

CLASSIFICATION OF EXTRAGALACTIC X-RAY SOURCES USING
MACHINE LEARNING METHOD

by

Mehrdad Rostami Osanloo

A thesis submitted to the Graduate College of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Physics
August 2019

Committee Members:

Blagoy Rangelov, Chair

Donald Olson

Meysam Khaleghian

COPYRIGHT

by

Mehrdad Rostami Osanloo

2019

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgement. Use of this material for financial gain without the author's express written permission is not allowed.

Duplication Permission

As the copyright holder of this work I, Mehrdad Rostami Osanloo, authorize duplication of this work, in whole or in part, for educational or scholarly purposes only.

DEDICATION

This Thesis is wholeheartedly dedicated to my wife, for her kindness and devotion, and for her endless support, to my beloved mother who gives me emotional and moral strength and is dedicated to the memory of my father who left fingerprints of grace on my life.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor, Dr. Blagoy Rangelov, for the continuous support of my Masters study and related research, for his patience, motivation, feedback, accessibility, and immense knowledge. His guidance helped me in my research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee, Dr. Donald Olson and Dr. Meysam Khaleghian, for their insightful comments and encouragement, but also for the hard questions which incited me to widen my research and gain various perspectives.

Last but not least, I would like to thank my wife for supporting me spiritually throughout writing this thesis and my life in general. I share this accomplishment with my wife and thank her deeply.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	xi
I. Introduction	1
I.1 History of X-ray Missions	1
I.2 Common Types of X-ray Detectors	3
I.3 Astrophysical Mechanism for Producing X-rays	4
I.4 Astrophysical Sources of X-ray Radiation	6
I.5 Measurable Properties of Astrophysical Sources	9
I.6 Apparent Magnitude	10
I.7 Photometric Filters	11
II. Machine Learning Methods	13
II.1 Types of Machine Learning Methods	13
II.2 The Machine Learning Process	15
II.3 Features	17
II.4 Cross Validation	17
II.5 Overfitting and Underfitting in Machine Learning	17
II.6 Feature Importance	18
II.7 Decision Tree	19
II.8 Ensemble Learning and Random Forest	19
II.8.1 ML Python Libraries	21

III. Classification Results	22
III.1 Training Data Set	23
III.2 Oversampling	23
III.3 Missing Data	24
III.4 Normalization and Standardization	24
III.5 Feature Selection	25
III.6 Hubble Space Telescope Filter Transformation	26
III.7 Classifier Fitting	28
III.8 Verifying the Model	29
III.9 Classification Performance	29
III.10 Classification Accuracy	31
IV. Future Work	33
APPENDIX SECTION	35
REFERENCES	37

LIST OF TABLES

Table		Page
I.1	The X-ray satellites, their missions and the scientific highlights. . . .	3
I.2	The Sloan Digital Sky "ugriz" System.	12
III.1	Training dataset of 5 different object classes for the RF classification.	23

LIST OF FIGURES

Figure	Page
I.1	Chandra’s X-ray mirror assembly focuses X-rays onto a detector to produce an image. (Credit: NASA/CXC/D. Berry) 2
I.2	Chandra X-ray Observatory(Credit: NASA). 3
I.3	NuSTAR X-ray Observatory (Credit:NASA). 3
I.4	Accelerated Charges and Bremsstrahlung Process. (Credit: Chandra X-ray Observatory) 5
I.5	Spectral forms expected from different astrophysical processes. (Adapted from Seward and Charles 2010) 7
I.6	U, B, and V filters: the wavelength to which the standard filters are transparent. (Credit: Freedman et al. 2011) 12
I.7	This flux (or color) ratio is small for hot, blue stars but large for cool, red stars.(Credit: Freedman et al. 2011) 12
II.1	K-fold cross-validation (Credit: Researchgate) 18
II.2	Examples of overfitting and underfitting in ML (Credit: Machine-learningmedium.com) 19
II.3	The decision tree is typically read from top (root) to bottom (leaves). A question is asked at each node (split point) and the response to that question determines which branch is followed next. The prediction is given by the label of a leaf. (Credit: DisplayR) 20
II.4	Random Forest Procedure (Credit: Medium.com). 21
III.1	Importance of features used by the classification ML algorithm. 27

III.2	Throughput HST and Pan-STARRS filters are shown with solid and dashed lines, respectively. Three black body models (for 3,000, 5,000, and 15,000 K, scaled to fit the figure) are shown for illustration purposes.	28
III.3	Importance of Features for the HST Filter Transformation.	30
III.4	Confusion matrix for the classification training dataset. The numerical labels correspond to AGN (0), CVs (1), HMXBs (2), LMXBs (3), and Stars (4).	31
III.5	Same as III.4 but for the validation training dataset shown in logarithmic scale for better visualization.	32
IV.1	<i>Left:</i> Existing <i>HST</i> coverage of M31. The WFPC2 pointings are shown in blue, the ACS in orange, and the WFC3 in purple. <i>Center:</i> The Chandra coverage of M31. ACIS-I pointings are shown with blue, ACIS-S with red, HRC-I with green, and HRC-S with magenta. The circle has a radius of 1 degree. <i>Right:</i> A deep optical image of M31 overplotted with the XMM-Newton fields of the XMM survey. The area covered by individual EPIC observations is approximated by circles with 14' radius.	34

ABSTRACT

Only a small fraction of extragalactic X-ray sources have reliable classifications. Although a large amount of X-ray data exists in the archives, the X-ray data alone are not enough to reveal the nature of the X-ray sources, and multi-wavelength data is the only way to make progress toward this goal. Therefore, creating an automated Machine Learning (ML) tool for classification of extragalactic X-ray sources with multi-wavelength data will enable us to understand X-ray source populations in a plethora of nearby galaxies. Modern ML methods can be used to quickly analyze the vast amount of multi-wavelength data for these unclassified sources providing both the classifications and their confidences. To this end, we have created a ML pipeline to classify extragalactic X-ray sources, which can utilize the large amount of existing archive data taken with Hubble Space Telescope. The use of the Hubble Space Telescope is essential when dealing with extragalactic sources, and we have adopted our pipeline accordingly. The tool that we have developed will open new avenues to explore extragalactic astronomy.

I. Introduction

I.1 History of X-ray Missions

Our knowledge about the Universe originates in studies of astronomical sources across the electromagnetic (EM) spectrum, from radio waves ($\sim 10^{-11}$ eV) to cosmic rays ($\sim 10^{12}$ eV). X-ray astronomy focuses on phenomena that emit radiation at X-ray energies (typically ~ 0.1 – 10 keV). However, ninety percent of the photons in a beam of 3 keV X-rays are absorbed by traveling through just 10 cm of air at sea level (blocked by the atmosphere molecules). Therefore, the earth's atmosphere is opaque to the high energy radiation such as X-ray and γ -ray, and such photons can be detected only from space (above the atmosphere) thanks to the development of advanced space technologies.

Due to the very high energies of the X-ray photons, X-ray telescopes are uniquely designed with barrel-shaped mirrors, required to focus the photons onto the detectors (Figure I.1). The X-ray “mirrors” have to be oriented at a very shallow angle such that the mirror surfaces are nearly parallel to the incoming light. The incoming photons are eventually focused onto the detectors where their energy is recorded.

The first discovery of the X-ray sky occurred in 1949 when X-rays from the Sun were discovered by rocket-borne experiments. The journey of X-ray astronomy began in the early 1970's by the advent of satellite observatories. Some of the major satellite X-ray missions are briefly described below to illustrate the vital developments achieved during the past five decades.

- In the 1970's Uhuru was the first earth-orbiting mission dedicated entirely to X-ray astronomy.
- In 1990, Roentgen Satellite (ROSAT), a joint project of Germany, the United States, and Great Britain, was launched. It operated over eight years and

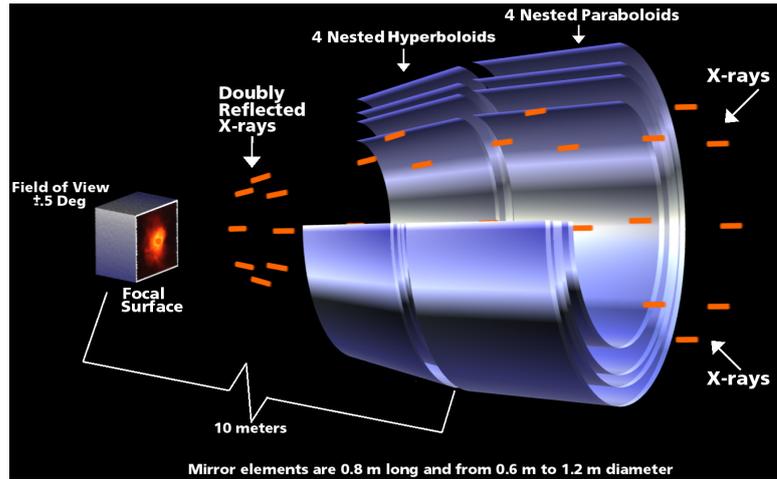


Figure I.1: Chandra's X-ray mirror assembly focuses X-rays onto a detector to produce an image. (Credit: NASA/CXC/D. Berry)

finally shut down in 1999. It was instrumental in the discovery of X-ray emissions from comets and conducted an all-sky survey in the X-ray region of the spectrum.

- In 1995, NASA orbited the Rossi X-ray Timing Explorer (RXTE) to study the variations in the emission of such X-ray sources as black hole (BH) candidates, active galactic nuclei (AGN), white dwarf stars, neutron stars (NS), and other high-energy sources.
- In 1999 the Chandra X-ray Observatory (Chandra) was deployed from a shuttle and boosted into a high earth orbit (Figure I.2). Chandra has changed the way we study population of X-ray point sources thanks to its unsurpassed (to date) resolving power. The X-ray Multimirror Mission (XMM-Newton) also launched in 1999 by the European Space Astronomy satellite (ESA), carries an optical-ultraviolet telescope in addition to the three parallel mounted X-ray telescopes.
- NuSTAR (launched in 2012, Figure I.3) is dedicated to observing hard X-rays. NuSTAR is seeking out BHs and other collapsed stars in our galaxy, mapping material in young supernova remnants, and studying relativistic jets in active galactic nuclei.

