HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

# Using Supervised Machine Learning Models to Find Active Galactic Nuclei in Multiwavelength Datasets

Rushabh Jain[1] and Antonio Rodriguez[#]

[1]Dubai College, United Arab Emirates
[#]Advisor

## ABSTRACT

Active galactic nuclei (AGN) are unique astronomical sources that emit intense radiation over the entire electromagnetic (EM) spectrum from radio to gamma ray frequencies. They are estimated to be vastly under classified with only 0.1% of them predicted to have been discovered (Padovani). A major contributing factor to this is that in the past high-quality data was not available for sources across multiple wavelengths. It is often not possible to classify a body as an AGN looking at just one section of the EM spectrum due to the variance in emissions from different sources. This implies looking at just one wavelength will lead to under-coverage in all but one type of AGN. With the emergence of new telescopes and technologies, this paper has managed to collate multi-wavelength data on galactic and extragalactic sources from the XMM-Newton and Gaia, and eROSITA telescopes. Applying various supervised learning models, such as random forests and histogram-based boosting on these sources, has allowed us to classify the AGN in these surveys with a ~97% accuracy using only emission data i.e. without including redshifts. These findings indicate that applying similar models to data collected from these telescopes should help overcome the current under-coverage in AGN.

## Introduction

AGN are extremely bright regions at the center of certain galaxies, much brighter than can be explained by the stellar population alone (ESA/Hubble). The brightness is explained by a supermassive black hole at the center of the galaxy that emits radiation across all parts of the EM spectrum as it accelerates and then accretes matter near it. These black holes can often outshine the luminosity of the entire galaxy.

There are subtle differences in the different classifications of AGN such as which emission lines are present. However, most of the classification can be explained by the relative orientation between the observer and the accretion disc as well as the presence or lack of relativistic jets. For the purpose of this paper, we will ignore the specific classifications of objects and focus on classifying objects into AGN or not AGN only.

AGN are only slightly understood and explained by our current understanding of physics. They are one of the most extreme objects in the universe and if understood could help develop future models of physics. Despite their importance, we have only managed to discover 0.1% of these predicted to be in our observable universe (Padovani). As such, a better multi-wavelength automated model to predict the presence of an AGN is necessary to the development of physics.

This paper will focus on classifying these objects into AGN or other based on multi-wavelength datasets from the Gaia and Erosita telescopes as well as the XMM-Newton survey. These datasets consist of high-quality measurements of astronomical objects including their motion and position assembled using the TOPCAT software. This data is first cleaned to filter out any objects with luminosity too low to be an AGN. Then, based

on the initial data exploration, random forest models and histogram based boosting classifiers will be trained using features selected from this data.

The models will be evaluated on how accurately they can classify AGN and a final model with an accuracy of over 95% on the testing dataset would be considered suitable for these purposes. Additionally, for objects where a spectroscopic redshift has been calculated, the model can be analyzed as the initial data cleaning would filter out any faraway objects that are not AGN. The model can therefore be evaluated by its performance on classifying faraway objects (z>1e-4) as AGN.

## Background

### Description and Classification of AGN

AGN are classified into various categories but mainly consist of the same structure. As shown in figure 1, the AGN consists of a supermassive black hole around which matter is accelerated in an accretion disc. As this matter is accelerated and absorbed by the black hole, the matter is heated to high temperatures due to viscosity (Peterson et al. 465-480). This disc emits the measured radiation up to the optical-ultraviolet waveband. A hot corona is formed above this disc which can also emit radiation in the X-ray waveband (Marinucci et al. 44). This rather exotic feature, the AGN corona, bears strong resemblance to the corona seen much closer to home, in our Sun. They are both regions above (or farther out from) the main emitting region, but are super-heated ("What Is the Sun's Corona?") to millions of degrees Kelvin. The reasons for this heating are still an outstanding question in astrophysics. Although some of the radiation is absorbed by interstellar gas and dust, including that in the dusty torus surrounding the accretion disc, this is re-emitted in the other wavelengths, most frequently the infrared spectrum ("What Are Active Galactic Nuclei?").

Some AGN also produce twin relativistic jets which appear in opposite directions from the accretion disc. These consist of outflows that are accelerated to near the speed of light and can stretch for thousands of light years in both directions. If these jets are pointed towards the observing telescope, the AGN appears extremely bright across the entire EM spectrum.

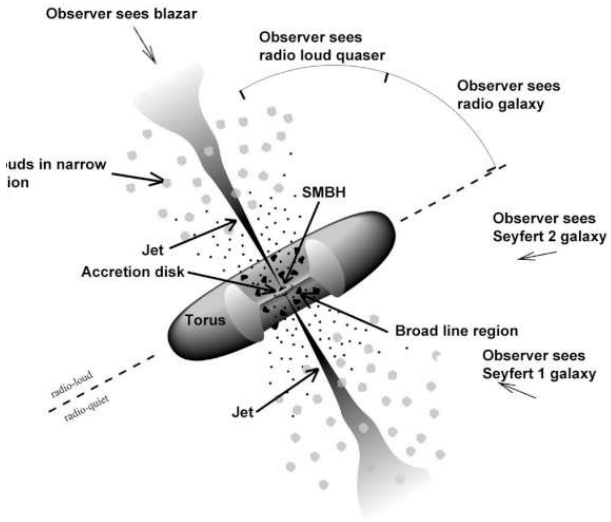


**Figure 1.** Diagram of a typical AGN and how it would be classified based on the orientation of the observer ("Gamma-ray Space Telescope: Exploring the Extreme Universe")

As indicated on the diagram the AGN can appear different based on the relative angle between the observer and the accretion disc and thus may be classified differently despite having the same base structure as other AGN.

## Issues Classifying AGN

In previous investigations of AGN, the radio wavelength has been thoroughly explored. However, AGN without jets are relatively difficult to classify looking at just one waveband as emissions are very faint as shown in Figure 2. Although analysis in this waveband is useful it is typically not enough to find AGN on a large scale without the inclusion of other measurements (Padovani).

