Emission Line-Independent Method to Classify Seyfert Galaxies

By Andrew Rebello, Vineet Burugu, Neerav Mula

Author Bio

Andrew Rebello is a junior at Staples High School. Andrew has been curious about Astronomy for longer than he can remember -- in fact, he still falls into awe when viewing the night sky. He is interested in pursuing Engineering and Physics as well as conducting more research in the future.

Neerav Mula is a junior at Round Rock High School. Neerav has been fascinated by Astronomy ever since he could read about it. He is also interested in History, Saxophoning, Math, and Computer Science.

Vineet Burugu is a senior at Westwood High School. Vineet loves science. His favorite activity is stargazing through his telescope and viewing faraway planets and galaxies.

Abstract

This study aims to classify Seyfert Type 1 and Seyfert Type 2 galaxies by differences other than the ratio of the strengths of their emission lines, which may allow researchers to discover something new about them that could not be discovered from spectral lines alone. We made use of 43,029 Seyferts from SIMBAD and used Python to analyze 5 properties of each galaxy: spatial distribution, redshift, color-magnitude, morphology, and luminosity. Analysis was done via inspection of graphs and descriptive statistics. Significant differences were found in luminosity, redshift, and color-magnitude. Based on these differences, we trained Decision Tree and Random Forest models to classify a given set of Seyferts as Sy1 and Sy2. The classification was accurate for 76.5% and 78.8% of the testing set for the respective models. Based on our findings, it can be concluded that our model could provide information about Seyfert properties independent of their emission lines.

Keywords: Seyfert Galaxies, Machine Learning, Active Galactic Nuclei, Galaxy Classification, Computational Astrophysics, Astronomy, Emission Lines, Decision Tree, Random Forest

Introduction

Seyfert Galaxies are a particular type of Active Galactic Nuclei(AGN). AGNs are nuclei of galaxies that spew high amounts of energy in the form of electromagnetic radiation. Such galaxies derive their intense activity from the matter of an accretion disk that surrounds a massive black hole. When this matter falls towards the black hole, friction heats it up, causing intense light emissions.

Seyfert Galaxies, discovered in 1943 by Carl K. Seyfert after analyzing NGC 1068 and galaxies with similar properties, have a bright nucleus and wide hydrogen emission lines (Seyfert, 1943). The forbidden lines of Seyferts, on the other hand, are much narrower than the long hydrogen emission lines. In 1967, astronomer Benjamin Markarian created a catalog, the Markarian Catalogue. The galaxies in the Catalogue were selected for their unique ultraviolet emission lines. Naturally, some selections were Seyfert galaxies – about 10%. Although the initial positions of the galaxies in the Catalogue were not initially accurate, they would improve six years later (Weedman, 1977).

Later, in 1974, Khachikian and Weedman created the two main classes used today by Astronomers: Seyfert Type 1 galaxies (Sy1) and Seyfert Type 2 galaxies (Sy2). Sy1 galaxies' Balmer lines can range to around "7500 km/sec in total breadth", making them much broader than their forbidden lines (Seyfert 1943). On the other hand, both the Balmer and forbidden lines are approximately the same width for Sy2 galaxies, with the width of the lines at half-maximum approximately ranging from 500 to 1000 km/s (Chen & Hwang, 2017). Objects with a mix of broad and narrow HI emission-line profiles cannot be classified as entirely Sy1 or Sy2. This led to the work of Osterbrock in 1987: the creation of the Seyfert 1.2, 1.5, 1.8, and 1.9 classifications (Antonucci 1993).

Antonucci theorized a Unified Model for AGN, arguing that the scientific community's classification of different types of AGN results not from intrinsic galactic properties, but from different viewing angles. For instance, Sy1 galaxies may be Seyferts whose galactic plane is viewed face-on whereas Sy2 galaxies are Seyfert galaxies whose plane is viewed edge-on. Antonucci found that NGC 1068, widely considered a Sy2 galaxy by the scientific community, showed broad emission lines, a property typical of Sy1, in polarized spectroscopic observations. His results were strong support for the Unified Model. The shape of an AGN's light emissions is caused by the way gas clouds are distributed or the uneven emission of light. A thick accretion disk or a torus of dust would therefore cause an anisotropic spectrum and explain the difference between Seyfert spectra according to the Unified Model (Antonucci 1993).

