RANDOM FOREST CLASSIFICATION OF ACTIVE GALACTIC NUCLEI WITH OPTICAL AND INFRARED DATA

by

Jacob Matthew McKee

HONORS THESIS

Submitted to Texas State University
in partial fulfillment
of the requirements for
graduation in the Honors College
May 2022

Thesis Supervisor:

Blagoy Rangelov

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Blagoy Rangelov for his support and mentorship in the process of completing this thesis. I would also like to thank my family and friends for supporting me throughout my undergraduate education. This work was completed under the supervision of my advisor and the Texas State University Honors College.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation

1. AGN ..... Active Galactic Nuclei
2. SMOTE ..... Synthetic Minority Over-sampling Technique
3. SMBH ..... Supermassive Black Hole
4. QSO ..... Quasi-Stellar Object (Quasar)
5. LINER ..... Low Ionization Nuclear Emission Line Region
6. BL LAC ..... BL Lacertae
7. WISE ..... Wide-field Infrared Survey Explorer
8. TP ..... True Positive
9. FP ..... False Positive
10. FN ..... False Negative

ABSTRACT

This project develops and evaluates a pipeline for automatic classification of active galactic nuclei (AGN) using the Random Forest algorithm. AGN are galaxies where the active accretion of gas and dust onto the disk around the galaxy's central supermassive black hole is generating enough energy to outshine the stars in the galaxy. Here we use Random Forest, a supervised, decision-tree-based algorithm to classify AGN using only optical and infrared photometry from the *Gaia* and WISE space telescope datasets. We train and test the algorithm on 5 classes of AGN, twice each with magnitude and color, using both the original data and data modified using the Synthetic Minority Over-sampling Technique (SMOTE). These 4 models have total classification accuracies of 90-93%, but f1-scores for each class varying from 0.44 to 0.97 across all models, with only quasars being classified highly accurately with f1-scores of 0.96-0.97. This may be due to an overabundance of quasars in the sample. This method can be explored further with more photometric data from other telescopes in other wavelength ranges, more algorithm parameters, other classification algorithms, and inclusion of other classes.

I. INTRODUCTION

A. Galaxies and Active Galactic Nuclei

Galaxies are gravitationally bound objects containing stars, gas and dust, and dark matter that first formed in the early universe as gas clouds collapsed to form stars. Galaxies were first classified by Edwin Hubble morphologically, i.e., based on shape (Ryden, Barbara, and Peterson, 2011). The main classifications of galaxies are elliptical, spiral, and irregular (Figure 1). Elliptical galaxies look like elliptical blobs of stars, while spiral galaxies, like our own Milky Way, have a central bulge or bar, with spiral arms protruding from the center which contain stars and star-forming regions of gas and dust (Ryden, Barbara, and Peterson, 2011). Irregular galaxies are of course irregular in shape, with lots of gas and dust to fuel star formation (Ryden, Barbara, and Peterson, 2011).

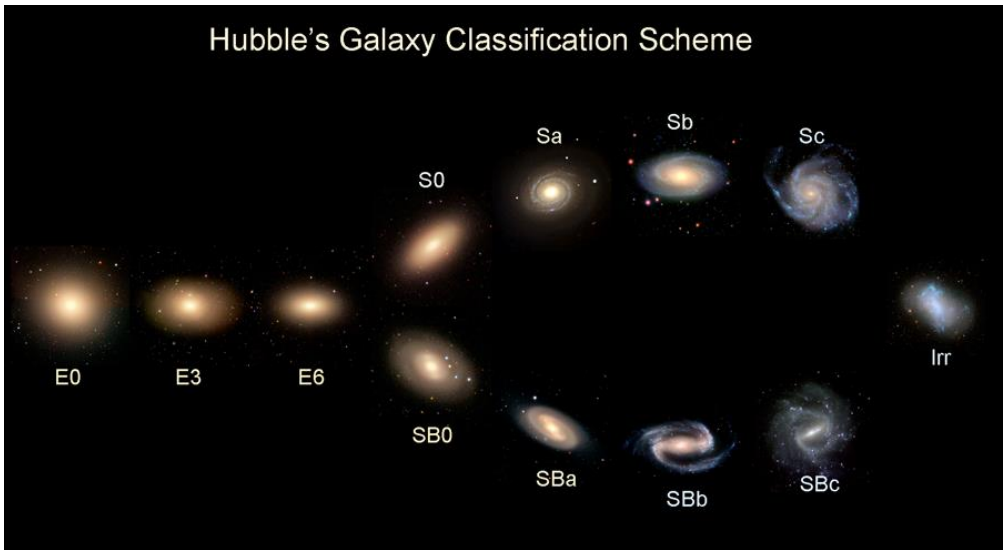

Figure 1. Hubble Morphological Classification of Galaxies. Elliptical (E#), Spiral (S_), Spiral Bar (SB_), and Irregular Galaxies. (Credit: University of Iowa, http://astro.physics.uiowa.edu/ITU/labs/foundational-labs/classifying-galaxies/part-1-hubbles-tuning-fork.html)

Galaxies have supermassive black holes (SMBHs) at their centers that can range from hundreds of millions to billions of times the mass of the Sun. Active Galactic Nuclei

(AGN, Figure 2) exist in galaxies where there is a nucleus of activity around the SMBH where gas and dust from the inner galaxy is being pulled onto a disk around the black hole, called an accretion disk (Ryden, Barbara, and Peterson, 2011). This disk is mostly thermal in nature, generating a blackbody spectrum. A blackbody is an ideal object that radiates light with a distribution of wavelengths called a continuum, which is dependent on temperature. We can approximate thermal sources of light as blackbodies. For AGN, deviation from this continuum can tell us about the molecules present. Stars are also approximated as blackbody objects, and galaxy spectra usually have a complex shape due to the combination of many stars at various temperatures, However, AGN are usually so bright that it dominates the spectrum of the galaxy, so the contributions of the nucleus form the main features of the spectrum. We can observe the spectra of AGN and classify them based on their spectral features.