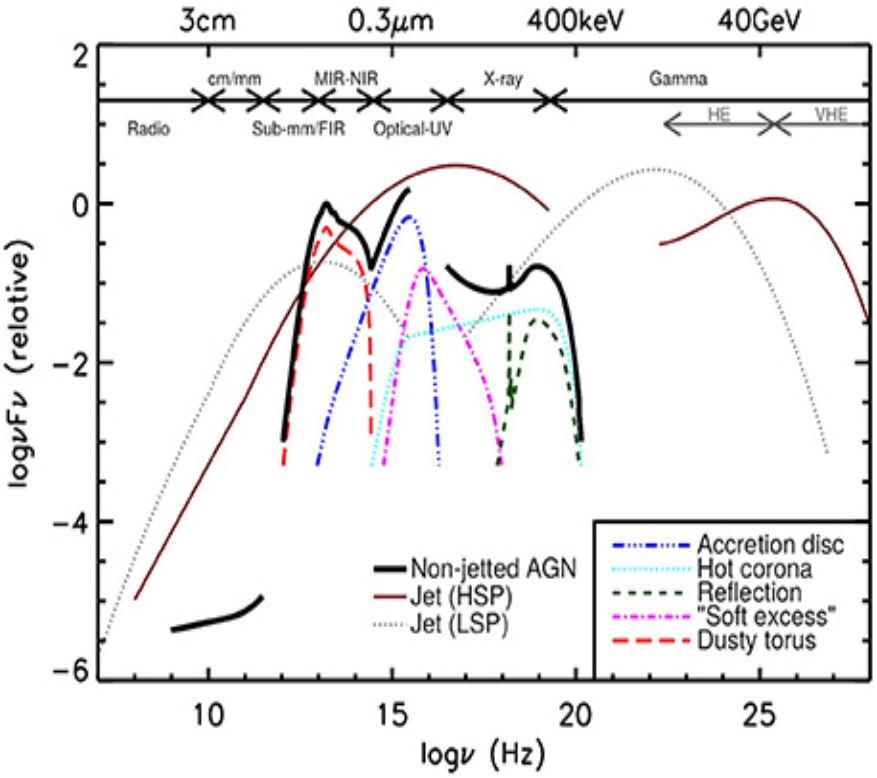


**Figure 2.** Graph representing the typical relative emissions of different types of AGN across the various wavebands of the EM spectrum (Padovani)

The infrared waveband is perhaps the most important to classifying AGN, especially non-jetted specimens. In addition to the natural emission in this waveband, other emissions such as ultraviolet are also often absorbed by the dusty torus and re-emitted in infrared. Another advantage of analysis in this waveband is that it is reliable for both obscured and unobscured AGN, allowing us to classify AGN that may have otherwise been missed in other wavebands. However, issues arise classifying jetted AGN as low accretion rate jetted sources often do not have any dust which would create the luminosity in this waveband. Additionally, for deep surveys high reliability in classification can only be achieved by having a low completeness (missing AGN above a certain energy limit) (Padovani).

As mentioned previously, optical and ultraviolet emissions are mainly released from the accretion disc. In many cases, this can be absorbed by the dust surrounding the AGN, and analysis in this waveband provides a very biased view of AGN. Although it is useful for actual understanding of the AGN, it is typically limited in its usefulness for classifying AGN as it only helps discover unobstructed sources (Padovani).

X-rays are perhaps the most important waveband in classifying AGN. They are mainly released from the corona (the atmosphere formed above the accretion disc). They then interact with the matter in the accretion disc and dusty torus and are scattered, reflected or absorbed. X-ray analysis, therefore, theoretically involves all AGN but these are often systematically missed and certain types of AGN tend to be missed due to their weak X-ray emissions (Padovani).

The gamma ray waveband is relatively unexplored and is only significantly present in jetted AGN. It has very few sources detected in it and is not suitable to detect AGN en masse.

Overall, there is a large plethora of surveys detecting AGN in one specific waveband such as the following survey using infrared data (Poleo et al.). Despite this, there is a distinct lack of analysis across multiple wavelengths which could help overcome some of the shortcomings of the individual wavebands discussed in this section. This paper will attempt to use multiwavelength analysis to help aid in classifying AGN using an automated process, despite the characteristics that make it difficult to categorize.

## Dataset

The machine learning models trained in this paper are based on two datasets. Both have data on astronomical objects including measured emissions, motion, spectroscopic redshifts, and previous classifications where possible. All data was combined using the TOPCAT software and the datasets used have been filtered so only objects above a certain luminosity are kept as any below this threshold cannot be AGN.

Both datasets include a variety of features and measurements. They include the right ascension and declination of the object which are used to identify the object and act similarly to coordinates.

The datasets additionally include the parallax as well as the standard error of this measurement for each object. The proper motion (ascension and declination) as well as the standard error associated with each of these measurements have also been collected from the Gaia telescope and included.

The GALEX telescope has also been used for its measurements of the far-ultraviolet and near-ultraviolet emissions of these objects. The associated standard error of these measurements has also been included in the dataset.

Near-infrared brightness and the related standard error for the emissions of each of these objects has been collected from the VISTA telescope. Three bands have been created in this waveband to provide more analysis for the model which are, in increasing order of energy, Ks, H and J.

The measurements for the mid-infrared brightness as well as the standard error are taken from the WISE telescope. Similar to the near-infrared brightness, these are split into 4 sections to provide more distinction. These have been labeled, in increasing order of energy W4, W3, W2 and W1.

The DeCals telescope has also been used for its measurements of the optical emissions of these objects. These have also been split into three categories, namely z, r and g in order of increasing brightness (there is also the i and y categories in the latter dataset). The associated standard error of these measurements has also been included in the dataset.

Lastly, the dataset includes the object's brightness in the radio waveband and the standard error of this measurement. This data has been collected from the Very Large Array Telescope.

In addition to all these measurements the datasets include extra fields based on previous literature. Firstly, for objects where it has been calculated using emission spectroscopy, the spectroscopic redshift is recorded. Additionally, the datasets also include the classification (i.e. star, exoplanet etc.) of an object if one has been assigned to it in previous literature. If such a classification exists, the dataset also includes the number of references to the object under this classification.

### Efeds Dataset

This dataset is based on a survey by the EROSITA telescope that measures the X-ray emissions of objects in a 140 square degree section of the sky (out of 42,000 square degrees). This dataset then includes the aforementioned categories taken from their respective telescopes as well two additional categories. The first measures the quality of the measurements on a scale from 1-4 and is labeled CTP_quality. The other is a classification of the object as galactic or extragalactic for objects where such a classification has been made.

The dataset is relatively small and contains only ~27,000 samples, however, this is likely sufficient to train a model to a relatively high level of accuracy. Additionally, another issue with the dataset is the lack of values for a significant proportion of objects for some features, namely near-infrared brightness.

Despite this the dataset has values across the entire spectrum of AGN. As illustrated by the histograms in figures 3 and 4, there is a wide spread of data that is sufficient to train the model for all types of AGN despite any confounding factors.

