

# Machine Learning Approaches for Classifying Star-Forming Galaxies and Active Galactic Nuclei from MIGHTEE-Detected Radio Sources in the COSMOS Field

Walter Silima,<sup>1,2</sup> Fangxia An,<sup>3,4,1\*</sup> Mattia Vaccari,<sup>2,1,5</sup> Eslam A. Hussein,<sup>1</sup> S. Randriamampandry<sup>6</sup>

<sup>1</sup>*Inter-University Institute for Data Intensive Astronomy (IDIA), Department of Physics and Astronomy, University of the Western Cape, 7535 Bellville, Cape Town, South Africa*

<sup>2</sup>*Inter-University Institute for Data Intensive Astronomy (IDIA), Department of Astronomy, University of Cape Town, 7701 Rondebosch, Cape Town, South Africa*

<sup>3</sup>*Yunnan Observatories, Chinese Academy of Sciences, Kunming 650216, People's Republic of China*

<sup>4</sup>*Purple Mountain Observatory, Chinese Academy of Sciences, 10 Yuanhua Road, Qixia District, Nanjing 210023, People's Republic of China*

<sup>5</sup>*INAF - Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy*

<sup>6</sup>*A&A, Department of Physics, Faculty of Sciences, University of Antananarivo, P.O. Box 906, Antananarivo 101, Madagascar*

Accepted 2025 September 26. Received 2025 September 25; in original form 2025 April 10

## ABSTRACT

Radio synchrotron emission originates from both massive star formation and black hole accretion, two processes that drive galaxy evolution. Efficient classification of sources dominated by either process is therefore essential for fully exploiting deep, wide-field extragalactic radio continuum surveys. In this study, we implement, optimize, and compare five widely used supervised machine-learning (ML) algorithms to classify radio sources detected in the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE)–COSMOS survey as star-forming galaxies (SFGs) and active galactic nuclei (AGN). Training and test sets are constructed from conventionally classified MIGHTEE–COSMOS sources, and 18 physical parameters of the MIGHTEE-detected sources are evaluated as input features. As anticipated, our feature analyses rank the five parameters used in conventional classification as the most effective: the infrared-radio correlation parameter ( $q_{\text{IR}}$ ), the optical compactness morphology parameter ( $\text{class\_star}$ ), stellar mass, and two combined mid-infrared colors. By optimizing the ML models with these selected features and testing classifiers across various feature combinations, we find that model performance generally improves as additional features are incorporated. Overall, all five algorithms yield an  $F1$ -score (the harmonic mean of precision and recall)  $> 90\%$  even when trained on only 20% of the dataset. Among them, the distance-based  $k$ -nearest neighbors classifier demonstrates the highest accuracy and stability, establishing it as a robust and effective method for classifying SFGs and AGN in upcoming large radio continuum surveys.

**Key words:** methods: observational – software: machine learning – galaxies: evolution – galaxies: formation – radio continuum: galaxies

## 1 INTRODUCTION

The radio continuum emission from galaxies is powered by star formation (SF) and black hole accretion, the two dominant physical processes that drive galaxy evolution. SF-related radio emission originates from supernova-accelerated cosmic ray (CR) electrons gyrating within galactic magnetic fields, producing non-thermal synchrotron radiation, and from Coulomb scattering between free ions and electrons in HII regions, resulting in thermal free-free emission. Both synchrotron and free-free emissions remain unaffected by dust obscuration, which is critical for obtaining an unobstructed view of SF in galaxies (see [Condon 1992](#), for a review). The synchrotron emission from relativistic jets and outflows powered by black hole (BH) accretion dominates the radio emission of luminous radio sources, namely radio galaxies ([Sadler et al. 1989](#); [Miley & De](#)

[Breuck 2008](#)). The feedback processes associated with BH accretion play a crucial role in regulating galaxy growth, with jets and outflows potentially expelling star-forming gas from galactic bulges and quenching star formation in galaxies. Consequently, distinguishing between SF-dominated and active galactic nuclei (AGN)-dominated radio emission is crucial for utilizing the radio continuum in exploring cosmic evolution.

Newly constructed and upgraded radio interferometric arrays in the past two decades, such as the Australian Square Kilometre Array Pathfinder (ASKAP, [Hotan et al. 2021](#)), Murchison Widefield Array (MWA, [Lonsdale et al. 2009](#)), MeerKAT ([Jonas & MeerKAT Team 2016](#)), the Low Frequency Array (LOFAR, [van Haarlem et al. 2013](#)), and upgraded Giant Metrewave Radio Telescope (uGMRT, [Swarup et al. 1991](#)), have led to a new generation of large extragalactic radio continuum surveys (e.g., [Ocran et al. 2020](#); [Ishwara-Chandra et al. 2020](#); [Heywood et al. 2022](#); [Best et al. 2023](#); [Hale et al. 2025](#)). Some of these deep surveys achieve an angular resolution of  $\lesssim 5''$

\* E-mail: anfangxia@ynao.ac.cn; fangxiaan@gmail.com

(e.g., Smolčić et al. 2017; Jiménez-Andrade et al. 2024), while the majority of current extragalactic radio continuum surveys operate at resolutions of  $\sim 6\text{--}10''$ , with sensitivities reaching the  $\mu\text{Jy}$ -level. Combined with survey areas covering dozens of square degrees, these surveys have led to an exponential increase in the number of radio sources detected over the past two decades (Norris 2017; Best et al. 2023; Hale et al. 2025). This rapid expansion in data volume demands the development of efficient and automated techniques to classify the sources detected from these surveys as SF-dominated or AGN-dominated before further investigating their physical nature.

Machine learning (ML) is firmly established in astronomy and has been widely used in various research areas, such as galaxy (morphology) classification (e.g., Ball & Brunner 2010; An et al. 2018), discovery/prediction of astrophysical activities (e.g., Florios et al. 2018; Mahabal et al. 2019), estimation of photometric redshifts (e.g., Li et al. 2023), noise analysis in gravitational wave detection (e.g., Biswas et al. 2013; George et al. 2018), and for many other applications (see Fluke & Jacobs 2020, for a review). Automated classification has recently been adopted in SFG-AGN separation but with only one particular ML algorithm (Karsten et al. 2023).

In this work, we implement and optimize five widely used supervised ML algorithms, namely Logistic Regression (LR, Menard 2010), Support Vector Machine (SVM, Cristianini et al. 2000), K-Nearest Neighbour ( $k\text{NN}$ , Peterson 2009), Random Forest (RF, Breiman 2001), and Extreme Gradient Boosting, commonly known as XGBoost (XGB, Chen & Guestrin 2016), to classify SF-dominated or AGN-dominated radio sources from the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) survey (Jarvis et al. 2016; Heywood et al. 2022; Hale et al. 2025). Since these ML models are based on distinct algorithmic approaches, we also aim to assess their relative effectiveness in classifying star-forming galaxies (SFGs) and AGN from radio continuum surveys. Sources detected from the MIGHTEE-COSMOS early science data have been classified as SF-dominated or accretion-dominated following the traditional SFG/AGN classification diagnostic (Whittam et al. 2022). We use these classifications to construct the training set and optimize the different ML algorithms.

The successful adoption of ML will efficiently provide accurate SFGs/AGN samples, which is essential for scientific studies based on recently completed or ongoing high-sensitivity and wide-field extragalactic radio continuum surveys, and eventually, the surveys conducted by the Square Kilometre Array (SKA, Dewdney et al. 2009), next-generation *Karl G. Jansky* Very Large Array (ngVLA, Murphy et al. 2018), and the Five-hundred-meter Aperture Spherical Radio Telescope (FAST) Core Array (Jiang et al. 2024).

We describe the MIGHTEE-COSMOS data as well as the ancillary data used in this work in Section §2. The data analyses and feature selection of ML are described in Section §3. We show the results of our ML application in Section §4. Our results are discussed and summarized in Sections §5 and §6 respectively. Throughout this paper, we adopt the AB magnitude system (Oke 1974) and assume a flat  $\Lambda\text{CDM}$  cosmological model with the Hubble constant  $H_0 = 67.27 \text{ km s}^{-1} \text{ Mpc}^{-1}$ , matter density parameter  $\Omega_m = 0.32$ , and cosmological constant  $\Omega_\Lambda = 0.68$  (Planck Collaboration et al. 2016).

## 2 MIGHTEE-COSMOS DATA

The MIGHTEE survey is one of the MeerKAT large survey projects, conducted by an international collaboration of researchers. MIGHTEE targets four extensively studied extragalactic fields: the Cosmological Evolution Survey (COSMOS) field, the Extended Chandra

Deep Field-South (E-CDFS), the European Large Area Infrared Survey South 1 (ELAIS-S1) field, and the XMM-Newton Large Scale Structure (XMM-LSS) field, covering a total of 20 square degrees with  $\mu\text{Jy}$ -level sensitivity (Jarvis et al. 2016). The survey includes deep GHz radio continuum (Heywood et al. 2022; Hale et al. 2025), spectral line (Maddox et al. 2021), and polarization (Taylor et al. 2024) observations, aimed at exploring cosmic evolution.

This work utilizes MIGHTEE early science radio continuum data in the COSMOS field, as released and fully described by Heywood et al. (2022). The COSMOS field was observed for a total of 17.45 hours on source between 2018 and 2019 using MeerKAT’s L-band receivers (856–1712 MHz), with a single pointing centered at  $\text{RA}=10^{\text{h}}00^{\text{m}}28.6^{\text{s}}$ ,  $\text{Dec}=+02^{\text{d}}12^{\text{m}}21^{\text{s}}$ . The MIGHTEE-COSMOS early science radio data were processed with Briggs’ robust weighting values of 0.0 and -1.2 (Briggs 1995). The former yielded more sensitive imaging data with a thermal noise of  $1.7 \mu\text{Jy beam}^{-1}$  and a circular synthesized beam size of  $8.6'' \times 8.6''$ . It is important to note that the high-sensitivity data are limited by classical confusion at the center, increasing the mean noise to  $4\text{--}5 \mu\text{Jy beam}^{-1}$  (Heywood et al. 2022). While the full coverage of the MIGHTEE-COSMOS early science data spans  $1.6 \text{ deg}^2$ , we restrict the analyses within the central  $0.86 \text{ deg}^2$ , where the radio data are deepest and multi-wavelength cross-matching has been completed for the MIGHTEE sources (Whittam et al. 2024).

### 2.1 MIGHTEE-COSMOS Multi-wavelength catalogue

As described in (Whittam et al. 2024), there are 6102 radio components with peak brightnesses that exceed the local background noise by  $5\sigma_{\text{local}}$  within the central  $0.86 \text{ deg}^2$  of the MIGHTEE-COSMOS field. Whittam et al. (2022, 2024) identified the host galaxy for 5223 out of 6102 radio-detected sources by visual cross-matching the MIGHTEE sources with  $K_s$ -band-detected sources from the fourth data release (DR4) of the UltraVISTA survey (Bowler et al. 2020; Adams et al. 2021). Details of the visual cross-matching are presented in Whittam et al. (2024).

Using the position of the host galaxies, Whittam et al. (2022, 2024) also identified multi-wavelength counterparts for the 5223 radio sources detected in the MIGHTEE survey. Here, we briefly summarize the identified multi-wavelength counterparts of MIGHTEE sources as reported by Whittam et al. (2022, 2024). Of the 5223 MIGHTEE sources with UltraVISTA  $K_s$ -band counterparts, 572 (11%) were detected in X-ray observations. The optical, near-infrared (NIR) counterparts of MIGHTEE sources were identified using the optical and near-infrared broad-band photometric catalogue created by Adams et al. (2021), which include  $YJHK_s$ -band data from the UltraVISTA DR4, as well as *grizy*-bands data from Hyper Suprime-Cam Subaru Strategic Program (HSC SSP; Tanaka et al. 2017), deep  $u^*$ -band data from the Canada–France–Hawaii Telescope Legacy Survey (CFHTLS; Cuillandre et al. 2012), and mid-infrared (MIR) data from the Spitzer Infrared Array Camera (IRAC) at 3.6 and  $4.5 \mu\text{m}$ . Whittam et al. (2022) also used the high-resolution *Hubble Space Telescope* (HST) Advanced Camera for Surveys (ACS) *I*-band imaging data (Scoville et al. 2007) and found 4697 out of 5223 MIGHTEE sources have HST *I*-band counterparts. Furthermore, to identify the MIR counterparts of MIGHTEE sources, Whittam et al. (2022) used data at  $5.8$  and  $8.0 \mu\text{m}$  from the Spitzer Large-Area Survey with Hyper-Suprime-Cam (SPLASH; Steinhardt et al. 2014), accessed through the COSMOS2015 catalog (Laigle et al. 2016). As a result, 4815 of the 5223 MIGHTEE sources have detections at  $5.8$  and  $8.0 \mu\text{m}$ . Far-infrared (FIR) counterparts were identified using data from the Herschel Extragalactic Legacy Project (HELP;

Vaccari 2015; Shirley et al. 2021). Among the 5223 MIGHTEE sources, 4540 were detected at 24, 100, and 160  $\mu\text{m}$  using the Multi-band Imaging Photometer (MIPS; Rieke et al. 2004) on the Spitzer Space Telescope and the Photodetector Array Camera and Spectrometer (PACS; Poglitsch et al. 2010). Additionally, 4957 out of the 5223 sources were detected at 250, 350, and 500  $\mu\text{m}$  using the Spectral and Photometric Imaging Receiver (SPIRE; Griffin et al. 2010) on Herschel. Furthermore, Whittam et al. (2022) cross-matched the optical positions of MIGHTEE sources with very long baseline interferometry (VLBI) observed sources, finding that 255 of the 5223 sources have VLBI detections (Herrera Ruiz et al. 2017).

