Journal of High School Science

Using machine learning to approximate asteroid composition with meteorite types

Trundle K.R[1], Garcia M.D[2]

Submitted: July 24, 2023, Revised: version 1, September 17, 2023
Accepted: September 22, 2023

Abstract
Studying the composition of asteroids is important to understanding the history and composition of our solar system. The spectral analysis of reflected sunlight serves to inform conventional asteroid composition research. Since asteroids have differing physical features, spectral data can result in biased and unreliable conclusions regarding an asteroid's composition. Using meteorites to study asteroid composition is a relatively reliable solution to this challenge, since the meteorite composition can be studied directly as opposed to studying the asteroids composition remotely. Therefore, finding a way to automate the connection between asteroid types and meteorite types is extremely beneficial to understand asteroid composition. We accomplished this using machine learning, or more specifically a densely connected neural network (DNN), to create AMCA, or the Asteroid Meteorite Connection Algorithm. The AMCA could find connections between asteroid and meteorite types using asteroid orbital and spectral characteristics. Our database was constructed using asteroid- and meteorite-type connections based on VISNIR (visual and near infrared) spectroscopy. After training, our algorithm achieved training and testing accuracies of 87.75% and 80.42% respectively. The asteroid SMASS spectral type, semi-major axis, eccentricity, perihelion distance, aphelion distance, and orbital period were the most influential in determining asteroid-meteorite type connections. The AMCA is anticipated to increase asteroid knowledge by using their orbital and spectral characteristics to determine which type of meteorite best represents the asteroid's composition.

Keywords
Machine Learning, TensorFlow, Astrogeology, Asteroids, Classification, Meteorites, DNN, SMASS classification, Orbital features

[1]Corresponding author: Kane R. Trundle, North Monterey County High School, 13990 Castroville Blvd, Castroville, CA 95012, USA. kanetrundle@gmail.com
[2]Mauricio Davalos Garcia, North Monterey County High School, 13990 Castroville Blvd, Castroville, CA 95012, USA. Mauriciogarcia0119@gmail.com

Introduction:

Studying asteroids is a critical and vital way to determine how our solar system has evolved and how it might change over time (1,2) One facet of asteroid research is studying their composition. Traditionally, asteroid composition and classification are based primarily on broadband color photometry and reflectance spectroscopy, which can be used to determine its surface composition. The challenge with these methods is that they use reflected sunlight from an asteroid, which can be altered due to its size, location, weathering, albedo, and temperature (2). These variations can result in bias and unreliable conclusions regarding composition (1). This is often why full surveys are carried out when studying an asteroid; however, instrumental, and physical anomalies are often difficult to remove (2). Thus, using other forms of measurement is anticipated to result in more accurate and reliable conclusions regarding an asteroid's composition. One solution to this issue is using meteorites to study asteroid composition. This additional form of measurement allows scientists to study both the interior and exterior composition of an asteroid remotely. However, for such a method to be useful, common attributes between asteroid and meteorite types need to be identified. Previous studies have been performed to find these connections; however, these are often found solely using visual and near-infrared spectroscopy. This results in similar issues caused by physical and instrumental anomalies. Therefore, finding a way to incorporate additional measurements to find these connections may result in more useful and accurate results.

Similar attempts to solve this issue have been made regarding the discovery of non-empirical connections between asteroid and meteorite reflectance spectra. For instance, Dyar, et al. used a linear regression algorithm to classify asteroids based on spectroscopic similarities between existing meteorite types (3). In addition, Miyamoto, et al. used various statistical methods including principal component analysis and other dimensional reduction algorithms to cluster meteorite and asteroid reflectance spectra (4). Both solutions leveraged unique approaches to discovering new connections between asteroids and meteorites. Building on this research, we designed a new approach for finding these connections. Instead of using only reflectance spectral data, which, as stated above is often unreliable due to space weathering and various instrumental differences, we incorporated asteroid orbital data into our algorithm to add dimensionality as well as include a different reference point for comparison. The use of orbital data allowed for new connections to be made; for instance, how certain asteroid semi major axis corresponded to certain meteorite types. These connections added more complexity to the model and included a unique way to categorize and connect data points. It is important to note that we decided not to exclude spectral data as the connections between asteroids and meteorites are primarily made through spectral analysis, and removing spectral data would make determining our model's accuracy difficult. The effect of removing the spectral data inputs from our model is presented in the Results and Discussion section.

When collecting the data for our project, we used a hybrid data set constructed from the NASA small body database and a dataset of asteroid-meteorite-type connections provided by Dr. Francesca DeMeo. These connections were based primarily on chi squared analysis, which calculated the distance between each asteroid and meteorite spectral graph to determine the closest asteroid-meteorite type match. The final dataset consisted of 1,437 asteroids, which we used to train the algorithm. When deciding what algorithm to use, we decided on a basic Keras Sequential Model using a ReLU function as an activation function for the input and hidden layers and SoftMax for the output layer. For the loss, we used categorical cross-entropy and Adam as our optimizer (we discuss the rationale for choosing the model in the Methods section).

To evaluate the model, we used Categorical Accuracy and Categorical Cross-Entropy to measure accuracy and precision. The loss function was based on the distance between the output values and the desired output. This distance was then used to calculate the loss and adjust the weights and biases of the algorithm. Based on the loss values and accuracy, we could determine how accurate the model was at predicting the corresponding meteorite type based on the physical characteristics of the asteroid, as well as what inputs were most influential in determining these connections.

Using a Deep Neural Network (DNN) is anticipated to find connections between various asteroid features and meteorite subtypes. These methods of finding connections may provide an alternative to finding connections based solely on spectral data. Hence, this project may increase the ability to find relationships between asteroids that may not have enough information to determine meteorite connections based on spectral data alone. Meteorite connections can thus be found with asteroids that are either too small or too far away to be classified using spectroscopy.

Methods

All tests were performed on a desktop computer with an 11th Gen Intel(R) Core(TM) i5-11400F @ 2.60GHz 2.59 GHz CPU, 32.0 GB (31.8 GB usable) of RAM, and an NVIDIA GeForce RTX 3060 GPU.

We began constructing our dataset by downloading the NASA small body database with Small Main-Belt Asteroid Spectroscopic Survey (SMASS) spectral types defined. The NASA small body database is a database of asteroids and comets within our solar system; this database includes orbital and spectral information (not all entries have spectral data, hence we filtered by SMASS type defined). The SMASS classification system is a taxonomic system based on an asteroid's color, albedo, and spectral features (5). We decided to use SMASS as our asteroid classification system as the data set provided to us by Dr. Francesca DeMeo used SMASS as the basis for asteroid-meteorite type connections. We decided to include spectral data. Without this data, the model was unable to reach desirable levels of performance due to the relatively small size of the database. (See the Conclusion section for recommended improvements).

