UNIVERSITY of the
WESTERN CAPE

# Machine Learning Approaches to Study Star Formation and Black Hole Accretion in the MeerKAT/MIGHTEE survey

Walter Silima

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science (MSc) in the Department of Physics and Astronomy at the University of the Western Cape.

Supervisor: Prof Mattia Vaccari
Co-supervisor: Dr Fangxia An

May 2023

# Abstract

Galaxy formation and evolution are driven by two main physical processes: star formation and black hole accretion. Both processes can be traced via the synchrotron emission at radio wavelengths. However, a reliable classification of radio sources as star-formation-dominated sources (or Star-Forming Galaxies, SFGs) and black-hole-accretion-dominated sources (or Active Galactic Nuclei, AGN) is non-trivial and often requires extensive use of multi-wavelength data. Although significant effort has been put into classifying radio sources as SFGs or AGN over the decades, the rapid growth of radio data available from facilities such as the South African MeerKAT telescope, the Australian Square Kilometre Array Pathfinder (ASKAP), and eventually the Square Kilometre Array (SKA) requires the development of efficient and reliable automated classification techniques. In this study, we implement, optimise and compare five supervised machine learning (ML) algorithms, namely Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbour ($k$NN), Random Forest (RF) and XGBoost (XGB), to classify radio sources detected in the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE)–COSMOS survey as SFGs and AGN. We first select and analyse the input features to train and test machine learning algorithms to optimally classify sources detected in radio continuum images as SFGs or AGN. We find that the quantities widely used in the literature, such as the mid-infrared colours, the Infrared Radio Correlation parameter ($q_{\rm IR}$), the optical morphology, and stellar mass, are also some of the most effective features for training machine learning models, and that excellent performance can also be achieved without X-ray observations. The feature importance analyses show that the $q_{\rm IR}$ is the most efficient feature in this classification, followed by optical morphology, then combined mid-infrared colours and stellar mass. We then combine the six selected features to train and optimise each of the adopted machine learning models using the F1-score performance metric. We find that supervised machine learning models perform very well in classifying SFGs and AGN detected in radio continuum surveys. All algorithms yield an F1-score $> 90\%$ even when only 20% of the data is used for training using only the $q_{\rm IR}$ feature. We also show that class imbalance has no significant impact on the performance of the machine learning methods since the F1-score is still above 90% with a small training set size. The $k$NN, RF and XGB produce state-of-art results when three or more features are used, while the LR and the SVM perform more poorly for more than two features. We show that X-ray data are not required in order to achieve such an excellent classification performance. We also show that if we remove the mid-infrared colour features log(S8/S45) and log(S58/S36) and only use the log(S45/S36) feature, the performance of $k$NN, RF and XGB drops substantially. This implies that future radio surveys in fields where no deep 5.8 µm and 8.0 µm observations are available will be at a disadvantage when it comes to classifying radio sources as SFGs and AGN.

# Declaration

I, Walter Silima (Student No: 4177331), hereby declare that this work entitled *Machine Learning Approaches to Study Star Formation and Black Hole Accretion in the MeerKAT/MIGHTEE survey*, submitted in partial fulfilment of the requirements for the degree of Master of Science (MSc) in the Department of Physics and Astronomy at the University of the Western Cape, is an original work carried out by me under the supervision of Prof Mattia Vaccari and Dr Fangxia An. It has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold academic ethics and honesty. Whenever an external information, statement, or result is used that has been duly acknowledged and cited.

Signature Date

## Acknowledgement

First, I would love to thank God for the health, strength, patience and wisdom he provided me to complete this work. I warmly thank my supervisor, Prof. Mattia Vaccari, for guiding me throughout the project. His guidance and input have made my work look possible and exciting. I also thank my co-supervisor, Dr Fangxia An, for her constant support. I am forever grateful for her willingness to explain and advise whenever needed. I also appreciate Mr Eslam Hussein for always assisting during the pipeline development. I thank Mr Lennart Heino for his insightful discussion and contribution to this work.

I thank my Mom, Grace Tinswalo Mabasa, for her prayers, support and encouragement. Your prayers for me sustained me this far. I thank Mr Ntsieni Murathi for always being there for me. I will always appreciate the help and advice. I express my appreciation to my love, Muvhulawa Mulaudzi, for her endless moral support and understanding throughout this journey. I thank my brothers, Promise, Vhutali and Witness, for their encouragement and continued support.

I thank the National Research Foundation (NRF), National Astrophysics and Space Science Programme (NASSP), and the Inter-university Institute for Data Intensive Astronomy (IDIA) for providing me with funding to support this work. I also thank IDIA for access to the ilifu cloud computing facility, which provided the computing resources required for this work. Lastly, I thank the University of the Western Cape, its Department of Physics and Astronomy and its Astronomy Group for the valuable discussions and suggestions.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Every galaxy like our own Milky Way is a massive system of billions of stars, gas and dust bound by gravity. Galaxies are the fundamental component of our Universe (Carroll and Ostlie, 2017). Some galaxies have relatively basic structures, contain only stars, and lack distinctive individual characteristics. Some are built entirely of inert gas. Some galaxies are intricate systems of numerous interacting parts, including stars, neutral and ionised gas, dust, molecular clouds, magnetic fields, and cosmic rays (Schneider, 2006). At the centre of many galaxies is a supermassive black hole (SMBH), a compact nucleus that may sometimes be so bright that it overwhelms all the normal radiation of the galaxy. SMBHs cause pairs of the twin lobes of radio galaxies (Donley et al., 2012). Galaxies can also exist in galaxy clusters bound together by the force of gravity in space, just like stars are bound in star clusters (Karttunen et al., 2007).

A galaxy is called a *normal* galaxy if its main energy source is not powered by a central supermassive black hole. Most of the radio emission of a normal galaxy comes from synchrotron radiation released by relativistic electrons and the free-free emission from the HII regions. Normal galaxies have a 1.4 GHz radio power in the range $10^{18}h^{-2}WHz^{-1} \leq L_{source} \leq 10^{-23}h^{-2}WHz^{-1}L$ (Condon, 1992). Galaxies that generate a significant fraction of luminosity from gravitational energy released as matter accretes onto a supermassive black hole are known as Active Galactic Nuclei (AGN) (Sparke and Gallagher III, 2007).

Understanding the detailed formation and evolution of galaxies is one of the significant challenges and fundamental to understanding the history of the universe (Mo et al., 2010). The formation and evolution of galaxies are driven by two main physical processes: star formation and black hole accretion. Both of these processes dominate the radio continuum and can be traced via synchrotron emission at radio wavelengths (Padovani et al., 2017). However, a reliable classification of these processes in the radio continuum is non-trivial and often requires extensive use of multiwavelength data. The star-formation-dominated radio sources are referred to as Star-Forming

Galaxies (SFGs), and black-hole-accretion-dominated sources are called AGN.
Reliable classification of these processes is essential to answer fundamental science questions in modern galaxy formation and evolution studies. According to Way et al. (2012), this simple operation is at the heart of the reductionist process in science: such assignment provides simplification and understanding and enables the derivation of valuable quantities. Classification of radio processes, which is, in this study, a classification of radio continuum sources as SFGs or AGN, is also essential for studying them both individually and statistically. This is an important consideration that motivated the execution of large-area radio surveys over the decades. The well-classified SFGs and AGN, especially those at high redshift, are also essential to constrain galaxy formation and evolution models.
Astronomers have long relied on conventional wisdom and ”gut feeling” to classify radio sources, and much progress has been made in this field to develop several classification criteria based on multi-wavelength galaxy morphology, photometry and spectroscopy. However, in the modern era, astronomical and digital technologies are advancing rapidly, which results in a rapid increase in the amount of data produced by radio telescopes. Radio surveys undertaken with the MeerKAT[1] telescope, the Australian Square Kilometre Array Pathfinder (ASKAP)[2], and eventually the Square Kilometre Array (SKA)[3] will result in an exponential increase in data. In contrast to about 2.5 million known radio sources, the Evolutionary Map of the Universe (EMU) survey on ASKAP is e.g. predicted to find roughly 70 million galaxies, and the SKA is expected to detect 1 billion radio sources in a survey area of $3\pi$ steradians (Bourke et al., 2015). The conventional classification techniques currently in use will not be efficient for such large samples. This necessitates the development of methods such as machine learning techniques to automate the tedious tasks that astronomers have long carried out on their own, hence our study.
The current extragalactic radio surveys with MeerKAT provide unprecedented radio continuum, spectral line, and polarisation information for us to study the formation and evolution of galaxies in the distant Universe. The rapid growth of radio data from MeerKAT surveys thus provides an excellent opportunity to apply machine-learning approaches to classify radio sources.
The following sections in this chapter are arranged in the following order. Section 1.1 describes SFGs with their properties. We then briefly discuss AGN properties and the different subclasses in subsection 1.2. Section 1.3 details the traditional techniques used to distinguish between AGN and SFGs and their limitations. Section 1.4 discussed the Source counts and their importance in radio astronomy.

[1] MeerKAT https://www.sarao.ac.za/gallery/meerkat/

[2] ASKAP https://www.atnf.csiro.au/projects/askap/index.html

[3] SKA https://www.skatelescope.org/the-ska-project/

## 1.1 Star Forming Galaxies

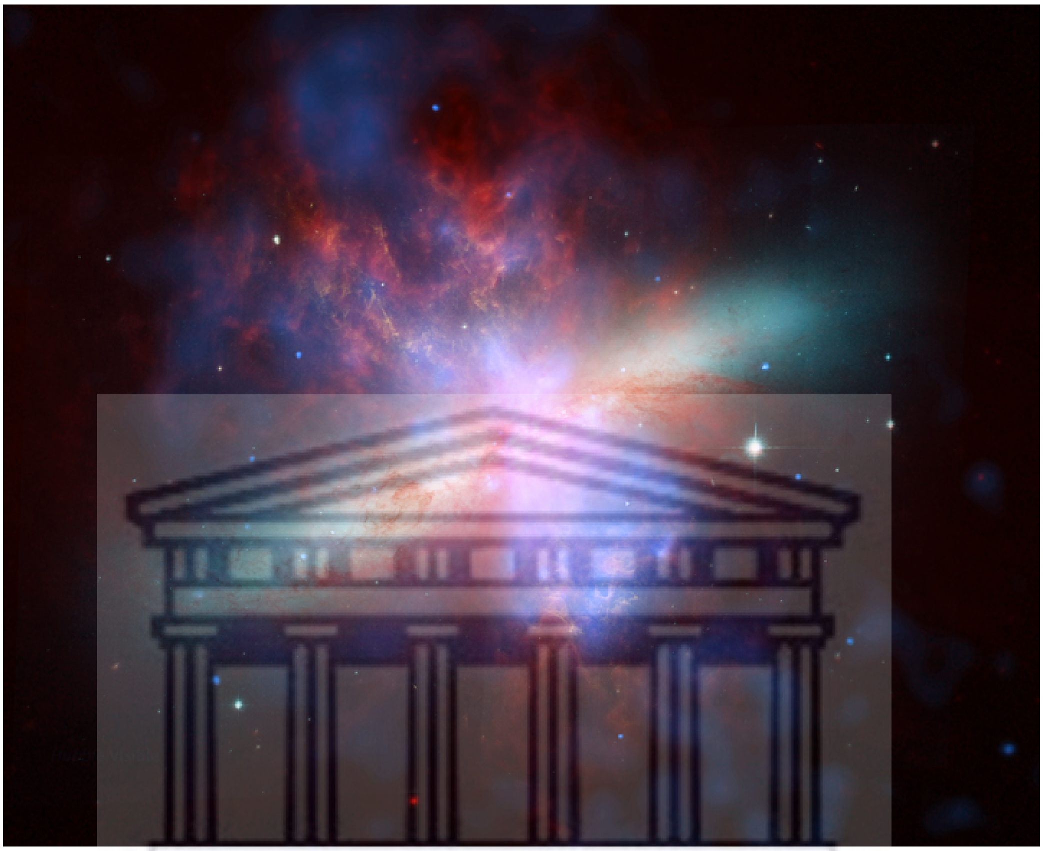


Figure 1.1: The multiwavelength view of the star-forming galaxy M82. Image adapted from https://chandra.harvard.edu/photo/2006/m82/

Star-forming galaxies are *normal* galaxies with a high rate of ongoing star formation (SF) activity. Figure 1.1 shows the multiwavelength view of the star-forming galaxy M82. A star-forming galaxy comprises low-frequency radio emission dominated by synchrotron radiation[4] from relativistic electrons, accelerated by supernovae and their remnants, interacting with the galactic magnetic field (Bonato et al., 2021). Figure 1.2 shows the typical relative intensities of synchrotron radiation, free to free emissions[5], and dust re-radiation[6] over the radio or far-infrared (FIR) spectrum of the M82 galaxy. The spectral energy distribution (SED) of an SFG can be understood as a superposition of emissions from the synchrotron, free-free and dust re-radiations.

[4] synchrotron radiation; refers to the emission of electromagnetic radiation by high-energy charged particles moving through magnetic fields in space

[5] free-free radiation; https://www.cv.nrao.edu/~sransom/web/Ch4.html.

[6] dust re-radiation; https://ned.ipac.caltech.edu/level5/Phinney/Phinney1.html

Star-forming galaxies are characterised by steep GHz radio spectra (spectral index $(\alpha) = 0.7$) dominated by synchrotron emission but also have a flat free-free component, which becomes predominant at frequencies $\geq$ 30GHz (Condon, 1992).

Since the synchrotron emission emitted by SFGs results from relativistic plasma accelerated in supernova remnants associated with massive star formation (Condon, 1992), thus radio observations probe recent star formation activity and, to some extent, pinpoint its location. The far infrared-radio correlation, one of the tightest correlations in observational astrophysics, supports this. The far infrared and radio emission are firmly and nearly linearly correlated in a variety of star-forming sources, and this correlation is driven by recent star formation activity (Padovani, 2016a).
SFGs are rich in gas and dust and are believed to dominate at higher redshifts ($z > 1.2$), where star-formation activity is dominant. A clumpy structure at these redshifts characterises the SFGs since their disks are richer in gas than the local universe (Genzel et al., 2014). It is proposed that star formation activities occur via two modes: a starburst triggered by major mergers; and another way is that the energy released by evolved stars and supernovae can affect the interstellar medium and shock

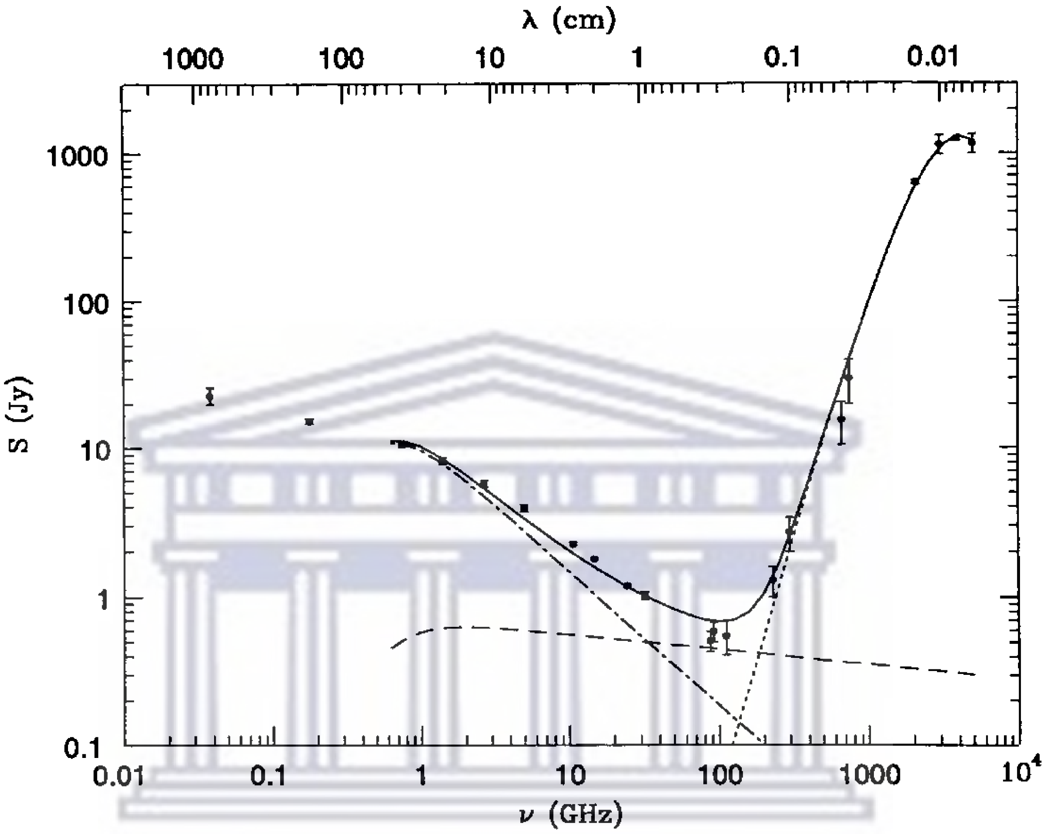


Figure 1.2: The Observed radio/FIR spectrum of M82 is the sum (*solid line*) of synchrotron (*dot-dash line*), free-free (*dashed line*) and dust reradiation *(dotted line)* components. Image adapted from (Condon, 1992).

it to induce the formation of new stars (Padovani, 2016b).

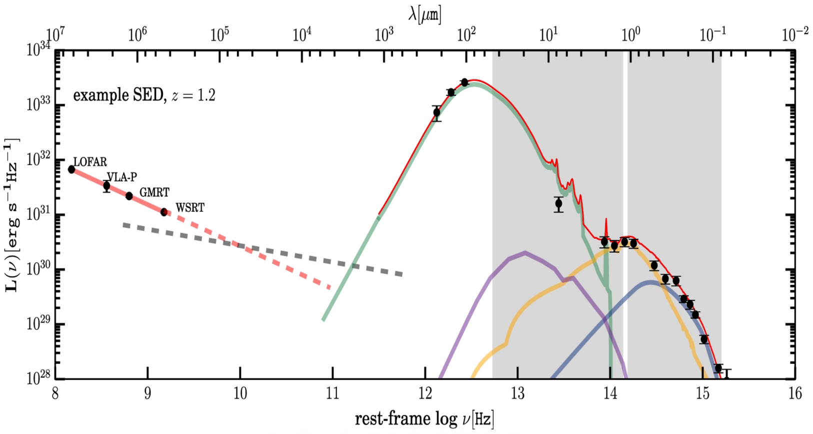


Figure 1.3: Example SED of a dusty SFG at $z = 1.2$. The multiwavelength coverage is from the low-frequency radio to the UV regime. The left rectangle covers the integration limits for the main IR components: the galactic cold dust emission (green line) and the hot dust emission from the AGN torus (purple line). The right rectangle is the integration area for the optical main components: the stellar emission (orange line) and the accretion disk emission (blue line). The thin red line represents the total fitted SED. The dashed red and grey lines correspond to the synchroton and thermal contributions to the radio emission, respectively. The image is acquired from Calistro Rivera et al. (2017).

## 1.2 Active Galactic Nuclei

An active galactic nucleus (AGN) is a compact region at the galaxy's centre with excess non-stellar luminosity over some part of the electromagnetic spectrum (Sparke and Gallagher III, 2007). Figure 1.4 shows a schematic representation of an AGN spectral energy distribution (SED)(e.g. Elvis et al. (1994); Richards et al. (2006)). An AGN is powered by the material from the surrounding interstellar medium accreting onto a central supermassive black hole, which results in the surrounding gas photoionised by the photons produced via the accretion mechanism (Peruzzi et al., 2021).

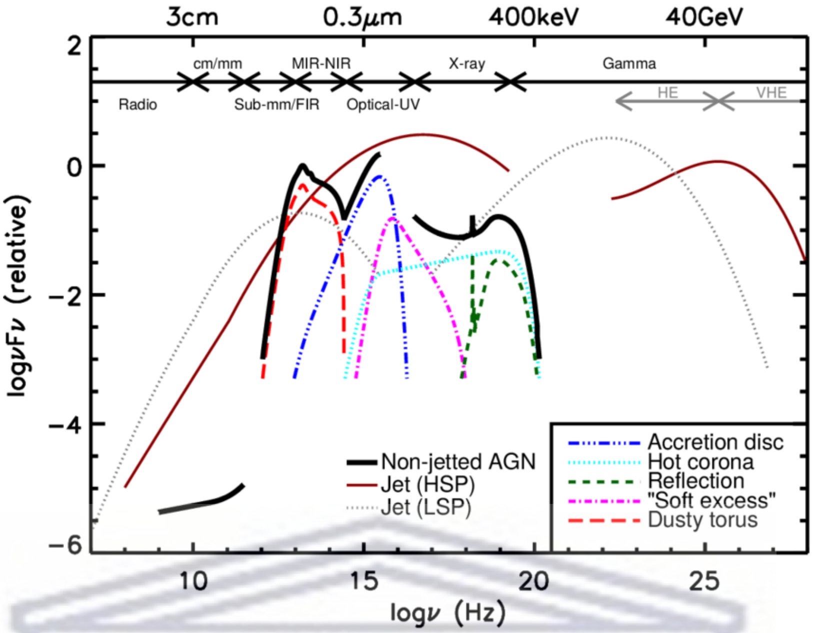


Figure 1.4: A schematic representation of an AGN spectral energy distribution (SED), (e.g. Elvis et al. (1994); Richards et al. (2006)). The black solid curve represents the total emission, and the various coloured curves (shifted down for clarity) represent the individual components. The image is adapted from (Padovani et al., 2017).

Observations of AGN at different wavelengths and variations over time reveal their different structures. AGN emits radiation over all the electromagnetic bands, even at the X-ray and gamma ($\gamma$)-ray part of the electromagnetic spectrum where most normal galaxies hardly emit radiation (Padovani et al., 2017). The population of AGN is different and shows different properties, and as a result, they have historically been described in terms of several subclasses of objects depending on the electromagnetic band they were discovered in; see Figure 1.5 for the various populations of AGN.
The AGN unification scheme has been developed over the past decades, and Urry and Padovani (1995) developed the sketch that illustrates the ideas behind the unification scheme as shown in Figure 1.6. The sketch indicates the composite AGN phenomenon; black hole, disk, torus, clouds and jet. According to Figure 1.6, the appearance of an AGN depends crucially on the orientation of the observer to the symmetry axis of the accretion disk (Dermer and Giebels, 2016). Thus, the similar intrinsic AGN may be confused with different radio sources.
AGN identification is not always clear-cut. The numerous methods used to identify AGN include classification based on emission line spectra, morphologies, and frequency. The observational classification of AGN is dominated by the dichotomy between radio-quiet (RQ) and radio-loud (RL) classes.

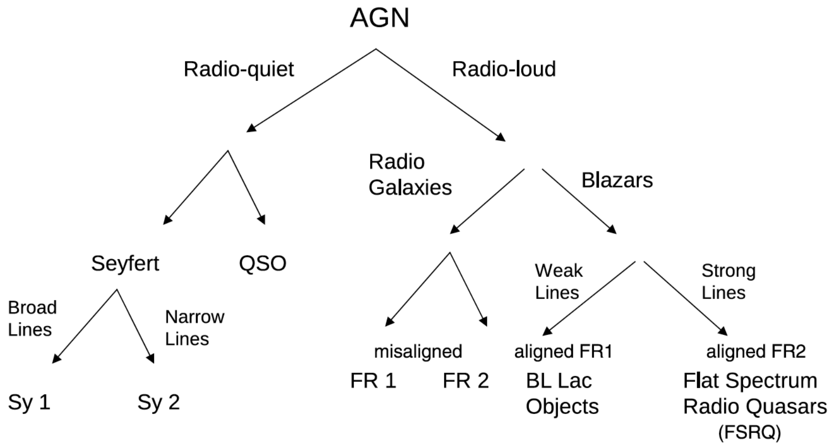


Figure 1.5: Observational classification of active galaxies. Image adapted from (Dermer and Giebels, 2016)

Thermal emission, directly or indirectly associated with the accretion disk, dominates the RQ AGN's multiwavelength emissions (Padovani et al., 2017). Quasars are the most common RQ AGN class, the most potent kind. Although quasars were originally discovered due to their radio emissions, only about a small fraction of quasars have substantial radio emissions. These quasars are now called radio-loud quasars. Quasars are characterised by higher luminosities that outshine the whole galaxy system. Very luminous active nuclei, such as the quasars, were far more common when the Universe was 20–40% of its present age than they are today (Sparke and Gallagher III, 2007).
The largest class of RQ AGN are Seyfert galaxies. Seyfert galaxies are distinguished from quasars via the energy emitted by the central compact region. Seyfert's energy emissions in the visible wavelength are equivalent to the energy emitted by stars in the galaxy. At the same time, quasars are brighter than the energy emitted by stars by a factor of about 100 or more (Peterson, 1997). Seyfert galaxies are subdivided into two subclasses; *Type 1* and *Type 2*. The physical explanation of the difference in the two Seyfert subclasses is provided by the *Unified Model*, which states that for *Type 1* Seyferts, the observer looks directly into the unobscured accretion disk by the fast-moving gas and for *Type 2* the observer's line-of-sight to the accretion disk is blocked by an obscuring medium (Padovani et al., 2017).

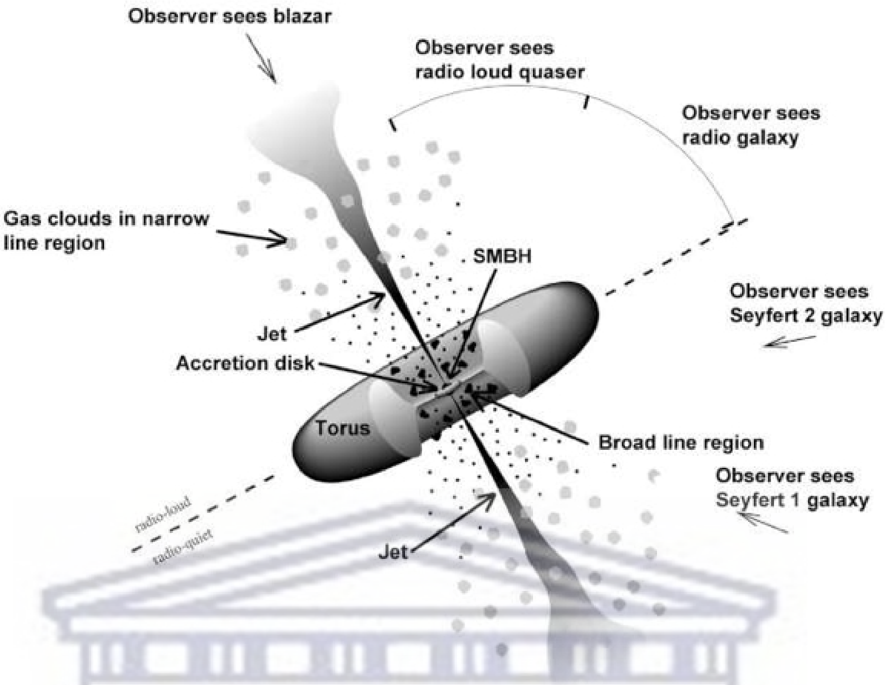

Observer sees blazar
Observer sees radio loud quaser
Observer sees radio galaxy
Gas clouds in narrow line region
SMBH
Observer sees Seyfert 2 galaxy
Jet
Accretion disk
Torus
Broad line region
Observer sees Seyfert 1 galaxy
radio-loud
radio-quiet
Jet

Figure 1.6: The unification model for AGN. Different viewing angles can be understood as the cause of different types of AGN. Image sourced from https://fermi.gsfc.nasa.gov/science/eteu/agn/

A RL class is made of radio galaxies that emit a significant fraction of their energy non-thermally and in association with the powerful relativistic jets. RL AGN is subdivided into the Blazars and Radio galaxies. B.L Fanaroff and J.M Riley identified radio galaxies with active nuclei as Fanaroff-Riley class 1 *(FR-I )* galaxy or Fanaroff-Riley class 2 *(FR-II )* galaxy. The classification scheme is based on the fact that radio galaxies have different morphologies depending on their radio power; The Fanaroff-Riley classification scheme has been explained in detail in section 1.3.1.
There are also other classes of RL AGN not shown in Figure 1.5, such as *low- and high-* excitation AGN class, grouped according to the optical spectroscopic properties, where objects without and with high-excitation emission lines in their optical spectra are referred to as low-excitation galaxies (LEGs) and high-excitation galaxies (HEGs), respectively (Padovani et al., 2017).
For this study, we consider any radio source in the MIGHTEE-COSMOS catalogue belonging to any AGN subclass as an AGN and use machine learning techniques to separate them from star-forming galaxies by considering their multiwavelength measurements.

## 1.3 Conventional Source Classification

Classification is often one of the first steps taken in a scientific endeavour and usually precedes – and indeed stimulates – understanding of the physical causes of the phenomenon in question (Tadhunter, 2008). Galaxy classification emerged after Edwin Hubble confirmed that some nebulae are external galaxies, i.e. the Galaxy is only one of the countless galaxies in the universe. After this realisation, the work began determining the physical properties of galaxies (Carroll and Ostlie, 2017). One task was to classify galaxies based on their intrinsic properties. This section discusses two classification schemes developed to classify the extragalactic sources in the radio continuum and their limitations. Each classification method played an essential role in the development of astronomy, although all were developed using heuristic methods.

### 1.3.1 Fanaroff-Riley Classification

In 1974 B.L Fanaroff and J.M Riley created a classification scheme that identifies the radio galaxies with active nuclei as Fanaroff-Riley class 1 (FR-I) galaxy or Fanaroff-Riley class 2 (FR-II) galaxy (Fanaroff and Riley, 1974). The classification scheme was based on the fact that radio galaxies have different morphologies depending on their radio power. Radio sources that belong to the FR-I were brighter towards the centre and fainter toward the edge. The spectra of FR-I galaxies are very steep, which indicates that sources giving rise to emissions have aged. On the other hand, FR-II sources are brighter towards the edge and are more luminous than their counterparts, with bright hotspots at the ends of their lobes (Kembhavi and Narlikar, 1999). Various properties of sources in the two classes are different, meaning that the distinction between FRI and FRII is important since it presents a direct link between the galaxy's luminosity and how energy is transported from the central region and converted to radio emission in the outer parts (Kembhavi and Narlikar, 1999). Figure 1.7, shows radio images of FR-I (left) and FR-II(right).

