frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 11 March 2022
doi: 10.3389/fspas.2022.816268

Taxonomy of Asteroids From the Legacy Survey of Space and Time Using Neural Networks

A. Penttilä[1]*, *G. Fedorets*[2] *and K. Muinonen*[1]

[1]*Department of Physics, University of Helsinki, Helsinki, Finland,* [2]*Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, United Kingdom*

The Legacy Survey of Space and Time (LSST) is one of the ongoing or future surveys, together with the Gaia and Euclid missions, which will produce a wealth of spectrophotometric observations of asteroids. This article shows how deep learning techniques with neural networks can be used to classify the upcoming observations, particularly from LSST, into the Bus-DeMeo taxonomic system. We report here a success ratio in classification up to 90.1% with a reduced set of Bus-DeMeo types for simulated observations using the LSST photometric filters. The scope of this work is to introduce tools to link future observations into existing Bus-DeMeo taxonomy.

Keywords: legacy survey of space and time, asteroids, spectra, taxonomy, neural network

OPEN ACCESS

Edited by:
Robert C. Allen,
Johns Hopkins University,
United States

Reviewed by:
Jean Baptiste Marquette,
Laboratoire d'astrophysique de Bordeaux (LAB), France
Valerio Carruba,
São Paulo State University, Brazil

***Correspondence:**
A. Penttilä
antti.i.penttila@helsinki.fi

Specialty section:
This article was submitted to Astrostatistics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *16 November 2021*
Accepted: *10 February 2022*
Published: *11 March 2022*

Citation:
Penttilä A, Fedorets G and Muinonen K (2022) Taxonomy of Asteroids From the Legacy Survey of Space and Time Using Neural Networks. Front. Astron. Space Sci. 9:816268. doi: 10.3389/fspas.2022.816268

INTRODUCTION

Spectroscopy is the primary technique for the precise physical characterization of asteroids. Its use is hindered by the requirements of time-consuming dedicated observations. Therefore, up-to-date spectroscopy is only available for some thousands of asteroids out of around a million currently known. However, the corpus of data required for physical characterization of asteroids can be and has been greatly amended by less accurate albeit far more abundant multi-filter photometric observations. In particular, sky surveys provide a plentitude of broad-band spectrophotometric data of small Solar System bodies, despite their main purpose often being in other fields of astronomy. In particular, the Sloan Digital Sky Survey (SDSS, York et al., 2000), by-design a survey for galaxies, has been a valuable resource for asteroid characterization. DeMeo and Carry (2013) were able to classify tens of thousands of asteroids from the SDSS Moving Object Survey database (Ivezić et al., 2002) into the Bus-DeMeo taxonomy, and recently Sergeyev and Carry (2021) performed probabilistic classifications for almost 400,000 asteroids in the SDSS data. These results have been essential in mapping the spectral distribution of asteroids in the main belt, opening insights to the mechanisms sculpting the Solar System (e.g., Raymond and Nesvorný, 2021, and references therein).

The next generation of synoptic sky surveys is headed by the Vera Rubin Observatory's Legacy Survey of Space and Time (LSST; Ivezić et al., 2019), going two magnitudes deeper compared to SDSS. One of four major science goals for LSST will be the inventory of the Solar System (Jones et al., 2009; Schwamb et al., 2018). Currently expected to commence its nominal decade of operations in 2024, LSST is anticipated to discover 5.5 million new small Solar System objects (LSST Science Collaboration 2009, Chapter 5) from close-approaching near-Earth asteroids and objects inside Earth's orbit all the way to the distant realms of the transneptunian object population.

The number of asteroid discoveries and observations by LSST depends heavily on the survey cadence of LSST, which is under discussion as of late 2021. From the point of view of Solar System

Science Collaboration, the so-called Northern Ecliptic Spur has been identified as the primary requirement to amend the nominal southern-sky (Wide-Fast-Deep) survey cadence to reach the scientific goals. The Northern Ecliptic Spur is a crescent-shaped area of 5,800 square-degrees north of the celestial equator covering the ecliptic in its entirety with 10-degree margins in latitude.