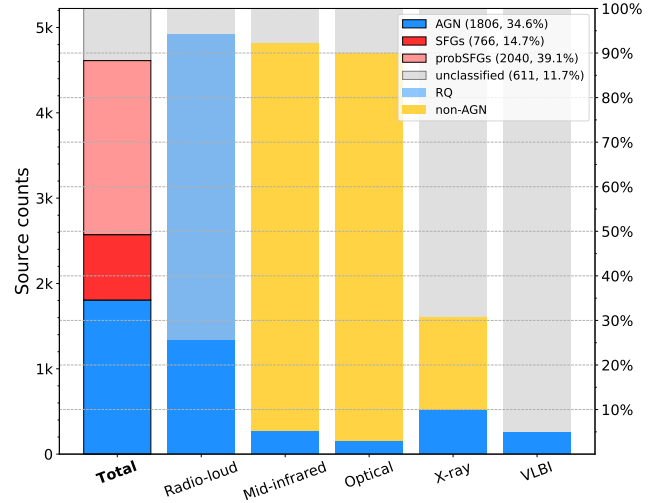
This work also uses the estimated redshift and stellar mass of MIGHTEE sources from the MIGHTEE-COSMOS multi-wavelength catalogue. As reported by Whittam et al. (2024), 2427 of the 5223 MIGHTEE sources have spectroscopic redshifts compiled from the literature. For the remaining 2796 sources, their photometric redshifts were determined by Hatfield et al. (2022) using a hierarchical Bayesian approach that integrates two distinct methodologies, as detailed by Duncan et al. (2018). The stellar masses of MIGHTEE sources were estimated using AGNFITTER SED-fitting. Details of the estimation and comparisons of stellar mass estimates across different SED-fitting codes are presented in Whittam et al. (2022).

## 2.2 MIGHTEE-COSMOS Conventional Classification

Utilizing the well-matched MIGHTEE-COSMOS multi-wavelength catalogue, Whittam et al. (2022) classified AGN and SFGs from the MIGHTEE-COSMOS survey by using five conventional classification techniques: radio excess, MIR colour-colour, optical morphology, X-ray luminosity, and the VLBI criteria (Table 1, Whittam et al. 2022).

Radio-excess AGN were identified as sources with significantly more radio emission than expected from star formation alone, determined by the infrared-radio correlation (IRRC) quantified as  $q_{\text{IR}}$  in Whittam et al. (2022). The MIR colour-colour diagram defined by Donley et al. (2012) was used to identify sources exhibiting power-law emissions from the torus, classifying them as MIR AGN in Whittam et al. (2022). Optical point-like AGN were identified using *HST* ACS *I*-band imaging, based on the principle that the emission from the nucleus outshines that of the host galaxy. Sources with a Source-Extractor (SExtractor) compactness parameter,  $\text{class\_star} \geq 0.9$  were classified as optical point-like AGN in Whittam et al. (2022). As some of the brightest AGN exhibit characteristic accretion-related X-ray emissions, X-ray AGN were identified by applying a rest-frame (0.5–10 keV) X-ray luminosity threshold of  $L_x \geq 10^{42} \text{ erg s}^{-1}$  (Szokoly et al. 2004). Finally, VLBI AGN were classified as sources with a brightness temperature exceeding that of typical SFGs. We refer the reader to Whittam et al. (2022) for details about the conventional classification of MIGHTEE-COSMOS radio sources.

Whittam et al. (2022) classified a source as an AGN if it satisfied any one (or more) of AGN criteria. Sources that did not meet any of the AGN criteria across all five diagnostic methods were classified as SFGs. However, due to the limited depth of the X-ray observations, only sources with  $z < 0.5$  could be confidently classified as ‘not X-ray AGN’ if they were undetected in the X-ray. For X-ray undetected sources with  $z > 0.5$ , their potential X-ray luminosity might be above the classification threshold of  $L_x \geq 10^{42} \text{ erg s}^{-1}$ , and therefore, are unable to fulfill the ‘not X-ray AGN’ criteria. If these sources were classified as ‘not AGN’ based on the other four diagnostics, Whittam et al. (2022) introduced an additional category, namely ‘probable SFG’.



**Figure 1.** The bar plot illustrates the completeness of the overall classification (total) and the completeness for each diagnostic method of MIGHTEE-COSMOS detected radio sources. The categories of sources are colour-coded, with AGN shown in blue, SFGs in red, probable SFGs in light red, radio quiet (RQ) in light blue, non-AGN in yellow, and unclassified sources in grey.

**Table 1.** Number of MIGHTEE-COSMOS sources per class

| Overall Class | Number of Sources |
|---------------|-------------------|
| AGN           | 1806              |
| SFG           | 2806              |
| Unclassified  | 611               |

Figure 1 shows the overall classification completeness and the completeness for each diagnostic method of MIGHTEE-COSMOS detected radio sources. Due to the low completeness and unpredictability of the X-ray and VLBI classifications, these two features are excluded from our ML analysis. Consequently, we merge the ‘probable SFG’ category into the ‘SFG’ class in this work. Table 1 summarizes the number of sources in each class used in this work.

## 3 ANALYSES

A successful supervised classification relies on the quality and representativeness of the dataset used in model development, the careful selection and tuning of adjustable model parameters, and the robustness of the evaluation criteria (Section §3.1) employed to assess the model’s performance. To ensure optimal results, we adhere to the standard workflow for supervised ML, which consists of the following steps:

- (i) feature analysis and selection (Sections §3.2, §3.3, and §3.4),
- (ii) building a training set using the selected features (Section §3.5), training of the ML model to create a classifier, and hyper-parameter optimization (Section §3.5.1), and
- (iii) applying the classifier to predict the class labels of the test sample. However, in this study, we evaluate the performance of ML classification using only the validation dataset (Section §4).

**Table 2.** Confusion matrix for binary classification. Actual represents the actual labeled class (in our case, the manually labeled class, i.e. AGN or SFG), while predicted represents the class predicted by the ML algorithms.

|           |          | Actual              |                     |
|-----------|----------|---------------------|---------------------|
|           |          | Positive            | Negative            |
| Predicted | Positive | True Positive (TP)  | False Positive (FP) |
|           | Negative | False Negative (FN) | True Negative (TN)  |

### 3.1 Evaluation Metrics

Evaluation metrics are used to evaluate the performance of ML models. In this work, we adopt the classification metrics *Precision*, *Recall* and *F1-score*, with the latter being particularly effective for imbalanced datasets (Hossin & Sulaiman 2015; Yadav & Bhole 2020).

Table 2 outlines the confusion matrix for a binary classification scenario. If we consider the AGN class as positive and the SFG class as negative, a True Positive (TP) refers to the number of labeled AGN that are correctly classified by the ML models. Conversely, False Positives (FP) represent cases where SFGs are incorrectly identified as AGN. Similarly, False Negatives (FN) occur when AGN are misclassified as SFGs, while True Negative (TN) refers to the number of SFGs that are accurately classified by the ML models.

The classification metrics are derived from the confusion matrices. *Precision* quantifies the accuracy of positive predictions made by the ML models, defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (1)$$

whereas *Recall* evaluates the model's ability to minimize false negatives, expressed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (2)$$

Lastly, the *F1-score*, which represents the harmonic mean of *Precision* and *Recall*, offers a comprehensive measure of the model's performance. It is defined as:

$$F1 = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})}, \quad (3)$$

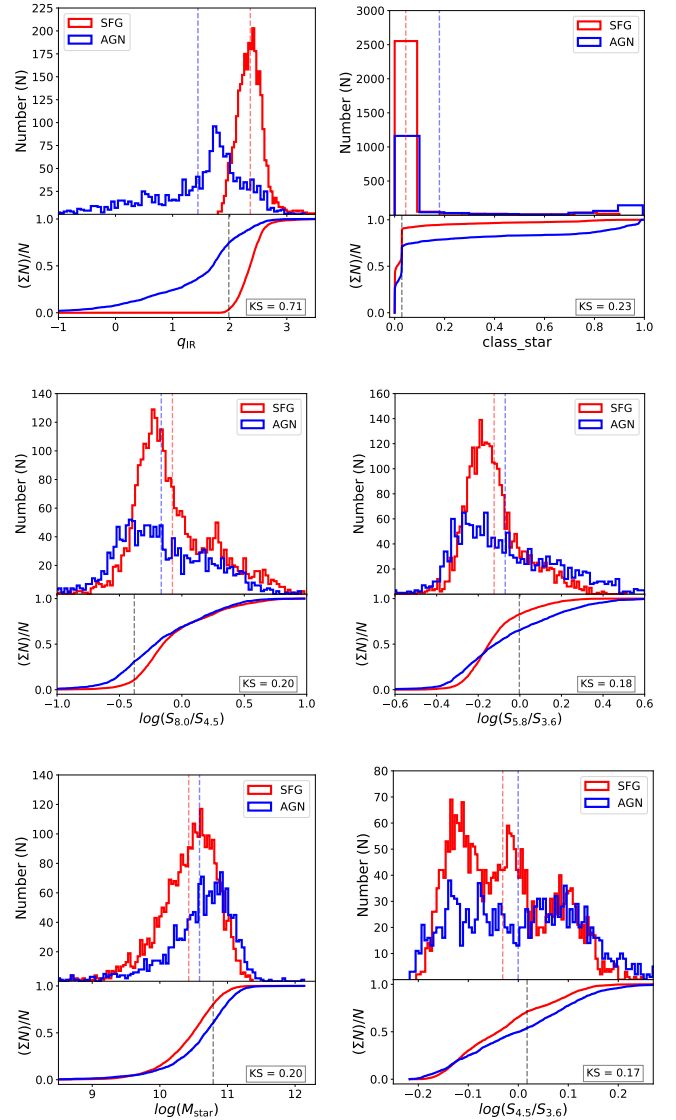
and utilised as the main evaluation metric in this work (Section §4).

### 3.2 Feature Analysis

As outlined in Table 1, Whittam et al. (2022) classified a total of 1806 AGN and 2806 SFGs from the MIGHTEE-COSMOS survey, utilizing five conventional classification diagnostics. This labeled sample of 4612 sources is used to construct training and test datasets, enabling the evaluation of ML models based on various input features. As described in Appendix A and shown in Figure A1, in this study, we incorporate all available photometric data from the MIGHTEE-COSMOS multi-wavelength catalogue (Whittam et al. 2024), alongside the conventional classification diagnostics used in Whittam et al. (2022), to evaluate their efficiency in distinguishing between SFGs and AGN in the MIGHTEE-COSMOS survey.

#### 3.2.1 One Dimensional Analysis

The performance of supervised ML models is highly dependent on the selection of features used for training. Effective feature selection reduces the dimensions of the data, enabling the model to perform more efficiently. However, identifying the most critical features is a non-trivial task and often requires advanced ML techniques. This



**Figure 2.** Histograms (top) and Kolmogorov–Smirnov (K-S) test results (bottom) for AGN (blue) and SFGs (red), in the MIGHTEE-COSMOS catalog. Among the 18 parameters considered for selecting input features for ML, these six exhibit the highest significance based on the K-S statistic. The K-S value for each feature is displayed in the bottom-right corner of each panel. The features are sorted by the significance level of the K-S statistic (from left to right, top to bottom). In the top panels, vertical dashed lines indicate the mean of each distribution. In the bottom panels, the Y-axis shows the cumulative distribution function (CDF), with vertical dashed lines marking the point of maximum separation between the two distributions.

section examines two methods for analyzing a low-dimensional feature space.