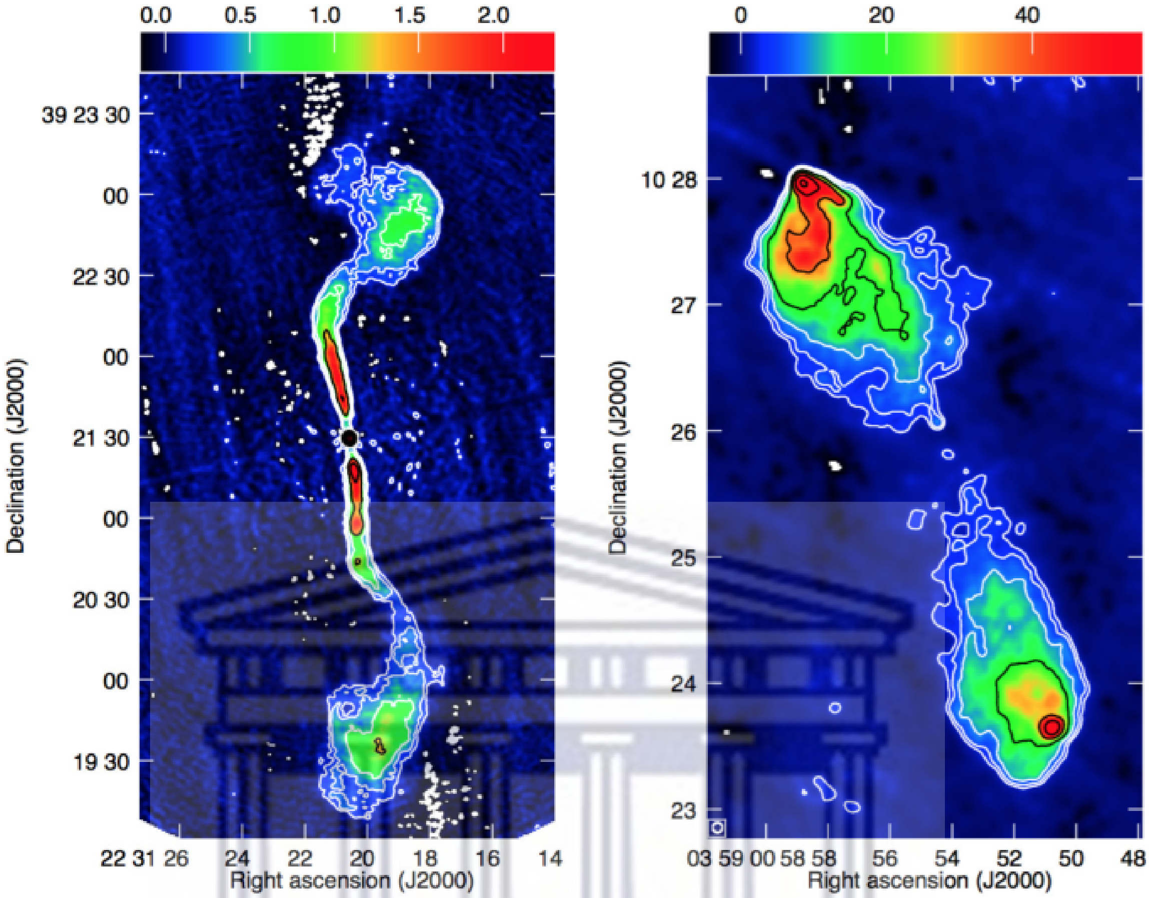


Figure 1.7: The FRI galaxy 3C 449 (left) and the FRII galaxy 3C 98 (right) illustrate the FR morphology. The red area indicates the brightest radio emission. Red represents the brightest radio emission, closer to the centre of FRI and at the edges of FRII. Image adapted from (Becker and Grobler, 2019)

The conclusion for the Fanaroff-Riley classification was based on a set of 57 radio galaxies and quasars from the complete revised version of the Third Cambridge catalogue (3CR catalogue (Bennett, 1962)) clearly resolved at 1.4 GHz or 5 GHz into two or more components. However, recent work by Mingo et al. (2019) has demonstrated, for a sample of 5,805 extended radio-loud AGN from the LOFAR Two-Metre Sky Survey (LoTSS), that radio luminosity does not reliably predict whether a source is edge-brightened (FRII) or centre-brightened (FRI). Further Mingo et al. (2022) found that the relationship between accretion mode and radio morphology is very indirect, with the host-galaxy environment controlling these two critical parameters differently. The existence of radio sources with hybrid FRI/FRII morphologies further complicates the distinction between FRI and FRII morphologies (Gopal-Krishna and Wiita, 2000). Although figuring out radio morphologies is a beneficial place to start when understanding radio galaxies better, multiwavelength observations are necessary to more precisely locate the centres of radio galaxies and investigate their physical characteristics (Prescott et al., 2018).

### 1.3.2 Multi-Wavelength Classification

The Fanaroff-Riley classification scheme is a physical-based classification schema that relies upon subjective and potentially incorrect model interpretation. There is a vast difference between what is observed and what is inferred to cause what is observed, the latter deriving from potentially several interconnected and complex physical processes (Way et al., 2012). Over the decades since two basic radio galaxy morphologies were recognised, several findings in different wavebands have reported properties on different scales (Saripalli, 2012). Currently, astronomers use the available multiwavelength observations to develop a radio source classification scheme based on the physical processes responsible for the formation and evolution of the radio sources.

A radio spectrum is typically a power-law spectrum with little information we can learn about sources of interest (Condon, 1992). The sensitivity of radio, infrared, optical and X-ray telescopes has increased. The overlap between radio, infrared, optical and X-ray surveys has grown to the point where many radio galaxies are detectable in the infrared, infrared galaxies are detectable in the optical, optical galaxies are detectable in the X-ray, and optical galaxies are observable in the radio, and vice versa (Norris, 2017). Combining measurements from different wavelength bands helps distinguish whether a source is an SFG or an AGN. It is also used to find redshifts of sources and extract information about the host galaxy, such as star formation rate, stellar mass, etc. (Padovani, 2016a).

As indicated in section 1.2, AGN has interesting properties and has a history of being discovered in all spectral bands. Different methods are deployed in each spectral band to identify AGN, and we learn different AGN physics from each spectral regime. For example, the infrared (IR) band is primarily sensitive to obscuring material and dust. The optical/ultraviolet (UV) band is related to emission from the accretion disk. In contrast, the X-ray band traces the emission of a (putative) corona. $\gamma-ray$ and (high flux density) radio samples, on the other hand, preferentially select AGN emitting strong non-thermal (jet [or associated lobe] related) radiation Padovani et al. (2017).

A summary of the main indicators used to classify faint radio sources in the radio continuum given by Padovani (2016a) includes *far-infrared-radio correlation*, *X-ray power*, *IRAC colour-colour diagram*, *X-ray spectrum and variability*, *optical indicators*, *VLBI detection* and *radio polarisation*. A summary of the classification scheme used to classify the MIGHTEE-COSMOS catalogue is given in section 3.5, however for a detailed explanation regarding each classification scheme, we refer the reader to (Padovani, 2016a).

Figure 1.8 shows the typical spectrum of a quasar (3C273) compared to that of an elliptical galaxy. Radiation emitted from the quasar spans the entire electromagnetic

spectrum while the radiation from an elliptical galaxy is concentrated over a small range of frequency (Schneider, 2006).
Using the available multiwavelength data to classify observed radio galaxies as AGN or SFG was significant progress and had several advantages. The advantages are that; classes are directly related to the actual physical properties of galaxies (essential), a reduced misclassification, clear distinction of AGN subclasses, etc. Adopting multiwavelength classification also significantly impacts extragalactic astronomy. Radio sources are now classified according to their physical processes rather than based on the method or band they are discovered at. When working with deep radio images, a multiwavelength strategy is essential because most sources are unresolved and not much information can be gleaned from a single frequency catalogue.
The main challenges to the conventional multiwavelength classification are that an experienced astronomer has to apply several methods independently and manually, which is often ineffective for large data sets. This also requires astronomers to distribute the data to different experts for classification, after which they will compare the classification results from each expert. Another concern is that each AGN indicator used for classification leaves a significant fraction of sources unclassified; even when all the classifications from each indicator are combined, many sources still need to be classified. Figure 3.6 shows the completeness of MIGHTEE-COSMOS sources classified as AGN or SFGs using four conventional techniques. This work uses the accurately derived multiwavelength information to automate source classification in the radio continuum using supervised machine learning methods.

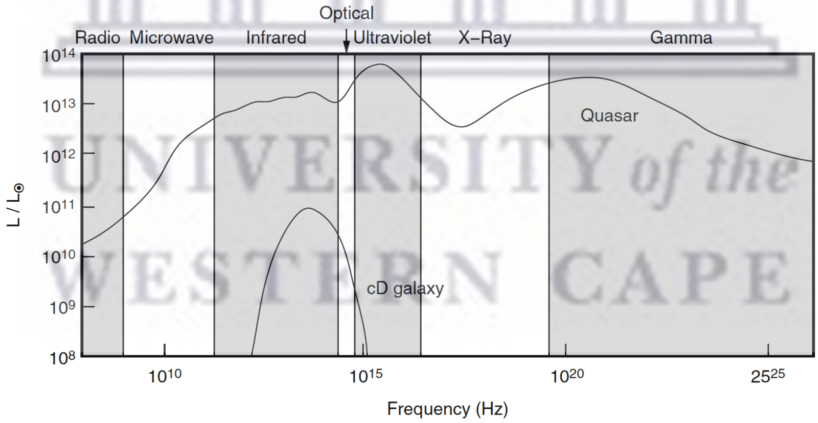


Figure 1.8: The spectrum of a quasar (3C273) [an AGN] as compared to that of an elliptical galaxy [a normal galaxy] in terms of the ratio $\frac{vL_v}{L_\odot}$. The radiation emitted by an elliptical galaxy is concentrated in a narrow range spanning less than two decades in frequency, while the emission from the quasar is observed over the full range of the electromagnetic spectrum, and the energy per logarithmic frequency interval is roughly constant (Schneider, 2006).

## 1.4 Radio Source Counts

According to Prandoni et al. (2001), radio source counts are the most immediate product, which can be derived from a radio survey and reflect the statistical properties of the radio source populations. The source counts distribution of radio sources from a radio-astronomical survey is the cumulative distribution of the number of sources (N) brighter than a given flux density (S). It is usually plotted on a log-log scale. It is one of several cosmological tests developed in the 1930s to evaluate and contrast new cosmological hypotheses (Kellermann and Wall, 1987).
The shape of the source counts is tightly related to the sources' evolutionary properties and the geometry of the Universe (Padovani, 2016a). The relative contribution of various sources at each flux, which is merely the luminosity functions at multiple redshifts, is primarily determined by the slope of the source counts. The source counts, therefore, represent the primary observational constraint to evolutionary models of radio sources. Accurately determining the source counts, their slope and normalisation are necessary to make this helpful constraint. The source counts, together with measured local radio luminosity functions, can then be used to make predictions about the redshift distribution, for instance, (Prandoni et al., 2001). The normalized source counts derived recently from a compilation of several deep radio surveys at 1.4 GHz are shown in Figure 1.9. This plot illustrates the long-standing issue of the large scatter in the counts present at flux densities ≤500 mJy.

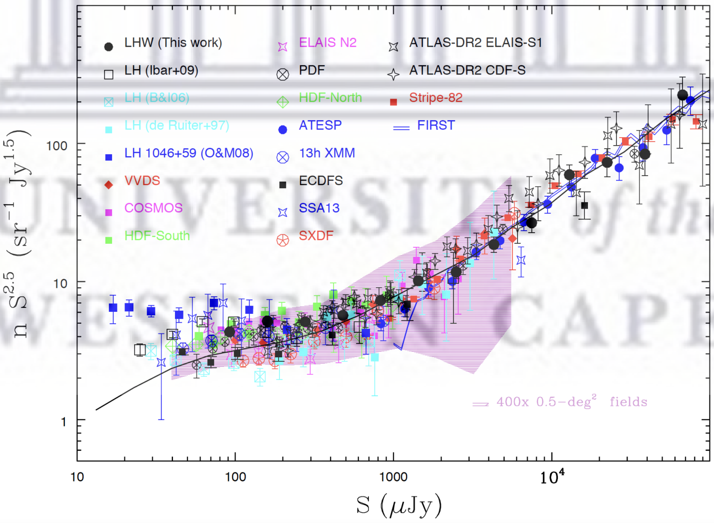


Figure 1.9: Differential source counts derived from deep 1.4 GHz surveys, and normalized to a non-evolving Euclidean model ($nS^{2.5}$) as a function of flux (S). Different samples are indicated in the figure and the text with different colours. The image is adapted from (Prandoni et al., 2018)

## 1.5 Aim of Study

Despite significant advancements in radio source classification, the volume of radio data produced by facilities like MeerKAT, ASKAP, and, ultimately, the SKA will outweigh the efficiency of the currently used manually applied methods. This work aims to build an effective and trustworthy automatic classification pipeline in response to the data influx from such facilities. We aim to implement and optimise supervised machine learning techniques to classify radio sources detected in the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) survey as an AGN or SFG. MIGHTEE is a project undertaken by a South African-led international collaboration of researchers to explore SFG and AGN evolution over cosmic time with the MeerKAT telescope. The project focuses on four well-studied extragalactic deep fields; COSMOS, XMM-LSS, ELAIS-S1 and E-CDFS, making up to 20 $deg^2$ at µ$Jy$ sensitivity at Giga-Hertz frequencies (Jarvis et al., 2016). MIGHTEE observations combined with the excellent multiwavelength data in these fields can provide a significant step forward in our understanding of galaxy evolution (Whittam et al., 2022). We take advantage of the MIGHTEE-COSMOS multiwavelength catalogue, whose manual classification is given by Whittam et al. (2022) to derive a machine learning pipeline that will be useful to classify billions of radio sources detected in future radio surveys.

## 1.6 Thesis Layout

This work aims to implement an automated method for classifying radio sources as star-formation-dominated or black-hole-accretion-dominated by exploring multiwavelength information in the MIGHTEE-COSMOS field. A summary of the chapters that follow is provided as follows.

- **Chapter 2**
  This chapter briefly introduces our machine learning approach for people with or without a background in machine learning. It reviews the general approach to most supervised machine learning classification problems. It details the necessary steps to acquire good data for supervised machine learning training, such as data preprocessing and feature selection. We also introduce the five supervised machine learning models or classifiers; Logistic Regression, Support Vector Machines, $k$-Nearest Neighbour, Random Forest and XGBoost. The approach each model adopts to determines the class of a source discussed. Hyperparameter optimisation has been described briefly, including different techniques for

selecting hyperparameters. The chapter concludes by giving examples of machine learning studies adopted in astronomy.

- **Chapter 3**
  This Chapter presents the radio and the MIGHTEE-COSMOS catalogue used in this work. MIGHTEE-COSMOS catalogue is a multiwavelength dataset comprising several wavelength measurements from X-ray to radio. Since most sources in deep radio images are unresolved and little information can be learned from a single frequency collection, multiwavelength data is essential for classifying radio sources. The MIGHTEE survey is introduced briefly in this Chapter. This chapter details the COSMOS field background and why it is a focus of this study. Briefly introduced the MIGHTEE-COSMOS radio data, including the available corresponding multiwavelength equivalent. The conventional techniques used to produce this catalogue are described, and the number of radio sources per class is also presented. We have also described the features used to train supervised machine learning models.

- **Chapter 4**
  This chapter presents the main results of applying supervised machine learning to classify radio sources. It briefly compares the automated feature analysis techniques with the conventional methods. We presented and compared all supervised machine learning results thoroughly. We also analyse these findings and detail their implications in this chapter.

- **Chapter 5**
  This chapter gives a summary of the results and implications of this work. Additionally, promising prospects are proposed for the advancement of this work.

# Chapter 2

# Introduction to Machine Learning

Traditionally, to achieve an outcome in computing, one writes explicit programs/rules that take input and generate the output described in the program. Once the program is written, it often works within the specific domain it is designed for and cannot change unless manually changed by a programmer (Alpaydin, 2020). This necessitates in-depth domain expertise and regular system maintenance, both of which can be expensive and time-consuming. The vast development of technology has resulted in an exponential increase in the amount of data available to researchers over a short period of time. Traditional programming will not cope with such data, and as such fast, automatic data-driven techniques such as machine learning techniques are required to learn from the data (Wang and Alexander, 2016).

In contrast to traditional programming, machine learning models take the input data and the desired outcome and automatically generate the model to solve the problem. The resulting models are the programs generated automatically with little or no user programming guidance Alpaydin (2020). These techniques have several advantages with respect to traditional programming tools, such as; they can identify patterns that are too complex for humans to observe, can also make predictions based on a much larger data set than traditional methods, they are not biased by human emotions or subjective opinions; they can adapt to changing data patterns over time and make more efficient use of computing and storage resources (Olah et al., 2018).

The first machine learning algorithm was introduced in 1959 by Samuel (1959) and was applied to computer gaming. Samuel (1959) programmed a checkers-playing program that played 10,000 games of checkers against itself. In a live game, he could then figure out which board position was good or bad, depending on the wins and losses of that move. However, one significant disadvantage of early machine learning algorithms was their need for higher computer processing power than available at the time. Recently, due to the improvement of the processing power of computers, machine learning algorithms have been applied to a wide variety of problems and in different fields. For example, Moosavi et al. (2021) demonstrates machine learning techniques to study the uncertainty in numerical weather prediction models due to the interaction of multiple physical processes. Vijayakumar (2019) puts forth the capsule neural network, the machine learning system that can be trained using a less number of datasets, unlike convolutional neural networks and is sturdy against the rotation or the affine conversions, to identify the type of cancerous tumours in the brain at its early stage. Dixon et al. (2020) introduces fundamental concepts in machine learning for canonical modelling and decision frameworks in finance. For other examples of machine learning research, we refer a reader to visit Olivas et al. (2009). An application of machine learning in astronomy summary is given in this work in section 2.4.

In the previous chapter, we pointed out important traditional source classification techniques that astronomers have used to classify extragalactic radio sources. The method's ability and limitations are explained briefly. One standard limit of conventional classification techniques is that the methods will not scale with the large amounts of radio data readily available, especially for SKA surveys. It will be impractical to classify the astronomical data conventionally. Adding multi-wavelength data to large amounts of existing data further exacerbates the impracticality.

Astronomy continues to be a scientific discipline to experience a flood of data over the past few decades. The amount of data is pushing the limits of what is technically possible. The exponential increase of available data in this field requires adopting fast and automatic techniques to find patterns in data and translate these patterns into useful information, hence *machine learning* techniques.

Compared to the conventional classification techniques, supervised machine learning models classify every source in the data; this means that every observation in the catalogue is classified using the supervised machine learning models. The well-trained and optimised models can classify all the observations in a catalogue within minutes to hours instead of the weeks or months that astronomers can take to produce a catalogue's classes for sources. The conventional techniques consider several methods independently, and each method's outcome is combined to conclude the class. In contrast, we have demonstrated that each supervised machine learning method has a specific way of finding the relationship between the input data and the class in the training data from which it can predict the class of sources in the unseen test data.

This chapter will overview the machine-learning approach to radio source classification and perhaps inspire exciting new ideas and applications. Although this work focuses on classification, we regard machine learning algorithms as tools that, when applied correctly, have vast potential for creating valid scientific results. I briefly discuss the methodology applicable to any classification problem with supervised machine learning.

The section in this chapter are arranged as follows; Section 2.1 gives a general overview of machine learning. The supervised classification approach is discussed in detail in section 2.2. This includes significant processes such as data collection (2.2.1), data preprocessing (2.2.2), feature selection (2.2.3), selection of a supervised machine learning models (2.2.4) and the evaluation metrics. In section 2.3, we discussed hyperparameter optimisation and how important this is for machine learning models. The last section, 2.4, gives a general overview of the application of supervised machine learning techniques in astronomy.

## 2.1 Overview of Machine Learning

Machine learning (ML) is a field of study in Artificial Intelligence (AI, Winston (1984)) and one of the fastest-growing areas of computer science, with far-reaching applications Buradkar and More (2020). It is a method of data analysis that automates analytical model building by processing the available data and maximising a problem-dependent performance criterion. The automatic model-building process is called *training*, and the data used for training purposes is called *training data*. The trained model can provide new insights into how input variables (also referred to as features) are mapped to the output, and it can be used to make predictions for novel input values that were not part of the training data called the *test data* (Baştanlar and Özuysal, 2014). There are numerous uses for ML, but data mining is the most important (Kotsiantis et al., 2007).

There are three categories of ML techniques (Akinsola, 2017); **Supervised learning**, **Unsupervised learning** and **Semi-supervised learning**. Supervised learning is a technique in which every instance in the dataset has an associated output. In supervised learning, we wish to fit a model that relates the output to the observations (predictors) in the training data to accurately predict the output for future observations in the test data that were not part of the training data (Borne, 2013). Suppose the output we wish to predict is a categorical value like a class to which observation belong to. In that case, it is referred to as *classification*; otherwise, if the objective is to predict a numerical value, it is called *regression*. An unsupervised machine learning technique is one in which each observation or instance in the dataset is not associated with output. This type of learning is unsupervised because no output variable can supervise the analysis. The main applications of unsupervised learning include clustering, visualisation, dimensionality reduction, and finding association rules. Semi-supervised learning considers observations with the output and some observations with no output. We wish to use an ML model that can include the observations for which learning output measurements are available and other observations for which they are not. In this study, we use the words output, class label, target and response interchangeably; all these refer to the target variable, a class to which an observation belongs. Figure 2.1 shows three major categories of machine learning techniques that have been a focus of most research (Borne, 2013).

The machine learning approach in astronomy is applied in data characterisation, assigning the output to observed celestial sources and discovering the unknowns. Data characterisation is an unsupervised problem, classification is a supervised problem and finding the unknown is a semi-supervised problem (Borne, 2013). With the data avalanche that astronomers face, such tools will be essential to perform data analyses (Gareth et al., 2013). For this study, we limit our detailed discussion to supervised ML classification techniques.

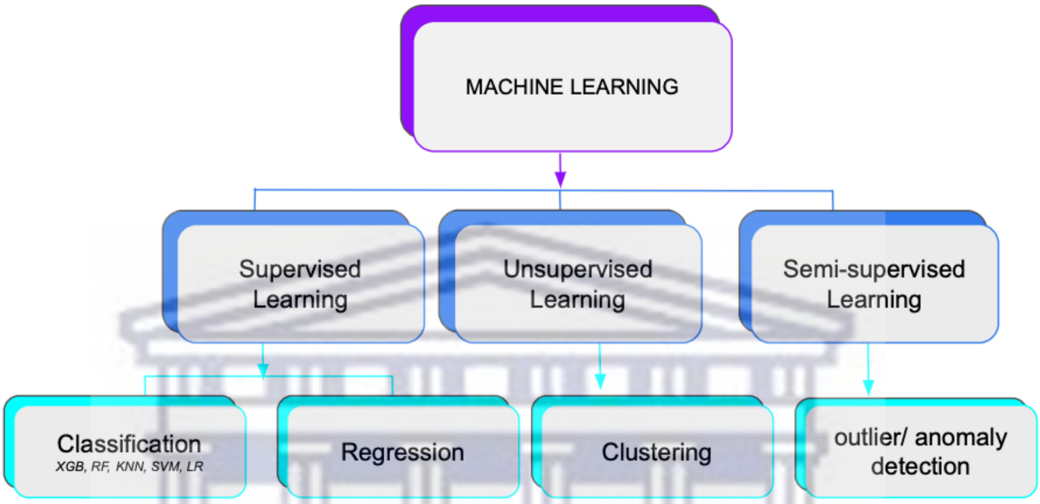


Figure 2.1: Classification of the machine learning approaches according to the learning.

## 2.2 Supervised Classification Approach

Classification is a supervised ML problem (see Figure 2.1). In this case, the target output is categorical such as classes to which observations in the dataset belong. The objective is that the algorithm should classify observations into two or more outputs. If the algorithm is classifying between two target outputs, it's called binary classification. And if the outputs are more than two, it is called a multi-classification problem. Both these classifications have been explored in astronomy, and section 2.4 gives some studies that have considered these classifications. In this work, we are focused on a binary classification since the task is to classify between SFGs and AGN.

In a Classification setting, the task is to assign a class label (y') to an unknown observation (X') based on the dataset $D((X_1, y_1), ..., (X_n, y_n))$ referred to as a training data. A supervised ML model learns a map that relates an instance (X) to a class label (y) such that when applied to a new sample (X') which was not part of training data, the model should assign the most likely class label (y'). In other words, the supervised classification objective is to build a concise model of the distribution of class labels in input features. The resulting well-trained classifier is then used to assign class labels to the testing samples where the values of the predictor features are known, but the value of the class label is unknown (Kotsiantis et al., 2007). The class labels provided on the training data are usually manually supplied by humans and are frequently imperfect. As such, if there are biases in a manually labelled training set, it will result in a biased machine-learning algorithm.

There are two different approaches in which a data classification task can be carried out: the first considers only a dichotomous distinction between the two classes. It assigns class labels 0 or 1 to an unknown source. The second attempts to model probability $P(y|X)$; this yields a class label for a source and a probability of class membership. The most prominent representatives of the first approach are Support Vector Machines (SVM). Logistic Regression (LR), $k$-Nearest Neighbours ($k$NN), and Decision trees all make use of the second approach. However, they vary considerably in building an approximation to $P(y|X)$ from data (Dreiseitl and Ohno-Machado, 2002). We adopt five different supervised classification algorithms; SVM, which considers the first classification approach, and four other models, which consider the second approach of estimating probabilities; Extreme Gradient Boosting commonly popular as XGBoost (XGB), Random Forest (RF), $k$NN and LR. The details on each model's specific approach to building an approximation to $P(y|X)$ from data are given in section 2.2.4.

This study aims to classify the MIGHTEE-detected radio sources as AGN or SFGs, and there are several steps that one needs to follow to produce good results for a classification problem. Figure 2.2 shows the supervised machine learning processes I will adopt to classify the MIGHTEE-COSMOS radio sources. Figure 2.2 establishes a layout of the steps necessary for any supervised learning task. The following subsections discuss the most critical steps concerning astronomy datasets for classification, which are *data collection* (2.2.1), *preprocessing* (2.2.2), *feature selection* (2.2.3), *algorithm selection* (2.2.4), *evaluation metrics* (2.2.5), and improving the results; *Hyperparameter optimisation* (2.3).

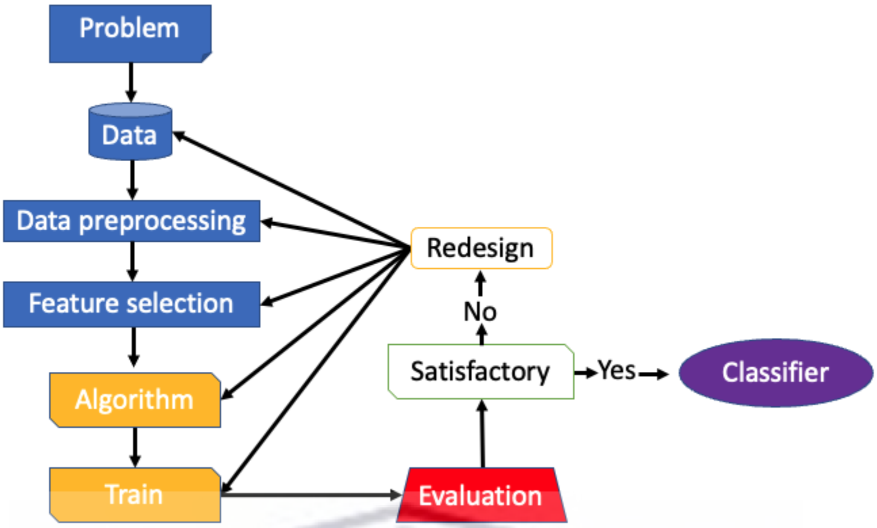


Figure 2.2: The steps involved in Supervised ML classification. The first four steps in blue are essential since the model performance mostly depends on them. The first step is identifying whether a *Problem* is feasible for classification or regression. The second step is data collection (2.2.1), and the third step is *preprocessing* the data (2.2.2). The fourth step is *feature selection* (2.2.3). The steps highlighted in yellow are the selection of algorithms (2.2.4) and training algorithms. The last step is *evaluation* (2.2.5) of the model's performance. If the classifier can make good classification, it is retained and used to classify observations not in the training data; otherwise, one or more steps in blue and yellow should be repeated, and the models should be retrained.

### 2.2.1 Data Collection

Data collection is the first process of all the steps needed to ensure the desired data is in digital format. Data collection methods include acquiring and archiving new observations, querying existing databases according to the science problem, and performing cross-matching or data combining, a process generically described as data fusion (Ball and Brunner, 2010).

A key focus of this work is to use the measurements observed at different wavelengths, often referred to as the multi-wavelength data, available in the COSMOS field to classify the detected radio sources. Combining observations spanning various electromagnetic bands using the positions of the sources (right ascension and Declination) is commonly known as cross-matching. These steps are already completed by Whittam et al. (2022) and *Prescott et al.* [submitted]. The same radio sources seen in various bands by other surveys are cross-matched using their unique identification. If there is no definitive identification for the source, the source's position in the sky with some astrometric tolerance would be used as a standard method for cross-matching Ball and Brunner (2010).

Cross-matching can introduce several challenges, including missing data, fuzzy matches, resolution of sources within or between datasets, differing survey footprints, survey masks, and large amounts of processing time and data transfer requirements when cross-matching large datasets. These challenges can drastically affect the performance of the supervised model. Hence, we must consider *data preprocessing* before applying ML models.

### 2.2.2 Data Preprocessing

Data preprocessing comprises several steps, some of which occur during data collection. It is often problem dependent, and one needs to apply it very carefully since it may affect the performance of the ML models. Data preprocessing is cleaning and organising the raw data to make it suitable for building and training ML models. In this work, we do data preprocessing to transform the data according to the selected algorithms. We will consider three steps to transform MIGHTEE-COSMOS catalogue data; Converting the categorical data into numerical data, handling the missing data and some invalid measurements. There are several steps one needs to consider for data preprocessing. See Pyle (1999), Zhang et al. (2003) for details regarding data processing.

The MIGHTEE-COSMOS catalogue comprises numerical and categorical data. The latter is a class of observed sources, such as an SFG or an AGN. We adopt scalarization, one of the standard methods to convert categorical data to numerical data. Scalarisation is a method in which various possible categorical attributes are given various numerical labels; for example, in a binary classification problem, 'SFG' and 'AGN' are labelled as the vectors [1,0] and [0,1], respectively.

Data will generally contain one or more types of incorrect or misclassified measurements. Common examples in astronomy include instances where the flux of an object at a certain wavelength has been set to -9999 or NaN. The value may be correct for some ML algorithms. Still, it has been flagged as bad, or the value is not bad in a formatting sense but is nonphysical, perhaps a magnitude of a high value that could not have been detected by the instrument (Ball and Brunner, 2010).

We transform the direct measurements in the catalogue (flux densities) to the logarithmic of ratio of two flux densities from different wavelength bands i.e. colours. In astronomy, to derive a colour is an ordinary operation to take a log of a division of flux densities in one waveband by another. These transformations can also introduce numerical issues, such as division by zero or loss of accuracy. Such measurements need to be removed either by simply removing the source containing them, ignoring the whole feature with those values but using the remaining data, or interpolating a value using other information.

Missing data is another common problem in almost all data types. Missing data could be attributed to three reasons; the corresponding survey did not observe a particular source, a source was detected in some wavelength bands but too faint to be detected in others, or there is no emission from that source in that band. Missing data can significantly affect the accuracy of results, attenuating the model's performance. Such datasets cause problems to most of the Scikit-learn (sk-learn, Pedregosa et al. (2011))[1] estimators that assume that all values in an array are numerical and that all have and hold meaning (Zhang et al., 2017). We replace or remove the missing values in the datasets. In this work, we limit our focus to only sources with valid measurements in each input feature used to train the ML algorithms since the fraction of sources with invalid measurements and missing values is less than 5%.

### 2.2.3 Feature Selection

The required input features for analysis are crucial since they will determine how well the ML models can assign class labels to the correct observations. Feature selection reduces the dimensions of the data, and as such, the models perform faster and efficiently (Kotsiantis et al., 2007).

Multi-wavelength astronomy has given rise to a vast range of measurements to use as input features. However, only a subset of features is helpful to a specific problem. Complete observable parameter space includes the source coordinates *(Right ascension and Declination)*, velocities, redshift, proper motions, *fluxes densities at different wavelengths*, surface brightness, and variability over a range of timescales (Zhang et al., 2008). Using all available features to train an ML model results in the so-called *curse of dimensionatily*. High-dimensional datasets cause machine learning models to yield worse outcomes. To avoid such a problem, one requires some dimension reduction, in which one wishes to retain as much information as possible but with fewer input features. The most trivial form of dimension reduction is to use one's judgement and select a subset of features. Another way one can usually take a more sophisticated and less subjective approach, such as principal component analysis (PCA) (Ball and Brunner, 2010).

The MIGHTEE-COSMOS catalogue comprises several measurements from the radio to X-ray part of the electromagnetic spectrum one can use as input features. We refer the reader to chapter 3 for the details regarding MIGHTEE-COSMOS data. This study considers one's judgement approach to feature selection and only the fluxes at different wavelengths used in radio astronomy for classification if SFGs and AGN are considered primary features to train the selected ML models. Other fluxes at other wavelengths have also been considered, and the results are reported in the appendix 5.3.

[1] SK-learn; https://scikit-learn.org/stable/

### 2.2.4 Selection of Algorithms

We adopt and briefly compare classification algorithms from the machine learning field; Logistic Regression (LR), Support Vector Machines (SVM), $k$-Nearest Neighbour ($k$NN), Random Forest (RF) and XGBoost. All these classifiers use some a-priori knowledge from spectroscopy or physical assumptions to deduce the function that maps the parameter space of observables onto the manual class label distribution (Norris et al., 2019). The implementations of these classification algorithms are available, both as free software in sk-learn[2] and *XGBoost*[3]. The quality of the results obtained using these models mainly depends on three factors: the quality of the data set employed in model-building, the care with which adjustable model parameters were chosen, and the evaluation criteria used to report the results of the modelling process (Dreiseitl and Ohno-Machado, 2002).
Each algorithm's technique to learn and make a prediction is discussed below.

#### Logistic Regression

Logistic regression (LR, Cramer (2002)) is a supervised learning classification approach that uses class labels (output) for building and uses a single multinomial logistic *regression model* with a single estimator (Osisanwo et al., 2017). It is like ordinary least squares regression and is one of the most commonly used tools for applied statistics and discrete data analysis. It is primarily applied to the binary classification problem. However, the technique can be extended to a multi-classification problem (James et al., 2013). The algorithm estimates the probability that an output belongs to a particular category based on the given dataset of independent variables (Osisanwo et al., 2017). LR's outcome is the probability, meaning that the predicted outputs are bounded between 0 and 1. The location of the class border is indicated by LR, which also shows how the class probabilities depend on distance from the boundary and how quickly they approach the extremes (0 and 1) as data is more significant.

