klimczak et al. (2021) were used. Precisely, we obtained the data set, consisting of 371 spectra from DeMeo et al. (2009) and 195 spectra from Binzel et al. (2019) that were available both in the visible and near-infrared ranges, giving a total of 0.45 μm–2.45 μm spectral range. Out of those, we removed the types that were highly underrepresented, meaning they contained less than 10 objects, which resulted in 504 objects split into 12 taxonomic types. The detailed description of objects per complex, taxonomic type, and data source is presented in Table 1.

The processing of the reflectance spectra was initiated by smoothing each spectrum with the Savitzky-Golay filter to decrease its noise without distorting the tendency of the signal. To understand how this filter works, one can think of the seven-point moving average filter, in which—for a set of the seven points—a straight line is fitted with least squares. Next, a signal for the middle point (which is the fourth point in our example) is replaced by the value taken from the fitted line. Finally, the seven-point "window" is advanced by one point, and the whole procedure is repeated. In the Savitzky-Golay approach, the fitting is done with higher-order polynomials, and the polynomial that best fits the data (in terms of the least squares) is used to obtain a new value for the middle point of the window.

Table 1
Number of Objects for Each Taxonomic Complex and Type used for Classification per Data Source

Complex	Type	DeMeo	Binzel	All
S	S	143	55	198
S	Sq	29	27	56
S	Sr	22	16	38
EM	L	22	8	30
EM	V	17	12	29
EM	D	16	3	19
EM	K	16	0	16
EM	Q	8	31	39
C	Ch	18	1	19
C	C	13	13	26
X	Xk	18	4	22
X	X	4	8	12

Note. DeMeo—(DeMeo et al. 2009), Binzel—(Binzel et al. 2019).

In the next step, each of the smoothed reflectance spectrum was fitted with the cubic spline (see Figure 1) from which we obtained reflectance values at the wavelengths used to map known features in the asteroid spectra. Those features, in a form of absorption bands, continuum slopes and bumps were identified in previous studies of the asteroid spectra and they listed in Table 2 (note that each feature is defined by two or three wavelengths). To that we also added reflectances extracted at the effective wavelengths λ_{eff} of the filters used by the SDSS, Vista, and Euclid surveys (see last rows of Table 2).

At this stage we had a choice of either converting the pairs of reflectances to color indices (expressed in magnitudes) or to compute reflectance differences $R(\lambda_{\mathrm{begin}}) - R(\lambda_{\mathrm{mid}})$, $R(\lambda_{\mathrm{mid}}) - R(\lambda_{\mathrm{end}})$ and $R(\lambda_{\mathrm{begin}}) - R(\lambda_{\mathrm{end}})$ (if no mid value for the feature was given). We chose the latter because we wanted to stay as close to the original spectra as possible. What is more, convertion to magnitudes requires a specification of the filters' passbands which would limit us to use the already defined photometric systems.

As a result we ended with a set of 31 reflectance differences, which we used in further analysis. Their positions on the spectrum are presented in the Figure 1.

3. Methodology

In this work we used five machine-learning algorithms: multinomial logistic regression, naive Bayes, support vector machines (SVM), gradient boosting, as well as multilayer perceptron. The selection of methods followed the methodology from Klimczak et al. (2021), where the detailed descriptions for each method are presented.

3.1. Feature Selection

The importance of each reflectance difference was quantified in the process of sequential feature selection (Whitney 1971). Starting with an empty feature set, each feature was independently added to the subset, which was used to train separate models. The results for models trained on each feature subset were then compared. The feature for which the performance improvement was the highest, was permanently added to the subset. In the following steps, the procedure was repeated for the remaining features, until the selected size of the feature subset was reached. This allowed to obtain feature

ranking, where the lower the average position, the earlier in the process this feature was selected, indicating higher importance in the decision process. It is important to mention that there is a possibility of cross-correlation between the features, which may affect the ability to find the best subset of features.

3.1.1. Feature Selection per Complex

We also decided to evaluate the importance of the reflectance differences for each complex separately, i.e., how a given feature contributes to discriminating a given complex from the remaining ones. This was obtained by the so called "one versus all" method, which works as follows. We perform four separate binary classification experiments, one for each complex. In a given experiment, the selected complex constitutes a "positive" class, while the remaining complexes are merged into a "negative" class. Then, the feature selection procedure described in the previous section is performed on such two-class data set until a target number of selected features (set to 10) is reached. This allows us to obtain the 10 most important features for each complex.

3.2. Evaluation Metrics

For our experiments, we are reporting the following evaluation metrics: accuracy Acc, balanced prediction accuracy BAcc, F1 score, and Matthews correlation coefficient (MCC; Kelleher et al. 2015). They are computed as follows:

$$\mathrm{Acc} = \sum_{i=1}^{k} \frac{TP_i}{N},$$

and balanced prediction accuracy:

$$\mathrm{BAcc} = \frac{1}{k}\sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i},$$

$$F_1 = \frac{1}{k}\sum_{i=1}^{k} \frac{TP_i}{TP_i + \frac{1}{2}(FP_i + FN_i)},$$

$$\mathrm{MCC} = \frac{N \times \sum_{i=1}^{k} TP_i - \sum_{i}^{k}(TP_i + FP_i) \times (TP_i + FN_i)}{\sqrt{\left(N^2 - \sum_{i}^{k}(TP_i + FP_i)^2\right) \times \left(N^2 - \sum_{i}^{k}(TP_i + FN_i)^2\right)}}$$

where k is the number of classes, N is the total number of all samples, TP_i ("true positives") is the number of correctly classified objects from class i, FP_i ("false positives") is the number of objects incorrectly classified to class i, and FN_i ("false negatives") the number of incorrectly classified objects from class i.