A straightforward method for identifying key features in binary classification is to examine the histograms of each feature for the two classes (Zhang et al. 2003), as illustrated in Figure 2. Greater separation between the distributions of the two classes indicates that the feature is more effective in distinguishing between them. A quantitative measure of this separation is the difference in the *means* of the two distributions (Figure 2). Additionally, in Figure 2, we also present the results of Kolmogorov–Smirnov (K-S) tests, which as-

sess the statistical differences between the two populations for each feature (Berger & Zhou 2014).

To determine the input features for training ML models, we first incorporate all twelve-band optical to MIR photometries from the MIGHTEE-COSMOS multi-wavelength catalogue (Section 2.1), including four HSC *griz*-band flux densities, four UltraVISTA *YJHK<sub>s</sub>*-band photometries, and four IRAC 3.6, 4.5, 5.8, and 8.0  $\mu\text{m}$ -band flux densities. From these photometries, we derive 15 color indices: three MIR colors ( $\log(S_{8.0}/S_{4.5})$ ,  $\log(S_{5.8}/S_{3.6})$ ,  $\log(S_{4.5}/S_{3.6})$ ) and 12 NIR and optical colors. A full description of these color indices is provided in Appendix A.

Additionally, we include other measurements available in the MIGHTEE-COSMOS multi-wavelength catalogue, particularly the conventional classification diagnostics used by Whittam et al. (2022), such as the IRRC parameter  $q_{\text{IR}}$ , stellar mass  $\log(M_{\text{star}})$ , and the optical compactness parameter  $\text{class\_star}$ . As discussed in Section 2.2, we exclude X-ray luminosity and VLBI detection from the input features due to their low completeness and the unpredictability.

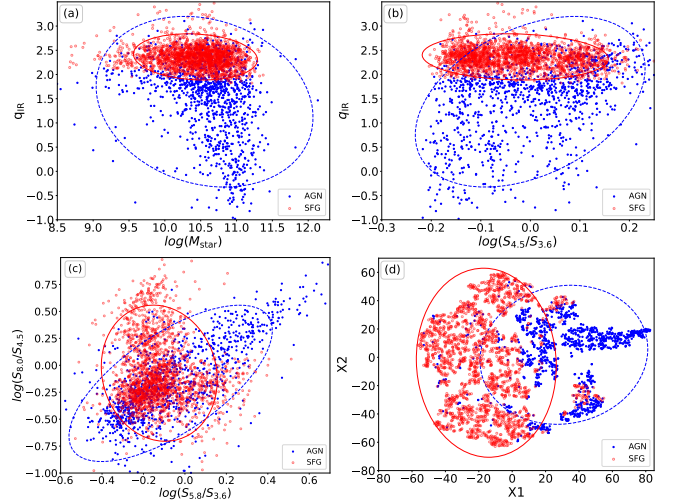
In total, we include 18 physical parameters in our analysis. Figure 2 highlights the six features with the greatest significance level based on the K-S statistic. The K-S value for each feature is displayed in the bottom-right corner of each panel in Figure 2, indicating that these six features exhibit the greatest differences in cumulative distribution functions (CDFs) between SFGs and AGN. As illustrated in Figure 2, both the histogram and K-S test results demonstrate that  $q_{\text{IR}}$  is the most discriminative feature for differentiating SFGs from AGN among the MIGHTEE-detected radio sources, followed by the optical compactness parameter  $\text{class\_star}$ . Stellar mass ( $\log(M_{\text{star}})$ ), along with three IRAC colours show slight variations in ranking between the two methods, yet consistently rank among the most effective features for classifying SFGs and AGN in the MIGHTEE dataset.

However, as shown in Figure 2, while clear differences exist between the distributions of SFGs and AGN across many features, substantial overlap occurs in individual features. Nonetheless, as demonstrated by our subsequent analyses and the results presented in Section 4, the performance of all ML models improves by incorporating multiple features when classifying radio sources as SFGs or AGN.

### 3.2.2 Feature Correlation (Two Dimensional analysis)

Since conventional classifications of SFGs and AGN, including that of Whittam et al. (2022), rely on combinations of multiple features, we generate all possible pairs of the six features shown in Figure 2 to investigate the correlation between features and their impact on the performance of ML models in classifying SFGs and AGN from the MIGHTEE-COSMOS survey. A total of 15 correlation plots are created, with three highlighted in Figure 3 and the remainder provided in Figure B1 (Appendix B).

Figure 3a presents the  $q_{\text{IR}}$  plots against stellar mass. Whittam et al. (2022) apply the mass- and redshift-dependent IRRC from Delvecchio et al. (2021) to identify radio-excess AGN. As a result, the combination of these two features tends to miss radio-quiet AGN, causing most SFGs to fall within the 95% confidence ellipse of AGN (indicated by the blue dashed ellipse in Figure 3a). Although combining  $q_{\text{IR}}$  with  $\log(S_{4.5}/S_{3.6})$  marginally enhances the classification of radio-quiet AGN, the confidence ellipses of the two populations remain significantly overlapped, as illustrated in Figure 3b. The Figure 3c (also see Figure B1b) illustrates the two populations plotted in the IRAC colour-colour feature space. Despite the substantial overlap between AGN and SFGs, the confidence ellipses exhibit different correlation directions: AGN display a positive correlation with the two



**Figure 3.** Feature correlation plots for pairs of features selected for classifying SFGs and AGN in the MIGHTEE-COSMOS radio source. Figure 3d shows two-dimensional feature space generated by t-SNE. The open red circles represent SFGs, while the blue dots represent AGN. The solid red ellipses outline the 95% confidence for SFGs, while the dashed blue ellipses represent the 95% confidence for AGN. The orientation and shape of each ellipse represent the strength and direction of the correlation between the paired features and the corresponding galaxy classifications.

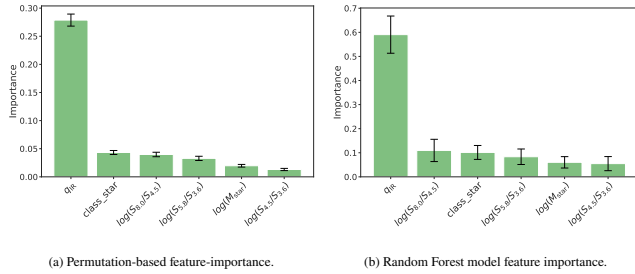
IRAC colours, whereas SFGs show a negative correlation. This suggests a potential association between the IRAC colour-colour and the two classes. Thus, combining the two IRAC colours improves the performance of ML models in distinguishing between SFGs and AGN among the MIGHTEE-detected radio sources, which may reflect the established MIR colour-colour classification diagnostic (Donley et al. 2012).

The remaining feature correlation plots, presented in Figure B1, suggest that in certain cases, combining two features may enhance the performance of ML models in classifying SFGs and AGN from the MIGHTEE-detected radio sources, despite the substantial overlap between the confidence ellipses of the two populations. As demonstrated in Section 4, the combination of multiple feature improves the ML models' ability to distinguish SFGs from AGN, motivating the use of t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton 2008) to reduce the six input features used in Section 4 to a two-dimensional feature space. The resulting correlation plot, shown in Figure 3d, reveals a remarkable improvement in the separation between the two populations when compared with the initial feature pairs presented in the first three plots of Figure 3 and the correlation plots in Figure B1.

However, since t-SNE features are generated through an unsupervised process and do not lend themselves to straightforward physical interpretation (Balamurali & Melkumyan 2016), these features were not used to train the ML models in this study. In the subsequent sections, we further explore the significance of features selected from the MIGHTEE-COSMOS multi-wavelength catalogue through automated techniques to assess their contribution to model performance.

### 3.3 Automated Feature Analysis

In this section, we employ three automated methods that do not rely on the built-in feature importance algorithms of the ML model, making them independent of the ML model's internal mechanisms. These



**Figure 4.** Feature importance estimated by the Permutation (*left*) and RF (*right*) models. The importance in the Permutation model is derived from the mean scores based on 1000 permutations. For the RF model, importance is computed by measuring the reduction in impurity within a decision tree node when a specific feature is used to split the data. The evaluation metric used is the  $F1$ -score.

methods include permutation and RF feature importances (detailed in Section 3.3.1), sequential feature importance (Section 3.3.2), and the receiver operating characteristic (ROC) curves (Section 3.3.3), which are used to assess the significance of the selected features in classifying SFGs and AGN from radio sources. For comparison, in this section, we utilize the RF ML model, which has built-in algorithms to measure the importance of features, allowing us to evaluate its results alongside the permutation method (illustrated in Figure 4).

### 3.3.1 Feature Importance

Permutation feature importance is defined as the decrease in an ML model score (we use  $F1$ -score as the evaluation metric) when a single feature values are randomly shuffled. This shuffling disrupts the true relationship between the feature and the target variable, resulting in degraded model performance. The extent of the performance drop reflects the feature’s importance, i.e., features causing greater drops when their relationship is disrupted are considered more significant. A detailed mathematical explanation of this method is provided in Molnar (2025).

Figure 4a presents the results of permutation feature importance in distinguishing SFGs and AGN from the MIGHTEE-COSMOS survey. The importance score is defined as the mean performance score obtained over 1,000 permutations. In this section, we present the permutation feature importance for the six most effective features. Complete results for all 18 features selected from the MIGHTEE-COSMOS multi-wavelength catalogue are shown in Figure A1 and are discussed in Section §5. Notably, Figure 4a shows that permutation feature importance yields a ranking consistent with that of the one-dimensional analysis (Section §3.2.1), further confirming these features’ efficiency in classifying radio-detected sources as SFGs or AGN.

Figure 4b illustrates the feature importance determined by the RF model. The built-in RF importance is computed using two methods: *Gini importance* (also known as mean decrease impurity (MDI)) and *Mean Decrease Accuracy* (MDA). In this study, we employ MDI since MDA closely mirrors the Permutation feature importance method. The details of RF MDI can be found in Li et al. (2019). In brief, this importance is calculated by evaluating the reduction in impurity (or randomness) within a decision tree node when a specific feature is used to split the data. The RF model also identifies  $q_{\text{IR}}$  as the most effective feature, though it switches the ranking of  $\text{class\_star}$  and  $\log(S_{8.0}/S_{4.5})$ . However, the difference in importance scores between these two features is negligible.

**Table 3.** Sequential feature importance results

| $N^a$ | $M^b$ | Features selected                                                                                                        |
|-------|-------|--------------------------------------------------------------------------------------------------------------------------|
| 6     | 1     | $q_{\text{IR}}$                                                                                                          |
| 6     | 2     | $q_{\text{IR}}$ and $\text{class\_star}$                                                                                 |
| 6     | 3     | $q_{\text{IR}}$ , $\text{class\_star}$ , and $\log(S_{8.0}/S_{4.5})$                                                     |
| 6     | 4     | $q_{\text{IR}}$ , $\text{class\_star}$ , $\log(S_{8.0}/S_{4.5})$ , and $\log(M_{\text{star}})$                           |
| 6     | 5     | $q_{\text{IR}}$ , $\text{class\_star}$ , $\log(S_{8.0}/S_{4.5})$ , $\log(M_{\text{star}})$ , and $\log(S_{5.8}/S_{3.6})$ |

<sup>a</sup>  $N$  represents the initial set of features, where  $N = 6$  in this study.

<sup>b</sup>  $M$  is the reduced set of features,  $M < N$ .

### 3.3.2 Sequential Feature Importance

In this subsection, we apply a sequential feature selection approach to determine and evaluate the importance of multiple features. This method reduces the initial set of  $N$  features to  $M$  features, where  $M < N$ . The selected  $M$  features are optimized and used as input for the ML models. For more details on the implementation of sequential feature selection, please refer to the official documentation at [scikit-learn: Sequential Feature Selection](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html)<sup>1</sup>.

In this study, the six most effective features identified in the previous subsections are used to run the sequential feature selection model five times, with  $M$  ranging from 1 to 5. As shown in Table 3, the model ranks  $q_{\text{IR}}$ ,  $\text{class\_star}$ , and  $\log(S_{8.0}/S_{4.5})$  as the three most essential features, respectively. Compared to the results from the permutation and RF feature importance models, the feature selection model alters the ranking of  $\log(M_{\text{star}})$  and  $\log(S_{5.8}/S_{3.6})$ . This is likely due to the sequential permutation method, which tends to consider only one feature when two or more features are highly correlated.