[2] SK-learn; https://scikit-learn.org/stable/
[3] XGBoost; https://xgboost.readthedocs.io/en/stable/

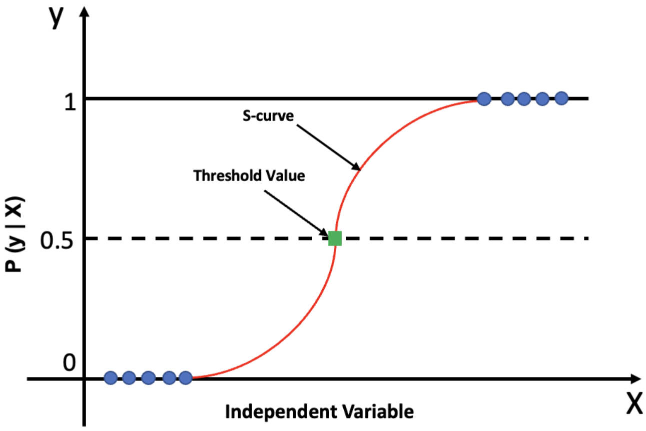


Figure 2.3: An example of the predicted LR probabilities fitted to some input variable X. The blue dots indicate the 0/1 values coded for an input variable X (e.g. one of the derived colours (SFGs or AGN)). X represents the input feature, $P(y|X)$ is the LR model probability; All probabilities lie between 0 and 1. The threshold value defines the probability of either 0 or 1. Values above the threshold value tend to 1, and those below the threshold value tend to 0. The figure was inspired by https://www.javatpoint.com/logistic-regression-in-machine-learning

### Support Vector Machines

The support Vector Machine (SVM, Evgeniou and Pontil (1999)) approach was developed in the 1990s, and its popularity has since been increasing due to the algorithm's potential. The SVM algorithm aims to create a decision boundary referred to as a hyperplane that can distinguish n-dimensional space into classes so a new data point is assigned to the correct category in the future (Cristianini et al., 2000). SVM chooses the extreme points called vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. Figure 2.4 shows two categories that are classified using a hyperplane. Support vector machines are intended for binary classification with two observation classes (James et al., 2013). We refer the reader to James et al. (2013), Cristianini et al. (2000), sk-learn SVM[4] for further reading about the SVM.

[4] SVM; https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

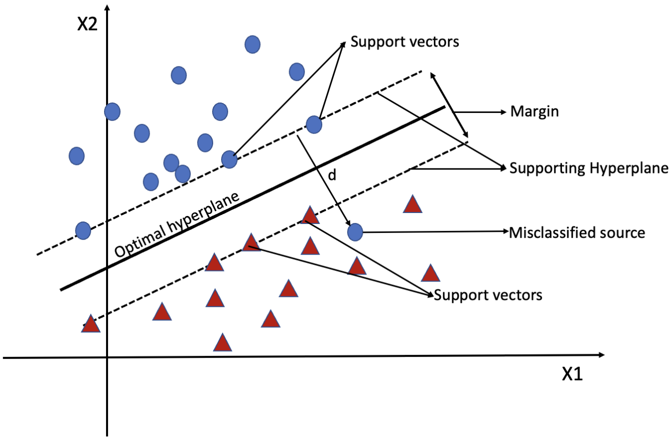


Figure 2.4: The SVM approach: The two classes of observations are illustrated by blue circles and red triangles. The classes are separated by the optimal hyperplane (solid line) parallel to the two supporting hyperplanes (dashed lines) on which the support vectors lie. Margin is the distance from the optimal hyperplane to either supporting hyperplane. The distance d between the supporting hyperplane and the misclassified source indicates "slack variables". Image modified from (Cardoso-Fernandes et al., 2020).

### K-Nearest Neighbour

$k$-Nearest Neighbour ($k$NN, Peterson (2009)) approach follows a different technique compared to the LR and SVM because it uses the data directly for classification, without being modelled first (Dreiseitl and Ohno-Machado, 2002). $k$NN classifier approximates the association between independent variables and the class labels (dependent variables/target) by averaging the observations in the same neighbourhood (Borne, 2013). The $k$NN algorithm decides the class of an unknown observation based on the majority of $k$-nearest neighbour category (Alaliyat, 2008). The size of the neighbourhood ($k$) needs to be set by the analyst or use cross-validation[5] approach to select the $k$-value from an array that yields high performance. $k$NN is considered one of the simple machine learning algorithms. It works based on the minimum distance from the unknown source to the training samples to determine the $k$ nearest neighbours. When $k$ is known, we take the simple majority of these $k$ nearest neighbours to predict the class of unknown observation, as illustrated in figure 2.5.
The algorithm depends on the user's ability to calculate the distance between database objects in any multivariate parameter space. However, the *Euclidean distance* is commonly used to measure neighbourhood (Alaliyat, 2008). The value of $k$ is usually chosen to be the square root of the total number of objects in the training set, and it should be an odd number to avoid the tie.

[5] Cross validation; https://scikit-learn.org/stable/modules/cross_validation.html

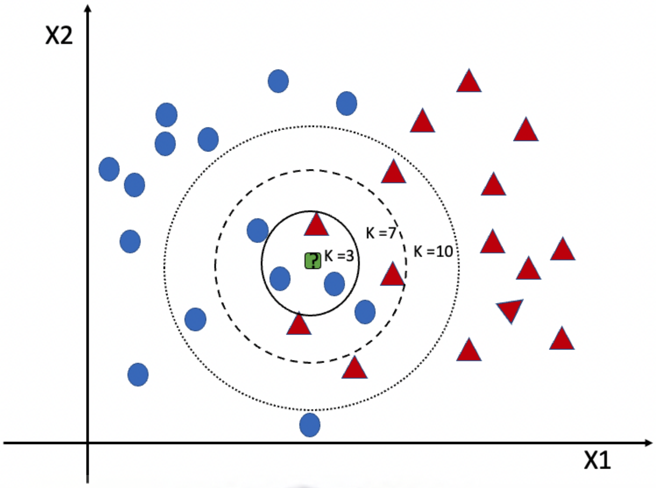


Figure 2.5: $k$NN algorithm illustration. X1 and X2 represent two input attributes. The blue circles and red triangles represent observations in the training set belonging to two classes. The green squares show an unknown observation that the algorithm must classify as one of the shown classes. The small solid circles show where the number of neighbours ($k$) considered to classify the analysed source equals 3. The dashed and dotted circles represent $k$=7, $k$=10, respectively. Image Modified from (Alaliyat, 2008)

### Random Forest

Random Forest (RF, Breiman (2001)) is an ML model that builds an entire forest of random uncorrelated *decision trees* and then uses a voting method to decide on the target variable (Pham et al., 2021). To get the most understanding of how random forest works, one needs to understand how *decision tree* algorithm works, which is not a focus of this work. The RF builds a forest of independent random decision trees to predict the best results. Every decision tree in an RF has special rules, and randomness is added to each tree. As opposed to the decision tree algorithm, which considers all the attributes, the RF only considers a finite number of selected features, hence the term' random' forest. For more details about the random forest model refer to (Breiman, 2001), sk-learn RF[6].

[6] RF; https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

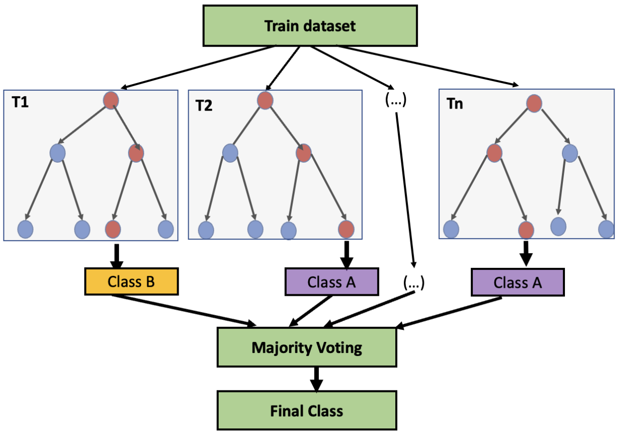


Figure 2.6: Sketch of RF approach. T1, T2, and Tn represent random uncorrelated decision trees from 1 to n. Each decision tree yields its class label from which the most dominant class label would be considered.

### XGBoost

The eXtreme Gradient Boosting (XGBoost, Chen and Guestrin (2016)) model implements machine learning models under the *Gradient Boosting* framework. The XGBoost model has been widely applied in many data science challenges and has shown state-of-the-art results on a wide range of problems (Chen and Guestrin, 2016). XGBoost is an iterative decision tree algorithm with multiple decision trees added via continuous splitting attributes. Every tree is learning from the residuals of all previous trees. Unlike the RF algorithm where most voting output result is adopted, the XGBoost predicted outcome is the sum of all the results from each tree (Wang et al., 2019). Figure 2.7 shows a sketch of the XGBoost classification approach. The different K trees from which each outcome and the residuals are predicted are represented by grey rectangles. The different colours of circles in each tree represent a leaf node associated are obtained after training, and the features of prediction samples in each tree and each leaf node corresponds to a score. The final score, as shown, corresponds to the scores of each tree added up.

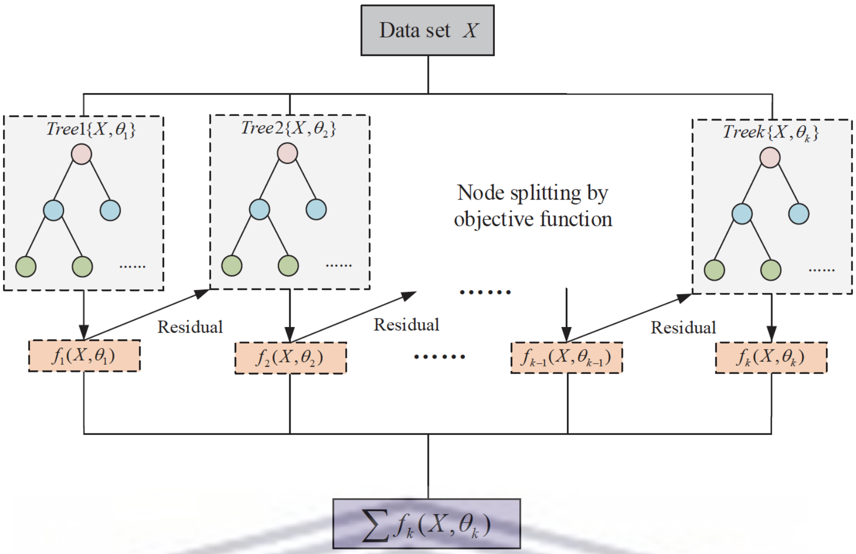


Figure 2.7: Sketch of XGBoost approach. Image adapted from (Guo et al., 2020)

### 2.2.5 Evaluation Metrics

We assess the performance of supervised ML models on the data which was not part of the training data, i.e. the unseen test data using the classification evaluation metrics. Evaluation metrics are used to evaluate and measure the performance of a specified model and can be used to compare to another model. There are several metrics to evaluate the accuracy of a model. These metrics can be used to assess models' performance for both classification and regression. This part will focus on the classification metrics. We adopt the classification metrics *Precision*, *Recall* and *F1-score*. For binary classification, the confusion matrix scheme is shown in table 2.1.

Table 2.1: Confusion matrix for binary classification. Actual represents the actual labelled class (in our case, the manually labelled class, i.e. AGN or SFG), while predicted represents the class predicted by the algorithm.

| | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted | Positive | True Positive | False Positive |
| | Negative | False Negative | True Negative |

If we consider that an AGN class is positive and that of SFG is Negative, then True Positive (TP) refers to the number of sources predicted as AGN and labelled as AGN manually. Similarly, True Negative (TN) refers to the number of sources predicted as SFGs and labelled as SFGs manually. The false Positive (FP) value is

the number of actual SFGs classified as AGN, and the False Negative (FN) value is the number of actual AGN classified as SFGs.
We define Precision, recall and F1 score as follows. *Precision* indicates how good the classifier is when identifying the true positives (TP). A good precision score indicates a high fraction of positive identifications, and a low precision indicates a low fraction of positive identifications.

$$Precision = \frac{TP}{(TP+FP)}$$

The second metric is *recall*, which assesses how well the algorithm minimises false negatives. A low recall for an individual class would indicate it is often misclassified as another.

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

Lastly, The *F1-score* is the harmonic mean of Precision and recall and gives the overall performance of the ML model.

$$F1 = \frac{2TP}{(2TP+FP+FN)}$$

## 2.3 Hyperparameter Optimization

The machine learning models comprise two parameters; *hyperparameters* and the model parameters. Hyperparameters are all the parameters the user can arbitrarily set before training the model; this includes parameters such as n-neighbours in $k$NN and n-estimators in RF. In contrast, model parameters are learned during the model training. Hyperparameters determine how our model is structured before training, and tuning them is an optimisation problem.
Currently, it is most likely that there is no best set of hyperparameters or standard fixed that will yield the best results. Finding the best collection of hyperparameters is still a try-and-catch process. Two common approaches to search for the best set of hyperparameters from which the model will yield good performance are *gridsearch parameter tuning and random search parameter tuning.* Grid Search is a technique that will try all possible parameter combinations and return a model with great accuracy. Grid search will build a model, evaluate it in every variety possible and choose the most accurate one by the end. Random Search is a technique used to generate random combinations of the hyperparameters to find the best fit for the model. Random Search uses a subset of random parameters from which the classifier that yields great accuracy scores will be considered.
In this work, we perform a k-fold cross-validation while trying to find the optimal hyperparameters. Cross-Validation (CV) is a data resampling method to assess the

generalization ability of predictive models and to prevent overfitting (Berrar et al., 2019). It works by splitting the data into k-folds (subsets or partitions) where each fold is used for testing the model while the remaining k-1 folds are used for training the model. For example, if k=3, then a dataset will be divided into 3 folds, one fold will be used as test data and the remaining two folds as training data which implies that about 33% of the data is test data and 67% is training data. This process is repeated until each fold is used as test data. In the end, three scores will result from each iteration and the mean of the scores will then be reported as a final score. We extend the CV by doing a 3-fold grid search CV while attempting to determine the optimal hyperparameters. All hyperparameters are used to train many models, from which the model with the greatest score has the best hyperparameters, which are referred to as optimal hyperparameters. We have highlighted the different sets of parameters for each model in section C of the appendix. Figure 2.8 depicts the three-fold cross-validation used in this work for training and for the hyperparameter selection of the model.

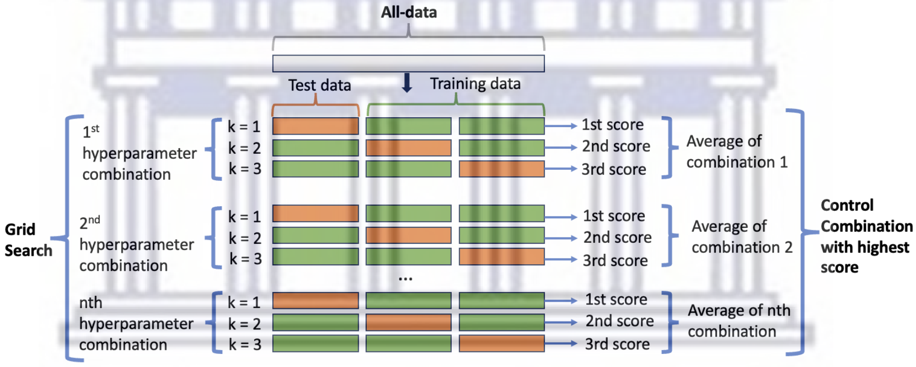

All-data
Test data
Training data
Grid Search
1st hyperparameter combination
2nd hyperparameter combination
nth hyperparameter combination
k = 1
k = 2
k = 3
1st score
2nd score
3rd score
Average of combination 1
Average of combination 2
Average of nth combination
...
Control Combination with highest score

Figure 2.8: Three-fold cross validation hyperparameter tuning using GridSearch method. Image inspired by (Shatnawi et al., 2022)

## 2.4 Machine Learning Application in Astronomy

The recent exponential increase in astronomical data necessitates the need for powerful computing capabilities, the use of data mining and the employment of machine learning techniques which can quicken the tedious work that astronomers go through. The application of machine learning in fundamental astronomical research spans seven main categories of activity: classification, regression, clustering, forecasting, generation, discovery, and the development of new scientific insight (Fluke and Jacobs, 2020). Classification is a common first step in the scientific process because it organises data so that hypotheses can be generated and data can be compared to models.

This section reviews some of the applications of supervised machine learning classification techniques in astronomy to classify stars, galaxies and AGN.

Astronomy has many classification problems, including the classification of star-forming galaxies and active galactic nuclei. One of the earliest classifications of galaxies using machine learning began in the 1980s when Whitmore (1984) used principal component analysis (PCA) for the morphological classification of spiral galaxies. In the early 1990s, astronomers began to use more complex methods that required labelled training sets. The first or early algorithms to be applied for classifying stars or galaxies are decision trees (Weir et al., 1995) and artificial neural networks(Odewahn et al., 1992). By the 2000s, the technique was widely used, and RF began to dominate (Fluke and Jacobs, 2020). Li et al. (2019), Golob et al. (2021), and Hughes et al. (2022) makes use of the XGBoost algorithm to classify the celestial objects. Fadely et al. (2012) applied the SVM to distinguish stars from galaxies. Since supervised machine learning classification algorithms have been widely adapted to classify celestial objects (Sen et al., 2022). Li et al. (2008) uses the $k$NN algorithm for the classification of multi-wavelength astronomical objects. Other studies such as Bai et al. (2019), Xiao-Qing and Jin-Meng (2021), Peng et al. (2013), Hughes et al. (2022) explore various supervised algorithms for the classification of celestial objects.

Machine learning algorithms are efficient and have been providing amazing results in this field. Clarke et al. (2020) uses the trained and optimised RF to classify the 111 million sources in the SDSS photometry catalogue with no spectroscopic observations as stars, galaxies, or quasars. Cavanagh et al. (2021) makes use of deep learning methods to predict the morphologies of galaxies. They train and test several convolutional neural networks (CNN) architectures to classify the morphologies of galaxies in both a 3-class (elliptical, lenticular, and spiral) and a 4-class (+irregular/miscellaneous) schema with a data set of 14 034 visually classified SDSS images. The Kim and Brunner (2017) model has performed over 99% accuracy for classifying stars and galaxies. Peng et al. (2013) integrated $k$-NN into SVM and created a new method called SVM-$k$NN. A newly created model SVM-$k$NN reached much higher performance, and the performance was slightly improved than that of SVM alone. For better reviews of machine learning applications in astronomy, we refer the reader to read Ashai et al. (2022), Sen et al. (2022), Ball and Brunner (2010), Borne (2013).

Much effort has been made to distinguish/ separate the stars, galaxies and AGN. However, the classification of radio sources in the radio continuum is still a significant classification problem. Various algorithms have been explored, and impressive results have been achieved to classify objects as galaxies, Stars and Quasars. Achieving better results in the radio continuum domain requires scrutinising suitable models to make the end prediction with the utmost accuracy possible. In the radio continuum domain, most of the classification tasks have been carried out manually by experts in the field.

In this study, we adopt five machine-learning techniques; LR, SVM, $k$NN, RF and

XGB to classify sources in the faint radio sky. We compare the scores of all algorithms and make use of the *Jack Knife* method (Miller (1974)) to estimate each model's uncertainty. The jackknife also known as the delete one is a method used to estimate the variance and bias of a large population, and it is still not commonly applied in the astronomical domain. It works by repeatedly estimating a statistic (F1-score in this work) of interest by systematically leaving out one observation from the dataset at a time and recalculating the statistic. This process is performed for each observation in the dataset, resulting in a collection of estimated statistics. This method has been used by Hussein et al. (2021) in the classification of 51 fermented (FR) and 47 unfermented (UFR) rooibos samples after extraction using water and methanol. Since in most radio surveys, the number of SFGs and AGN are not equal. Usually, the SFGs are more than the AGN. We have tried to balance the number of SFGs and AGN in the training data and evaluate the trained models to determine whether the ML models will better classify the majority population. We use the LR model as a baseline algorithm for this study. LR is among the simplest, interpretable algorithms; however, they are not widely adopted by many studies in astronomy. As such, we wish to find out how sophisticated algorithms perform when compared to LR. We aim to produce a comprehensive machine-learning pipeline to classify the radio sources in the faint radio sky.

# Chapter 3

# MIGHTEE-COSMOS Data

This study uses early science radio continuum data from the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) survey and rich ancillary data in the COSMOS field to explore the physical processes of the radio emission from these extragalactic survey. The MIGHTEE survey has the potential to significantly advance our understanding of galaxy evolution due to its exceptional depth over a wide area and great multiwavelength coverage. The COSMOS field is one of the well-studied sky areas, close to the celestial equator, and is visible by all astronomical facilities. Bright radio, ultraviolet (UV), and X-ray sources are absent from the COSMOS, which makes it an excellent sky area for deep surveys.

The sections in this chapter are organized as follows. Section 3.1 briefly introduces the MIGHTEE survey. We give the details and an overview of the COSMOS field in section 3.2. Section 3.3 details the radio data catalogue of the COSMOS field undertaken with the MeerKAT telescope. In section 3.4, we give the details of the MIGHTEE COSMOS data used in this work and the equivalent multiwavelength data. Section 3.5 details the conventional classification techniques employed to classify MIGHTEE-detected radio sources. The last section briefly discusses the limitations of conventional techniques employed to produce the MIGHTEE-COSMOS catalogue labels.

## 3.1 The MIGHTEE Survey

The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE, Jarvis et al. (2016)) is a project currently undertaken by a South African-led international collaboration of researchers to explore cosmic evolution of galaxies and active galactic nuclei by using radio continuum, polarisation, and spectral line data taken by MeerKAT Telescope. MIGHTEE covers four well-studied extragalactic deep fields, i.e., the Cosmological Evolution Survey (COSMOS) field, the Extended Chandra

Deep Field Source (E-CDFS), the European Large Area Infrared Survey field South 1 (ELAIS-S1), and the XMM-Newton Large Scale Structure field (XMM-LSS), making up to 20 deg$^2$ deep fields with a central frequency of 1284 MHz (Jarvis et al., 2016) observed with MeerKAT's L-band (870 – 1670 MHz) receivers for about 1000 hours. The fields were selected because there are best-studied extragalactic regions with extensive multiwavelength data from existing surveys and potential future surveys (Heywood et al., 2022b).

The MIGHTEE survey goals include enhancing our understanding of; 1) active galactic nuclei and star-formation activity progression across cosmic time as a function of stellar mass and environment, free of dust obscuration, 2) the evolution of neutral hydrogen in the Universe, how it passes through the molecular phase to become stars, and how effectively this might feed active galactic nuclei activity, and 3) the properties of cosmic magnetic fields and how they evolve in clusters, filaments and galaxies (Jarvis et al., 2016).

The total intensity continuum images will reach the classical confusion limit of MeerKAT at a depth of 2 $\mu$Jy/beam sensitivity at Giga-Hertz frequencies (Jarvis et al., 2016). The MIGHTEE survey has the potential to significantly advance our understanding of galaxy evolution due to its exceptional depth over a wide area and great multiwavelength coverage (Whittam et al., 2022). By Adopting supervised machine learning techniques, this work uses MIGHTEE Early Science observations in the COSMOS field (Heywood et al. (2022b)) to Study Star Formation and Black Hole Accretion.

## 3.2 The COSMOS Field

The Cosmic Evolution Survey (Scoville et al., 2007, COSMOS) aims to investigate the associated evolution of galaxies, star formation, active galactic nuclei (AGN), and dark matter (DM) with large-scale structure. The survey combines Hubble Space Telescope (HST)[1] imaging, multiwavelength imaging, and spectroscopy spanning a two deg$^2$ area from X-ray to radio wavelengths (Scoville et al., 2007). An illustration of the COSMOS field relative to the full moon and the Galaxy Evolution from Morphologies and SEDs (GEMS), the Great Observatories Origins Deep Survey (GOODs), and Hubble Ultra-Deep Field (HDUF) fields is shown in Figure 3.1. To ensure visibility by all astronomical facilities, the COSMOS field is close to the celestial equator. The field must be easily viewable by all major optical/IR telescopes due to the time needed for deep imaging and spectroscopy over a total area of 2 deg$^2$ containing more than a million galaxies. No prominent radio, UV, or X-ray sources are in the field. The Galactic extinction in COSMOS is remarkably low and homogeneous compared to other equatorial fields. More details about the COSMOS project

[1] HST; https://hubblesite.org

are available at https://cosmos.astro.caltech.edu/.

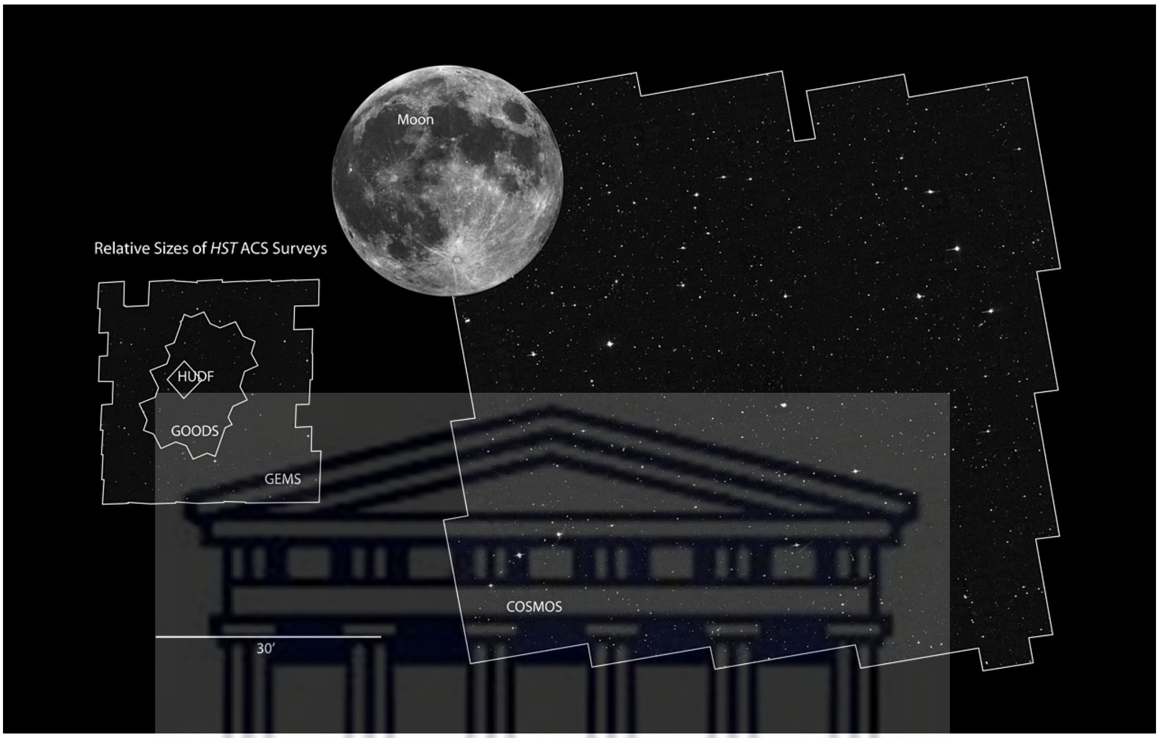


Figure 3.1: Illustration of the COSMOS field (right) relative to the full moon and the GEMS, the GOODs, and HDUF fields. Image adapted from https://cosmos.astro.caltech.edu/page/public

There are numerous observations of the COSMOS field from space and ground-based telescopes, see Figure 3.2 for examples of the COSMOS field's multiwavelength coverage. The HST conducted various observations with different bands in the COSMOS (Weaver et al., 2022), and to date, COSMOS is one of the largest sky areas ever surveyed with the HST. In 2006, some of the first broad- and narrow-band ground-based observations with the Subaru Supreme-Cam were made over the entire region, resulting in one of the most extensive imaging data sets at the time (Capak et al., 2007). Using the Spitzer Space Telescope[2], mid-infrared images of the whole COSMOS field were also made (Sanders et al., 2007). The other major astronomical facilities pointed to the COSMOS field include the Galaxy Evolution Explorer (GALEX)[3], the Herschel Space Observatory[4], the Chandra X-ray Observatory[5], and

[2] https://www.spitzer.caltech.edu

[3] http://www.galex.caltech.edu

[4] https://sci.esa.int/web/herschel

[5] https://chandra.harvard.edu

the XMM-Newton satellite[6]. COSMOS now possesses one of the largest, most comprehensive, standardized data sets throughout the electromagnetic spectrum, from X-ray to radio. We refer the reader to Weaver et al. (2022) for more details on the photometric catalogues available in the COSMOS field.
The COSMOS field is one of the best-studied regions of the sky, with an incredible multiwavelength data archive. This work uses the multiwavelength COSMOS data complied by *Prescott et al.*, [submitted]. The data comprises the near-infrared YJHK-[7]-band data from the UltraVISTA Survey, optical $u^*$-band measurements from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS)[8], grizy-band data from Hyper Supreme Cam (HSC) images, and grizy-band data, 3.6 and 4.5-µm infrared data from the Spitzer Extended Deep Survey.

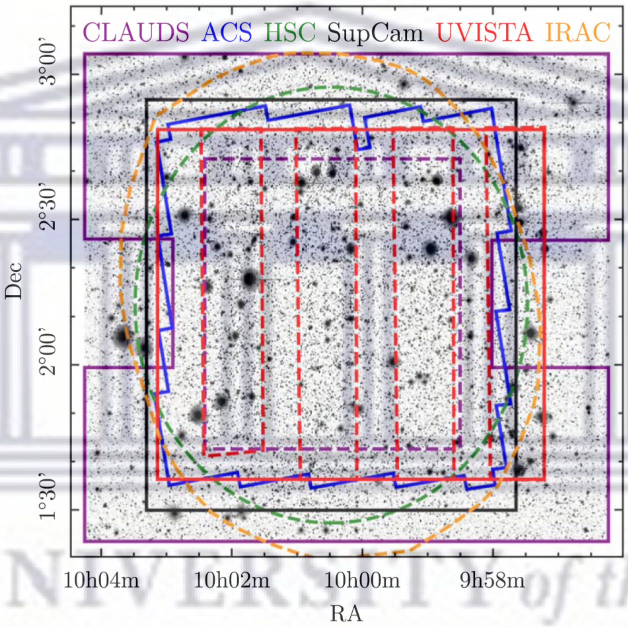


Figure 3.2: Examples of the COSMOS field's multiwavelength coverage. The background image matches the detection image from izYJHK wavelength bands. The dotted lines indicate the deepest portions of the image, while the solid lines denote the survey limits. Image is taken from (Weaver et al., 2022).

[6] https://www.cosmos.esa.int/web/xmm-newton

[7] YJHK-UGRIZ: is different atmospheric transmission windows centred at different wavelengths in the near-infrared and optical part of the electromagnetic spectrum respectively, for further reading we refer the reader to visit Bessell (2005)

[8] CFHTLS; https://www.cfht.hawaii.edu/Science/CFHLS/

## 3.3 MIGHTEE-COSMOS 1.3 GHz Data

The COSMOS field's MIGHTEE Early Science radio continuum data release covers 1.6 deg$^2$ and is described by Heywood et al. (2022b). The COSMOS was observed for 17.45 hours between 2018 and 2019 using MeerKAT's L-band (900–1670 MHz) receivers with a single field of view centred on RA 10h00m28.6s, Dec +02d12m21s. The effective frequency across the MIGHTEE data is not constant. However, it gradually decreases from the centre to the edges since MeerKAT's L-band receivers are wide bandwidth and the primary beam response changes with frequency. The MIGHTEE flux densities and radio luminosities are scaled to 1.4 GHz using the effective frequency map published by Heywood et al. (2022b), assuming a spectral index of -0.7. This ensures that observations are at a constant frequency. All radio luminosities are k-corrected using the same assumed spectral index.

The MIGHTEE-COSMOS Early Science data comprises two versions of the images processed with Briggs' (1995) robust weighting values of 0.0 and -1.2. The Briggs' robustness parameter 0.0 is optimised for sensitivity but has a lower resolution. In contrast, Briggs' robustness parameter -1.2 results in higher resolution and lower sensitivity images (Heywood et al., 2022b). The maximum-sensitivity image, shown in Figure 3.3, has a circular synthesised beam size of 8.6 arcsec and thermal noise of 1.8 Jy/beam. It is limited by classical confusion at approximately 4.5 µJy/beam. As it contains more sources than a high-resolution image, this was used as the primary radio source catalogue by *Prescott at al.* [submitted]. The Whittam et al. (2022) study provides the classification labels of the catalogue using conventional astronomical techniques.