These metrics are selected in alignment with Klimczak et al. (2021) and Klimczak et al. (2022). The straightforward interpretability of prediction accuracy (Acc) was the reason for selecting this metric, whereas other metrics commonly used in the case of imbalanced problems, such as balanced accuracy (BAcc), F1 score and MCC (Kelleher et al. 2015), are also measured to accommodate for class imbalance.

3.3. Experimental Setup

The experimental setup was equivalent to that described in Klimczak et al. (2021). That is, our results are based on an average of 10 runs of five-fold cross validation to compensate for the randomness of train/test splits. In each of the folds, the external parameters of each method were tuned by splitting the training fold further into two parts. Separate models for each
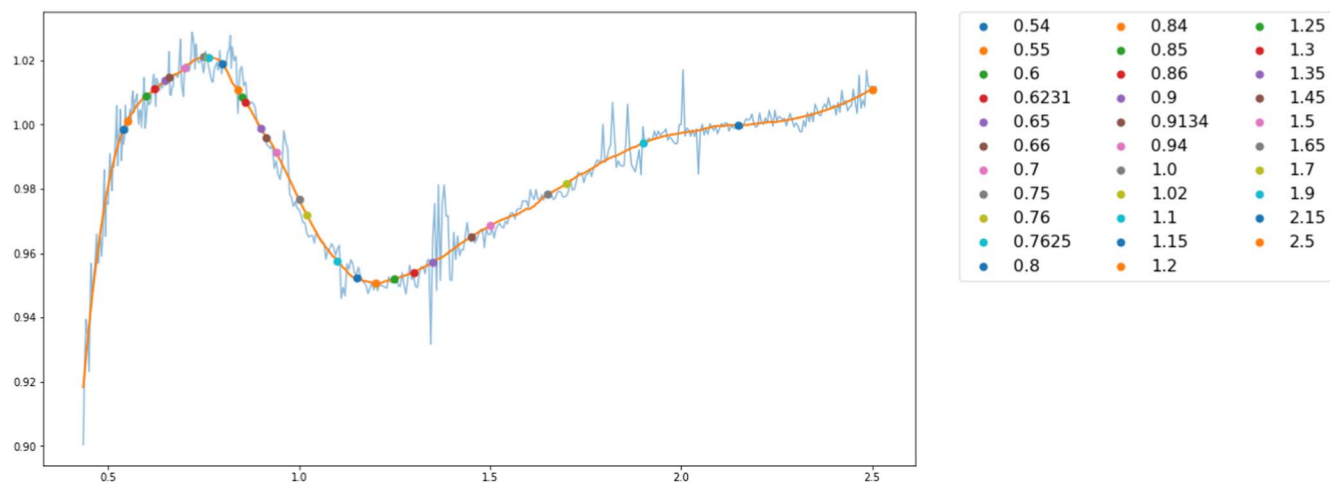
Figure 1. The plot presents the spectrum before processing (blue line), the cubic spline of smoothed spectra (orange line) and sampled 32 wavelengths used for calculating reflectance differences (dots).

hyperparameter sets were trained on the first part and evaluated on the second part of the training fold. Then, the best model was selected and evaluated on the test fold. Those were the final results reported.

Only up to 10 sets of hyperparameters were selected for each classifier, as multiple runs of cross validation for each of the methods required significant amount of computational power. It is worth noting that in order to achieve the final model performance, 500 models had to be trained for each method.

Specific hyperparameters for individual models were chosen as follows:

1. Multiclass logistic regression—inverse regularization strength in the range from 5 to 60; regularization norm either L_1 or L_2.
2. Naive Bayes—variance smoothing parameter in the range of $1e - 10$ to $1e - 6$.
3. Support vector machine—kernel: *RBF* or *linear*; gamma parameter: *scale* or *auto* (only for RBF kernel); regularization parameter C from 6 to 24.
4. Gradient boosting—the number of estimators from 50 to 500; maximum tree depth from 3 to 15; learning rate from 0.01 to 0.1; subsampling parameter 0.75 or 1.
5. Multilayer perceptron—2 to 3 hidden layers with 32 or 64 neurons each; batch size 32 or 64; the optimizer was set as SGD or Adam, learning rate ranged from 0.001 to 0.1.

All the aforementioned methods are from the scikit-learn package. A more detailed description of each hyperparameter can be found in Buitinck et al. (2013).

Taxonomic types were ranked during the sequential forward feature selection. While it is a computationally complex process, it was carried out only for the multilayer perceptron, the model with the highest Acc in predicting taxonomic types. During five-cross validation, the process was performed in each fold with the best set of parameters from previous experiments and then the results for all runs were averaged.

One versus all sequential feature selection per complex was performed in the manner described in Section 3.1.1. Similarly as for the sequential feature selection above, the model selected for this experiment was the multilayer perceptron with most commonly chosen parameters, and the results were averaged over five runs of the procedure.

4. Results

4.1. Classification with Reflectance Differences

The experiments were carried to assess how much of the information from the spectral reflectance data is retained in the derived reflectance differences, and how it affects the prediction Acc. For that reason, we used a wide range of spectral features calculated from each object's reflectance spectrum and described in Section 2 and Table 2. The results for classification into taxonomic types are given in Table 3. Surprisingly, all the methods produced predictions not worse, and sometimes even slightly better, than those obtained from automatically selected spectral features and five main principal components as in Klimczak et al. (2021). The best method—the multilayer perceptron—correctly classified 81% of the objects, as measured with the average of recall for each type. Furthermore, the average BAcc for classification into complexes (Table 3) has consistently improved for all models, and the Acc of multilayer perceptron reached almost 91%, while the support vector machine was the best performing model for this task, reaching 92%. These promising results indicate that spectral features, selected manually according to the criteria used in definition of the Bus-DeMeo taxonomic classes, lead to a data representation even more suitable for classification than the spectral values selected by machine-learning algorithms. We believe that the reason for our models performing better on the smaller feature set is the phenomenon described as the curse of dimensionality (Bellman 2010). It states that with the increase in the number of features, hence the number of dimensions in the feature space, the amount of data points required to train a model grows exponentially. As we were struggling with a small number of samples, especially for some classes, the larger number of features increased the difficulty of the problem, and as a results the models were not trained well enough. By handpicking wavelengths that represent characteristic features of the spectra, we performed feature engineering that reduced the complexity of the problem, and improved the final results. With a larger number of data points, and longer training, we expect that this difference would be eliminated.