### 3.3.3 ROC Curve

The ROC curve is a graphical tool used to assess the performance of ML classifiers across all classification thresholds. It plots the true positive rate (TPR or *Recall*, as defined in Equation 2) against the false positive rate (FPR), which is defined as:

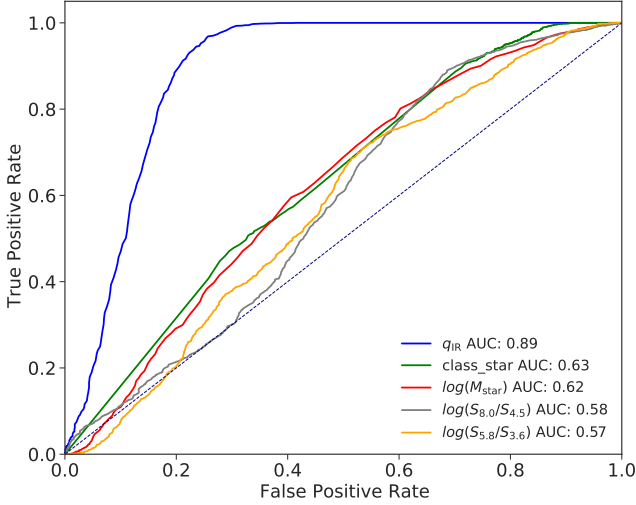
$$FPR = \frac{FP}{FP + TN}. \quad (4)$$

In this subsection, we also begin with a total of 18 features, comprising both the multi-wavelength photometric measurements from the MIGHTEE-COSMOS catalogue and the conventional classification diagnostics previously employed by Whittam et al. (2022). The ROC curves, computed based on thresholds applied individually to each feature, are presented in Figure 5. The area under the curve (AUC) is used to evaluate the effectiveness of these features in distinguishing SFGs from AGN in the MIGHTEE-COSMOS survey.

To mitigate potential bias arising from highly informative features, such as  $q_{\text{IR}}$ , which may dominate the feature space and obscure the contribution of other variables, we also implement an iterative feature-ranking procedure. In each iteration, the most dominant feature, based on AUC performance, is removed from the feature set, and the process is repeated on the remaining features. This process continues until a complete ranking is established. The AUC values of the top-ranked features at each iteration are summarized in Table 4.

As expected, the  $q_{\text{IR}}$  achieves the maximum AUC, indicating it is the most significant feature for distinguishing SFGs and AGN from

<sup>1</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.SequentialFeatureSelector.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html)



**Figure 5.** Receiver Operating Characteristic (ROC) curves for the five selected features, namely, the  $q_{\text{IR}}$  (blue),  $\text{class\_star}$  (green),  $\log(M_{\text{star}})$  (red),  $\log(S_{8.0}/S_{4.5})$  (grey), and  $\log(S_{5.8}/S_{3.6})$  (yellow).

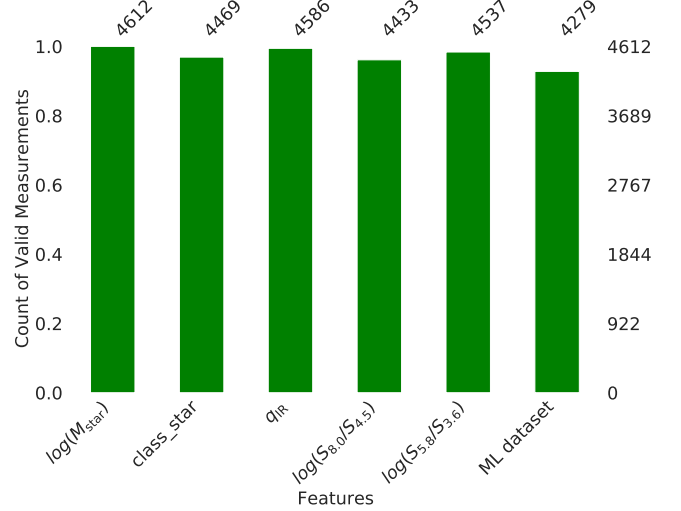
**Table 4.** AUC values of the top-ranked features identified at each iteration

| Iteration run | Number of total features | Top-ranked features     | AUC   |
|---------------|--------------------------|-------------------------|-------|
| 0             | 18                       | $q_{\text{IR}}$         | 0.886 |
| 1             | 17                       | $\text{class\_star}$    | 0.630 |
| 2             | 16                       | $\log(M_{\text{star}})$ | 0.619 |
| 3             | 15                       | $\log(S_{8.0}/S_{4.5})$ | 0.580 |
| 4             | 14                       | $\log(S_{5.8}/S_{3.6})$ | 0.574 |
| 5             | 13                       | $\log(i/z)$             | 0.570 |
| ...           | ...                      | ...                     | ...   |
| 9             | 9                        | $\log(S_{4.5}/S_{3.6})$ | 0.542 |
| ...           | ...                      | ...                     | ...   |
| 17            | 1                        | $\log(H/K_s)$           | 0.500 |

the radio sources, followed by  $\text{class\_star}$ . Contrary to the findings in previous subsections, the ROC curves suggest that  $\log(M_{\text{star}})$  is the third most important feature. The IRAC colour indices,  $\log(S_{8.0}/S_{4.5})$  and  $\log(S_{5.8}/S_{3.6})$ , are identified as the fourth and fifth most informative features, respectively, although their individual discriminative power remains low, with  $\text{AUC} \leq 0.6$ . In contrast,  $\log(S_{4.5}/S_{3.6})$  fails to rank among the top six features, displaying an AUC score close to 0.5, indicating a performance akin to random classification.

### 3.4 Feature Selection

Sections §3.2 and §3.3 detail our feature analyses, which combine one-dimensional, two-dimensional, ML-independent, and ML-dependent analyses to identify the most effective features for classifying SFGs and AGN among the radio-detected sources in the MIGHTEE-COSMOS survey. Across all feature analysis methods, the  $q_{\text{IR}}$  parameter consistently emerges as the most significant feature. This is likely because the majority (74%) of AGN in our sample are radio-excess AGN, which are conventionally distinguished from SFGs based on their  $q_{\text{IR}}$  values. The optical compactness parameter,  $\text{class\_star}$ , is consistently ranked among the top three. In addition,  $\log(M_{\text{star}})$  and two IRAC colours, namely,  $\log(S_{8.0}/S_{4.5})$



**Figure 6.** The completeness of the five features selected to train ML models from the MIGHTEE-COSMOS multi-wavelength catalogue. The left vertical axis indicates the completeness fraction for each feature, while the right vertical axis displays the corresponding numbers. The bottom horizontal axis lists the names of the selected features, and the top horizontal axis represents the total number of valid measurements for each feature. ML dataset bar represents the number of sources in the final sample used for ML.

and  $\log(S_{5.8}/S_{3.6})$ , are generally among the five most important features across most analyses. By contrast, the remaining IRAC colour,  $\log(S_{4.5}/S_{3.6})$ , although occasionally ranked sixth, is shown by the ROC-based AUC metric, which is a model-independent measure of feature discriminative power, to perform comparably to random classification.

Another crucial factor in feature selection is completeness, defined as the fraction of sources with measured values for the chosen features. As discussed in Section §2.2, while X-ray and VLBI detections are highly effective diagnostics for identifying AGN, their limited completeness results in approximately 70% of MIGHTEE sources remaining unclassified if only these two features are used (as shown in Figure 1).

Balancing completeness and classification efficiency, we select five key features for the subsequent ML analyses: the IRAC parameter ( $q_{\text{IR}}$ ), optical compactness ( $\text{class\_star}$ ), stellar mass ( $\log(M_{\text{star}})$ ), and two IRAC colours:  $\log(S_{8.0}/S_{4.5})$  and  $\log(S_{5.8}/S_{3.6})$ . A detailed description of these features is provided in Section §2.1, where they are outlined as conventional diagnostics frequently employed in the literature to classify sources as SFGs or AGN. The effectiveness of these features in ML-based classification of SFGs and AGN among MIGHTEE-detected sources is thoroughly evaluated in Sections §3.2 and §3.3.

As shown in Figure 6, for the 4612 labeled sources in Whittam et al. (2022), the completeness of these selected features  $> 96\%$  ( $4433/4612$ ). Due to the unpredictable nature of some of these astronomical features, we include 4279 sources with valid measurements for all five features in our ML analyses. This approach is justified, as sources with missing measurements account for approximately 7% of the entire dataset. As shown in Figure A2, the inclusion of additional features does not improve the performance of ML classification but slightly reduces the completeness of the dataset available for ML analysis.



**Figure 7.** The three-fold cross-validation hyperparameter tuning using grid search technique. The data is split into three folds with each fold used for testing (highlighted in red) while the remaining two folds are used for training (green). Image inspired by (Shatnawi et al. 2022).

### 3.5 Supervised ML Classification

As illustrated in the previous subsection, the dataset used for ML analyses consists of 4279 sources, with 1,526 classified as AGN and 2753 as SFGs according to Whittam et al. (2022). This sample is used to construct the training and test datasets and to optimize the ML models.

For binary classification, two distinct approaches can be employed: the first involves a straightforward dichotomous distinction between the two classes, where class labels 0 or 1 are assigned to an unknown source. The second approach models the probability  $P(y|X)$ , providing both a class label and the probability of class membership for a given source. In this study, we implement five different supervised classification algorithms. SVM uses the first approach, while the remaining four, namely, LR,  $k$ NN, RF, and XGB, adopt the second approach, estimating class probabilities.

The implementations of these classification algorithms are readily available through widely-used open-source libraries such as `scikit-learn`<sup>2</sup> and `XGBoost`<sup>3</sup>. Therefore, detailed descriptions of these algorithms are omitted here, and the focus of this work is on optimizing these models for classifying SFGs and AGN from the radio-detected sources.

#### 3.5.1 Hyperparameter Optimization

Hyperparameters determine the structure and behavior of an ML model before training, and optimizing them remains a trial-and-error process. Two common approaches to identifying the optimal set of hyperparameters for maximizing model performance are grid search and random search parameter tuning. Grid search exhaustively explores all possible combinations of hyperparameters, systematically evaluating each one to identify the best performance configuration. In contrast, random search generates random combinations of hyperparameters and evaluates a subset of them. The classifier with the best accuracy from the random search is then considered optimal.

In this work, we perform  $k$ -fold cross-validation using the grid search technique to identify the optimal hyperparameters for each ML model. Cross-validation (CV) is a resampling method used to evaluate the generalization ability of predictive models and to prevent overfitting (Berrar 2018). The data is split into  $k$  folds with each fold used for testing while the remaining  $k-1$  folds are used for training. For this work, we use  $k = 3$ , as illustrated in Figure 7. We highlight some optimized hyperparameters for different ML models in Appendix C.

**Table 5.** Five feature combinations used for training the ML models

| Name of combination | Features                                                                                                                 |
|---------------------|--------------------------------------------------------------------------------------------------------------------------|
| F1                  | $q_{\text{IR}}$                                                                                                          |
| F2                  | $q_{\text{IR}}$ and $\text{class\_star}$                                                                                 |
| F3                  | $q_{\text{IR}}$ , $\text{class\_star}$ , and $\log(M_{\text{star}})$                                                     |
| F4                  | $q_{\text{IR}}$ , $\text{class\_star}$ , $\log(M_{\text{star}})$ , and $\log(S_{8.0}/S_{4.5})$                           |
| F5                  | $q_{\text{IR}}$ , $\text{class\_star}$ , $\log(M_{\text{star}})$ , $\log(S_{8.0}/S_{4.5})$ , and $\log(S_{5.8}/S_{3.6})$ |

## 4 RESULTS

In this section, we present the results of ML approaches for classifying SFGs and AGN among radio sources detected by the MIGHTEE survey using selected input features and optimized ML models. As outlined previously, five widely used supervised ML models, namely, LR, SVM,  $k$ NN, RF, and XGB, are employed for this classification task. Our feature analyses suggest that combining multiple features improves the performance of ML models in distinguishing between SFGs and AGN in radio-detected sources (Section §3.2.2). We, therefore, create five distinct feature combinations (Table 5), guided by the ROC-based AUC metric, which offers a model-independent evaluation of feature importance.

### 4.1 Cross-validation Results

The performance of ML models in classifying SFGs and AGN from radio surveys is systematically evaluated through cross-validation techniques. We use random stratified sampling to partition our ML dataset, comprising 4279 MIGHTEE sources, into *training* and *validation* sets. Specifically, we implement several training-to-validation splits, i.e., [1:4], [2:3], [3:2], and [4:1], to assess model performance across a range of training data ratios.

These variations in the sizes of the training and validation datasets are motivated by the anticipated scale of future radio continuum surveys, such as those planned for SKA1 and the full SKA, which are projected to detect billions of radio sources. In contrast, current radio surveys have identified only tens of thousands of sources. As a result, the labeled data available for training ML models is likely to constitute merely a small fraction of the total dataset expected from future surveys. Therefore, it is imperative to evaluate the models' ability to maintain robust performance in scenarios where training data is limited.

