A Novel AI-Powered Approach to Spectral Asteroid Classification to Identify Optimal Mining Targets

Nayan A. Patel

Cary Academy

Abstract

The mining of Earth's finite metals has devastating environmental and humanitarian impact. Asteroid mining provides an elegant alternative to the limited resources on Earth, as well as playing a critical role in the quest for interplanetary travel. Thus, a system to reliably determine the composition of asteroids is imperative to seek out the correct targets. Currently, asteroid spectral graphs are matched using mathematical error methods with known spectra of elements to determine their composition. Unfortunately, these approaches are unidimensional and blind to other factors that influence asteroid composition, breaking down when faced with more unique asteroid spectra. In this paper, a novel approach to asteroid spectral analysis is taken by using a convolutional neural network (CNN), engineered to analyze patterns within asteroid spectral graphs supplemented by auxiliary asteroid data – diameter, absolute magnitude, and albedo – to provide insight not found in current asteroid classification methods. Data was sourced from MIT's Small Main-Belt Asteroid Spectroscopic Survey, Phase II (SMASSII) and the space-rocks database. A program was created to automatically aggregate data from the two sources, filling empty values with simulated data. Data augmentation algorithms were created and employed to balance the dataset, ensuring sufficient unique data was present in each spectral class for the CNN to learn from. Reclassifying images of spectral graphs into 18 common spectral classes from the Bus-DeMeo taxonomic system, the CNN predicted the correct asteroid class 98.8% of the time – a 13.7% improvement from a model solely using spectral graphs. Thus, this study not only showcases the accuracy of CNNs in analyzing asteroid spectral graphs, but also underscores how auxiliary data enhances asteroid composition prediction.

1. Introduction

The mining of asteroids is one of the most potentially groundbreaking advances in the history of humanity. Earth's resources are scarce, and for mankind to succeed in interplanetary colonization, additional resources will be needed. For example, there will be a need for more gold to build circuits, more cobalt to build batteries, more titanium to build rockets. Further, the extraction of said scarce materials is a laborious process with adverse humanitarian and environmental effects, with mineral mining described as employing "modern slavery" in Congo cobalt mines while encompassing almost 7% of global greenhouse gas emissions. The only choice, then, is the vast expanse of our celestial neighborhood. Take 16 Psyche, an asteroid located in our own asteroid belt estimated to contain precious metals worth more than the global economy (Sohn, 2023). Harvesting metals from an asteroid such as 16 Psyche would revolutionize innovation and accelerate space exploration to unforeseen levels. While all nearly all asteroids hold value through their mineral-rich contents, not all asteroids are created equal; the ability to discern between metallic, carbonaceous, and silicate-rich asteroids, among others, is critical when searching for mining targets. One way of discovering asteroid composition is through spectroscopy – a method that measures light reflectance based on different wavelengths. As different materials reflected different amounts of light throughout the electromagnetic spectrum, an asteroid's reflectance to various wavelengths of light reveals much about its composition. Thus, DeMeo created the Bus-DeMeo (B-DM) asteroid taxonomy, a series of asteroid classifications based on their spectral data, indirectly grouping asteroids based on their estimated composition, each holding a different value for mining purposes (DeMeo, 2009). For example, water-abundant carbonaceous C-type asteroids may be valuable for lengthy trips where life-

sustaining substances such as water and phosphorous are critical. On the other hand, the rare mineral-rich M-type asteroids (the classification 16 Psyche falls under) hold the most monetary value for immediate extraction. Overall, the B-DM asteroid taxonomy system contains 25 unique asteroid classes, each differentiated solely by their spectral reflectance over different wavelengths. Fig. 1 displays aggregated data from 18 asteroids of the most common B-DM asteroid types and their spectra plotted on standardized graphs for comparison. Differences between some classes are not clear-cut, however; humans may struggle to distinguish between L and P-type asteroids, for example. While DeMeo did develop a mean spectrum for each class, the introduction of newly discovered asteroids daily warrants a new, more consistent classification system to distinguish minute similarities between asteroids to better classify type. Additionally, factors such as diameter, absolute magnitude, and albedo also play a key role in the composition of an asteroid, factors that are difficult to weigh individually and have not been considered by current methods of classification. A Convolutional Neural Network (CNN) was created in this paper to classify asteroids and provide a foundation for extremely precise and accurate estimation of asteroid composition.