This work is restricted to the central part of the COSMOS Early Science image, with a diameter of 0.9 degrees. This is the region over which multiwavelength counterparts for the radio sources have been identified and was a focus of Whittam et al. (2022) study. The radio measurements from this region were cross-matched with the X-ray, Optical, near-, mid-, and far-infrared observations from other surveys using the right ascension and declination. *Prescott at al.* [submitted] and Whittam et al. (2022) provide a comprehensive description of how cross-matching sources have been achieved. The Python Blob Detector and Source Finder (PyBDSF; Mohan and Rafferty (2015)) source with the default source extraction parameters were used to find the radio sources from the images. The initial sample contains 6,263 radio sources with $S_{1.28\,\mathrm{GHz}} > 20$ Jy in the central part of the COSMOS field. For further details regarding the source finding, radio flux measurements, optical spectra and redshifts, refer to Heywood et al. (2022b), Whittam et al. (2022) and *Prescott at al.* [submitted].

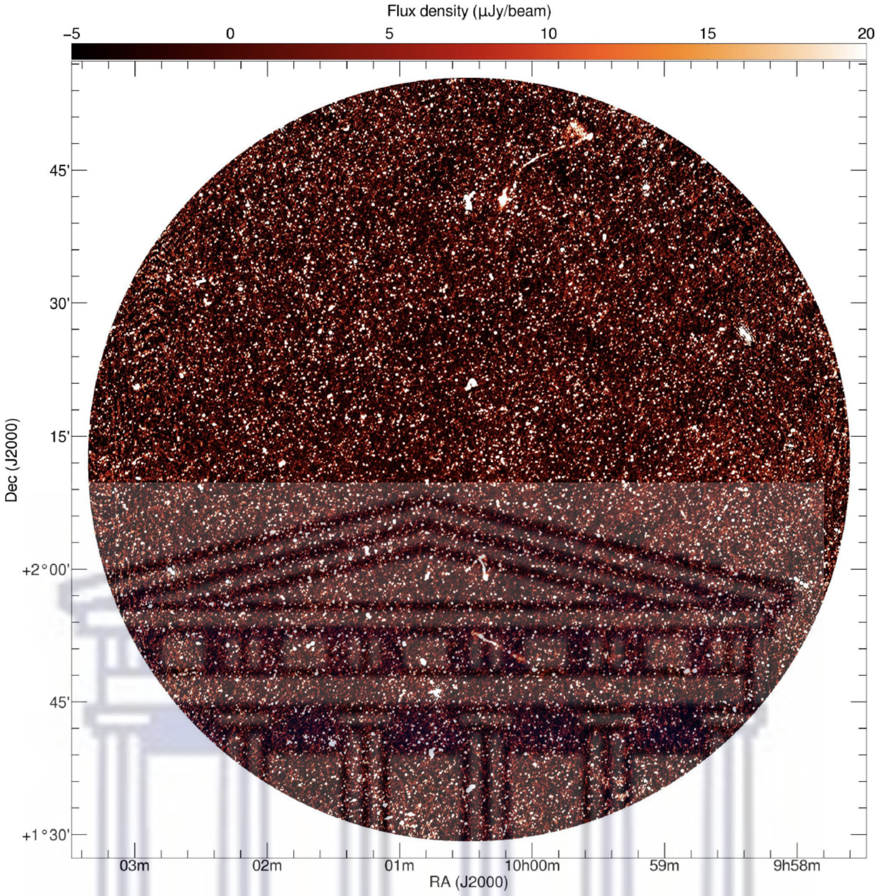


Figure 3.3: The Early Science image of the MIGHTEE-COSMOS field. The image covers 1.6 deg$^2$ with nearly 10,000 radio components with peak brightness that is greater than the $\sigma_{\mathrm{local}}$ (Heywood et al., 2022a).

Figure 3.4 shows the radio luminosity as a function of redshift for sources in the MIGHTEE-COSMOS. Investigating the universe up to redshift 5 using this sample is possible. Using the MIGHTEE survey, we can probe low-powered radio galaxies across cosmic time, providing new insight into AGN activity (Whittam et al., 2022). The top subplot of Figure 3.4 shows the redshift distributions of the MIGHTEE-COSMOS samples. About 45 per cent of the sources have a spectroscopic redshift value from the Deep Extragalactic Visible Legacy Survey (DEVILS; Davies et al. (2018)). In contrast, the remaining sources have photometric redshift estimates. Photometric redshifts are estimated using two techniques; machine learning and template fitting. Template fitting estimates were computed using the Spectral Energy Distribution (SED) template fitting code known as Le Phare (*Prescott et al.* [submitted]). In contrast, the machine learning approach uses the GPz algorithm to compute redshifts (Almosallam et al., 2016).

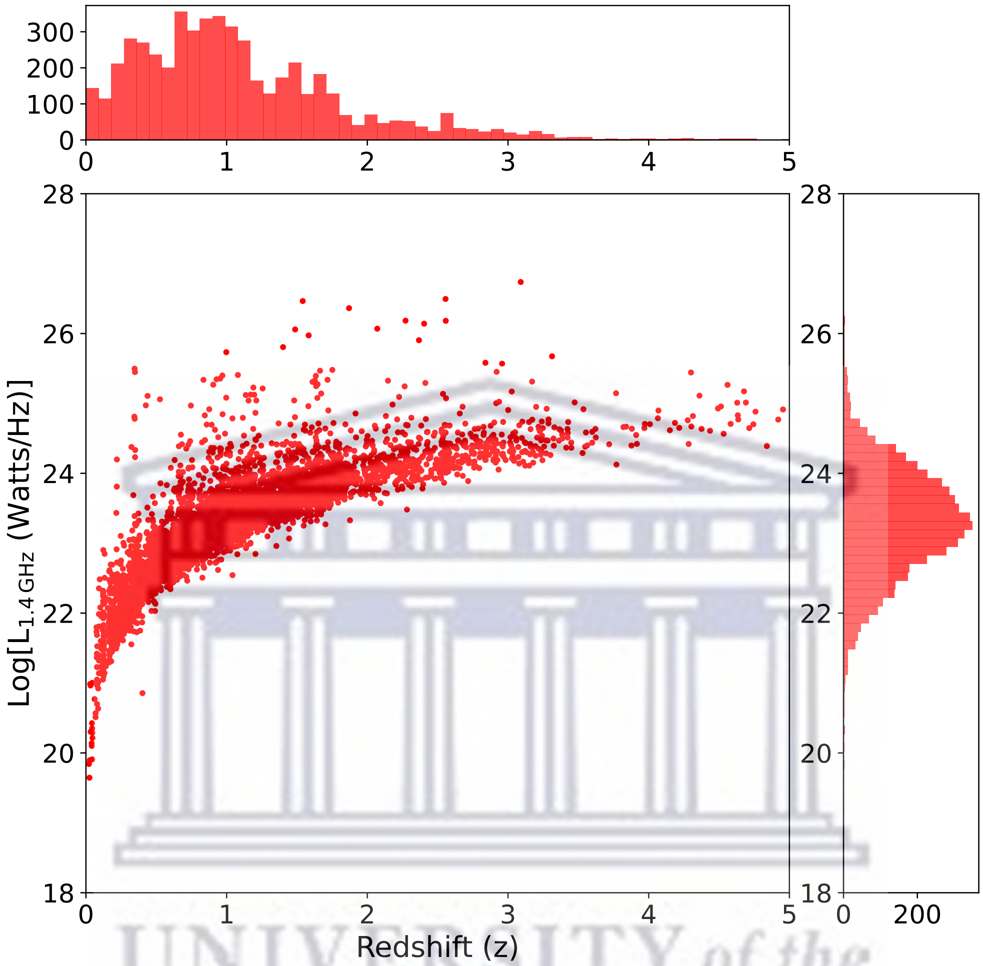


Figure 3.4: The 1.4-GHz radio luminosity-redshift distribution of the well-matched MIGHTEE-COSMOS sample. The main plot represents the radio luminosity of sources as a function of photometric and spectroscopic redshift. The top subplot depicts the redshift distribution of the MIGHTEE-COSMOS sources, and the right subplot shows the distribution of 1.4-GHz radio luminosity. Figure inspired by *Prescott et al.* [submitted]

## 3.4 MIGHTEE-COSMOS Multiwavelength Counterparts

The MIGHTEE-COSMOS multiwavelength catalogue comprised radio and matched measurements from optical, near-, mid- and far-infrared, and Xray information for the radio sources in the central part of the MIGHTEE Early Science Data in the COSMOS field (*Prescott et al.* [submitted]) and is assembled as follows.

The MIGHTEE-COSMOS radio catalogue was produced by Heywood et al. (2022b), and a brief introduction to the data is discussed in section 3.3. The radio catalogue was cross-matched with the optical and near-infrared counterparts adopted from Bowler et al. (2020) catalogue by *Prescott et al.* [submitted]. The optical and near-infrared catalogue comprised the near-infrared imaging in the *YJHKs* band; the optical measurements in the *grizy* bands from Hyper Suprime-Cam Subaru Strategic Program (HSC SSP); and deep optical imaging from CFHTLS's $u^{*}griz$ bands. The method followed to identify the host galaxy of each radio source is available at *Prescott et al.,* [sumitted] study and is also summarised by Whittam et al. (2022). The host galaxy of 5,224 radio sources has been identified. Thus a sample of 5,224 radio sources had an optical and near-infrared counterpart. This sample was further cross-matched with measurements from other surveys by Whittam et al. (2022) as follows. Whittam et al. (2022) used the positions of the optical host galaxies to find X-ray counterparts to the MIGHTEE-COSMOS radio sources using the optical and infrared counterpart of the Chandra COSMOS-Legacy survey catalogue presented in Marchesi et al. (2016). Of 5,224 radio sources, 572 ($\sim$ 10%) were detected in X-ray observations.

The Mid-Infrared (MIR) counterparts were then added to the catalogue. The MIR data was taken from the COSMOS2015 (Laigle et al., 2016) catalogue. This catalogue provides the SPLASH_1_FLUX (ch1), SPLASH_2_FLUX (ch2), SPLASH_3_FLUX (ch3), SPLASH_4_FLUX (ch4) which are 3.6 µm, 4.5 µm, 5.8 µm, 8.0 µm fluxes respectively. Often the 3.6 µm, 4.5 µm, 5.8 µm, 8.0 µm will be referred to as the IRAC1, IRAC2, IRAC3, IRAC4 flux densities and the colours derived from these measurements will be referred to as the IRAC (The Infrared Array Camera) colours. The Herschel Extra-galactic Legacy Project (HELP; Vaccari (2015)) provided the far-infrared data. The observations come from the Multiband Imaging Photometer (MIPS) instrument on the Spitzer Space Telescope, the Photodetector Array Camera and Spectrometer (PACS) on Herschel, and the Spectral and Photometric Imaging Receiver (SPIRE) on Herschel. The MIPS provide 24 µm data, PACS produce 100 µm and 160 µm data, and SPIRE provides the 250 µm, 350 µm, and 500 µm data. Four thousand five hundred forty-one radio sources are identified in the MIPS and PACS data, and 4,958 in the SPIRE data.

## 3.5 MIGHTEE-COSMOS Classification

Using the well-matched MIGHTEE-COSMOS multiwavelength catalogue, Whittam et al. (2022) then classified the MIGHTEE-COSMOS radio sources as SFGs or AGN using four conventional classification techniques; infrared-radio correlation ($q_{\mathrm{IR}}$), mid-infrared colour-colour, optical morphology and X-ray luminosity including the VLBI criteria (Table1, Whittam et al. (2022)). Figure 3.5 shows the completeness of the overall classification of the MIGHTEE-COSMOS catalogue using conventional techniques. A radio source that meets one or more AGN criteria is considered an AGN. A source that could be securely classified as not being an AGN using all four criteria is classified as an SFG. The X-ray criterion is limited to classifying sources with redshifts $z < 0.5$. Another additional classification of 'probable SFG' was introduced for sources with redshifts $z > 0.5$ that cannot fulfil the 'not X-ray AGN' criteria but are classified as 'not AGN' using the other three criteria. For more details regarding the MIGHTEE-COSMOS catalogue and classification criteria, we advise the reader to visit Whittam et al. (2022). The four conventional techniques used are described below.

### 3.5.1 Mid-Infrared/ The Infrared Array Camera (IRAC) Colors

The Mid-infrared colours defined by Donley et al. (2012) are used to identify galaxies that display the power-law emissions from the torus as mid-infrared AGN (midIRAGN). The region is defined on the diagram of the ratio $S_{8.0\mu\mathrm{m}}/S_{4.5\mu\mathrm{m}}$ versus $S_{5.8\mu\mathrm{m}}/S_{3.6\mu\mathrm{m}}$. Any radio source that falls within this region is regarded as a midIRAGN. (See (Donley et al., 2012)).

### 3.5.2 X-ray

X-ray observations are considered accurate and beneficial for identifying AGN. This is because some of the brightest AGN display characteristic accretion-related X-ray emissions. An observed galaxy is considered to be an X-ray AGN (XAGN) if it's unobscured (0.5 – 10 keV) X-ray luminosity $L_x \geq 10^{42}$ erg/s (Szokoly et al., 2004). X-ray observation can be more complex as some X-ray emissions are absorbed and reprocessed into infrared emissions by dust.

### 3.5.3 Infrared – Radio Correlation ($q_{\mathrm{IR}}$ Parameter)

The infrared–radio correlation is used to identify galaxies with more radio emissions that can not be understood as a sum of the emissions of its stars. The sources which display a radio excess above what would be expected from star formation alone are called radio-loud AGN (RLAGN). This relation is quantified as $q_{\rm IR}$ and is defined as the logarithmic ratio of the infrared and radio luminosities:

$$q_{IR} = \frac{log(L_{IR}[\mathrm{W}]/3.75 \times 10^{12}[\mathrm{Hz}])}{log(L_{1.4\mathrm{GHz}}[\mathrm{WHz}^{-1}]})\tag{3.1}$$

In (Whittam et al., 2022) classification, the $L_{IR}$ is the total infrared luminosity between 8 - 1000 µm, estimated by AGNfitter. This is divided by a nominal central frequency of 3.75 × $10^{12}$Hz, corresponding to a nominal wavelength of (80µm), so that $q_{\rm IR}$ is dimensionless (See also Delvecchio et al. (2021)).

### 3.5.4 Optical Morphology

The last AGN diagnostic considered separating the star-forming galaxies from active galactic nuclei using optical morphologies. In optical imaging, the emission from the AGN nucleus outshines the whole galaxy system and becomes a point-like source. For MIGHTEE, they used Hubble Space Telescope (HST) Advanced Camera for Surveys (ACS) -band image. Any source with the compactness of ≥ 0.9 in Source-Extractor (SExtractor), i.e class_star ≥ 0.9 would be point-like in the image as an optical AGN. Else sources with class_star < 0.9 are classified as 'not optical AGN'.

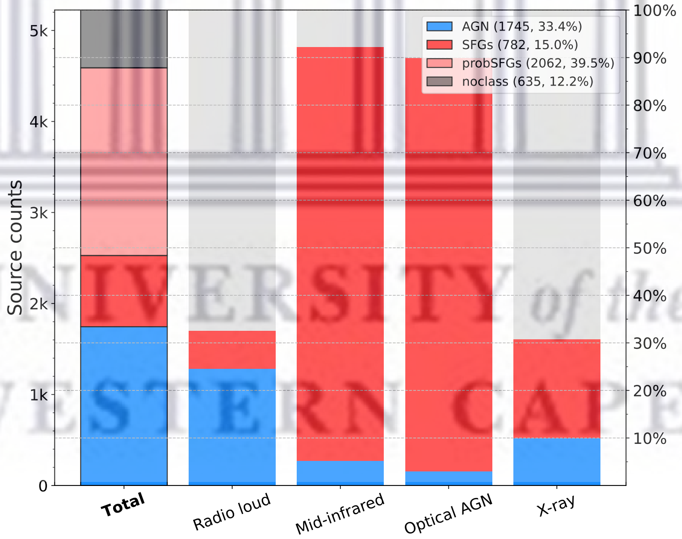


Figure 3.5: A bar plot showing the completeness of MIGHTEE-COSMOS overall classification by different criteria. The first bar shows the total MIGHTEE sources classified as AGN (blue), SFGs (red), probSFGs (light red) and unclassified sources (grey). The remaining bars represent the sources that are classified as AGN (blue), not AGN (red) and not classified (light grey).

This work uses a MIGHTEE-COSMOS multiwavelength catalogue which is slightly different from the published version by Whittam et al. (2022). The only difference between the two versions of the catalogues is that the final catalogue was updated during the course of our work to include the VLBI classification criteria (refer to Whittam et al. (2022)), which slightly changed the number of radio sources per class. However, the available multiwavelength measurements between the two versions of the catalogue are the same only a slight change in the number of classified radio sources. This slight difference is not likely to affect the classification pipeline we aim to develop in this work; hence we stick to the previous version of the catalogue. Table 3.1 summarises each corresponding class's total number of sources.

Table 3.1: Number of MIGHTEE-COSMOS sources per class

| Overall Class | Number of Sources |
|---:|---:|
| AGN | 1,745 |
| SFG | 782 |
| probSFG | 2,062 |
| Unclassified | 635 |

Of all the available measurements in the catalogue, this work focuses on only a subset of the MIGHTEE-COSMOS multiwavelength catalogue described below. This study uses the four IRAC flux densities in 3.6 µm, 4.5 µm, 5.8 µm, and 8.0 µm wavelengths, L14, LIR_WHz, $M_\star$, class_star, and $q_{\mathrm{IR}}$. L14 is the radio luminosity at 1.4GHz. The $L_{\mathrm{IR}}$ is the total infrared luminosity between 8 - 1000 µm, estimated by AGNfitter. $M_\star$, class_star, and $q_{\mathrm{IR}}$ are total stellar mass, optical morphology and the infrared–radio correlation, respectively. Chapter 4 discusses the various preprocessing techniques applied to these features, including the three IRAC colours derived. We build the automated classification pipeline based on these features. However, we have tried other input features, such as the Near and far infrared and optical physical properties (the results are shown in the appendix) found that these additional measurements show no improvements to the performances of the machine learning models.

## 3.6 Conventional Techniques Challenges

In this section, we select two conventional classification criteria; MIR colour cuts and the $q_{\mathrm{IR}}$-parameter mentioned in section 3.5 to illustrate that each conventional technique can only yield a reasonable classification for the specific subclass of AGN. However, the method dramatically fails when used with other AGN subclasses. For example, a class of AGN predicted using the $q_{\mathrm{IR}}$-parameter can not be constrained within the mid-infrared AGN region and vice versa.

Figure 3.6 shows the MIR colour cuts (left) and the $q_{\mathrm{IR}}$-parameter (right) used to classify MIGHTEE-detected radio sources as SFGs or AGN. In both plots, all sources detected in the MIGHTEE survey are shown in blue circles, and the mid-infrared AGN (midIRAGN) are represented in red squares. The optical AGN (optAGN), radio-loud AGN (RLAGN), and X-ray AGN (XAGN) are depicted in magenta crosses, orange triangles and green stars, respectively. Only midIRAGN fits well within the *donley-wedge* (shown by a solid black line) in the left plot. In contrast, other sources classified as AGN by other methods do not fall within this region. Similarly, the plot on the right shows the Delvecchio et al. (2021) + $3\sigma$ fit at redshift (z) equal to one, below which sources are considered as RLAGN. Only RLAGN meets this requirement. Other AGN classes, however, do not adhere to this pattern.

Figure 3.6 gives a clear picture of classifying radio sources using conventional techniques. It is demonstrated that more than one method is required to securely classify a radio source as an AGN. One has to constantly adapt various techniques to accurately classify an AGN, which is a severe concern when faced with the massive amount of data or billions of sources to classify.

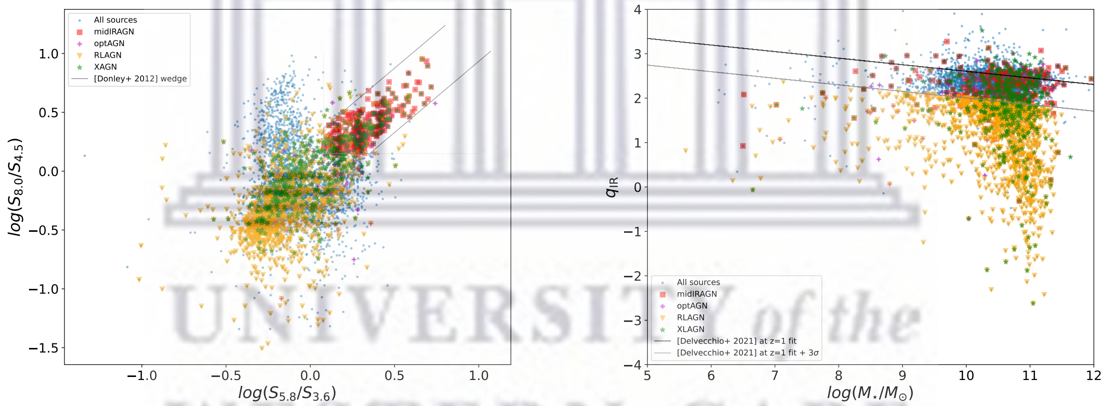


Figure 3.6: The MIR colour cuts (left) and the $q_{\mathrm{IR}}$-parameter (right) diagram. The solid black line on the left plot shows the Donley wedge (midIRAGN) region. The solid black and grey lines on the right plot show the Delvecchio et al. (2021) fit and Delvecchio et al. (2021) + $3\sigma$ at z=1. All MIGHTEE detected sources, midIRAGN, optAGN, RLAGN, and XAGN, are represented by blue circles, red squares, magenta crosses, orange triangles and green stars.

# Chapter 4

# Classification of SFGs and AGN with ML

This chapter details the results of applying machine learning (ML) algorithms to classify the MIGHTEE-COSMOS radio sources as Star-Forming Galaxies (SFGs) or Active Galactic Nuclei (AGN). We adopt the five supervised ML methods detailed in Chapter 2. The MIGHTEE-COSMOS catalogue considered in this study has been introduced in Chapter 3.

We adopt the conventional diagnostics of classifying SFGs and AGN as ML models' input features. The six features adopted from conventional classification indicators are the far-infrared-radio correlation parameter ($q_{\mathrm{IR}}$), the optical morphology (class_star), the total stellar mass ($\mathrm{M}_{\star}$), and the three IRAC colors; log(S8/S45), log(S58/S36), and log(S45/S36). The IRAC colour; log(S45/S36) is derived from the ch2 and ch1 of the IRAC flux densities introduced in section 3.4. Since MIGHTEE-COSMOS multiwavelength catalogue spans several electromagnetic bands, which implies several input features available for training ML models, we also try to add more features, such as optical and near-infrared magnitudes and colours, to test the performance of the ML models. The results of these tests are presented in Appendix B.2.

We begin this chapter by selecting the input features to train ML algorithms for classifying SFGs and AGN from the MIGHTEE-COSMOS catalogue. We describe the details of feature selection in Section 4.1. Sections 4.2, 4.3, 4.4, 4.5 and 4.6 detail the results of applying various techniques to assess the importance of the selected features. In section 4.7, we discuss and investigate the effect of unbalanced data towards the model's classification for the two classes of radio sources under study. We conclude the section by showing the results of applying and comparing various ML techniques on the derived feature combinations in section 4.8.

## 4.1 Feature selection and analyses

The MIGHTEE-COSMOS radio catalogue used in this work constituted 6,263 radio sources, 5,224 of which have optical and near-infrared counterparts in the COSMOS multi-wavelength catalogue compiled by Bowler et al. (2020). The 5,224 resulting sources were cross-matched with the mid-infrared and X-ray measurements from COSMOS2015 and the Chandra COSMOS-Legacy project by Whittam et al. (2022). The MIGHTEE-COSMOS multiwavelength catalogue was then classified manually by Whittam et al. (2022), and 1,745 sources are classified as AGN, 2,062 as probable SFGs, and 782 as SFGs. In this work, we group probable SFGs and SFGs and just refer to them as SFGs.
We adopt six features traditionally used as diagnostic criteria to classify sources as SFGs and the AGN in the astronomical literature. The features are the far-infrared radio correlation parameter ($q_{\mathrm{IR}}$), the compactness in optical imaging (class_star), the total stellar mass ($\mathrm{M}_\star$), and the three IRAC colours, i.e., log(S8/S45), log(S58/S36), and log(S45/S36), details of these features are described at 3.4. Several works use the far-infrared radio correlation (FIRRC), the ratio between total (rest-framw 8-1000 µ$m$) infrared and radio (rest-frame 1.4 GHz) luminosity in SFGs, to calibrate the radio emission as a star formation rate (SFR) indicator. Delvecchio et al. (2021) calibrates the $q_{\mathrm{IR}}$ as a function of both stellar mass ($\mathrm{M}_\star$) and redshift (z). The $q_{\mathrm{IR}}$ parameter is widely applied to categorize radio sources as either SFGs or radio-loud AGN because it is typical of AGN activities to create an excess of radio emission. The compactness in optical imaging could indicate whether a source is an optical AGN because the AGN is intrinsically more compact than a SFG. In Whittam et al. (2022), a source is an optical AGN if it has a class_star value of 0.9 or greater. Class_star is a parameter that measures the compactness of sources in SExtractor. Normal SFG spectral energy distributions (SEDs) show a dip between the 1.6 µm stellar bump and long-wavelength emission from star formation hot dust. The superposition of blackbody emission from the AGN-heated dust will fill in the dip in the galaxy's SED and form a red, power-law-like thermal continuum across the IRAC bands if the AGN is sufficiently bright in comparison to its host galaxy. The IRAC criteria are explained in detail by Donley et al. (2012); see also Figure 1 of their study. In this study, we adopt two colours; 4.5µm-8µm (log(S8/S45)), and 3.6µm-5.8µm (log(S58/S36)) used by Whittam et al. (2022) to classify radio sources as midIRAGN or non-midIRAGN. In addition, we derive the third IRAC colour, 3.6µm-4.5µm (logS45/S36), using the IRAC1 and IRAC2 flux density. Future surveys may not have measurements in the IRAC3 and IRAC4 wavelengths. As such, we investigate the performance of the ML model with the two IRAC colours not used as input features; we show the results in section 4.8.
Whittam et al. (2022) also used X-ray and optical spectral line measurements to classify SFGs and AGN from the MIGHTEE-detected radio sources in the COSMOS

field. In this work, we combine the classification results (i.e. the labels) from Whittam et al. (2022) and our selected six features to train and test the performance of ML models in classifying SFGs and AGN from the MIGHTEE-COSMOS radio catalogue. In our study, however, we decided not to use X-ray or optical spectral line measurements as the input features because only 10% of the MIGHTEE sources have X-ray detections, and a very small number of sources have got reliable optical spectral line measurements. Since the missing value in a dataset can significantly affect the accuracy of results, attenuating the ML models' performance and such datasets cause problems for most of the ML estimators that assume that all values in an array are numerical and that all have and hold meaning. Thus this study is limited to only features whose missing data is less or equal to 10%.

We created a new sample of the MIGHTEE-COSMOS radio sources comprising the selected features with the class labels, from which all the analyses in this section are based. Hereafter, we will refer to it as MIGHTEE-COSMOS data/sample or just data/sample. The completeness of the selected features is provided in Figure 4.1. The last bar shows the total of 4,589 labelled sources in the MIGHTEE-COSMOS catalogue, and the height of the first six bars shows the completeness of the selected features.

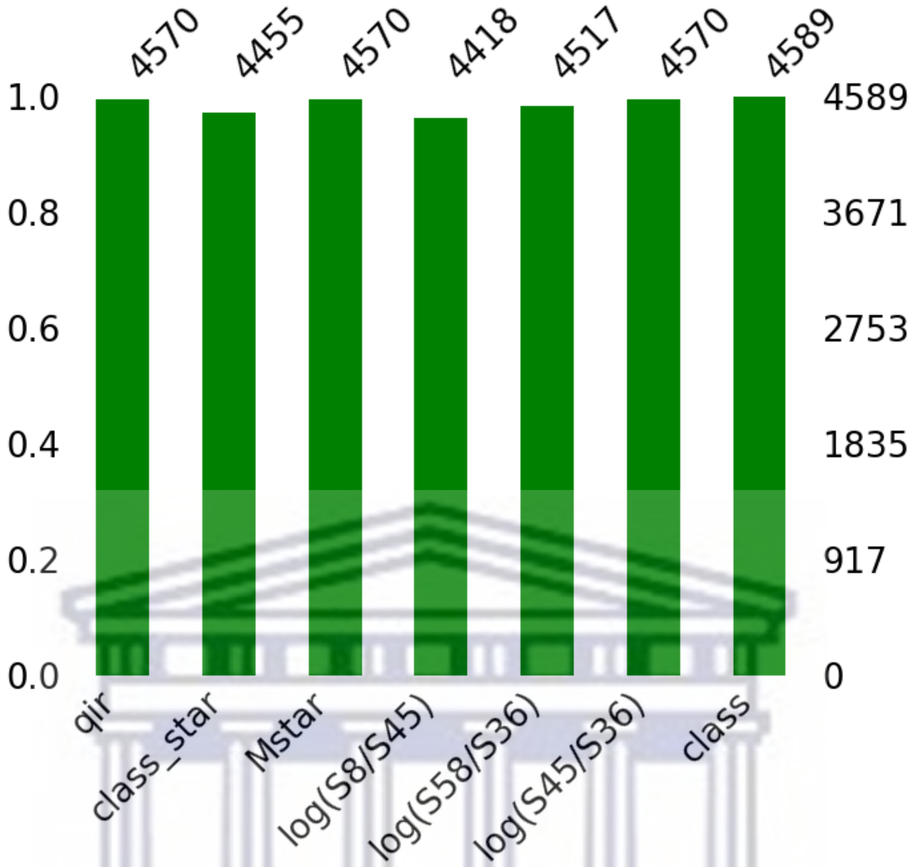


Figure 4.1: The completeness of the six features selected to train ML models from the MIGHTEE-COSMOS catalogue. The left vertical axis scale ranges from 0.0 *to* 1.0 and shows the fraction of completeness of each feature, where 1.0 depicts 100% data completeness (No missing data). The right vertical axis represents the number of observations available from 0 to the total number of observations. The bottom and the top horizontal axis represent the name of features and the total number of valid measurements in each feature, respectively.

As shown in Figure 4.1, the intense multiwavelength data in the COSMOS field resulted in the features' highest completeness. The IRAC colour log(S8/S45) has the least valid measurements. This study considered sources with valid measurements in any of the six features. Thus, if a source has an invalid measurement in any features, it will be discarded from the analysis. This is reasonable since sources with invalid measurements are less than 5% of the whole dataset. The clean MIGHTEE-COSMOS data with no missing data comprises 4,418 sources, from which 1,484 and 2,789 are classified as AGN and SFGs, respectively by Whittam et al. (2022).
The final MIGHTEE-COSMOS data were randomly divided into a training (75%) and a test data (25%). The training data comprises 1,113 AGN and 2,091 SFGs, totalling 3,204. The test contains 371 AGN and 698 SFGs, which sum up to 1,069 sources. The size and number of sources in test data were fixed and not involved in all the analysis except when testing the algorithm to ensure no data leakage in all the

calculations and training models.

The process of the supervised ML method is described well in section 2.1. A supervised classification depends on the quality of the data set employed in model-building, the care with which adjustable model parameters were chosen, and the evaluation criteria used to report the modelling process results. To achieve good results, we follow the fundamental steps in supervised ML 1. feature selection, including feature analysis, 2. building a training set using the selected features, 3. training the ML model to build a classifier, 4. hyperparameter optimization, and 5. application of the classifier to predict the class of the test sample.

## 4.2 One Dimensional Feature Analysis

As mentioned in Chapter 2, supervised ML model performance depends strongly on the features selected to train the model. Feature selection reduces the dimensions of the data so that the model can perform faster and more efficiently. However, selecting the essential feature is not trivial and sometimes requires sophisticated ML methods. This section considers two methods to analyse a low-dimensional feature space. One is called the histogram method and suggested by Zhang et al. (2003); the other is called the Kolmogorov–Smirnov test (KS test).

Zhang et al. (2003) demonstrated that histograms could extract essential features for classification. Figure 4.2 shows a histogram of each feature per class (AGN or SFG). A perfect feature for separating AGN from SFGs is one in which the distributions of the two classes are completely separate (with significant differences between the *means* of the two classes) such that one can infer a single threshold between the distributions, i.e., if the class distributions are completely separable, the corresponding feature is essential.