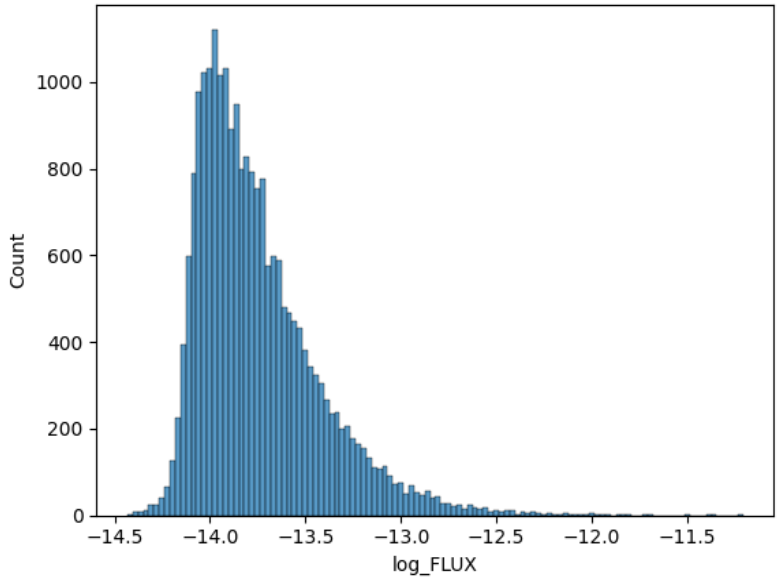


**Figure 3.** A histogram plot of the number of objects against their X-ray luminosity measured in a logarithmic scale. Only shown for the objects where this value has been measured and this value is not null.

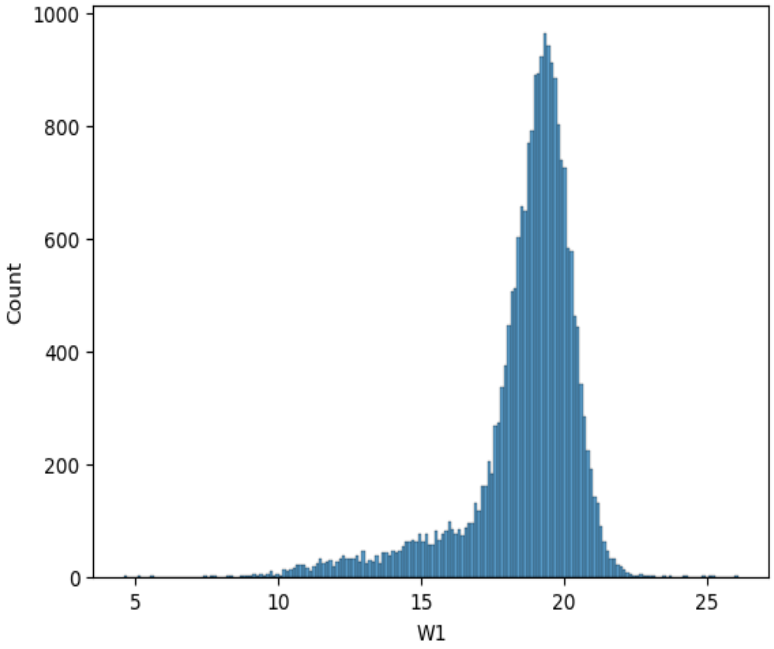


**Figure 4.** A histogram plot of the number of objects against their mid-infrared brightness measured in the W1 waveband. Only shown for the objects where this value has been measured and this value is not null.

The data presents some initial patterns that should be noted. With the use of colors (the difference in magnitude between two different wavebands), AGN typically presents much lower than other objects when looking at g-r in the optical waveband. Additionally, they are much higher than other objects when looking at the W1-W2 color. This can be seen in figure 5 which shows the color against the X-ray brightness of all objects and colors them based on their known classification (this value is explained in the methodology).

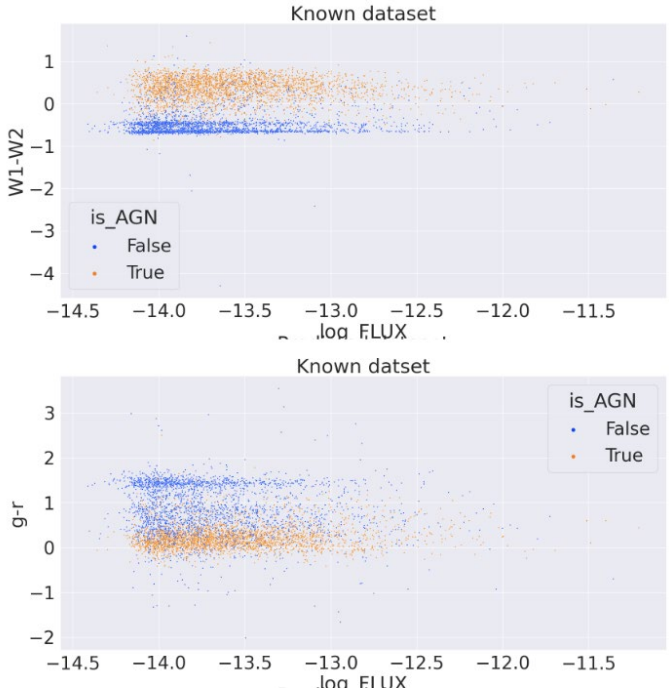

**Figure 5.** Scatter plots representing the distribution of AGN against other objects with the g-r and W1-W2 colors against the X-ray luminosity. AGN are represented in orange and other objects in blue.

The features selected from this dataset were the mid-infrared, optical and x-ray wavelengths. These were suitable for this purpose as they contained values for most objects and were not null. Other features, such as the near-infrared were not considered as they contained many null values and thus would significantly reduce the number of samples that could be used to train a random forest model. In addition to this, the CTP_quality and classification as well as the proper motion and parallax were used to clean the data and classify objects as galactic or extragalactic. Lastly, the classification and references in previous literature are used to classify some objects into AGN or not AGN which are then used as a key to train a supervised machine learning model. The details of this are included in the methodology.

## XMM Newton Dataset

The data from the XMM telescope comes from the XMM Serendipitous Point Source Catalog, which was formed from the entire collection of its observations. These observations are quasi-random, having been compiled from many proposals to observe different science targets and with different observing strategies. This differs strongly from how the eFEDS dataset was created, because the eFEDS dataset only comprises a single type of observation over a small patch of sky. The XMM dataset, however, contains patches all across the sky (but does not cover the entire sky uniformly). This data covers 1200 square degrees of the sky. The key features comparing both datasets can be found in table 1.

This dataset is significantly larger and contains an excess of 200,000 samples. This provides significantly more data and should provide a more complete model that is less likely to mis-predict or overfit to the training data. Although this dataset includes significant null values, these can be solved by using models that can deal with null values or simply ignoring these objects altogether as the larger quantity of data allows for the model to view patterns even if these values are ignored.