Justification of Research Topic

According to a summary of Seyfert research that introduces a paper by Chen, Seyferts have always been classified based on the ratio of the strengths of their emission lines. From Khachikian and Weedman to Osterbrock, classification has always meant looking at the ratio of the strengths of the spectral lines (Chen & Hwang, 2017). Our study proposes to classify based on other properties, which may allow us to potentially discover something new about Seyferts unable to be seen by solely focusing on the spectral lines. The study will also create a unique Seyfert classification model for future astronomers to not only use but also improve.

Sources of Data and Methods

Overview

We searched the SIMBAD database for Seyfert galaxies and found 43,029 Seyferts. We obtained equatorial coordinates, distance, redshift, morphological type, and color-magnitude values for each Seyfert. We solely used the Python programming language and its libraries for data analysis. Queries, data analysis, and graphs can be found on the linked Github.

Morphology

We searched SIMBAD's Seyferts and obtained 2 datasets with different morphological classifications: Hubble Tuning Fork and Hubble Stage T classifications. We used Pandas to clean the data and organize the Seyferts by morphological type. To graph the data, we used MatPlotLib.

Luminosity

Taking the data from SIMBAD, we used Pandas to clean and organize the data, and we used NumPy to format it such that we could graph the data with MathPlotLib.

Color-Magnitude

Two methods were used to analyze color-magnitude: the analysis of statistics and the construction of diagrams. For the raw statistics, we used Pandas to clean, organize, and graph the data into boxplots. For the color-magnitude diagrams, we used Pandas once again for cleaning and organizing the data but converted subsets of the data to NumPy and used MatPlotLib to graph color-magnitude diagrams of the Seyferts in the dataset.

Spatial Distribution

To make an Aitoff Projection of all the Seyferts, we used Pandas to clean the data and MatPlotLib to graph the data. To make a 3-D Model of the Seyferts, we used Pandas to organize and clean the data, Astropy to help compute the X, Y, and Z coordinates in parsecs from Earth, and Plotly to make an interactive 3-D diagram of the data. Using the X, Y, and Z coordinates obtained with Astropy, we used Pandas to procure descriptive statistics and boxplots of the dataset.

Redshift

We solely used Pandas to organize, clean, and graph the data for redshift analysis.

Machine Learning Models

We believed the Decision Tree and Random Forest models were the best models to use because, for our purposes, they were quick to implement among other benefits. Both models tend to be accurate on unseen datasets because they tend to avoid overfitting training sets. Furthermore, Random Forest models benefit from an ensemble of Decision Trees, which, in theory, should lead to better performance. We decided to train both and compare the two as we wanted to see whether more Decision Trees working in comparison would aid classification. Using the three properties in which there was a significant difference, we used Pandas and NumPy to clean the data and split it into training and testing sets, scikit-learn to create the models, and MatPlotLib to graph the Confusion Matrix. Note that for each model, we split the 843 Seyferts into 758 Seyferts for our training set and 85 Seyferts for the testing set – training our models with ~90% of the data and testing it with ~10% of the data. Seyfert type was proportionally represented in the testing and training sets.

Analysis

We analyzed 5 properties of Seyfert galaxies in isolation to determine which of the 5 would pose significant differences.