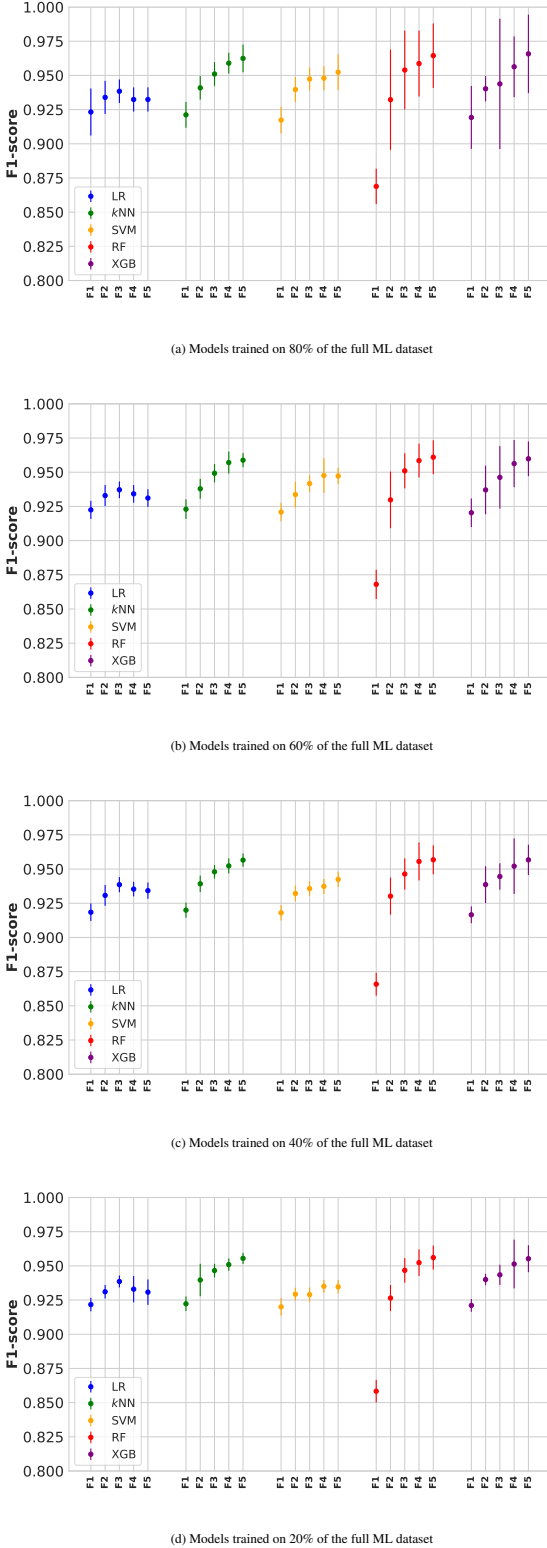
Figure 8 shows the performance of LR,  $k$ NN, SVM, RF, and XGB models trained on input feature combinations described in Table 5 in distinguishing SFGs from AGN from the *validation data*. Figures 8a, 8b, 8c and 8d show the results of models trained on 80%, 60%, 40% and 20% of the complete ML dataset respectively. All the classifiers are evaluated using the  $F1$ -score. The error bars shown in Figure 8 are the standard deviations of the  $F1$  score calculated using jackknife resampling.

#### 4.1.1 Results based on different feature combinations

For each training data size, all five ML models demonstrate strong performance across various feature combinations, consistently achieving  $F1$ -scores  $> 90\%$ , with the exception of the RF model trained using only the  $q_{\text{IR}}$  feature. The models exhibit slight variations in performance depending on the specific feature combinations. Notably, the LR model attains the highest  $F1$ -score when using the combination of  $q_{\text{IR}}$ ,  $\text{class\_star}$ , and  $\log(M_{\text{star}})$ . Introducing additional

<sup>2</sup> scikit-learn: <https://scikit-learn.org/stable/>

<sup>3</sup> XGBoost: <https://xgboost.readthedocs.io/en/stable/>



**Figure 8.** Evaluation of the performance of five supervised ML models, Logistic Regression (blue),  $k$ -Nearest Neighbour (green), Support Vector Machine (yellow), Random Forest (red) and XGBoost (purple), in classifying SFGs and AGN from the *validation data*. The  $F1$ -score is used as the evaluation metric. Feature combinations F1 through F5, as outlined in Table 5, are used for training the ML models, respectively. Subplots (a), (b), (c), and (d) display the results of all five models trained on 80%, 60%, 40%, and 20% of the complete ML dataset, respectively. Error bars represent the standard deviation derived from jackknife resampling.

input features does not enhance the LR model’s performance and instead leads to a slight decline, indicating that the inclusion of the two IRAC colours,  $\log(S_{8.0}/S_{4.5})$  and  $\log(S_{5.8}/S_{3.6})$ , may introduce redundancy or noise, thereby diminishing its discriminative effectiveness.

In contrast, the remaining four ML models benefit from the inclusion of the two IRAC colour indices, although the improvement for the boundary-based SVM classifier is generally marginal. For the other three models, excluding these IRAC colours leads to a noticeable decline in classification accuracy. This highlights the importance of these IRAC colour features for effective classification of SFGs and AGN in the MIGHTEE survey. Therefore, the absence of 5.8 and  $8.0\mu\text{m}$  observations will be a disadvantage for ML approaches in classifying radio-detected sources from future radio continuum surveys.

#### 4.1.2 Results based on different ML models and training sets

We also evaluate the performance of all five ML models and compare their results across different training sets. For clarity, the  $F1$ -score of the LR classifier trained on the feature combination F1 ( $q_{\text{IR}}$ ) is used as *baseline* (Figure 9). Figure 8 and Figure 9 show that, in classifying SFGs and AGN from the radio-detected sources,  $k\text{NN}$ , RF, and XGB perform slightly better than the LR and SVM classifiers, particularly when trained on the feature combinations of F3, F4, and F5. However, the jackknife scatter for the two decision-tree-based models (RF and XGB) is notably higher. Therefore, among the five ML models considered, the  $k\text{NN}$  classifier, which determines the membership of the class based on distance metrics (e.g. Euclidean distance) to identify the nearest neighbors, offers the most sustainable and interpretable approach. Its consistent performance and low variance make it a compelling choice for classifying SFGs and AGN in current and future radio continuum surveys.

Figure 9 further demonstrates that as the size of the training dataset decreases, the performance of all ML classifiers experiences a slight decline. Nonetheless, all models achieve an  $F1$ -score  $> 90\%$  across various feature combinations, even when trained with only 20% of the available data (with the exception of the RF model trained solely on the  $q_{\text{IR}}$  feature). This outcome underscores the robustness of ML approaches in classifying SFGs and AGN, even with limited training data.

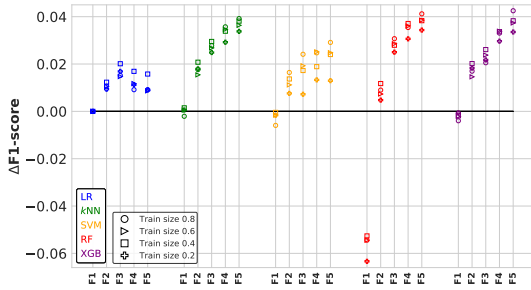
Overall, these findings demonstrate the effectiveness of ML techniques in the classification of radio sources, thereby reinforcing their promise for application in forthcoming large-scale radio continuum surveys to be conducted with next-generation interferometric facilities, such as the SKA and the ngVLA.

## 5 DISCUSSIONS

In this study, we assess the performance of five widely used supervised ML algorithms in classifying SFGs and AGN from the MIGHTEE-COSMOS radio continuum survey. To construct training and test datasets, we use SFGs and AGN that have been conventionally classified from the MIGHTEE-COSMOS (Whittam et al. 2022). Additionally, we incorporate all available photometric data from the MIGHTEE-COSMOS multi-wavelength catalog (Whittam et al. 2024), alongside conventional classification diagnostics, to inform the selection of ML input features. As expected, the five parameters used in conventional classification prove to be the most effective. Although the other two conventional classification features, X-ray

**Table 6.** Recalls of X-ray-only and VLBI-only AGN

| ML models                             |                          | LR         | kNN         | SVM         | RF          | XGB         |
|---------------------------------------|--------------------------|------------|-------------|-------------|-------------|-------------|
| trained on 80% of the full ML dataset | Recall of X-ray-only AGN | (3.3±0.1)% | (16.7±0.2)% | (26.7±0.3)% | (20.0±0.3)% | (13.3±0.2)% |
|                                       | Recall of VLBI-only AGN  | 0          | (33.3±1.3)% | (66.7±1.3)% | (33.3±1.3)% | (33.3±1.3)% |
| trained on 20% of the full ML dataset | Recall of X-ray-only AGN | (6.0±0.1)% | (8.5±0.1)%  | (16.2±0.1)% | (14.5±0.1)% | (12.0±0.1)% |
|                                       | Recall of VLBI-only AGN  | 0          | 0           | (16.7±0.3)% | (16.7±0.3)% | (8.3±0.3)%  |

**Figure 9.** This figure mirrors Figure 8, but it uses the  $F1$ -score of the LR classifier trained on feature combination  $F1$  ( $q_{\text{IR}}$ ) as a baseline, shown as a solid black line. Distinct symbols represent different training set sizes, providing a clear visualization of the impact of training set size on the performance of ML models in classifying SFGs and AGN from radio-detected sources.

luminosity and VLBI detection, are excluded from our ML analyses due to their limited completeness, the classification still achieves an  $F1$ -score  $> 90\%$ . Consequently, the selection of input features is guided by an optimal balance between classification efficiency and completeness.

In this section, we first examine the ability of the ML models to recover AGN that are identified exclusively by their X-ray luminosity or VLBI detection, despite the exclusion of these features from the input set. In Section §5.2, we further explore the characteristics of the selected input features for AGN versus SFG classification and evaluate the impact of incorporating additional features from the MIGHTEE-COSMOS multi-wavelength catalog on the performance of ML models. We then apply dimensionality reduction techniques and evaluate their influence on model performance (Section §5.3), followed by an examination of the impact of data normalization (Section §5.4). We also address the issue of class imbalance (Section §5.5), which may influence certain supervised ML algorithms, potentially causing them to neglect the minority class. Finally, we discuss the limitations of using ML approaches to classify SFGs and AGN from the extragalactic radio continuum survey.

### 5.1 X-ray and VLBI classifications

As presented in Section §3, the training set for our ML models is based on conventional classifications from Whittam et al. (2022), which include X-ray and VLBI classifications. However, due to the limited completeness and unpredictability of these X-ray and VLBI classifications, they are not used as input features for the ML classification. In our full ML dataset, there are 146 AGN identified solely based on their X-ray luminosity, while 15 are classified exclusively through VLBI detection. Although the number of VLBI-only AGN is negligible, X-ray-only AGN constitute approximately 10% of the

total AGN sample. Table 6 presents the recalls for these two AGN subpopulations as achieved by each ML model. Notably, when training with 20% of the dataset, only about 10% of the X-ray-only AGN are successfully recovered. This recovery fraction increases to approximately 20% when 80% of the dataset is used for training, with the exception of the LR and XGB models.

This result is significant because obtaining deep and wide X-ray data will remain a challenge for at least the next 15 years, until the launch of ESA’s Athena X-ray observatory<sup>4</sup>. During the operational periods of MeerKAT and SKA1, the classification of radio continuum sources will, therefore, often proceed without X-ray data. Our ML approach indicates that incorporating even limited X-ray observations into model training can marginally improve classification recall.

### 5.2 Input Features

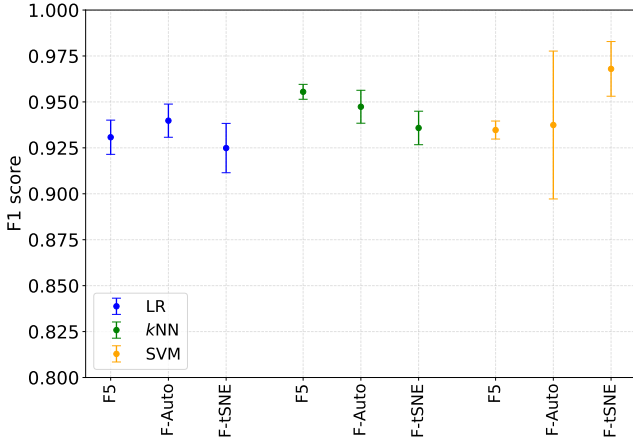
As shown in Section §3, we select the five most effective features in classifying SFGs and AGN from the MIGHTEE-COSMOS survey. For the five selected features, our feature analyses consistently indicate that the  $q_{\text{IR}}$  parameter is the most effective feature in distinguishing between the two classes of radio sources. This is likely due to the fact that the majority (74%) of AGN in the sample are radio-excess AGN, which are traditionally separated from SFGs using  $q_{\text{IR}}$ . However, as evidenced by the ML cross-validation results (Section §4), all ML models, except the RF, achieve  $F1$ -scores exceeding 90% when trained only with  $q_{\text{IR}}$ .