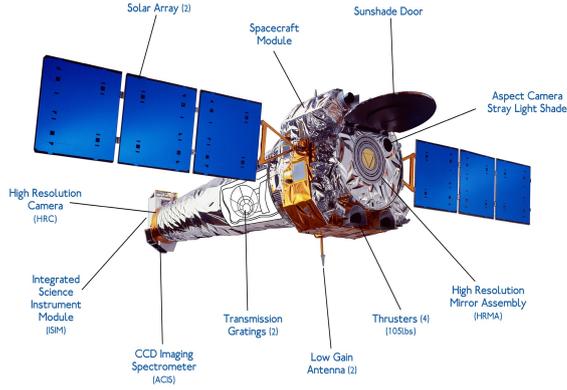


Figure I.2: Chandra X-ray Observatory (Credit: NASA).

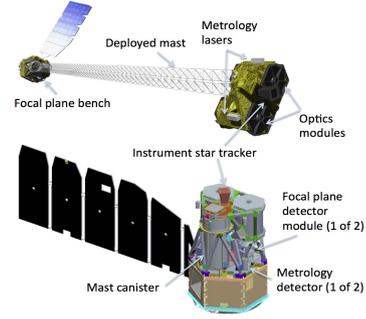


Figure I.3: NuSTAR X-ray Observatory (Credit: NASA).

	Lifetime	Energy Range (KeV)	Scientific Highlights
Uhuru	1970-1973	2-20	First Comprehensive and Uniform Sky Survey
ROSAT	1990-1999	0.1-2.5	X-ray All Sky Survey Catalogue
RXTE	1995-2012	2-250	X-ray Observed by Blackholes and Neutron Stars
Chandra	1999-Present	0.2-10	Observed the Region Around Supermassive Blackholes
NuSTAR	2012-Present	3-78	Deep Observation of The Extragalactic Sky

Table I.1: The X-ray satellites, their missions and the scientific highlights.

The X-ray satellites' discoveries and their scientific highlights are summarized in Table I.1.

I.2 Common Types of X-ray Detectors

X-ray detectors usually detect each individual photon and measure photons' properties such as energy, time of arrival, position and (sometimes) polarization. The first X-ray electronic detectors for detecting X-rays were gaseous detectors. Proportional counters are a type of gaseous detectors with a signal output that is

proportional to the photon energy of the incidence radiation. Imaging gas proportional counters were employed on the X-ray satellites Einstein and ROSAT. They had a moderate energy resolution with an acceptable background rejection efficiency.

Silicon-based (or Solid State) detectors were the next generation of X-ray detectors. Large Charge Coupled Device detectors (CCDs) are currently used whenever a high spatial and spectral resolution required. In fact, CCDs are an array of linked coupled capacitors. In these detectors, photoelectric interaction of arrived X-rays with silicon atoms generate electron-hole pairs (only 3.7 eV is needed to create an electron-hole pair), then an applied electric field collects and stores electrons as pixels which are coupled and can transfer their stored charge to the neighboring pixels. Eventually, stored charged transfer to an output amplifier. Semiconductor detectors have some intrinsic advantages compared to gaseous detectors. Higher density of detector material ($\rho_{Silicon}/\rho_{Gas} \approx 100$) yields higher quantum efficiency and lower required energy for generating an electron hole pair yields better energy resolution. However, gaseous detectors have an excellent background noise rejection capability.

I.3 Astrophysical Mechanism for Producing X-rays

X-rays are produced when an object is heated to very high energies (about a million degrees). By studying these X-rays, scientist can obtain valuable information and answer fundamental questions about the origin, evolution and density of the universe.

The principle method of measurement in X-ray astronomy is to detect individual photons in order to determine the complete set of four properties: direction on the sky, energy of the photon, the arrival time of the photon, and its polarization angle. These features allow us to expand our knowledge of the universe.

X-rays can be generated through three main physical processes which are:

- **Bremsstrahlung:** Thermal radiation produced by charged particles as they

are accelerated through Coulomb interaction with other charged particles (Figure I.4). This type of radiation can be seen in a very hot ionized plasma such as X-ray binaries with a temperature of about 10^7 K and the hot intergalactic medium in cluster of galaxies at very high temperature (10^8 K).

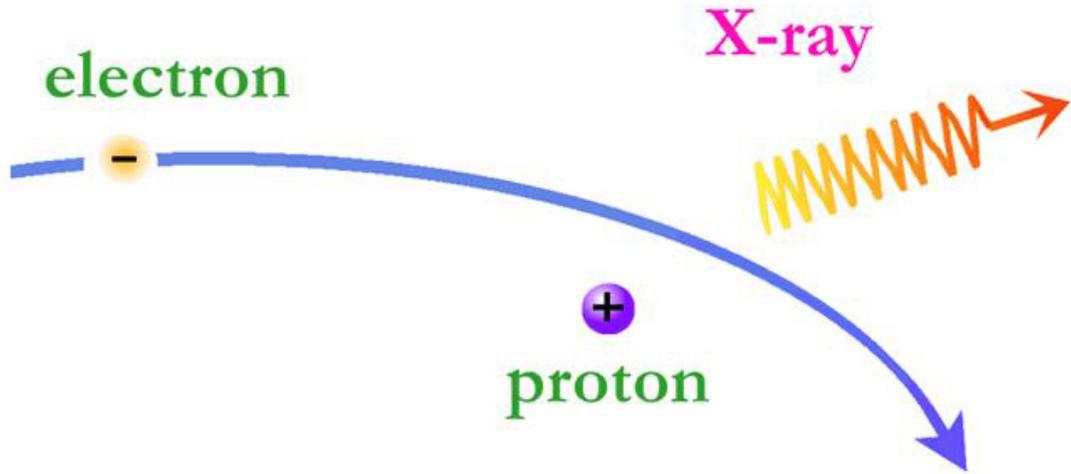


Figure I.4: Accelerated Charges and Bremsstrahlung Process. (Credit: Chandra X-ray Observatory)

- **Synchrotron radiation from relativistic electrons:** Synchrotron radiation is a non-thermal emission from relativistic particles in the presence of magnetic field (or magnetic bremsstrahlung). The production of Synchrotron radiation could be a strong indicator of the presence of violent particle acceleration in a huge magnetic field. This phenomenon can be found within the Galaxy, in the supernova remnants, external active galactic nuclei (AGN), pulsars, other radio sources, and γ -ray bursts with a power-law spectrum ($I(E) = AE^{-\alpha}$). It should be noted that the electrons which produce Synchrotron X-ray must have energies about 10^4 GeV, so the presence of Synchrotron X-ray confirms the existence of the super energetic electrons.
- **Inverse Compton Scattering:** Compton scattering of longer wavelength electromagnetic radiation by high-energy electrons that results in higher-

energy photons, commonly called Inverse Compton scattering is another primary X-ray production mechanism (Gorenstein, 2010). When the electron has significant kinetic energy as compared to the energy of the photon, the electron's energy can be transferred to the photon, resulting in the scattered photon having a higher frequency.

- **Blackbody radiation from star-like objects:** The thermal EM radiation comes from an object in thermal equilibrium. Such emission has specific spectrum and intensity that depends only on the object's temperature following Planck's law.

Stars radiate as blackbodies in a wide variety of temperatures, ranging from 2500 K, such as red dwarfs, to 40,000 K like O type stars. For instance, the surface of a NS can have a temperature of few million degrees, with the peak of the blackbody radiation observed in the X-ray range, which can be determined by:

$$\lambda_{max}T = 2.898 \text{ [mm K]} \tag{I.1}$$

where T is the temperature in K and λ_{max} is the peak wavelength in mm.

Each of above processes has its own spectral signature that determines the characteristics of the X-ray sources (Figure I.5; Seward and Charles 2010).

I.4 Astrophysical Sources of X-ray Radiation

A brief list of objects that emit X-rays is provided below.

- **Active Galactic Nuclei (AGN):** AGNs are among the most energetic sources in the universe. The galactic center, a small region compared to the entire galaxy, emits tremendous amount of energy in the form of radio, optical, X-ray, and gamma radiation as well as high-speed particle jets. There are

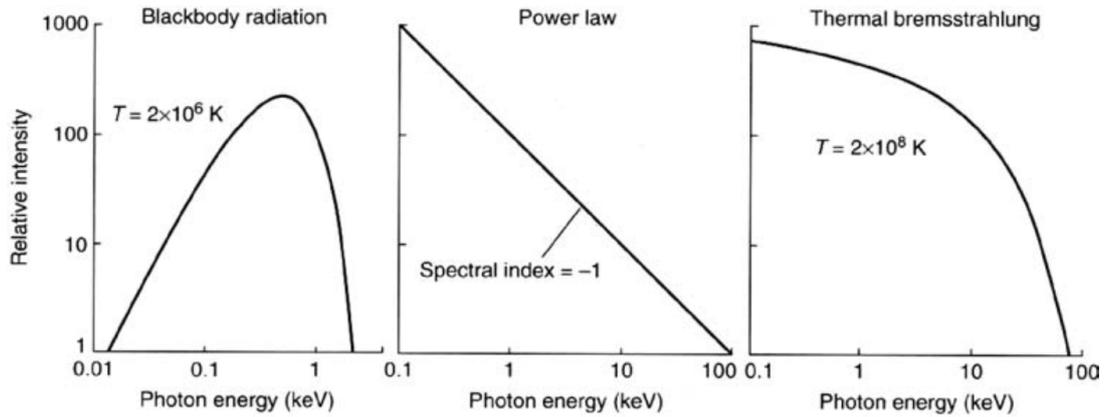


Figure I.5: Spectral forms expected from different astrophysical processes. (Adapted from Seward and Charles 2010)

different kind of active galaxies, such as Quasars, Seyfert galaxies, and Radio galaxies, but for simplicity we will refer to all of them as AGN. Although the AGN emission is spread widely across the electromagnetic spectrum, often peaking in the UV, there is significant luminosity in the X-ray bands. The X-ray radiation from AGN is due to accretion of matter by a supermassive BH at the center of the host galaxies. Therefore, the characteristics of an AGN depend on different factors such as the mass of the central BH, the rate of gas accretion onto the BH, the amount of obscuration of the nucleus by dust, and the AGN geometric orientation.