Morphology

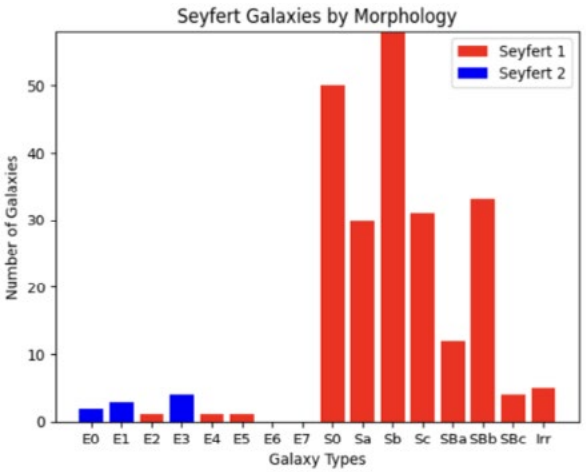


Figure 1. Frequency graph for both Seyfert types by morphology, utilizing already classified galaxies based on the SIMBAD-provided Hubble Tuning Fork classifications.

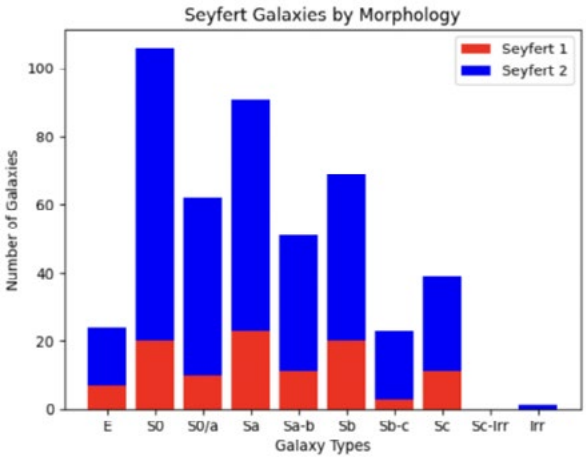


Figure 2. Stacked bar frequency graph for the Seyfert types, using SIMBAD's Hubble Stage T data converted to the Tuning Fork classification.

As seen from the graphs (Figure 1), both sets of data were so imbalanced in terms of morphology across Seyfert Types that we could not draw any significant conclusions about a relationship between Seyferts and morphology based on the SIMBAD dataset alone, causing us to disclude this data from our models.

Luminosity

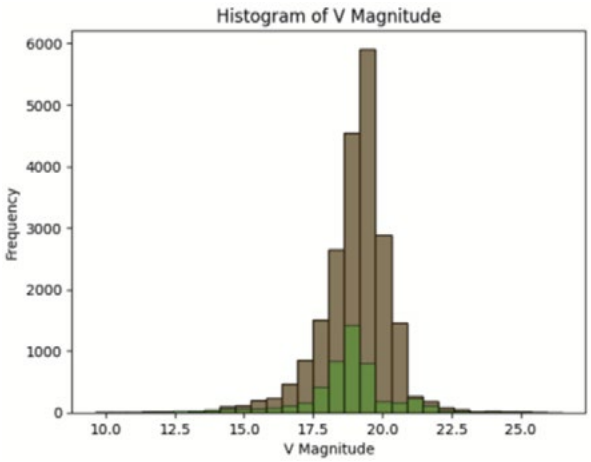


Figure 2. V magnitude histogram for Sy1 and Sy2 Galaxies. Brown bins represent Sy1 data while green bins represent Sy2 data

	#	Mean(μ)	Std. Dev(σ)	Min	Q1	Median	Q3	Max	Range
Sy1	21543	19.03	1.21	9.63	18.49	19.19	19.7	26.5	16.87
Sy2	4980	18.73	1.86	6.84	18.2	18.86	19.38	28.34	21.5

Table 1. V magnitude table of descriptive statistics by Seyfert type

Using V magnitudes as a measure of galactic luminosity, we found a significant difference between Sy1 and Sy2 to be the dispersion of magnitudes: the V magnitudes of Sy1 galaxies tend to be much more spread out, which can help the models classify the Seyferts with V magnitudes closer to the extremes (Figure 2). The mean and median of Sy1 V magnitudes are slightly greater than Sy2 V magnitudes (Table 1). These differences make V magnitude a differentiating factor that should be included in our model.