In this paper, a description of a novel framework for collecting and processing asteroid spectral data will first be shared. Next, an explanation of CNNs is demonstrated alongside the specific architecture of the model used for this research. Different methods of data augmentation and aggregation are then discussed alongside their implications. The results of the CNN will next be shared across various trials. The significance of this research will then be discussed alongside conclusions and further research opportunities.

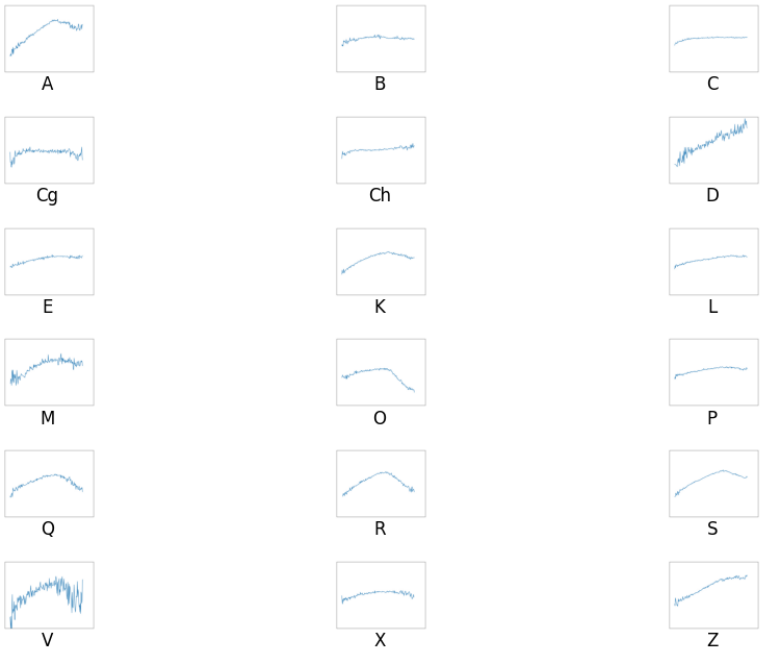

Figure 1: Spectra of asteroids that fall under 18 of the most common B-DM asteroid types. Each graph is scaled the same with the x-axis being the wavelength of light being from 0.4 µm to 1 µm. On the y-axis the reflectance normalized to unity at 0.55 µm plotted from 0.5 to 1.5 units.

2. Methods

Before being used for training, spectral data must first be preprocessed to be standardized and suitable for machine learning. Additionally, the model itself must be constructed.

2.1 Asteroid Concepts

To create a classification algorithm with more accuracy than previous methods, data beyond just asteroid spectra was used for training. This way, the model can find patterns within other factors that influence asteroid composition to obtain more insight on an asteroid's classification. Such factors are: diameter, absolute magnitude, and albedo. A brief explanation for each piece of auxiliary data is listed below alongside the impact it brings to asteroid composition.

Diameter – The diameter of an asteroid often hints at the composition. For example, more primitive, carbonaceous C-type asteroids such as Ceres and Pallas tend to be larger – Ceres is currently the largest discovered asteroid.

Absolute Magnitude (H) – The absolute magnitude is a measure of brightness; specifically, it describes how bright an object would appear if it were located 10 parsecs (about 32.6 light years) from the observer. It is calculated by the equation

$$\boldsymbol{M = m - 5\log_{10}(\frac{d}{10})}$$

where M is the absolute magnitude, m is the apparent magnitude (magnitude from the current observing distance), and d is the distance at which the apparent magnitude was observed at. A lower value equates to a brighter object. The standardization that absolute magnitude provides allows for consistent comparison between asteroids, allowing the machine learning model to

consider grouping brighter and dimmer asteroids, a factor which is influenced by that asteroid's composition.

Albedo – The albedo of an asteroid is a measure of reflectance. An albedo is a value from 0 to 1 that represents the portion of light reflected by an asteroid where 0 is perfect absorption (no reflected light) and 1 is perfect light reflection. Albedos offer a strong indication of asteroid composition. Carbonaceous C-type asteroids are notable for their darkness and low albedos, while metallic M-type and silicate-rich S-type asteroids tend to be very reflective, sporting high albedos.