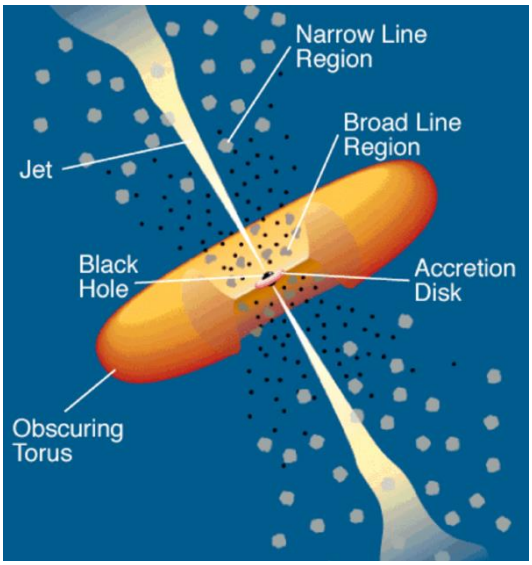

Figure 2. Unified Model of AGN. Adapted from Urry and Padavani 1995.

B. Physics of Light and Galactic Gas and Dust

Spectra show the intensity of light at different wavelengths. Spectroscopy is done

by taking incoming light and putting it through a prism or a diffractive grating in a telescopes optical system to diffract photons of different wavelengths at different angles. The spectrum shows spectral lines at different wavelengths, which can indicate abundances of atoms or molecules in astronomical objects. There are a few key properties of these lines. There is redshift, in which the expansion of the universe causes the galaxies' light to Doppler shift to higher wavelengths, which can help determine the distance of a galaxy. Another factor is the height and width of the lines. The height of a line is the intensity or flux of the line, which is related to the number of photos measured at that wavelength, while width can be determined by the motion or other properties of the gas in the galaxy. These spectral lines come from electron and molecular energy transitions.

Electrons in an atom can only exist at certain discrete energy levels. The normal state of an atom is that every electron is in the lowest possible energy state, also called its ground state. The valence electrons of an atom are the only electrons that can be excited into a higher energy state. This happens when a photon interacts with the electron. The photon must be at a specific energy that corresponds to the energy difference between the current energy of the electron and one of the higher energy levels. The photon's energy is transferred to the electron, which now exists in this higher energy state. These higher energy states are very unstable, and so almost immediately the electron will fall back into its ground state. It can do this by the same transition that just excited it, reemitting a photon of the same energy of the incoming photon, or through multiple intermediate transitions, emitting multiple photons of lower energies. These interactions are called absorption and emission. In the spectrum, emission lines are peaks over the continuum and absorption lines are valleys beneath the continuum.

Molecules also emit and absorb photons, but their spectra are much more complex. This is because in addition to electronic transitions, molecules also have rotational and vibrational modes of energy. The changes of energy of these modes are also quantized and interact with photons similarly to the electrons of single atoms. These transitions can also be caused by kinetic energy transfer in molecular collisions. This is the dominant cause of the thermal emission in AGN. Rotational energy is energy due to the rotation of the whole molecule, and vibrational energy is determined by the motions of the atoms in a molecule with respect to each other. In the simplest case, a diatomic molecule like H_2 or CO, has one mode of vibrational energy, the stretching of the bond, and one mode of rotational energy. Compared to atomic spectra, molecular spectra tend to be much richer in spectral lines, depending on the complexity and energy configuration of the molecule. In a galactic spectrum, all the spectral lines of all the atoms and molecules present in the galaxy are overlaid, creating a very messy spectrum.

C. AGN Classification

There are many spectral classifications of AGNs depending on the variance and presence of specific spectral features (Figure 3). The most general classification is radio loud and radio quiet, i.e., whether or not there is significant radio emission. In this project, we use the classes quasar (QSO), Seyfert Type 1 and Type 2, BL Lacertae, and LINER (Low Ionization Nuclear Emission Line Regions). Classifications of AGN can depend on several factors, such as the inclination of the object galaxy, or differences in activity in the AGN, such as whether the galaxy has a relativistic jet or not, or the presence of ionized particles.

Quasars, originally named quasi-stellar objects for their similarity in appearance to

stars, have very high luminosities and strong emission lines (Ryden, Barbara, and Peterson, 2011). Seyfert galaxies are spiral galaxies with narrow and broad emission lines from gas in different regions around the SMBH. Type 1 have both narrow and broad lines, while Type 2 only have narrow lines as a result of the inclination of the galaxy. LINERs are less luminous nuclei with strong lines of weakly ionized (+1) atoms, and weak lines of strongly ionized (+2 or higher) atoms. BL Lacertae AGN are a type of blazar. Blazars are active galaxies oriented so that their relativistic jets are pointing almost directly at Earth. Because we don't really see the disk, only the spectrum of the jet, these objects tend to have a non-thermal optical and infrared spectrum, are radio loud, and highly variable, meaning their emission changes relatively quickly.