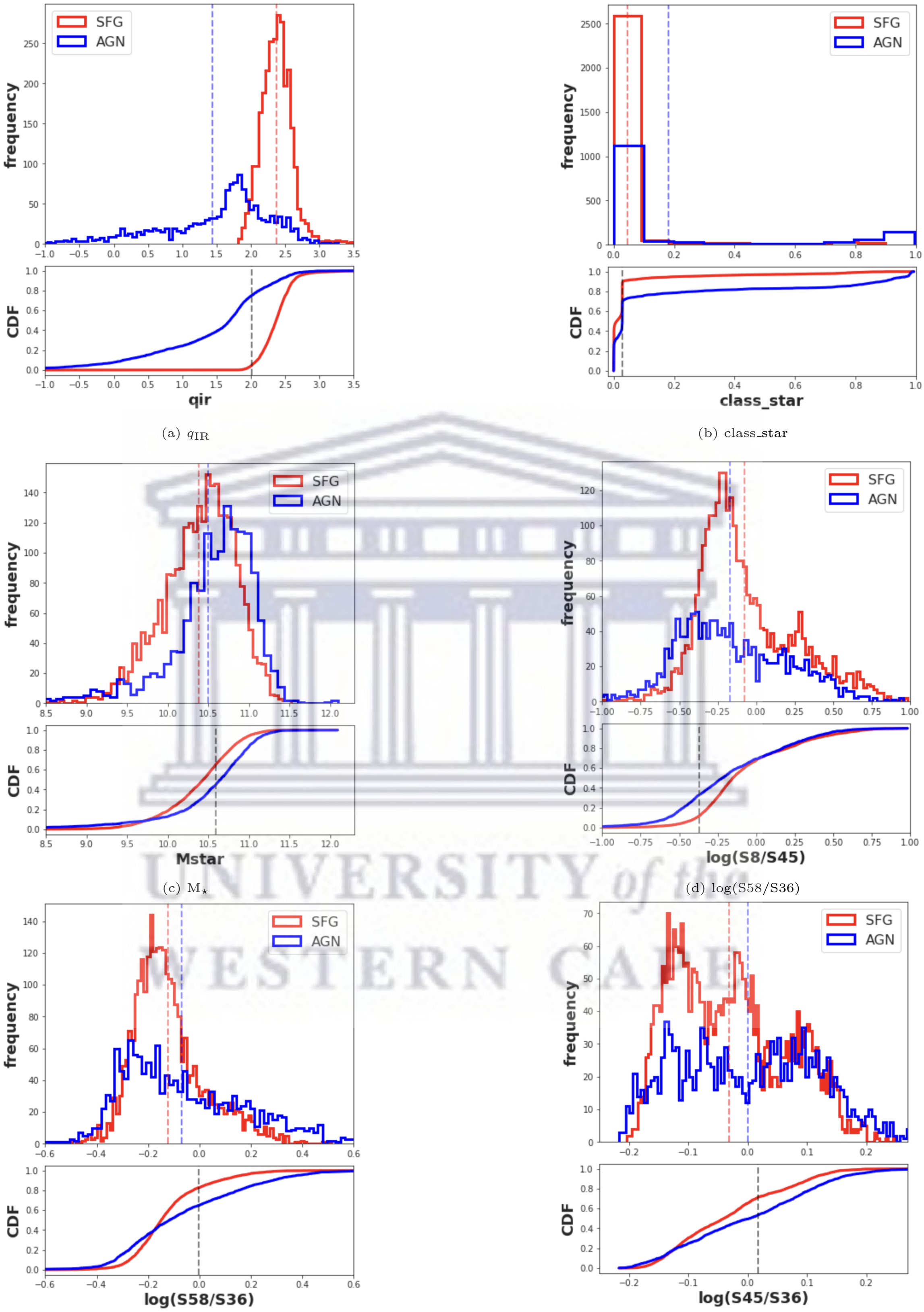
The KS test is a nonparametric goodness-of-fit test used to determine whether two distributions differ or whether an underlying probability distribution differs from a hypothesized distribution (Berger and Zhou, 2014). The KS test checks the difference in the overall shape of the two sample distributions or compares it to the expected statistical distribution. The KS test can be implemented from the scipy.stats.ks_2samp[1] python method. The KS test is based on the maximum distance between the two distributions, referred to as KS-stats. If this distance is zero, then the two distributions are the same. Otherwise, if the KS-stats is 1, then the two distribution differs. The black dashed line in the bottom plots of Figure 4.2 shows an x-axis value where the maximum distance between two distributions exists.

Figure 4.2a shows the histogram (top) and KS test (bottom) of SFGs and AGN over the $q_{\mathrm{IR}}$ space. Both plots indicate that two classes are separable using this feature.

[1] KS test; https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html

For example, in the histogram, the *means* of SFGs and AGN are not identical, and to some extent, we may approximate that most of the SFGs have the $q_{\mathrm{IR}}$ value greater than 2 while the AGN has the $q_{\mathrm{IR}}$ value less than 2. Similarly, the KS test plot shows a great separation between the SFGs and AGN curves, implying that the two populations differ. This is because $q_{\mathrm{IR}}$ is used to identify sources with significantly more radio emission than would be expected from star-formation alone Whittam et al. (2022). This correlation is parameterised via the far-infrared-to-radio luminosity ratio discussed in section 3.5.3. Figure 4.2b shows similar results for the class_star indicator. In the histogram, the two classes, in this case, are not separable below 0.9; however, there are more AGN than SFGs above this threshold. The class_star feature could be used to categorize SFGs and AGN. Because the emission from AGN outshines that of the host galaxy, AGN sources in optical imaging will be more compact. In Whittam et al. (2022), this indicator was used to identify optical AGN with a value of class_star greater than 0.9 as Optical AGN. We can also infer from the plot that only the minority of the AGN meets this criterion. Figure 4.2c represents the analysis of SFGs and AGN over stellar mass ('$\mathrm{M}_{\star}$'). Most of the two classes overlap in the histogram, so estimating the clear-cut between them would not be easy. This is also indicated in the KS test results; the AGN and SFGs curves are close. Figure 4.2d, Figure 4.2e, and Figure 4.2f depict the analysis of AGN and SFGs for the three IRAC colours, log(S8/S45), log(S58/S36), and log(S45/S36) respectively. In all three plots, estimating a threshold for discriminating between SFGs or AGN is problematic since the two classes are distributed within the same range, i.e. the two distributions have nearly identical means.

The histogram and KS test methods for analysing the selected features indicate that the $q_{\mathrm{IR}}$, compared to the other five features, is the best indicator for classifying SFGs or AGN. The second best indicator is the class star, which tends to work well for the minority of the AGN, and this may be due to the manual classification mentioned in Chapter 3. When used alone, the stellar mass and derived IRAC colours are not good indicators of the AGN or SFG since the two populations are highly overlapped.



(a) $q_{\rm IR}$

(b) class_star

(c) $M_{\star}$

(d) log(S58/S36)

(e) log(S8/S45)

(f) log(S45/36)

Figure 4.2: Histograms (top) and KS test (bottom) results of the six selected features of labelled AGN (blue) and SFGs (red) in the MIGHTEE-COSMOS data. Each diagram shows the distribution of AGN and SFGs and the KS test results in each feature. For Histograms, the vertical axis represents the number of radio sources. If the means (represented by a vertical dashed line) of the distribution of two classes are entirely different, we consider that the two classes are separable by the corresponding feature. However, if the means of the two distributions are identical, we assume that the two classes are indistinguishable using that feature. Similarly, in the KS test, the vertical axis shows cumulative distribution function (CDF), and if the curves in the KS test plots are close, then the two classes are indistinguishable. The $q_{\rm IR}$-parameter plot shown in Figure 4.2a has a substantial difference between the mean of AGN and the mean of SFGs as such, it is essential to discriminate SFGs from AGN. AGN and SFGs on the Figure 4.2b, 4.2c, 4.2d, and 4.2e has nearly identical means. Figure 4.2f shows that the distributions of AGN and SFGs in the log(S45/S36) space are similar.

## 4.3 Feature Correlation (Two Dimensional feature analysis)

In this section, we make joint feature correlation plots using all possible pairs of features for the two classes of radio sources under study. Since we have six features and want to make a combination of 2, using equation 4.1 will have 15 correlation plots as shown in Figure 4.3. Figure 4.3 shows all feature correlation plots. The AGN and SFGs are shown in blue and red, respectively, with the corresponding 95% confidence ellipses.

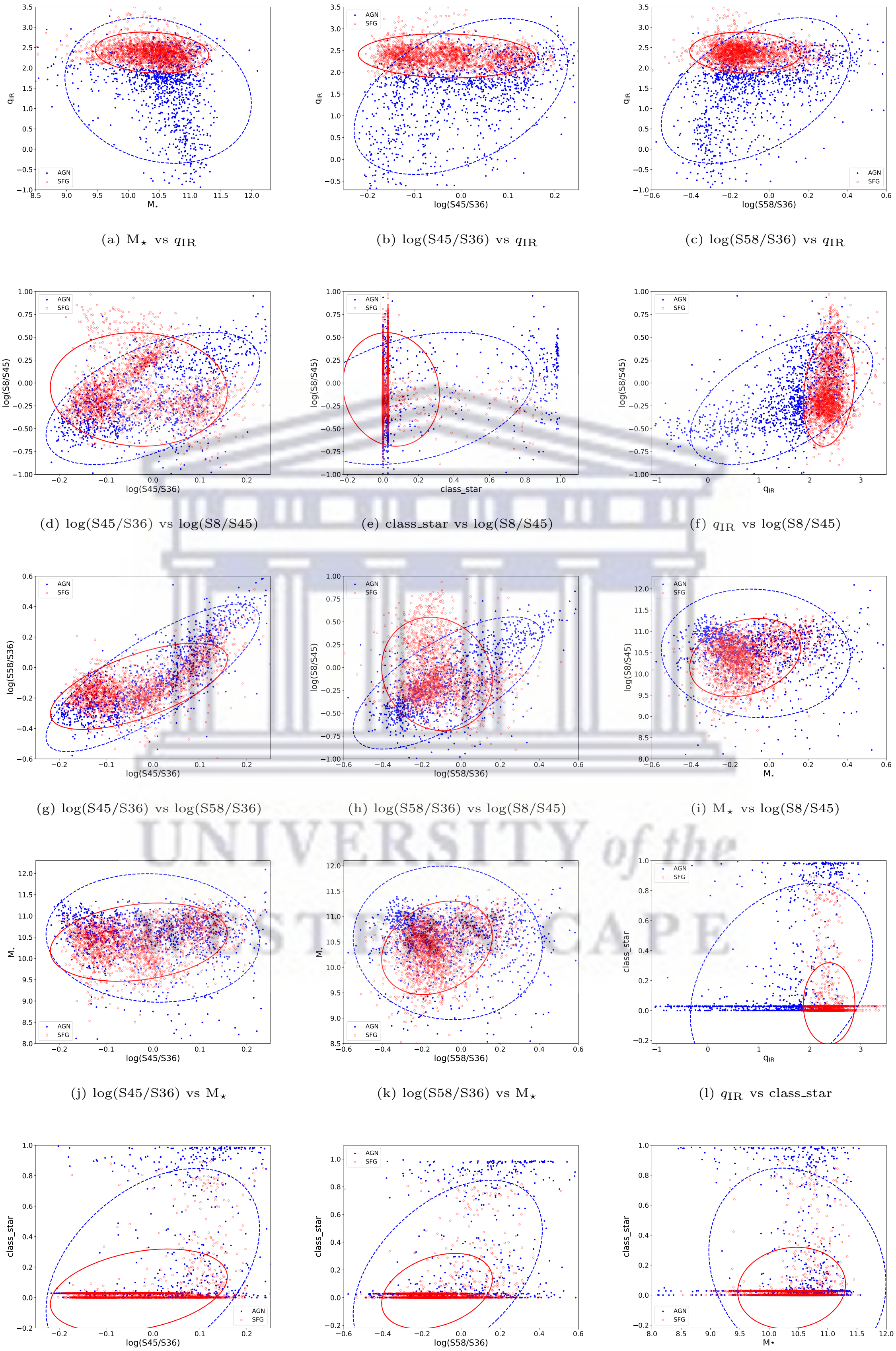
$$C(n,r) = \frac{n!}{(r!(n-r)!)} \tag{4.1}$$

Figure 4.3a depicts the $q_{\rm IR}$ parameter plotted as a function of stellar mass for both SFGs and AGN. In this feature space, if we consider the vertical axis, the SFGs could be securely approximated to be above $q_{\rm IR}$ value of 2. However, there is no cutoff we can approximate on the horizontal axis. This feature combination does show that two classes may be separated, but the confidence level ellipse of AGN is much broader than that of the SFGs, such that SFGs falls within the confidence level of the AGN. Due to this, we can say that these two features fail to distinguish the two classes. The SFGs extend slightly outside the AGN ellipse in both 4.3b and 4.3c; however, most of the sources still lie within the AGN ellipses, making it difficult to classify the SFG and AGN using feature combinations in the first row of Figure 4.3.

Figure 4.3d shows the two classes plotted over the IRAC colour-colour feature space. Similarly, the majority of SFGs lie within the AGN ellipse. The remarkable result is that the two ellipses tend to take slightly different correlation directions, indicating some association between the two IRAC colours and the two classes. It is shown in Figure 4.2 above that using only one of the IRAC is ineffective for classifying these sources, but adding another IRAC colour has a slight improvement for classification. Figure 4.3e shows that the features log(S8/S45) and class_star fail to discriminate between SFGs and AGN. Similarly, the features log(S8/S45) and $q_{\rm IR}$ shown in Figure 4.3f cannot separate the two classes.

In row three, Figure 4.3g shows the two classes plotted over the log(S58/S36) vs log(S45/S36). A positive correlation exists between the two features of SFGs and AGN. SFGs and AGN overlap within the same space, making this feature combination ineffective for separating the two classes. Figure 4.3h shows log(S8/S45) as a function of log(S58/S36) for both SFGs and AGN. The two IRAC colours show a positive correlation for the AGN and a negative correlation for the SFGs. These two IRAC colours are the best indicators of AGN and SFGs. The two IRAC colours are defined by Donley et al. (2012) as the indicators of whether a radio source is a mid-infrared AGN or not, refer to section 3.5.1 or Donley et al. (2012) for more details regarding this indicator. The features log(S8/S45), and $M_\star$ shown in Figure 4.3i fail to distinguish the two classes.

The remaining diagrams show that the corresponding feature pairs can not effectively classify SFGs and AGN. This method indicates that adding more features could improve the classification results. However, making a correlation pair plot for a high-dimensional dataset might not be effective since the more features are added, the more plots are to be analysed. As described in Chapter 3, the conventional techniques for classification follow such methods with additional derived relations or assumptions. Some sources cannot be securely classified because they overlap in some feature space, which would require more information before confirmation.



(a) M$_\star$ vs $q_{\rm IR}$

(b) log(S45/S36) vs $q_{\rm IR}$

(c) log(S58/S36) vs $q_{\rm IR}$

(d) log(S45/S36) vs log(S8/S45)

(e) class_star vs log(S8/S45)

(f) $q_{\rm IR}$ vs log(S8/S45)

(g) log(S45/S36) vs log(S58/S36)

(h) log(S58/S36) vs log(S8/S45)

(i) M$_\star$ vs log(S8/S45)

(j) log(S45/S36) vs M$_\star$

(k) log(S58/S36) vs M$_\star$

(l) $q_{\rm IR}$ vs class_star

(m) log(S45/S36) vs class_star

(n) log(S58/S36) vs class_star

(o) M$_\star$ vs class_star

Figure 4.3: The feature correlation of each pair of features we selected to classify SFGs and AGN from the MIGHTEE-COSMOS radio sources. The red and blue points represent the SFGs and AGN distribution labelled in the MIGHTEE-COSMOS catalogue in Whittam et al. (2022) respectively. The solid red ellipse shows the 95% confidence level of SFGs, while the blue dashed line shows the 95% confidence level of AGN. The direction of the ellipse indicates the correlation of the two features per class.

## 4.4 t-SNE feature analysis

t-distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised and non-linear dimensional reduction technique introduced by Laurens van der Maaten and Geoffrey Hinton in 2008. t-SNE is widely applied to visualise high-dimension data. It works by converting the high-dimensional data into two or three-dimensional data that can be displayed in a scatterplot while preserving the local and global structure of the data. It is particularly effective in revealing clusters, patterns, and relationships in complex datasets Van der Maaten and Hinton (2008). We refer the reader to visit the Van der Maaten and Hinton (2008) study for details of t-SNE model.
In this study, we adopt and apply the t-SNE model to reduce the six-input features in the training data to a two-dimensional feature space. We also use the correlation technique used in section 4.3 to deduce if the new t-SNE features are able to separate between the SFGs and AGN. Figure 4.4 shows the correlation plot of t-SNE two-dimensional features mapped from the six MIGHTEE features in the training data. The AGN and SFG in this feature space are separable although there is an overlap between the two 95% confidence ellipses. Since t-SNE features are generated through an unsupervised process and are unlikely to have a direct physical interpretation (Balamurali and Melkumyan, 2016), we do not use the t-SNE features to train the ML models in this study. We further investigate the importance of the MIGHTEE-selected features using automated techniques in the following sections.

## 4.5 Automated Feature Importance

We realised in section 4.2 and 4.3 that we were able to get some knowledge on which features or pair of features is essential for classifying the SFGs and AGN by just plotting a histogram of sources per class, KS test and by making use of correlation plots. We found that the most crucial feature for classification is $q_{\mathrm{IR}}$ and the combination of the two IRAC colours: log(S58/S36) and log(S45/S36). The two methods, however, are time-consuming and do not scale when one has a high-dimensional dataset. In this section, we adopt three automated approaches that are not dependent on the build-in feature importance algorithm of the ML model (ML-independent model); *permutation feature importance, sequential feature importance and ROC curves* to assess the significance of the selected features in classifying SFGs and AGN from the radio sources. We chose the random forest model to use as the ML model for both permutation and sequential feature importance because the random forest model has a built-in component of feature importance that we can confirm some results with.

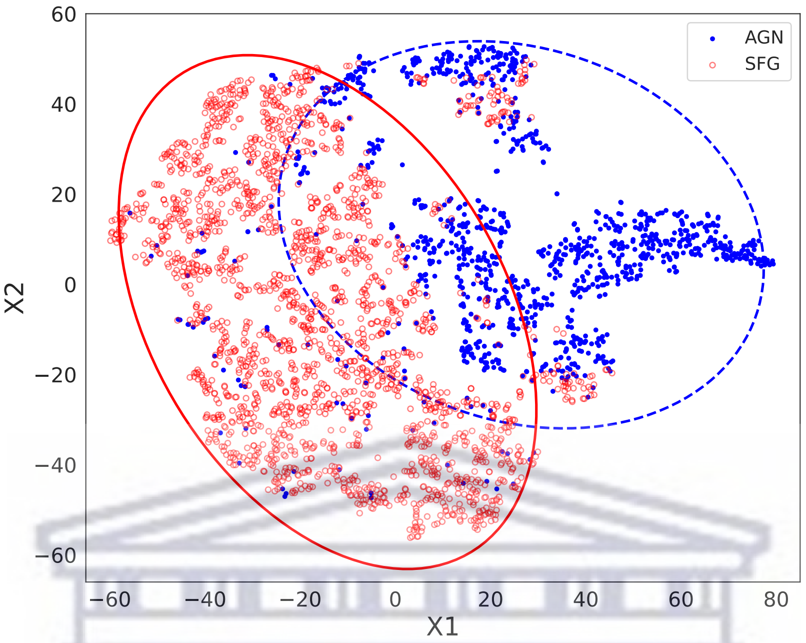


Figure 4.4: Visualization of MIGHTEE feature space in a two-dimensions using t-SNE. The vertical and horizontal axis represents the two t-SNE features converted from the six MIGHTEE features. The red and blue points represent star-forming galaxies (SFGs) and active galactic nuclei (AGN) distribution labelled in the MIGHTEE-COSMOS catalogue in Whittam et al. (2022) respectively. The solid red and blue dashed ellipse shows the 95% confidence level of SFGs and AGN respectively. The direction of the ellipse indicates the correlation of the two features per class.

### 4.5.1 Permutation Feature Importance

Permutation feature importance is a technique that can inspect a model for any fitted ML model when the data is tabular. The permutation procedure breaks the relationship between the feature and the output or class label, leading to poor performance of the ML model. The more the model performance drops, the more important the feature is. Thus permutation feature importance is defined as the decrease in an ML model score when a single feature value is randomly shuffled. The details about the models mathematics can be found in interpretable-ml-book[2] and the implementations are available at sk-learn permutation importance[3]

Figure 4.5 shows the permutation feature importance method's results for assessing the derived features' significance in classifying SFGs and AGN. The importance is the mean scores calculated by using 1000 permutations. The results indicate that the $q_{\mathrm{IR}}$ is an essential feature for this classification, consistent with the one-dimensional

[2] https://christophm.github.io/interpretable-ml-book/feature-importance.html
[3] https://scikit-learn.org/stable/modules/permutation_importance.html

analysis described in section 4.2. The $q_{\mathrm{IR}}$ is ranked more than 25% important for this classification, with the remaining features rated below 5% important. These results complement the above conventional techniques; the $q_{\mathrm{IR}}$ histogram and KS test results showed that the two classes could be separated by $q_{\mathrm{IR}}$ than other features. The method ranks the class_star as the second most important feature. The two IRAC colours, log(S58/S36) and log(S45/S36), which gave positive outcomes when combined in section 4.2, are rated as the third and fourth important features, respectively. The method suggests that the stellar masses and the IRAC colour log(S45/S36) are the least significant features with an importance of about 2%.

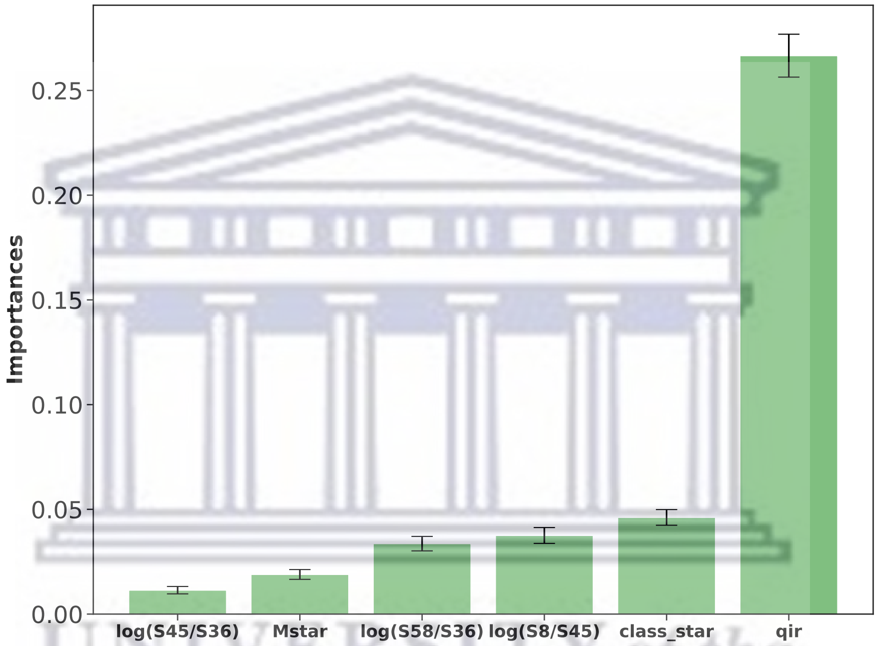


Figure 4.5: Permutation-based feature-importance measures for the explanatory features included in the Random Forest model for the MIGHTEE-COSMOS data using F1-score (defined in section 2.2.5) as the evaluation metric. The importance is the mean scores calculated by using 1000 permutations. A feature with a higher value should be considered the most important feature for classification.

### 4.5.2 Sequential Feature Importance

The sequential feature selection approach adds or removes features from the dataset sequentially. The method is used to reduce initial N features to M features where M<N. The M features are optimized for the performance of the model. This study uses a sequential feature importance approach to determine and evaluate multiple feature importance. The details regarding the implementation of sequential feature

selection can be found sk-learn; sequential feature importance [4].
In our study, we had six features and decided to run the model five times from M 1-5, as shown in Table 4.1. Since the algorithm returns important M features, we can compare our results to the permutation and ML models' results in the following sections. Table 4.1 below shows the results for each trial. This method ranks $q_{\rm IR}$, class star, and log(S8/S45) as the three most essential features, respectively. Compared to the permutation importance method, the sequential permutation method rates the $M_\star$ as the 4th important feature, not the log(S58/S36); instead, the log(S58/S36) is considered the least important feature. This is likely because the sequential permutation method is designed to consider only one if two or more features are highly correlated. The correlation between these two features is evident in Figure 4.3h.

Table 4.1: Sequential feature importance results for MIGHTEE-COSMOS data

| N | M | Features selected |
|---|---|---|
| 6 | 1 | $q_{\rm IR}$ |
| 6 | 2 | $q_{\rm IR}$ and class_star |
| 6 | 3 | $q_{\rm IR}$, class_star and log(S8/S45) |
| 6 | 4 | $q_{\rm IR}$, class_star, log(S8/S45) and $\mathrm{M}_\star$ |
| 6 | 5 | $q_{\rm IR}$, class_star, log(S8/S45), $\mathrm{M}_\star$ and log(S58/S36) |

### 4.5.3 ROC Curve

We conclude the automated feature analysis with the receiver operating characteristic (ROC) curve. The ROC curve is a graph showing the performance of a classification model at all classification thresholds. We plot the *true positive rate* (TPR) as a function of the *false positive rate* (FPR), where TPR is the probability that an actual positive will test positive (given by equation 4.2). FPR is the probability that a true negative will test negative (given by equation 4.3). We consider the area under the ROC curve (AUC) to understand which feature performance is great. The more area under the curve, the more accurate the model implies, and the more critical the feature is for classification, which means the ideal classification with TPR=1 and FPR is 0. Therefore, the closer the AUC is to 1, the better the ML performance. The minimum AUC is 0.5, which means a random classification. Therefore the closer the model is to an AUC of 0.5 it is considered a random classification model.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2}$$

$$\text{FPR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.3}$$

where TP, FN, TN and FP are defined in section 2.2.5.
Figure 4.6 shows the ROC curves based on thresholds imposed individually on six features. The AUC scores for $q_{\text{IR}}$, class_star, $\text{M}_{\star}$, log(S8/S45), log(S58/S36), and log(S45/S36) are 0.88, 0.62, 0.62, 0.59, 0.54 and 0.57 respectively. The $q_{\text{IR}}$ curve (shown in green) has more area under the curve when compared to other features, implying that it is the most important feature for classification.

It is necessary to emphasize that no ML-dependent or based feature importance technique has been adopted to this point. Still, the results show that the selected features are crucial for distinguishing the SFGs and AGN, in which the $q_{\text{IR}}$ is the most significant feature. The ROC curves are calculated from the threshold imposed individually on each feature with no ML model involved, but already we can achieve AUC scores close to 90%. The approach of imposing a threshold on a feature when

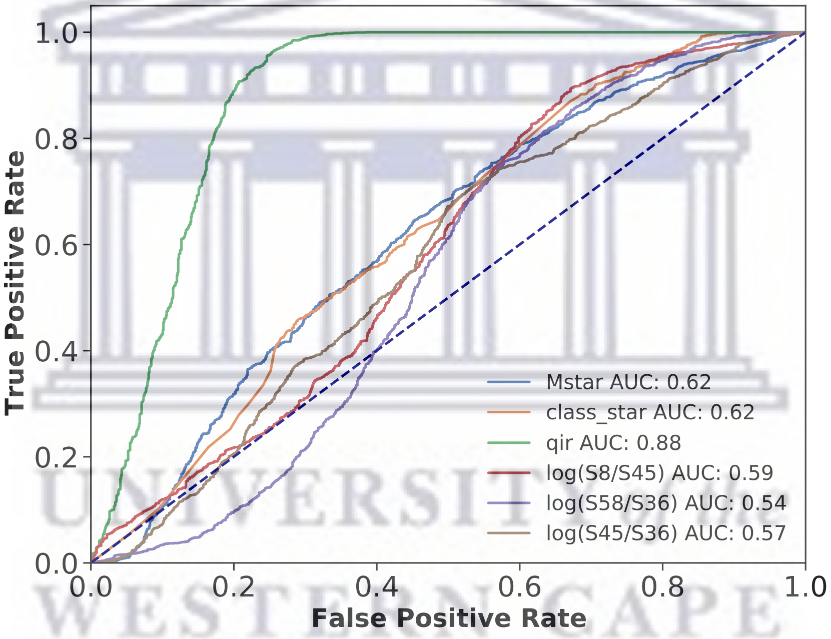


Figure 4.6: Receiver Operating Characteristic (ROC) curve for the six selected MIGHTEE-COSMOS features, namely, the $q_{\text{IR}}$ (green), class_star (yellow), $\text{M}_{\star}$ (blue), log(S8/S45) (purple), log(S58/S36) (violet), and log(S45/S36) (brown). The higher the AUC score, the greater the model's performance and thus, the corresponding feature is essential.

only one feature is considered for classification is very similar to how the logistic regression model sets a threshold for a single feature to make a classification. However, the one considered here is different from the ML method. I should expect that the logistic regression results for $q_{\rm IR}$ only be very close to the predicted AUC for $q_{\rm IR}$ in Figure 4.5.

## 4.6 Random Forest Feature Importance

In this section, we compute the feature importance using the random forest build-in feature importance model. We compare the techniques employed for feature selection with the model's built-in importance approach. The built-in random forest importance is computed in two ways: *Gini importance* (or mean decrease impurity (MDI)) and *Mean decrease accuracy* (MDA). We chose the MDI method since the MDA is not implemented in the SK-learn package and is very similar to the permutation importance method. More details regarding the MDI implementation are available at Feature importance with a forest of trees[5].
Figure 4.7 represents the importance estimated for each feature by the built-in random forest feature importance model. The $q_{\rm IR}$ is rated as the most crucial feature with about a 60% score. The remaining features are ranked as class_star, $M_{\star}$, log(S8/S45),log(S58/S36), and log(S45/S36) which are consistent with our previous no-ML-dependent assessments. These features are ranked in the same as the permutation importance method.

We have employed several techniques to assess the importance of the selected features. All the feature analysis methods rank the $q_{\rm IR}$ parameter as the most crucial feature, whereas the IRAC colour log(S45/S36) is the least important feature. The other features change the order in which they are ranked depending on the method employed. The emission of local star-forming galaxies follows the far-infrared-radio correlation, initially discovered at rest-frame 1.4 GHz and covers three orders in the magnitude of radio continuum luminosity (Schober et al., 2022). The infrared emission is caused by dust that relatively massive OB stars have heated. The radio emission results from relativistic cosmic ray electrons (CRe) accelerated by shock waves produced when massive stars explode as supernovae. Delvecchio et al. (2021). The AGN activities accelerate the CRe electrons and cause a radio excess. Such sources with AGN activities will have significantly more radio emission than would be expected from star formation alone. This is reflected by the $q_{\rm IR}$ parameter ranked as the most significant feature in this study. Most models rated the class_star as the second best feature, selected because the source with an AGN activity displays the point-like morphology

[5] https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

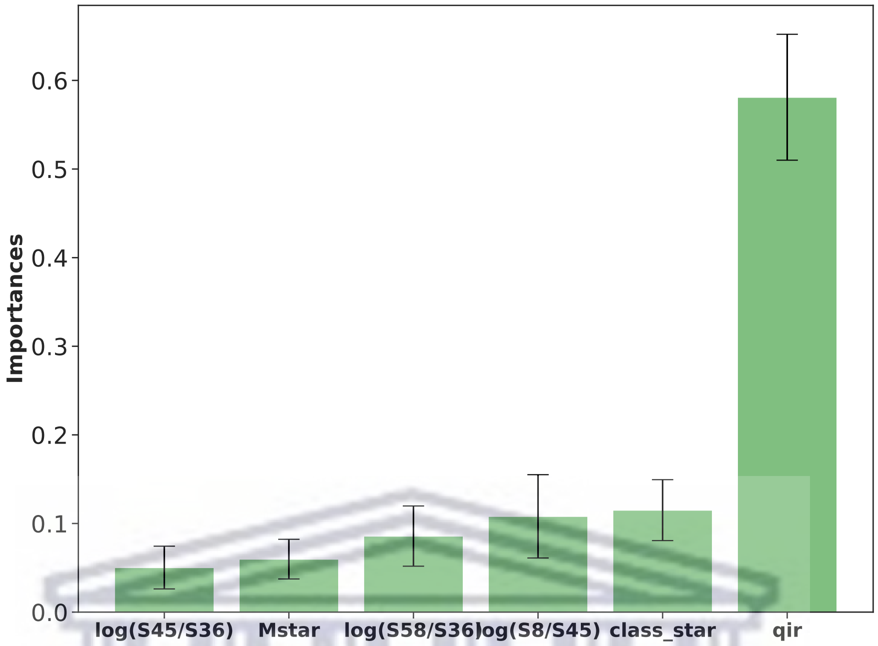


Figure 4.7: The feature importance results of the random forest model for our selected six features. The F1-score is used as the evaluation. A feature with a higher value should be considered the most important feature for classification

as the emission from the nucleus outshines the emissions from the host galaxy. The two IRAC colours log(S8/S45) and log(S58/36) are shown in a correlation scatter plot that they are essential for classification when used together, which is found by (Donley et al., 2012) that the two IRAC colours can identify the mid-infrared AGN. As shown in the histogram plots, the total stellar mass of radio sources can not alone be a reliable feature for classifying SFGs and AGN, possibly because the stellar mass of an SFG and AGN is not systematically different. Still, when we consider it with other features such as $q_{\rm IR}$, we can securely identify the SFGs as shown in Figure 3.6. We have also derived the third IRAC colour log(S45/S36) to investigate if fields not covered by IRAC3 and IRAC4 observations could be a disadvantage for source classification purposes. We found that all techniques consider the log(S45/S36) colour the least vital feature to discriminate SFGs from AGN.