With these sources of supplemental data into the model, the CNN can employ a multi-faceted approach to asteroid classification, allowing for a more versatile, accurate, and nuanced model.

2.2 Data

To be used for machine learning, large sums of spectral data of each asteroid type must be found and classified already. Thus, the first priority was finding an adequate dataset with both breadth of asteroid type and depth for each classification. MIT's Small Main-Belt Asteroid Spectroscopic Survey, Phase II (SMASSII) displayed both traits, containing 1341 main-belt asteroid spectra from 0.435 μm to 0.925 μm. Unfortunately, SMASSII did not classify the asteroids in their dataset. Thus, space-rocks – an open-source asteroid data repository in Python – was used to access the asteroid classification. Additionally, auxiliary data for each asteroid – diameter, absolute magnitude, and albedo – were accessed through space-rocks. The exact process of data generation began with the SMASSII spectroscopy files, each containing two data columns: the first represents the wavelength in μm and the second displays the corresponding

normalized reflectance (unified at 0.55 μm) of light to that certain wavelength. By graphing the corresponding points on a connected scatter plot with wavelength on the x-axis and normalized reflectance on the y-axis, a graph displaying the spectral characteristics of an asteroid is created. To find the known class of the graphed asteroid, the asteroid was searched in the space-rocks database to find the class of the desired asteroid. The asteroids in the SMASSII dataset were distinguished by the asteroid ID number found in each file name, each file representing an individual asteroid. The ID was then queried in the space-rocks database and the newly created graph is automatically labeled by that asteroid's class, diameter, absolute magnitude, and albedo. Fig. 2 illustrates the full data collection process.