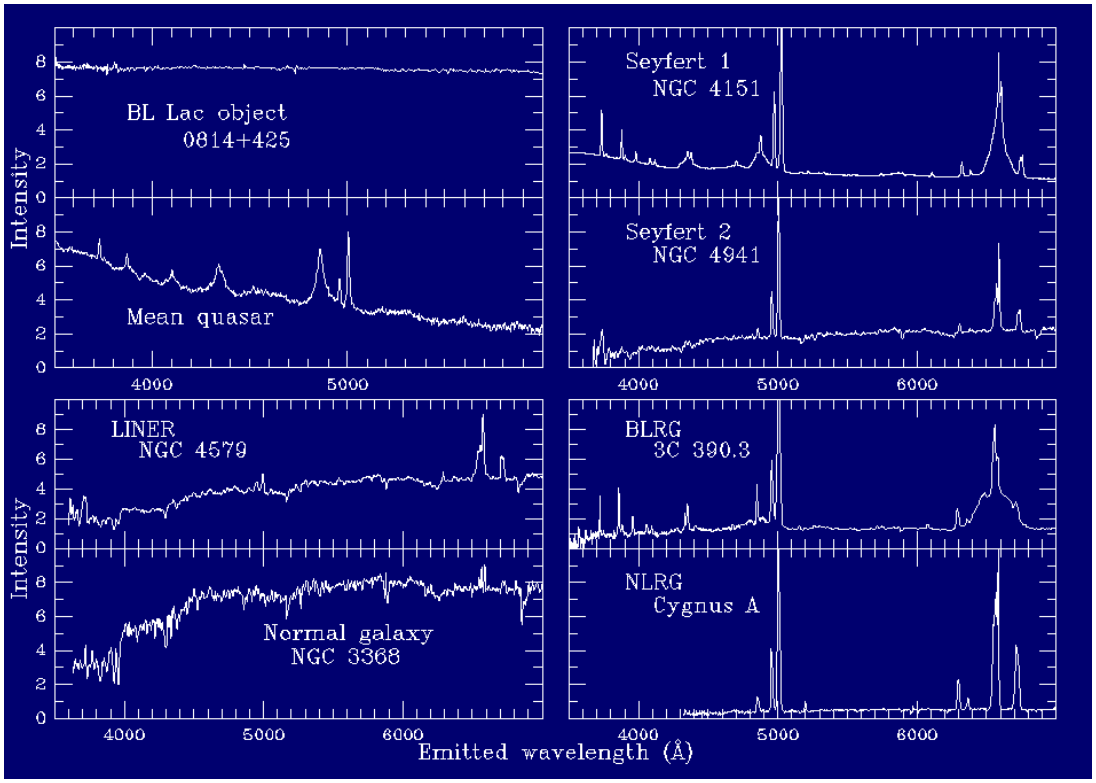

Figure 3. Examples of optical spectra of the principal AGN types. (Credit: William Keel, https://pages.astronomy.ua.edu/keel/agn/spectra.html)

Instead of spectroscopy, this project uses photometry. Photometric astronomy is done by putting light through a special filter with physical properties that define the range of wavelengths that can pass through. Usually, a few different filters are available on each telescope. In this work, *Gaia* and WISE (Wide-field Infrared Survey Explorer) photometry is used. *Gaia* is a space telescope designed for the study of the motion of astronomical objects, but it also takes photometric data in the visible and near-infrared, and while WISE is an infrared survey telescope. The data used is in the form of magnitude, which is a measure of the luminosity (energy output) of the galaxy. In this work, both the magnitudes themselves and colors are used. Colors are simply the differences in magnitudes that indicate how bright the galaxy is in one band compared to another.

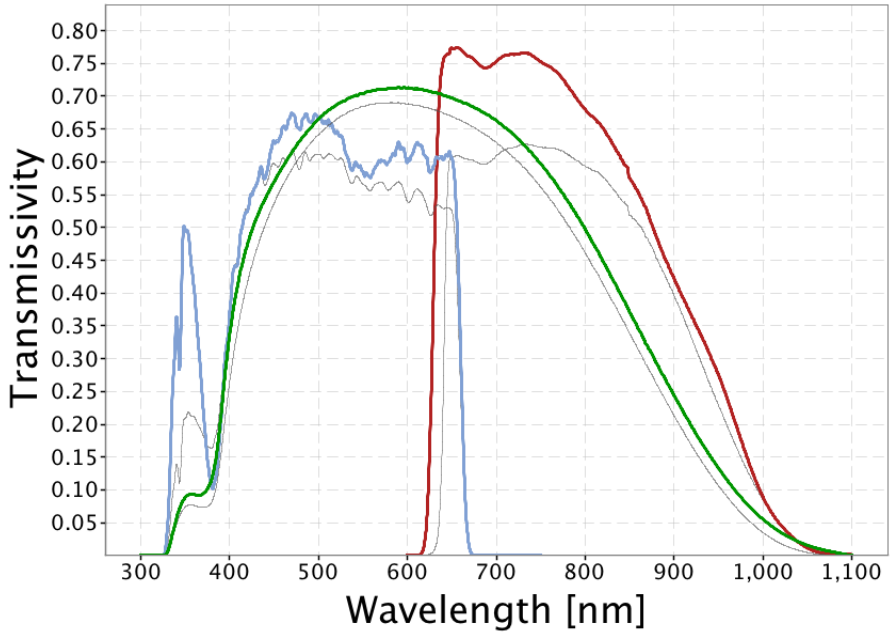

Figure 4. *Gaia* DR2 Passbands. The green line = g band, blue line = bp band, and the red line = rp band (Credit: https://www.cosmos.esa.int/web/gaia/iow_20180316).

Gaia photometry is in the visible and near-infrared, with its filters transmitting light between about 300 and 1100 nanometers (Figure 4). Visible light is between about

300 and 750 nanometers, while the infrared region spans about 750 nanometers (0.75 microns) up to submillimeter wavelengths. The WISE photometry is in the mid-infrared region, measuring wavelengths ranging from about 3 microns to almost 30 microns (Figure 5).