The classification utility of  $q_{\text{IR}}$  arises from AGN-dominated sources exhibiting substantially more accelerated CR electrons than would be expected from star formation alone, producing an observable ‘excess’ in radio emission relative to infrared emission. Although this excess may vary with redshift, stellar mass, or radio spectral index (e.g., Delvecchio et al. 2021; An et al. 2021),  $q_{\text{IR}}$  remains a robust and effective parameter to distinguish between SF- and AGN-dominated radio sources.

The optical compactness parameter,  $\text{class\_star}$ , also ranks among the top three features across all feature selection methods, as discussed in Section §3. This parameter is particularly useful for identifying optical point-like AGN among radio sources, offering a straightforward approach to differentiation. While the IRAC colour index may not be the most individually effective feature for classification, our two-dimensional feature analyses underscore the significance of combining the two IRAC colours for improved separation of AGN from SFGs.

Overall, as shown in Figure 8, we observe improvements in ML model performance with additional features incorporated into the training dataset. However, beyond the selected five features, adding further optical or NIR photometric data or colours does not improve

<sup>4</sup> <https://www.the-athena-x-ray-observatory.eu/en>



**Figure 10.** Comparison of  $F1$ -scores for feature combination  $F5$ ,  $q_{IR}$  with autoencoder-compressed features (F-Auto), and  $q_{IR}$  and  $t$ -SNE-projected features (F-tSNE) across LR,  $kNN$ , and SVM classifiers.

classification accuracy but slightly reduces dataset completeness, as shown in Appendix A.

### 5.3 Feature Space Dimensionality Reduction

Our analysis reveals that  $q_{IR}$  serves as the most discriminative feature for distinguishing between SFGs and AGN in radio continuum surveys. While the inclusion of four additional features enhances the classification performance of ML models, it also introduces increased model variance in most cases (Figure 8). To address this, we implement a two-step dimensionality reduction approach: (1) feature selection to retain the most informative predictors, as described in Section §3.2; (2) non-linear compression of the feature space using either:

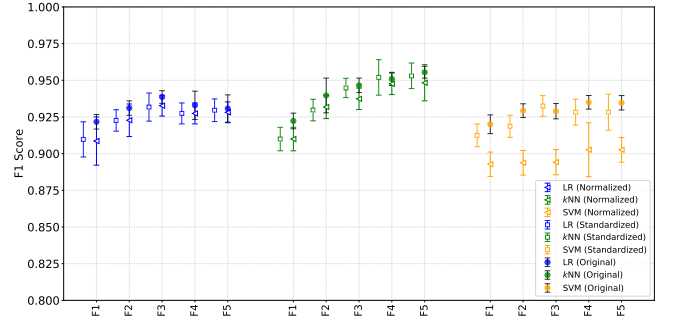
- Autoencoder: A lightweight symmetric autoencoder trained over 10,000 epochs to minimize reconstruction error ( $MSE=0.32$ ), compressing the selected feature set into a two-dimensional latent space; or
- $t$ -SNE: As a comparative method, we also apply  $t$ -distributed stochastic neighbor embedding to project the same feature set into two dimensions.

Using the five most informative features, we train the LR,  $kNN$ , and SVM classifiers on two compressed feature sets:

- F-Auto:  $q_{IR}$  combined with two autoencoder-derived latent dimensions (Auto1, Auto2),
- F-tSNE:  $q_{IR}$  combined with  $t$ -SNE-projected dimensions (t-SNE1, t-SNE2).

Contrary to expectations, both dimensionality reduction methods increase model variance rather than stabilizing performance (Figure 10). While the SVM classifier achieves a modest  $\sim 4\%$  improvement with F-tSNE, the performance of  $kNN$  and LR declined, suggesting that the two-dimensional projections may oversimplify complex non-linear relationships or that the limited input set ( $N = 5$ ) constrains the extraction of meaningful latent structure.

To further investigate, we conduct two additional tests. First, we increase the latent dimensionality (e.g.,  $n = 3$ ), but observed similar performance degradation. Second, we expand the feature set to include the top nine features (excluding  $q_{IR}$ ), identified via ROC-based importance metrics:  $[class\_star, \log(M_{star}), \log(S_{8.0}/S_{4.5})$ ,



**Figure 11.**  $F1$ -score performance for the LR (blue),  $kNN$  (green) and SVM (yellow) classifiers trained on the original (solid circles), min-max normalized (open triangles), and z-score standardized (open squares) datasets, across various feature combinations. The results indicate that feature scaling has a negligible impact on the performance of the LR and  $kNN$  models. In contrast, the SVM demonstrates a statistically significant decrease in performance when trained on the normalized dataset.

$\log(S_{5.8}/S_{3.6})$ ,  $\log(i/z)$ ,  $\log(r/z)$ ,  $\log(g/z)$ ,  $\log(Y/H)$ ,  $\log(S_{4.5}/S_{3.6})$ , and repeated the compression experiments. In this case as well, both autoencoder and  $t$ -SNE transformations generally reduced classifier performance across LR,  $kNN$ , and SVM, with the only exception being a modest gain for SVM when combined with F-tSNE.

Taken together, these results indicate that dimensionality reduction is not effective in our case, possibly due to the relatively small sample size and the dominance of a single very prominent feature ( $q_{IR}$ ). We therefore conclude that retaining the original five-feature combination (F5), without additional dimensionality reduction, provides the most reliable classification performance for distinguishing SFGs and AGN in our MIGHTEE-COSMOS survey.

### 5.4 Feature Scaling

Data scaling is an important preprocessing step in data analysis and ML (Korobchynskiy & Nadraga 2025). It involves transforming features into a consistent scale or format to enhance the efficiency and performance of computational models. This step is particularly essential when raw datasets contain variables with heterogeneous units, scales, or distributions, which may negatively impact model training and convergence (Ali et al. 2014). Common techniques include (Mahmud Sujon et al. 2024):

- min-max scaling (normalization), which rescales values to a fixed range (typically  $[0,1]$ ),
- z-score standardization, which centers data to zero mean and unit variance, and
- robust scaling, which uses medians and interquartile ranges to mitigate the influence of outliers.

In our study, we adopt min-max normalization and z-score standardization to all features to harmonize the feature space and assess its effect on model performance. We focus on LR,  $kNN$ , and SVM models since they are known to be sensitive to feature scaling (Mahmud Sujon et al. 2024). Both LR and SVM depend on the orientation of the decision boundary in the feature space, which can be skewed by unscaled inputs. Similarly,  $kNN$  depends on raw distance metrics (e.g., Euclidean distance), making it susceptible to variations in feature scales. RF and XGB, on the other hand, are tree-based models and do not rely on distance calculations or gradient-based updates that depend on feature scale. These models split nodes based on

**Table 7.** ROC-based AUC values based on balanced dataset

| Feature ranking | Input features          | AUC values |
|-----------------|-------------------------|------------|
| 1               | $q_{\text{IR}}$         | 0.886      |
| 2               | class_star              | 0.630      |
| 3               | $\log(M_{\text{star}})$ | 0.621      |
| 4               | $\log(S_{8.0}/S_{4.5})$ | 0.574      |
| 5               | $\log(S_{5.8}/S_{3.6})$ | 0.574      |

feature thresholds, so they are largely invariant to monotonic transformations like normalization.

Figure 11 indicates that both implemented data scaling techniques have only a marginal effect on the performance of the  $k$ NN and LR models. In contrast, the SVM exhibits a statistically significant degradation in performance when trained on the normalized dataset. Given this counterintuitive outcome, and in light of the established scale invariance of tree-based ensemble methods such as RF and XGB, we conclude that feature scaling provided no substantive benefit to our modeling framework. Consequently, we opt to employ the original, unscaled dataset in all our analyses to avoid introducing unnecessary preprocessing artifacts while maintaining the integrity of the underlying feature distributions.

### 5.5 Class Imbalance

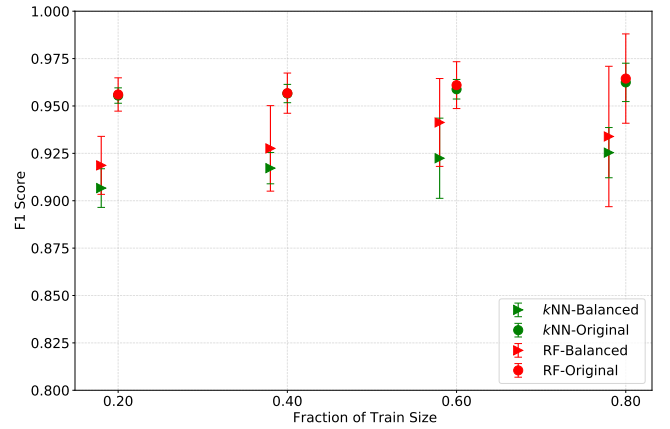
Imbalanced data refers to datasets with a pronounced skew in the distribution of class labels. This imbalance can affect many supervised ML algorithms, often causing them to overlook the minority class. This issue is particularly concerning, as predictions for the minority class are typically of the greatest importance (Das et al. 2022).

The typical approach to addressing data imbalance is to resample the training data randomly. The two standard methods are *Under-sampling* and *Oversampling*. Undersampling reduces the number of sources in the majority class, while oversampling duplicates examples from the minority class. In this study, we used undersampling to balance training data by removing some SFGs, resulting in an equal number of AGN and SFGs. This approach is appropriate here, as AGN typically constitute the minority class in most radio continuum surveys.

We first assess the effect of class imbalance on our feature selection by comparing ROC-based AUC metrics derived from both the original and balanced datasets. As summarized in Table 7, the balanced dataset yields a feature importance ranking consistent with that of the original dataset (Table 4), albeit with minor variations in absolute AUC values.

Secondly, we evaluate the performance of ML models in classifying SFGs and AGN from radio continuum survey data using both the original and balanced datasets. As shown in Figure 12, we train the  $k$ NN and RF models on varying fractions of the training data using the feature combination F5 ( $q_{\text{IR}}$ , class\_star,  $\log(M_{\text{star}})$ ,  $\log(S_{8.0}/S_{4.5})$ ,  $\log(S_{5.8}/S_{3.6})$ ) and assess performance based on the  $F1$ -score. For both the original and balanced datasets, model performance remains consistently high, exceeding 90% even when only 20% of the training data is used. In the case of the balanced dataset, we further validate model robustness by using the remaining SFGs not included in the training set as an independent test sample. The resulting performance metrics remain high, with a recall of  $(96 \pm 0.01)\%$  and an  $F1$ -score of  $(97 \pm 0.01)\%$ , regardless of whether 20% or 80% of the dataset is used for training.

As also illustrated in Figure 12, both ML models achieve slightly



**Figure 12.** Comparison of  $F1$ -score performance for the  $k$ NN and RF models trained on the original (circles) and class-balanced (triangles) datasets, shown as a function of the fraction of training data used. For visual clarity, data points corresponding to the balanced dataset are slightly offset leftward along the X-axis.

higher performance when trained on the imbalanced (original) dataset. This does not imply that imbalance is intrinsically beneficial, but rather reflects the fact that SFGs dominate the true underlying class distribution in deep radio continuum surveys. Although a more robust approach in ML classification is to train on balanced data to obtain well-calibrated probability models and subsequently incorporate prior information (such as the natural dominance of SFGs in this case), for simplicity and to remain consistent with the intrinsic survey distribution, we adopt the original dataset in our main analyses.

### 5.6 Limitations

While supervised ML models deliver state-of-the-art classification results for MIGHTEE-COSMOS radio sources, several limitations merit attention, such as the quality of training data and challenges posed by missing or invalid measurements.

The ML algorithm's mapping accuracy depends significantly on the quality of the labels in the training data. However, these outputs, derived using conventional methods, may be subject to biases and imperfections. For instance, our training set is based on the MIGHTEE-COSMOS multi-wavelength catalogue, where Whitam et al. (2022) employed five conventional techniques to label MIGHTEE-COSMOS radio sources. Although each diagnostic was applied independently, they are constrained by observational data quality, including data depth, coverage, and photometric accuracy.

Another challenge in ML-based classification of SFGs and AGN in radio continuum surveys is missing data or invalid measurements. Due to the unpredictability of astronomical properties, such as X-ray luminosity, VLBI detection, and optical or NIR photometry, we opted not to estimate these missing values using statistical imputation techniques (e.g., Pelckmans et al. 2005). While XGB can manage missing data, it does so by inferring values based on the available measured features. Consequently, we restricted our ML analysis to samples with valid measurements across all five selected input features. This choice excluded approximately 7% of radio sources with conventional labels in the MIGHTEE-COSMOS catalogue from the ML classification. Addressing such gaps will be a persistent challenge in applying ML classification to MIGHTEE and upcoming radio surveys.