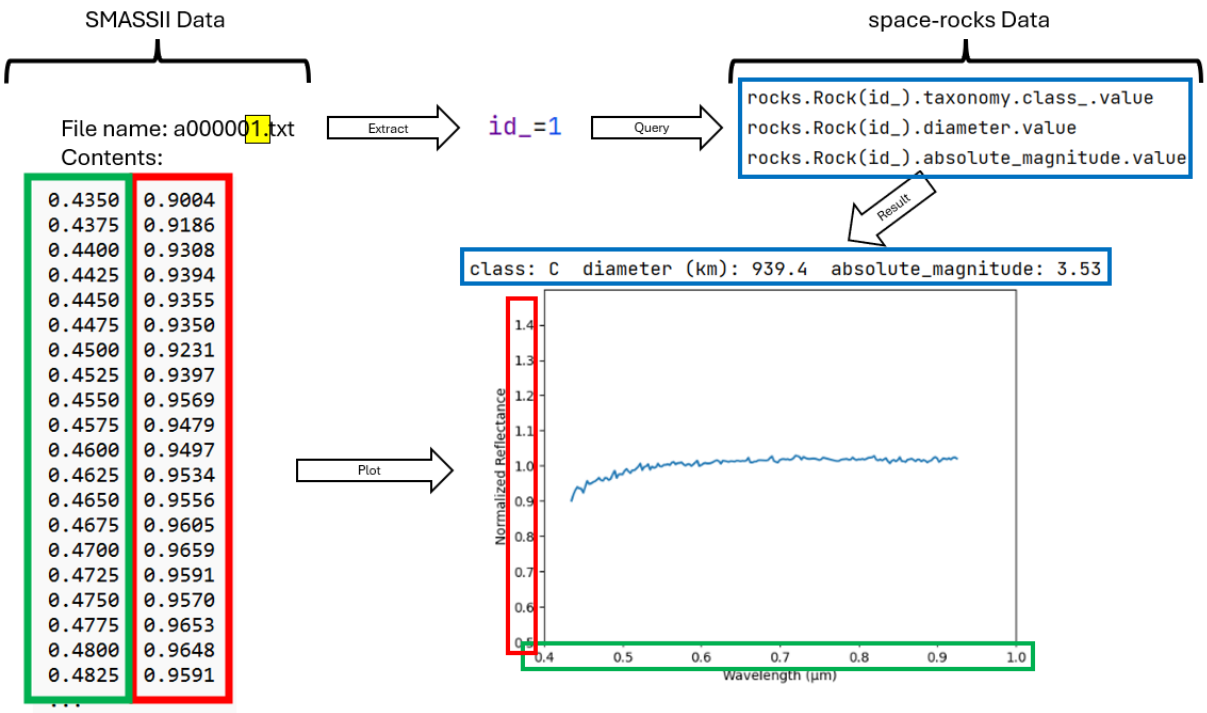

Figure 2: Map of the process of data aggregation. An example file is provided on the right. The asteroid number is extracted from the file name and queried through the space-rocks database to find the B-DM taxonomy, diameter, absolute magnitude, and

albedo of the asteroid. The asteroid type is then labelled to a graph that plots normalized reflectance vs. wavelength, valued by the two columns in the SMASSII data files.

A Python program quickly executed these tasks for the 1341 SMASSII files and, after removing the asteroids that were classless in the space-rocks database, resulted in 1263 unique asteroid graphs, the distribution of which are displayed below in Table 1. While only SMASSII data was used in this research, the dataset is easily expandible with this framework due to the standardized naming and data-format conventions of spectral asteroid data. Additionally, the dataset grows as new asteroids are classified and added the space-rocks database.

Class	A	B	C	Cg	Ch	D	E	K	L
Number	24	6	184	1	156	6	9	20	24

Class	M	O	P	Q	R	S	V	X	Z
Number	84	1	107	3	8	577	42	2	9

Table 1: Number of spectral graphs for each taxonomic class.

As seen in Table 1, certain classes are far more common than others. If this data were to be inputted in a model, it would result in unbalanced training and an ineffective model. Thus, data augmentation was used to effectively expand the dataset. To do so, a program was created to randomly select an asteroid from a desired class for augmentation, then access its original SMASSII raw spectroscopy data. The program then randomly alters normalized reflectance

values by ±0.01 units, preserving the overall spectral behavior (and therefore spectral type), but still altering the graph enough to act as new data within the same spectral class. Thus, new data is effectively simulated to balance the data for machine learning purposes. Fig. 3 displays the impact of data augmentation on the full dataset for class V asteroids.

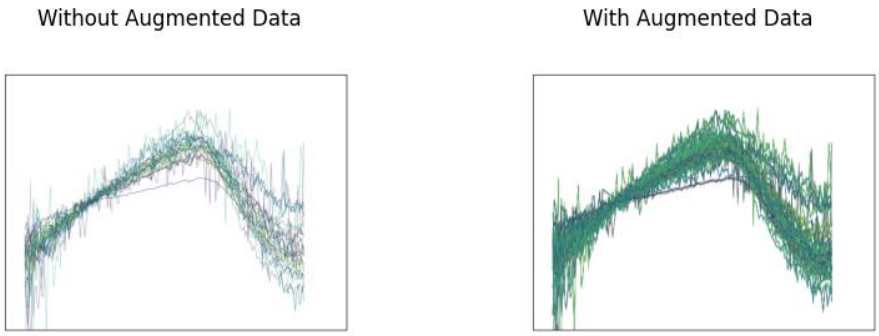

Figure 3: The left figure is a plot of all V-type spectra in the SMASSII dataset without augmentations. The right figure contains the same data as the left in addition to the 158 augmentations that form the custom dataset.

Similar operations were done to diameters, absolute magnitudes, and albedos, where augmented asteroid values were randomly altered from the original asteroid's values by ±1.5 km, ±1 units, and ±0.01 units, respectively. For the rare values that have not yet been documented in the space-rocks database, a value was estimated by taking an average of the statistic for that asteroid's taxonomic class, then randomly varying it by the above values depending on the statistic.

Each class was augmented until it had at least 200 graphs, and the S class was pared to 200 graphs. Before inputting into the model, the images were scaled to 128x128 resolution and made grayscale to make the model more efficient, thus finalizing the custom-built dataset associating asteroid spectral characteristic to their B-DM spectral type and, by extension, composition.