As a rule, LSST data will be released using two different mechanisms (Jurić et al., 2017). Objects deemed to have high importance will be released as alerts 60 s after the observation. Generally, all observed and calibrated data will be distributed through annual data releases. In addition to alerts and data releases for all LSST data, Solar System observations will be submitted daily to the Minor Planet Center (e.g., Jurić et al., 2021). These daily submissions will be resubmitted upon recalibration of astrometry and photometry with annual data releases. The bulk of new asteroid discoveries is anticipated during the first year of LSST operations, but the photometric corpus of observations will be updated throughout the planned 10-year survey.

It would be extremely interesting to classify the asteroids observed by LSST using a taxonomic system, namely the Bus-DeMeo system. Having at least a preliminary classification for these millions of asteroids would greatly improve our understanding of the distribution of different materials and evolutional history in the asteroid population. We have similar new opportunities also with the upcoming data from the Gaia mission by the European Space Agency. The Data Release 3 (expected in 2022) and the data releases thereafter will include low-resolution spectral data from hundred thousand or so asteroids with wavelengths of 0.33–1.05 μm. Penttilä et al. (2021) showed that by using a neural network it is possible to obtain the Bus-DeMeo taxonomic classification for the Gaia spectral observations, even though the wavelength ranges for the Bus-DeMeo system (0.45–2.45 μm) differ from the Gaia wavelengths. Since neural networks are basically very flexible nonlinear approximators to any function, they are suitable for various classification tasks. In this article, we will study how a neural network, similar to that in Penttilä et al. (2021), could classify the LSST spectrophotometric observations into the Bus-DeMeo system.

MATERIALS AND METHODS

Combined Asteroid Dataset With Vis-NIR Spectra and Bus-DeMeo Taxonomic Classifications

The spectroscopic dataset utilized in this study is a combination of spectroscopic observations used in DeMeo et al. (2009) and observations from the MIT-Hawaii Near-Earth Object Spectroscopic Survey (MITHNEOS; Binzel et al., 2019). The Bus-DeMeo data has observations of 371 asteroids and the MITHNEOS has 316. Both datasets also contain the taxonomic classification of the objects together with their Vis-NIR spectra. When the datasets are combined, there are 602 unique asteroids. However, after further inspection, there is a total of 591 asteroids with spectral observations that could reliably be resampled into the wavelength range of 0.45–2.45 μm. Finally, three of the asteroids are of unknown taxonomy, and there is only a single asteroid for the taxonomic types O and R, making these impossible to use when both training and validating a classifier. Thus, the combined spectral dataset for this study has 586 asteroids from 11 taxonomic types. For more details about the processing of the data, see Appendix A from Penttilä et al. (2021).

Simulating the Legacy Survey of Space and Time Observations

The LSST survey will observe with six photometric filters. The filters *u*, *g*, *r*, *i*, *z*, and *y* and their passbands are described in the LSST webpage,[1] and the filter transmission curves can be downloaded from the LSST GitHub site.[2] The half-maximum transmission wavelength ranges for the filters are repeated here in **Table 1**. To simulate how the LSST survey would see the asteroids in our combined dataset, we convolve the spectral data with the filter transmission curves to produce six "colors." Since the sizes of the asteroids are generally not known, the absolute albedo information is not available in the future LSST observations. Thus, we should only use normalized colors in taxonomic classification.

We need to select one filter where the colors are normalized to unity. The *u* filter is below the range of the Bus-DeMeo wavelengths and cannot be used in Bus-DeMeo classification. Thus, selecting the *u* filter for normalization does not decrease the number of filters left for classification. On the other hand, the spectra in our combined dataset are already normalized at 550 nm. For the classification it should not matter if the normalization is done at 550 nm or at 375 nm (the center of the *u* filter), so we can use the combined dataset and the colors (without *u*) produced by the convolutions as they are in this example study. It should be noted, however, that if we want to build a taxonomic classifier for the actual LSST observations following the example in this study, we will need to augment our combined dataset with observations down to the *u* filter and normalize there to produce the correct numerical values for the coefficients in the neural network classifier.