Similar to the other dataset, this data also includes objects across the entire spectrum of AGN as illustrated by figure 6. This helps train the models as it gives them information on all types of objects they may encounter.

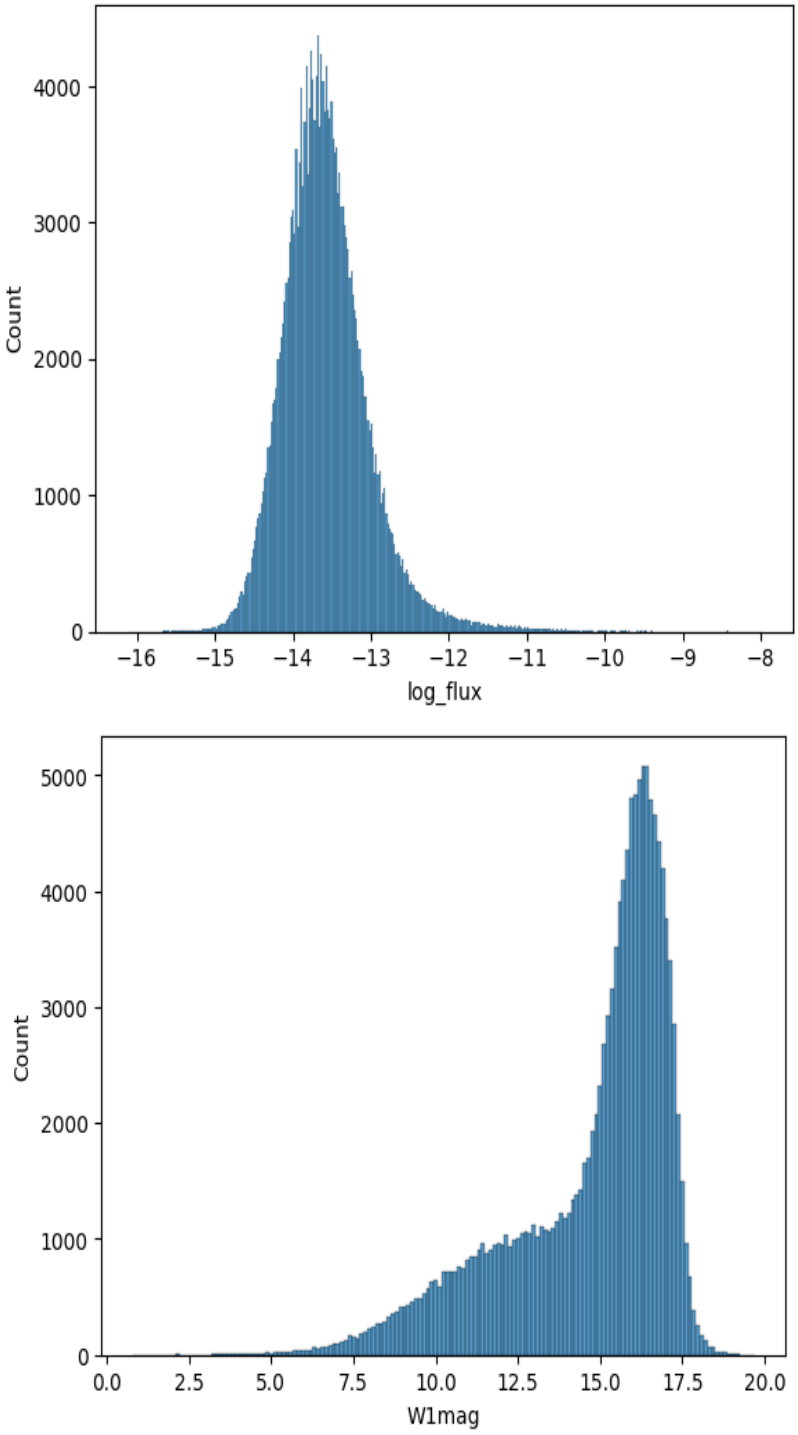


**Figure 6.** Histograms of the number of objects against the X-ray flux(erg/s/cm^2) measured in a logarithmic scale as well as the mid-infrared brightness in the W1 waveband.

This dataset also presents the same patterns in the colors as described by the other dataset. This can be seen by the scatter plots in figure 7. As shown, AGN are typically higher in the W1-W2 color and are typically lower in the g-r color.

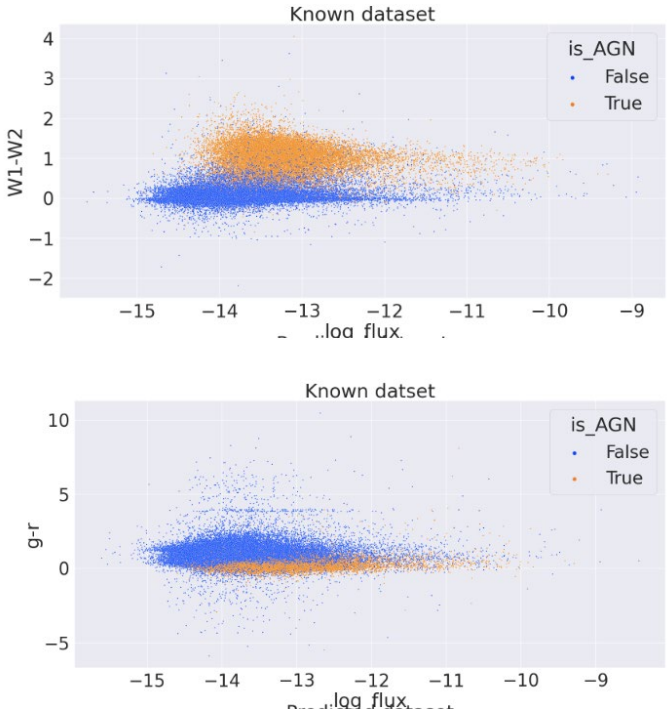


**Figure 7.** Scatter plots representing the distribution of AGN against other objects with the g-r and W1-W2 colors against the X-ray luminosity from the XMM Newton dataset. AGN are represented in orange and other objects in blue.

Following the same reasoning as the previous dataset, only the mid-infrared, x-ray and optical wavelengths were used as features to train the model. Similar to the simbad dataset, the proper motion and parallax were used to classify objects as galactic and extragalactic. Lastly, the previous classification and references in literature were used to classify some objects as AGN, which are used to train the supervised learning models.