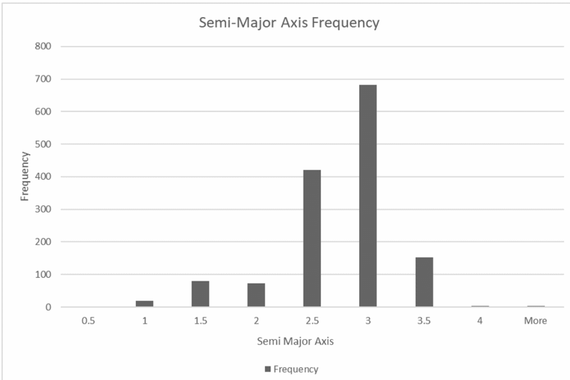

Figure 1. Distribution of the Semi-Major Axis frequency

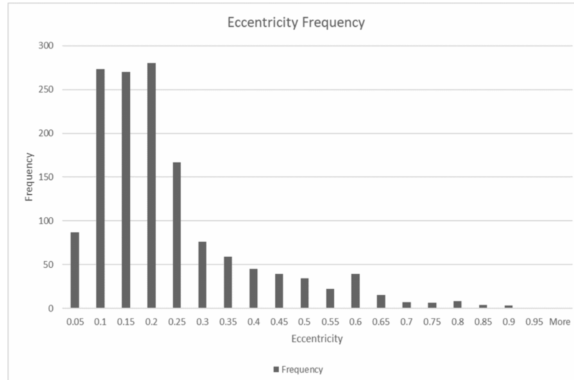

Figure 2. Distribution of the Eccentricity frequency

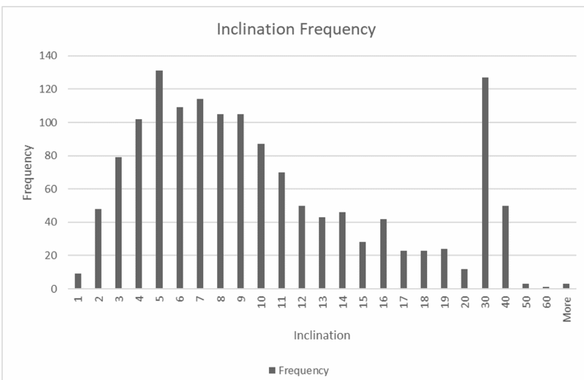

Figure 3. Distribution of the inclination frequency

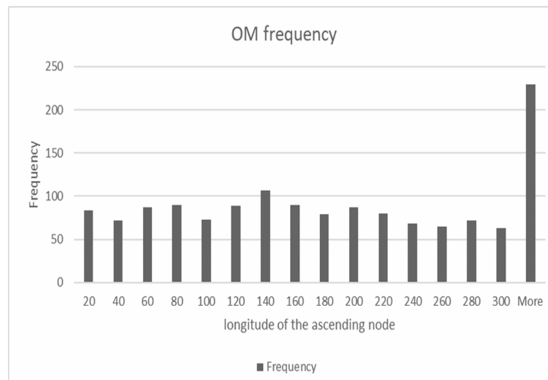

Figure 4. Distribution of the OM frequency

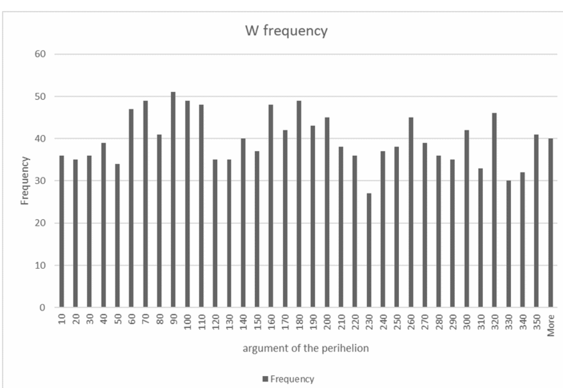

Figure 5. Distribution of the W frequency

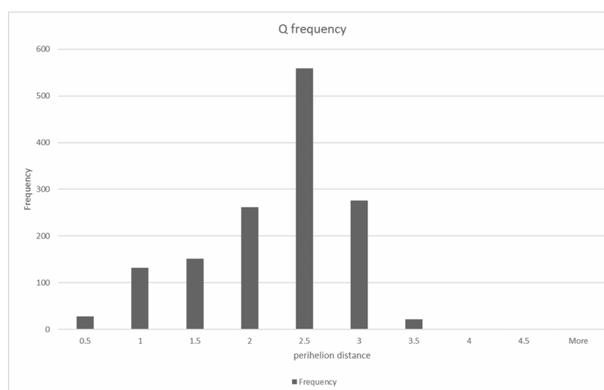

Figure 6. Distribution of the Q frequency

To determine which parameters to include in our model, we graphed the data based on frequency and type distribution—the graphs are shown in Figures 1-8. Graphing the parameters based on frequency helped us identify outliers, which we then removed from the dataset to ensure that the model would not overfit the data (6). During this process, we also removed parameters that had defined data for less than 100 asteroids, e.g., Albedo. Additionally, we observed patterns in the data. For instance, we noticed that the distribution of the semimajor axis matched the positions of the main asteroid belt (1.9-3.2 AU) and the Trojan asteroids (5.2 AU) (7). Additionally, we noticed that aphelion and perihelion distances had distributions closely centered around the mean, which illustrated the high concentration of asteroids in the main asteroid belt.

We also graphed asteroid orbital parameters in relation to their spectral type. As shown in figures 9-16, these orbital-spectral graphs show how spectral types align with each orbital parameter based on the relative frequency of each spectral type. We created these graphs using the Orange graphing software which uses machine learning and algorithms to assist in data visualization and compilation. Using the Orange graphing software, we determined the orbital-spectral combinations with the most distinct groupings: perihelion distance (q), aphelion distance (ad), and orbital period (per_y). These distinct groupings were determined heuristically and therefore our conclusions could be prone to error. We hypothesized that these parameters would have the most significant impact on model performance as their distinct clusters could potentially facilitate the identification of connections between orbital and meteorite spectral characteristics, which could enhance the model's performance and predictability.