The average spectral curves for each Bus-DeMeo taxonomical type are shown in **Figure 1** together with the bars indicating the LSST filters *g*, *r*, *i*, *z*, and *y*. All the subtypes of S, C, and X-complexes are shown in the same subfigure with their main type. By mere visual inspection, one can see that separating the subtypes in a complex using only the LSST filters would be a hard task, especially when remembering that the curves in the figures show only the average behavior without any natural variation within a subtype. Perhaps the subtype B for the C-complex with its negative spectral slope can be recognized, therefore in our first try on the classification, we will keep the B type but simplify all the

[1]https://www.lsst.org/about/camera/features
[2]https://github.com/lsst-pst

TABLE 1 | The six photometric filters used in the LSST survey, and their half-maximum transmission wavelength (HMTW) ranges.

LSST filters	u	g	r	i	z	y
HMTW range (nm)	350–400	400–552	552–691	691–818	818–922	948–1,060

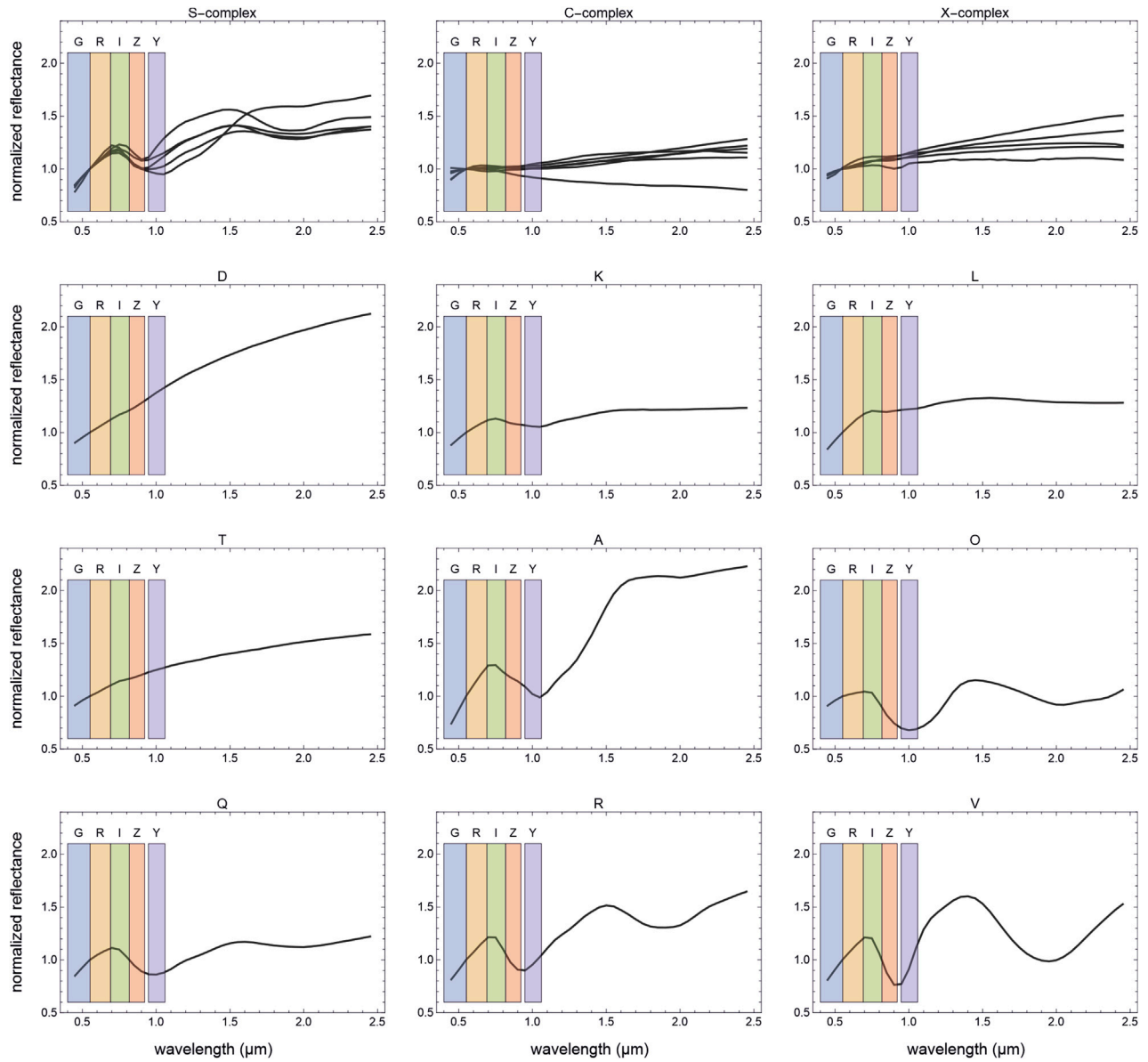

FIGURE 1 | Spectral behavior of the three taxonomic complexes and the nine endmember types in the Bus-DeMeo taxonomy. The average spectral behaviors of the types are shown with black solid lines. The wavelength ranges from which the LSST filters *g*, *r*, *i*, *z*, and *y* integrate the signal are shown in the background with colored rectangles. The spectral curves are normalized to unity at 0.55-μm wavelength.

other subtypes to their main type of S, C, or X. We will also keep all the endmember types in the taxonomy. Therefore, our dataset for testing the taxonomical classification on LSST observations consists of five normalized colors and a taxonomic type with 11 categories for 586 objects.

Neural Network Classifier