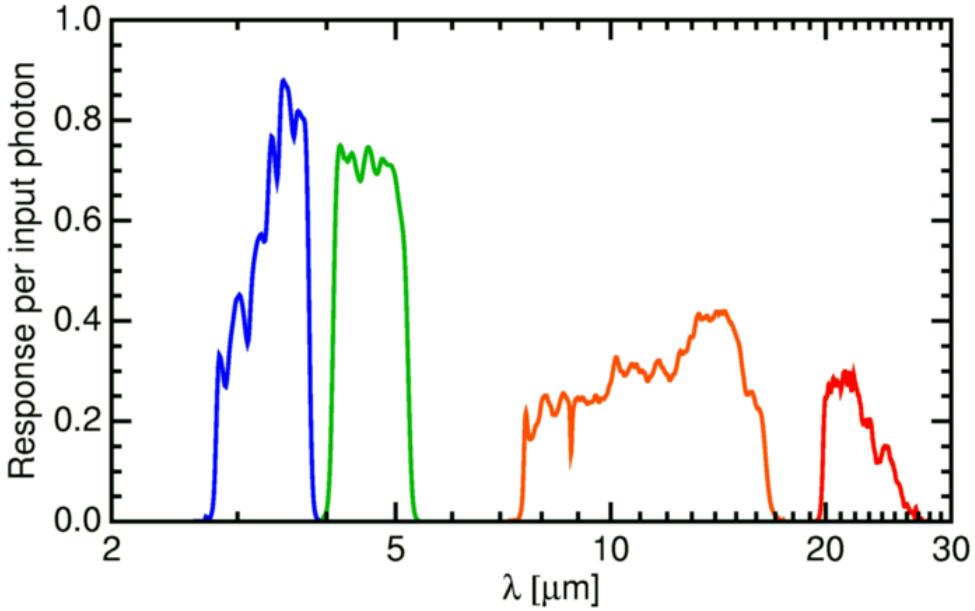

Figure 5. WISE passbands. In order from left to right, W1, W2, W3, and W4 passbands (Credit: https://www.astro.ucla.edu/~wright/WISE/passbands.html).

D. Machine Learning and the Random Forest Classifier

Machine learning tools are becoming very important in helping astronomers quickly analyze large datasets. One of the principal uses of machine learning in astronomy is classification. In this case, we want to create a model that can classify AGN with some degree of confidence. There are two types of classification algorithms: supervised and unsupervised. Unsupervised algorithms classify objects in a dataset by finding groups of objects with similar properties. Supervised algorithms use training datasets with known classes and data to "learn" the qualities of objects in each class and generate a model that could classify an unknown object. The effectiveness of the model is tested with a validation

dataset. This project uses the Random Forest Classifier, which is a supervised machine learning algorithm. It is fed a training dataset of classified AGN objects and magnitude data and returns a model to verify. Usually, the training and validation datasets are randomized subsets of the whole dataset at a given ratio, with most data used for training and the rest for validation.

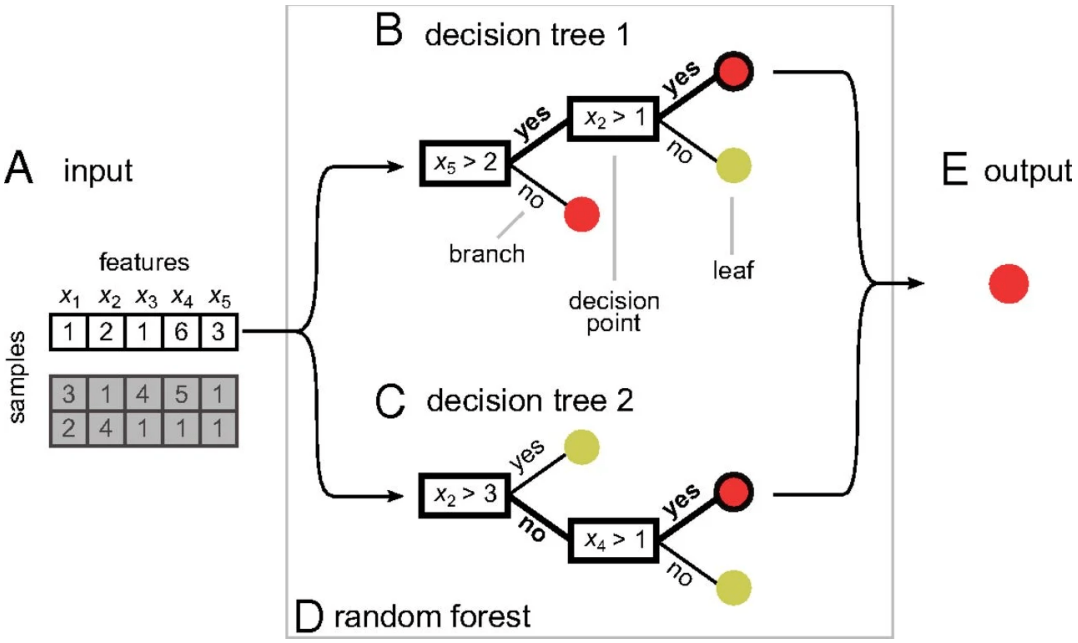

Figure 6. Diagram of a Random Forest. Rectangles are decision nodes, and the colored circles are leaf nodes (Denisko and Hoffman 2018).

A Random Forest is a group of decision trees in which each tree takes a given data point and will classify it into one of the possible classes (Figure 6). A decision tree is a series of decision points that progressively evaluate aspects of the input data. Each decision point is a node that assesses whether an object fits some criteria, and has two branches, a True and a False branch. The tree tests the criteria and proceeds through the corresponding branch to the next decision node. This continues until a leaf node is reached, which gives a classification for the object. Every tree must have at least one path to each class. The

whole forest is considered, and the class with the most trees' votes will be returned as the output class. These forests are built on the training data in such a way that if you have the model classify the training data, it should ideally be 100% correct, though this may not always be the case. In this project, the Random Forest is trained on galaxy magnitudes or colors and the output of the forest is an AGN class.