We have demonstrated that the employed techniques agree with ranking the features according to their importance. We can reproduce the significance of features or feature combinations used in literature and asses which ones are important without using any assumption or physics regarding the SFGs or AGN.

## 4.7 Class Imbalance

Imbalanced data refers to the data with a severe skew in the class label distribution, such as 1:100 or 1:1000 examples in the minority class to the majority class (Krawczyk, 2016). This bias can influence many supervised ML algorithms, leading some to ignore the minority class entirely. This is a problem as it is typically the minority class in which predictions are most important (Das et al., 2022).
The typical approach to addressing the data imbalance problem is to resample the training data randomly. The two standard methods are; *Undersampling* and *Oversampling*. The undersampling procedure removes some sources from the majority class, while the oversampling technique duplicates the examples from the minority class. In this study, we choose the undersampling technique to remove some SFGs in the training data to equal the number of AGN and SFGs. Considering the undersampling approach for this case makes sense since the number of AGN is usually in the minority in most radio continuum surveys.
We consider several fractions of the complete training set to train the RF model and evaluate it using the F1-score. Figure 4.8 shows how the F1-score per class varies as a function of the fraction of the complete training data used to fit the model. The model's performance for the two categories is more than 90% for training set sizes greater or equal to 20% of the complete training set. Above a training size of 20%, the increase in the train size has no significant impact on the model's performance. This is indicated by an almost constant F1-scores for training set sizes above 20%. Below the training set size of 20% there is an overall non-linear increase of F1-scores, which results because the model has not captured enough information regarding the relationship between input features and target classes.
In the complete training set data, the AGN to SFGs ratio is about 1: 2. Thus, we have approximately two times more SFGs than AGN. We repeat the above procedure, balancing the two classes in the training dataset whilst maintaining the same unbalanced test dataset gives the result shown by a dashed line in Figure 4.8. To balance the data, we have adopted the imlearn - RandomUnderSampler[6] method. The model's resulting F1-scores are slightly lower for all classes, decreasing by an identical fraction for all fractions of the training dataset.
Overall, the class imbalance in our dataset does not affect the model's performance in this study because of the relatively modest class imbalance. Most SFGs in the training dataset do not necessarily imply that SFGs will be classified more accurately than AGN. The good F1-score of SFGs classification shown in figure 4.8 might be related to other observational or intrinsic properties of SFGs, which is not a focus of this study.

[6] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

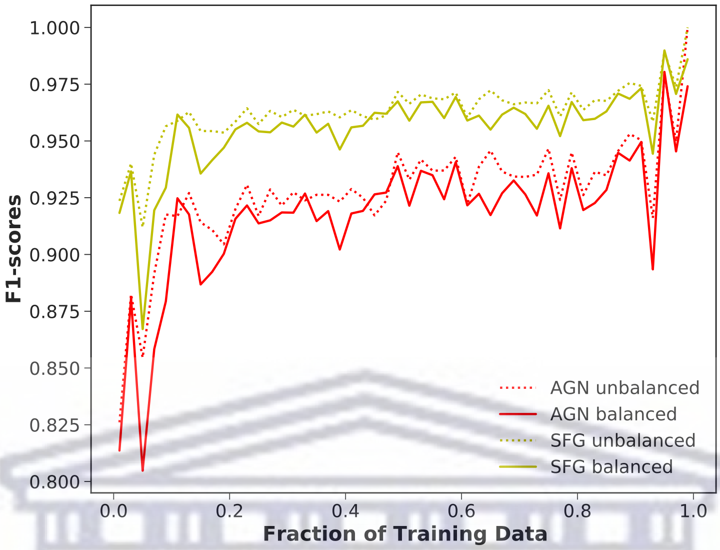


Figure 4.8: Results of applying Random Forest ML method to classify SFGs and AGN from the MIGHTEE-COSMOS survey. A plot of the Random Forest F1-score per class label against the fraction of the training data used to train the random forest. The data was balanced using the imlearn - RandomUnderSampler Method. The complete training data is divided into training and validation data. The training data is further reduced using the RandomUnderSampler method such that AGN and SFGs are equal in the train data. The model was evaluated on the validation dataset of spectroscopically confirmed sources without balancing the classes.

## 4.8 Machine Learning Application

In this section, we show the results of adopting five supervised ML models, namely, LR, SVM, KNN, RF and XGB on classifying SFGs and AGN from the MIGHTEE-COSMOS radio sources. With seven different feature combinations, we compare the performances of different ML algorithms on SFGs versus AGN classification based on training data sizes of 80%, 60%, 40%, and 20%, respectively. The different approaches that models consider for classification are explained in section 2.2.4. We train, validate and test the models on the same number of features to ensure that we make apples-to-apples comparisons. The seven input feature combinations, F1, F2, F3, F4, F5, F6 and F7, are described in Table 4.2. The first six feature combinations, F1-F6, are grouped according to their importance as ranked with the feature importance techniques discussed in the above sections. The seventh feature combination, F7 ($q_{\rm IR}$, class_star, $M_{\star}$, log(S45/S36)) was derived because future observations may not have the measurements on IRAC3 and IRAC4 bands. We want to investigate if the classifiers could still yield significant results or be disadvantaged if IRAC3 and IRAC4 measurements are unavailable to train the models.

Supervised ML models are structured such that they require training and testing datasets. Before testing, we run some cross-validation steps that ensure that the ML models are not overfitting, i.e. the model yields accurate results on the train data but yields worse results when applied to the unseen test data. We adopted a method by (Hussein et al., 2021) where alternative cross-validation configurations were tried, and *jackknife* estimation was used to determine the variability of each alternative.

We use random stratified sampling to split the MIGHTEE-COSMOS catalogue into a training (75%) and testing (25%) dataset. The complete training dataset has been used in the previous sections for feature and class imbalance analysis; however, the unseen test data has not been involved in any experiments. The test set size was fixed throughout all calculations.

We split the complete training dataset into *train* and *validation* data four times. The split ratios of train and validation data are [1:4], [2:3], [3;2] and [4:1], respectively. Dividing the complete train data reduces the available data to train the ML models. The split [1:4] indicates that only 20% of the complete train data is used to train the model while 80% of the data is kept as validation data. The second split [2:3] means that only 40% of the complete training data is used to train the model; in contrast, the remaining 60% is used as a validation dataset. Future radio facilities such as the SKA will detect billions of radio sources compared to the tenth of thousands of sources detected in current radio surveys. Thus the available conventional detected radio sources may be a small fraction of data to train the model to classify the billions of future radio sources. As such, it is very significant for this study to figure out how well the models perform when the training data is limited. We give the models different fractions of examples to learn the relationships between the input features and the class label and evaluate if there is a significant loss of accuracy as examples decrease.

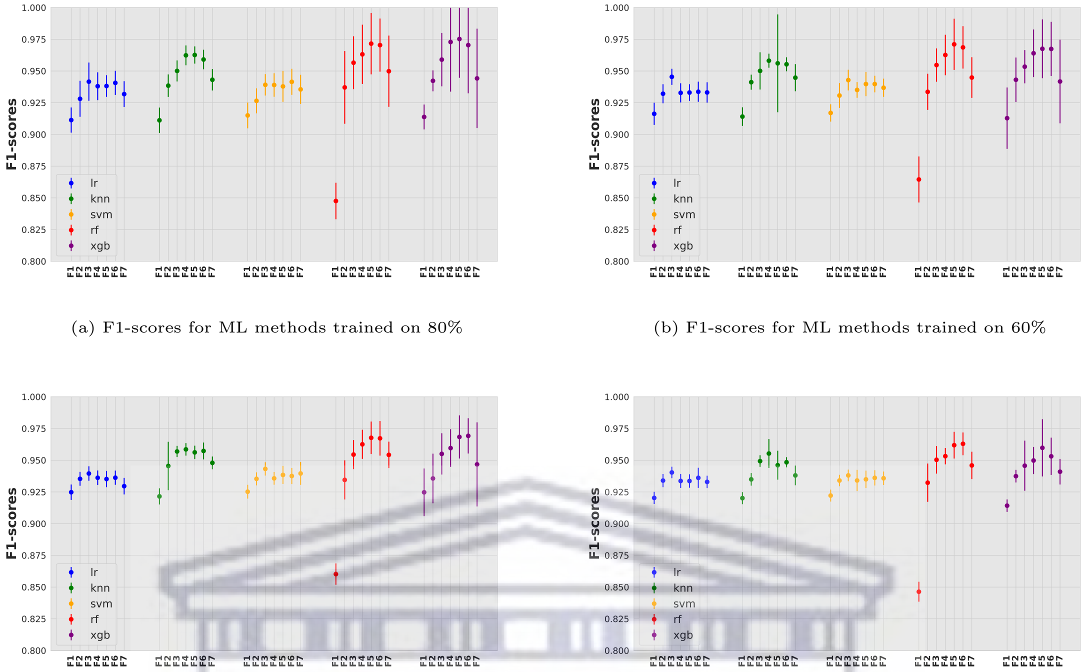
Table 4.2: Seven feature combinations

| Name of combination | Features |
|---|---|
| F1 | $q_{\rm IR}$ |
| F2 | $q_{\rm IR}$ and class_star |
| F3 | $q_{\rm IR}$, class_star and log(S8/S45) |
| F4 | $q_{\rm IR}$, class_star, log(S8/S45) and $M_\star$ |
| F5 | $q_{\rm IR}$, class_star, log(S8/S45), log(S58/S36) and $M_\star$ |
| F6 | $q_{\rm IR}$, class_star, log(S8/S45), log(S58/S36), $M_\star$ and log(S45/S36) |
| F7 | $q_{\rm IR}$, class_star, $M_\star$ and log(S45/S36) |

Figure 4.9 shows the F1-scores results of applying the LR, *k*NN, SVM, RF, and XGB models trained on input feature combinations described in Table 4.1 to separate SFGs from AGN evaluated on the *validation data*. We trained the models on 80%, 60%, 40% and 20% of the complete training data, respectively. All the classifiers were evaluated using the F1-score. Figures 4.9a, 4.9b, 4.9c and 4.9d show the results of models trained on 80%, 60%, 40% and 20% of the complete training data respectively. For each training data size, all models perform well for each feature combination, yielding F1-scores greater than 90% except for the random forest model on the F1 ($q_{\rm IR}$). As the training dataset decreases, all models' scores slightly drop, as seen in Figures 4.9a to 4.9d. This global trend is also indicated more clearly in Figure 4.12a. ML models yield scores above 90% for each feature combination even when trained only with 20% of the data. The LR and the SVM models obtain the highest score for F3, the combination of $q_{\rm IR}$, class_star, and log(S8/S45) features. However, both models' performances drop for F4, F5, F6, and F7, suggesting that adding more features to F3 does not improve the ability of these models to generalise the relationship between the input features and output (AGN or SFG). The *k*NN, RF and XGB models show some improvements as more features are added to train the model up to feature combination F5, beyond which the model's performance decreases. This suggests that the sixth added feature to the model is not essential and also confuses the model's ability to discriminate between SFGs and AGN. For all the training sets, the results indicate that the feature combination F7 yields poor results compared to F3 and F4, which means that sources without IRAC3 and IRAC4 measurements would be at a disadvantage. The derived uncertainties suggest that adding more features to the models has no significant impact on the overall performance of a model.

We applied the same trained and optimised models evaluated on the validation data as shown in Figure 4.9 to the fixed unseen test datasets. Figure 4.10 shows the F1-score results of all five models based on seven different feature combinations. Figures 4.10a, 4.10b, 4.10c and 4.10d show the performance of all models trained on 80%, 60%, 40% and 20% of the complete train data respectively. Compared to the validation dataset results, the models still yield comparable performances with scores above 90% in all four train splits. F1-scores decrease slightly with a decreased training data size (see Figure 4.12b for a clear trend), but all scores are still above 90%. The scores for all data splits have dropped somewhat slightly compared to the validation scores.

Figures 4.9 and 4.10 show that the models perform well in classifying MIGHTEE-COSMOS sources as SFGs or AGN. The models yield more than 90% F1-scores even when only trained on the $q_{\rm IR}$ parameter and only trained on 20% of the data. The LR and SVM have small error bars meaning the two models have precise results. The extensive RF and XGB uncertainties imply no significant difference in the performance of a simple LR and other sophisticated models for this classification.



(a) F1-scores for ML methods trained on 80%

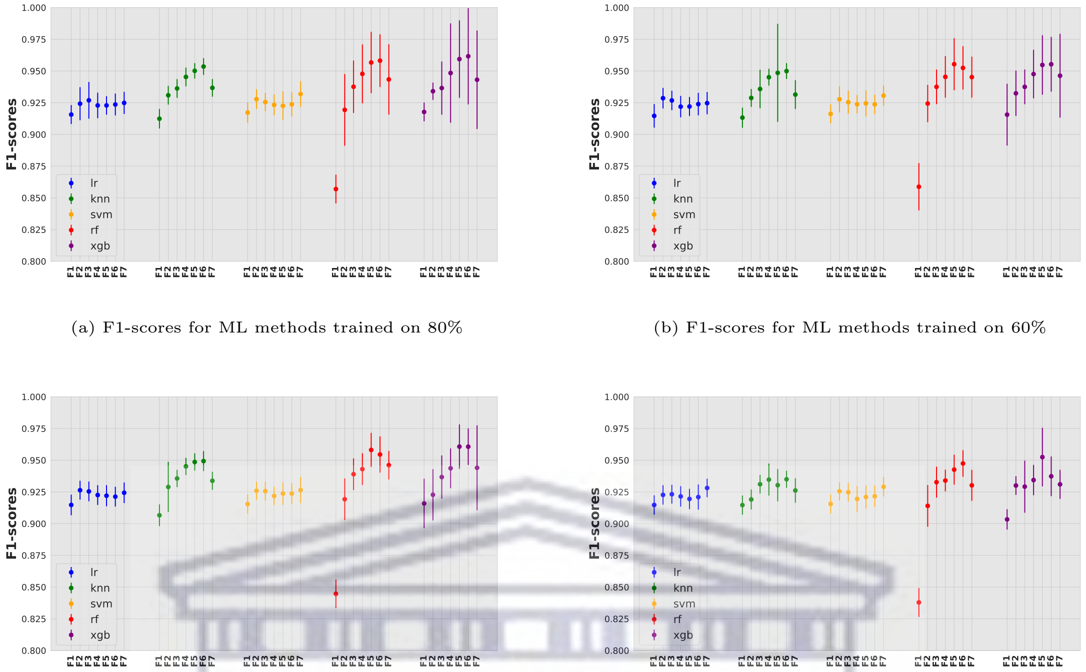
(b) F1-scores for ML methods trained on 60%

(c) F1-scores for ML methods trained on 40%

(d) F1-scores for ML methods trained on 20%

Figure 4.9: Final F1-scores results of applying supervised ML models; Logistic Regression (blue), k-Nearest Neighbour (green), Support Vector Machine (yellow), Random Forest (red) and XGBoost (magenta) based on seven feature combinations, evaluated on the *validation data.* Subplots a, b, c and d show the results of all five models trained on 80%, 60%, 40%, and 20% of the complete train data, respectively. The error bars represent the standard deviation calculated using the jackknife method.

We then considered the LR F1-score for feature combination F1 ($q_{\mathrm{IR}}$) on each complete train data split as a *baseline*. A baseline model should be a simple model that acts as a reference in an ML project. Its main function is to contextualize the results of trained models. Baseline models usually lack complexity and may have little predictive power.
Logistic regression is this study's simple and most interpretable ML model. Hence it is considered as a baseline. Because several feature combinations are available to train the model, we have considered the results of feature combination F1 as a baseline. Thus the baseline in this study is the LR results for $q_{\mathrm{IR}}$ in each data split. In each data split, we calculated the differences between the F1-score of LR for feature combination F1 and the F1-score for all feature other combinations and models. We compare the derived differences in Figure 4.11. Similar to Figures 4.9 and 4.10, the five models are shown in different colours, with the baseline F1-score highlighted by a solid horizontal black line.
In each of the splits, We calculate the F1-score differences between every model's



(a) F1-scores for ML methods trained on 80%

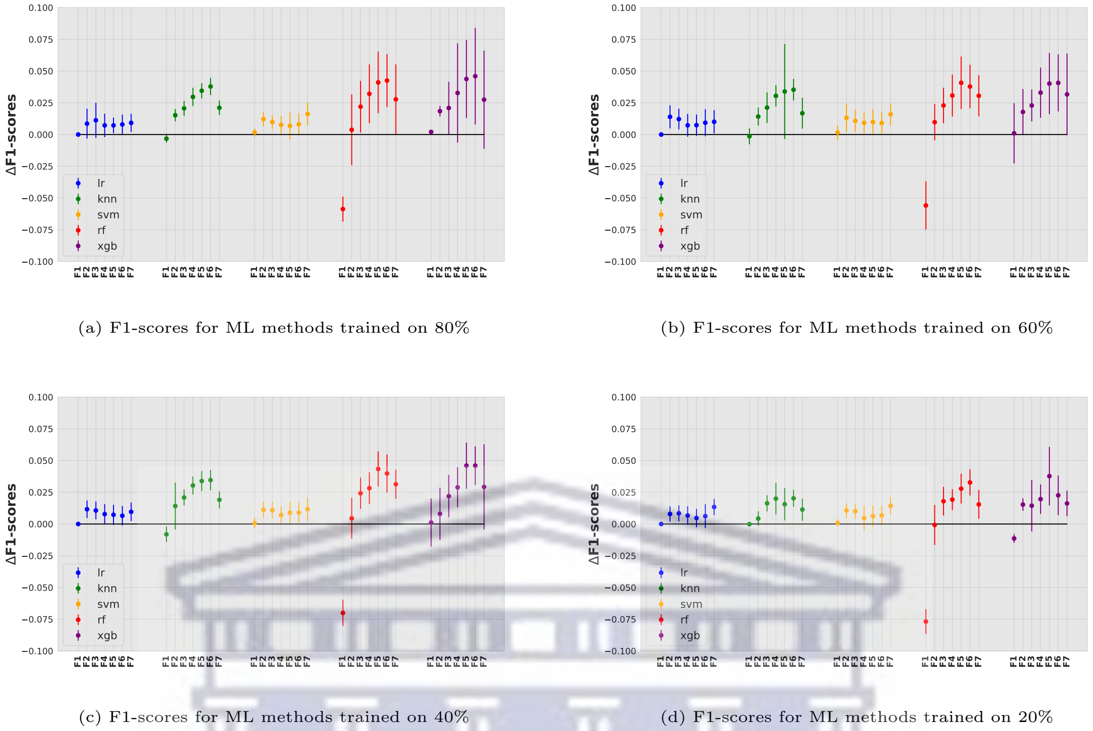
(b) F1-scores for ML methods trained on 60%

(c) F1-scores for ML methods trained on 40%

(d) F1-scores for ML methods trained on 20%

Figure 4.10: Final F1-scores results of applying supervised ML models; Logistic regression (blue), k-nearest neighbour (green), Support vector machine (yellow), Random forest (red) and XGBoost (magenta) based on seven feature combinations, evaluated on the *unseen data*. Subplots a, b, c and d show the results of all five models trained on 80%, 60%, 40%, and 20% of the complete train data, respectively. The error bars represent the standard deviation calculated using the jackknife method.

results for each feature combination and the LR model results on $q_{\mathrm{IR}}$, the F1-score differences is shown in the vertical axis of Figure 4.11. All models yield similar results when only $q_{\mathrm{IR}}$ is used as an input feature. The LR and SVM models produce identical results for all of the features considered in this study, and this may result from the fact that we have only considered the linear kernel for the SVM model. Figure 4.11 results suggest no significant difference between the performance of the LR model and that of he other four more sophisticated ML methods. However, the performance of logistic regression drops as features get more than four.



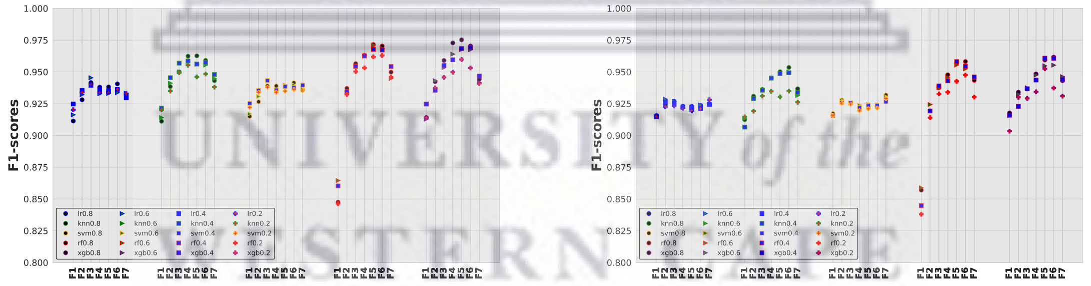
(a) F1-scores for ML methods trained on 80%

(b) F1-scores for ML methods trained on 60%

(c) F1-scores for ML methods trained on 40%

(d) F1-scores for ML methods trained on 20%

Figure 4.11: Final F1-scores differences of all supervised ML models compared to a *Logistic Regression baseline* (represented by a solid black line). In each split, we use the *Logistic Regression* F1-score obtained with the F1 feature combination as a baseline from which all differences are calculated.



(a) Final F1-scores results of applying ML models evaluated on the validation data

(b) Final F1-scores results of applying ML models evaluated on the unseen data.

Figure 4.12: Overall performance of ML models in validation and test data. The feature combination and models are the same as in figure 4.9 and 4.10, however, we show the results of all five ML models trained on 80% (black circles), 60% (green triangle), 40% (blue squares), and 20% (red plus) of the complete train data as shown in subplots 4.9a, 4.9b, 4.9c and 4.9d for validation data and in subplots 4.10a, 4.10b, 4.10c and 4.10d for unseen data, respectively. The error bars are omitted so that the global trend of results is clear.

# Chapter 5

# Conclusion

## 5.1 Summary

In this study, five supervised ML classification models, namely LR, SVM, $k$NN, RF and XGB, are applied and compared to classify radio sources as star-formation-dominated or black-hole-accretion-dominated sources, i.e. SFGs or AGN. We use a sample of 4,418 MIGHTEE-COSMOS radio sources classified as SFGs or AGN by Whittam et al. (2022) to train and test the ML models. The radio sources are from the MIGHTEE-COSMOS Early Science Data Release (Heywood et al., 2022b). *Prescott et al.* [Submitted] identified optical and NIR counterparts for these radio sources by using the multiwavelength catalogue by Bowler et al. (2020). Whittam et al. (2022) added the X-ray, MIR, and FIR measurements by Marchesi et al. (2016), Laigle et al. (2016), and Vaccari (2015), respectively, and used five multiwavelength diagnostics to perform the source classifications.

We use this radio sample and the associated multiwavelength measurements to evaluate the performance of ML models in classifying radio sources as SFGs and AGN;

1. We select and analyse the best features to train and test ML models. Our one-dimension, two-dimension, ML-independent, ML-dependent, and ROC curves analyses show that the FIR-Radio correlation parameter, $q_{\rm IR}$, is the most efficient feature in classifying SFGs and AGN, followed by the optical morphology parameter class_star and the combined MIR colours log(S8/S45) and log(S58/S36). The IRAC12 MIR log(S45/S36) colour is found to be the least important feature within those considered in this study.

2. We train and test the ML models with selected features. The features are grouped according to their importance as ranked by the feature importance

analysis. Our results show that most ML models perform very well with a combination of multiple features when classifying radio sources as SFGs and AGN.

3. Since the performance of ML depends on the sample size of the training data, we use 20%, 40%, 60% and 80% of the full data to train ML models. All models yield F1-scores greater than 90% with any training set size except for the RF model when trained with the single feature $q_{\mathrm{IR}}$ and with a training set size of 20%.

4. We also investigate the class imbalance bias in the MIGHTEE-COSMOS data. Our results show that in our case the class imbalance does not affect the performance of ML models.

5. Overall, our results show that all ML models perform very well in classifying SFGs and AGN from the radio sources with an F1-score greater than 90% even when using a small training set and only using the single most efficient input feature.

## 5.2 Discussion

### 5.2.1 Feature Selection

We select six features and derive seven feature combinations to train and evaluate ML models for classifying the radio sources detected in the MIGHTEE-COSMOS survey as SFGs and AGN. Whittam et al. (2022) provided the MIGHTEE-COSMOS radio sources manual classification using five conventional techniques, including X-ray measurements. X-ray observations are beneficial for identifying AGN because some of the brightest AGN display characteristic accretion-related X-ray emission. However, in this study, the X-ray luminosity is not included as the input feature because only 10% of the MIGHTEE-COSMOS have X-ray detections. If we removed sources with no X-ray measurements, a large proportion of the sample in the MIGHTEE-COSMOS catalogue would be removed. Instead, the X-ray measurements were not part of our analysis but we still achieved excellent performance. This is an important conclusion as deep and wide X-ray observations will be difficult to obtain at least over the next 15 years, i.e. before the launch of ESA's Athena X-ray observatory space mission[1]. Therefore, during the lifetime of MeerKAT and Phase 1 of the SKA, the classification of radio continuum sources will have to be carried out without the availability of X-ray data, and our results demonstrate that an accurate classification will still be possible. All ML-independent feature importance techniques show that

[1] https://www.the-athena-x-ray-observatory.eu/en

the $q_{\mathrm{IR}}$ parameter is the most efficient feature in separating the two classes of radio sources. These results are shown in Figure 4.5, 4.6 and 4.7 for the sequential feature importance (4.5.1), permutation feature importance (4.5.2) and the ROC-curves (4.5.3). The same insight is represented by a histogram and KS test in figure 4.2a. The random forest built-in feature importance model also returns the $q_{\mathrm{IR}}$ as the most significant feature. The significance of this feature has also been demonstrated when training the ML models. All the ML models can achieve F1-scores above 90% even when only trained with the $q_{\mathrm{IR}}$. As discussed, the $q_{\mathrm{IR}}$ parameter is derived from FIR and radio luminosities. Both luminosities trace star formation rates (SFRs) in a way largely immune from dust obscuration. The infrared emission is caused by dust that relatively massive OB stars have heated, while radio emission results from relativistic cosmic ray electrons (CRe) accelerated by shock waves produced when massive stars explode as supernovae (Delvecchio et al. (2021)). AGN activities accelerate the CRe electrons and cause a radio excess. Such sources with AGN signatures will have significantly more radio emission than would be expected from star formation alone. This makes the $q_{\mathrm{IR}}$ parameter the most accurate single predictor of the source class. Other features are also significant for classification, as shown by all the feature importance techniques. We see improvements in ML models when these features are added to the training data. Furthermore, compared to traditional input features, adding other derived colours from multiwavelength measurements, such as near-infrared, far-infrared, and optical photometry, did not improve the classification performance of ML models.

### 5.2.2 Feature Importance

The features are grouped according to their importance as ranked by the discussed feature importance analysis techniques (4.5.1, 4.5.2, 4.5.3, and 4.6). The resulting seven feature combinations were further used as input features to train each ML model over the four training set sizes of 80%, 60%, 40% and 20%. The LR, SVM, $k$NN, RF and XGB results for the validation and unseen test data are shown in figures 4.9 and 4.10, respectively. All models yield F1-scores of about 90% or more for all training set sizes except for the RF model when trained only with the F1 features. The LR and SVM models produce identical results in each training set size considered. This may be because we only considered the linear kernel of the SVM model, which is similar to the LR model approach for classification. Other kernels of the SVM were tried in a hyperparameter optimization setting using only an 80% training set size for each combination of features. The differences in the obtained results were insignificant; hence, we kept working only with linear kernels in this work. LR and SVM models' performances decrease when four or more attributes are used to train the models. Conversely, the $k$NN, RF, and XGB models yield great performances as more attributes are added, i.e. there is a slight increase in these models' performance from feature combination F1 to F6. Furthermore, all models show a slight decrease in performance when feature combination F7 is used as input to train the models. This implies that sources not observed or detected in the IRAC3 and IRAC4 bands but only in the IRAC1 and IRAC2 are at a disadvantage when classification is carried out with such supervised ML methods. The LR, SVM and $k$NN models yield a consistent F1-score as indicated by the error bars in figures 4.9 and 4.10. However, there are larger error bars for the RF and XGB. The error bars suggest that there is not a substantial advantage when using the more sophisticated techniques ($k$NN, SVM, RF and XGB) rather than the more interpretable LR model. We can therefore use the LR model to separate the SFGs from AGN by considering three attributes $q_{\mathrm{IR}}$, class_star and log(S8/S45) and still yield excellent F1-scores.

### 5.2.3 Class Imbalance

This work uses the *undersampling* approach to address the class imbalance. Data imbalance is a concern for our work as well as for future machine learning studies of radio continuum sources since star-forming galaxies usually dominate faint radio continuum samples. Our sample's AGN to SFGs ratio is about 1:2. Thus, our sample has two times more SFGs than AGN. The two classes are balanced in the training dataset, whilst the test dataset was unbalanced. We show the results in 4.8. The different models' resulting F1-scores are lower for all classes, decreasing by an identical fraction for all fractions of the training dataset. Overall, the data imbalance does not affect the models' performance, meaning that having more SFGs than AGN in the training dataset does not necessarily imply that SFGs will be classified more accurately than AGN. The better F1-score of SFGs classification shown in 4.8 is thus related to either observational or intrinsic properties of SFGs.

### 5.2.4 The need to apply ML techniques

The radio continuum surveys undertaken with large radio telescopes such as the MeerKAT telescope, ASKAP, and eventually, the SKA will result in an exponential increase in data. In phase one, the SKA is expected to detect 1 billion radio sources in a survey area of $3\pi$ steradians (Bourke et al., 2015). In contrast to the ~2.5 million presently-known radio sources, the Evolutionary Map of the Universe (EMU) survey on ASKAP is predicted to find roughly 70 million galaxies. Next-generation radio surveys will detect millions of radio sources while improving resolution, sensitivity to extended emission, and the ability to quantify spectral index and polarization for the strongest sources Norris et al. (2015). Future radio continuum surveys will regularly monitor the full sky, in the process producing extremely large data volumes.

Conventional data processing and visualization techniques will not be practical for such large datasets, but we will rather require accurate, efficient and automated approaches to analyze them. Between these approaches, machine learning algorithms provide some of the best opportunities to find patterns in digital data and translate these patterns into useful information (Ball and Brunner, 2010). Machine learning algorithms are better capable of handling extreme-scale data, and after their initial training is done they can be run with relatively limited computing resources.

This work implements and optimizes supervised ML models to accurately classify radio sources in the MIGHTEE-COSMOS catalogue as SFGs or AGN. The results demonstrate that supervised machine learning algorithms can learn the relationship between the data and human-provided radio source classes/labels and then apply the learned relationship to provide the classes/labels to the new unseen test data. Furthermore, we attempted to reduce the amount of data used to train the machine learning models to investigate if the training data size significantly impacts the overall classification ability. We showed that for the classification of MIGHTEE-COSMOS radio sources, reducing the size of the training set has no significant impact on the overall classification performance.