After visualizing the parameters and removing outliers, we compiled the final dataset, which consisted of 1437 asteroids. Each asteroid was labeled, indicating its matched meteorite type, based on Francesca DeMeo's dataset. These labels were then one hot encoded using the Sklearn preprocessing library.

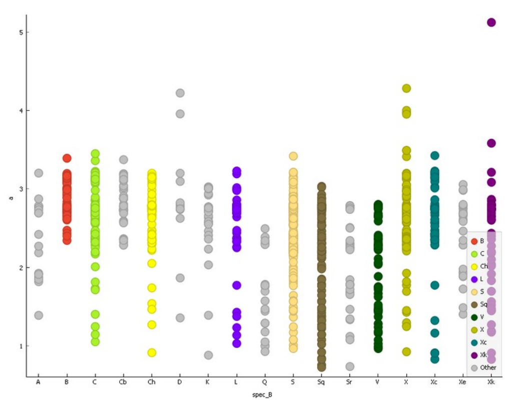

Figure 9: Spectral types with corresponding semi-major axes

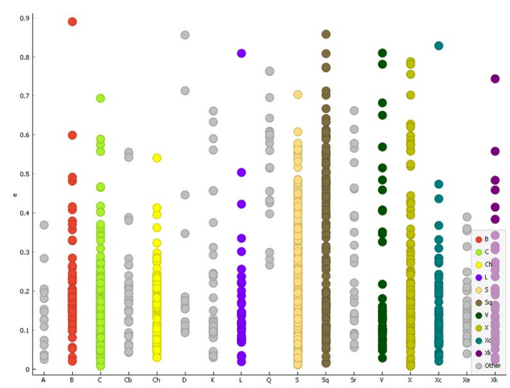

Figure 10: Spectral types with corresponding eccentricities

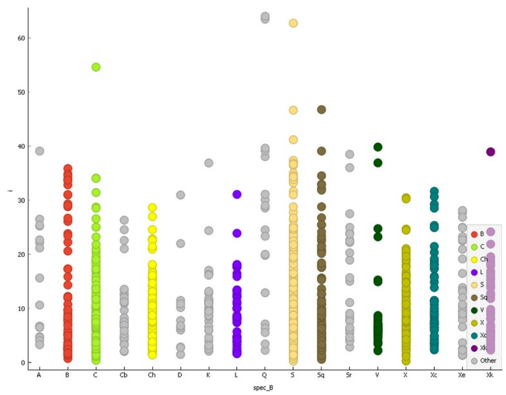

Figure 11. Spectral types with corresponding inclinations

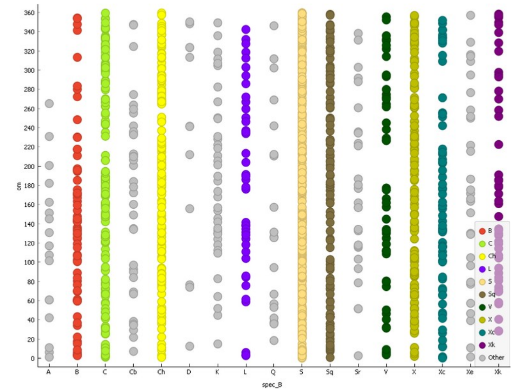

Figure 12. Spectral types with corresponding longitudes of the ascending mode

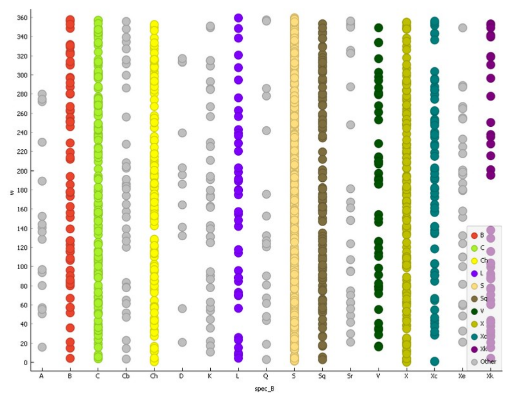

Figure 13. Spectral types with corresponding arguments of the perihelion

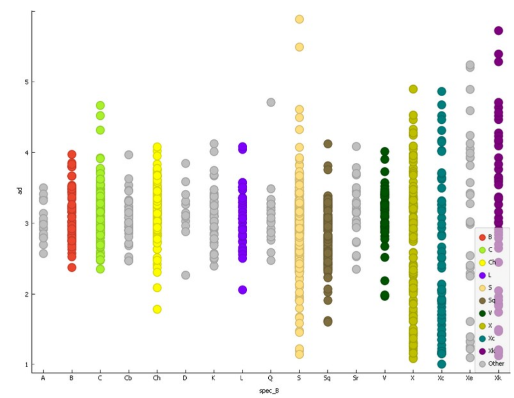

Figure 14: Spectral types vs perihelion distance scatter plot

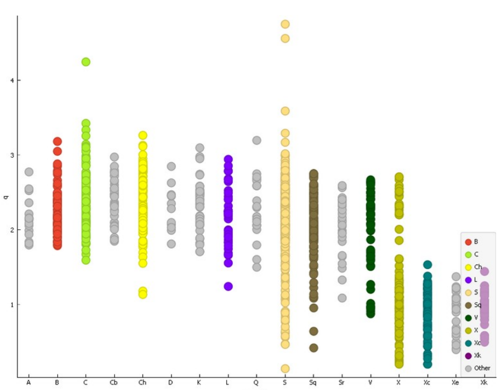

Figure 15. Spectral types with corresponding aphelion distances

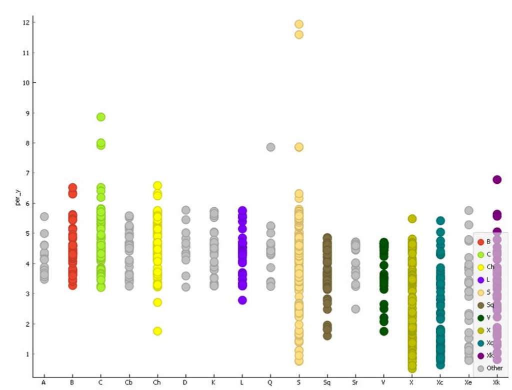

Figure 16. Spectral types with corresponding orbital periods

The inputs of our algorithm included 1437 asteroids. Each asteroid included its matching meteorite type, SMASS type, semi-major axis measured in AU (a), eccentricity (e), inclination measured in degrees (i), longitude of the ascending node measured in degrees (om), argument of the perihelion measured in degrees (w), perihelion distance measured in AU (q), aphelion distance measured in AU (ad), and orbital period measured in years (per_y). It is important to note that all our orbital parameters described different aspects of an asteroid's orbit. For example, eccentricity is the measurement of how circular an orbit is. An eccentricity of 0 describes a perfect circle and any number between 0 and 1 describes an elliptical orbit. An eccentricity of 1 creates a parabola which is not an orbit, but rather an escape trajectory (8). Inclination is the angle between the orbital plane (such as asteroids) and the ecliptic plane, which is an invisible plane that maps out the Earth's orbit around the sun (8). The longitude of the ascending node is an angle in the ecliptic plane that is between the inertial frame x-axis and the line through the ascending node (8). The argument of the perihelion is the angle between the ascending node and the perihelion (8). The perihelion and the aphelion describe the nearest and farthest distance respectively of an orbiting object from the sun (8). And lastly the orbital period is the time it takes for an object to complete one orbit (8).

For our spectral data we used the SMASS types. Our most common SMASS type was S and other sub-S types, which we originally thought would cause a decrease in model generalization, but after rigorous testing we concluded that the S types did not negatively affect our model's generalization (see Results & Discussion sections for model performance). Overall, we included 17 SMASS types: A, B, C, Cb, Ch, D, K, L, Q, S, Sq, Sr, V, X, Xc, Xe and Xk. These types were selected based on frequency, with types defining less than 10 asteroids being removed from the dataset. A-type asteroids are somewhat rare, found in the inner-belt and are classified as nearly pure olivine conglomerates based on their spectral characteristics (9). B-type asteroids are related to carbonaceous chondrites composed of carbonaceous minerals and phyllosilicates. High levels of carbonaceous minerals are responsible for B types' low albedo and broad spectral properties (10). C-type asteroids are the most common asteroid. C-types likely consist of clay and silicate rocks and have a dark appearance and are distinguished by their low albedo due to carbon making up a large part of their composition (11). D type asteroids have low albedos which makes them difficult to detect and study from earth, consequently their origin and composition is mostly unknown (12). K type asteroids are asteroids that have an S-type spectra in visible wavelengths and a C-type spectra in near infrared wavelengths (13). Q type asteroids are uncommon inner belt asteroids with a strong broad 1 micrometer olivine and pyroxene feature and spectral slope that indicates the presence of metal. S-types or "Stony"-types are asteroids that consist of silicate materials and nickel-iron (11). X-types or EMP asteroids are a group of asteroids that were at first identified as being indistinguishable due to their similar spectra, however they can have different compositions (14).