Color-Magnitude

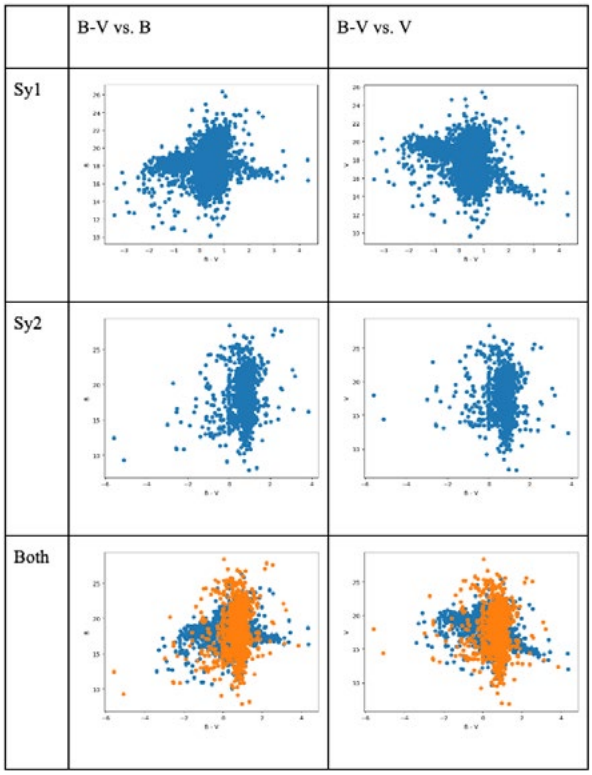


Figure 3. Color-magnitude diagrams of Sy1 and Sy2 galaxies. For the third row, a blue dot indicates a Sy1 galaxy and an orange dot indicates a Sy2 galaxy.

	#	Mean(μ)	Std. Dev(σ)	Min	Q1	Median	Q3	Max	Range
Sy1 Galaxies									
u-g	338	0.57	0.44	-1.52	0.28	0.44	0.78	2.57	4.09
g-r	338	0.42	0.33	-0.23	0.13	0.43	0.66	2.17	2.4
r-i	338	0.29	0.19	-0.56	0.17	0.32	0.43	0.81	1.37
i-z	338	0.19	0.21	-1.45	0.07	0.21	0.3	0.81	2.26
Sy2 Galaxies									
u-g	40	1.52	0.39	0.39	1.31	1.53	1.67	2.81	2.42
g-r	40	0.8	0.19	0.48	0.68	0.79	0.87	1.47	0.99
r-i	40	0.41	0.09	0.18	0.36	0.41	0.44	0.78	0.6
i-z	40	0.26	0.07	0.04	0.23	0.26	0.3	0.45	0.41

Table 2. Table of descriptive statistics for Sy1 and Sy2 color-magnitudes.

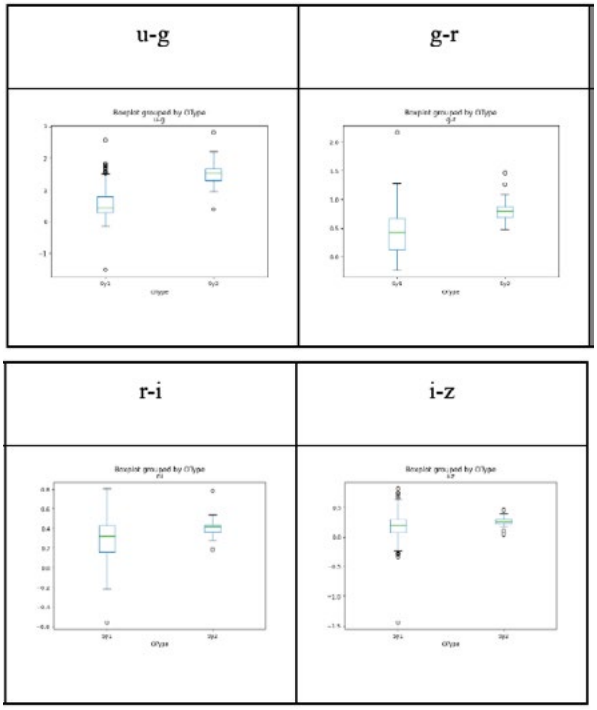


Figure 4. Side-by-side boxplots comparing color-magnitude values by Seyfert type.