Table 2
Spectral Features and Reflectance Differences used in Computations

Feature Type	λ_{begin} (μm,)	λ_{mid} (μm)	λ_{end} (μm)	Reasoning
NIR brightening	0.9	⋯	2.5	All asteroid types
0.9 μm absorption band	0.7	1.0	1.2	A type: deep broad absorption band with a minimum near 1.0 μm V type: strong and narrow 1.0 μm absorption band C type: may exhibit slight feature longward 1.0 μm K type: wide absorption line centered just longward of 1.0 μm O type very rounded and deep "bowl" shape feature at 1.0 μm Q type: distinct 1.0 μm absorption feature R type: deep 1.0 μm feature S type: moderate 1.0 μm feature Sa: deep and extremely broad 1.0 μm feature Sq: Wide and shallow 1.0 μm absorption band Sr and Sv: narrow 1.0 μm feature
2.0 μm absorption band	1.5	1.9	2.5	A type: may have a shallow 2-μm absorption around 2.0 μm V type: strong absorption band around 2.0 μm L type: There may or may not be 2.0 μm absorption feature O type: significant absorption feature at 2.0 μm Q type: 2.0 μm feature with varying depths between objects R type: deep 1.0 μm feature S type: moderate 2.0 μm feature Sr and Sv: 2.0 μm feature present
Slope	1.1	⋯	2.5	C type: low positive slope after 1.3 μm Cg: small-positive slope that begins at around 1.3 μm
0.6 μm bump	0.6	⋯	0.75	B type: round bump around 0.6 μm C type: slight round bump around 0.6 μm
Slope	1.3	⋯	2.5	C type: low positive slope after 1.3 μm Cg: small-positive slope that begins at around 1.3 μm
Slope	1.1	⋯	2.5	Cb type: linear positive slope that begins at 1.1 μm
0.7 absorption band	0.55	0.7	0.85	Cgh: shallow absorption band near 0.7 μm Ch: broad and shallow absorption band near 0.7 μm hydrated minerals
Kick around 1.5 μm	1.35	1.5	1.65	D type: gentle kick around 1.5 μm
Gentle concave up	1.0	1.5	1.7	L type: often gentle concave down curvature with a maximum at 1.5 μm
1.3 μm feature	1.15	1.3	1.45	Q type: evidence of feature around 1.3 μm Sq: shallow feature around 1.3 μm
0.55 absorption feature	0.55	⋯	0.7	Xe: absorption band feature shortward 0.55 μm
0.8 to 1.0 absorption feature	0.65	0.8	1.0	Xk: absorption band feature between 0.8 and 1.0 μm
Absorption feature	0.54	0.6	0.66	Hydrated minerals
Absorption feature	0.76	0.8	0.84	Hydrated minerals
Absorption feature	0.86	0.9	0.94	Hydrated minerals
0.9 μm feature	0.7	0.9	1.0	Xn: Narrow 0.9 μm feature
$R(\lambda_r) - R(\lambda_i)$	0.623		0.763	Equivalent of the respective color index
$R(\lambda_i) - R(\lambda_z)$	0.763		0.913	Equivalent of the respective color index
$R(\lambda_z) - R(\lambda_Y)$	0.913		1.020	Equivalent of the respective color index
$R(\lambda_Y) - R(\lambda_J)$	1.020		1.250	Equivalent of the respective color index
$R(\lambda_J) - R(\lambda_H)$	1.250		1.650	Equivalent of the respective color index
$R(\lambda_H) - R(\lambda_{Ks})$	1.650		2.150	Equivalent of the respective color index

Note. definitions of spectral features taken from Vilas & Gaffey (1989), Vilas (1994), Rivkin et al. (2002), DeMeo et al. (2009), and Binzel et al. (2019); and color indices from LSST Science Collaboration (2009) and Popescu et al. (2016).

4.2. Feature Selection

As the set of derived reflectance difference features still contained a large number of features with the wide spectral range (32 distinct reflectance differences in total), we ran the sequential forward feature selection to reduce the size of the feature set and assess the importance of the reflectance differences for classification of the asteroids into types and complexes. The averaged results from the five runs of feature selection per step are presented in Table 4 (separately for types and complexes). The tables should be interpreted as follows: the ith row ($i = 1,…,20$) contains the average (over five runs) classification Acc when only i features are used to describe the objects. It follows from these tables that only five features are sufficient to obtain the quality of predictions essentially equal to the one on the entire feature set. Furthermore, the BAcc does not improve significantly past a certain point—in the case of taxonomic types the improvement slows down at about eight steps, while it takes seven steps for the prediction of complexes. This dramatic reduction in the number of features shows that the entire spectrum is not necessary to make reliable prediction, and only a small number of coefficients computed from the spectra suffices.

In order to have more insight into which reflectance differences carry most information about the classes, Table 5 contains the average ranks obtained by each feature during the feature selection process. The rank of a given feature is the iteration number in which the feature was added to the selected subset (e.g., the feature added in the first iteration receives rank 1). This information allows us to quantify the importance of each feature for classification of taxonomic types and complexes: the lower the rank, the larger the importance, as the feature gets caught early in the selection process as the one improving the Acc the most. One can see that the top performing features are consistently selected in the first iterations in most of the experiments, thus forming a stable and informative subset, sufficient for reliable classification of the Bus-DeMeo taxonomy.