**Table 1.** A table illustrating key values to show the kind and amount of objects provided by both datasets. All units of flux are in (erg/s/cm^2). Other categories are magnitudes and therefore unitless.

| | eFEDS | XMM-Newton |
|---|---|---|
| Number of rows | 22256 | 221,124 |
| Area of sky | 140 square degrees | 1200 square degrees |
| Median X-ray flux | 1.5e-14 | 2.38e-14 |
| Median optical | g: 21.195345<br>r: 20.7580655<br>z: 20.342543 | g: 20.7677<br>y: 20.7677<br>r: 20.4854<br>z: 19.8567 |

| | | |
|---|---|---|
| | | i: 20.2589 |
| Median mid-infrared | W1: 19.09067<br>W2: 18.8890285<br>W3: 17.5750105<br>W4: 15.5127975 | W1: 15.595<br>W2: 14.855<br>W3: 12.11<br>W4: 8.904 |

## Methodology

### Efeds Dataset

Firstly, any objects with a CTP quality of 2 or below are removed to ensure only high-quality data is kept. After this, a new feature classifying objects as galactic or extragalactic is created. This is based on readings from the Gaia telescope and based on the CTP classification. If any object is classified as secure galactic or secure extragalactic based on the CTP classification, then this value is used. When this is not the case, the parallax and proper motion is checked. If the parallax divided by the parallax error is greater than 3, then it is classified as galactic. The same calculation is performed with both proper motions. At this point, any object that does not have a classification is stored as unknown.

Following this a new column containing a classification of an object as AGN, not an AGN, or unknown is created. This is done based on the previous classifications of an object. The dataset contains classifications of 6 different types of AGN, namely: QSO, Seyfert 1, Seyfert 2, BLLac, Blazar and RadioG. It also contains a general classification of AGN for any objects classified as AGN but not assigned a type. Any objects that have been classified under these categories in 3 separate pieces of previous literature are considered to be AGN. Any objects that have been classified in another category (i.e. star etc.) in 3 or more previous publications are considered to be not AGN. Any objects that have been previously classified as galactic are also considered to be not AGN in this stage of the process. Any remaining objects are classified as unknown. This process provides approximately 2,700 AGN, 2,600 non-AGN and 16,000 unknown objects.

At this point only known objects are considered and are used to train a random forest model. 80% of the data is used to train the model and 20% is used to test its accuracy. The random forest model works based on the process described in figure 8. The model builds multiple decision trees each of which predict an individual result for the parameters provided. Then these results are averaged, or in this case majority voting is applicable, and this is output as the final result.

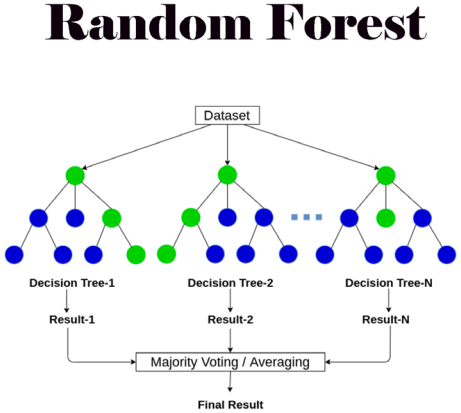

**Figure 8.** A schematic diagram of a random forest model ("Random Forest Algorithm Explained .")

Since a random forest model cannot handle null values when training this model, any values with null value are ignored. At this point, in order to find the ideal hyperparameters for the model, a grid search is used. This involves training a model with each combination of hyperparameters and evaluating each of them based on a performance metric, in this case accuracy. For this model, significant changes in accuracy will be caused by the number of estimators and the minimum samples for each split. The results of this search are shown in figure 9.

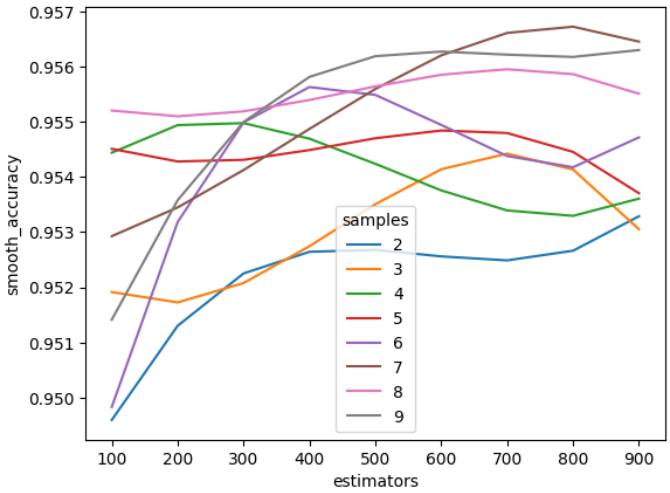


**Figure 9.** A smoothed line graph showing the accuracy of a random forest model based on the min_samples split and estimators.

Following this another refined grid search with 5-fold cross validation is used. This is a resampling technique to ensure that a certain model or set of parameters is not successful just due to the training data. This trains 5 separate models using a different training dataset each time and returns the model with the highest average accuracy as illustrated in figure 10. This helps ensure the model is applicable in practical situations.

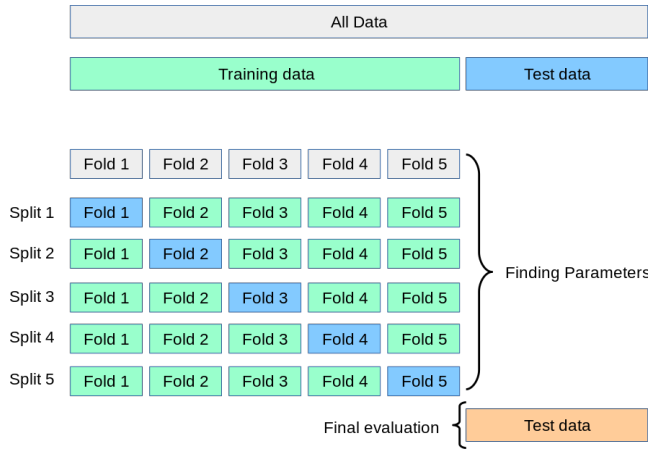


**Figure 10.** A graphical depiction of the cross validation algorithm (Scikit Learn)

After this another model is trained using the same process. Note that one major difference between this model and the previous one is that this model can handle null values. This makes it more useful practically as well as giving it access to more training data. Thus, a new training set is created for this model with the same features as before, but this time including rows with null values. This is a histogram based boosting algorithm. This works similarly to the random forest model in that it uses a series of decision trees. However, the difference in boosting algorithms is that the new trees are trained based on the weaknesses of the existing model which helps minimize the overall loss. This process is depicted here in figure 11 which shows how the models are trained to create an ensemble with a minimized loss.