The meteorite subtypes used in our database were selected using previously discovered connections between asteroid and meteorite types by Dr. DeMeo and her colleagues. It is important to note that when referring to meteorite types we are really referring to meteorite subtypes. The meteorite subtypes within our data set include Aubrite, CM, CV, H, Howardite, IA, LL, Other I, and R. Aubrites are primarily composed of orthopyroxene enstatite. Their igneous origin separates them from primitive enstatite achondrites implying that they originated from an asteroid (15). CM is a group of carbonaceous chondrites that are identifiable by small chondrules and refractory inclusions, abundant fine-grained matrix, and ample hydrated minerals (16). CV chondrites are identified by their large chondrules which are mostly surrounded by igneous rims, large refractory inclusions, and abundant matrix. CV types can be divided into oxidized and reduced subgroups (17). H-type meteorites contain a high content of free nickel-iron which is common among the meteorites in this group. H chondrites have a weight of “25 to 31%” of total iron of which “15 to 19%” is found in its reduced form. (18). Howardite are a large group of polymictic-breccia achondrites that show mixtures of eucrites and diogenites, collectively known as HED. The main minerals of Howardites are pyroxene and Na-poor plagioclase (19). IA types are iron sub meteorite type, they contain high amounts of Ni (20,21). Other I is an ungrouped meteorite type under the division of iron meteorites (20,21). LL ordinary chondrites can be identified by their low siderophile element content, rather large chondrules and oxygen isotope compositions (22).

Table 1: Training Data (First 8 Rows)

Matching Meteorite subtype	**SMASS type**	**a**	**e**	**i**	**om**	**w**	**q**	**ad**	**per_y**
Aubrite	Xc	3.142689	0.71644	10.23781	252.1502	115.1444	2.55424	546.7388	4551.705
IA	X	2.970584	0.174424	4.29593	260.8609	189.1633	13.05102	18.42403	62.43281
H	S	2.338112	0.224503	6.11948	44.09449	107.6547	8.460001	18.84644	50.44998
H	S	1.567623	0.316999	9.454138	124.432	256.7521	4.558825	5.887311	11.93704
H	S	2.29417	0.397083	24.90146	288.0327	169.1684	4.749654	5.49335	11.59055
CM	C	1.403684	0.329157	13.58967	64.69052	50.03044	4.245147	4.317847	8.859346
H	S	2.112514	0.430138	6.908318	134.1524	160.1979	1.211506	7.232728	8.675681

After compiling our database, we then began writing code for our algorithm. We prioritized simplicity and high flexibility. After experimenting with different models, we decided on the Keras Sequential model because of its simple implementation and flexibility in terms of model customization (23). Additionally, we split our database into training, validation, and training batches. The training batch included 70% of the data, the validation batch included 20% of the data, and the testing batch included 10% of the data.

After choosing a model, we then downloaded the TensorFlow (24) and Sklearn (25) libraries to begin programming our network. We also imported Keras as a high-level API and used TensorFlow as a backend. As shown in table 2, The overall model consisted of three hidden layers, each using a ReLU function, one input layer, and one output layer, which used a SoftMax activation function. The input layer used Keras's Flatten function, which converts the input into a one-dimensional vector. The first hidden layer consisted of 25 neurons or nodes, the second and third of 15 and 10 respectively. The output matched the number of meteorite types (9 total). The size of the layers was based primarily on allowing the model to have high flexibility while also limiting its ability to overfit the data. For instance, when experimenting with very large layers, we observed near perfect accuracy and loss for our training but very low accuracy and high loss for both the validation and testing data. Using this information, we scaled our model down until we obtained layer sizes that yielded the best performance.