The neural network (NN) classifier that we are testing in this study is a two-layer, all-to-all-connected feed-forward network that was presented in Penttilä et al. (2021). The input layer for the normalized photometric color data that we are using has five

TABLE 2 | The accuracy of the neural network classifier per taxonomic type and their share of the data. "Cc," "Sc," and "Xc" stand for C, S, and X-complexes.

Taxonomy	A	B	Cc	D	K	L	Q	Sc	T	V	Xc	Accuracy (%)
Recall (%)	85.7	41.7	93.4	81.8	40.0	57.6	86.0	92.9	0.0	92.9	80.4	85.8
Precision (%)	85.7	83.3	85.1	75.0	66.7	59.4	80.4	92.6	–	92.9	73.2	
Share (%)	1.2	2.0	10.4	3.8	2.6	5.6	7.3	52.9	0.7	4.8	8.7	100

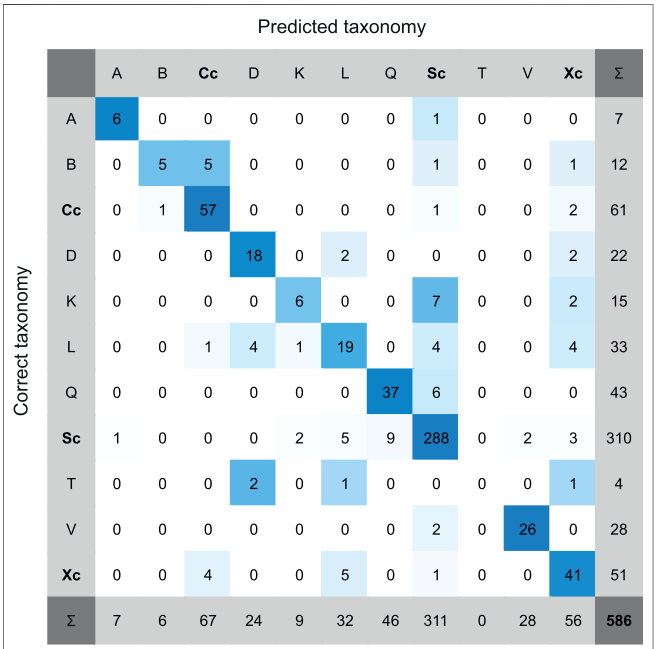

FIGURE 2 | Confusion matrix plot of the results of the NN taxonomic classification of simulated LSST observations. The correct taxonomic types are organized in rows and predicted types in columns. The rightmost column is a row sum, and the bottom row is a column sum of the asteroids. For example, for class D there are 22 asteroids of which 18 are correctly classified as D class, and 2 incorrectly as X-complex. In total, 586 asteroids are classified. The blue-to-white color of the cell indicates the share of the asteroids from that row having the predicted type of that column. A perfect classifier should have deep blue throughout the diagonal, and white elsewhere.

nodes. This input layer is connected to the first hidden layer with k nodes using hyperbolic tangent sigmoid (*tansig*) activation functions. The second hidden layer has as many nodes as there are classes in the data, starting with 11, using the so-called *softmax* activation function. The second hidden layer and the *softmax* function will produce the output, a single taxonomical type for the object.

