Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **537,** 1028–1055 (2025) https://doi.org/10.1093/mnras/staf080
Advance Access publication 2025 January 15

The VMC Survey – LI. Classifying extragalactic sources using a probabilistic random forest supervised machine learning algorithm

Clara M. Pennock [1,2★] Jacco Th. van Loon [1] Maria-Rosa L. Cioni [3] Chandreyee Maitra [4] Joana M. Oliveira [1] Jessica E. M. Craig [1] Valentin D. Ivanov,[5] James Aird [2] Joy O. Anih,[1] Nicholas J. G. Cross,[2] Francesca Dresbach,[1] Richard de Grijs[6,7,8] and Martin A. T. Groenewegen [9]

[1]*Lennard-Jones Laboratories, Keele University, Keele, ST5 5BG, UK*
[2]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK*
[3]*Leibniz-Institut für Astrophysik Potsdam, An der Sternwarte 16, D-14482 Potsdam, Germany*
[4]*Max-Planck-Institut für extraterrestrische Physik, Gießenbachstraße, D-85748 Garching, Germany*
[5]*European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching bei München, Germany*
[6]*School of Mathematical and Physical Sciences, Macquarie University, Balaclava Road, Sydney, NSW 2109, Australia*
[7]*Astrophysics and Space Technologies Research Centre, Macquarie University, Balaclava Road, Sydney, NSW 2109, Australia*
[8]*International Space Science Institute-Beijing, 1 Nanertiao, Zhongguancun, Beijing 100190, China*
[9]*Koninklijke Sterrenwacht van België, Ringlaan 3, B-1180 Brussels, Belgium*

Accepted 2025 January 10. Received 2025 January 10; in original form 2024 June 7

ABSTRACT

We used a supervised machine learning algorithm (probabilistic random forest) to classify ∼130 million sources in the VISTA Survey of the Magellanic Clouds (VMC). We used multiwavelength photometry from optical to far-infrared as features to be trained on, and spectra of active galactic nuclei (AGNs), galaxies and a range of stellar classes including from new observations with the Southern African Large Telescope (SALT) and South African Astronomical Observatory (SAAO) 1.9-m telescope. We also retain a label for sources that remain unknown. This yielded average classifier accuracies of ∼79 per cent [Small Magellanic Cloud (SMC)] and ∼87 per cent [Large Magellanic Cloud (LMC)]. Restricting to the 56 696 719 sources with class probabilities ($P_{\rm class}$) > 80 per cent yields accuracies of ∼90 per cent (SMC) and ∼98 per cent (LMC). After removing sources classed as 'Unknown', we classify a total of 707 939 (SMC) and 397 899 (LMC) sources, including >77 600 extragalactic sources behind the Magellanic Clouds. The extragalactic sources are distributed evenly across the field, whereas the Magellanic sources concentrate at the centres of the Clouds, and both concentrate in optical/IR colour–colour/magnitude diagrams as expected. We also test these classifications using independent data sets, finding that, as expected, the majority of X-ray sources are classified as AGN (554/883) and the majority of radio sources are classed as AGN (1756/2694) or galaxies (659/2694), where the relative AGN–galaxy proportions vary substantially with radio flux density. We have found >49 500 hitherto unknown AGN candidates, likely including more AGN dust dominated sources which are in a critical phase of their evolution; >26 500 new galaxy candidates and >2800 new young stellar object (YSO) candidates.

Key words: methods: data analysis – galaxies: active – Magellanic Clouds – galaxies: photometry.

1 INTRODUCTION

As the depth and field of view of survey telescopes continues to improve, it becomes increasingly more unfeasible for astronomers to manually verify every individual source. Separating the extragalactic from the non-extragalactic is important for population studies and for the study of individual systems. Identifying galaxies, and whether they are hosting an active galactic nucleus (AGN) or not, allows us to study how galaxies evolve over cosmic time and what role AGN have to play in this process. Galaxies, especially those that host an AGN that have the potential to produce emission across the entire electromagnetic spectrum (e.g. Padovani et al. 2017), are more easily identified from combinations of multiwavelength photometric survey data.

Different wavelength regimes probe different parts of the AGN's structure (e.g. Netzer 2015; Padovani et al. 2017; Hickox & Alexander 2018). The UV/optical are sensitive to the accretion disc and the infrared (IR) is sensitive to the dusty obscuring material surrounding the accretion disc. The emission of a corona is observed in X-rays, whilst possible non-thermal radiation (which often take the form of jets/lobes) is picked up in the radio. One of the most reliable methods of identifying an AGN is with optical spectroscopy, which can reveal the broad and/or narrow emission lines of an AGN. However, spectroscopically observing every source would be a time consuming process, which is why we often turn to photometric surveys, which can observe large areas of the sky much more quickly.

★ E-mail: cpennock@ed.ac.uk

© 2025 The Author(s).
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Spectral energy distributions can be produced from photometric surveys, which for an AGN would exhibit a 'big blue bump', due to the accretion disc and another bump in the mid-infrared (mid-IR) due to re-processed emission from dust heated by accretion from the central supermassive black hole, which is a feature that has been well used to select large samples of AGN (e.g. Lacy et al. 2004; Stern et al. 2005; Secrest et al. 2015; Assef et al. 2018). However, factors such as obscuration from dust, emission from the host galaxy and the abundance of stars in our field of view can make AGN harder to select in one wavelength band, hence the need to make use of multiple wavelength regimes to ascertain a source's true nature.