- **Pulsars and NSs:** NSs are rapidly rotating (as short as 1.6 milliseconds) compact objects with a radius of ~ 10 km. They have incredibly large ($\sim 10^{12}$ Gauss) magnetic fields with significant portion of the NS radiation emitted in a beam along the magnetic axis. If the beam sweeps across the line-of-sight of the observer, a pulsed emission is seen and such NS are also referred to as pulsars. A continuous X-ray emission can also be detected because of thermal and/or magnetospheric emission in NSs. The first X-ray pulsar to be discovered was Centaurus X-3, in 1971 with the Uhuru X-ray satellite.

- **X-ray Binaries:** X-ray binaries are among the most luminous galactic X-ray sources. They consist of normal star or a white dwarf (“donor” star) and a collapsed one, such as a NS or a BH. X-rays are released as material from the donor is accreted onto the collapsed companion (also called compact object). As the material spirals inward, it heats up (due to friction) to the very high temperatures up to 10^8 K. As a result, the gravitational potential energy of infalling matter is released in the form of X-ray radiation. X-ray binaries are classified into two major subclasses:

- **High Mass X-ray Binaries (HMXB):** In this group of X-ray binaries, the mass-losing star has a mass $> 10 M_{\odot}$ and they are optically luminous ($L_{X-ray}/L_{Optical} < 1$). The compact object in HMXBs most of the time are NS or BH. Due to the strong stellar wind of the donor star, the compact object can accumulate large amount of material throughout its orbit resulting in X-ray emission. The donor star, due to its large mass, has short life-span (< 100 Myr), and, therefore, HMXBs are associated with young stellar populations.
- **Low Mass X-ray Binaries (LMXB):** In this group of X-ray binaries, the mass-losing star has a mass of about $1 M_{\odot}$ and they do not have very strong stellar winds. Therefore, mass transfer can occur only after the donor star enter the last stages of its evolution and expands into a red giant. This takes long time (in the order of giga years) and LMXBs are associated with the older stellar population.

Both HMXB and LMXB both can exhibit variability in X-rays. The main phenomena that cause changes in the X-ray flux are (1) Eclipse: it happens when the compact object is hidden behind the normal star once in every orbital period (the binary orbit is seen edge on), and X-rays from the compact star are obscured; (2) X-ray dips: dips can be explained by absorption of the X-ray source by material passing in front of the source (White et al. 1995), or accretion disk instabilities, which cause the ejection of the inner part of

the disk (Greiner et al. 1996); and/or (3) pulsation: if the compact object in an X-ray binary is a pulsar, where the X-ray emitting hot spots move in and out of the line-of-sight as the NS spins, giving rise to regular X-ray pulses.

- **Cataclysmic Variables (CVs):** CVs are a type of binary system which consist of a normal star as a companion, and a White Dwarf (WD) as the compact object. Such systems have typical orbital periods between 75 min and 8 h (Ritter and Kolb, 2003). Gravitational potential energy is converted into X-rays during the accretion process similar to other XRBs, however, the thermal emission from associated with the compact object peaks at lower energies (UV) compared to other types of XRBs.

I.5 Measurable Properties of Astrophysical Sources

Observed flux is a measurement of the brightness of the source, usually given in units of photons $\text{cm}^{-2} \text{s}^{-1}$, or $\text{erg cm}^{-2} \text{s}^{-1}$. What is measured by the detector is the count rate, C , of the source, which is a product of the flux, F , detector efficiency, η , and detector area, A , integrated over the energy range of the detector:

$$C = A \int F(E)\eta(E)dE \quad (\text{I.2})$$

Luminosity, L , the intrinsic power of a source, is the total amount of energy emitted by a star each second and measured in units of erg s^{-1} . It relates to the flux through the square of the distance to the source, d :

$$F = \frac{L}{4\pi d^2} \quad (\text{I.3})$$

For example, the Sun's persistent X-ray luminosity, L_X , can reach $5 \times 10^{27} \text{ erg s}^{-1}$. In our Milky Way Galaxy, the brightest individual X-ray emitters, with persistent

LX of up to $\sim 10^{39} \text{ erg s}^{-1}$ (100 billion times greater than that of the Sun), are the XRBs, while in the range of 0.2–10 KeV, the most luminous X-ray source known is the quasar (a type of galaxy) PKS 2126-158 with an X-ray luminosity of $5 \times 10^{47} \text{ erg s}^{-1}$.

I.6 Apparent Magnitude

Historically, source brightness has been measured by astronomers in magnitudes. The apparent magnitude (m) of an astronomical object is logarithmic and a difference of 1 in magnitude corresponds to a change in brightness by a factor of 2.512. The dimmer an object appears, the higher the numerical value given to its apparent magnitude, which is given by:

$$m_x = -2.5 \log \frac{F_x}{F_{x,0}} \quad (\text{I.4})$$

where F_x is the observed flux using spectral filter x , and $F_{x,0}$ is the reference flux (zero-point) for that photometric filter.

The flux density (flux per wavelength) of any object varies across its spectrum. To fully understand the nature of a source, we need to gather its multi-wavelength (MW) information (sample the flux density at a large number of wavelengths). To achieve this, we often need to observe an object with several different telescopes and filters (X-ray, UV, optical, IR). The peak emission is related to the effective temperature of a thermal source (Wien's displacement law). For most of stars the wavelength of the peak flux is in or near the visible region of the spectrum. It is not true for other some objects such as AGNs, which emit significant energy over a wide range of wavelengths.

I.7 Photometric Filters

The process of measuring the brightness (or flux) of an object is called photometry. For example to measure color, the star's light is collected by a telescope through various color filters. Then, the filtered light will be collected by a CCD to get the star's image with each of the filters in the set. By utilizing photometry in different filters, which have well-defined pass-bands, we can gain information about the objects' temperature, distance, and their ages.

A number of astronomical photometric/filter systems have been introduced in the past with a variety of detectors and passbands. Here are some of the more commonly used filter systems:

- **UBV** : One of the first filter systems, and still commonly used, is the UBV, which was introduced in the early 1950s by Harold L. Johnson and William Wilson Morgan. The UBV system uses three pass-bands: one in the near-ultraviolet, one in the blue, and the other in the dominant visual range (Figure I.6). By measuring the apparent magnitudes in these filters, one can determine the surface temperature of a star. It is well known that a clear correlation exists between the star's temperature and the magnitude difference, or flux ratio, in the B and V filters (Figure I.7). If a star is very hot, its radiation peak is in the short, ultraviolet wavelength. In contrast, stars with lower temperature peak at longer wavelength.
- **The Sloan Digital Sky Survey System**: Each filter in the Sloan Digital Sky Survey (SDSS) system is designed to pass light around a specific wavelength, somewhat differently than the UBV system. Table I.2 shows the mid-point wavelengths at which each of the five SDSS filters. Sampling wider EM range allows astronomers to compile a crude spectral profile of the astronomical sources.

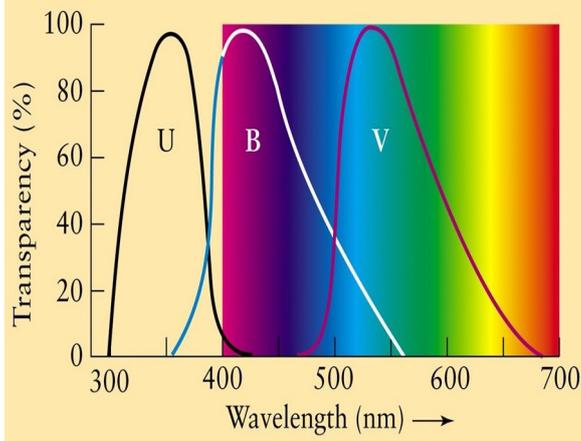


Figure I.6: U, B, and V filters: the wavelength to which the standard filters are transparent. (Credit: Freedman et al. 2011)

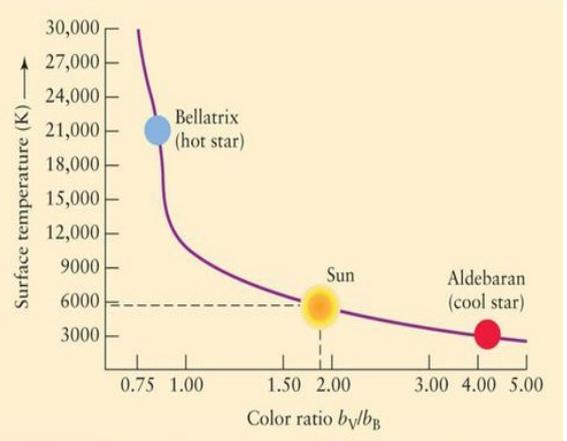


Figure I.7: This flux (or color) ratio is small for hot, blue stars but large for cool, red stars. (Credit: Freedman et al. 2011)

SDSS's	Wavelength (Angstrom)
ultraviolet (u)	3542
green (g)	4770
red (r)	6231
near infrared (i)	7625
infrared (z)	9134

Table I.2: The Sloan Digital Sky "ugriz" System.