2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are machine learning models that branch off traditional machine learning in their ability to analyze images and extract features through convolution. For a CNN to learn how to classify, it first must be trained on images that are pre-classified to learn what constitutes to each class. After each epoch – a round of training – the CNN tests itself on new, different data from its training data; it predicts the class for each image and compares it to the true classification, a process called validation. The loss – or error – of the machine learning model during the validation phase of training is often used to determine its performance. A successful CNN is one that minimizes loss.

CNNs analyze images through convolution, a process where the CNN tunes various hyperparameters to extract features from image pixels.

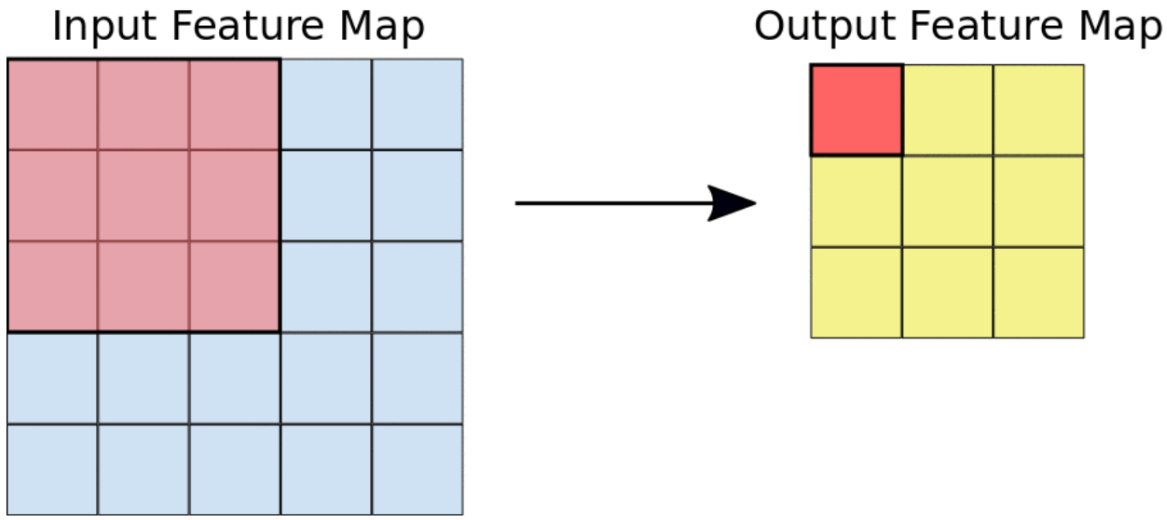

Figure 4 (Google, n.d.): Simplified map displaying convolution. The large red square is the convolutional filter corresponding to the smaller red square on the output feature map. The blue square is the original feature map.

Fig. 4 displays the premise of convolution. The red square represents a convolutional filter – essentially an array of numbers that the CNN tunes to manipulate images and extract features. This convolutional filter is passed over the original image – the blue square – and is combined with the original image through element multiplication. Each pixel of the original image is multiplied by the respective overlapping pixel of the filter. The resulting multiplications are added to form the resulting value in the output. As the convolutional filter slides rightward to reach the next set of filters, the resulting value will be in the next square on the right in the output feature map. Since the CNN has control over the multipliers within the convolutional filter, it can effectively manipulate the values within an image until it can discern between different images.

Another essential facet of CNNs is max pooling. Similar to convolution, a filter is run over a feature map. However, rather than executing calculations, a max pool filter simply outputs the maximum value within the filter. Similar to downscaling an image, max pooling is critical to CNNs for two main reasons. First, condensing the features of an image allows a model to get a better big-picture view of the image's features, allowing it to find large-scale patterns it is otherwise blind to at the cost of detail. Max pooling is analogous to a bird flying higher: the bird gains a better view of the full expanse of land but loses the precision and detailed look at ground-level. Second, max pooling helps reduce the computer power necessary to train the model, as reducing the feature map's size reduces the number of hyperparameters the model must train.

The final layer necessary for CNNs to classify images is the fully connected layer (also known as a "linear" or "dense" layer). Each neuron in a fully connected layer dissects and transforms input data according to an array of weights that are tuned by the CNN. Thus, each

neuron controls the shape of the output based on the array of weights it uses. For this reason, fully connected layers are critical for transitioning between image features and classification. In practice, a fully connected layer morphs the thousands of features extracted from images into the different classes desired.