V varies more than B for Sy1 but has about the same variation for Sy2 galaxies, so it makes sense when we observe that the B-V values vary more for Sy1 than Sy2 Galaxies (Figure 3). This alone is not yet enough of a difference to include color-magnitude in our model. However, the differences in filter magnitude are significantly greater for Sy2 galaxies (Table 2). This can be further confirmed as one examines the medians of u-g, g-r, r-i, and i-z data: Sy2 galaxies tend to have greater color-magnitude differences than Sy1 galaxies(Figure 4). Therefore, we can include color-magnitude values in our models due to this clear difference.

Spatial Distribution

	Effective Distance	X	Y	Z
Entire Dataset				
Discounting Outliers				

Figure 5: Side-by-side boxplots of Sy1 and Sy2 galaxies for the distance in the X, Y, and Z directions as well as total distance.

	#	Mean(μ)	Std. Dev(σ)	Min	Q1	Median	Q3	Max	Range
Sy1									
Distance	493	108883966	114780285	52	22400000	90360000	136500000	850200000	850199948
X	493	-7565753	105959540	-494035878	-57006810	-11340316	33888544	727273105	122130 8984
Y	493	-4240583	81497518	-444660717	-24427090	-413399	23872136	394510800	839171517
Z	493	22033319	81383224	-666317267	-4085946	9617738	42799914	584348086	125066 5353
Sy2									
Distance	1318	86779947	135999483	182	28750000	61745000	111192500	1773700000	177369 9817
X	1318	-18474862	115104569	-1596736969	-54357696	-17699133	13585283	1488000228	308473 7197
Y	1318	-5381265	89361075	-1595165166	-19219756	-992348	19602502	964135359	255930 0526
Z	1318	19278863	63681681	-868548189	-2928979	6906978	36177246	797260688	166580 8878

Table 3. Table of the descriptive statistics of the effective distance, X, Y, and Z positions of Sy1 and Sy2 galaxies.

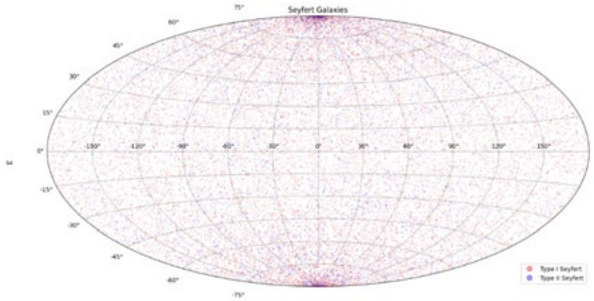


Figure 6. Aitoff projection of all Seyfert galaxies in the SIMBAD database. Red dots indicate Sy1 galaxies and blue dots indicate Sy2 galaxies.

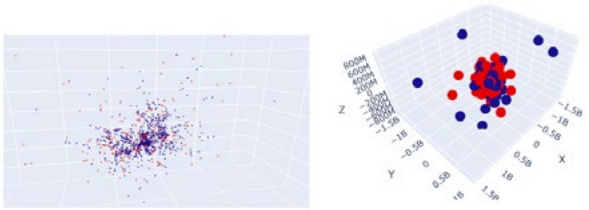
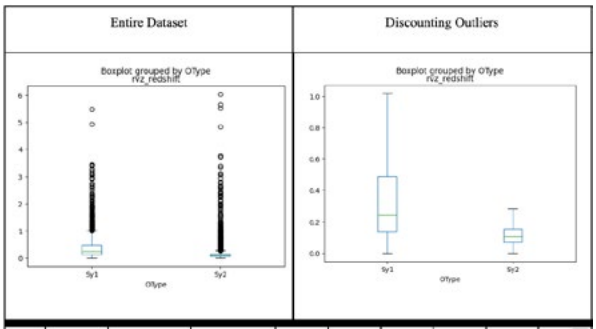


Figure 7. Snapshots of the interactive 3-D plot illustrating the spatial distribution of Sy1 and Sy2 galaxies. Red dots indicate Sy1 galaxies and blue dots indicate Sy2 galaxies.