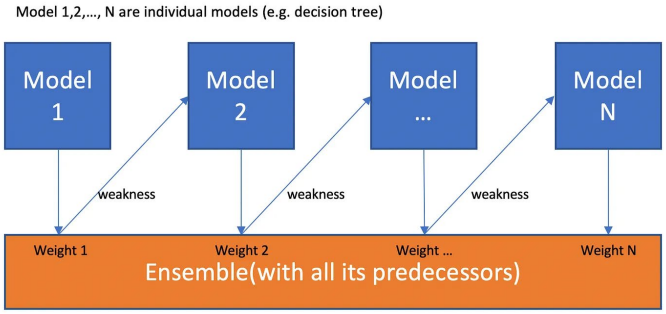


**Figure 11.** A graphical depiction of a general boosting algorithm and the process it uses to minimize loss (Zhang).

This model should typically generate better results than a random forest as it addresses weaknesses of previous models rather than building a completely new decision tree in each iteration. This can be seen in figure 12 which shows the grid search accuracies for the histogram-based model. As shown here, this reaches slightly higher accuracy than the random forest model but these are significant when applied on a large scale. In the grid search for this model, the main factors that are considered are the learning rate and the number of iterations (unique decision trees) that the model goes through.

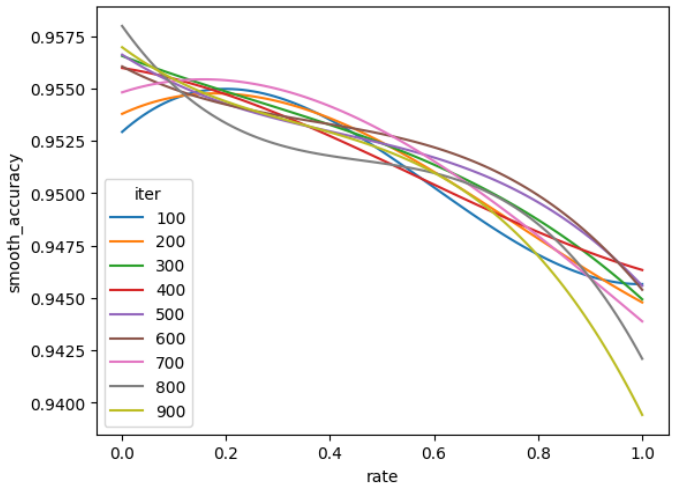


**Figure 12.** A smoothed line graph showing the accuracy of a model as the learning rate and iterations of a histogram based boosting model change.

After this both models are used to make predictions on the overall dataset. This is used to evaluate the final performance of the models and is how they would be applied in the real world. The details of this process are described in the results section of this paper.

## XMM Newton Dataset

This dataset does not contain the CTP classifications of different objects. The classification of objects in this case is done based on the parallax and motion only. For the parallax and the motion, if the measure value divided by the error is greater than 3 the object is classified as galactic, if this metric is between 1 and 3, it is classified as unknown and is otherwise classified as extragalactic. This is passed as a parameter to this model and is more useful than simply removing galactic objects as it gives the model more training information.

Following this, the same process as the Efeds dataset is used to classify these objects into AGN. Due to the much larger size of this dataset, this results in much more training data being available. After applying this algorithm, we find that the data contains approximately 23,000 AGN, 45,000 non–AGN and more than 150,000 unknown objects.

The same procedures are then used to create 2 different models from this dataset. The random forest model, again, can only deal with non-null values and does not improve on previous trees and as such has a slightly lower accuracy. The training parameters are then again adjusted using the grid search with cross validation method. The results of this can be seen in figure 13. This also shows how the increase in data helps create a more accurate model as these are approximately 1 percentage point higher in accuracy than the previous dataset. This suggests that the same methodology could be used to create better models when more astronomical surveys are performed.

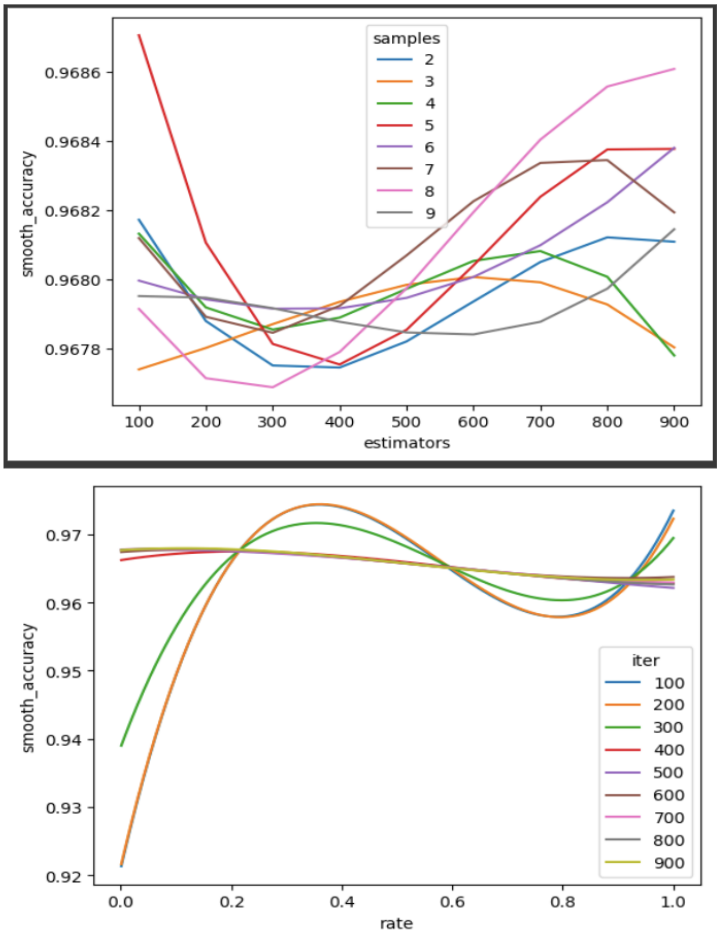


**Figure 13.** Smoothed line graph showing how different hyperparameters affect the accuracy of the random forest and boosting model respectively.

The evaluation of these models are again done by applying them to the overall dataset including the unknown values. This is done similarly to the previous dataset and is detailed in the results section.

## Results

Using the methodology detailed above we have managed to create 4 unique models to classify objects into AGN. The three main evaluation metrics are represented in table 2 for each of these models. The accuracy is calculated by dividing the number of true positives and the number of true negatives by the total number of objects. The precision is defined as the number of true positives divided by all positives. The recall is the number of true positives divided by the number of true positives and false negatives.