2.4 Use of a CNN to process asteroid spectra

For this experiment, the CNN was built to analyze spectral graphs from asteroids and predict the B-DM asteroid type. To do so, the model – constructed in PyTorch – first used a convolutional layer of input 128x128x1 (128x128 pixels and 1 depth color channel for grayscale images) to extract detailed features. The feature map was next inputted into a max pool layer to broaden the scope of the model. The image was then convoluted again to extract big-picture patterns and max-pooled once again. Finally, the data was flattened and inputted into a fully connected layer, outputting 18 values corresponding to each asteroid class. Fig. 5 displays the map of the architecture. Using the Adam optimizer and cross entropy loss, the model was trained for 10 epochs before validation and training loss reached its minimum.

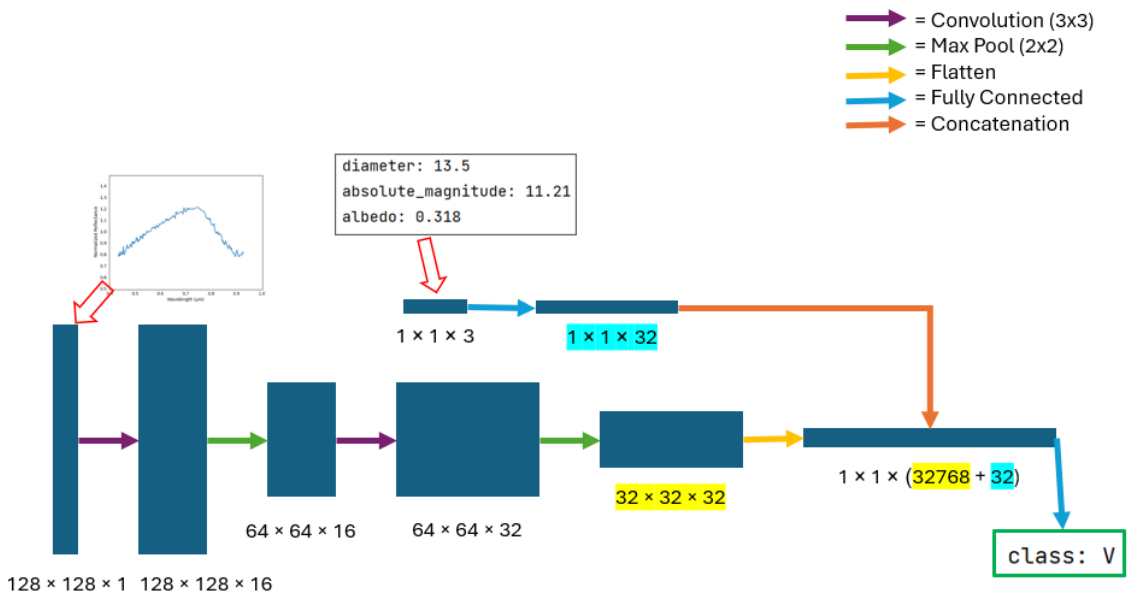

Figure 5: Map of the CNN architecture. The bottom row of boxes represents the image's path, and the top row of two boxes shows the numeric values' path. The two inputs are concatenated to form one large feature map, eventually being condensed to classify the asteroid.

3. Results

After training, the model was tested on the entire dataset and the accuracy, in percent, was calculated simply by the number of correct classifications divided by the size of the dataset. Overall, the model achieved 98.8% accuracy over the entire dataset. In comparison, the model achieved an 85.1% when the auxiliary data was omitted (only classifying based on spectral graph). Using the confusion matrix for the model with auxiliary data in Fig. 6, the specific accuracy of the model for classifying different asteroid taxonomies can be pinpointed.