For the purpose of defining spectrophotometric passbands for an efficient asteroid survey we extracted the top five features that were sufficient for obtaining taxonomic type with the BAcc level of 80% (they correspond to step 5 in Table 4). The names of those features can be extracted from Table 5 (the column referring to taxonomic types) and are as follows:

1. The $0.9\,\mu$m absorption band (best measured by a reflectance difference $R(1.0\,\mu\text{m}) - R(1.2\,\mu\text{m})$;
2. The $R(0.6\,\mu\text{m}) - R(0.75\,\mu\text{m})$ feature;
3. The $2.0\,\mu$m absorption band (best measured by reflectance difference $R(1.5\,\mu\text{m}) - R(1.9\,\mu\text{m})$);
4. The $R(\lambda_i) - R(\lambda_z)$ feature;
5. The $R(\lambda_r) - R(\lambda_i)$ feature.

Their positions in comparison to avarage asteroid spectra are presented in Figure 2. Obviously, those spectral reflectances should be first normalized to unity at $\lambda_n = 0.55\,\mu$m, so this wavelength is a mandatory addition to the set. As a result we end up with the following sequence of 10 wavelengths: 0.55, 0.6, 0.623, 0.75, 0.763, 0.913, 1.0, 1.2, 1.5, and $1.9\,\mu$m. As can be seen, there are two pairs of wavelengths which are very similar: (0.6, 0.623) and (0.75, 0.763). In practice, the reflectances at such close wavelength are almost the same so for the purpose of the band definition we can take their mean values: 0.611 and 0.756. After this operation our set is reduced to 8 values: 0.55, 0.611, 0.756, 0.913, 1.0, 1.2, 1.5, and $1.9\,\mu$m.

In the study on complexes, the top two spectral features resulted in 85% BAcc (they correspond to step 2 in Table 4). The names of those features can be extracted from Table 5 (the column referring to complexes) and are as follows:

1. The $0.9\,\mu$m absorption band (best measured by a reflectance difference $R(1.0\,\mu\text{m}) - R(1.2\,\mu\text{m})$;
2. Kick around $1.5\,\mu$m which is diagnostic for the identification of D type (best measured by a reflectance difference $R(1.5\,\mu\text{m}) - R(1.65\,\mu\text{m})$.

Also in this case we have to add the reflectance at $\lambda_n = 0.55\,\mu$m for normalization which results in a set of 5 wavelengths: 0.55, 1.0, 1.2, 1.5, $1.65\,\mu$m.

Based on the identified, optimal reflectance differences, we conclude that to replace spectroscopic identification of types and complexes by the reflectance spectrophotometry (at the 80% Acc level) a set of eight passbands (for taxonomic types) and five passbands (for complexes) would be needed. Unfortunately, adding more passbands for the taxonomic type determination does not increase the BAcc significantly (the BAcc grows from BAcc = 80% for step 5, to BAcc = 85% for step 14 and then fluctuates around this value). The situation is different for the complexes. There we can increase the BAcc from BAcc = 85% (for step 2) to BAcc = 92% for step 4. This corresponds to extending the filter set to 7 wavelengths: 0.55, 0.7, 1.0, 1.2, 1.25, 1.5, $1.65\,\mu$m. Adding more wavelengths does not improve the BAcc much (it grows from 92% to 95%) and is not useful. Their positions in comparison to avarage asteroid spectra are presented in Figure 3.

4.3. Prediction of Individual Complexes

Further experiments were conducted on a reduced problem, where the task for the models was to predict whether an object belongs to a certain complex or not. Table 4 presents the scores obtained during the Feature Selection process for C, S, X and EM complexes. Table 6 contains a ranking of features, with the lower average position indicating that the feature was selected earlier in the process, hence it has higher impact on the classification score. As previously mentioned, the results for the C complex indicate that there is only one feature $R(\lambda_i) - R(\lambda_z)$ needed to obtain 92% BAcc. Additional features provide further, slight improvement which saturates at 98% and three features $(R(\lambda_i) - R(\lambda_z)$, $R(0.7\,\mu\text{m}) - R(1.0\,\mu\text{m})$, $R(0.9\,\mu\text{m}) - R(2.5\,\mu\text{m}))$. Similarly high scores can be obtained for the S complex, one feature $(R(0.7\,\mu\text{m}) - R(1.0\,\mu\text{m}))$ contributing to 85% BAcc, with further improvements reaching 97% at four features $(R(0.7\,\mu\text{m}) - R(1.0\,\mu\text{m})$, $R(0.9\,\mu\text{m}) - R(2.5\,\mu\text{m})$, $R(1.0\,\mu\text{m}) - R(1.2\,\mu\text{m})$, $R(\lambda_r) - R(\lambda_i))$. The X and EM complexes present a slightly bigger challenge for the models. The X complex requires at least two features to reach 85% $(R(\lambda_i) - R(\lambda_z)$, $R(\lambda_r) - R(\lambda_i))$. This score can be steadily improved up to 95% at 7 features $(R(\lambda_i) - R(\lambda_z)$, $R(\lambda_r) - R(\lambda_i)$, $R(\lambda_J) - R(\lambda_H)$, $R(\lambda_H) - R(\lambda_{Ks})$, $R(1.0\,\mu\text{m}) - R(1.2\,\mu\text{m})$, $R(0.9\,\mu\text{m}) - R(2.5\,\mu\text{m})$, $R(0.7\,\mu\text{m}) - R(1.0\,\mu\text{m}))$. A similar behavior occurs for the EM complex, with the initial score of 75% for one feature: $R(0.9\,\mu\text{m}) - R(2.5\,\mu\text{m})$, going up to 95% at five features: $R(0.9\,\mu\text{m}) - R(2.5\,\mu\text{m})$, $R(\lambda_H) - R(\lambda_{Ks})$, $R(0.7\,\mu\text{m}) - R(1.0\,\mu\text{m})$, $R(\lambda_i) - R(\lambda_z)$, $R(1.0\,\mu\text{m}) - R(1.2\,\mu\text{m})$.