The Magellanic Clouds are an often overlooked but – whilst challenging – eminently suitable and worthwhile location to search for the extragalactic sources behind them. The Clouds span $\sim$ 100 sq. degrees on the sky that have been covered as part of all-sky multiwaveband surveys such as the optical *Gaia* (Gaia Collaboration 2021a), the near-infrared (near-IR) Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) and the mid-IR Wide-field Infrared Survey Explorer (WISE; e.g. Cutri et al. 2013) surveys. Furthermore, there have been Magellanic Cloud specific surveys, where depth and angular resolution are improved compared to all-sky surveys. They include ultraviolet (UV) with the *Galaxy Evolution Explorer* (*GALEX*) and the UltraViolet Imaging Telescope (*UVIT*; e.g. Kumar et al. 2012; Thilker, Bianchi & Simons 2014) and the optical Survey of the Magellanic Stellar History (SMASH; Nidever et al. 2017) and the MAGellanic Inter-Cloud Project (MAGIC; Noël et al. 2013, 2015; Carrera et al. 2017), as well as surveys observed with the *Sptizer Space Telescope* in the mid-IR as part of the *Sptizer* Agents of Galaxy Evolution (SAGE) survey of the Large Magellanic Cloud (LMC; Meixner et al. 2006) and Small Magellanic Cloud (SMC; Gordon et al. 2011), and the *Herschel Space Observatory* in the far-IR as part of the *HERschel* Inventory of The Agents of Galaxy Evolution (HERITAGE; Meixner et al. 2010). Additionally, there have been many observations in the radio domain (e.g. MOST, ATCA; Mauch et al. 2003; Murphy et al. 2010) and X-ray (*XMM–Newton*; Sturm et al. 2013). These galaxies are also located away from the Galactic Plane and Galactic Centre, reducing source confusion in the radio band and extinction at UV/optical/near-IR wavelengths. The main caveat for searching in the field of the Magellanic Clouds is the increased stellar confusion. The new and deeper surveys towards the Magellanic Clouds such as the near-IR VISTA Survey of the Magellanic Clouds (VMC; Cioni et al. 2011) and radio Evolutionary Map of the Universe (EMU) all-sky (Joseph et al. 2019; Pennock et al. 2021) survey greatly enhance such attempts.

The VMC ESO public survey showcases a great improvement in depth and angular resolution compared to previous near-IR surveys, and has detected stars encompassing most phases of evolution such as main sequence stars, sub-giants, upper and lower red giant branch (RGB) stars, red clump stars, RR Lyræ and Cepheid variables, asymptotic giant branch (AGB) stars, post-AGB stars, young stellar objects (YSOs), planetary nebulæ (PNe), and supernova remnants (SNRs) populations (e.g. Gullieuszik et al. 2012; Ripepi et al. 2015; Zivkov et al. 2018; Groenewegen et al. 2019, 2020; Zivkov et al. 2020; Choudhury et al. 2021; Cusano et al. 2021) that can be used to help assess the age, metallicity, 3D structure, etc., within the Magellanic systems. This survey has also had success in discovering background extragalactic sources (Cioni et al. 2013; Ivanov et al. 2016; Bell et al. 2019, 2020, 2022; Pennock et al. 2022).

Machine learning algorithms are a use of artificial intelligence to automate tasks, such as identification and classification, on large sets of data that would otherwise prove time-consuming. They can also replace subjective approaches that depend on user choices by objective approaches that are data driven. Another advantage of machine learning is that they can combine information from multiple data sets, effectively classifying within a highly multidimensional parameter space. Machine learning algorithms are usually divided into two types, supervised and unsupervised. Unsupervised machine learning is predominantly used for clustering and dimensionality reduction, where objects with similar properties are grouped together in 2D space to find patterns/trends in the data.

Supervised machine learning algorithms predict classifications/values based on example data with features (e.g. photometry) and known classifications/values. They do this by analysing a known data set, the training set, and producing a model from this data set, which can then be used to make predictions of the output of an unseen data set. A disadvantage of supervised learning is that it is only as good as the data it is trained upon, and is therefore not best suited to finding new or unusual objects. Furthermore, imbalanced training sets (when the amount of objects for one or more of the classes dominates the training set) can lead to poor performance of the classifier, though this can be mitigated by artificially balancing the training sets (e.g. Kinson, Oliveira & van Loon 2021, 2022).

The VMC survey consists of $\sim$130 million sources, the identities of the majority of which remain unknown. In Pennock et al. (2022), unsupervised machine learning was used to cluster similar sources together in the radio-detected population of the VMC near-IR sources (Pennock et al. 2021), showing that machine learning can be a valuable tool in separating objects into different classifications, especially for separating dusty/evolved stars (such as YSOs, PNe, and post-AGB/RGB) that are often confused with AGN and vice versa in the optical and IR. This work is a continuation of these studies, where we use supervised machine learning with multiwavelength data from UV to far-IR to classify all of the sources in the VMC survey. In a follow-on paper, we will apply an unsupervised approach.

This paper is laid out as follows: Section 2 outlines the photometric surveys used in this work from which the features for the machine learning algorithm are selected, as well as the spectroscopic data sets from which the data with known labels is selected as a training set. Section 3 describes the machine learning algorithm used in this work, the choice of parameters and how it was trained. In Section 4, we explore the spatial distributions of the newly classified sources across the VMC fields of the Magellanic Clouds. Then in Section 5, we test the classifier against sources with known classes that were not used in training in Sections 5.1 and 5.2, as well as exploring their distributions of the high-confidence classifications across colour–colour/magnitude diagrams in the optical, near-IR and mid-IR regimes in Section 5.3. Furthermore, in Sections 5.4 and 5.5, we use the radio and X-ray-detected sources as an independent check to test the classifications, as the majority of X-ray and radio-detected sources are expected to be extragalactic, and explore these populations. We explore the classifications of *Gaia* low-resolution spectroscopically confirmed QSOs from Storey-Fisher et al. (2023) in Section 5.6 and the classifications of spectroscopically and photometrically selected YSOs in the LMC from Kokusho et al. (2023) in Section 5.7. Then, in Section 5.8, we explore the sources that have been confidently classed as Unknown by using the known photometric selection techniques. Lastly, we summarize our results in Section 6.