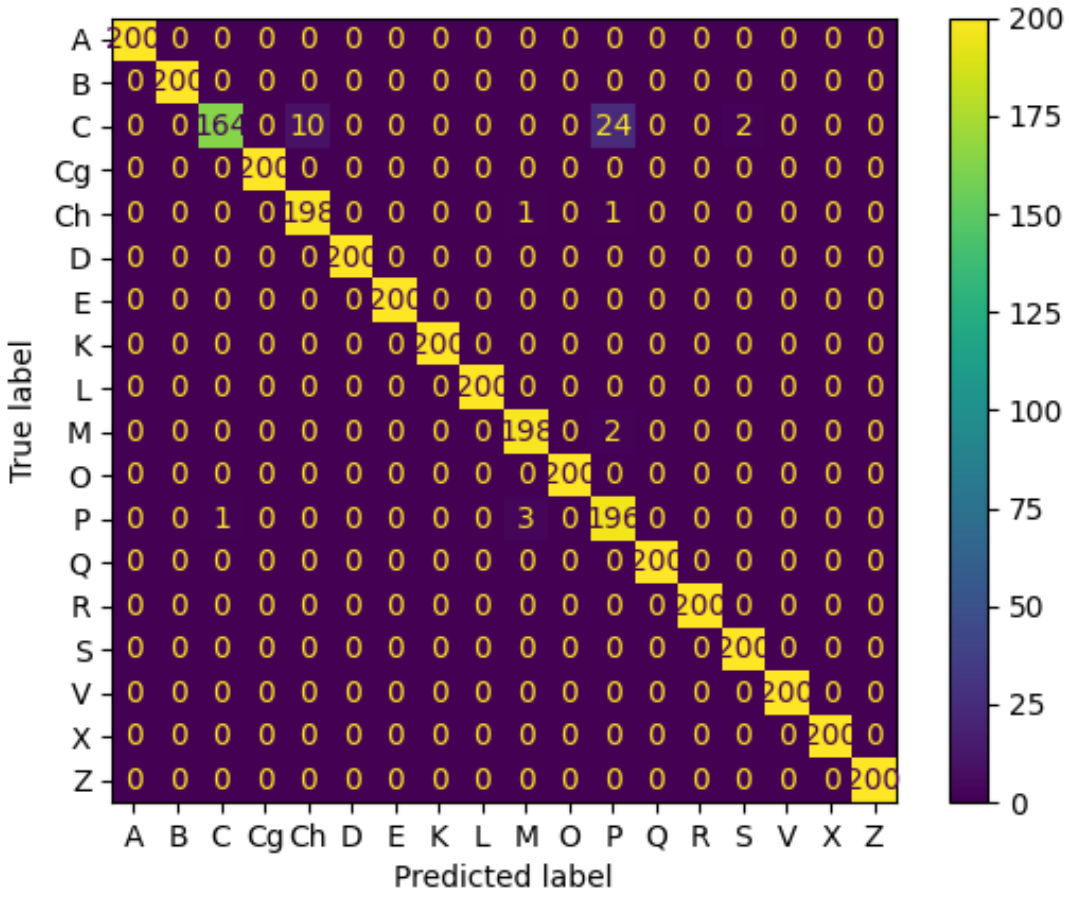

Figure 6: Confusion matrix of the CNN with auxiliary data. The x-axis shows the label the CNN predicted, and the y-axis displays the true label of the graph. The corresponding value is the number of occurrences.

According to the confusion matrix in Fig. 6, the model was very successful at most classifications, scoring perfect for many. Contrarily, the model had difficulty with the C class, only scoring about 82% correct, and missing mostly by incorrectly predicting the P class. This makes sense considering that both P and S class asteroids are largely made of carbon, resulting in a flat spectrum graph and low albedo and absolute magnitude values. This also opens the possibility that the model is predicting some asteroid compositions with more accuracy than the current consensus, which will be elaborated on in the conclusions.

Discussion and Conclusion

This research highlights the transformative potential of machine learning, specifically Convolutional Neural Networks (CNNs), in advancing asteroid classification beyond traditional approaches and provides a foundation for the future of asteroid identification. The novel approach of using multiple data points from a single asteroid to spectrally classify it was proven to be very successful. Moreover, this auxiliary data was proven to be a beneficial part of the model, experiencing a 13.7% increase in performance when used. However, since the data used in this experiment used traditional methods of unidimensional classification, there is a possibility that the model using auxiliary data was more accurate than the dataset it trained on. This is because current established asteroid types are unable to be empirically verified. Thus, factoring in diameter, absolute magnitude, and albedo may have given the CNN more insight on asteroid composition, insight that is unavailable to the current literature who exclusively uses spectroscopy to determine the consensus asteroid composition. This is represented by the CNN predicting various P-type asteroids to be C-type – two similar asteroids in composition and on a spectral graph, yet may yield different classifications when factoring in auxiliary data. While this is only speculation, it is certain that a CNN can surpass current methods of classification when provided enough data. For this reason, the model was built with the future in mind for simple, easy ramifications as the global dynamic dataset of asteroid data expands.