## 6 CONCLUSIONS

In this study, we adopt and compare five supervised ML classification models, namely LR, SVM,  $k$ NN, RF, and XGB, to classify star-formation-dominated or black-hole-accretion-dominated radio sources from the MIGHTEE-COSMOS survey. Using a sample of 4279 MIGHTEE-COSMOS radio sources labeled by Whittam et al. (2022) as either SFGs or AGN, along with their associated multi-wavelength measurements, we evaluate ML performance in classifying SFGs and AGN from radio continuum surveys. Our main conclusions are as follows:

(i) We analyze and select the most effective features for training and testing ML models. As expected, our one-dimensional, two-dimensional, ML-independent, ML-dependent, and ROC curve analyses indicate that the five parameters used in conventional classification prove to be the most effective. The IRRC parameter,  $q_{\text{IR}}$ , is the most effective feature for distinguishing between SFGs and AGN. The optical compactness morphology parameter,  $\text{class\_star}$ , consistently ranks among the top three most effective features across all selection methods. While the IRAC colour may not be individually impactful, two-dimensional feature analyses reveal the importance of combining two IRAC colours for improved AGN-SFG separation. Therefore, the five features we selected to train ML models are  $q_{\text{IR}}$ ,  $\text{class\_star}$ , stellar mass, and two IRAC colours ( $\log(S_{8.0}/S_{4.5})$  and  $\log(S_{5.8}/S_{3.6})$ ). The dataset completeness for sources with valid measurements across these five features is 93%.

(ii) We optimized the ML models using these selected features and evaluated classifiers with various feature combinations, guided by ROC-based AUC metric. Our results indicate that, for most models, ML performance generally improves as more feature combinations are included. Additionally, excluding the MIR colour features  $\log(S_{8.0}/S_{4.5})$  and  $\log(S_{5.8}/S_{3.6})$  leads to a noticeable performance drop for most ML models. This finding suggests that future radio surveys in regions lacking deep 5.8 and 8.0  $\mu\text{m}$  observations may experience a slight disadvantage in accurately classifying radio sources as either SFGs or AGN.

(iii) We assess ML classification performance dependency on training data size by using 20%, 40%, 60% and 80% of the full dataset. All models achieve  $F1$ -scores greater than 90% with any training size, except for the RF model when trained with the single feature  $q_{\text{IR}}$  and a training set size of 20%.

(iv) Due to the limited completeness and unpredictability of X-ray and VLBI classifications, we do not include them as input features for training the ML models. However, our ML approach indicates that incorporating even limited X-ray observations into model training can marginally improve classification recall.

(v) We assess the impact of dimensionality reduction strategies and feature scaling and find that neither provides substantive benefits to our modeling framework. We also examine the effect of class imbalance in the MIGHTEE-COSMOS data and find that class imbalance does not impact ML model performance in our case. We therefore conclude that the unscaled dataset, combined with the original five-feature set (F5) and without additional dimensionality reduction, yields the most robust and reliable classification of SFGs and AGN in the MIGHTEE-COSMOS survey.

(vi) Overall, our results demonstrate that all ML models perform well in classifying SFGs and AGN from radio sources, achieving  $F1$ -score  $> 90\%$  even with a small fraction (20%) of the training data and a few key input features. Among the models assessed, the distance-based  $k$ NN classifier consistently emerges as the most accurate and stable, making it a compelling choice for the classification of SFGs

and AGN in future large-scale radio continuum surveys, such as those by next-generation radio interferometric facilities.

## ACKNOWLEDGEMENTS

We are grateful to the anonymous referee for a detailed report and valuable comments that improved the quality of this work. FXA acknowledges the support from the National Natural Science Foundation of China (12303016) and the Natural Science Foundation of Jiangsu Province (BK20242115). WS is grateful for support from the South African National Research Foundation (NRF) and National Astrophysics and Space Science Programme (NASSP). WS and MV acknowledge financial support from the Inter-University Institute for Data Intensive Astronomy (IDIA - a partnership between the University of Cape Town, the University of Pretoria and the University of the Western Cape). MV acknowledges financial support from the South African Department of Science and Innovation's National Research Foundation under the ISARP RADIOMAP Joint Research Scheme (DSI-NRF Grant Number 150551) and the CPRR HIPPO Project (DSI-NRF Grant Number SRUG22031677). EH expresses gratitude for the valuable discussions with Prof. Chris Thron. FXA and WS sincerely thank Prof. Ian Smail and Prof. Seb Oliver for their insightful suggestions. We acknowledge the use of the [rooibosTea\\_classification](#) code described by Hussein et al. (2022) as a guideline in our work.

The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. We acknowledge the use of the ilifu cloud computing facility - [www.ilifu.ac.za](http://www.ilifu.ac.za), a partnership between the University of Cape Town, the University of the Western Cape, Stellenbosch University, Sol Plaatje University, the Cape Peninsula University of Technology and the South African Radio Astronomy Observatory. The ilifu facility is supported by contributions from IDIA, the Computational Biology division at UCT and the Data Intensive Research Initiative of South Africa (DIRISA).

## DATA AVAILABILITY

The MIGHTEE Early Science continuum data used in this work is extensively detailed in Heywood et al. (2022). The MIGHTEE-COSMOS conventional classification catalogue and the cross-matched multi-wavelength catalogue were released with Whittam et al. (2022) and Whittam et al. (2024), respectively. All codes used in our analyses are publicly available on GitHub: <https://github.com/pfunzowalter/mightee-class-pub>.

## REFERENCES

- Adams N. J., Bowler R. A. A., Jarvis M. J., Häußler B., Lagos C. D. P., 2021, *MNRAS*, **506**, 4933
- Ali P. J. M., Faraj R. H., Koya E., Ali P. J. M., Faraj R. H., 2014, *Mach Learn Tech Rep*, 1, 1
- An F. X., et al., 2018, *ApJ*, **862**, 101
- An F., et al., 2021, *MNRAS*, **507**, 2643
- Balamurali M., Melkumyan A., 2016, in *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part IV* 23. pp 565–572
- Ball N. M., Brunner R. J., 2010, *International Journal of Modern Physics D*, **19**, 1049

- Berger V., Zhou Y., 2014, Kolmogorov–Smirnov Test: Overview, doi:10.1002/9781118445112.stat06558.
- Berraz D., 2018, Cross-Validation, doi:10.1016/B978-0-12-809633-8.20349-X.
- Best P. N., et al., 2023, *MNRAS*, **523**, 1729
- Biswas R., et al., 2013, *Phys. Rev. D*, **88**, 062003
- Bowler R. A. A., Jarvis M. J., Dunlop J. S., McLure R. J., McLeod D. J., Adams N. J., Milvang-Jensen B., McCracken H. J., 2020, *MNRAS*, **493**, 2059
- Breiman L., 2001, *Machine Learning*, **45**, 5
- Briggs D. S., 1995, in American Astronomical Society Meeting Abstracts. p. 112.02
- Chen T., Guestrin C., 2016, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. p. arXiv:1603.02754 (arXiv:1603.02754), doi:10.48550/arXiv.1603.02754
- Condon J. J., 1992, *ARA&A*, **30**, 575
- Cristianini N., Shawe-Taylor J., et al., 2000, An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press
- Cuillandre J.-C. J., et al., 2012, in Peck A. B., Seaman R. L., Comeron F., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 8448, Observatory Operations: Strategies, Processes, and Systems IV. p. 84480M, doi:10.1117/12.925584
- Das S., Mullick S. S., Zelinka I., 2022, *IEEE Transactions on Artificial Intelligence*, **3**, 973
- Delvecchio I., et al., 2021, *A&A*, **647**, A123
- Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, *IEEE Proceedings*, **97**, 1482
- Donley J. L., et al., 2012, *ApJ*, **748**, 142
- Duncan K. J., Jarvis M. J., Brown M. J. I., Röttgering H. J. A., 2018, *MNRAS*, **477**, 5177
- Florios K., Kontogiannis I., Park S.-H., Guerra J. A., Benvenuto F., Bloomfield D. S., Georgoulis M. K., 2018, *Sol. Phys.*, **293**, 28
- Fluke C. J., Jacobs C., 2020, *WIREs Data Mining and Knowledge Discovery*, **10**, e1349
- George D., Shen H., Huerta E. A., 2018, *Phys. Rev. D*, **97**, 101501
- Griffin M. J., et al., 2010, *A&A*, **518**, L3
- Hale C. L., et al., 2025, *MNRAS*, **536**, 2187
- Hatfield P. W., Jarvis M. J., Adams N., Bowler R. A. A., Häußler B., Duncan K. J., 2022, *MNRAS*, **513**, 3719
- Herrera Ruiz N., et al., 2017, *A&A*, **607**, A132
- Heywood I., et al., 2022, *MNRAS*, **509**, 2150
- Hossin M., Sulaiman M. N., 2015, *International journal of data mining & knowledge management process*, **5**, 1
- Hotan A. W., et al., 2021, *Publ. Astron. Soc. Australia*, **38**, e009
- Hussein E. A., Thron C., Ghaziasgar M., Vaccari M., Marnewick J. L., Hussein A. A., 2022, *Plants*, **11**, 16
- Ishwara-Chandra C. H., Taylor A. R., Green D. A., Stil J. M., Vaccari M., Ocran E. F., 2020, *MNRAS*, **497**, 5383
- Jarvis M., et al., 2016, in MeerKAT Science: On the Pathway to the SKA. p. 6 (arXiv:1709.01901), doi:10.22323/1.277.0006
- Jiang P., et al., 2024, *Astronomical Techniques and Instruments*, **1**, 84
- Jiménez-Andrade E. F., Murphy E. J., Momjian E., Condon J. J., Chary R.-R., Taylor R., Dickinson M., 2024, *ApJ*, **972**, 89
- Jonas J., MeerKAT Team 2016, in MeerKAT Science: On the Pathway to the SKA. p. 1, doi:10.22323/1.277.0001
- Karsten J., et al., 2023, *A&A*, **675**, A159
- Korobchynskiy M., Nadraga V., 2025, in Lecture Notes in Data Engineering, Computational Intelligence, and Decision-Making, Volume 2: 2024 International Scientific Conference "Intelligent Systems of Decision-Making and Problems of Computational Intelligence", Proceedings. p. 206
- Laigle C., et al., 2016, *ApJS*, **224**, 24
- Li X., Wang Y., Basu S., Kumbier K., Yu B., 2019, *Advances in Neural Information Processing Systems*, **32**
- Li C., et al., 2023, *MNRAS*, **518**, 513
- Lonsdale C. J., et al., 2009, *IEEE Proceedings*, **97**, 1497
- Maddox N., et al., 2021, *A&A*, **646**, A35
- Mahabal A., et al., 2019, *PASP*, **131**, 038002
- Mahmud Sujon K., Binti Hassan R., Tusnia Towshi Z., Othman M. A., Abdus Samad M., Choi K., 2024, *IEEE Access*, **12**, 135300
- Menard S. W., 2010, *Logistic regression: From introductory to advanced concepts and applications*. Sage
- Miley G., De Breuck C., 2008, *A&ARv*, **15**, 67
- Molnar C., 2025, *Interpretable Machine Learning*, 3 edn. <https://christophm.github.io/interpretable-ml-book>
- Murphy E. J., et al., 2018, in Murphy E., ed., *Astronomical Society of the Pacific Conference Series Vol. 517, Science with a Next Generation Very Large Array*. p. 3 (arXiv:1810.07524), doi:10.48550/arXiv.1810.07524
- Norris R. P., 2017, *Nature Astronomy*, **1**, 671
- Ocran E. F., Taylor A. R., Vaccari M., Ishwara-Chandra C. H., Prandoni I., 2020, *MNRAS*, **491**, 1127
- Oke J. B., 1974, *ApJS*, **27**, 21
- Pelckmans K., De Brabanter J., Suykens J., De Moor B., 2005, *Neural Networks*, **18**, 684
- Peterson L. E., 2009, *Scholarpedia*, **4**, 1883
- Planck Collaboration et al., 2016, *A&A*, **594**, A13
- Poglitsch A., et al., 2010, *A&A*, **518**, L2
- Rieke G. H., et al., 2004, *ApJS*, **154**, 25
- Sadler E. M., Jenkins C. R., Kotanyi C. G., 1989, *MNRAS*, **240**, 591
- Scoville N., et al., 2007, *ApJS*, **172**, 1
- Shatnawi A., Alkassar H. M., Al-Abdaly N. M., Al-Hamdany E. A., Bernardo L. F. A., Imran H., 2022, *Buildings*, **12**, 550
- Shirley R., et al., 2021, *MNRAS*, **507**, 129
- Smolčić V., et al., 2017, *A&A*, **602**, A2
- Steinhardt C. L., et al., 2014, *ApJ*, **791**, L25
- Swarup G., Ananthakrishnan S., Kapahi V. K., Rao A. P., Subrahmanya C. R., Kulkarni V. K., 1991, *Current Science*, **60**, 95
- Szokoly G. P., et al., 2004, *ApJS*, **155**, 271
- Tanaka M., et al., 2017, *arXiv e-prints*, p. arXiv:1706.00566
- Taylor A. R., et al., 2024, *MNRAS*, **528**, 2511
- Vaccari M., 2015, in *The Many Facets of Extragalactic Radio Surveys: Towards New Scientific Challenges*. p. 27 (arXiv:1604.02353), doi:10.22323/1.267.0027
- Van der Maaten L., Hinton G., 2008, *Journal of machine learning research*, **9**
- Whittam I. H., et al., 2022, *MNRAS*, **516**, 245
- Whittam I. H., et al., 2024, *MNRAS*, **528**, 1171
- Yadav S., Bhole G., 2020, *International Journal of Recent Technology and Engineering (IJRTE)*, **8**, 1907
- Zhang S., Zhang C., Yang Q., 2003, *Applied artificial intelligence*, **17**, 375
- van Haarlem M. P., et al., 2013, *A&A*, **556**, A2