Table 2: Architecture of Sequential Model from Keras

Layer (type)	Output Shape	Param #
flatten(Flatten)	(None, 9)	0
dense(Dense)	(None, 25)	250
dense_1(Dense)	(None, 15)	390
dense_2(Dense)	(None, 10)	160
dense_3(Dense)	(None, 9)	99

In addition to the main structural features of the algorithm, we also focused on the optimizer and loss function. For the loss function, we chose categorical cross entropy, which measured the distance between the output and the one hot encoded label. For the optimizer, we chose Keras's Adam, which employed both classical stochastic gradient descent features as well as momentum gradient descent to avoid local minima.

In deciding what loss and activation function to use, we balanced memory efficiency and high accuracy/low loss. For our activation function, we chose ReLU for its higher convergence, which allowed for faster training (26). ReLU, or Rectified Linear Activation Unit (shown below), is a function that compresses the sum of all the weights and biases of the previous layer. This compression turns any input into a value between 0 and the input value (26). This avoids the vanishing gradient problem, which occurs when weight values are either very high or very low, in turn decreasing gradation and preventing the algorithm from learning properly (26). Additionally, the ReLU introduced sparsity into the model, which makes certain weights zero, allowing the model to focus on important parameters while avoiding overfitting and increasing predictability (26). When comparing ReLU to other activation functions such as Sigmoid or Tanh, we observed faster training times and lower computational costs with the former (26).

$$R = max(0, x)$$

$$x = \sum of\ weights \wedge biases$$

For our output layer, we used a SoftMax function (shown below), a type of logistic function that outputs a probability distribution over the meteorite types based on the output layer's output vector [output vector size is (0,9)] because there are nine meteorite types (27). The output of SoftMax is a vector of all the probabilities of each possible outcome or class, in our case meteorite types. Because SoftMax is continuous and differentiable at every interval, the optimizer can back propagate and adjust the models' weights and biases.

$$S(x) = \frac{e^{y}}{\sum_{j=1}^{n} e^{y_j}}$$

$$y = i^{th} element \in inputvector$$

$$\sum_{j=1}^{n} e^{y_j} = \sum of\ outputvales, adds\ 1$$

For the loss function, we used Categorical Cross entropy (shown below), which takes the output SoftMax probabilities as input and measures their distance from the desired output or the one hot encoded label vector (28,29).

$$CE = -\sum_{i=1} T_i \log(S_i)$$

$$T_i = i^{th} trueonehotencodedvalue$$

$$S_i = i^{th} outputtedsoftmaxvalue$$

$$-\log(S_i) = -logappliedsoftmaxinput$$

For our optimizer, we used Adam. Adam is a powerful and widely used optimizer in Deep Learning (30). Adam leverages adaptive learning rates to find unique learning rates for each parameter (30). It does this by using squared gradients to adjust the learning rate and using a moving average as the gradient. Overall, we selected Adam for its memory efficiency and momentum-based optimization (30). The performance of the optimizer can be viewed in the Results and Discussion section.

Results and Discussion

After constructing the algorithm, we began training and testing it. We tested our algorithm's performance using Categorical Cross Entropy and Keras Accuracy (23) (see the Methods section). We trained our algorithm for 80 rounds of 500 epochs and tested it using our testing database (see Methods section on data splitting). In addition, we also used a validation dataset to measure how accurate the algorithm's predictions were while training. We used this information to determine whether the dataset was overfitting while training.

For our first rounds of training, we included all our parameters, and after 80 rounds of 500 epochs, the algorithm achieved an 87.75% training accuracy and an 80.42% average test accuracy. To calculate the test accuracy, we took an average of 12 testing rounds with 1 randomized testing batch using our testing data set. This approach functioned like k-fold cross-validation; however, we did not split our testing data into k batches. Instead, we took the fluctuations in testing accuracy and loss output by the model and calculated the average and standard deviation to ensure the testing metrics were reliable (see Table 3 for averages and standard deviations). (All metrics of initial performance are shown in Table 3). Additionally, we graphed the model's Accuracy and Test data, shown in figures 17-18, to see whether the layer size and activation function were affecting training. We also created a confusion matrix and a Receiver Operating Characteristic Curve (ROC), with area under the curve (AUC) defined, for each meteorite type, as shown in figure 19-20. To create the confusion matrix and the ROC graph, we used the sklearn predict function using the testing dataset.

The graphs of accuracy and loss showed fast convergence as well as small fluctuations in accuracy and loss values. Additionally, the distance between training and validation loss and accuracy showed that the model performed with a relatively high prediction level and low overfitting. The algorithm was able to successfully learn the data set and make predictions with relatively high accuracy.