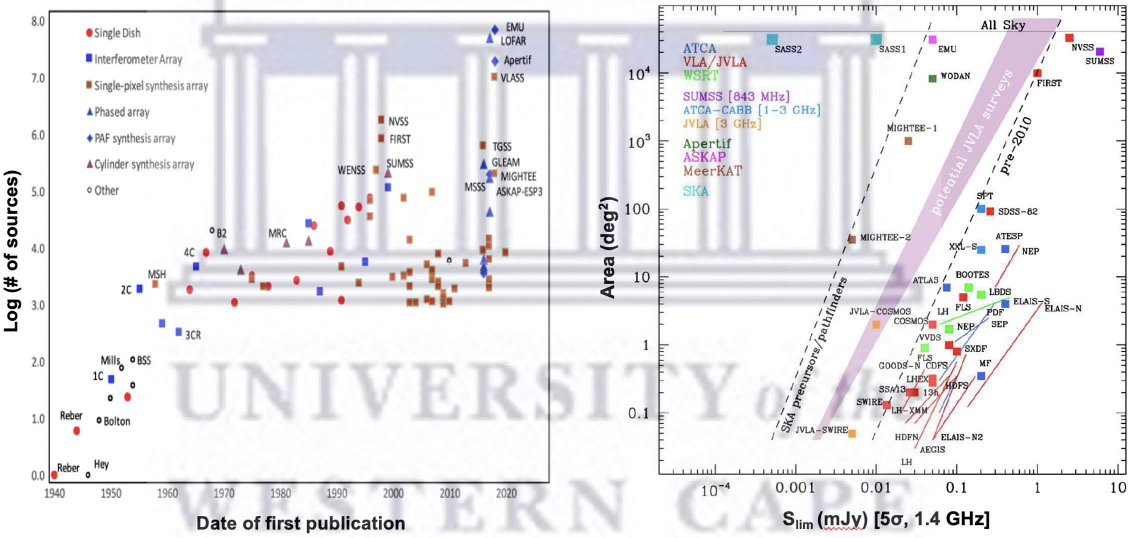


Figure 5.1: *Left plot*; A plot of the number of known extragalactic radio sources discovered by surveys as a function of time. An image adapted from (Norris, 2017). *Right Plot;* Deep 20 cm radio continuum surveys currently being conducted and planned (Norris et al., 2015).

### 5.2.5 Limitations of ML techniques

Although supervised machine learning models provide state-of-the-art results when classifying the MIGHTEE-COSMOS radio sources, some remaining possible limitations are worth pointing out. This section will limit the discussion to a few remaining challenges in classifying radio sources in the MIGHTEE-COSMOS sample as well as in future surveys.

As discussed in chapter 2, all supervised ML models deduce a function that maps the input data to the target output in the training data. A learned function is then used to predict the output for the test data, that were not part of the training data. Thus the accuracy of the mapping function of the algorithm depends on the quality of the target output provided in the training data. However, the target output in the training data is provided by conventional methods and may thus be imperfect and/or biased. For example, in order to produce our sample, Whittam et al. (2022) used five conventional techniques to assign manual labels to MIGHTEE-COSMOS radio sources. The different diagnostics were applied independently, which may have introduced some errors. Also, their capability to detect all AGN signatures in radio sources is limited by the observational data quality and by our current theoretical understanding of the AGN phenomenon. As a result, the excellent performance of our ML models may partially reflect our poor understanding and therefore our simplification of the phenomenon we are trying to study. This limitation may be partly addressed by developing sophisticated numerical simulations of SFG and AGN formation and evolution. However, the accuracy of such simulations is also limited by our theoretical understanding to date and by our current computational capabilities.

Another possible limitation is that missing data and invalid measurements may affect the accuracy and the general applicability of our results. Most ML assume that all values in an array are numerical and that all have and hold meaning. This problem can be avoided by removing the sample with the missing data or by replacing a missing value with e.g. the mean or median value over the full sample. However, this works well if the fraction of missing data is very small. Otherwise, this can introduce an observational bias in the data which will be reflected in our ML models. Our preliminary results showed that including more features in the analyses will improve the performance of the machine-learning model. However, the fraction of MIGHTEE-COSMOS sources with valid measurements in all features decreases with the number of additional features. Although the XGBoost algorithm can e.g. handle missing data, the algorithm still predicts the missing data using the measured features. Therefore, the performance of machine learning will still be affected by these missing data. In addition, limiting our analysis to radio sources with a large complement of multi-wavelength data will limit the applicability of our machine learning models. How to balance the performance of machine learning models and the completeness of our sample will be a continuing challenge for MIGHTEE and future radio surveys.

## 5.3 Future Plans

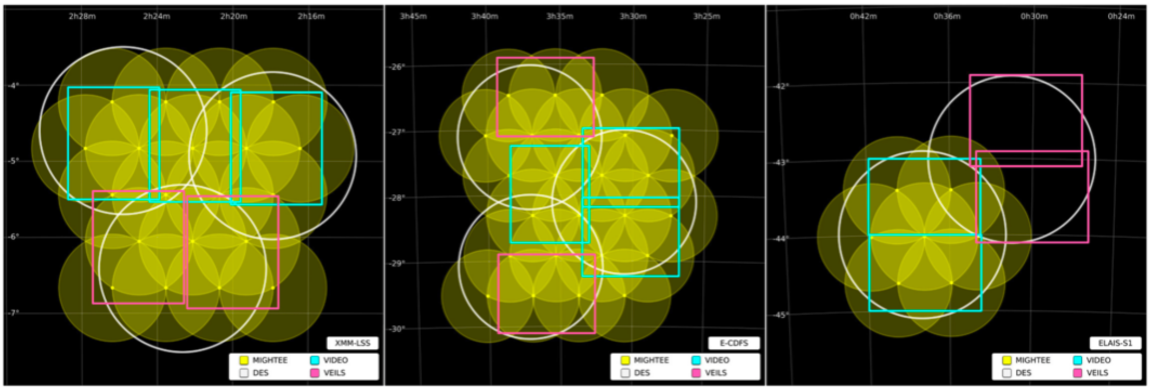


Figure 5.2: Pointing strategies for the three mosaics imaged MIGHTEE fields. Left: XMM-LSS (20 pointings, 6.7 deg$^2$). Middle: E-CDFS (24 pointings, 8.3 deg$^2$). Right: ELAIS-S1 (7 pointings, 1.6 deg$^2$). Image adapted from Jarvis et al. (2016).

Currently, our machine learning pipeline efficiently separates the radio continuum population in the COSMOS field with MIGHTEE Early Science Data as SFGs or AGN. However, some MIGHTEE radio continuum research goals require us to focus on AGN subclasses such as radio-loud AGN and radio-quiet AGN. As a result, we would like to extend the current machine learning pipeline to further classify AGN into the sub-classes of interest. Thus, we would like to expand the pipeline so that, given a sample of AGN, it could determine whether the corresponding AGN is e.g. radio-loud or radio-quiet or incorporate a more sophisticated AGN classification scheme.

The input features currently used by our classification pipeline are often derived from several independent original features such as flux density, morphology or other measurements. The derived quantities can introduce further biases and uncertainties or not make the most of the original features. This can also further reduce the sample size as multiple original features may be needed to obtain a single derived feature. To address this challenge, we will extend this work to include most of the original input features in our feature importance analysis.

The MIGHTEE-COSMOS Early Science Data Release and its cross-match with multiwavelength data has provided us with a sample of $\sim 4,000$ radio sources for our machine-learning analyses. The small data set may affect the performance of the machine learning analysis. The small size of the data is also arguably too small for running a deep learning pipeline. However, the entire MIGHTEE survey will soon reach deeper and wider in the COSMOS field as well as in three other survey fields; XMM-LSS, E-CDFS and ELAIS-S1. Following the multiwavelength identification of the MIGHTEE Final Data Release sample of radio sources, we will develop a deep learning pipeline and compare its performance with that of the more traditional machine learning algorithms used in this work.

# Bibliography

Akinsola, J. E. T. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48:128 – 138. 18

Alaliyat, S. (2008). Video-based fall detection in elderly's houses. Master's thesis, Gjovik University College. viii, 26, 27

Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., and Roberts, S. J. (2016). A sparse Gaussian process framework for photometric redshift estimation. *MNRAS*, 455(3):2387–2401. 39

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press. 16

Ashai, M., Mukherjee, R. G., Mundharikar, S. P., Kuanr, V. D., and Harikrishnan, R. (2022). Classification of astronomical objects using knn algorithm. In *Smart Intelligent Computing and Applications, Volume 1*, pages 377–387. Springer. 32

Bai, Y., Liu, J., Wang, S., and Yang, F. (2019). Machine Learning Applied to Star-Galaxy-QSO Classification and Stellar Effective Temperature Regression. *AJ*, 157(1):9. 32

Balamurali, M. and Melkumyan, A. (2016). t-sne based visualisation and clustering of geological domain. In *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part IV 23*, pages 565–572. Springer. 56

Ball, N. M. and Brunner, R. J. (2010). Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19(7):1049–1106. 21, 22, 23, 32, 74

Baştanlar, Y. and Özuysal, M. (2014). Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, pages 105–128. 18

Becker, B. and Grobler, T. (2019). Classification of fanaroff-riley radio galaxies using conventional machine learning techniques. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–8. IEEE. vi, 10

Bennett, A. S. (1962). The revised 3C catalogue of radio sources. *MmRAS*, 68:163. 10

Berger, V. W. and Zhou, Y. (2014). Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*. 50

Berrar, D. et al. (2019). Cross-validation. 31

Bessell, M. S. (2005). Standard Photometric Systems. *ARA&A*, 43(1):293–336. 37

Bonato, M., Prandoni, I., De Zotti, G., Brienza, M., Morganti, R., and Vaccari, M. (2021). New constraints on the 1.4 GHz source number counts and luminosity functions in the Lockman Hole field. *MNRAS*, 500(1):22–33. 3

Borne, K. (2013). Virtual Observatories, Data Mining, and Astroinformatics. In Oswalt, T. D. and Bond, H. E., editors, *Planets, Stars and Stellar Systems. Volume 2: Astronomical Techniques, Software and Data*, page 403. Springer. 18, 19, 26, 32

Bourke, T. L. et al., editors (2015). *Proceedings, Advancing Astrophysics with the Square Kilometre Array (AASKA14): Giardini Naxos, Italy, June 9-13, 2014*. SISSA. 2, 73

Bowler, R. A. A., Jarvis, M. J., Dunlop, J. S., McLure, R. J., McLeod, D. J., Adams, N. J., Milvang-Jensen, B., and McCracken, H. J. (2020). A lack of evolution in the very bright end of the galaxy luminosity function from z $\simeq$ 8 to 10. *MNRAS*, 493(2):2059–2084. 41, 47, 70

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32. 27

Buradkar, V. T. and More, M. (2020). Introduction to machine learning and its applications: A survey. *Journal of Artificial Intelligence, Machine Learning and Soft Computing*, 5(1):12. 18

Calistro Rivera, G., Williams, W. L., Hardcastle, M. J., Duncan, K., Röttgering, H. J. A., Best, P. N., Brüggen, M., Chyży, K. T., Conselice, C. J., de Gasperin, F., Engels, D., Gürkan, G., Intema, H. T., Jarvis, M. J., Mahony, E. K., Miley, G. K., Morabito, L. K., Prandoni, I., Sabater, J., Smith, D. J. B., Tasse, C., van der Werf, P. P., and White, G. J. (2017). The LOFAR window on star-forming galaxies and AGNs - curved radio SEDs and IR-radio correlation at 0¡z¡2.5. *MNRAS*, 469(3):3468–3488. vi, 5

Capak, P., Aussel, H., Ajiki, M., McCracken, H. J., Mobasher, B., Scoville, N., Shopbell, P., Taniguchi, Y., Thompson, D., Tribiano, S., Sasaki, S., Blain, A. W., Brusa, M., Carilli, C., Comastri, A., Carollo, C. M., Cassata, P., Colbert, J., Ellis, R. S., Elvis, M., Giavalisco, M., Green, W., Guzzo, L., Hasinger, G., Ilbert, O., Impey, C., Jahnke, K., Kartaltepe, J., Kneib, J. P., Koda, J., Koekemoer, A., Komiyama, Y., Leauthaud, A., Le Fevre, O., Lilly, S., Liu, C., Massey, R., Miyazaki, S., Murayama, T., Nagao, T., Peacock, J. A., Pickles, A., Porciani, C., Renzini, A., Rhodes, J., Rich, M., Salvato, M., Sanders, D. B., Scarlata, C., Schiminovich, D., Schinnerer, E., Scodeggio, M., Sheth, K., Shioya, Y., Tasca, L. A. M., Taylor, J. E., Yan, L., and Zamorani, G. (2007). The First Release COSMOS Optical and Near-IR Data and Catalog. *ApJS*, 172(1):99–116. 36

Cardoso-Fernandes, J., Teodoro, A. C., Lima, A., and Roda-Robles, E. (2020). Semi-Automatization of Support Vector Machines to Map Lithium (Li) Bearing Pegmatites. *Remote Sensing*, 12(14):2319. vii, 26

Carroll, B. W. and Ostlie, D. A. (2017). *An introduction to modern astrophysics*. Cambridge University Press. 1, 9

Cavanagh, M. K., Bekki, K., and Groves, B. A. (2021). Morphological classification of galaxies with deep learning: comparing 3-way and 4-way CNNs. *MNRAS*, 506(1):659–676. 32

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, page arXiv:1603.02754. 28

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., and Griguta, V. (2020). Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra. *A&A*, 639:A84. 32

Condon, J. J. (1992). Radio emission from normal galaxies. *ARA&A*, 30:575–611. vi, 1, 4, 11

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute*. 24

Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press. 25

Das, S., Mullick, S. S., and Zelinka, I. (2022). On supervised class-imbalanced learning: An updated perspective and some key challenges. *IEEE Transactions on Artificial Intelligence*, 3(6):973–993. 63

Davies, L. J. M., Robotham, A. S. G., Driver, S. P., Lagos, C. P., Cortese, L., Mannering, E., Foster, C., Lidman, C., Hashemizadeh, A., Koushan, S., O'Toole, S., Baldry, I. K., Bilicki, M., Bland-Hawthorn, J., Bremer, M. N., Brown, M. J. I., Bryant, J. J., Catinella, B., Croom, S. M., Grootes, M. W., Holwerda, B. W., Jarvis, M. J., Maddox, N., Meyer, M., Moffett, A. J., Phillipps, S., Taylor, E. N., Windhorst, R. A., and Wolf, C. (2018). Deep Extragalactic VIsible Legacy Survey (DEVILS): motivation,design, and target catalogue. *MNRAS*, 480(1):768–799. 39

Delvecchio, I., Daddi, E., Sargent, M. T., Jarvis, M. J., Elbaz, D., Jin, S., Liu, D., Whittam, I. H., Algera, H., Carraro, R., D'Eugenio, C., Delhaize, J., Kalita, B. S., Leslie, S., Molnár, D. C., Novak, M., Prandoni, I., Smolčić, V., Ao, Y., Aravena, M., Bournaud, F., Collier, J. D., Randriamampandry, S. M., Randriamanakoto, Z., Rodighiero, G., Schober, J., White, S. V., and Zamorani, G. (2021). The infrared-radio correlation of star-forming galaxies is strongly $M_\star$-dependent but nearly redshift-invariant since $z \sim 4$. *A&A*, 647:A123. ix, 43, 45, 47, 61, 72

Dermer, C. D. and Giebels, B. (2016). Active galactic nuclei at gamma-ray energies. *Comptes Rendus Physique*, 17(6):594–616. vi, 6, 7

Dixon, M. F., Halperin, I., and Bilokon, P. (2020). *Machine learning in Finance*, volume 1170. Springer. 17

Donley, J. L., Koekemoer, A. M., Brusa, M., Capak, P., Cardamone, C. N., Civano, F., Ilbert, O., Impey, C. D., Kartaltepe, J. S., Miyaji, T., Salvato, M., Sanders, D. B., Trump, J. R., and Zamorani, G. (2012). Identifying Luminous Active Galactic Nuclei in Deep Surveys: Revised IRAC Selection Criteria. *ApJ*, 748(2):142. 1, 42, 47, 53, 62

Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359. 20, 24, 26

Elvis, M., Wilkes, B. J., McDowell, J. C., Green, R. F., Bechtold, J., Willner, S. P., Oey, M. S., Polomski, E., and Cutri, R. (1994). Atlas of Quasar Energy Distributions. *ApJS*, 95:1. vi, 5, 6

Evgeniou, T. and Pontil, M. (1999). Support vector machines: Theory and applications. In *Advanced Course on Artificial Intelligence*, pages 249–257. Springer. 25

Fadely, R., Hogg, D. W., and Willman, B. (2012). Star-Galaxy Classification in Multi-band Optical Imaging. *ApJ*, 760(1):15. 32

Fanaroff, B. L. and Riley, J. M. (1974). The morphology of extragalactic radio sources of high and low luminosity. *MNRAS*, 167:31P–36P. 9

Fluke, C. J. and Jacobs, C. (2020). Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. *WIREs Data Mining and Knowledge Discovery*, 10(2):e1349. 31, 32

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger. 19

Genzel, R., Förster Schreiber, N. M., Lang, P., Tacchella, S., Tacconi, L. J., Wuyts, S., Bandara, K., Burkert, A., Buschkamp, P., Carollo, C. M., Cresci, G., Davies, R., Eisenhauer, F., Hicks, E. K. S., Kurk, J., Lilly, S. J., Lutz, D., Mancini, C., Naab, T., Newman, S., Peng, Y., Renzini, A., Shapiro Griffin, K., Sternberg, A., Vergani, D., Wisnioski, E., Wuyts, E., and Zamorani, G. (2014). The SINS/zC-SINF Survey of z ~2 Galaxy Kinematics: Evidence for Gravitational Quenching. *ApJ*, 785(1):75. 4

Golob, A., Sawicki, M., Goulding, A. D., and Coupon, J. (2021). Classifying stars, galaxies, and AGNs in CLAUDS + HSC-SSP using gradient boosted decision trees. *MNRAS*, 503(3):4136–4146. 32

Gopal-Krishna and Wiita, P. J. (2000). Extragalactic radio sources with hybrid morphology: implications for the Fanaroff-Riley dichotomy. *A&A*, 363:507–516. 10

Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J., and Gao, D. (2020). Degradation state recognition of piston pump based on iceemdan and xgboost. *Applied Sciences*, 10(18):6593. viii, 29

Heywood, I., Jarvis, M. J., Hale, C. L., Whittam, I. H., Bester, H. L., Hugo, B., Kenyon, J. S., Prescott, M., Smirnov, O. M., Tasse, C., Afonso, J. M., Best, P. N., Collier, J. D., Deane, R. P., Frank, B. S., Hardcastle, M. J., Knowles, K., Maddox, N., Murphy, E. J., Prandoni, I., Randriamampandry, S. M., Santos, M. G., Sekhar, S., Tabatabaei, F., Taylor, A. R., and Thorat, K. (2022a). MIGHTEE: total intensity radio continuum imaging and the COSMOS/XMM-LSS Early Science fields. *MNRAS*, 509(2):2150–2168. viii, 39

Heywood, I., Jarvis, M. J., Hale, C. L., Whittam, I. H., Bester, H. L., Hugo, B., Kenyon, J. S., Prescott, M., Smirnov, O. M., Tasse, C., Afonso, J. M., Best, P. N., Collier, J. D., Deane, R. P., Frank, B. S., Hardcastle, M. J., Knowles, K., Maddox, N., Murphy, E. J., Prandoni, I., Randriamampandry, S. M., Santos, M. G., Sekhar, S., Tabatabaei, F., Taylor, A. R., and Thorat, K. (2022b). MIGHTEE: total intensity radio continuum imaging and the COSMOS/XMM-LSS Early Science fields. *MNRAS*, 509(2):2150–2168. 35, 38, 41, 70

Hughes, A. C. N., Bailer-Jones, C. A. L., and Jamal, S. (2022). Quasar and galaxy classification using Gaia EDR3 and CatWise2020. *A&A*, 668:A99. 32

Hussein, E. A., Thron, C., Ghaziasgar, M., Vaccari, M., Marnewick, J. L., and Hussein, A. A. (2021). Comparison of phenolic content and antioxidant activity for fermented and unfermented rooibos samples extracted with water and methanol. *Plants*, 11(1):16. 33, 65

Hussein, E. A., Thron, C., Ghaziasgar, M., Vaccari, M., Marnewick, J. L., and Hussein, A. A. (2022). Comparison of phenolic content and antioxidant activity for fermented and unfermented rooibos samples extracted with water and methanol. *Plants*, 11(1):16. 91

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer. 24, 25

Jarvis, M., Taylor, R., Agudo, I., Allison, J. R., Deane, R. P., Frank, B., Gupta, N., Heywood, I., Maddox, N., McAlpine, K., Santos, M., Scaife, A. M. M., Vaccari, M., Zwart, J. T. L., Adams, E., Bacon, D. J., Baker, A. J., Bassett, B. A., Best, P. N., Beswick, R., Blyth, S., Brown, M. L., Bruggen, M., Cluver, M., Colafrancesco, S., Cotter, G., Cress, C., Davé, R., Ferrari, C., Hardcastle, M. J., Hale, C. L., Harrison, I., Hatfield, P. W., Klockner, H. R., Kolwa, S., Malefahlo, E., Marubini, T., Mauch, T., Moodley, K., Morganti, R., Norris, R. P., Peters, J. A., Prandoni, I., Prescott, M., Oliver, S., Oozeer, N., Rottgering, H. J. A., Seymour, N., Simpson, C., Smirnov, O., and Smith, D. J. B. (2016). The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey. In *MeerKAT Science: On the Pathway to the SKA*, page 6. xii, 14, 34, 35, 76

Karttunen, H., Kröger, P., Oja, H., Poutanen, M., and Donner, K. J. (2007). *Fundamental astronomy*. Springer. 1

Kellermann, K. I. and Wall, J. V. (1987). Radio Source Counts and Their Interpretation /. In Hewitt, A., Burbidge, G., and Fang, L. Z., editors, *Observational Cosmology*, volume 124, page 545. 13

Kembhavi, A. K. and Narlikar, J. V. (1999). *Quasars and active galactic nuclei : an introduction*. Cambridge University Press. 9

Kim, E. J. and Brunner, R. J. (2017). Star-galaxy classification using deep convolutional neural networks. *MNRAS*, 464(4):4463–4475. 32

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24. 18, 20, 23

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232. 63

Laigle, C., McCracken, H. J., Ilbert, O., Hsieh, B. C., Davidzon, I., Capak, P., Hasinger, G., Silverman, J. D., Pichon, C., Coupon, J., Aussel, H., Le Borgne, D., Caputi, K., Cassata, P., Chang, Y. Y., Civano, F., Dunlop, J., Fynbo, J., Kartaltepe, J. S., Koekemoer, A., Le Fèvre, O., Le Floc'h, E., Leauthaud, A., Lilly, S., Lin, L., Marchesi, S., Milvang-Jensen, B., Salvato, M., Sanders, D. B., Scoville, N., Smolcic, V., Stockmann, M., Taniguchi, Y., Tasca, L., Toft, S., Vaccari, M., and Zabl, J. (2016). The COSMOS2015 Catalog: Exploring the 1 ¡ z ¡ 6 Universe with Half a Million Galaxies. *ApJS*, 224(2):24. 41, 70

Li, C., Zhang, W. H., and Lin, J. M. (2019). Research on Star/Galaxy Classification Based on XGBoost Algorithm. *Acta Astronomica Sinica*, 60(2):16. 32

Li, L., Zhang, Y., and Zhao, Y. (2008). k-Nearest Neighbors for automated classification of celestial objects. *Science in China: Physics, Mechanics and Astronomy*, 51(7):916–922. 32

Marchesi, S., Civano, F., Elvis, M., Salvato, M., Brusa, M., Comastri, A., Gilli, R., Hasinger, G., Lanzuisi, G., Miyaji, T., Treister, E., Urry, C. M., Vignali, C., Zamorani, G., Allevato, V., Cappelluti, N., Cardamone, C., Finoguenov, A., Griffiths, R. E., Karim, A., Laigle, C., LaMassa, S. M., Jahnke, K., Ranalli, P., Schawinski, K., Schinnerer, E., Silverman, J. D., Smolcic, V., Suh, H., and Trakhtenbrot, B. (2016). The Chandra COSMOS Legacy survey: optical/IR identifications. *ApJ*, 817(1):34. 41, 70

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, 61(1):1–15. 33

Mingo, B., Croston, J. H., Best, P. N., Duncan, K. J., Hardcastle, M. J., Kondapally, R., Prandoni, I., Sabater, J., Shimwell, T. W., Williams, W. L., Baldi, R. D., Bonato, M., Bondi, M., Dabhade, P., Gürkan, G., Ineson, J., Magliocchetti, M., Miley, G., Pierce, J. C. S., and Röttgering, H. J. A. (2022). Accretion mode versus radio morphology in the LOFAR Deep Fields. *MNRAS*, 511(3):3250–3271. 10

Mingo, B., Croston, J. H., Hardcastle, M. J., Best, P. N., Duncan, K. J., Morganti, R., Rottgering, H. J. A., Sabater, J., Shimwell, T. W., Williams, W. L., Brienza, M., Gurkan, G., Mahatma, V. H., Morabito, L. K., Prandoni, I., Bondi, M., Ineson, J., and Mooney, S. (2019). Revisiting the Fanaroff-Riley dichotomy and radio-galaxy morphology with the LOFAR Two-Metre Sky Survey (LoTSS). *MNRAS*, 488(2):2701–2721. 10

Mo, H., van den Bosch, F. C., and White, S. (2010). *Galaxy Formation and Evolution*. Cambridge University Press. 1

Mohan, N. and Rafferty, D. (2015). PyBDSF: Python Blob Detection and Source Finder. Astrophysics Source Code Library, record ascl:1502.007. 38

Moosavi, A., Rao, V., and Sandu, A. (2021). Machine learning based algorithms for uncertainty quantification in numerical weather prediction models. *Journal of Computational Science*, 50:101295. 16

Norris, R., Basu, K., Brown, M., Carretti, E., Kapinska, A. D., Prandoni, I., Rudnick, L., and Seymour, N. (2015). The SKA Mid-frequency All-sky Continuum Survey: Discovering the unexpected and transforming radio-astronomy. In *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, page 86. xii, 73, 74

Norris, R. P. (2017). Extragalactic radio continuum surveys and the transformation of radio astronomy. *Nature Astronomy*, 1:671–678. xii, 11, 74

Norris, R. P., Salvato, M., Longo, G., Brescia, M., Budavari, T., Carliles, S., Cavuoti, S., Farrah, D., Geach, J., Luken, K., Musaeva, A., Polsterer, K., Riccio, G., Seymour, N., Smolčić, V., Vaccari, M., and Zinn, P. (2019). A Comparison of Photometric Redshift Techniques for Large Radio Surveys. *PASP*, 131(1004):108004. 24

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., and Zumach, W. A. (1992). Automated Star/Galaxy Discrimination With Neural Networks. *AJ*, 103:318. 32

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, 3(3):e10. 16

Olivas, E. S., Guerrero, J. D. M., Martinez-Sober, M., Magdalena-Benedito, J. R., Serrano, L., et al. (2009). *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI global. 17

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., and Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138. 24

Padovani, P. (2016a). The faint radio sky: radio astronomy becomes mainstream. *A&A Rev.*, 24(1):13. 4, 11, 13

Padovani, P. (2016b). The faint radio sky: radio astronomy becomes mainstream. *A&A Rev.*, 24(1):13. 5

Padovani, P., Alexander, D. M., Assef, R. J., De Marco, B., Giommi, P., Hickox, R. C., Richards, G. T., Smolčić, V., Hatziminaoglou, E., Mainieri, V., and Salvato, M. (2017). Active galactic nuclei: what's in a name? *A&A Rev.*, 25(1):2. vi, 1, 6, 7, 8, 11

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 23

Peng, N., Zhang, Y., and Zhao, Y. (2013). A SVM-kNN method for quasar-star classification. *Science China Physics, Mechanics, and Astronomy*, 56(6):1227–1234. 32

Peruzzi, T., Pasquato, M., Ciroi, S., Berton, M., Marziani, P., and Nardini, E. (2021). Interpreting automatic AGN classifiers with saliency maps. *A&A*, 652:A19. 5

Peterson, B. M. (1997). *An introduction to active galactic nuclei*. Cambridge University Press. 7

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883. 26

Pham, S. T., Vo, P. S., and Nguyen, D. N. (2021). Effective electrical submersible pump management using machine learning. *Open Journal of Civil Engineering*, 11(1):70–80. 27

Prandoni, I., Gregorini, L., Parma, P., de Ruiter, H. R., Vettolani, G., Wieringa, M. H., and Ekers, R. D. (2001). The ATESP radio survey. III. Source counts. *A&A*, 365:392–399. 13

Prandoni, I., Guglielmino, G., Morganti, R., Vaccari, M., Maini, A., Röttgering, H. J. A., Jarvis, M. J., and Garrett, M. A. (2018). The Lockman Hole Project:

new constraints on the sub-mJy source counts from a wide-area 1.4 GHz mosaic. *MNRAS*, 481(4):4548–4565. vii, 13

Prescott, M., Whittam, I. H., Jarvis, M. J., McAlpine, K., Richter, L. L., Fine, S., Mauch, T., Heywood, I., and Vaccari, M. (2018). The Stripe 82 1-2 GHz Very Large Array Snapshot Survey: multiwavelength counterparts. *MNRAS*, 480(1):707–721. 11

Pyle, D. (1999). *Data preparation for data mining*. morgan kaufmann. 22

Richards, G. T., Lacy, M., Storrie-Lombardi, L. J., Hall, P. B., Gallagher, S. C., Hines, D. C., Fan, X., Papovich, C., Vanden Berk, D. E., Trammell, G. B., Schneider, D. P., Vestergaard, M., York, D. G., Jester, S., Anderson, S. F., Budavári, T., and Szalay, A. S. (2006). Spectral Energy Distributions and Multiwavelength Selection of Type 1 Quasars. *ApJS*, 166(2):470–497. vi, 5, 6

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229. 16

Sanders, D. B., Salvato, M., Aussel, H., Ilbert, O., Scoville, N., Surace, J. A., Frayer, D. T., Sheth, K., Helou, G., Brooke, T., Bhattacharya, B., Yan, L., Kartaltepe, J. S., Barnes, J. E., Blain, A. W., Calzetti, D., Capak, P., Carilli, C., Carollo, C. M., Comastri, A., Daddi, E., Ellis, R. S., Elvis, M., Fall, S. M., Franceschini, A., Giavalisco, M., Hasinger, G., Impey, C., Koekemoer, A., Le Fèvre, O., Lilly, S., Liu, M. C., McCracken, H. J., Mobasher, B., Renzini, A., Rich, M., Schinnerer, E., Shopbell, P. L., Taniguchi, Y., Thompson, D. J., Urry, C. M., and Williams, J. P. (2007). S-COSMOS: The Spitzer Legacy Survey of the Hubble Space Telescope ACS 2 deg$^2$ COSMOS Field I: Survey Strategy and First Analysis. *ApJS*, 172(1):86–98. 36

Saripalli, L. (2012). Understanding the Fanaroff-Riley Radio Galaxy Classification. *AJ*, 144(3):85. 11

Schneider, P. (2006). *Extragalactic astronomy and cosmology: an introduction*, volume 146. Springer. vii, 1, 12

Schober, J., Sargent, M. T., Klessen, R. S., and Schleicher, D. R. G. (2022). A model for the infrared-radio correlation of main-sequence galaxies at GHz frequencies and its dependence on redshift and stellar mass. *arXiv e-prints*, page arXiv:2210.07919. 61

Scoville, N., Aussel, H., Brusa, M., Capak, P., Carollo, C. M., Elvis, M., Giavalisco, M., Guzzo, L., Hasinger, G., Impey, C., Kneib, J. P., LeFevre, O., Lilly, S. J., Mobasher, B., Renzini, A., Rich, R. M., Sanders, D. B., Schinnerer, E., Schminovich, D., Shopbell, P., Taniguchi, Y., and Tyson, N. D. (2007). The Cosmic Evolution Survey (COSMOS): Overview. *ApJS*, 172(1):1–8. 35

Sen, S., Agarwal, S., Chakraborty, P., and Singh, K. P. (2022). Astronomical big data processing using machine learning: A comprehensive review. *Experimental Astronomy*, 53(1):1–43. 32

Shatnawi, A., Alkassar, H. M., Al-Abdaly, N. M., Al-Hamdany, E. A., Bernardo, L. F. A., and Imran, H. (2022). Shear strength prediction of slender steel fiber reinforced concrete beams using a gradient boosting regression tree method. *Buildings*, 12(5):550. viii, 31