II. Machine Learning Methods

In the current era of big data, modern science requires an efficient way to enhance the computation accuracy, speed, and efficiency in a broad range of areas such as Quantum Computation, signal and image processing, Machine Vision, DNA sequencing, and, recently, in astrophysics. The field of machine learning (ML) is multi-disciplinary, closely related to statistics, numerical methods and data analysis. ML is one of the most broad and applicable fields of Artificial Intelligence (AI), in which computers learn to think and make decisions like humans. In order to achieve this goal, the computer uses data (provided by the user) to learn and improve its decision making algorithm. To create a robust ML method, one must have large enough data for the computer to learn from. This will also ensure higher result accuracy.

II.1 Types of Machine Learning Methods

There are four main types of ML models, which are used for the classification problems.

- **Supervised Learning:** In supervised ML, machines use labeled (already classified) training set of data, based on which a classification algorithm is generated. The training data set consists of object with already assigned label (or class) and a number of parameters (features) associated with a given object. The supervised approach is comparable to the human learning process in which students learn new things under the supervision of instructor and through good examples. As a result, students derive general rules from the given examples. The most popular supervised ML algorithms are: Random Forest (used in this work), Decision Tree, K-Nearest Neighbors (KNN), Linear Regression, Logistic Regression, Support Vector Machines, and Neural Networks.

- **Classification Task:** Classification is a powerful method where we classify data into a given number of classes. The main objective of a classification problem is to distinguish the class to which a new data will belong and it is used for the prediction of discrete responses. Classes are often called as targets or labels.
- **Regression Task:** Regression models a target prediction value based on independent variables and it is basically useful for discovering the connection between variables and forecasting. A regression task can be applied if the output variable is a real or continuous value.
- **Unsupervised Learning:** In unsupervised learning, the the data set is not labeled (i.e., objects are not classified), and machines are tasked to find classes without prior knowledge of the number of existing classes. The goal in such unsupervised learning problems is to find groups of similar instances within the data, where it is called clustering, or to realize how the data is distributed in the space, known as density estimation.
- **Semi-supervised Learning:** In many machine learning problems the domains data is cheap (or easily accessible), but labelled data is somewhat expensive. In these cases, semi-supervised machine learning could be a wise approach that is to apply a small labeled training set to build a cost effective and initial model, which is then enriched using the unlabeled data to increase the size of training dataset. Another example of semi-supervised machine learning is image recognition in Astronomy so that algorithm automatically recognizes that the same object A shows up in a group of photos (say 1, 3, 5) while another object B shows up in another group of photos (say 2, 4, 6). This is the unsupervised part of the algorithm (e.g., Clustering). The next step is for the user to provide information about the nature of the objects.
- **Reinforcement Learning:** This method allows machines the most freedom

to employ trial and error to discover the actions that yield the greatest rewards for the system. Errors help our system to learn better because they have some negative consequences such as loss of time, and cost. The algorithm reports the answer when it is right, but it does not provide any feedback of finding the right and more efficient answer.

In this thesis, we will focus on Random Forest decision tree algorithm, which is a Supervised Learning method.

II.2 The Machine Learning Process

All ML algorithms (see previous Section) respond to the same logic tasks. The idea behind an ML algorithm is using a mathematical formulation (e.g., function) that algorithms do not know at the beginning but will deduce them via some input data over time as training examples. Each learning algorithm follows the same specific steps toward achieving the right prediction. The most common steps include: Data Collection, Data Preparation and Preprocessing, Algorithm selection and Model Training, Model Evaluation, and Performance improvement.

The most important part of any efficient ML algorithm is data collection. In Astrophysics, we use catalogues of literature verified sources from an archival database (discussed in Chapter 3). Both the quality and the quantity affect the model performance and accuracy. It is very difficult to determine the amount of data that might be relevant. One also has to make sure that the data set is clean and there are as few as possible missing values. Therefore, finding a high-quality data with an appropriate size will significantly increase the model efficiency and model prediction.

Preprocessing of the data refers to any required task that should be done on the data set to make it clean, unbiased, weighted, and normalized. Moreover, the data should be balanced so that there is about the same amount of data in each class during the training and testing process. The last step involves splitting data into the following three different sets:

- **Training set:** Training set are a set of data that a system uses train from to create the model.
- **Validation set:** A validation dataset is a dataset of examples used to tune the model parameters of the classifier. Generally, the validation data set is used to compare the performance of the potential algorithms which have been suggested by training set and then decide which model has been trained more properly.
- **Test set:** Testing set is used to evaluate the performance of our model. Overall, test set is set of examples which are used to evaluate the performance of a fully-trained classifier in order to estimate the error rate after we have chosen the final model. The main difference between the test set and validation set is that the validation is used for the model evaluation during tuning the parameters and data preparation while the test provides a standard for the model evaluation and it can only used once a model is completely trained by the training and validation sets.

There are many ML models that one can may choose. A good model should be simple, accurate, and applicable to the big data sets. Model Evaluation is an essential part of the ML process because every system needs to be evaluated for the accuracy and precision on data that it was not trained on. Selection of appropriate accuracy score is a convenient and reliable technique that we can use to judge our model precision. Different classes of evaluation metrics are available, such as classification accuracy, F-1 Score, mean squared errors, mean Absolute Error, and Confusion Matrix (CM). In this thesis we applied the CM to find out whether the performance of our model is good.

The CM demonstrate how comprehensive is the performance of the model. It consists of a square matrix containing both horizontal (Actual Values) and vertical (Predicted Values) directions with a list of all classes. To calculate the accuracy of CM one can use the following equation:

$$Accuracy = \frac{TP + FN}{Total\ Number\ of\ Samples} \quad (II.1)$$

where TP and FN are true positives and false negatives, respectively.

II.3 Features

A feature defines a relevant object in our dataset that sits between raw data and models in the machine learning pipeline (Zheng and Casari, 2018). It is a specific property or characteristic of an object or phenomenon which has been observed. Choosing the proper features of a dataset may significantly affect the classification accuracy of the ML pipeline.

II.4 Cross Validation

A validation set can help us to improve and fine-tune the ML procedure. The cross validation is a great way to evaluate the model. Cross validation based on k-fold relies on random splitting, but it divides a given data into k-folds. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set and eventually the average error across all k trials is computed (Figure II.1). One important benefit of k-fold cross validation is that it matters less how the data gets divided. The drawback of this technique is that the training algorithm must be rerun from scratch k times, which means the execution time get k times longer. A solution for this problem could be randomly splitting the data into a test and training set k different times.

II.5 Overfitting and Underfitting in Machine Learning

Overfitting happens when an ML model is closely fit to a limited dataset. Overfitting negatively impacts the performance of the model on new data and any noise in our data will be interpreted as a significant deviation from the model.

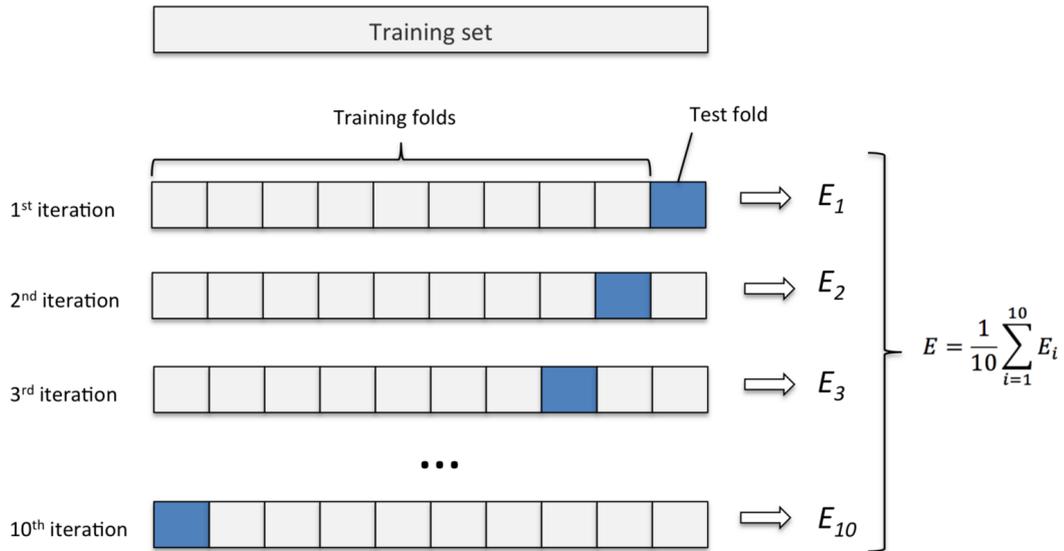


Figure II.1: K-fold cross-validation (Credit: Researchgate)

This makes the model more complicated so that it cannot easily be generalized. Because of the algorithm flexibility of nonlinear and nonparametric models (such as a decision tree), they are subjected to the overfitting more than other type of models.

An underfit ML algorithm is not an appropriate model and displays a very low qualified performance of the model. The solution for the underfitting problem is to apply alternate machine learning algorithms. In contrast to overfitting problem, an underfitted ML model is less flexible and cannot account for the data. Figure II.2 shows examples of overfitting and undefitting. An optimum hypothesis in the Figure II.2 is the trade off between variance and bias.

II.6 Feature Importance

Feature importance helps summarize how important each feature is for model. It represents each feature by a number between 0 and 1 (feature importance always sums to one), where 1 means “perfectly predicts the target” and 0 means “not used at all”.