The network is initialized with random weights and biases and trained using the Adam algorithm, a stochastic first-order gradient-based optimization (Kingma and Ba, 2015). We did validation rounds leaving randomly one-fourth of the data for validation and trained the network with the rest of the data. We increased the number of the nodes k in the first hidden layer, and the number of the training rounds in the optimization until the network accuracy was not increasing significantly. This process led us to have 20 nodes in the first hidden layer, trained with 10,000 rounds with the Adam algorithm.

TABLE 3 | The accuracy of the neural network classifier per taxonomic type with simplified taxonomy. The "Cc" stands for the C-complex, now also including the B-type. "D*" is the D-type and the T-type. "Sc*" is the S-complex with the K and L-types. "Xc" is the X-complex.

Taxonomy	A	Cc	D*	Q	Sc*	V	Xc	Accuracy (%)
Recall (%)	85.7	93.2	76.9	83.7	93.6	89.3	74.5	90.1
Precision (%)	85.7	91.9	80.0	78.3	93.6	92.6	77.6	
Share (%)	1.2	12.5	4.4	7.3	61.1	4.8	8.7	100

RESULTS

The final evaluation of the accuracy of the NN classifier was obtained using leave-one-out cross-validation. We trained the NN 586 times, each time leaving one asteroid out from the training data. This asteroid was then classified with the trained network, and the predicted taxonomic type was recorded.

The overall cross-validated accuracy, the fraction of correctly classified objects to all objects, of the NN predicting 11 taxonomic types using only the five normalized LSST colors was 85.8%. This performance, considering that only five filters (plus one for normalization) was used, is promising. The accuracy can be compared to possible upper limit from Penttilä et al. (2021), where they used the same NN classifier but for full spectral data between 0.45 and 2.45 μm, reaching 90.6% accuracy. On the other hand, a lower limit of 52.9% accuracy can be reached with the most simplistic possible approach, i.e., classifying all objects as S-complex since it is the most common one in the data.

In **Table 2** we show the NN recall and precision for each type or complex, together with their share in the data. Recall is defined to be the fraction of correctly classified objects to all objects of that class. Precision is the fraction of correctly classified objects to all objects that were classified to that class. All the complexes S, C, and X were quite well predicted with recall from 93.4% for the C-complex to 80.4% for the X-complex. Several endmember types, A, D, Q, and V were also well predicted. Quite poor recall was received for B, K, L, and T-types. Precision scores were not as low for any type, but the lowest was for predictions into L-type, of which only 59.4% were actually from L-type.