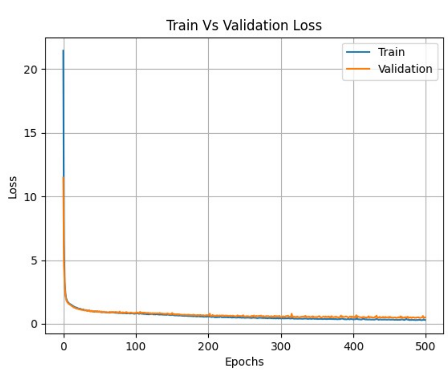

Figure 17. Validation loss across epochs

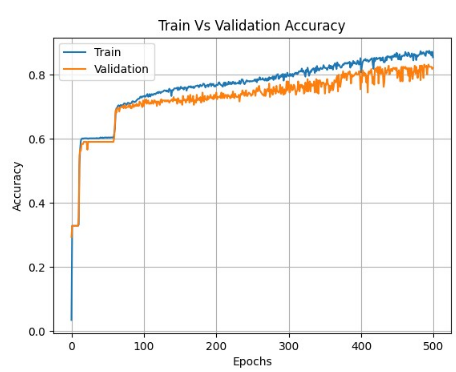

Figure 18. Validation accuracy across epochs

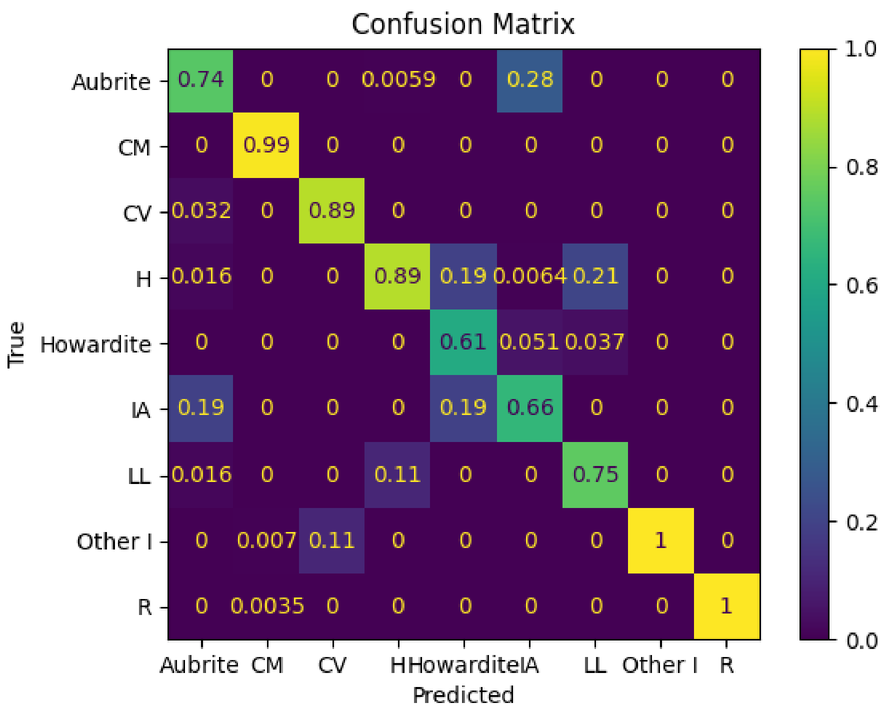

Figure 19. Confusion matrix

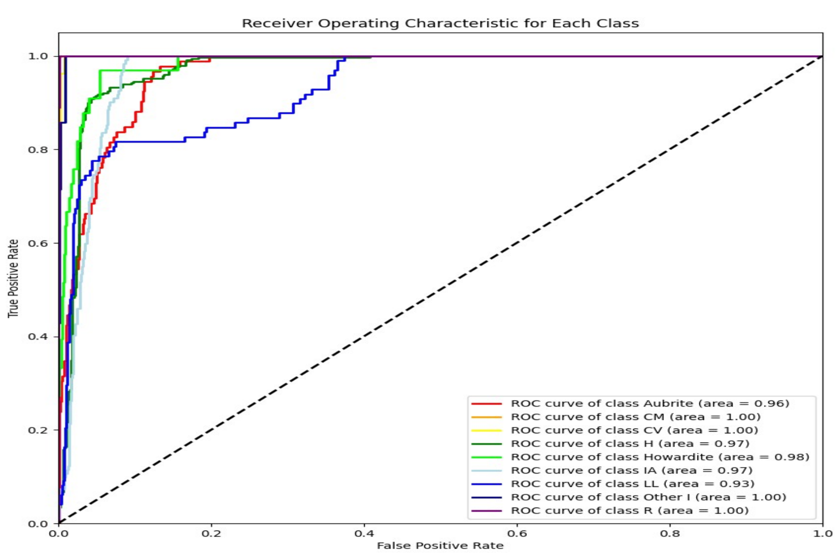

Figure 20. Receiver Operating Characteristic curve

From the ROC graph, we were able to identify the model's performance based on each classes' sensitivity and specificity for each classification threshold value (essentially where the minimum value needed for an asteroid to relate to a certain meteorite type). Based on the ROC curves, our model performed significantly better than the Random Classifier curve (the gray dotted line in the middle of the graph). This implied that our model performed significantly better than random chance indicating that it was able to learn from the dataset and accurately predict asteroid-meteorite type connections. Additionally, we were able to assess the model's ability to predict different meteorite types. This was important as our dataset was imbalanced and not all meteorite types were equally represented.

From the confusion matrix, we were able to pinpoint what meteorites the model had difficulty classifying. The model had the most difficulty correctly identifying Howardite and IA meteorite types, most often confusing IA with Aubrite and Howardite with H and IA.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

To further analyze our model's accuracy, we calculated the sensitivity and specificity for each meteorite type using the same prediction data from our confusion matrix and ROC graph. We calculated the specificity and sensitivity using their respective formulas (shown below) for each meteorite type. Calculating the specificity and sensitivity for each meteorite type gave us further insight into the specifics of our model's accuracy regarding the true positive rate and the true negative rate. From our calculations we were able to determine that our model was competent at predicting CM, CV, Howardite, Other I, and R. Interestingly the model was better at identifying which meteorite types did not correspond to a certain asteroid (given by the specificity or true negative rate). This may potentially be an upside to our model as instead of directly finding a corresponding meteorite, our model could instead eliminate the meteorites that did not correspond to the characteristics of the input asteroid.

Table 3: Initial Performance (Without Any Parameter Modification)

Training Loss	**Training Accuracy**	**Validation Loss**	**Validation Accuracy**	**Avg. Test Loss & (Standard Deviation)**	**Avg. Test Accuracy & (Standard Deviation)**
0.3185	87.75%	0.4163	82.87%	0.4499 (0.0254)	80.42% (1.5415)

Table 4: Specificity and Sensitivity for Each Meteorite Type

	Aubrite	CM	CV	H	Howardite	IA	LL	Other I	R
Sensitivity	0.71	0.98	0.91	0.70	0.84	0.84	0.76	0.96	1.00
Specificity	0.96	1.00	0.99	0.98	1.00	0.95	0.95	1.00	1.00

Overall, the model's performance and ability to make accurate predictions regarding what meteorite type would best represent the asteroid's composition was relatively high. In future iterations, we would like to include more SMASS types as well as meteorite types (see the Conclusion section for more information regarding future improvements).

To further analyze our model, we re-trained and tested the model using only spectral data. After training and testing the model, the model demonstrated 100% training accuracy and near 0 loss. This implied that the model was nearly perfect. We assumed this was due to the model's labeling system which was based on previously discovered asteroid-meteorite connections. This allowed the model to create a 1-1 correspondence between asteroid-meteorite, resulting in near error-free classification. We must state that the limited size of our data restricted our ability to make absolute claims regarding this method. Furthermore, the purpose of our research was to discover whether using orbital characteristics in addition to spectral classification represented a viable method (see the Conclusion section for more information regarding this issue). Therefore, the spectral only method was not directly relevant to our approach.

After testing the SMASS only model, we decided to test the effect of the individual orbital parameters on the model's predictions. Additionally, we decided to remove SMASS as well, even though we tested it alone prior, to see how well the model would perform with only orbital data. To do this we removed each parameter at one time to observe its effect on training, validation, and testing losses (see Table 4). We based the effect of each parameter on how far the loss values were from the standard error of the mean (SEM) of the loss values from the initial training. We decided to base our evaluation using loss values, because this represented a measure of how far the output was from the desired output and therefore provided a more accurate depiction of model performance (5).