The obvious advantage of using spectrophotometry in asteroid taxonomy determination is the efficiency of this

Table 3
The Results for the Classification of Taxonomic Types and Complexes on Reflectance Differences

Model	Accuracy	Balanced_accuracy	F1	MCC
		Taxonomic Types		
Logistic regression	0.845 ± 0.01	0.807 ± 0.011	0.804 ± 0.01	0.807 ± 0.013
Naive Bayes	0.763 ± 0.01	0.726 ± 0.013	0.74 ± 0.014	0.717 ± 0.012
SVM	0.848 ± 0.009	0.808 ± 0.014	0.801 ± 0.014	0.81 ± 0.011
Gradient boosting	0.798 ± 0.009	0.766 ± 0.009	0.748 ± 0.009	0.747 ± 0.012
MLP	0.858 ± 0.018	0.811 ± 0.025	0.807 ± 0.024	0.824 ± 0.022
		Complexes		
Logistic regression	0.783 ± 0.006	0.788 ± 0.011	0.765 ± 0.01	0.618 ± 0.01
Naive Bayes	0.83 ± 0.004	0.78 ± 0.007	0.784 ± 0.007	0.71 ± 0.007
SVM	0.944 ± 0.005	0.921 ± 0.01	0.918 ± 0.009	0.903 ± 0.009
Gradient boosting	0.921 ± 0.009	0.898 ± 0.012	0.892 ± 0.012	0.863 ± 0.016
MLP	0.938 ± 0.012	0.909 ± 0.023	0.906 ± 0.023	0.893 ± 0.02

Table 4
Average BAcc for the Classification of Different Complexes per Step of Sequential Feature Selection

Step	Average BAcc					
	C	S	X	EM	Taxonomies	Complexes
1	0.919 ± 0.004	0.847 ± 0.001	0.538 ± 0.001	0.750 ± 0.002	0.346 ± 0.003	0.647 ± 0.001
2	0.975 ± 0.001	0.931 ± 0.006	0.850 ± 0.001	0.817 ± 0.003	0.609 ± 0.009	0.845 ± 0.011
3	0.988 ± 0.001	0.956 ± 0.010	0.917 ± 0.006	0.892 ± 0.009	0.691 ± 0.010	0.894 ± 0.005
4	0.988 ± 0.001	0.971 ± 0.008	0.938 ± 0.009	0.945 ± 0.006	0.773 ± 0.014	0.922 ± 0.003
5	0.988 ± 0.001	0.976 ± 0.003	0.949 ± 0.011	0.949 ± 0.003	0.801 ± 0.011	0.934 ± 0.003
6	0.988 ± 0.001	0.977 ± 0.003	0.948 ± 0.007	0.950 ± 0.004	0.814 ± 0.006	0.939 ± 0.002
7	0.988 ± 0.001	0.978 ± 0.004	0.956 ± 0.011	0.953 ± 0.002	0.832 ± 0.009	0.941 ± 0.002
8	0.988 ± 0.001	0.979 ± 0.002	0.950 ± 0.006	0.954 ± 0.008	0.841 ± 0.009	0.944 ± 0.003
9	0.988 ± 0.001	0.977 ± 0.002	0.957 ± 0.006	0.957 ± 0.010	0.844 ± 0.008	0.945 ± 0.002
10	0.987 ± 0.001	0.977 ± 0.003	0.956 ± 0.006	0.958 ± 0.008	0.845 ± 0.008	0.944 ± 0.003

Note. Columns 2–5 refer to the specified complexes, column 6 refers to the overal Bacc for all taxonomies, and column 7 for the overal Bacc for all complexes.
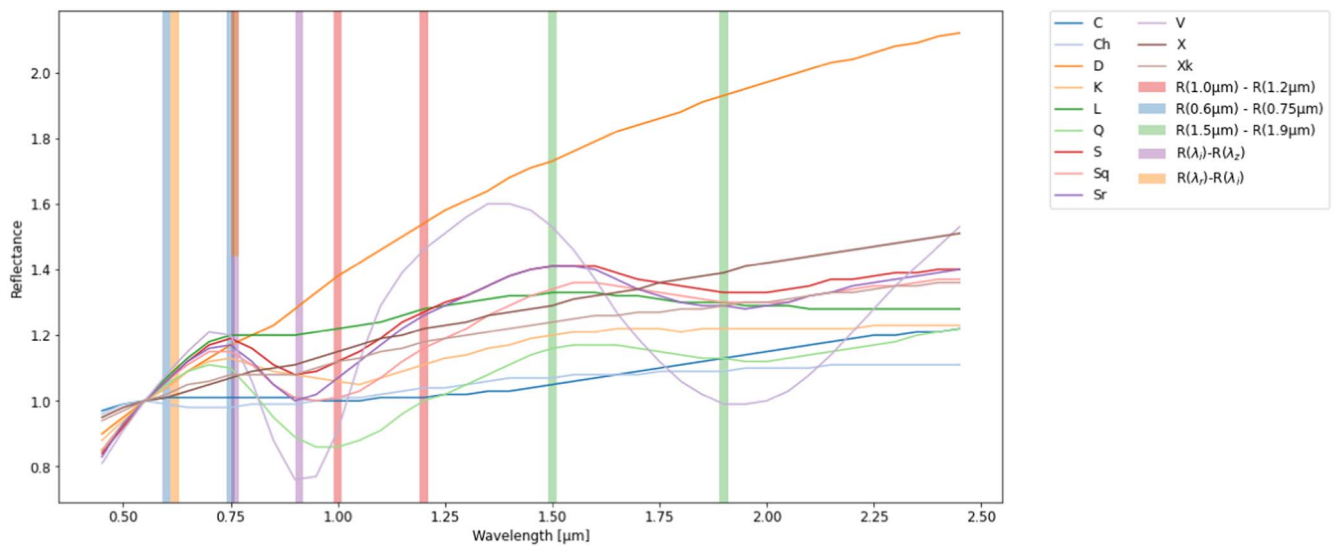

Figure 2. The most important reflectance differences obtained for the prediction of taxonomic types used in classification are marked as different colors. If the wavelenght is used for two reflectance differences it is marked by two colors.

technique. A single wide-field, deep sky image of modern sky surveys records brightnesses of thousands of asteroids—this cannot be obtained with spectroscopy.