It should be considered that if a feature has a low feature importance value, this cannot be concluded that it is useless and uninformative. It could mean that the

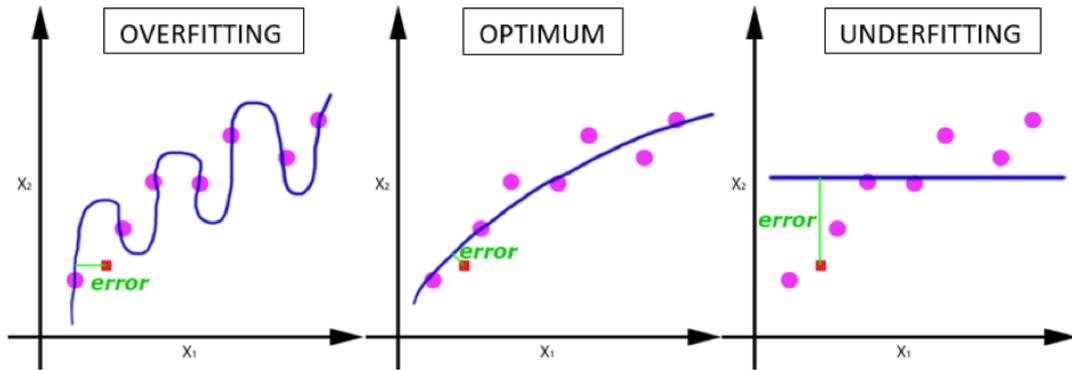


Figure II.2: Examples of overfitting and underfitting in ML (Credit: Machinelearningmedium.com)

feature was not picked by the model, likely because another feature provides the same information. In fact, the feature importance tell us how much a particular feature is important, but not whether that feature is indicative of a sample being benign or malignant (Muller and Guido, 2016).

II.7 Decision Tree

A decision tree is a supervised predictive model. It makes predictions based on a series of questions related to the features. The outcome of each particular question defines which branch of the tree to follow. It can learn to forecast outputs by answering questions based on the values of the received inputs. Root node, leaf node, internal node, and branches form the basis of tree-based algorithms that facilitate the rules' recognition to classify or predict events and variables (discrete or continuous). Moreover, unlike linear or logistic regression, decision trees optimized for both regression and classification problems. An example of decision tree is shown in Figure II.3.

II.8 Ensemble Learning and Random Forest

Ensemble learning improves ML results by combining several models (algorithms), which yield better predictive performances that could be obtained by each ML algorithm alone. Ensemble learning method decreases bias and variances. Random

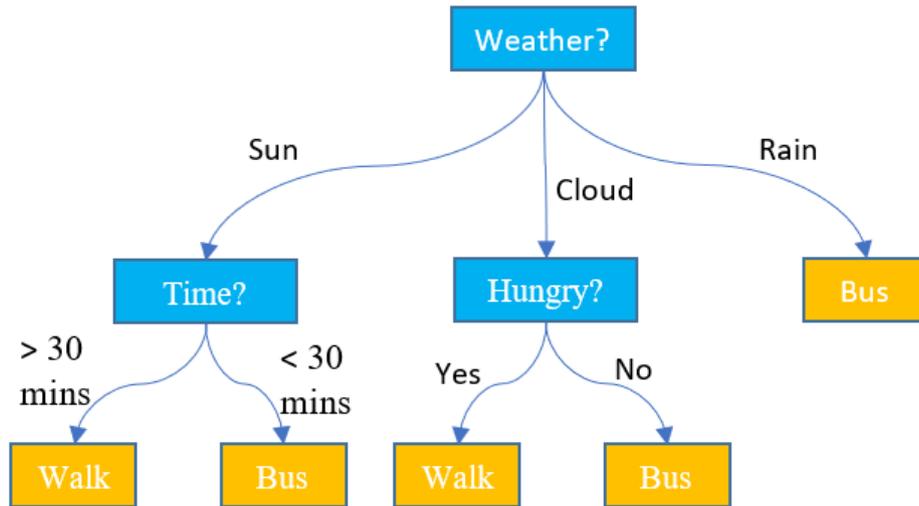


Figure II.3: The decision tree is typically read from top (root) to bottom (leaves). A question is asked at each node (split point) and the response to that question determines which branch is followed next. The prediction is given by the label of a leaf. (Credit: DisplayR)

Forest (RF) generate a family of techniques that consist in building an ensemble (or forest) of decision trees grown from a randomized variant of the tree induction algorithm. They are an example of an ensemble learner built on decision trees, meaning it relies on aggregating the results of an ensemble of simpler estimators and eventually select the prediction result with the most votes as the final prediction (Figure II.4). In this thesis we will use RF classifier for our analysis (see next chapter for specific details).

A main drawback of decision trees is overfitting the training data. Therefore, the RF classifier is a great way to alleviate overfitting. It makes a collection of decision trees, where each tree is a bit different from the others, and each tree will likely overfit one part of the data. But by creating many trees the amount of overfitting will decrease (because of averaging all results). To accomplish this, each tree should do an acceptable task of predicting the target and should also be different from the other trees.

Similarly to the decision tree, the RF provides feature importance, which are

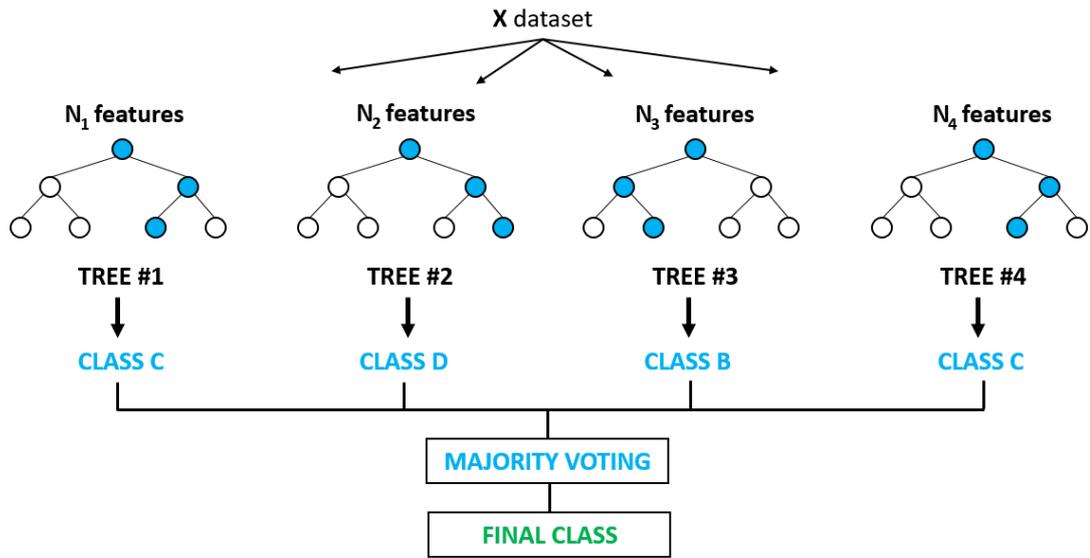


Figure II.4: Random Forest Procedure (Credit: Medium.com).

computed by adding the feature importance over the trees in the forest. Therefore, one could say that the feature importance created by the RF classifier are more reliable than the ones provided by each tree alone.

II.8.1 ML Python Libraries

The ML pipeline utilized the existing Python library Scikit-learn, which is a very popular tool, and probably the most prominent Python library for ML. It contains a number of state-of-the-art ML algorithms, as well as comprehensive documentation about the algorithms. Scikit-learn works very well with almost all scientific Python tools. The RF Decision Tree that we are using is a part of Supervised Machine Learning algorithms, which is accessible via Scikit-learn library.

To manage the project, we utilized several other Python libraries which contain mathematical and statistical tools for the implementation of our machine learning model. Each of these essential libraries play an important role in each portion of our analysis. Some of these libraries are NumPy, SciPy, Pandas, PySynphot, and Matplotlib.

III. Classification Results

Currently, only a small fraction of extragalactic X-ray sources have reliable classifications. The X-ray data alone are not enough to reveal the nature of the X-ray source. Therefore, creating an automated ML tool for classification of extragalactic X-ray sources with MW data is essential to understand X-ray source populations in a plethora of nearby galaxies. Here we describe the specifics of the ML model we used to accomplish this goal.

It is important to note that we can only use the HST to study the optical/NIR counterparts of extragalactic X-ray sources (these sources are too far from us and ground based observatories lack the resolving power and sensitivity to detect them). At the same time, there are not enough verified X-ray sources observed with HST to build a training dataset from the HST catalog. Therefore, we used training dataset created based on galactic X-ray sources detected with ground based instruments and transformed the dataset to HST parameter space.

This chapter explores our ML results, which have been obtained with the RF classifier. As we have discovered so far, there are different approaches for analyzing a big set of data, however it is crucial to know the exact nature of our data before diving into data exploration and manipulation. Our goal is to create a ML classifier capable of identifying X-ray sources using astrophysical measurements from different observatories probing large range in the EM spectrum.

We split our data-set into a training set and a test set, where training set assists us to build our model, and the test set applies for the evaluation process in order to see how well our model will generalize to a new-unseen data. Finally, we checked the accuracy of our model using Confusion Matrix and Score method to see how good our model (RF decision tree classification algorithm) fit our test set.

	AGN	HMXB	LMXB	CV	STAR
Number	3856	13	20	105	536

Table III.1: Training dataset of 5 different object classes for the RF classification.

III.1 Training Data Set

The pipeline makes use of the RF decision tree classifier and relies on a large training dataset of literature verified sources. The training data set we used has $\sim 6,000$ sources with the following object classes: (1) stars (Samus et al., 2009), (2) AGN (Véron-Cetty and Véron, 2010), (3) LMXBs (Liu et al., 2007), (4) HMXBs (Liu et al., 2006), and (5) CVs (Downes et al., 2001). The source breakdown is illustrated in Table III.1. All sources in the dataset (with exception of AGN) are Galactic sources.