2 DATA

Machine learning algorithms require 'features', for each object, such as photometric measurements in various wavebands. Supervised machine learning requires an additional training set of known objects

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

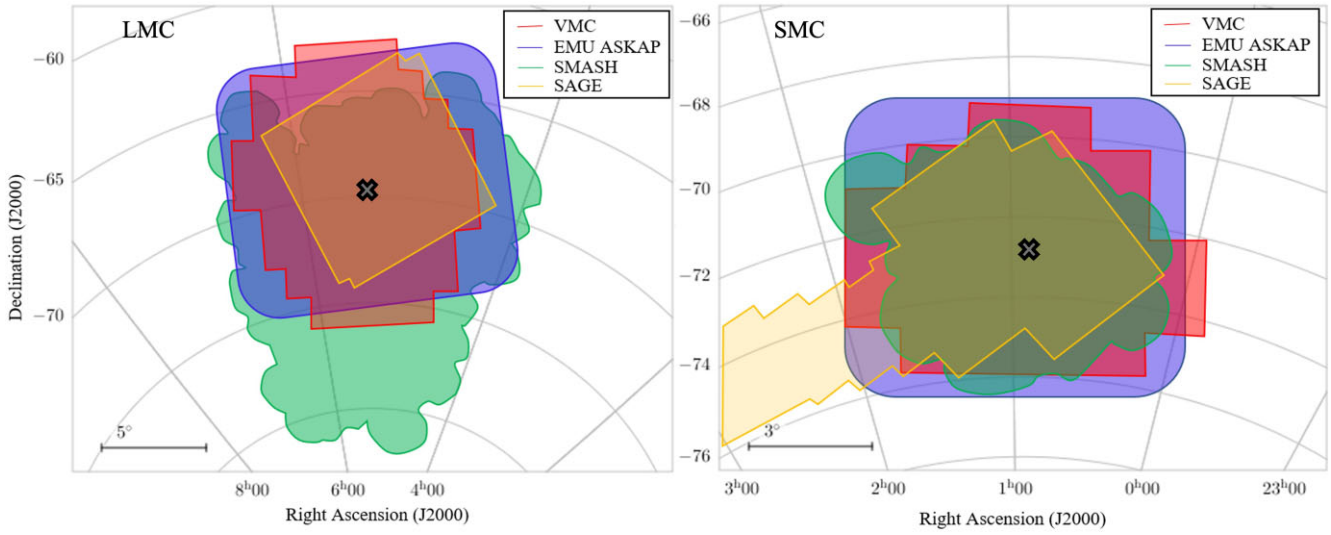

Figure 1. Area of the sky covered by the VMC (near-IR, *red*), EMU ASKAP (radio, *blue*), SMASH (optical, *green*), SAGE (mid-IR, *yellow*) surveys of the LMC (left) and SMC (right). The approximate centres of the Clouds are marked with a black 'X'.

with labels that can be used to learn from, in order to classify unseen data. Here, the photometry and the training sets are outlined.

2.1 Photometry

In this study we use photometry obtained from dedicated observations of the Magellanic Clouds, from the optical SMASH (Nidever et al. 2017), VMC (Cioni et al. 2011), *Spitzer* SAGE (Meixner et al. 2006; Gordon et al. 2011), *Herschel* HERITAGE (Meixner et al. 2010) and *XMM–Newton* (Sturm et al. 2013) X-ray and Australian Square Kilometre Array Pathfinder (ASKAP) radio (Joseph et al. 2019; Pennock et al. 2021) imaging surveys. Fig. 1 shows the comparisons between the area covered by the VMC, EMU ASKAP, SMASH, and SAGE surveys of the SMC and LMC. Below we describe each of the individual survey data sets used in this work. Details of how the UV/optical/IR surveys are combined to produce the multiwavelength catalogue for the machine learning process will be given in Section 3.1.

The X-ray and radio catalogues are not used in the machine learning, but as an independent check of the final source classifications, as radio/X-ray sources tend to be extragalactic in origin, as opposed to stars. Also, they are only a small fraction of the total sources so would introduce a lot of missing data if used as features.

2.1.1 *Optical SMASH survey*

The Dark Energy Camera (DECam; Schumacher et al. 2010) on NOAO's 4-m Blanco telescope was used as part of the SMASH (Nidever et al. 2017) to map 480 square degrees of sky to depths of *ugriz* $\sim$ 23.9, 24.8, 24.5, 24.2, 23.5 mag (Vega) at median seeing of 1.22, 1.13, 1.01, 0.95, 0.90 arcsec, respectively. The main goal of this survey was to identify broadly distributed, low surface brightness stellar populations associated with the stellar halos and tidal debris of the Magellanic Clouds. The catalogue contains $\sim$ 360 million objects in 197 fields. Note that Nidever et al. (2017) adjusted the DECam *ugriz* photometry to be comparable to SDSS and are therefore 'pseudo-SDSS', *ugriz* bands.

2.1.2 *Near-IR VISTA magellanic clouds survey*

The Visible and Infrared Survey Telescope for Astronomy (VISTA; Dalton et al. 2006) is a 4.1-m near-IR optimized telescope, which is equipped with the VISTA InfraRed CAMera (VIRCAM; Emerson, McPherson & Sutherland 2006) which is composed of a large array of 16 detectors that fill about a 1.5 square degree field.