II. METHODOLOGY

We classify AGN utilizing the Random Forest Classifier included in the Scikit-learn[1] python package for machine learning applications. The magnitude data from *Gaia* and WISE is taken with TOPCAT[2] along with the name of each object and the object's AGN classification. Any object containing a null value for any of the seven bands is removed from the dataset. The dataset contains 25014 objects. We have two pipelines, one using magnitude and one using colors. The magnitudes and colors' values z are standardized by

$$z = \frac{x - u}{s}$$

where x is the sample, u is the mean of the sample, and s is the standard deviation of the sample. Then the set is randomly divided into a training dataset and a verification dataset with a 70/30 distribution. We then create a copy of the training dataset and modify it using the Synthetic Minority Over-Sampling Technique (SMOTE). This resamples the dataset and creates additional synthetic objects for each of the undersaturated classes so that we have a new training dataset with the same number of objects in each class. We invoke the

[1] scikit-learn.org
[2] http://www.star.bris.ac.uk/~mbt/topcat/

Random Forest Classifier and fit it using these training datasets and have all models predict the classes and class probabilities for the test dataset. We also have the models predict the classes of unclassified AGN.

The models are evaluated in a few ways. The first interesting result to consider is which photometry filters the Random Forest relied on the most in creating the model. This feature importance score is given in percentage and tells us what features we might use in further study. There are a few specific statistical figures we can use as well. Precision, recall, and f1-score are helpful statistical indicators based on the True Positives, False Positives, and False Negatives that are specific to individual classes so we can compare the performance of the models on each class. Precision (P) is a ratio or percent of how many objects in a determined class are correctly classified:

$$P = \frac{TP}{TP + FP}$$

Recall (R) represents how many objects of a true class are correctly classified:

$$R = \frac{TP}{TP + FN}$$

The f1 score is the harmonic mean of precision and recall:

$$f1 = \frac{2PR}{P + R}$$

The accuracy score is simply a sample-wide classification accuracy given by the ratio of the number of objects correctly classified against the total number of samples:

$$A = \frac{TP_{all}}{total}$$

These are all on the unit scale, i.e. 1.0 is a perfect score. Another useful indicator of the performance of a model is a confusion matrix. A confusion matrix shows the intersections of determined class and true class. Each row represents the true class, and each column

represents the determined class. The diagonals are true positives, i.e., the amount of correctly classified objects, while the other entries in the row are false negatives, and the entries in each column are false positives. We also examine the class probability score distributions, i.e., the distributions of the confidences of the classifications, as well as the accuracy at increasing lower limits of confidence.

III. RESULTS AND ANALYSIS

Our main statistical results are given by the classification reports for each model (Figure 7). The validation set contains 7505 objects, 6152 of which are QSOs. The accuracy scores of the models trained on magnitude were 0.92 and 0.90 for original and SMOTE trained models respectively. QSO was the most successful classification with f1-score of 0.96 for both models. BL LAC was the least with 0.44 in both. The macro average of f1 was 0.67 and 0.66 with a weighted average of 0.91 and 0.90.

	precision	recall	f1-score	support
BL LAC	0.83	0.30	0.44	80
LINER	0.63	0.50	0.56	24
QSO	0.95	0.98	0.96	6152
SEYFERT TYPE 1	0.75	0.60	0.67	983
SEYFERT TYPE 2	0.71	0.70	0.70	266
accuracy			0.92	7505
macro avg	0.77	0.62	0.67	7505
weighted avg	0.91	0.92	0.91	7505

	precision	recall	f1-score	support
BL LAC	0.90	0.55	0.68	80
LINER	0.69	0.46	0.55	24
QSO	0.95	0.98	0.97	6152
SEYFERT TYPE 1	0.79	0.65	0.72	983
SEYFERT TYPE 2	0.73	0.78	0.75	266
accuracy			0.93	7505
macro avg	0.81	0.68	0.73	7505
weighted avg	0.92	0.93	0.92	7505

	precision	recall	f1-score	support
BL LAC	0.37	0.55	0.44	80
LINER	0.44	0.62	0.52	24
QSO	0.97	0.95	0.96	6152
SEYFERT TYPE 1	0.66	0.68	0.67	983
SEYFERT TYPE 2	0.67	0.78	0.72	266
accuracy			0.90	7505
macro avg	0.62	0.72	0.66	7505
weighted avg	0.91	0.90	0.90	7505

	precision	recall	f1-score	support
BL LAC	0.53	0.79	0.63	80
LINER	0.45	0.54	0.49	24
QSO	0.97	0.95	0.96	6152
SEYFERT TYPE 1	0.69	0.70	0.70	983
SEYFERT TYPE 2	0.67	0.78	0.72	266
accuracy			0.91	7505
macro avg	0.66	0.75	0.70	7505
weighted avg	0.92	0.91	0.91	7505

Figure 7. Full classification reports for original (top) and SMOTE models (bottom) for magnitude (left) and color (right)

The color accuracy scores were also high at 0.93 and 0.91, and QSOs had f1-scores of 0.97 and 0.96. BL LAC had f1-scores of 0.68 and 0.63, and LINER had f1s of 0.55 and

0.49. The macro averages were 0.73 and 0.70 with weighted averages of 0.92 and 0.91.

The confusion matrix for each model shows the distribution of correct and erroneous classifications (Figure 8). QSOs are most often mislabeled as Seyfert 1 while Seyfert 1 are most often mislabeled as QSO. The SMOTE models classify the BL LAC objects more correctly than the original models, but also misidentifies other objects as BL LAC more often. This is also true for LINER, and Seyfert 1 and 2.