Since we are interested in X-ray sources, we require all X-ray sources in the training dataset to be detected in X-ray with XMM-Newton. Also, since we have to make a transition to HST filter system, we also require all sources to have a Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) counterpart. The resulting training dataset uses the following MW parameters: X-ray fluxes in two bands (0.5–2 and 2–7 keV) and two hardness ratios from the XMM-Newton Observatory, fluxes from the Pan-STARRS catalog in five bands (g , r , i , z , and y), NIR fluxes from 2MASS in three bands (J , H , and K), and IR fluxes in three bands (W1, W2, and W3). We then constructed various combinations of flux ratios.

We set 25% of the data aside as a validation set, to be used for checking the accuracy of the algorithm.

III.2 Oversampling

Due to observational biases, we have more AGN in the catalog than any of the other classes, with LMXBs and HMXBs being the least represented. This makes the

training data unbalanced, which can reduce the performance of our classification ML algorithm. To avoid this, we applied the SMOTE (Synthetic Minority Over-sampling Technique) method (part of Scikit-learn) to introduce synthetic sources with similar properties to the class that is being underrepresented. This fixes the weight issue with the RF algorithm. By applying oversampling methods all categories have equal amount of records simply by oversampling instances of the minority class. Fitting SMOTE is a great way to fix imbalanced our data sets and makes the same amount of sources as the number of AGN sources are (here 4478 sources).

III.3 Missing Data

There are many reasons that could result in missing values in the dataset: the source being too faint to be detected, limited observing coverage of the observatory, low signal-to-noise, etc. Missing values are commonly handled via imputation, where the blank spaces or NAN values can be filled in with inferred values, such as mean or median. While appropriate for some domains, this approach is not well suited to astronomical data sets, because a missing value may well be physically meaningful. Therefore, missing values cannot be eliminated easily from our data set because they may contain important information that we would get back to them for further analysis.

III.4 Normalization and Standardization

The distance to all sources in the training dataset is not always known, which introduces bias and can confuse the algorithm because we cannot determine the intrinsic power of the source (just measure its flux). For example, if two identical light bulbs are placed at different distances from the observer, the closer light bulb will appear brighter than the other, even though they are exactly the same (e.g., same class). To account for this, we normalize all fluxes (for a given object) by the total X-ray flux.

We then normalize each feature (e.g., parameter) so that it will have the properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$, where μ and σ are the mean (average) and the standard deviation respectively. The Z-score normalization can be written as:

$$z = \frac{x - \mu}{\sigma} \tag{III.1}$$

III.5 Feature Selection

Feature selection is a proper method for removing irrelevant and non-related information that must be applied before the implementation of algorithm and can be expected to result an improvement in learning accuracy and comprehensiveness. While there are numerous attributes that can play a role in the analysis of astronomical data, only a fraction of them are helpful in the model making and statistical predictions.

To determine which features should be eliminated, all features are ranked according to their predictive power and those with very low importance are removed. The scores can be measured by t-statistics and F-statistics, which are two well-known examples of features selection method. Here we apply F-statistics in evaluating the model accuracy.

By making different flux ratios, we can create a large number of new features. While selected few have been shown historically to provide useful information (in terms of separating classes of sources), many of these features are unlikely to provide new, useful insight. To determine which features contribute the most, we have analyzed the importance of different features. We can improve the model by focusing on the most important features and removing those that are insignificant and have not been made a serious impact on our prediction. Moreover, shorter training time will be obtained, which can be very important when we are dealing with high dimensional datasets. We implemented the feature importance method

using Scikit learn tools, and we have optimized the classification pipeline by analyzing the feature importance to select the most optimal parameters. An example of how individual features are evaluated is shown in Figure III.1.

The features include normalized (to the total X-ray) fluxes and flux ratios (also referred to as optical colors for Hubble Space Telescope (HST) filters, e.g., F475W-F110W), and X-ray hardness ratios (HR2 and HR4). In the example shown in Figure III.1, four features, HR4-norm, HR2-norm, FX27-norm, and FX052-norm) are the most informative while three (FU_norm, FG_norm, and FZ_norm) are very insignificant and can be removed from consideration without lowering the model’s predictive accuracy.

III.6 Hubble Space Telescope Filter Transformation

As mentioned earlier, we have to convert the training dataset (constructed from catalogs via ground based observatories) to HST parameter space. This is necessary because both Pan-STARRS and 2MASS (ground based observatories) use different filter system than HST, and, therefore, they each observe slightly different part of the EM spectrum (the difference in filter transmission is illustrated in Figure III.2). However, the conversion is very model dependant. Since we are transforming the training dataset, for which the sources are known, we can estimate the overall spectral profiles for each class. To calculate conversions between ground-based telescopes and HST, we use the Python package PySynphot. We use derived fluxed ratios between similar filters (e.g., J and F110W) to make the transformation. To model these spectral profiles, we use black body models for the following classes: HMXBs (15,000 K), CVs and LMXBs (3,000K), and stars (5,000 K). These models are shown on Figure III.2 (scaled to fit the Figure). For AGN we use a power law model with slope of -1.2 .

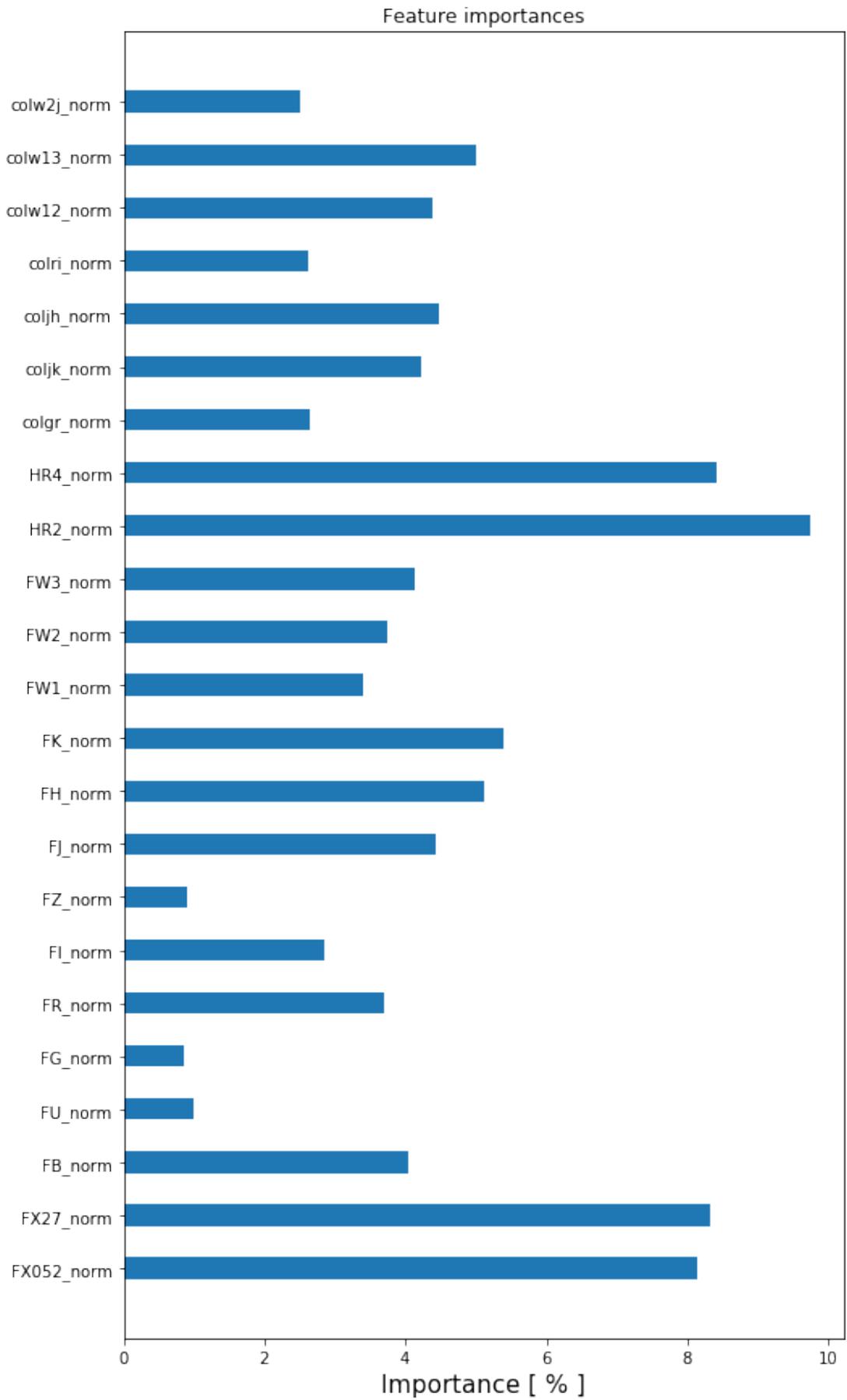


Figure III.1: Importance of features used by the classification ML algorithm.