The widths of the passbands should be chosen in such a way, that they do not smooth the reflectance spectrum too much (e.g., are not too wide), and on the other hand allow enough
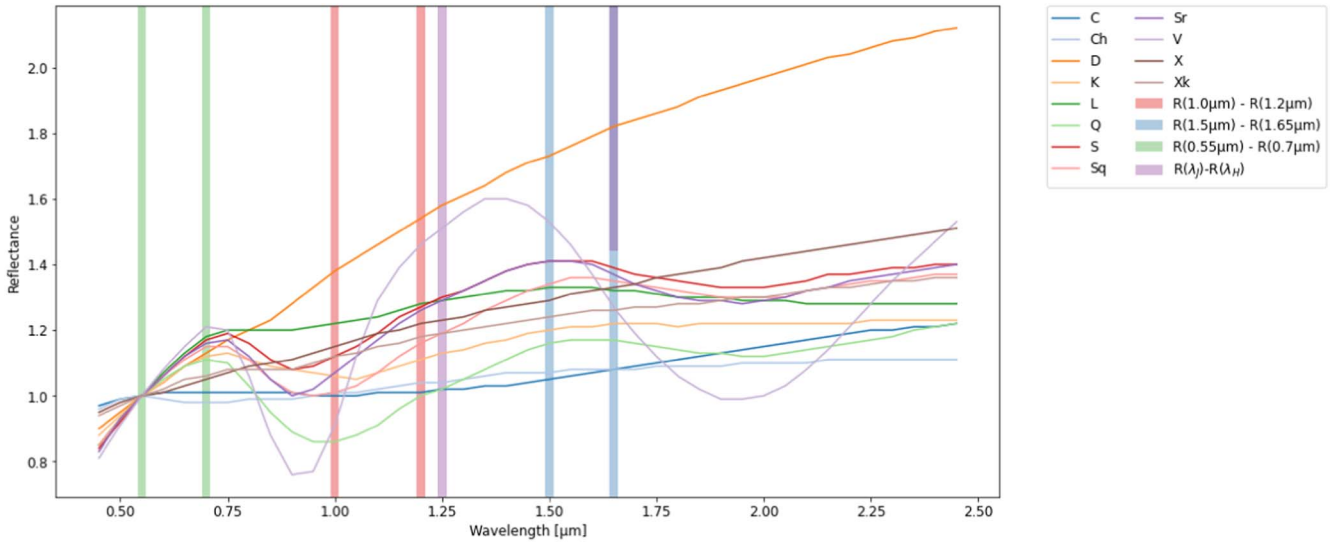
Figure 3. The most important reflectance differences obtained for the prediction of complexes used in classification are marked as different colors. If the wavelenght is used for two reflectance differences it is marked by two colors.

Table 5
Average Rank per Feature for the Classification of Taxonomic Types and Complexes during Sequential Feature Selection

Taxonomies		Complexes	
Feature	Position	Feature	Position
$R(1.0\ \mu m) - R(1.2\ \mu m)$	8.4	$R(1.0\ \mu m) - R(1.2\ \mu m)$	3.0
$R(0.6\ \mu m) - R(0.75\ \mu m)$	8.6	$R(1.5\ \mu m) - R(1.65\ \mu m)$	6.2
$R(1.5\ \mu m) - R(1.9\ \mu m)$	9.4	$R(0.55\ \mu m) - R(0.7\ \mu m)$	7.6
$R(\lambda_i) - R(\lambda_z)$	9.8	$R(\lambda_J) - R(\lambda_H)$	8.6
$R(\lambda_r) - R(\lambda_i)$	10.6	$R(1.5\ \mu m) - R(1.9\ \mu m)$	10.6
$R(1.5\ \mu m) - R(1.65\ \mu m)$	10.6	$R(0.7\ \mu m) - R(1.0\ \mu m)$	12.0
$R(0.7\ \mu m) - R(0.85\ \mu m)$	10.8	$R(1.5\ \mu m) - R(1.7\ \mu m)$	13.0
$R(\lambda_Y) - R(\lambda_J)$	11.6	$R(\lambda_i) - R(\lambda_z)$	13.2
$R(0.55\ \mu m) - R(0.7\ \mu m)$	12.2	$R(\lambda_Y) - R(\lambda_J)$	14.4
$R(1.1\ \mu m) - R(2.5\ \mu m)$	13.0	$R(\lambda_J) - R(\lambda_{Ks})$	14.6
$R(1.9\ \mu m) - R(2.5\ \mu m)$	13.2	$R(1.35\ \mu m) - R(1.5\ \mu m)$	15.2
$R(0.7\ \mu m) - R(1.0\ \mu m)$	14.2	$R(1.0\ \mu m) - R(1.5\ \mu m)$	15.4
$R(0.9\ \mu m) - R(2.5\ \mu m)$	15.4	$R(1.15\ \mu m) - R(1.3\ \mu m)$	15.6
$R(1.35\ \mu m) - R(1.5\ \mu m)$	15.8	$R(0.7\ \mu m) - R(0.85\ \mu m)$	17.4
$R(1.0\ \mu m) - R(1.5\ \mu m)$	16.0	$R(\lambda_r) - R(\lambda_i)$	17.4
$R(\lambda_J) - R(\lambda_{Ks})$	16.2	$R(\lambda_H) - R(\lambda_{Ks})$	17.8
$R(1.15\ \mu m) - R(1.3\ \mu m)$	16.4	$R(0.6\ \mu m) - R(0.75\ \mu m)$	18.0
$R(\lambda_J) - R(\lambda_H)$	16.8	$R(0.9\ \mu m) - R(2.5\ \mu m)$	20.4
$R(1.3\ \mu m) - R(2.5\ \mu m)$	17.0	$R(1.1\ \mu m) - R(2.5\ \mu m)$	21.0
$R(1.5\ \mu m) - R(1.7\ \mu m)$	17.8	$R(1.3\ \mu m) - R(2.5\ \mu m)$	21.0
$R(\lambda_H) - R(\lambda_{Ks})$	20.0	$R(1.9\ \mu m) - R(2.5\ \mu m)$	21.0