The VMC survey (Cioni et al. 2011) is a near-IR deep, multi-epoch and wide-field study of the Magellanic Clouds, covering an area of about 170 deg^2. VISTA observations for the VMC main survey started in November 2009 and ended in October 2018. It has a spatial resolution of 1.0–1.1, 0.9–1.0, and 0.8–0.9 arcsec in the YJK_s filters, respectively, where the two values specified for seeing indicate maximum allowed seeing for crowded and uncrowded regions, respectively. It also reaches a sensitivity at 5σ level of about 22, 22 and 21.5 mag (Vega; in AB this is 22.5, 22.9, and 23.4 mag) in the YJK_s bands, respectively. Its depth and coverage can be compared to the VISTA Deep Extragalactic Observations (VIDEO; Jarvis et al. 2013) survey, which was specifically designed to study galaxy and cluster/structure evolution up to $z \sim 4$ in a 12 deg^2 area, reaching depths of about 24.5, 24.4, and 23.8 mag (AB) at 5σ detection level in the YJK_s bands, respectively. The VMC data provide an opportunity to expand on the effort of the VIDEO survey and cover more area to better overcome cosmic variance, and has already proven successful in discovering more AGN (e.g. Ivanov et al. 2016). This however comes with the caveat of increased stellar confusion with the presence of the LMC and SMC.

The catalogues created from the VMC survey provide both aperture and PSF photometry, where PSF photometry reaches sources on average 0.3 magnitudes fainter than aperture photometry. The PSF catalogue is created as described in Rubele et al. (2015) and are publicly available as part of VMC DR6.[1,2] The magnitudes in each band are calculated from deep tile images, which are a combination of single exposure images from different epochs. Due to the PSF photometry's increased depth and ability to distinguish sources in crowded regions, it is the PSF photometry that is used in this work.

[1] http://archive.eso.org
[2] http://vsa.roe.ac.uk

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

2.1.3 Infrared SAGE and HERITAGE surveys

The Magellanic Clouds were observed by *Spitzer* as part of the SAGE survey of the LMC (Meixner et al. 2006) and SMC (Gordon et al. 2011) which map 49 and 30 deg^2, respectively. It produced a list of about 8.4 million sources taken with IRAC filters 3.6, 4.5, 5.8, 8.0 μm with an angular resolution of 2 arcsec. The faint limits for SAGE are 18.3, 17.7, 15.7, and 14.5 mag, respectively.

The *Herschel* Space Observatory (Pilbratt et al. 2010) was a 3.5-m IR telescope that was active from 2009 to 2013 and was sensitive to the far-IR and submillimetre wavebands (55–672 μm). *HERschel* Inventory of The Agents of Galaxy Evolution (HERITAGE; Meixner et al. 2010) used the *Herschel*'s Photodetector Array Camera and Spectrometer (PACS, 100 and 160 μm; Poglitsch et al. 2010) and the Spectral and Photometric Imaging REceiver (SPIRE, 250, 350, 500 μm; Griffin et al. 2010) bands to image the LMC, SMC, and Magellanic Bridge. This survey is complementary to the SAGE survey.

2.1.4 All-sky surveys

Various all-sky surveys have also observed the Magellanic Clouds. However, this comes with the caveat of not reaching the same depths as the Magellanic specific surveys. All-sky surveys used in this work include optical *Gaia* (Gaia Collaboration 2023) and mid-IR AllWISE and unWISE (Wright et al. 2010; Cutri et al. 2013; Schlafly, Meisner & Green 2019) surveys.

The *Gaia* mission (Gaia Collaboration 2016) was launched on 19 December 2013, with the aim of measuring the 3D spatial and velocity distribution of stars, as well as determine their astrophysical properties. The *Gaia* on-board system is designed to detect point-like sources, but can detect extragalactic sources (Gaia Collaboration 2023) if their central region is sufficiently bright and compact. The latest data release, DR3 (Gaia Collaboration 2023), is based on 34 months of *Gaia* operations. The catalogue provides celestial positions, proper motions, parallaxes, and broad band photometry in the wide G (centred on 650 nm), blue-enhanced G_{BP} (centred on 360 nm), and red-enhanced G_{RP} (centred on 750 nm) pass-bands. This data release also includes class probabilities (QSO, galaxy, or stellar source) for 1.5 billion sources.

WISE (Wright et al. 2010) is a telescope launched in 2009 to repeatedly map the entire sky in IR. WISE mapped the whole sky in four bands *W*1, *W*2, *W*3, *W*4 centred at 3.4, 4.6, 12, and 22 μm, respectively, using a 40-cm telescope feeding arrays with a total of four million pixels. The sensitivities of *W*1, *W*2, *W*3, and *W*4 correspond to Vega magnitudes of 16.5, 15.5, 11.2, and 7.9, respectively, in the all-sky WISE survey. The AllWISE (Cutri et al. 2013) programme extended the work of the WISE survey mission by combining *W*1 and *W*2 data from the cryogenic and post-cryogenic survey phases to form the most comprehensive view of the mid-IR sky currently available. *W*3 and *W*4 measurements remain unchanged from the All-Sky Release because no additional data were included in those bands.

Further addition to the WISE mission, is the unWISE (Schlafly et al. 2019) catalogue, which used the deep unWISE coadded images built from five years of publicly available WISE imaging, as well as improved modelling of crowded regions. This resulted in a catalogue of ∼2 billion unique objects detected in the *W*1 and/or *W*2 channels, reaching depths ∼0.7 mag fainter than those achieved by AllWISE.

2.1.5 X-ray XMM–Newton

An SMC-survey point-source catalogue was created from archival *XMM–Newton* data with additional newer observations from the same facility (Sturm et al. 2013), which covers 5.6 deg^2, including the bar and eastern wing of the SMC. The catalogue contains 3053 unique X-ray sources with a median position uncertainty of 1.3 arcsec down to a flux limit of $\sim 10^{-14}$ erg cm^{-2} s^{-1}. The majority of the sources are expected to be AGN. One limitation of this survey is that it only covers the central part of the SMC, and therefore does not cover the same breadth as the VMC survey.