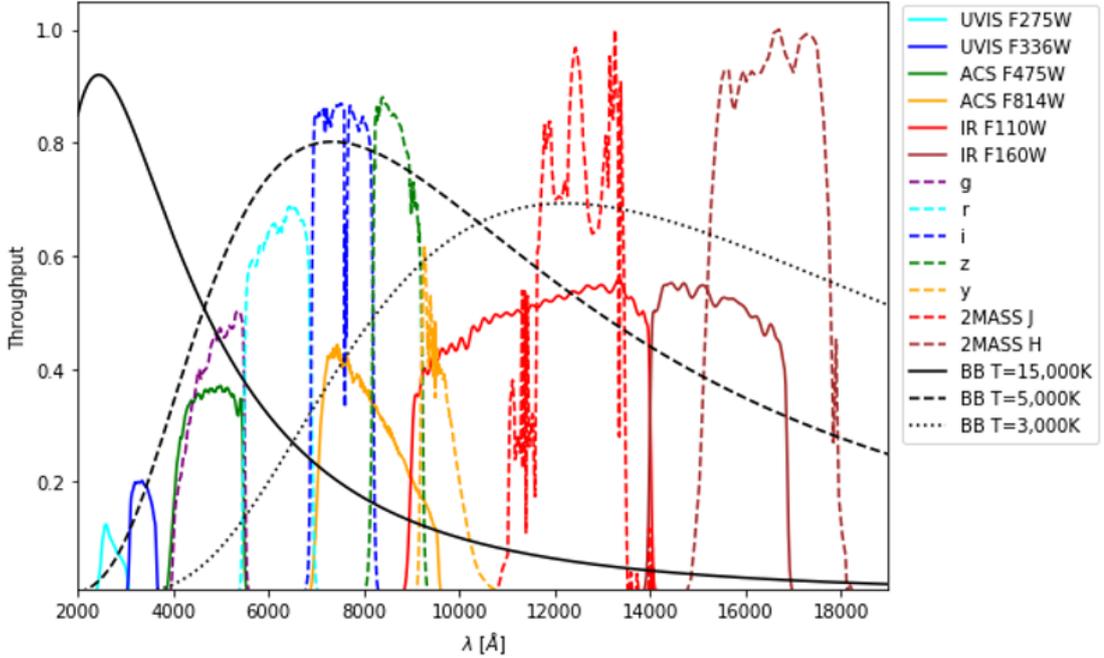


Figure III.2: Throughput HST and Pan-STARRS filters are shown with solid and dashed lines, respectively. Three black body models (for 3,000, 5,000, and 15,000 K, scaled to fit the figure) are shown for illustration purposes.

III.7 Classifier Fitting

For the classification part we used RF, which creates decision trees on randomly selected data samples, fits several decision tree classifiers on various sub-samples of the dataset and ultimately gets prediction from each tree and selects the best solution by averaging. Some of the RF classifier parameters control the size of the trees, such as `max_depth`, and `Class_weight` that may finally cause fully grown trees which results in very large trees on some data sets and increased over-fitting. After some testing, we conclude that leaving these parameters at default provide good combination of model accuracy and computational speed.

Two other parameters can be optimized: the number of trees (`n_estimator`) and the number of variables at each node (`min_sample_leaf`). We have found that the `n_estimator=1000` is the optimal number of trees to create an accurate RF model.

III.8 Verifying the Model

The 25% of the data that was initially set aside, the validation set, is now used to check the accuracy of the model. The metric module implements functions to estimate prediction error for specific purposes. Here, we used scoring parameters such as *accuracy_score* where this function computes subset accuracy in multilabel classification. The model accuracy can be expressed by *classification_report* function which shows the main classification metrics.

The F1 score can be interpreted as a weighted average of the precision and recall (F1 score is between 0 and 1). The recall represents the ability of the classifier to find all the positive samples and the support is the number of occurrences of each class in *target_test*. The scores corresponding to every class tells us the accuracy of the classifier in classifying the data points in that particular class compared to all other classes. The F1 score can be calculated as follow:

$$F1\ Score = 2 \cdot \left(\frac{precision \cdot recall}{precision + recall} \right) \quad (III.2)$$

III.9 Classification Performance

A confusion matrix is a table we used here to describe the performance and precision of our classifier on the validation (or training) data for which the true values are known. The diagonal elements of the confusion matrix indicate the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. Figure III.4 shows the confusion matrix for the classification training after applying the SMOTE procedure. Almost all 5 classes of object have been accurately labeled. The color bar beside the confusion matrix is an indicator of how many objects of each class (balanced class) are precisely labeled.

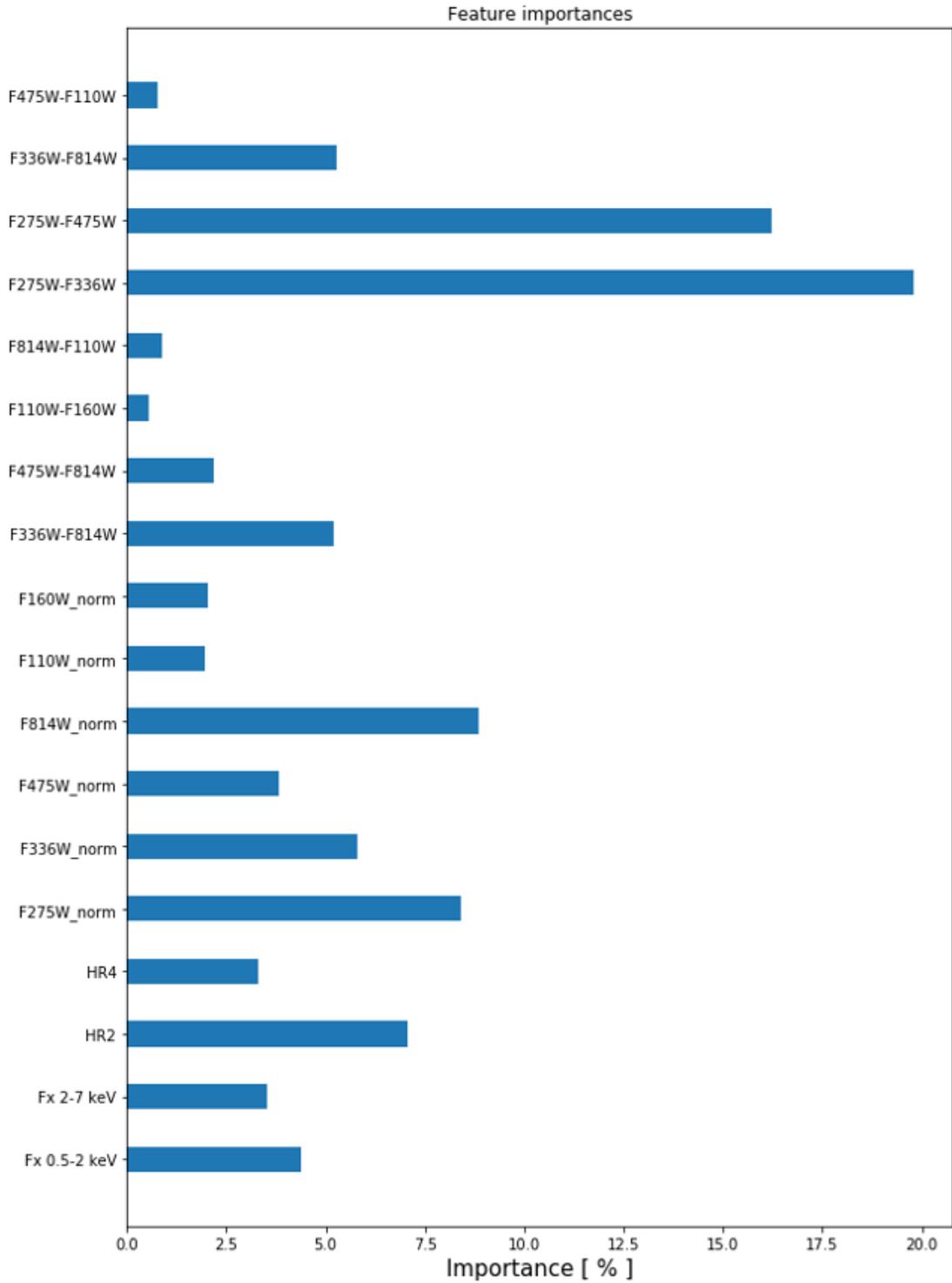


Figure III.3: Importance of Features for the HST Filter Transformation.

Similar confusion matrix, but for the validation set, is shown in Figure III.5. While the training dataset is balanced because of the SMOTE procedure, the validation set does not contain synthetic sources, and therefore is unbalanced (e.g., more AGN than other classes). Because of this, we have plotted the Figure III.5

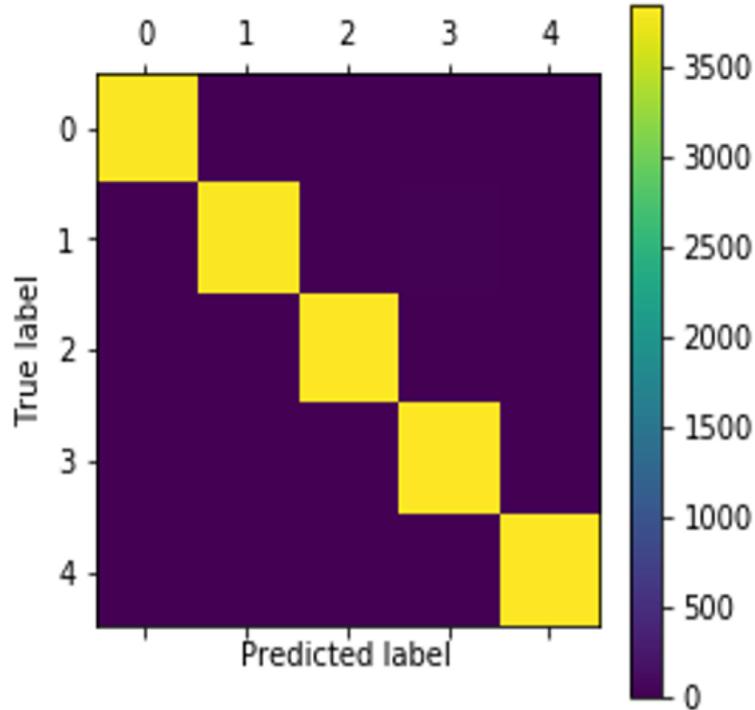


Figure III.4: Confusion matrix for the classification training dataset. The numerical labels correspond to AGN (0), CVs (1), HMXBs (2), LMXBs (3), and Stars (4).

in logarithmic scale to help with the visualization.