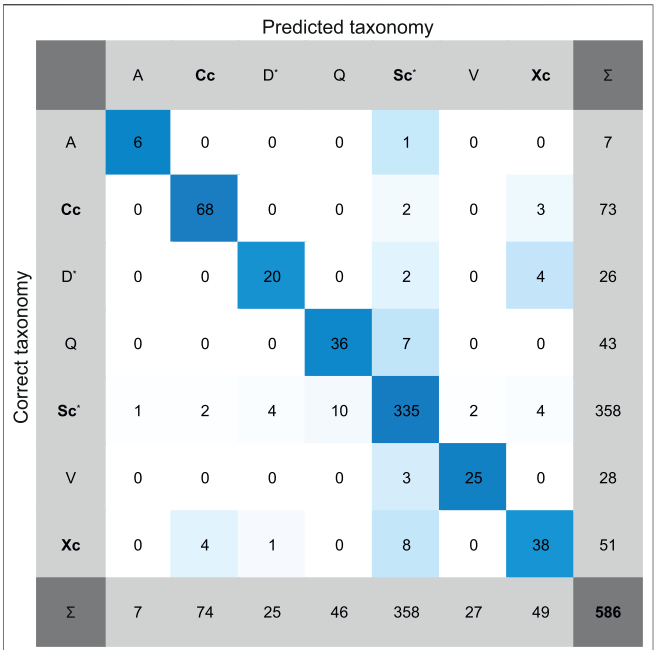

FIGURE 3 | Confusion matrix plot of the results of the NN taxonomic classification with simplified taxonomy of simulated LSST observations. See **Figure 2** for the explanation of the matrix, and **Table 3** for simplified taxonomy labels.

The results are more accurately presented in **Figure 2** with a so-called confusion matrix. In this matrix, one can see into which types the misclassifications are addressed. For example, none of the four T-type asteroids are correctly classified, instead, two of them are predicted as D, and one as L and X each. B-type is often misclassified as C, which is not surprising, and K as S. If the Q-type is misclassified, it is always misclassified as S. This is also quite natural, since the Q-type is thought to be a fresh, non-space-weathered counterpart of S.

Even though the overall accuracy of the NN classifier was quite good, it is evident that some of the types cannot be predicted well using only the LSST filters. These are the B, K, and T-types with accuracies below 50%. Therefore, we will test the classification again but with some types merged. We will merge the B-type into the C-complex, the K-type with the S-complex, and the T-type with the D-type. The results with this simplified taxonomy showed that the accuracy of the L-type did not improve but decreased, so finally we also merged the L-type with the S-complex.

The cross-validated accuracy of the NN trained with the simplified taxonomical types was 90.1%. All the individual types had more than about 75% accuracy, see **Table 3**. The confusion matrix in **Figure 3** shows how the misclassifications are distributed among the types. Overall, we find the results very good.

DISCUSSION

The results presented here show that a feed-forward neural network can be used to classify asteroid observations made with photometric filters matching the ones planned for the LSST survey. The classification into Bus-DeMeo taxonomy (without subtypes of S, C, and X-complexes) has an 85.8% accuracy, and by simplifying the taxonomic types we can reach 90.1% accuracy. We think that this method is promising to be used with the LSST asteroid data, at least for giving a preliminary Bus-DeMeo taxonomic classification.

We note that our tests with the method assume that we have the lowest-wavelength *u* filter observed for the asteroids with the LSST, and that we can reliably augment the existing training data that we used here into the same wavelengths. This is because we need to have one filter for spectral normalization. If, for some reason, either the upcoming observations or our training data will not have the *u* filter values, we need to normalize using the five other filters. This will still be possible, but it will have a decreasing effect on the accuracy. We predict that the effect will be small enough to keep the proposed method still useful, but another study verifying this assumption would be needed if the situation would be realized.

The neural network designed for classification can also give probability estimates for all the types of an object. Studying these can be useful, and perhaps objects with relatively high probabilities of the second-best taxonomic type can be marked for further verifications. If the probability estimates will be used, we would like to propose to study how robust the probability estimates are. The trained neural network has always some randomness since the training starts with random initial node weights and biases, and the training algorithm is not guaranteed to find the global minimum of the loss function. When comparing multiple identical networks that are trained similarly but from random initial parameter values, the actual classifications should not vary much, but the probability estimates for the types might vary more. This can be studied, and if there is significant variation, we would suggest using the method of multiple independent neural network "voters," as discussed in Penttilä et al. (2021). In short, one would train several independent classifiers, and form the final output probabilities by taking the (trimmed) mean over the voters. This method will tackle the possible problem of random variation in the probability estimates.

While the training of a neural network can be a somewhat lengthy process demanding computer time, the evaluation using the trained network is not an especially heavy task computationally. Therefore, the proposed neural network classifier should be possible to apply automatically to all LSST survey asteroids, if so desired.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

AP is responsible of the article concept, the programming, simulations and the analysis of the results, and writing the article. GF and KM have participated in writing the article.