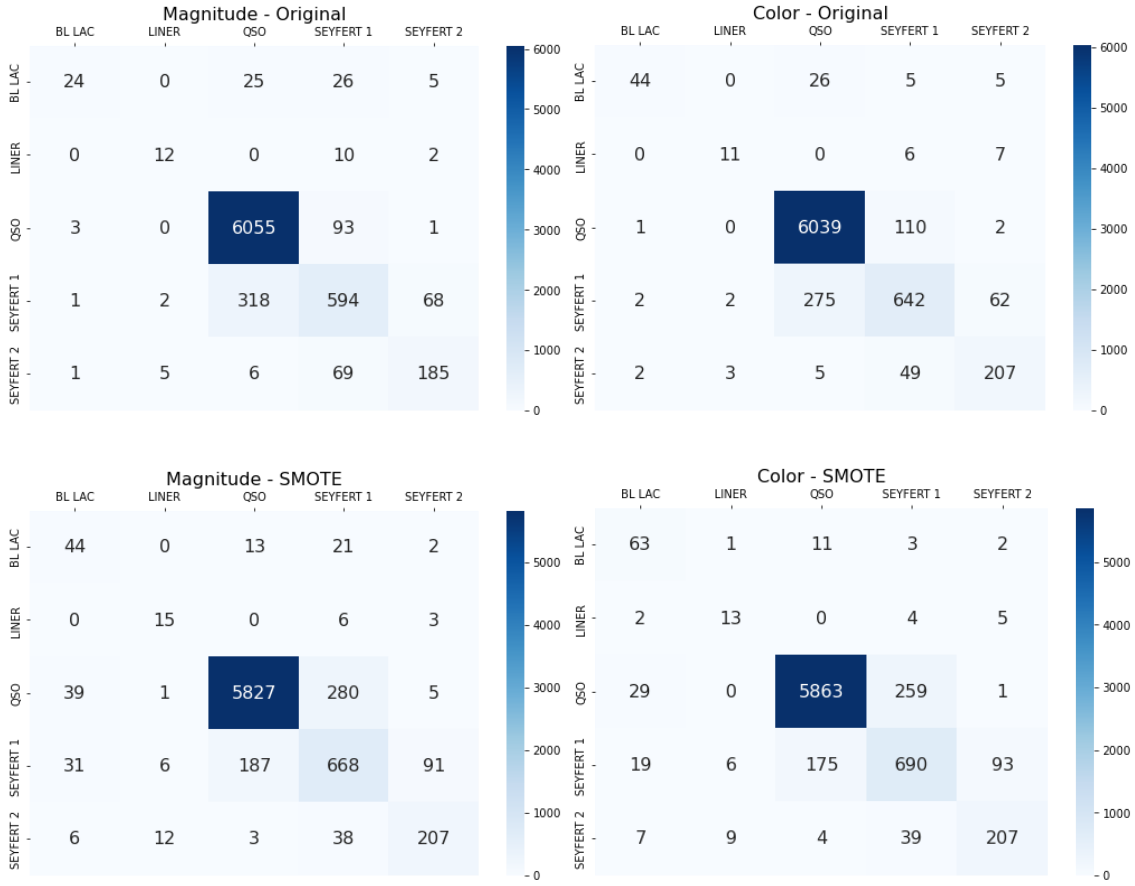

Figure 8. Confusion Matrices, where rows are True class and columns are Classified class.

The disagreement rate of the magnitude models on the test dataset is 6.76%, while the color models had a similar disagreement rate of 5.28%. To further test the agreement of the models, they were run on a set of 1553 Unclassified AGN. Of these, 154 or 9.91% were disagreed upon by the magnitude models and 114 or 7.34% by the color models.

All four models have some similar results. Nominally, they performed reasonably well with high accuracy, however, this is likely skewed due to the saturation of quasars in the validation set, comprising 82% of the sample. For all models, QSO performed the best by far in all metrics, and while the precision for BL LAC and LINER is very low in the SMOTE models compared to the original models, their recall is somewhat higher. The Seyfert 1 and 2 classes performed consistently at around 70% for every model and metric. It is interesting to note that the only constant in these models is that LINER are never mistaken for quasars.

For both original and SMOTE magnitude models, the g and W1 bands were the most important features at >0.20 (Figure 9). The two most important colors were g-bp at 0.173 and 0.141 and g-rp at 0.166 and 0.099.

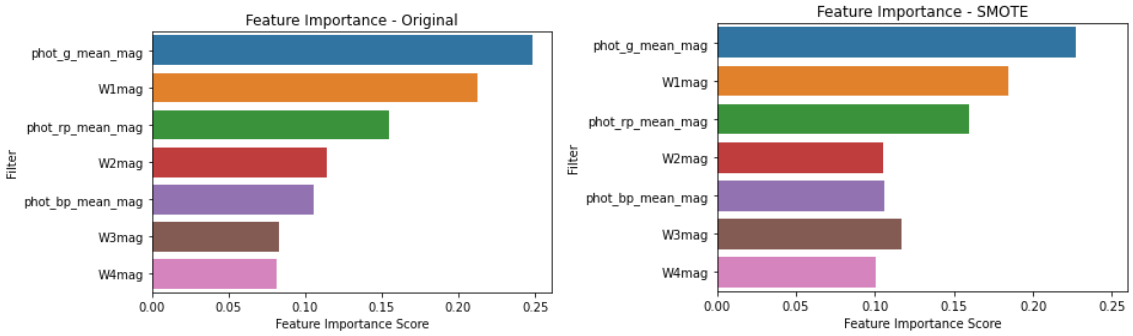

Figure 9. Feature importance scores for magnitude models.

The SMOTE models changed the feature importances slightly, with g and W1 for magnitude and g-bp and g-rp being less relied upon, while some others increase, with notably almost all the WISE-only colors increase significantly. This might mean that the saturation of QSO in the original training set is causing an overestimation of the importance of g and W1 and g-bp and g-rp, which might mean they are one of the main factors in identifying quasars.

The color importance scores may indicate some specific spectral features that significantly affect the magnitude differences among the classes, especially in quasars (Figure 10). The WISE colors' importance increase in the SMOTE model could mean that these are more important to classifying the underrepresented classes of AGN.