Based on the boxplots of X, Y, and Z coordinates, there doesn't seem to be a significant difference between the two types of Seyferts concerning their position (Figure 5). The descriptive statistics showed negligible differences between

Sy1 and Sy2 galaxies in terms of position (Table 3). In terms of clustering, Seyfert galaxies do clump in occasional groups of 2s and 3s, but they are not organized into clusters and predominantly tend to be field galaxies scattered throughout space (Figure 6 + Figure 7). Also, including spatial distribution in our models could be a source of error should the models recognize differences that are not representative of reality. Therefore, coordinates and distance will not be included in our model.

Redshift



	#	Mean(μ)	Std. Dev(σ)	Min	Q1	Median	Q3	Max	Range
Sy1	737	0.28	0.26	0.0	0.1	0.2	0.42	3.41	3.41
Sy2	262	0.16	0.3	0.0	0.04	0.08	0.13	2.65	2.65

Figure 8. (Top) Side-by-side boxplots of Sy1 and Sy2 galaxies' redshift values. (Bottom)Table that compares the descriptive statistics of Sy1 and Sy2 galaxies' redshift values.

The minimum redshift of both Seyferts was approximately the same(z=0). However, the dispersion, measured via standard deviation and interquartile range, was much higher in Sy1 galaxies than in Sy2 galaxies discounting outliers. If we count outliers, the standard deviations are approximately the same. However, Sy1 galaxies typically have a greater redshift than Sy2 galaxies for both the mean and median of both data (Figure 8). This difference will enable us to include redshift data in our model.

Summary of Analysis

In sum, luminosity, color-magnitude, and redshift are all distinctive properties of Sy1 and Sy2 galaxies that we will use in our models.

Machine Learning Models

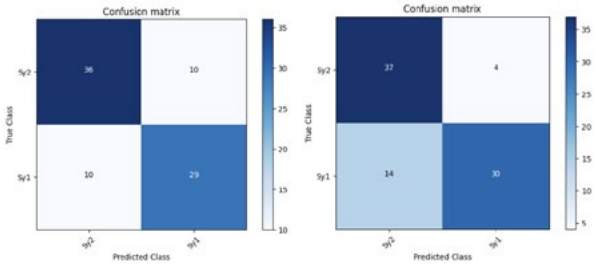


Figure 9. (Left) Confusion Matrix for the Decision Tree model. (Right) Confusion Matrix for the Random Forest model. Note that for both diagrams, numbers in the center of the squares represent the number of Seyferts that had the corresponding true and predicted classes.

We queried SIMBAD for all Seyferts in the database that had data for redshift, luminosity, and color-magnitude. After cleaning the data, 843 Seyferts remained. We used the data to build Decision Tree and Random Forest models, which classified 76.5% and 78.8% of all Seyfert galaxies in the testing set respectively. As the Random Forest was only marginally better, we concluded that both models may serve as equally appropriate classifiers. Furthermore, the results can be more closely analyzed on the Confusion Matrices (Figure 9).

Conclusion

Astronomers have historically classified Seyferts based on emission line ratios: from Carl Seyfert to Khachikian and Weedman to Osterbrock. However, this study has found 3 significant differentiating properties and accordingly created two models that both yielded above 75% accuracy, demonstrating that astronomers can classify Seyferts based on redshift, color-magnitude, and luminosity. In other words, Seyferts can be classified on more than just the strengths of their emission lines. We assume the Unified Model of AGN is false. However, we can still conclude information about the viewing angle and the environment around Seyferts even if the Unified Model is correct. Therefore, this study and the produced models would, whether the Unified Model is correct or not, provide information about a certain Seyfert's properties based on existing information. In terms of sources of error, because we only used SIMBAD, the sample size for our models was small. Therefore, there may not have been enough data for the models to fully realize the true extent to which

Sy1 and Sy2 galaxies differ. Another source of error that stems from using one database is if there was any inherent bias in SIMBAD sampling methods, it would bias our results because we would only be analyzing Seyferts from a certain direction, luminosity, etc. should SIMBAD have such biases.