Sparke, L. S. and Gallagher III, J. S. (2007). *Galaxies in the universe: an introduction*. Cambridge University Press. 1, 5, 7

Szokoly, G. P., Bergeron, J., Hasinger, G., Lehmann, I., Kewley, L., Mainieri, V., Nonino, M., Rosati, P., Giacconi, R., Gilli, R., Gilmozzi, R., Norman, C., Romaniello, M., Schreier, E., Tozzi, P., Wang, J. X., Zheng, W., and Zirm, A. (2004). The Chandra Deep Field-South: Optical Spectroscopy. I. *ApJS*, 155(2):271–349. 42

Tadhunter, C. (2008). An introduction to active galactic nuclei: Classification and unification. *New Astronomy Reviews*, 52(6):227–239. 9

Urry, C. M. and Padovani, P. (1995). Unified Schemes for Radio-Loud Active Galactic Nuclei. *PASP*, 107:803. 6

Vaccari, M. (2015). The Spitzer Data Fusion: Contents, Construction and Applications to Galaxy Evolution Studies. In *The Many Facets of Extragalactic Radio Surveys: Towards New Scientific Challenges*, page 27. 41, 70

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11). 56

Vijayakumar, T. (2019). Classification of brain cancer type using machine learning. *Journal of Artificial Intelligence*, 1(02):105–113. 17

Wang, L. and Alexander, C. A. (2016). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, 1(2):52–61. 16

Wang, Y., Pan, Z., Zheng, J., Qian, L., and Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. *Ap&SS*, 364(8):139. 28

Way, M. J., Scargle, J. D., Ali, K. M., and Srivastava, A. N. (2012). *Advances in Machine Learning and Data Mining for Astronomy*. taylor20 & Francis. 2, 11

Weaver, J. R., Kauffmann, O., Ilbert, O., McCracken, H. J., Moneti, A., Toft, S., Brammer, G., Shuntov, M., Davidzon, I., Hsieh, B.-C., et al. (2022). COSMOS2020: A Panchromatic View of the Universe to z 10 from Two Complementary Catalogs. *ApJS*, 258(1):11. viii, 36, 37

Weir, N., Fayyad, U. M., Djorgovski, S. G., and Roden, J. (1995). The SKICAT System for Processing and Analyzing Digital Imaging Sky Surveys. *PASP*, 107:1243. 32

Whitmore, B. C. (1984). An objective classification system for spiral galaxies. I. The two dominant dimensions. *ApJ*, 278:61–80. 32

Whittam, I. H., Jarvis, M. J., Hale, C. L., Prescott, M., Morabito, L. K., Heywood, I., Adams, N. J., Afonso, J., An, F., Ao, Y., Bowler, R. A. A., Collier, J. D., Deane, R. P., Delhaize, J., Frank, B., Glowacki, M., Hatfield, P. W., Maddox, N., Marchetti, L., Matthews, A. M., Prandoni, I., Randriamampandry, S., Randriamanakoto, Z., Smith, D. J. B., Taylor, A. R., Thomas, N. L., and Vaccari, M. (2022). MIGHTEE: the nature of the radio-loud AGN population. *MNRAS*, 516(1):245–263. x, xii, 14, 21, 35, 38, 39, 41, 42, 43, 44, 47, 48, 49, 51, 55, 57, 70, 71, 75, 96

Winston, P. H. (1984). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc. 18

Xiao-Qing, W. and Jin-Meng, Y. (2021). Classification of star/galaxy/QSO and star spectral types from LAMOST data release 5 with machine learning approaches. *Chinese Journal of Physics*, 69:303–311. 32

Zhang, S., Li, X., Zong, M., Zhu, X., and Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):1–19. 23

Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381. 22, 50

Zhang, Y., Zheng, H., and Zhao, Y. (2008). Knowledge discovery in astronomical data. In Bridger, A. and Radziwill, N. M., editors, *Advanced Software and Control for Astronomy II*, volume 7019 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 701938. 23

# Appendix A

# Machine Learning Pipeline

The machine-learning Python pipeline developed in this study can be accessed on the GitHub link attached below. The underlying data is still private, and we would like to keep the code also private until we publish the paper. However, we will make this available upon request.
Username: pfunzowalter
Account Holder: Walter Silima
GitHub: https://github.com/pfunzowalter/MIGHTEE-CLASSIFICATION

This study adopted several supervised machine learning algorithms such as Logistic Regression, Super Vector Machine, K-nearest Neighbour, XGBoost, and Random forest to classify the radio sources detected in MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) survey as Star-Forming Galaxies or Active Galactic Nuclei. We take advantage of the available multiwavelength measurements and use them as input features to train the supervised machine learning algorithms. This pipeline was strictly developed based on the MIGHTEE-COSMOS measurements. However, the analysis approach used to develop this pipeline can be adapted to solve any binary classification problem with machine learning.
The pipeline workflow was divided into a series of notebooks, each with its own focus. The experiments were organised as follows;

- EXP1 - This is a data preprocessing and feature visualization notebook - I have, in this notebook, After an understanding of the data, sample the data that is clean and free from missingness and define the spectroscopic features that we going to use as input features.

- EXP2 - This is a Feature Analysis notebook. In this section, we perform feature analysis using Correlation Matrix, SeaBorn Pair plot, Permutation Importance, and ROC curves.

- EXP3 - This is a Data Correlation notebook. We make a plot of two features for each target label. Using simple statistics such as Pearson correlation coefficients, we asses the correlation of each pair of features.

- EXP4 - We now adopt the machine learning algorithms mentioned above and do classification. We also went on to balance the data and redo the calculation to asses if data imbalance significantly affects this classification problem

- EXP5. In this section, I adopted the Hussein et al. (2022) method to classify MIGHTEE sources using simple statistics such as Mahalanobis distance.

- EXP6 reference - This serves as a machine-learning reference notebook. If you would like to know how the code is run to produce results in "EXP 6-final notebook", it could be worth looking, but the results are also in "EXP 6-final notebook". **EXP6 Final **- This is an important notebook where we compare all the supervised machine-learning algorithms. We applied different machine learning algorithms in different combinations of features and different train sizes to classify AGN or SFGs.

- Appendix Notebook - This notebook contains several results, including the plots in section B

- MIGHTEE-CRITERIA - In this notebook, we analyse the MIGHTEE-COSMOS data, reproducing some plots used in literature to classify radio sources as star-forming galaxies or active galactic nuclei.

- Hyper notebook - focuses on yielding the results for hyperparameter-optimisation. It uses a function in 'sources' which takes the hyperparameter and the model you would like to optimise as input and then yield different results for different options or values of that hyperparameter with other models' hyperparameters as defaults. The resulting plot shows which values of hyperparameters give better results.

An important directory *sources* is also available, which contains all Python scripts with the important functions used in all the experiments notebooks.

# Appendix B

# Additional Input Features

In this chapter, we introduce other features that were available to use as input features to train the machine learning models. The MIGHTEE-COSMOS spans several wavebands from radio to X-rays .ie MIGHTEE comprises a lot of measurements one can experiment with for classifying radio sources. However, almost all algorithms assume that every input feature in the training set is vital and has a valuable physical meaning to solve the problem. Assuming that all features are vital is a problem because not all features are significant. Some input features need to be carefully filtered or removed from a training set since they introduce the curse of dimensionality. We hope the results obtained in this section justify why some attributes were not considered while constructing the pipeline. This section is divided into two sections; Section B.1 uses only direct physical measurement (flux densities) as an input feature to train the models, while section B.2 uses the derived colours as an input feature for the ML-models.

## B.1 Direct Measurement

Several direct measurements, the radio emission detected by MeerKAT and the well-matched multi-wavelength information for MIGHTEE sources available in the MIGHTEE catalogue are used as the input features to train the RF model. We combine the fluxes that derive attributes from conventional classification plus the infrared and optical flux densities. As a result, a total of sixteen attributes; COS_best_z_v5, L14, LIR_WHz, $M_\star$, class_star, four HSC optical flux densities (flux-G, flux-R, flux-I, flux-Z), four mid-infrared flux densities (SPLASH_1_FLUX, SPLASH_2_FLUX, SPLASH_3_FLUX, SPLASH_4_FLUX) was then available to use as a training set. These inputs have the redshift (COS_best_z_v5) information we do not include in the derived experiments since the $q_{\rm IR}$ is derived from redshift. The results of applying the

RF model to classify the SFGs and AGN using these measurements as input features are shown in figure B.1.

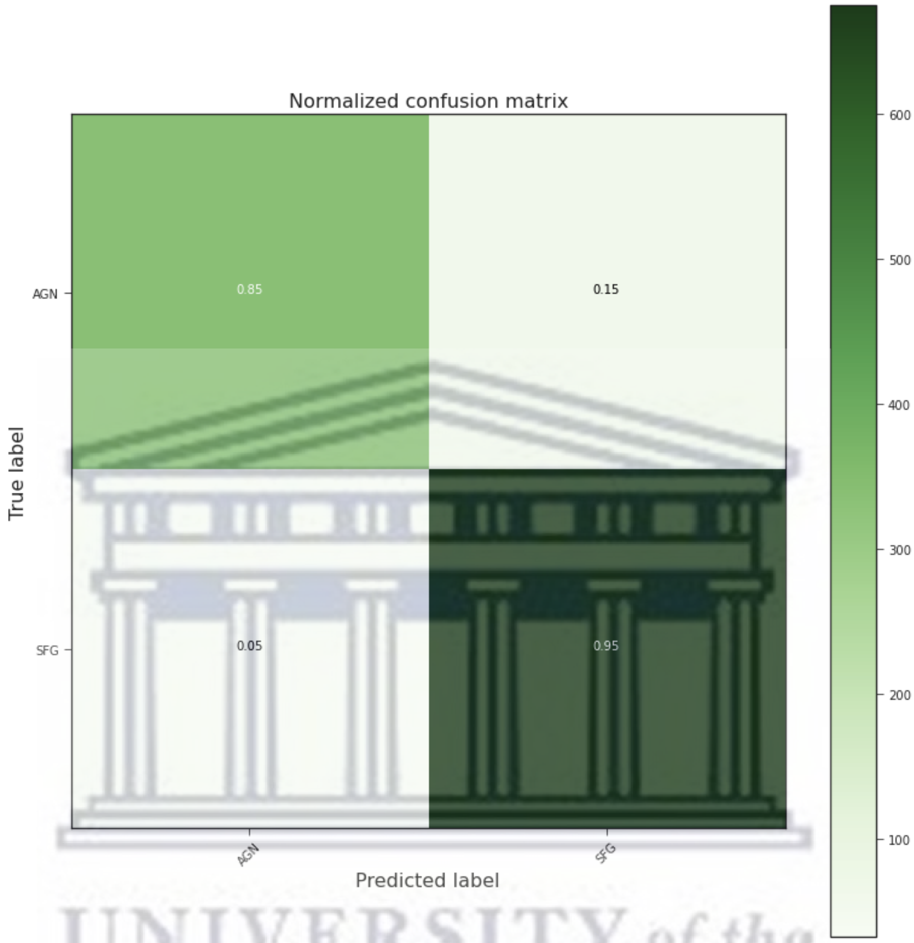


Figure B.1: Normalized Confusion Matrix.

## B.2 Derived Measurement

In this section, we derive several *colours* using the flux densities in the Radio, Far-infrared, Mid-infrared and near-infrared and optical wavelengths. The details of the catalogues from which the measurements are adopted have been described fully in section 3.4. Combining all these measurements, we had twelve flux densities; four HSC optical flux densities (flux-G, flux-R, flux-I, flux-Z), four mid-infrared flux densities (SPLASH_1_FLUX, SPLASH_2_FLUX, SPLASH_3_FLUX, SPLASH_4_FLUX)

and far infrared flux densities (flux_Y, flux_H, flux_Ks, flux_J) in different electromagnetic bands. In addition, we use the other measurements available in the MIGHTEE-COSMOS catalogue, such as $q_{\rm IR}$ and class_star already derived within the catalogue. Using the available measurements we end up with 17 colours; class_star, $q_{\rm IR}$, log(S8/S45), log(S58/S36), log(S45/S36), log(g/r), log(r/i), log(i/z), log(g/i), log(g/z), log(r/z), log(y/j), log(j/h), log(h/k), log(y/h), log(y/k), log(j/k) that can be used as input features. For example, feature log(S45/S36) is an IRAC colour derived from IRAC flux densities SPLASH_1_FLUX and SPLASH_2_FLUX, feature log(g/z) is an optical colour derived from the HSC flux-G and flux-Z, the feature log(h/k) is the colour derived from flux_H and flux_Ks, and so on.
After combining all these colours into train data, we compute Pearson correlation coefficients using https://seaborn.pydata.org[1] to inspect any correlation between the derived attributes. The colours derived from flux densities in the same window of the electromagnetic spectrum show a very high correlation. A high correlation between two or more attributes implies that using either one or all to train the ML models has no difference in the performance of models.

[1] https://seaborn.pydata.org

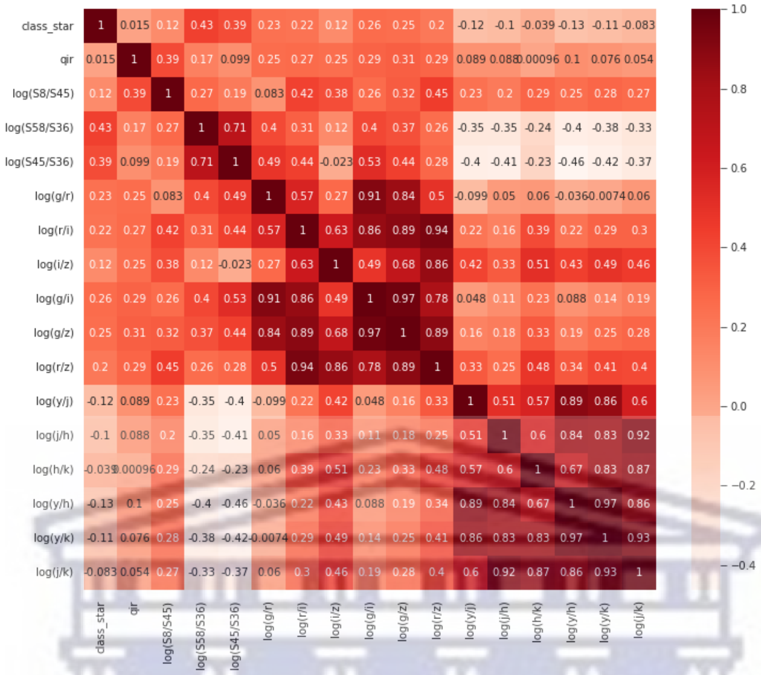


Figure B.2: The heatmap of Pearson correlation coefficients between the derived input features. A value of 1 (also the reddest) between two attributes implies a strong positive correlation. A value of -1 would mean a strong negative correlation between two attributes. A value of 0 would mean no correlation between the two attributes.

We also make a scatter plot using two attributes for each class. We had seventeen attributes and wanted to make a combination of 2. Using equation 4.1, we end up with 136 pair plots to visualize. Figure B.3 shows a scatter plot of all 136 pairs of features. The AGN and SFGs are shown in blue and red, respectively. This pair plot provides a statistical analysis of how variables in a dataset relate to each other and how those relationships depend on other variables. If we first consider the diagonal plots of figure B.3 we can tell that the distribution of most of the sources in each colour is almost the same. The two population of radio sources is distributed within a similar space or has an identical mean, which implies that most of the colours are not suitable separators of the AGN or SFGs. The same trend is observed in the scatter plot of the colours. The star-forming galaxies and active galactic nuclei are mostly within the same feature space in these plots except when '$q_{IR}$', 'IRAC colours' and 'class_star' has been used.

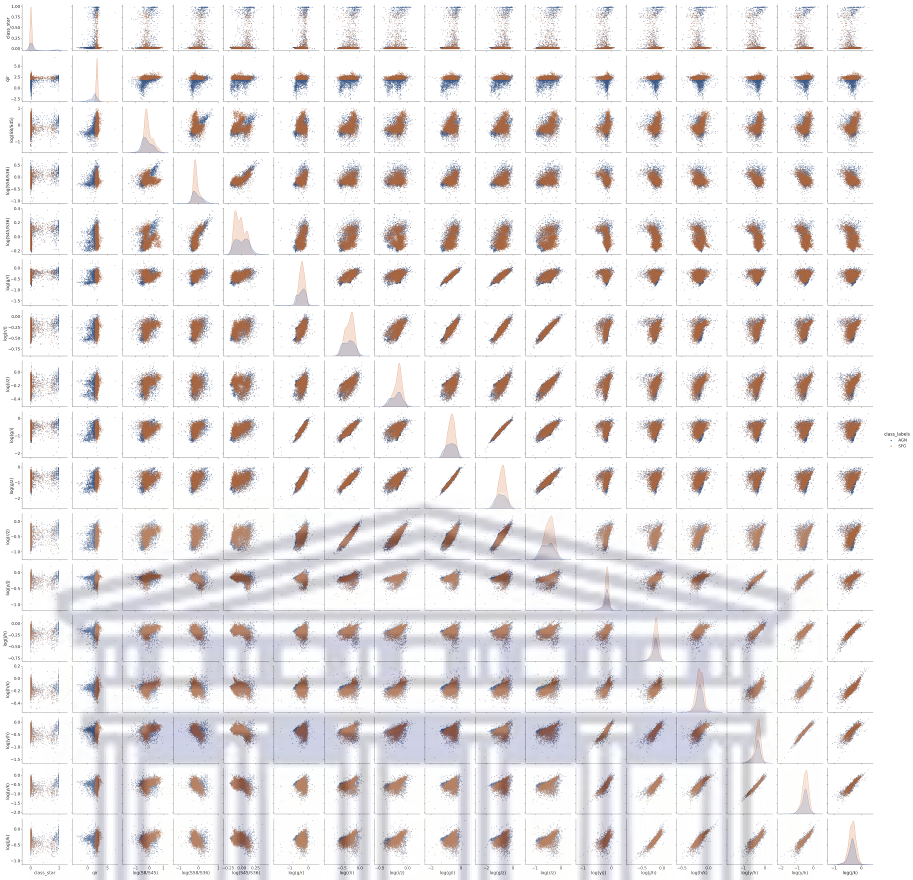


Figure B.3: Pair plot of different colours derived from the MIGHTEE-COSMOS catalogue. The AGN (blue) and SFG (red) represent active galactic nuclei and star-forming galaxies labelled in the MIGHTEE-COSMOS catalogue in Whittam et al. (2022), respectively. The diagonal plots show a distribution of the two classes of radio sources on each feature (one-dimensional). The off-diagonal plots show a scatter of the two sources in a two-dimensional feature space.

We then adopt the permutation feature importance method to assess the importance of the derived features. Figure B.4 shows the results of applying the permutation importance algorithm to the derived input features. The additional features derived from the far-infrared and optical flux densities are raked below the 2%. The importance of less than 2 % indicates that these features do not add knowledge about the distinction between AGN and SFGs. These results agree with what we learn from figure B.3.

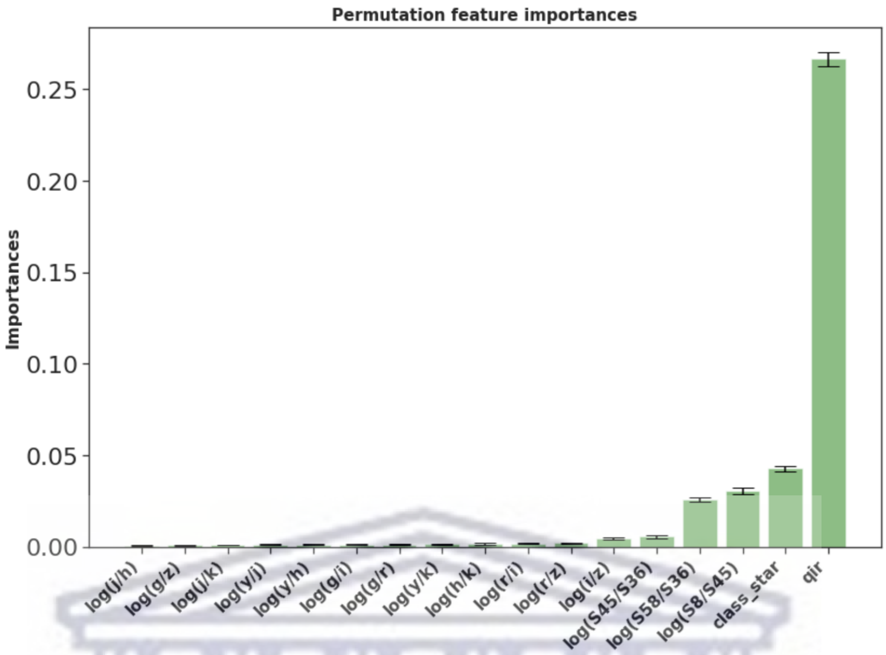


Figure B.4: Permutation-based feature importance for the explanatory features included in the Random Forest model for the derived input features calculated using F1-scores. The plot indicates the relative importance of the derived colours as input features rated by the random forest model.

.

Finally, we show the results of applying the $k$NN model to the derived features. The $k$NN is selected in this case because it gave better results in the classification of radio sources under study, refer to Figure 4.9 and 4.10. We made three combinations of features, namely, F4 (class_star, $q_{\mathrm{IR}}$, log(S8/S45), log(S58/S36)), F4+optical colours (class_star, $q_{\mathrm{IR}}$, log(S8/S45), log(S58/S36), log(g/r), log(r/i), log(i/z), log(g/i), log(g/z), log(r/z)) and F4+optical+NIR (class_star, $q_{\mathrm{IR}}$, log(S8/S45), log(S58/S36), log(g/r), log(r/i), log(i/z), log(g/i), log(g/z), log(r/z), log(y/j), log(j/h), log(h/k), log(y/h), log(y/k), log(j/k)). The data was split into 80% training data and 20% test data and was evaluated using the F1-score as the classification metrics. The performance of $k$NN model is reported in Figure B.5. It is clear in Figure B.5 that the additional optical and NIR colours do not improve the performance of the ML model, which complement the permutation-based feature importance shown in Figure B.4

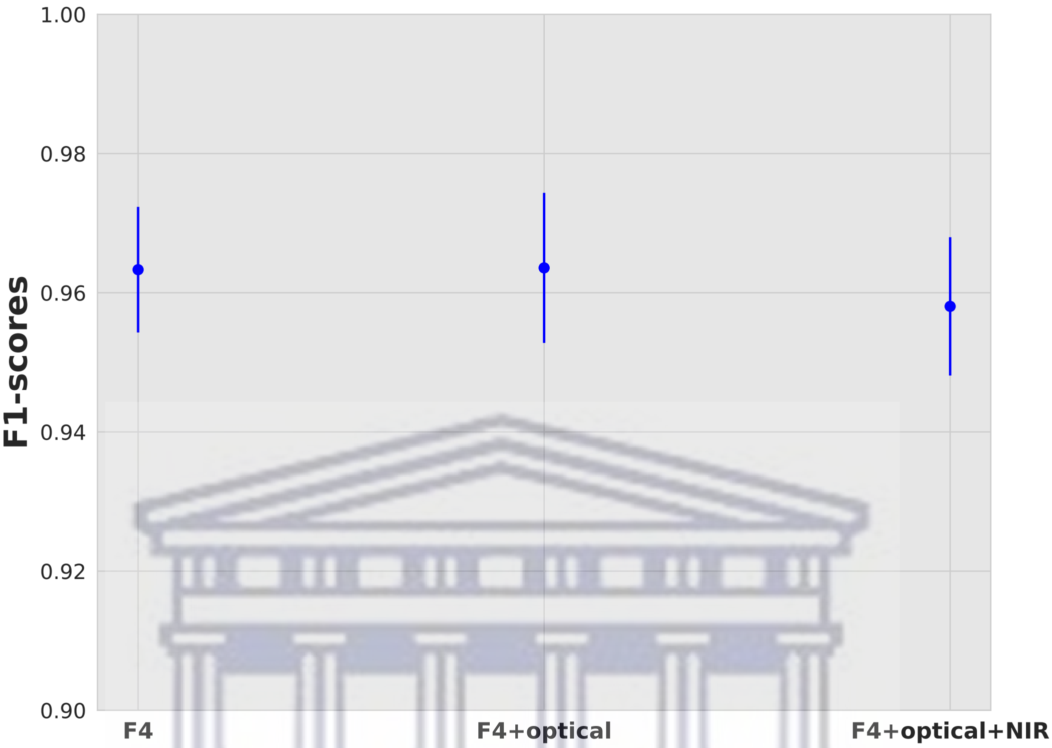


Figure B.5: The results of applying the $k$NN classifier to distinguish between AGN and SFGs trained using three different feature combinations, namely, F4, F4+optical and F4+optical+NIR. The vertical axis shows the F1-scores and a score of 1 indicates a perfect classification.

# Appendix C

# Hyperparameters

Hyperparameter optimization is one factor that plays a role in improving the performance of machine learning models. I have discussed the implications of hyperparameter tuning in section 2.3. In this section, I give a set of parameters used to tune each model used in this study. Due to the computational complexity that comes when more hyperparameters are included for optimization, We only select a few values per hyperparameter to optimise. We do this by using a plot of the f1-score and the different values of the hyperparameter. The parameters that give better results then be selected for further optimization when training the models. Figure C.1 shows the f1-score results for different n_*estimator* hyperparameters. The n_*estimator* hyperparameter indicates the number of trees to include in the forest and is one of the important hyperparameters of this model. The same results can be used for different parameters and different models. Below, we show the parameters we tune for each model during training.

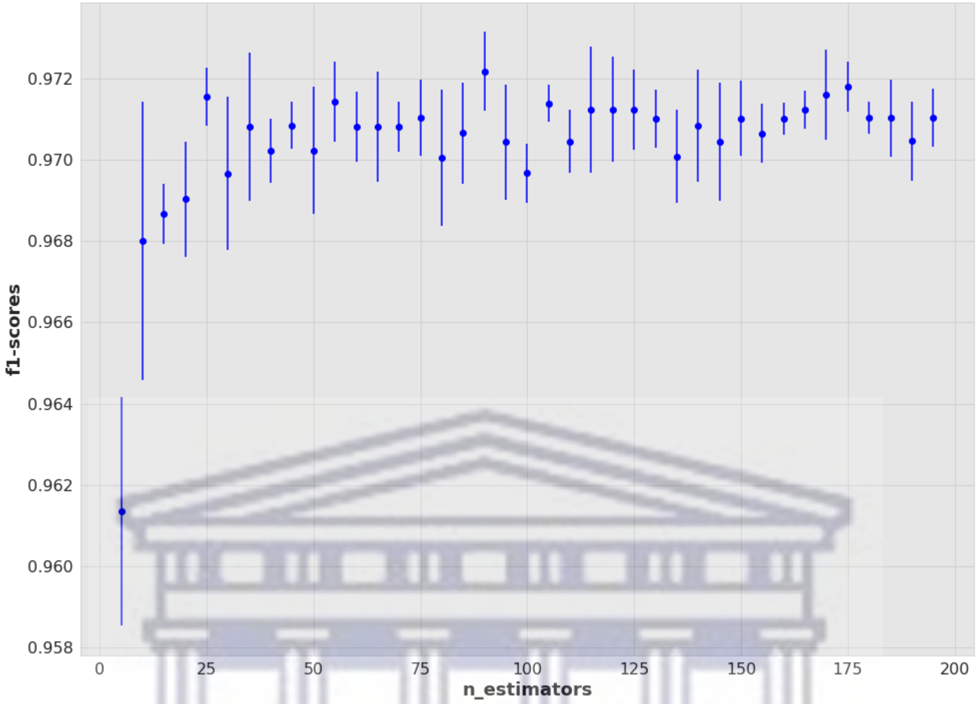


Figure C.1: The results of applying the random forest model to MIGHTEE-COSMOS varying the number of trees in the forest

.

- **Logistics Regression**
  [https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
  'solvers' = ['newton-cg', 'lbfgs', 'liblinear']

  'penalty' = ['l2']

  'c_values' = [1000, 100, 10, 1.0, 0.1, 0.01, 0.001]

- **Support Vector Machines**
  [https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)
  'C': [0.1, 1, 10, 100, 1000]

  'gamma': [1, 0.1, 0.01, 0.001, 0.0001]

  'kernel': ['linear', 'rbf', 'poly', 'sigmoid' ]

- **K-Nearest Neighbour**
  https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
  'n_neighbors' : [5, 10, 15]

  'p':[1, 2]

  'weights' : ['uniform', 'distance']

- **Random Forest**
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
  'n_estimators' = [50, 100, 150, 200, 300, 450]

  'max_features' = ["sqrt", "log2", "NONE"]

  'max_depth' = [5, 10, 30, 40, 50, 60, 70]

  'min_samples_split' = [2, 3, 5]

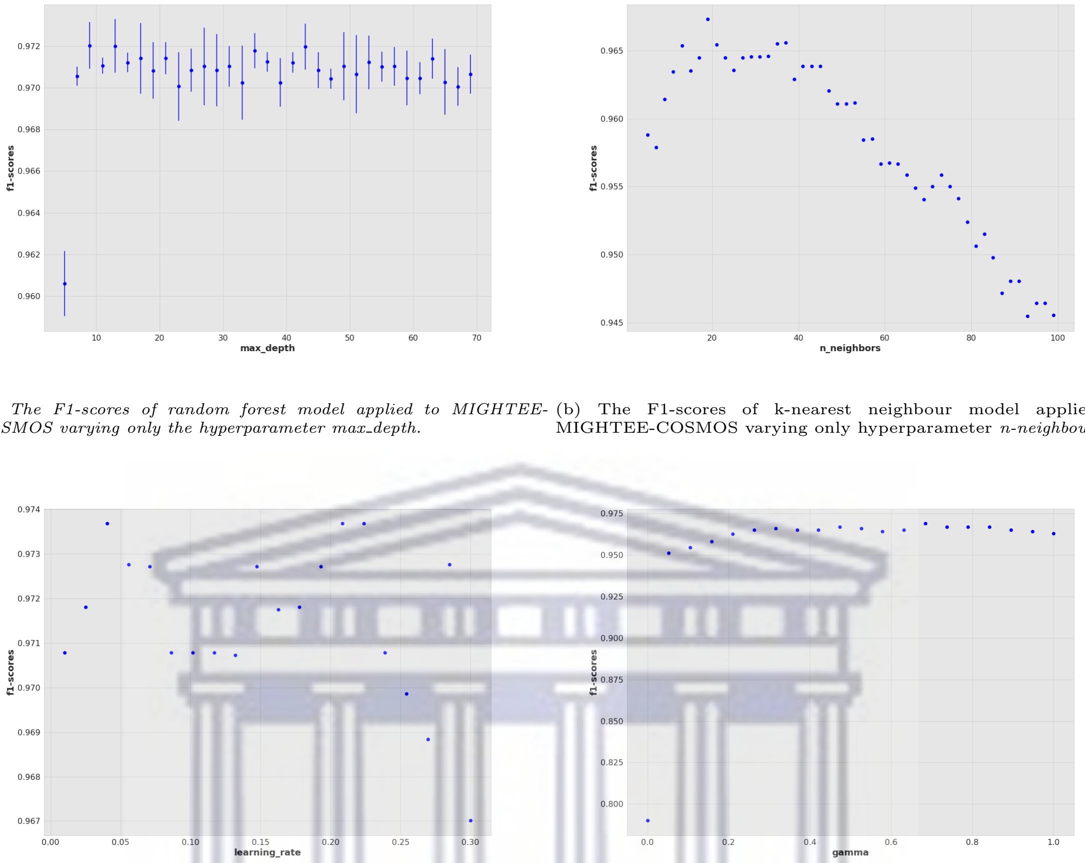
  'min_samples_leaf' = [1, 3]

  'bootstrap' = [True, False]

- **XGBoost**
  https://xgboost.readthedocs.io/en/stable/parameter.html
  'learning_rate': [0.01, 0.15, 0.23, 0.3]

  'max_depth': [3, 7, 10]

  'n_estimators' : [50, 100, 150, 200, 300]



(a) *The F1-scores of random forest model applied to MIGHTEE-COSMOS varying only the hyperparameter max_depth.*

(b) The F1-scores of k-nearest neighbour model applied to MIGHTEE-COSMOS varying only hyperparameter *n-neighbours.*

(c) *The F1-scores of XGBoost model applied to MIGHTEE-COSMOS varying only hyperparameter learning rate.*

(d) *The F1-scores of Support Vector Machines model applied to MIGHTEE-COSMOS varying only hyperparameter gamma.*

Figure C.2: The result of varying a single hyperparameter of different supervised machine learning models.