There is no similar X-ray catalogue specifically for the LMC. There is, however, an *XMM–Newton* serendipitous source catalogue (Webb et al. 2020), which is a collection of the sources detected in all the publicly available *XMM–Newton* observations. This catalogue includes observations in the direction of the LMC.

2.1.6 Radio ASKAP survey

The EMU (Norris et al. 2011) is a wide-field radio continuum survey which uses the ASKAP (Johnston et al. 2008; Hotan et al. 2021) telescope. EMU's primary goal is to make a deep (RMS ∼ 10 μJy beam^{-1}) radio continuum survey of the Southern sky, extending as far north as +30° declination, with a resolution of 10 arcsec. It is expected to catalogue about 70 million galaxies, including AGN up to the edge of the visible Universe.

Two radio continuum images from the ASKAP survey in the direction of the SMC were taken as part of the EMU Early Science Project (ESP) survey of the Magellanic Clouds (Joseph et al. 2019). The two source lists that were produced from these images by Joseph et al. (2019) contain radio continuum sources observed at 960 MHz (4489 sources) and 1320 MHz (5954 sources) with a bandwidth of 192 MHz and beam sizes of 30 arcsec × 30 arcsec and 16.3 arcsec × 15.1 arcsec, respectively. The median RMS noise values were 186 μJy beam^{-1} (960 MHz) and 165 μJy beam^{-1} (1320 MHz). The observations of the SMC were made with only 33 per cent and 44 per cent (for 960 and 1320 MHz, respectively) of the full ASKAP antenna configuration and 66 per cent of the final bandwidth that was available in the final array, with which the LMC was observed, so the resolution and depth is not as good as for the LMC observation.

The LMC was observed at 888 MHz (Pennock et al. 2021) with a bandwidth of 288 MHz taken on 2019 April 20 using ASKAP's full array of 36 antennas (scheduling block 8532). The LMC was observed as part of the ASKAP commissioning and early science (ACES, project code AS033) verification (DeBoer et al. 2009; Hotan et al. 2014; McConnell et al. 2016) in order to investigate issues that were found in higher-frequency higher-spectral-resolution Galactic-ASKAP (GASKAP; Dickey et al. 2013) survey observations, as well as to test the rapid processing with ASKAPsoft (Whiting 2020). The observations cover a total field of view of 120 deg^2, with a total exposure time of $\sim 12^h40^m$. They were compiled by four pointings ($\sim 3^h10^m$ each) with three interleaves,[3] each to result in more uniform depth across the field – effectively 12 pointings. The three interleaves overlap by ∼ 0.5° to improve the uniformity of sensitivity across the field. The largest angular scales that can be recovered in this survey are 25–50 arcmin (McConnell et al. 2020).

2.2 Training sets for machine learning

A set of known sources is required to train a supervised machine learning classifier. We focused on using sources with spectroscopic observations. The total number of sources for each class can be seen

[3]interleaves are overlapping pointings where the telescope slews between them at a more rapid cadence.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

in Table 1. The training sets for the SMC and LMC are made available alongside this paper as online supplementary material.

We chose to focus on ten classes: AGN; galaxies; stars of O and B type (OB); RGB stars; AGB; red supergiants (RSG), post-AGB and post-RGB stars (pAGB/RGB); PNe, YSOs and compact H II regions (H II/YSOs) and Milky-Way high proper-motion stars (PM). These classes were chosen for having larger numbers of sources with classifications based on spectroscopy and/or due to their tendency to be mistaken for AGN (dusty and/or emission-line sources) and vice versa.

Using spectroscopically observed sources, however, introduces bias into the training sample, since the sources observed tend to be chosen based on colour cuts that similar previously observed objects conform to. Furthermore, there is also a bias in magnitude, as the faintest sources would be too faint for spectroscopy. This therefore leaves the rarer/unusual and fainter versions of each class to not be observed, which would make the machine learning algorithm less certain about these sources.

A potential problem could be that classes that have little to no spectroscopic observations, and therefore not trained upon, could be misclassified as one of the classes trained upon if they are similar enough. A class that encompasses these sources that are not part of the known classes could be needed to prevent confusion.

2.2.1 South African Astronomical Observatory 1.9 m

We observed 174 new optical spectra (see online Appendix Section A for full list) at the South African Astronomical Observatory (SAAO) 1.9-m telescope with SpUpNIC (Spectrograph Upgrade: Newly Improved Cassegrain; Crause et al. 2019) during observing runs in 2019 and 2021. Grating 7 (grating angle of 16°) and the order blocking 'BG38' filter were used, delivering a resolving power $R = \frac{\lambda}{\Delta\lambda} \sim 500$ over a wavelength range of 3800 Å–9000 Å. Dome-flats and bias images were taken at the beginning of each night. The CuAr lamp was used for wavelength calibration. Three 600 s (300 s for sources brighter than $\sim$ 16 mag) exposures were obtained for each source. The standard stars (EG 21, Feige 110 or LTT 1020; Hamuy et al. 1994) were observed on the same night under the same conditions for 30 s.

The data was processed using the standard IRAF[4] tools (Tody 1986, 1993).

The sources that we observed with the 1.9m telescope, and have been classified based on their optical spectroscopy, were added to the training set. This added 26 sources (18 AGN, 7 galaxies, 1 HII/YSO).

2.2.2 SALT

We also observed 40 sources with the Southern African Large Telescope (SALT) (Buckley, Swart & Meiring 2006), located in Sutherland, South Africa that has an effective diameter of 7–9 m. SALT was used to observe AGN candidates that had the potential to be similar to SAGE0536AGN (Pennock et al. 2022; Pennock et al., in preparation). The Robert Stobie Spectrograph (RSS; Burgh et al. 2003; Kobulnicky et al. 2003) was used, a combination of three CCD detectors with total 3172 × 2052 pixels and spatial resolution of 0.1267 arcsec per pixel. We used the long-slit with width 1.5 or 1.25 arcsec, grating PG0300 or PG0900 and an Argon arc lamp. Initial processing (basic CCD data reductions) was done automatically by the SALT pipeline (Crawford et al. 2010). We processed these data by performing cosmic ray removal, wavelength calibration and source extraction also using also using the standard IRAF tools (Tody 1986, 1993).