**Table 2.** A table describing the success metrics of the different models developed for this paper.

| Dataset | Model | Accuracy | Precision | Recall |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| eFEDS | Random Forest | 0.957 | 0.955 | 0.967 |
| | Histogram based boosting | 0.960 | 0.951 | 0.973 |
| XMM-Newton | Random Forest | 0.968 | 0.943 | 0.956 |
| | Histogram based boosting | 0.968 | 0.944 | 0.964 |

As shown, the models developed here have between a 96-97% accuracy, which is close to perfect when compared to other models without the use of spectroscopic redshift. The improved accuracy of the models with more training data using the latter dataset can be seen as they have on average one percentage point higher accuracy. Additionally, the advantages of the histogram based boosting model are prominent as, despite the similar accuracy scores, the recall is significantly higher for these models. For practical purposes, the ratio of true positives compared to false negatives are at least as important, if not more, than the accuracy of the model. This is because the aim of the model is to find new AGN and is not concerned with false positives as significantly.

The exact values of these results can be seen in figure 14 in which the heatmaps for the most successful model, the histogram-based boosting using the XMM-Newton dataset are shown.

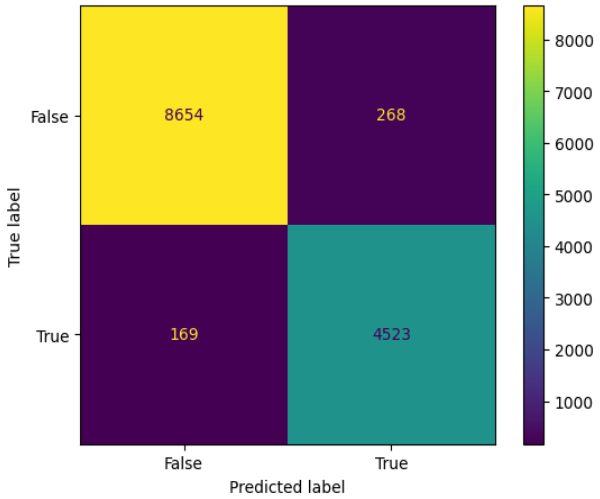


**Figure 14.** A heat map representing the success and failures of the most successful of the four models that were trained.

The success of the model can also be seen from the patterns observed in the dataset section using the W1-W2 and g-r colors. This predictive algorithm is applied to the unknown values in the dataset and is used to assign a classification on these objects. When these are graphed in the same manner as previously, with the color

against the flux of the object we can see the same patterns emerge. This can be seen in figures 15 and 16 where the most successful model has been used for this purpose.

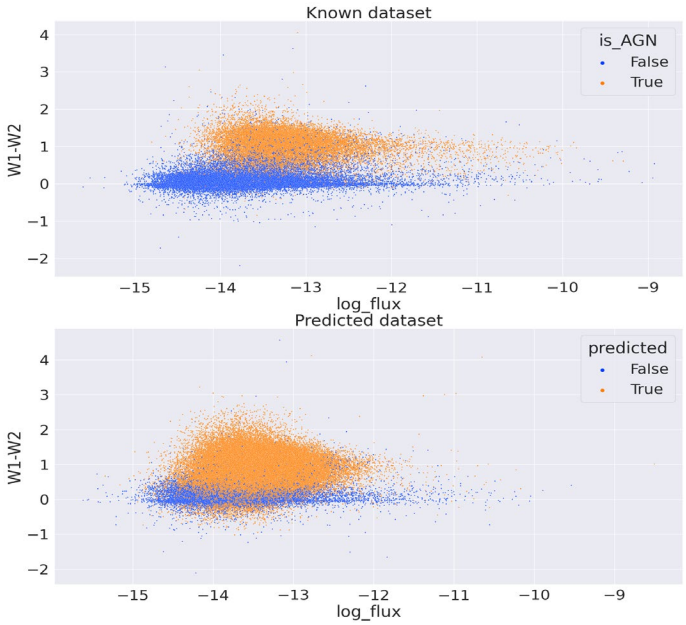


**Figure 15.** A comparison of values of the W1-W2 color compared to the log flux of an object for both the known and predicted dataset generated using the histogram based boosting model.

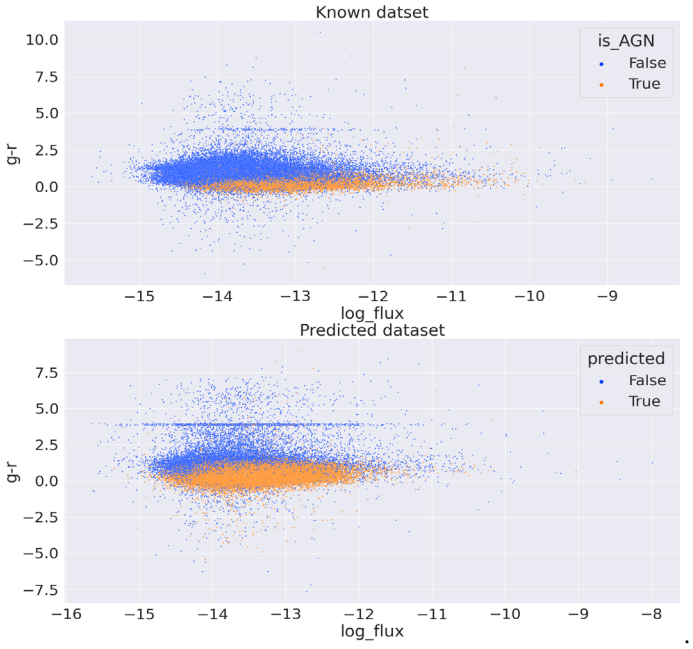


**Figure 16.** A comparison of values of the g-r color compared to the log flux of an object for both the known and predicted dataset generated using the histogram based boosting model.

As visible in these plots, the model has picked up the same patterns that were visible in the initial exploration of the data. This shows the suitability of the models and how they are likely to be successful in application.

The last check performed to check the reliability of the model was to check its predictions for values with a known redshift. During the formation of the datasets, any objects with luminosity too low to be AGN were filtered out. As such any object that is far away is very likely to be an AGN. Therefore, any object with redshift greater than 1E-4 can be considered to be an AGN for these purposes. It is important to note that the model itself did not have access to these redshift values and they were added retrospectively. The predictions from this model for objects that have a recorded red shift can be seen in figure 17.

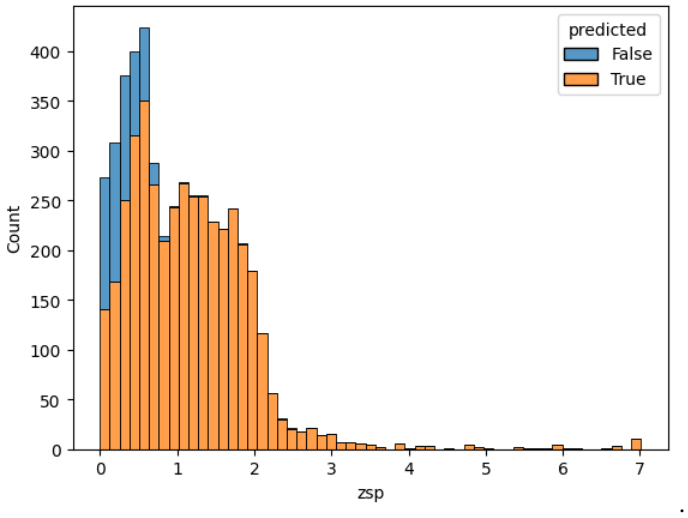


**Figure 17.** A histogram of the redshift of an object with the hue representing the prediction of the histogram based model trained on the XMM Newton dataset.