FUNDING

This work was supported by the Academy of Finland projects Nos. 325805, 336546, and 345115. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101032479.

ACKNOWLEDGMENTS

We acknowledge the computational resources provided by CSC—IT Center for Science Ltd., Finland. We thank Francesca DeMeo for providing us the original asteroid spectra that was used in creating the Bus-DeMeo taxonomy.

REFERENCES

Binzel, R. P., DeMeo, F. E., Turtelboom, E. V., Bus, S. J., Tokunaga, A., Burbine, T. H., et al. (2019). Compositional Distributions and Evolutionary Processes for the Near-Earth Object Population: Results from the MIT-Hawaii Near-Earth Object Spectroscopic Survey (MITHNEOS). *Icarus* 324, 41–76. doi:10.1016/j.icarus.2018.12.035

DeMeo, F. E., Binzel, R. P., Slivan, S. M., and Bus, S. J. (2009). An Extension of the Bus Asteroid Taxonomy into the Near-Infrared. *Icarus* 202, 160–180. doi:10.1016/j.icarus.2009.02.005

DeMeo, F. E., and Carry, B. (2013). The Taxonomic Distribution of Asteroids from Multi-Filter All-Sky Photometric Surveys. *Icarus* 226, 723–741. doi:10.1016/j.icarus.2013.06.027

Ivezić, Ž., Jurić, M., Lupton, R. H., Tabachnik, S., and Quinn, T. (2002). "Asteroids Observed by the Sloan Digital Sky Survey," in *Survey and Other Telescope Technologies and Discoveries*. Editors JA Tyson and S Wol. (Proceedings of the SPIE), 4836, 98–103.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., et al. (2019). LSST: from Science Drivers to Reference Design and Anticipated Data Products. *Astronomical J.* 873 (2), 111. doi:10.3847/1538-4357/ab042c

Jones, R. L., Chesley, S. R., Chesley, S. R., Connolly, A. J., Harris, A. W., Ivezic, Z., et al. (2009). Solar System Science with LSST. *Earth Moon Planet.* 105, 101–105. doi:10.1007/s11038-009-9305-z

Jurić, M., Holman, M., Eggl, S., Lackner, M., Moeyens, J., Pan, M., et al. (2021). DEx1: The First LSST-MPC Data Exchange Challenge Report. Available at: https://dmtn-180.lsst.io/ (Accessed 1.9.2021).

Jurić, M., Kantor, J., Lim, K-T., Lupton, R. H., Dubois-Felsmann, G., Jenness, T., et al. (2017). "The LSST Data Management System," in Astronomical Data Analysis Software and Systems XXV ASP Conference Series, Sydney, Australia, October 25–29, 2015. Editors NPF Lorente, K Shortridge, and R Wayth, 512, 279–289.

Kingma, D., and Ba, J. (2015). "Adam: A Method for Stochastic Optimization," in Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, May 7–9, 2015.

LSST Science Collaboration (2009). *LSST Science Book*. arXiv:0912.0201.

Penttilä, A., Hietala Hand Muinonen, K. (2021). Asteroid Spectral Taxonomy Using Neural Networks. *Astron. Astrophysics* 649, A46. doi:10.1051/0004-6361/202038545

Raymond, S. R., and Nesvorny, D. (2021). *Origin and Dynamical Evolution of the Asteroid Belt*. arXiv:2012.07932.

Schwamb, M. E., Jones, R. L., Chesley, S. R., Fitzsimmons, A., Fraser, W. C., Holman, M. J., et al. (2018). Large Synoptic Survey Telescope Solar System Science Roadmap. arXiv:1802:01783.

Sergeyev, A. V., and Carry, B. (2021). A Million Asteroid Observations in the Sloan Digital Sky Survey. *Astron. Astrophysics* 652, A59. doi:10.1051/0004-6361/202140430

York, D. G., Adelman, J., Anderson, J. E., Jr, Anderson, S. F., Annis, J., Bahcall, N. A., et al. (2000). The Sloan Digital Sky Survey: Technical Summary. *Astronomical J.* 120, 1579–1587. doi:10.1086/301513

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.