The sources we observed with SALT that have been classified were added to the training set. This numbered 22 sources: 1 AGB, 1 H II/YSO, 1 galaxy, and 19 AGN.

[4]IRAFis distributed by the National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy, Inc., under cooperative agreement with the National Science Foundation.

Table 1. The number of sources for each class and the region of the Clouds they were spectroscopically observed in, as well as the references of the literature they originated from. (1) From own observations using SALT or SAAO's 1.9-m telescope; (2) *Spitzer*-spec surveys (Ruffle et al. 2015; Jones et al. 2017); and (3) From a Simbad (Wenger et al. 2000) search of 'PM' stars in the VMC footprint that had listed spectral type and reference given.

Class	SMC	LMC	References
AGN	303	639	(1); (2); Kozłowski et al. (2012); Flesch (2019); Geha et al. (2003); Kozłowski et al. (2013); Esquej et al. (2013); Ivanov et al. (2016); Ivanov et al. (2024)
Galaxies	124	430	(1); (2); Jones et al. (2009)
OB	417	1073	(2); Walborn et al. (2014); Evans et al. (2015a,b); Lamb et al. (2016); Grin et al. (2017); Roman-Duval et al. (2019); Dorigo Jones et al. (2020)
RGB	519	489	(2); Cole et al. (2005); Neugent et al. (2020) Parisi et al. (2009, 2010, 2022); De Bortoli et al. (2022)
H II/YSOs	86	459	(2); Seale et al. (2009); Oliveira et al. (2011, 2013); Oliveira et al. (2019); van Gelder et al. (2020)
PNe	53	50	(2); Shaw et al. (2001)
AGB	165	221	(2); van Loon et al. (1998); Groenewegen & Blommaert (1998); van Loon, Zijlstra & Groenewegen (1999a); van Loon et al. (1999b, 2005, 2006, 2008); Kamath, Wood & Van Winckel (2014)
RSG	44	70	(2); Neugent et al. (2020)
pAGB/RGB	46	33	(2); van Loon et al. (2008); Kamath et al. (2014)
PM	78	303	(3)

2.2.3 SAGE-spec

The Infrared Spectrograph onboard the *Spitzer* Space Telescope was used to observe the LMC and SMC in low and high resolution modes for the wavelength range of 5–38 μm. The resolving power varies between 60 and 130 for low resolution mode, whereas high-resolution mode has resolving power of $\sim$ 600.

All the spectra obtained by *Spitzer* within the SAGE footprint were looked at as part of the SAGE-Spec project. In the SMC (Ruffle et al. 2015), this survey found 58 AGB stars, 51 YSOs, 4 post-AGB objects, 22 RSGs, 27 undefined stars (of which 23 are dusty OB stars), 24 PNe, 10 Wolf-Rayet (WR) stars, 3 H II regions, 3 R Coronæ Borealis (R CrB) stars, 1 blue supergiant and six other objects.

In the LMC (Jones et al. 2017), this survey observed $\sim$ 800 sources, the majority of which are YSO and H II regions and (post-)AGB stars, PNe and massive stars. Also observed were two SNRs, a nova and several background galaxies.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

2.2.4 *Extragalactic classes*

There are 657 spectroscopically observed AGN from the Milliquas catalogue of Flesch (2015, b) in the field of the VMC footprint of the LMC. The largest contributions are from Kozłowski et al. (2012, 2013), Geha et al. (2003), Esquej et al. (2013) and Ivanov et al. (2016), contributing 547, 24, 23, and 10 objects, respectively.

There are 240 spectroscopically observed AGN from the Milliquas catalogue in the field of the VMC footprint of the SMC. The largest contributions are from Kozłowski, Kochanek & Udalski (2011), Kozłowski et al. (2013), and Ivanov et al. (2016), contributing 194 and 10 objects, respectively.

Galaxies (with no signs that they are hosting AGN) were taken from the 6dFGS survey (Jones et al. 2009). The observations for this survey were carried out using the Six Degree Field (6dF) fibre-fed multi-object spectrograph at the UK Schmidt Telescope (UKST; Siding Spring Observatory, Australia) over 2001 May to 2006 January (Jones et al. 2005). Target fields covered $\sim 17\,000$ deg^2 of the southern sky more than 10° from the Galactic Plane. This survey data however comes with the caveat that it is limited to the brightest (it is complete to total extrapolated 2MASS magnitude limits of 13.75, 12.95, 12.65 mag for J, H, and K, respectively) and closest (median redshift of whole survey is $z \sim 0.053$) of galaxies.

To further augment the extragalactic sample, galaxies and AGN were added from the Galaxy and Mass Assembly (GAMA Driver et al. 2011; Hopkins et al. 2013) survey, specifically the GAMA09 region, which was also observed with VISTA as part of the VIKING survey (de Jong et al. 2017)[5] The SDSS DR16 survey (Blanton et al. 2017; Ahumada et al. 2020) covered this region, and spectroscopically observed extragalactic sources and further separated them into galaxies and AGN, adding 8504 and 2337 sources, respectively. For these sources *Sptizer* IRAC band and *Herschel* 250 *um* were left as missing data.

2.2.5 *Galactic and magellanic classes*

For the training sample, sources that are often mistaken for AGN were needed, whilst also including other sources that are more distinct from AGN that are prevalent throughout the Magellanic Clouds. There have been many spectroscopic surveys of the LMC and SMC, generally looking at specific types of stellar objects that can be found within the Magellanic Clouds. This led to the accumulation of 1490 OB stars, 1008 RGB stars, 545 YSO or compact H II regions, 386 AGB stars, 114 RSGs, 103 PNe, 79 pAGB/pRGB and 382 high proper-motion/foreground stars (see Table 1 for details). A large part of these sources were from the SAGEspec surveys (Ruffle et al. 2015; Jones et al. 2017), which observed 209 and 862 sources in the direction of the SMC and LMC, respectively.