III.10 Classification Accuracy

We have presented here the preliminary results from our completed ML classification pipeline. Comparing the accuracy's results reveals that the accuracy of the training sets is 99.7% where the accuracy of validation test set is about 97.8%. This proximity of the predicted values shows that the model is accurately predicting the “new”, validation set. While further tests need to be made (using different extragalactic data), we can conclude that we have successfully constructed a ML method that can automatically classify X-ray sources using existing archival data.

Finally, the full code is available at https://www.dropbox.com/s/irxs4e32x1dnmld/classification_code.pdf?dl=0.

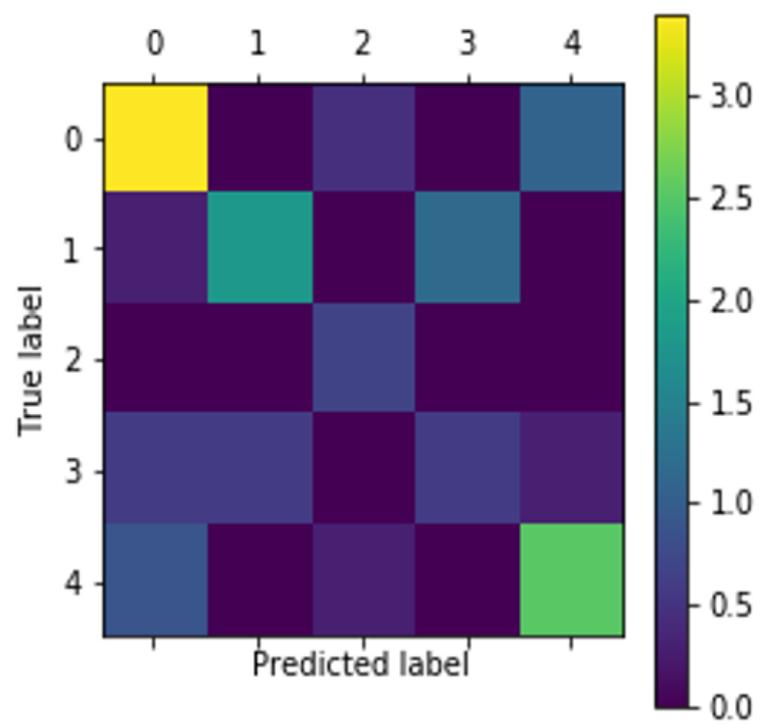


Figure III.5: Same as III.4 but for the validation training dataset shown in logarithmic scale for better visualization.

IV. Future Work

Over the last decades astrophysics has been in a unique period of its history with observatories viewing the sky over a multitude of wavelengths, which leads to a fast increase in the number of detected sources. We have seen that creating an automated ML tool for classification of extragalactic X-ray sources will enable future studies of X-rays sources in a large number of nearby galaxies. This will significantly increase the number of classified X-ray sources, enabling a statistical study of the X-ray source populations. This in turn will provide greater insight about galaxy evolution.

One does not need to look further than the two closest large galaxies, Andromeda (M31) and Triangulum (M33) to see that, even though roughly $\sim 2,000$ X-ray sources have been detected in M31 and M33, only $\sim 25\%$ have been successfully classified (see Figure IV.1). The majority of these X-ray sources lack astrophysical identification, and, thus we are unable to study the formation and evolution of the population of X-ray sources in these galaxies. The situation is not that dissimilar in other nearby galaxies.

Going forward, the ML classification tool that was developed here will be used to first explore the nature of X-ray sources in M31, where large amount of archival X-ray and HST data exists. This will further test and optimize the ML pipeline. Moreover, this will enable the classification of thousands of X-ray sources.

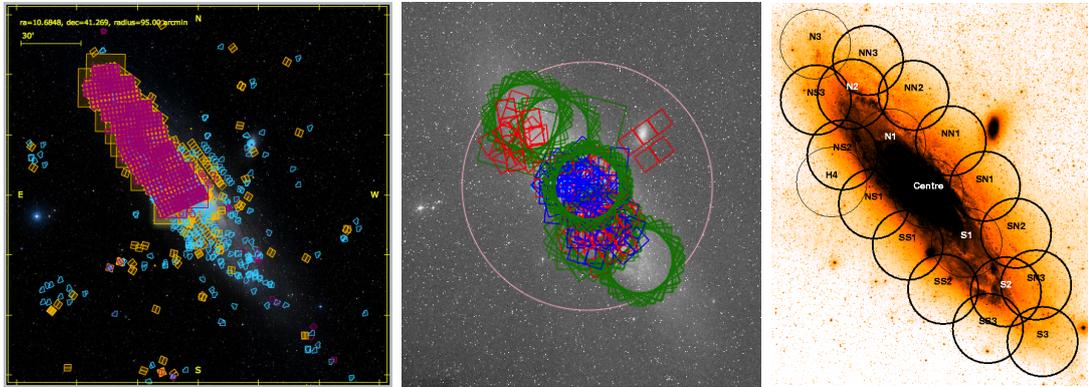


Figure IV.1: *Left*: Existing *HST* coverage of M31. The WFPC2 pointings are shown in blue, the ACS in orange, and the WFC3 in purple. *Center*: The Chandra coverage of M31. ACIS-I pointings are shown with blue, ACIS-S with red, HRC-I with green, and HRC-S with magenta. The circle has a radius of 1 degree. *Right*: A deep optical image of M31 overplotted with the XMM-Newton fields of the XMM survey. The area covered by individual EPIC observations is approximated by circles with $14'$ radius.

APPENDIX SECTION

APPENDIX A: Essential Libraries and Tools

A Python library is a reusable fragment of code that people may use to include in their projects for specific purposes such as data analyses, machine learning tasks, mathematical computation, and so on. The most important Python packages are listed below.

- **NumPy:** NumPy is one of the essential packages for the scientific computing in Python that contains sophisticated functions such as linear algebra operations, Fourier transform, high-performance multidimensional array object, and powerful N-dimensional array objects.
- **Pandas:** Pandas is another core library in the Python and the best Python tool for the data wrangling and data analysis. It provides users a vast range of methods to manipulate and operate on the table for doing inquiries over rows and columns, by using Python two-dimensional data structure (DataFrame). In contrast to NumPy in which all entries in an array should be of the same type, Pandas allows each column to have different types. Pandas also has ability to accord with a great range of formats such as Excel files, SQL, and Comma-Separated Values.
- **Scikit-Learn:** Scikit-learn is a very popular tool and probably the most prominent Python library for ML. It contains several state-of-the-art ML algorithms, as well as comprehensive documentation about the algorithms. Scikit-learn works very well with almost all scientific Python tools. The RF Decision Tree that we are going to use in this chapter is a part of Supervised Machine Learning algorithms which is accessible via Scikit-learn library.
- **Matplotlib:** Matplotlib is a comprehensive 2D and 3D plotting tool in the Python which gives users important insights to the data and provides great

functions for making precise and diverse visualizations such as line charts, histograms, scatter plots, and so on.

- **SMOTE Algorithm:** Heavily unbalanced training sets can reduce the performance of our classification ML algorithm. To rectify this problem, we oversampled the 4 most underrepresented classes using the SMOTE algorithm. We used `imb_os.SMOTE` which is a class that is located in the `imblearn` package and is a part of `smotefamily` library in the Scikit-Learn and gives us a powerful tool to perform oversampling. By applying oversampling methods all categories have equal amount of records simply by oversampling instances of the minority class. Fitting SMOTE is a great way to fix imbalanced our data sets and make the same amount of sources as the number of AGN sources are.

We used the aforementioned libraries, specifically Scikit-learn libraries, packages, and modules for the data preprocessing, classification problem, and data analysis. The code was written using the Jupyter Notebook.

REFERENCES

- Downes, R. A., Webbink, R. F., Shara, M. M., Ritter, H., Kolb, U., and Duerbeck, H. W. (2001). A Catalog and Atlas of Cataclysmic Variables: The Living Edition. , 113(784):764–768.
- Gorenstein, P. (2010). Focusing X-Ray Optics for Astronomy. *X-Ray Optics and Instrumentation*, 2010.
- Greiner, J., Morgan, E. H., and Remillard, R. A. (1996). Rossi X-Ray Timing Explorer Observations of GRS 1915+105.
- Liu, Q. Z., van Paradijs, J., and van den Heuvel, E. P. J. (2006). Catalogue of high-mass X-ray binaries in the Galaxy (4th edition). , 455:1165–1168.
- Liu, Q. Z., van Paradijs, J., and van den Heuvel, E. P. J. (2007). A catalogue of low-mass X-ray binaries in the Galaxy, LMC, and SMC (Fourth edition). , 469:807–810.
- Muller, A. and Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, 1st edition.
- Ritter, H. and Kolb, U. (2003). Catalogue of cataclysmic binaries, low-mass x-ray binaries and related objects (7th edition). , 404:301–303.
- Samus, N. N., Kazarovets, E. V., Durlevich, O. V., Kireeva, N. N., and Pastukhova, E. N. (2009). VizieR Online Data Catalog: General Catalogue of Variable Stars (Samus+, 2007-2017). *VizieR Online Data Catalog*, page B/gcvs.
- Seward, F. D. and Charles, P. A. (2010). *Exploring the X-ray Universe*, volume 4 of 10. Cambridge University Press, 2nd edition.
- Véron-Cetty, M.-P. and Véron, P. (2010). A catalogue of quasars and active nuclei: 13th edition. , 518:A10.
- White, N. E., Nagase, F., and Parmar, A. N. (1995). The properties of X-ray binaries. *X-ray Binaries*, pages 1–57.
- Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O’Reilly Media, 1st edition.