signal to be recorded (not too narrow). If the filter is too narrow, then the exposure time must be longer, and the influence of the asteroid rotation is greater (this can be overcome with a simultaneous exposure through two filters in a beam splitter configuration). If the filter is too wide, then the individual bands (which are often close to each other and describe various bends and bumps) can overlap and the BAcc of sampling the reflectance spectrum will be lower.

Since modern interferometric filters with widths of the order of 10 nm can be made almost top-hat like, with very steep edges, we can characterize their passbands by specifying the central wavelength λ_0 and the total width. Because of that we can assign the flux received in such filter to the flux recorded in a spectrum at λ_0.

For the purpose of demonstration, we present a photometric system with five passbands (B1–B5), suitable for classification of asteroid complexes.

1. B1: $\lambda_0 = 0.55\ \mu m$, range [0.52, 0.58] μm;
2. B2: $\lambda_0 = 1.00\ \mu m$, range [0.94, 1.06] μm;
3. B3: $\lambda_0 = 1.20\ \mu m$, range [1.14, 1.26] μm;
4. B4: $\lambda_0 = 1.50\ \mu m$, range [1.44, 1.56] μm;
5. B5: $\lambda_0 = 1.65\ \mu m$, range [1.59, 1.71] μm.

Note that the widths of the B2—B5 bands are two times bigger than the yellow B1 band. This is because the solar flux in the infrared is smaller than in the visible range.

In some cases a more specific goal of the asteroid survey can be considered. For example, there are near-Earth asteroids which can be easily accessed by space missions (the so called low delta-v NEAs; Xie et al. 2021). The Ch type asteroids, included in the C complex, contain hydrated minerals and are considered as good targets for water extraction. Water is important for life support in human missions or *in situ* propellent production. A preselection of possible C complex, low delta-v NEAs can be based on spectrophotometric observations with a 3 band system ($\lambda = 0.55$, 0.763, 0.913 μm, see Table 6). This would give the classification BAcc of BAcc = 92% (first row of Table 4). In fact such a goal appeared during the planning of the Osiris-rex mission to asteroid Bennu, when the C type backup targets had to be selected (see, for example, Christou et al. 2012). It should be emphasized that some NEAs, during their Earth fly-by's, display large sky motions which makes it difficult to observe them with a typical slit spectrograph. For example, the low resolution spectrograph installed at the LCOGT 2-m telescope at Haleakala allows to get spectra of NEAs with the sky motions less than $15''$ minute^{-1} (J. Chatellain 2023, private communication). In such cases a spectrophotometric system shows its advantage.

Table 6
Average Rank per Feature for the Classification of the C, S, X, and EM Complexes during Sequential Feature Selection

C Type		S Type		X Type		EM Type	
Feature	Position	Feature	Position	Feature	Position	Feature	Position
$R(\lambda_i) - R(\lambda_z)$	2.0	$R(0.7~\mu m) - R(1.0~\mu m)$	1.0	$R(\lambda_i) - R(\lambda_z)$	4.4	$R(0.9~\mu m) - R(2.5~\mu m)$	1.0
$R(0.7~\mu m) - R(1.0~\mu m)$	4.6	$R(0.9~\mu m) - R(2.5~\mu m)$	2.0	$R(\lambda_r) - R(\lambda_i)$	5.0	$R(\lambda_H) - R(\lambda_{Ks})$	2.0
$R(\lambda_Y) - R(\lambda_J)$	5.0	$R(1.0~\mu m) - R(1.2~\mu m)$	5.4	$R(\lambda_J) - R(\lambda_H)$	7.2	$R(0.7~\mu m) - R(1.0~\mu m)$	5.4
$R(\lambda_H) - R(\lambda_{Ks})$	5.4	$R(\lambda_r) - R(\lambda_i)$	7.4	$R(\lambda_H) - R(\lambda_{Ks})$	7.2	$R(\lambda_i) - R(\lambda_z)$	6.6
$R(\lambda_J) - R(\lambda_{Ks})$	6.2	$R(\lambda_i) - R(\lambda_z)$	10.0	$R(1.0~\mu m) - R(1.2~\mu m)$	7.2	$R(1.0~\mu m) - R(1.2~\mu m)$	7.0
$R(1.5~\mu m) - R(1.9~\mu m)$	9.0	$R(\lambda_H) - R(\lambda_{Ks})$	10.0	$R(0.9~\mu m) - R(2.5~\mu m)$	8.8	$R(\lambda_r) - R(\lambda_i)$	9.2
$R(0.9~\mu m) - R(2.5~\mu m)$	10.8	$R(\lambda_J) - R(\lambda_{Ks})$	10.2	$R(0.7~\mu m) - R(1.0~\mu m)$	9.8	$R(\lambda_J) - R(\lambda_H)$	10.4
$R(\lambda_r) - R(\lambda_i)$	11.0	$R(\lambda_J) - R(\lambda_H)$	10.4	$R(\lambda_Y) - R(\lambda_J)$	10.2	$R(\lambda_Y) - R(\lambda_J)$	10.8
$R(\lambda_J) - R(\lambda_H)$	11.0	$R(\lambda_Y) - R(\lambda_J)$	11.0	$R(\lambda_J) - R(\lambda_{Ks})$	11.0	$R(\lambda_J) - R(\lambda_{Ks})$	10.8
$R(1.0~\mu m) - R(1.2~\mu m)$	11.0	$R(1.5~\mu m) - R(1.9~\mu m)$	11.0	$R(1.5~\mu m) - R(1.9~\mu m)$	11.0	$R(1.5~\mu m) - R(1.9~\mu m)$	11.0