When evaluating the effect of each parameter, if the loss values were below the standard error, we designated that parameter to have a negative effect on the model's performance since it decreased the loss (if the loss became lesser, this implied that the parameter was keeping the loss higher when it was included, i.e., lowering its performance). If the loss values were above the standard error, we designated the parameter to have a positive effect on the model's performance as it increased the loss (if the loss became higher, this implied that the parameter was keeping the loss lower when it was included in the first training round, i.e., increasing its performance). If the loss values with the parameter removed were like those with the parameter included, we designated it to have a neutral effect on the dataset.

Table 5: Model Performance (With Parameter Modification)

	SMASS type Removed	**a removed**	**e**	**i**	**om**	**w**	**q**	**ad**	**per_y**
Training Loss	1.2603	0.3601	0.4021	0.3047	0.1949	0.2063	0.4019	0.3203	0.5768
Training Accuracy	54.33%	86.53%	83.88%	88.8%	92.83%	92.04%	86.53%	87.40%	78.22%
Validation Loss	1.4552	0.5225	0.5670	0.4537	0.4431	0.3407	0.4802	0.6611	0.6744
Validation Accuracy	46.15%	79.72%	75.52%	82.52%	83.22%	85.31%	79.08%	79.37%	72.73%
Avg. Test Loss	1.3897	0.5544	0.6041	0.4602	0.4402	0.4039	0.5536	0.7800	0.7344
Avg. Test Accuracy	48.25%	77.62%	73.43%	82.33%	83.22%	83.31%	76.02%	77.62%	69.93%

Table 6: Parameter Standard Error Difference

	SEM of original loss values	Parameter removed							
		SMASS	a	e	i	om	w	q	ad	per_y
Training Loss distance	0.3304 ±0.0138	+0.9161	+0.0159	+0.0579	-0.0119	-0.1217	-0.1103	+0.0577	0	+0.2602
Validation Loss distance	0.4763 ±0.0183	+0.9606	+0.0279	+0.0724	-0.0137	-0.015	-0.1173	0	+0.1665	+0.1798
Testing Loss	0.4941 ±0.0267	+0.8689	+0.0336	+0.0833	0	-0.0272	-0.0541	+0.0128	+0.2592	+0.2136
Caused over fitting	N/A	1	1	0	0	2	1	1	2	0
Average Deviation	N/A	+0.9152	+0.0258	+0.0712	-0.0085	-0.0546	-0.0939	+0.0235	+0.1419	+0.2179

For testing the loss standard error, we took the mean of 10 test averages. The bottom number represents SEM. (+) implies over SEM, (-) implies under, and (0) implies in between. For overfitting, column 0 represents little to no overfitting, column 1 represents little to medium overfitting, and column 2 represents high overfitting (based on visual inspection of loss graphs). The average deviation was calculated by taking the average of the training, validation, and testing loss distances.

Based on the results of removing parameters, we found that SMASS type, semi-major axis (a), eccentricity (e), perihelion distance (q), aphelion distance (ad), and orbital period (per_y) had a positive effect on model performance. In order of impact on loss, SMASS had the largest effect with an average deviation of +0.9152; hence SMASS type played the largest role in model performance in terms of loss. The orbital period had the second largest impact with an average deviation of +0.2179; hence the orbital period had the second largest effect on model performance. The orbital period was followed by aphelion distance (+0.1419), eccentricity (+0.0712), semi-major axis (+0.0258), and perihelion distance (+0.0235). The parameters that had a negative impact on model performance were inclination (i), longitude of the ascending node (om), and argument of the perihelion measured in degrees (w). The longitude of the ascending node had the largest negative effect on model accuracy.

We noticed that both the longitude of the ascending node and the aphelion distance produced significant overfitting. We are unsure as to why this may be, but we hypothesize that it may be an anomaly in the data or due to the model's structure and having fewer parameters. However, this does not explain why other parameters did not yield similar overfitting patterns. We concluded that more testing would be needed (see the Conclusion section for more information).

Overall, we determined that the parameters that were most influential on our model's performance were SMASS type, semi-major axis, eccentricity, perihelion distance, aphelion distance, and orbital period. Aphelion distance and orbital period did have a relatively substantial impact on model performance, which matched our previous predictions. Surprisingly, our original hypothesis that perihelion distance would have a large impact on the model's performance was disproven by our results. We are unsure of the cause, especially since it had one of the most distinct groupings. However, we assume that this may also be an anomaly or an effect of the model's architecture, in which removing certain parameters may have caused underfitting.

Conclusion

In conclusion, we were able to create an algorithm that found connections between asteroids and meteorites using both asteroid orbital and spectral features. For this project, we collected a series of asteroids and meteorite samples (SMASS types) from professional websites and sources and used an algorithm to group them. The process of matching asteroids and meteorites is often done through spectral analysis; however, spectral analysis can be unreliable due to differing albedos and distances. We hence trained the algorithm to find connections between the orbital and spectral characteristics between asteroid and meteorite types instead of using spectral analysis alone. Based on the results, our algorithm could more accurately predict asteroid composition and add to the toolkit of existing methods.

To address the results of removing all orbital and/or spectral data from our model, we would first like to preface the limitations of our data. Due to the relatively small number of data points in our database, the performance of our model and the conclusions we could extrapolate from our results are limited. Therefore, the reasons behind the significant difference in performance between the model with no orbital data and the model with no spectral data cannot be determined absolutely. Furthermore, the model with only spectral data defeats the purpose of our model using both orbital and spectral data, which is to move away from spectral data due to challenges related to asteroid spectral data. This means that our model using both orbital and spectral data, which may be technically worse than the model with only spectral data, is still useful as it incorporates non-spectral data and may imply that a model using only orbital data is possible. Thus, the need for spectral data from asteroids may not be necessary when finding relevant connections. Therefore, the value of our model is not only its performance but also the fact that it can perform relatively accurately in the first place. It is important to note that the non-spectral and hybrid orbital models are less intuitive than the actual connections between asteroids and meteorites, as their connections are based only on composition similarities. As such, the connections found between non-spectral, and hybrid orbital models are purely mathematical.

In addition, the non-intuitive findings concerning the ascending node and the aphelion distance were most likely due to the model's architecture and random data sorting, which took place prior to each training cycle. Future tests may be performed to determine the causes of these deviations.