The more common stars (e.g. main-sequence M-type stars 2700–3800K) however tend not to be spectroscopically observed as they are an already well observed/studied class of objects (in the Milky Way), whereas spectroscopic studies preferentially target (and confirm the identities of) the rarer less well-studied stellar classes, therefore leading to a lack of main-sequence stars for training. They are however distinct from AGN, with a lack of emission in the IR, so should not be mistaken for extragalactic sources, and be preferentially associated with the stellar sources based on proper motions and colours.

The sources with high-proper motion and that are in the foreground were identified using a Simbad (Wenger et al. 2000) search for 'PM' stars in the VMC footprint that had a listed spectral type and reference given. This yielded 78 and 303 sources for the SMC and LMC, respectively, with a cross-match with a source in the VMC catalogue with a $1''$ search radius.

2.2.6 *Unknown class*

Not all classes can be accounted for, as some classes have too few spectroscopically observed sources to create a robust training sample from, and others are simply unknown. However, all the sources must be classified as one of the classes it has been trained upon, which would inevitably lead to contamination within each class. Therefore, to ensure that these sources are not classified incorrectly, an Unknown class is created for the LMC and SMC. This was done by randomly selecting a number of sources (same amount as the largest training set, galaxies, at 9118) from the VMC catalogues for both the LMC and SMC. This creates a sample of sources that have no structure in feature space as it includes a mix of everything. This should allow the clearly defined classes in feature space (collection of features, in this case photometry, that are used to characterise the different classes) to be classed correctly whilst setting the sources that lay away from the known classes to be set as Unknown. Note that none of the spectroscopically observed sources are in this class.

We found that creating this class was necessary to ensure that faint and/or difficult to classify sources were not (erroneously) allocated to one of the other classes by the machine learning algorithm (i.e. that this class is needed to fully capture our remaining uncertainty/ignorance).

3 PROBABILISTIC RANDOM FOREST

A random forest (Breiman 2001) is a supervised machine learning algorithm that can be used for both classification and regression problems. The algorithm builds several decision trees (a decision tree consists of a series of nodes where at each node a condition is given that is either true or false) independently and then averages the predictions of these to obtain the final prediction, as well as the probability the prediction is correct from the fraction of trees that agree with the final prediction. This reduces variance over using a single estimator and creates an overall more stable model. It is called a random forest because randomness is injected into the training process of each individual tree via a method called 'bagging'. This method splits up the training set into randomly selected subsets, and each decision tree is then trained on one of those subsets. Furthermore, at each node of the decision tree, only a randomly selected subset of the features is considered.

The Probabilistic Random Forest[6] (PRF; Reis, Baron & Shahaf 2018) is a random forest algorithm that can handle and take into account measurement uncertainties and missing data (where data is missing that is required for a condition at a node, the PRF will propagate down both true and false paths with equal weight given to the probability that either path is correct). Compared to an ordinary random forest it has been proven to provide an up to 10 per cent increase in classification accuracy with noisy features and proven to be more accurate than the original random forest when up to 45 per cent objects in the training set are misclassified.

This algorithm was created in Python and requires the Python module SCIKIT-LEARN (Pedregosa et al. 2011) to run.

[5]Based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme ID 179.A-2004.

[6]PYTHON code can be found here: https://github.com/ireis/PRF

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Table 2. Parameters taken from various surveys to act as features in the PRF algorithm. For each parameter the error on the values is also taken from the corresponding surveys. New features were created by subtracting each feature from all the other features to create colours. This did not include the sharpness (a measure of the difference between the observed width of the object and the width of the PSF model, where stars should have a sharpness value of ∼0 and resolved objects values of >0. Sharpness values < 0 indicate artefacts such as bad pixels or cosmic ray impacts) and proper motions in RA and DEC (pmRA and pmDEC, respectively).

Parameter	Units	Survey
Y PSF	mags (Vega)	VMC
J PSF	mags (Vega)	VMC
K_s PSF	mags (Vega)	VMC
Y sharp PSF	–	VMC
J sharp PSF	–	VMC
K_s sharp PSF	–	VMC
u	mags (Vega)	SMASH
g	mags (Vega)	SMASH
r	mags (Vega)	SMASH
i	mags (Vega)	SMASH
z	mags (Vega)	SMASH
sharp	–	SMASH
pmRA	mas/yr	Gaia DR3
pmDEC	mas/yr	Gaia DR3
G	mags (Vega)	Gaia DR3
G_{BP}	mags (Vega)	Gaia DR3
G_{RP}	mags (Vega)	Gaia DR3
IRAC 3.6 μm	mags (Vega)	SAGE
IRAC 4.5 μm	mags (Vega)	SAGE
IRAC 5.8 μm	mags (Vega)	SAGE
IRAC 8.0 μm	mags (Vega)	SAGE
unW1	mags (Vega)	unWISE
unW2	mags (Vega)	unWISE
$W1$	mags (Vega)	AllWISE
$W2$	mags (Vega)	AllWISE
$W3$	mags (Vega)	AllWISE
$W4$	mags (Vega)	AllWISE
SPIRE PSW 250 μm	MJy	HERITAGE

3.1 Creating the multiwavelength data set

The base of the multiwavelength data set is the near-IR VMC PSF survey catalogue. All coordinate matchings were made to the VMC coordinates. We matched the VMC catalogue with SMASH (Nidever et al. 2017), *Gaia* DR3 (Gaia Collaboration 2023), SAGE (Meixner et al. 2006; Gordon et al. 2011), unWISE (Schlafly et al. 2019) and AllWISE (Cutri et al. 2013) using TOPCAT (Taylor 2005). The parameters of different surveys in the data set can be seen in Table 2. The cross-matchings between all catalogues were done with a 1″ search radius. Note that the parameters from X-ray and radio surveys have not been used as features, and will be used as an independent check (e.g. they should favour extragalactic sources) for the classifications.