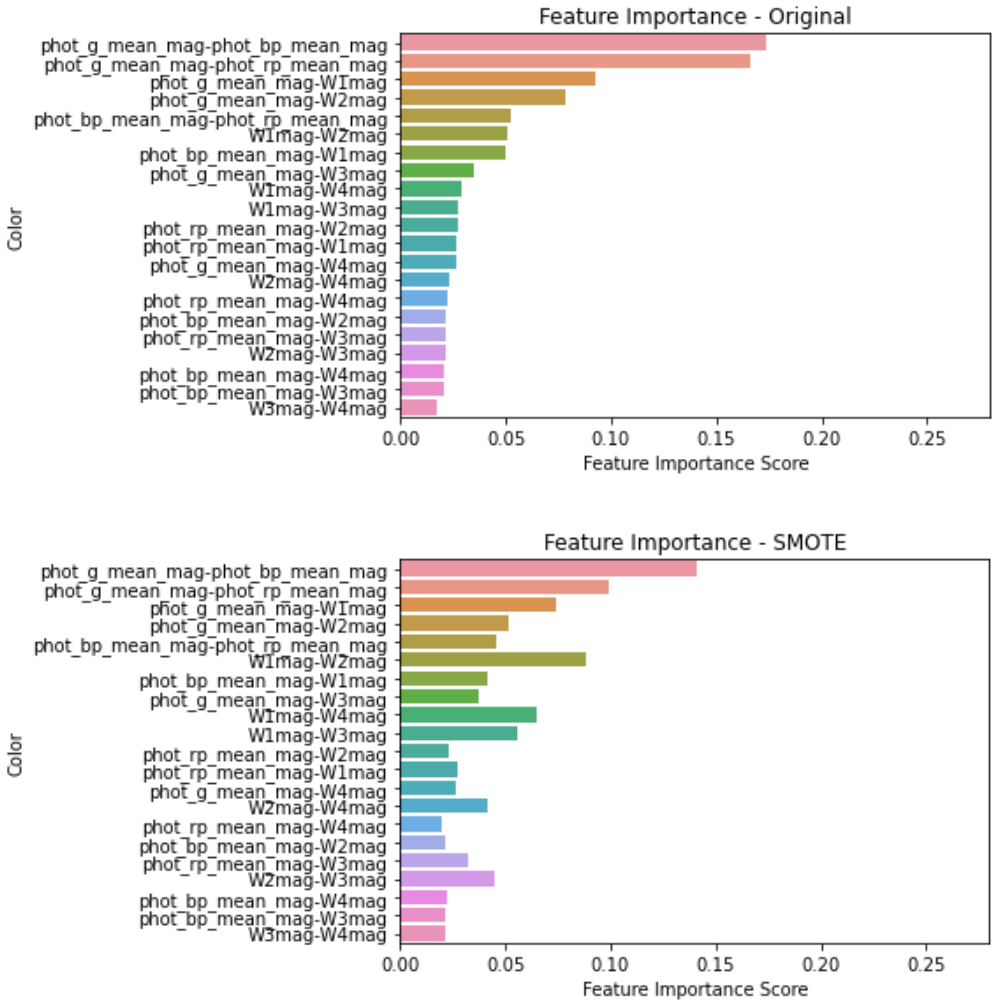

Figure 10. Feature importance scores for color models.

The next major point of analysis that we consider is the confidence of each classification. We define the confidence to be the class probability of the determined class. Any given object in the validation set has a set of class probabilities, which is the percentage of trees that voted for each class, where the determined class is the class with

the most votes. It should be noted that two classes can tie for the highest class probability, but only if that tie is at or under 50% (0.50). Here we consider the confidence distribution of the True Positives and the False Negatives (Figure 10). Due to the QSO abundance, we separate for clarity the QSO confidence distributions from the other classes for the TP case.

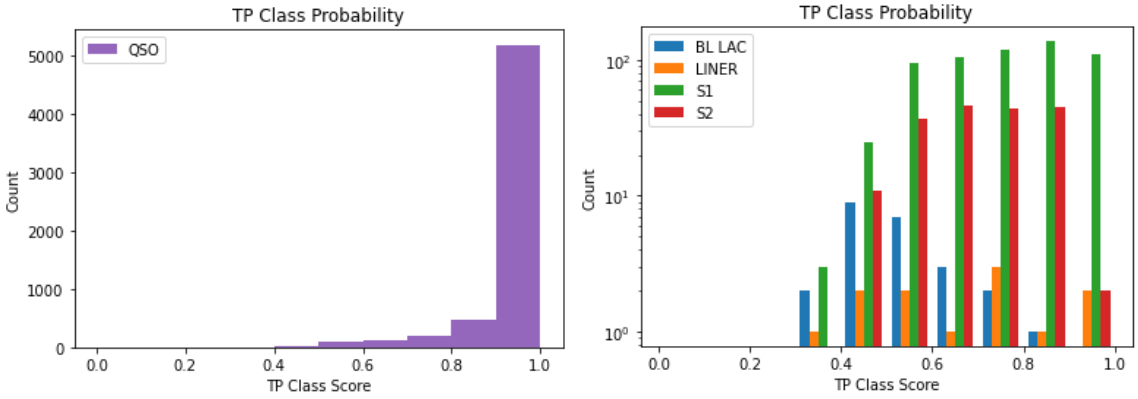

Figure 11. Distribution of class probability scores for correctly identified AGN for the original magnitude model.

The correctly identified quasars are almost entirely above 0.80 class probability score, with most above 0.90 (Figure 11). This very high confidence is likely a direct result of the abundance of these objects in the training dataset. When compared to the confidence distributions of the other classes, we see that the other classes are much more distributed. The Seyfert classes both have relatively even distributions across the approximate range of 0.5 to 0.9, BL LAC peaks at 0.4-0.5 class probability, indicating a generally low confidence in these classifications. LINER has a somewhat even distribution, but it should be noted that LINER and BL LAC both have low TP counts.