5. Conclusions

The aim of this paper was the assessment of prediction BAcc for the Bus-DeMeo taxonomy using reflectance differences at selected wavelengths. The experiments have proven that color indices, after convertion to reflectance differences, were a promising resource for classification and the results were slightly better than those reported in Klimczak et al. (2021) for the asteroid spectra. For the presented experiments, the best model was a multilayer perceptron with the architecture of (64, 64) neurons. The batch size was set to 32, the adaptive learning rate was initialized to 0.1 and the solver was Stochastic Gradient Descent. We quantified the importance of each spectral feature from the wide spectral range, and selected a limited number of spectral differences, which provide a high significance for determination of the asteroid taxonomic types and complexes. It appeared that a set of eight spectrophotometric bands can be used in a survey to asses 80% BAcc in predicting the asteroid type. With a set of five bands slightly higher, 85% BAcc can be obtained in determination of the asteroid complexes. In special cases, in which a specific complex is searched for (as sometimes is in the case of space missions), even smaller band sets are suitable. For example, to check if a given asteroid belongs to the C complex (with the BAcc = 92%) a set of only three bands is required.

ORCID iDs

Tomasz Kwiatkowski ● https://orcid.org/0000-0003-3462-5714
Francesca DeMeo ● https://orcid.org/0000-0002-8397-4219

References

Bellman, R. 2010, Dynamic Programming (New York: Dover Publications)
Binzel, R., DeMeo, F., Turtelboom, E., et al. 2019, Icar, 324, 41
Buitinck, L., Louppe, G., Blondel, M., et al. 2013, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning (Leuven: Ku Leuven), 108
Carry, B. 2018, A&A, 609, A113
Carvano, J., Hasselmann, P., Lazzaro, D., & Mothé-Diniz, T. 2010, A&A, 510, A43
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
Christou, A. A., Kwiatkowski, T., Butkiewicz, M., et al. 2012, A&A, 548, A63
DeMeo, F. E., Binzel, R. P., Carry, B., Polishook, D., & Moskovitz, N. A. 2014, Icar, 229, 392
DeMeo, F. E., Binzel, R. P., Slivan, S. M., & Bus, S. J. 2009, Icar, 202, 160
DeMeo, F. E., Polishook, D., Carry, B., et al. 2019, Icar, 322, 13
Hietala, H. 2020, Asteroid Spectra and Machine Learning, MSc Thesis, Univ. of Helsinki
Hromakina, T., Birlan, M., Barucci, M. A., et al. 2021, A&A, 656, A89
Ivezić, Ž., Lupton, R. H., Jurić, M., et al. 2002, AJ, 124, 2943
Jedicke, R., Nesvorny`, D., Whiteley, R., Ivezić, Ž., & Jurić, M. 2004, Natur, 429, 275
Kelleher, J. D., Namee, B. M., & D'Arcy, A. 2015, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (Cambridge, MA: MIT Press)
Klimczak, H., Kotłowski, W., Oszkiewicz, D., et al. 2021, FrASS, 8, 216
Klimczak, H., Oszkiewicz, D., Carry, B., et al. 2022, A&A, 667, A10
Licandro, J., Popescu, M., Morate, D., & de León, J. 2017, A&A, 600, A126
LSST Science Collaboration 2009, arXiv:0912.0201
Morate, D., Marcio Carvano, J., Alvarez-Candal, A., et al. 2021, A&A, 655, A47
Moskovitz, N. A. 2012, Icar, 221, 63
Moskovitz, N. A., Jedicke, R., Gaidos, E., et al. 2008, Icar, 198, 77
Oszkiewicz, D., Kwiatkowski, T., Tomov, T., et al. 2014, A&A, 572, A29
Parker, A., Ivezić, Ž., Jurić, M., et al. 2008, Icar, 198, 138
Peña, J., Fuentes, C., Förster, F., et al. 2020, AJ, 159, 148
Penttilä, A., Fedorets, G., & Muinonen, K. 2022, FrASS, 9, 816268
Popescu, M., Licandro, J., Carvano, J., et al. 2018, A&A, 617, A12
Popescu, M., Licandro, J., Morate, D., et al. 2016, A&A, 591, A115
Rivkin, A., Howell, E., Vilas, F., & Lebofsky, L. 2002, Asteroids III, 1, 235
Rivkin, A. S. 2012, Icar, 221, 744
Sergeyev, A. V., & Carry, B. 2021, A&A, 652, A59
Sergeyev, A. V., Carry, B., Onken, C. A., et al. 2022, A&A, 658, A109
Sykes, M. V., Cutri, R. M., Fowler, J. W., et al. 2000, Icar, 146, 161
Szabó, G. M., Ivezić, Ž., Jurić, M., & Lupton, R. 2007, MNRAS, 377, 1393
Vaduvescu, O., Macias, A. A., Tudor, V., et al. 2017, EM&P, 120, 41
Vilas, F. 1994, Icar, 111, 456
Vilas, F., & Gaffey, M. J. 1989, Sci, 246, 790
Whitney, A. 1971, ITCmp, C-20, 1100
Xie, R., Bennett, N. J., & Dempster, A. G. 2021, AcAau, 181, 249