Looking forward, our model may be improved using a more robust and/or increased data set, different models, and a longer training period. Using a more robust dataset may allow the model to find more complex patterns as well as increase model reliability and predictability. A more robust and larger dataset may also determine whether an algorithm using both orbital and spectral data and or an algorithm using only orbital data is technically worse than an algorithm that groups only using spectral data.

Acknowledgments
We would first and foremost like to thank Dr. Francesca DeMeo for providing us with her data set as well as recommendations and ideas for our research and paper—we are deeply thankful for her support. Additionally, we would like to thank our teacher Heather Tynan for supporting our research as well as giving us the much-needed mathematical knowledge for this ambitious project.

References:

1. DeMeo, F. E., et al. (2022, March 12). *Connecting Asteroids and Meteorites with Visible and Near-Infrared Spectroscopy*. Icarus. https://www.sciencedirect.com/science/article/abs/pii/S0019103522000896?via%3Dihub

2. Krämer Ruggiu, et al. (2021, March 1). *Visible-Infrared Spectroscopy of Ungrouped and Rare Meteorites Brings Further Constraints on Meteorite-Asteroid Connections*. Icarus. https://www.sciencedirect.com/science/article/abs/pii/S0019103521000828?via%3Dihub

3. M. Darby Dyar, et al. (2023, December). *A machine learning classification of meteorite spectra applied to understanding asteroids*. Icarus. https://www.sciencedirect.com/science/article/abs/pii/S0019103523002956?via%3Dihub

4. Hideaki Miyamoto, et al. (2018). *Reflectance spectra of Asteroids and Meteorites: their classifications and statistical comparisons*. IOP. https://iopscience.iop.org/article/10.1088/1742-6596/1036/1/012003

5. DeMeo, F. E., et al. (2009, February 9). *An extension of the Bus asteroid taxonomy into the Near-Infrared*. Internet archive: Wayback Machine. https://web.archive.org/web/20140317200310/https://www.tara.tcd.ie/bitstream/2262/43276/1/PEER_stage2_10.1016/j.icarus.2009.02.005.pdf

6. Khamis, A., et al. (2005, January). *The Effects of Outliers Data on Neural Network Performance.* researchgate.net. https://www.researchgate.net/publication/26568300_The_Effects_of_Outliers_Data_on_Neural_Network_Performance

7. Brownlee, D. E. (2007). *Trojan asteroids*. Trojan Asteroids - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/physics-and-astronomy/trojan-asteroids#:~:text=Comets%20are%20generally%20considered%20to,with%20Jupiter%20near%205.2%20AU.

8. NASA. (n.d.). *Glossary*. NASA. https://cneos.jpl.nasa.gov/glossary/

9. Reddy, V. (2005). *Mineralogic and temperature-induced spectral investigations of A-type ...* lpi.usra.edu. https://www.lpi.usra.edu/meetings/lpsc2005/pdf/1375.pdf

10. Alí-Lagoa, V. (2013, June 5). *Physical Properties of B-type Asteroids From Wise Data*. aanda.org. https://www.aanda.org/articles/aa/pdf/2013/06/aa20680-12.pdf

11. NASA. (2021, July 19). *Asteroids*. NASA. https://solarsystem.nasa.gov/asteroids-comets-and-meteors/asteroids/in-depth/#:~:text=The%20C%2Dtype%20(chondrite),silicate%20materials%20and%20nickel%2Diron

12. Gartrelle, G. M. (2017). *Spectral variations of D-type asteroids at different ... - USRA*. hou.usra.edu. https://www.hou.usra.edu/meetings/lpsc2017/pdf/1088.pdf

13. Granahan, J. C., Smith, G., & Bell, J. F. (1993, March). *New K type asteroids*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/1993LPI....24..557G/abstract

14. Clark, B. E., et al. (2004, July 29). *Spectroscopy of X-type Asteroids.* Iopscience. https://iopscience.iop.org/article/10.1086/424856/pdf

15. The Meteoritical Society. (n.d.). *Recommended classifications*. Meteoritical bulletin: Recommended classifications. https://www.lpi.usra.edu/meteor/metbullclass.php?sea=Aubrite

16. Mindat. (n.d.). *CM chondrite meteorite*. mindat.org. https://www.mindat.org/min-49552.html

17. Mindat. (n.d.). *CV chondrite meteorite*. mindat.org. https://www.mindat.org/min-49577.html

18. *Ordinary Chondrites*. Meteorite.fr. (n.d.). http://www.meteorite.fr/en/classification/ordinarychon.htm

19. Mindat. (n.d.). *Howardite meteorite*. mindat.org. https://www.mindat.org/min-49878.html

20. Benedix, G. K., et al. (1998, February 3). *A petrologic study of the IAB iron meteorites: Constraints on the formation of the IAB-Winonaite parent body.* NASA. https://web.archive.org/web/20120301102551/http://trs-new.jpl.nasa.gov/dspace/bitstream/2014/19251/1/98-0545.pdf

21. NASA. (n.d.). *Meteorite Classification Chart*. NASA. https://curator.jsc.nasa.gov/education/classification.cfm

22. Mindat. (n.d.). *LL chondrite meteorite.* mindat.org. https://www.mindat.org/min-49745.html

23. Team, K. (n.d.-b). *Keras Documentation: The Sequential Class*. keras.io. https://keras.io/api/models/sequential/

24. Abadi, M., et al. (2015, November 9). *Large-Scale Machine Learning on Heterogeneous Distributed Systems*. tensorflow.org. https://www.tensorflow.org/static/extras/tensorflow-whitepaper2015.pdf

25. Bioinformatics Laboratory, U. of L. (n.d.-b). *Documentation*. Orange Data Mining - Data Mining. https://orangedatamining.com/docs/

26. Agarap, A. F. (2019, February 7). *Deep Learning Using Rectified Linear Units (ReLU)*. arXiv.org. https://arxiv.org/abs/1803.08375

27. Koech, K. E. (2021, November 18). *Softmax Activation Function-How It Actually Works*. Medium. https://towardsdatascience.com/softmax-activation-function-how-it-actually-works-d292d335bd78

28. Scikit-Learn. (n.d.). *6.3. Preprocessing Data*. scikit-learn.org. https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing

29. Koech, K. E. (2022, July 16). *Cross-Entropy Loss Function*. Medium. https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e

30. Kingma, D. P., & Ba, J. (2017, January 30). *Adam: A Method for Stochastic Optimization*. arXiv.org. https://arxiv.org/abs/1412.6980