The models produced in this study as well as stronger models from future work can assist large databases in classifying Seyfert galaxies. The models prove resourceful if such databases cannot provide complete spectral information but instead have complete information about certain properties. In addition, recognizing such a difference could help scientists discover more about this special type of active galaxy.

Comparative Analysis

When we compared our results to other studies in the literature, a study by Chen used a Convolutional Neural Network (CNN), a type of deep learning algorithm, to differentiate Seyfert 1.9 spectra from Seyfert 2 spectra. He obtained 91% precision in classifying Seyfert 1.9 spectra. The cleaned dataset that we used to train the model was composed of 844 Seyferts, while Chen's study consisted of 341 Seyfert 1.9 galaxies and 53,494 Seyfert 2 galaxies. As such, his methods were different: he was able to classify with better accuracy because he had more data and was able to therefore utilize a stronger deep learning model to make better predictions (Chen, 2021). A CNN is superior to a decision tree or random forest, especially with more data, because it independently creates its own categories rather than being assigned categories. In other words, the model may be able to find patterns not initially apparent to human researchers. We used characteristics not often utilized to classify Seyferts (redshift, color-magnitudes, and luminosity). As their differences aren't as stark as emission line strength ratios, it is readily apparent why Chen's model may be more accurate.

Future Work

There is controversy over the morphology of Seyfert galaxies. In terms of future work, our data supports the theory that Sy1 galaxies are typically spiral while Sy2 galaxies are typically elliptical. It would therefore be understandable why Sy2 galaxies vary in luminosity more than Sy1 galaxies: elliptical galaxies vary more in luminosity than spiral galaxies. Furthermore, Sy2 were found to be redder than Sy1 galaxies, corroborating our theory: if Sy1 galaxies are typically spiral, they will be bluer than elliptical Sy2 galaxies. As we were not able to produce experimental results from this theory using SIMBAD data, future work must be done to confirm such conclusions. We may also be able to improve our results if we use a deep learning model coupled with more data from other databases like NED.

Acknowledgments

We would like to thank Dr. Shyamal Mitra for teaching us and supervising the premise of our research – his guidance was vital to our research process. We would also like to thank the Geometry of Space research program at the University of Texas at Austin for providing us with support for our research paper. This research made use of the SIMBAD database, operated at CDS, Strasbourg, France.

References

Antonucci, R. (1993). Unified Models for Active Galactic Nuclei and Quasars. Annual Review of Astronomy and Astrophysics, 31(1), 473–521. doi:10.1146/annurev.aa.31.090193.002353

Chen, Y. C. (2021). Classifying Seyfert galaxies with deep learning. The Astrophysical Journal Supplement Series, 256(2), 34.

Chen, Y. C., & Hwang, C. Y. (2017). Morphology of Seyfert galaxies. Astrophysics and Space Science, 362(12), 230.

Netzer, H. (2015). Revisiting the unified model of active galactic nuclei. Annual Review of Astronomy and Astrophysics, 53, 365-408.

Osterbrock, D. E. (1987). Seyfert galaxies: Classification, morphology, observations at optical wavelengths, environmental factors. Symposium - International Astronomical Union, 121, 109–118.

Seyfert, C. K. (1943). Nuclear emission in spiral nebulae. The Astrophysical Journal, 97, 28.

Shields, G. A. (1999). A brief history of active galactic nuclei. Publications of the Astronomical Society of the Pacific, 111(760), 661.

Simkin, S. M., Su, H. J., & Schwarz, M. P. (1980). Nearby Seyfert galaxies. Astrophysical Journal, Part 1, vol. 237, Apr. 15, 1980, p. 404-413., 237, 404-413.

Weedman, D. W. (1977). Seyfert galaxies. Annual Review of Astronomy and Astrophysics, 15(1), 69–95.