## APPENDIX A: ADDITIONAL FEATURES

To efficiently classify SFGs and AGN from the radio continuum survey, we derive colour indices using flux densities in the MIR, NIR, and optical wavelengths from the MIGHTEE-COSMOS multi-wavelength catalogue (Whittam et al. 2024). Details on these multi-wavelength data are provided in Section §2.1. Briefly, we use twelve photometric bands, including HSC *griz*-band, IRAC 3.6, 4.5, 5.8, and 8.0  $\mu\text{m}$  data, along with UltraVISTA *YJHK<sub>s</sub>*-band photometries. In addition, other measurements available in the MIGHTEE-COSMOS catalogue, such as  $q_{\text{IR}}$ ,  $\text{class\_star}$ , and stellar mass are incorporated.

From these data, we select the most effective input features for ML analyses from a total of 18 parameters:  $q_{\text{IR}}$ ,  $\text{class\_star}$ ,  $\log(M_{\text{star}})$ , three MIR colours ( $\log(S_{8.0}/S_{4.5})$ ,  $\log(S_{5.8}/S_{3.6})$ ,  $\log(S_{4.5}/S_{3.6})$ ), and 12 NIR and optical colours ( $\log(g/r)$ ,  $\log(r/i)$ ,  $\log(i/z)$ ,  $\log(g/i)$ ,  $\log(g/z)$ ,  $\log(r/z)$ ,  $\log(Y/J)$ ,  $\log(J/H)$ ,  $\log(H/K_s)$ ,  $\log(Y/H)$ ,  $\log(Y/K_s)$ ,  $\log(J/K_s)$ ).

Figure A1 illustrates the permutation importance of these features, with our selected five features demonstrating the highest effectiveness in classifying SFGs and AGN among radio-detected sources.

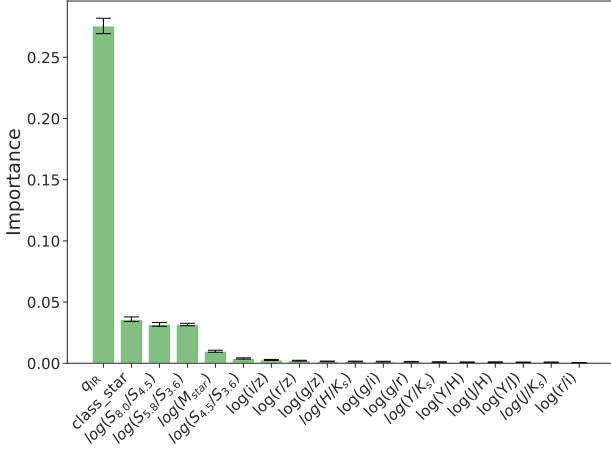


Figure A1. Permutation feature importance of 18 measurements.

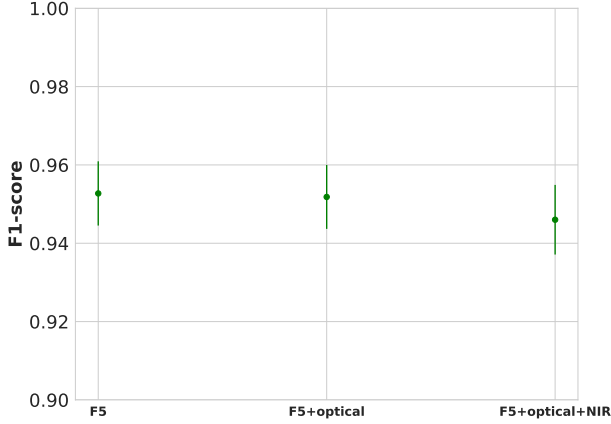


Figure A2. The results of applying the  $k$ NN classifier to distinguish between AGN and SFGs trained using three different feature combinations, namely, F5, F5+optical and F5+optical+NIR. The evaluation metric is the  $F1$ -score, with error bars representing the standard deviation obtained through jackknife resampling.

In addition, we incorporate these optical and NIR colours as input features to assess the performance of ML models. Using the  $k$ NN model as an example, we present results in Figure A2. Three feature combinations are used to train the  $k$ NN model: 1) F5, which includes  $\text{class\_star}$ ,  $q_{\text{IR}}$ ,  $\log(S_{8.0}/S_{4.5})$ ,  $\log(S_{5.8}/S_{3.6})$ ; 2) F5 + optical colours ( $\log(g/r)$ ,  $\log(r/i)$ ,  $\log(i/z)$ ,  $\log(g/i)$ ,  $\log(g/z)$ ,  $\log(r/z)$ ); and (3) F5 + optical + NIR colours ( $\log(Y/J)$ ,  $\log(J/H)$ ,  $\log(H/K_s)$ ,  $\log(Y/H)$ ,  $\log(Y/K_s)$ ,  $\log(J/K_s)$ ). The data are randomly split into 80% for training and 20% for testing, with model performance evaluated using the  $F1$ -score as the classification metric. As shown in Figure A2, adding these features does not improve or even slightly decrease the performance of the  $k$ NN classifier. This outcome is likely due to the additional features introducing confusion, which hampers the model's ability to effectively distinguish between SFGs and AGN. Furthermore, the completeness of the ML dataset is marginally reduced if all optical and NIR photometric measurements are required.

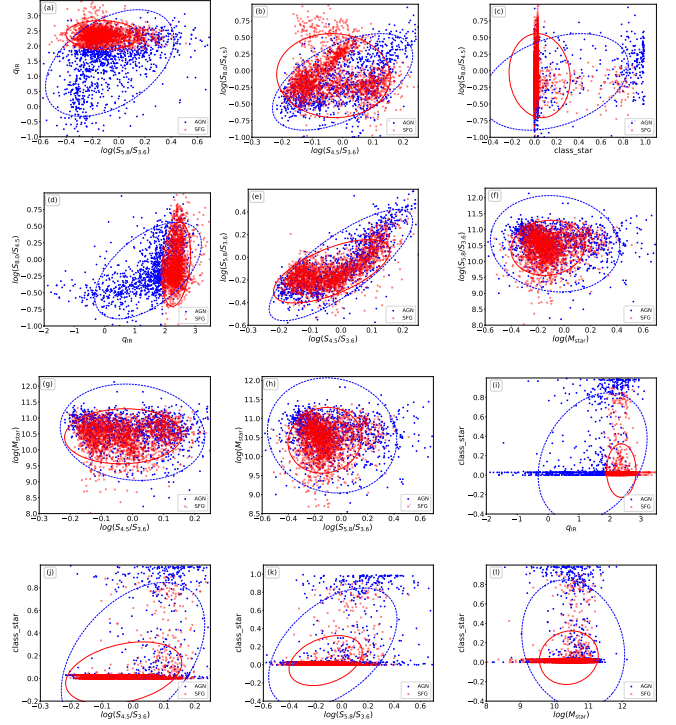


Figure B1. As with Figure 3, the feature correlation plots for the remaining 12 feature pairs are shown.

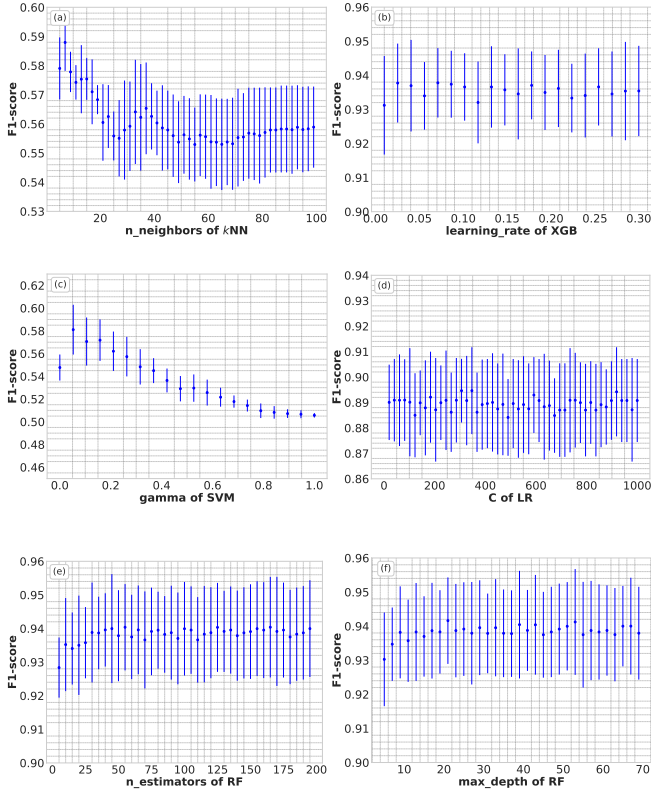
## APPENDIX B: FEATURE CORRELATION PLOTS

As described in Section §3.2.2, we examine the correlations among the six features shown in Figure 2, resulting in 15 correlation plots. Three of these plots are presented in Figure 3, with the remainder shown in this section (Figure B1). As further discussed in Section §3.2.2, combining certain feature pairs can, in some cases (for instance, the two IRAC colours), improve the performance of ML models for classifying SFGs and AGN from the radio continuum surveys, despite significant overlap between the confidence ellipses of these two populations.

## APPENDIX C: HYPERPARAMETERS

Hyperparameter optimization is a key step in ML classification. Section §3.5.1 details the methods employed for hyperparameter tuning in this study. Here, we provide examples illustrating the optimization of hyperparameters across various ML models.

As outlined in Section §3.5.1, we perform a three-fold split of the sample and apply a grid search technique to identify the optimal hyperparameters for each ML model. This process involves adjusting one hyperparameter at a time, while holding the others at their default values, and evaluating performance changes based on the  $F1$ -score. Figure C1 illustrates examples of how ML model performance varies with specific hyperparameters. For instance, in the  $k$ NN classifier, performance decreases as the *Number of Neighbors* increases, leading us to select a value of  $< 15$  for this hyperparameter. Figure C1c demonstrates that SVM performance improves with an increase in  $\gamma$ , a parameter used in the Radial Basis Function (RBF) kernel and other nonlinear kernels. Higher  $\gamma$  values increase the flexibility of the decision boundary, allowing it to adapt more closely to individual data points.



**Figure C1.** Six examples illustrate how the performance of ML models varies with specific hyperparameters, evaluated using the  $F1$ -score. For each model, the selected hyperparameter is varied while others remain at their default values. Error bars represent the  $F1$ -score standard deviation, calculated via jackknife resampling. This analysis provides insights into the sensitivity of model performance to parameter tuning, highlighting optimal configurations and trade-offs for each model type.

Not all hyperparameters, however, exhibit a monotonic relationship with model performance. Figure C1b shows fluctuations in XGB performance in response to the learning rate parameter, while Figure C1c suggests that LR classification performance remains largely unaffected by variations in the  $C$ -value, which controls regularization strength by balancing the trade-off between model fit and weight minimization to prevent overfitting. Additionally, Figures C1e and C1f demonstrate that RF classifier performance reaches its maximum when the hyperparameters  $n\_estimators$  and  $max\_depth$  exceed values of 40 and 10, respectively. We point out that this section provides only selected examples of hyperparameter optimization for ML models. For a comprehensive list of hyperparameters for each model, we refer readers to the [scikit-learn](#) and the [XGBoost](#).

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.