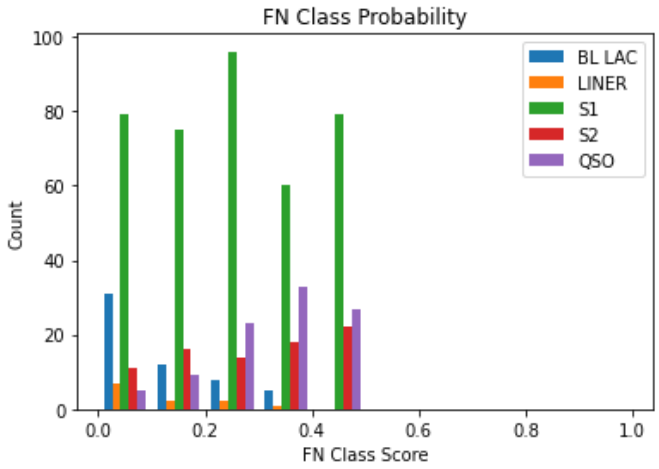

Figure 12. FN Confidence distribution for original magnitude model

Considering the FN case (Figure 12), we see that when quasars are incorrectly classified, their probability score tends upwards towards 0.3-0.5. Thus, quasars are generally correctly classified with high confidence, and incorrectly classified with low confidence. The Seyferts again have a distributed confidence here, while BL LAC and LINER both tend towards a very low class probability. This could be a result of the severe underrepresentation in the training set.

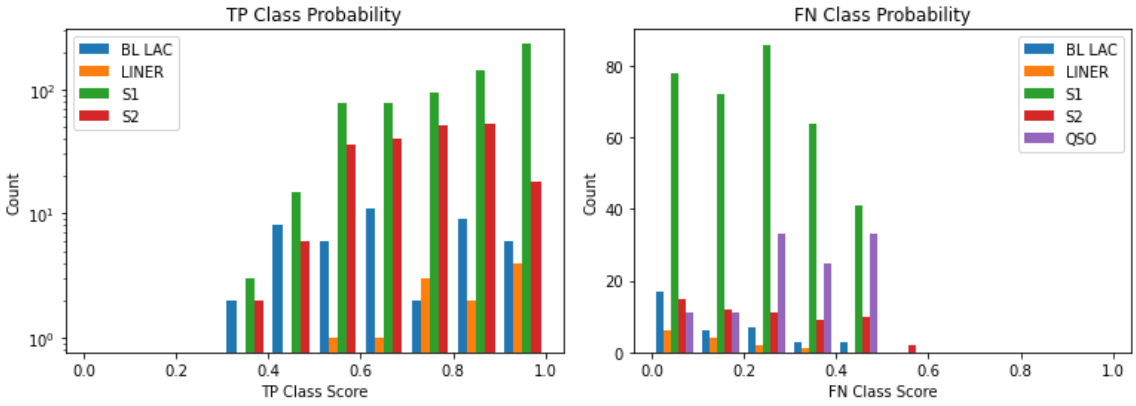

Figure 13. TP and FN Confidence distributions for the original color model. (QSO TP not shown)

The confidence distributions for the color model are very similar to the magnitude model (Figure 13). There are more medium-high confidence True Positive BL LACs and

LINERs, but otherwise there are no significant differences. Not shown is the quasar distribution, it is a similar distribution as the original magnitude model.

Accuracy as a Function of Confidence Limit

Confidence	MOW	MOW/O	MSW	MSW/O	COW	COW/O	CSW	CSW/O
0.5	0.923819	0.614516	0.912363	0.703974	0.930198	0.678125	0.918008	0.739199
0.6	0.944103	0.644444	0.935777	0.737817	0.943767	0.692448	0.939463	0.777882
0.7	0.957646	0.645161	0.956508	0.770886	0.957494	0.715576	0.955815	0.809028
0.8	0.971041	0.635974	0.972872	0.804233	0.970930	0.729688	0.970332	0.836364
0.9	0.982909	0.551220	0.987654	0.849650	0.982558	0.736544	0.984450	0.865196

Table 1. Accuracy as minimum confidence limit increases. M = Magnitude trained model, C = Color trained model, O = Original training set, S = SMOTE training set, W = QSO included, W/O = QSO not included.

The last thing we checked is the accuracy at minimum confidence limits (Table 1). This is done by filtering the entire validation set by the confidence score of the determined classes (TP and FP class probabilities) and using the accuracy formula. Again, for clarity we determined a score including true quasars and one without for each model, as without is where the differences are much clearer. We found that as the minimum confidence increases, the accuracy generally increases as expected. However, we also found that the SMOTE trained models outperformed the original models for non-quasars by more than 0.08 at confidence limits >0.6, with a larger margin on the magnitude models than the color models.

IV. CONCLUSION

Astrophysicists need ways to analyze large datasets quickly and effectively. In this work, we created and tested a pipeline for the automatic classification of active galactic nuclei using the Random Forest Classifier with optical and infrared photometry from the *Gaia* and WISE space telescopes. The source dataset was comprised mostly of quasars, which had a noticeable effect on the statistical validation. The classifier performed extremely well on quasars, while for blazars and LINERS the models were at best unsuccessful, and the Seyfert classes were somewhat successful. We could expect that given a higher saturation of Seyfert AGN in the sample, their performance could be much improved. The use of SMOTE resulted in a tradeoff of precision in favor of recall for non-QSO classes, a change in feature importances compared to the original models, and increased the accuracy of high confidence classifications. The feature importance scores might indicate that the optical is an important place in the spectrum for identification of quasars, while the infrared may be important for the other classes. This could be explored by examining and characterizing specific spectral features typical of each class and comparing those characteristics with the passbands used here and the feature importance scores. This method could be further tested with the use of different photometric bands to explore other wavelength regions, the use of a more even or more representative dataset, and with more classifier parameters or other types of classifying algorithms. It could also be useful to test limited magnitude bands and classes to more explicitly explore the relationship between classes.

REFERENCES

Denisko, D, and M.M Hoffman. "Classification and interaction in random forests." *Proceedings of the national Academy of Sciences*, 2018.

Ryden, Barbara, and Bradley M. Peterson. *Foundations of Astrophysics.* Pearson, 2011.

Urry, C M, and P Padovani. "Unified schemes for radio-loud active galactic nuclei." *Publications of the Astronomical Society of the Pacific* 107, no. 715 (1995): 803.