Another facet of the CNN presented in this paper is the standardized method for spectral data processing, allowing for any spectral study to be uploaded to the data generator constructed in this research. The data can be automatically converted to a spectral graph and can be classified for machine learning. This allows for a dynamic machine learning framework designed around future research and continual improvement. Future open-source availability for these algorithms

can encourage collaboration between institutions attempting to plan future space missions or develop projects regarding space mining.

The use of spectral graphs rather than the raw spectral numbers is another decision made in this paper that departs from the norm previous studies such as DeMeo et al (2009, 2019). The justification for this decision is compatibility. Often, different studies and data surveys use different equipment and therefore extract different data. Thus, sometimes data that traces out the same spectral graph may be slightly different if different increments or range of wavelength were used, which may confuse a CNN if data from multiple sources are used. As a result, a spectral graph – nearly standardized for all studies – was used to maximize the expandability of the CNN framework. Additionally, using images as model input gives CNNs the freedom for multiple levels of precision and pattern-recognition. Convoluting an image before and after max pooling, for example, allows a CNN to identify detailed features while also noticing general trends. This dual level analysis – both detailed feature identification and broad pattern recognition – is not feasible with purely numerical data, and this research showcases the versatility of CNNs in handling different data structures.

This research demonstrates a substantial leap in asteroid classification offering a more accurate, flexible, and future ready approach. Analysis of spectral data by CNNs sets a new standard in the field that promises significant growth as future data is generated and forwarded to the model. The end goal of future work to this research is to amend this CNN to predict asteroid composition with increased accuracy, allowing for more confident missions and mineral extraction, ultimately ending the horrors that are occuring today in the field of Earth’s mineral extraction.

References

Apparent and absolute magnitudes. (1998, April 10)

Binzel, R. P., DeMeo, F. E., Turtelboom, E. V., Bus, S. J., Tokunaga, A., Burbine, T. H., Lantz, C., Polishook, D., Carry, B., Morbidelli, A., Birlan, M., Vernazza, P., Burt, B. J., Moskovitz, N., Slivan, S. M., Thomas, C. A., Rivkin, A. S., Hicks, M. D., Dunn, T., … Kohout, T. (2019, May). *Compositional distributions and evolutionary processes for the near-earth object population: Results from the MIT-Hawaii near-earth object spectroscopic survey (MITHNEOS)*. NASA/ADS

Birlan, M., Popescu, M., Irimiea, L., & Binzel, R. (2016, October). *M4AST - a tool for asteroid modelling*. NASA/ADS

Bus, S. J., & Binzel, R. P. (2002). *Phase II of the small main-belt asteroid spectroscopic survey. A feature-based taxonomy*. NASA/ADS

DeMeo, F. E., Binzel, R. P., Slivan, S. M., & Bus, S. J. (2009, July). *An extension of the bus asteroid taxonomy into the near-infrared*. NASA/ADS. DeMeo, F. E., Polishook, D., Carry, B., Burt, B. J., Hsieh, H. H., Binzel, R. P., Moskovitz, ,Stephen M. Slivan,Schelte J. Bus

N. A., & Burbine, T. H. (2019, April). *Olivine-dominated A-type asteroids in the main belt: Distribution, abundance and relation to families*. NASA/ADS

Google. (n.d.). *ML Practicum: Image Classification | machine learning | google for developers*. Google

OpenStax. (n.d.). *13.1 asteroids*. Astronomy

Penttilä, A., Hietala, H., & Muinonen, K. (2021, May 7). *Asteroid spectral taxonomy using neural networks*. Astronomy & Astrophysics

Sohn, R. (2023, October 12). *Metal asteroid psyche has a ridiculously high "value." but what does that even mean?* Space.com