Overall, the model has been quite successful and passed most tests. The accuracy of this methodology is likely to keep improving as more data is collected, especially in the wavelengths that were not considered here. Despite this, this model is already quite suitable for these purposes and could be applied on future surveys using these telescopes.

## Discussion

It can be concluded that the models presented here found the same patterns that astrophysicists themselves use to discover new AGN. In addition, the high recall and accuracy of the models mean that very few AGN are likely to be missed using these models. This means that after manual screening we can use the histogram based boosting model in particular to significantly improve on our current under-coverage of AGN.

As shown here, this model has been quite successful overall. It can accurately classify objects across a wide range, with the furthest away object in this dataset being almost 13 billion light years away. The relatively high accuracy rate of approximately 97% means when future surveys are conducted using the same telescopes the same model can be applied. This is essential to the development of modern physics as it enables an automatic discovery of one of the most extreme objects in the universe. Additionally, as pointed out here, the large range of these telescopes means that this technology can help understand how AGN developed in conjunction with the universe as the travel time of the radiation allows us to explore earlier states of the universe.

Despite this, the model is not perfect at classifying AGN, especially those closer to our galaxy. As shown in Figure 17, the accuracy rate of the model significantly improves when the redshift of the object is greater than 1. This is most likely caused by a lack of sufficient data on extragalactic objects that are relatively close to our galaxy leading to inaccuracies in the model. This issue could be corrected in future studies of this issue as more data is collected.

An additional limitation of the model is that due to lack of sufficient data it was only possible to analyze certain parts of the EM spectrum whilst the rest were ignored. Further measurements, especially in the near infrared brightness are likely to bring improvements in accuracy to the model as a large amount of information can be found in this waveband as explained in the background. Furthermore, improvements in other fields of AI and physics have led to large developments in calculating photometric redshifts using neural networks instead of the spectroscopic method (Henghes et al.). The addition of information on redshift of an object would lead to an almost perfect accuracy and would eliminate the issue with this model discussed previously. The calculation of photometric redshift from telescopic data, when sufficiently advanced, should be included in the feature list provided to this model in order to create a better performance.

Next steps involving this data include using physical models to interpret these AGN in the context of their environment. This could then be used to solve existing mysteries in our understanding of these black holes such as the "big blue bump" (Huang et al.) or the reason for the existence of the relativistic jets produced. This allows the machine learning model to help improve the basis of theoretical physics understanding by offering more objects for observation and understanding.

## Conclusion

Overall, the model produced in this paper is successful and suitable for its purpose. It achieved a 96.8% accuracy and a 96.4% recall – this clearly illustrates that it can classify AGN accurately at a high rate. Although this can be improved by the addition of more wavebands within the EM spectrum, using the current features, the model could be used to aid in classifying AGN in future surveys with the XMM-Newton or Erosita telescopes or those with similar methodologies. This presents a novel opportunity in modern astrophysics to have a large amount of data to study one of the least understood concepts, AGN. As more data becomes available, and the resolution of telescopes improve, this model, and the use of machine learning in this field will become vital to our understanding of this concept.

## Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

## References

ESA/Hubble. "Active Galactic Nucleus | ESA/Hubble." *ESA/Hubble*, https://esahubble.org/wordbank/active-galactic-nucleus/. Accessed 28 October 2023.

"Gamma-ray Space Telescope: Exploring the Extreme Universe." *Fermi Gamma-ray Space Telescope: Exploring the Extreme Universe*, 17 February 2016, https://fermi.gsfc.nasa.gov/science/eteu/agn/. Accessed 29 October 2023.

Henghes, Ben, et al. "Deep learning methods for obtaining photometric redshift estimations from images." *Monthly Notices of the Royal Astronomical Society*, vol. 512, no. 2, 2022, pp. 1696–1709, https://academic.oup.com/mnras/article/512/2/1696/6536923.

Huang, Shuang-An, et al. “Statistics of the Big Blue Bump Feature in a Low Redshift Sample of Active Galactic Nuclei.” *Astrophysics and Space Science*, vol. 280, 2002, pp. 301-316, https://link.springer.com/article/10.1023/A:1015537228840#citeas.

Marinucci, Andrea, et al. “Hot Coronae in Local AGN: Present Status and Future Perspectives.” *Galaxies*, no. 6, 2018, p. 44, https://doi.org/10.3390/galaxies6020044.

Padovani, Paolo. “Active Galactic Nuclei at All Wavelengths and from All Angles.” *Frontiers in Astronomy and Space Sciences*, vol. 4, 2017, https://www.frontiersin.org/articles/10.3389/fspas.2017.00035. Accessed 28 10 2023.

Peterson, B.M, et al. “Quasars.” *Encyclopedia of Physical Science and Technology (Third Edition),*, Academic Press, 2003, pp. 465-480, https://doi.org/10.1016/B0-12-227410-5/00629-3.

Poleo, Valentina, et al. “Identifying Active Galactic Nuclei at z ∼ 3 from the HETDEX Survey Using Machine Learning.” *The Astronomical Journal*, vol. 165, no. 4, 2023, https://iopscience.iop.org/article/10.3847/1538-3881/acba92.

“Random Forest Algorithm Explained .” *Anas Brital*, 20 September 2021, https://anasbrital98.github.io/blog/2021/Random-Forest/. Accessed 6 November 2023.

Scikit Learn. “3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.3.2 documentation.” *Scikit-learn*, https://scikit-learn.org/stable/modules/cross_validation.html. Accessed 11 November 2023.

“What Are Active Galactic Nuclei?” *Webb Space Telescope*, 17 March 2021, https://webbtelescope.org/contents/articles/what-are-active-galactic-nuclei. Accessed 29 October 2023.

“What Is the Sun's Corona?” *NASA Space Place*, https://spaceplace.nasa.gov/sun-corona/en/. Accessed 19 November 2023.

Zhang, Zixuan. “Boosting Algorithms Explained. Theory, Implementation, and… | by Zixuan Zhang.” *Towards Data Science*, 26 June 2019, https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30. Accessed 11 November 2023