Some of the parameters required calculation. For instance, the unWISE catalogue only provided fluxes rather than Vega magnitudes like the rest of the catalogues. For consistency the fluxes were converted using the method recommended in the notes of the table on CDS[7] (Centre de Données astronomiques). The fluxes in Vega nanomaggies (nMgy; Finkbeiner et al. 2004) were converted to Vega magnitudes using $m = 22.52.5 \log(\text{flux})$. These fluxes showed slight discrepancies with the AllWISE values and a correction was applied (see Schlafly et al. 2019) of subtracting 0.004 mag and 0.032 mag from unWISE $W1$ and unWISE $W2$, respectively. The differences in the values of $W1$ and $W2$ bands between the AllWISE and unWISE catalogues is, as expected, centred around 0. Differences beyond this could be explained by variability. Where there are no differences, this should not affect the classifications as no new information is being provided, so using the feature again, once from AllWISE and then once from unWISE, would make no difference in splitting up the data.

Other parameters that had to be calculated were colours between all photometry bands and their corresponding errors, which were calculated with standard propagation of errors. Note that if, for example, $Y - J$ was calculated, the reverse, $J - Y$, would not be calculated and added as a feature. This led to a total of 237 features.

The far-IR measurement is taken from the SPIRE PSW 250 μm images of the SMC and LMC that were taken with *Herschel*, that are first smoothed with a 2D box kernel across ten pixels. After smoothing the image, the flux values were taken from the image at each co-ordinate, where the median within a 5 arcsec radius was taken. This provides a measurement of the background flux instead of the individual source flux, where the flux is expected to be higher when looking through the Magellanic Clouds.

Parameters which were not observed for a source in a given survey were assigned null values within their respective survey data bases, which differs between surveys. For example, the VMC applies the large negative value of -9.99999×10^8, whereas SMASH represents a null value as 99. It is also commonplace to leave null parameters unassigned or to assign a value of 0. Therefore, these data must be homogenized prior to the implementation of the PRF. To achieve this, we assigned the standard null value, 'NaN', to all null values across our input data, regardless of their origin.

3.2 Extinction

The effect of extinction from the Magellanic Clouds (e.g. Bell et al. 2019, 2020, 2022) is a non-explicit factor that is included in the photometry/colours that help to separate the stellar from the extragalactic. Galactic extinction from the Milky Way (e.g. Schlegel, Finkbeiner & Davis 1998; Schlafly & Finkbeiner 2011), though minor in this case compared to the Clouds, is also a factor that needs to be taken into consideration.

However, it is not straightforward to correct for LMC/SMC extinction, because it is unknown in advance which objects are extragalactic and which belong to the LMC or SMC. The idea of this classifier is that it can take the raw photometric data, with examples of sources from across the Clouds affected by different amounts of extinction, and learn from this. This is why it is important to have spectroscopically classified extragalactic sources right across the Clouds, so that the PRF is appropriately trained to recognise such populations even in cases of substantial foreground reddening.

As shown in Section 3.3.6 and Fig. 4, the far-IR HERITAGE SPIRE PSW 250 μm average flux density at each source position is shown to be the most important feature for the classifier. This band traces the emission from cold dust across the Clouds (responsible for the extinction of other bands) and provides our classifier with information related to the reddening that is likely being used to aid the classification of different source classes, including extragalactic sources that lie behind the central regions of the Clouds. Furthermore, a higher average flux density would tend to be found in areas of high star-formation (found in the centre of the Magellanic Clouds), therefore the far-IR feature would most likely bias the classifier

[7] https://cds.unistra.fr

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

positively towards young stellar populations, and negatively bias against background galaxies and AGN.

3.3 Training

In this section, we discuss the various aspects of the training of the classifiers, including the configuration of the input training set and the PRF parameter choices.

3.3.1 Inputs

The data were arranged in three configurations, individual 'LMC' and 'SMC' data sets, as well as a further 'MC' (LMC and SMC) data set. Each configuration had two versions, one with no colour features and another with colour features. Individual classifiers are trained for the SMC and LMC due the different stellar populations, population histories, extinction distributions and metallicities (e.g. Rubele et al. 2012; Rezaeikh et al. 2014; Rubele et al. 2018; Bell et al. 2019, 2020, 2022) between the two Clouds. Therefore, the Magellanic classes were specifically trained for their respective galaxy, but the same extragalactic and foreground stars training sets were used for training both classifiers.

For training the classifiers, data sets were split into features, X, errors on features, dX, and class, y.

To ascertain the accuracies of the trained classifiers each data set was split into training and testing sets, where 75 per cent of the data were trained on and 25 per cent were retained to test the classifier on. For each of the training runs the data split was randomized. Note that when testing a machine learning model a training data set is often split into training, validation and test set. A validation set is used to tune the parameters of the model, whilst a test set is used to test the final model. Both these data sets are not trained upon. In the interest of not splitting the different sets into too small groupings, and therefore not providing a good overview of how well the trained classifier works, we combined the validation and test sets together into an overall 'test' set. The configuration of the training and test sets were then randomized over multiple runs of tuning and testing the classifier, so that the classifier is not overfitting to one training set.

3.3.2 Probability threshold parameter

The probabilistic random forest classifier has parameters that can be varied. Most parameters are set to default.

The probability threshold parameter, $p_{\rm th}$, determines the probability threshold at which to stop propagating along a branch. In an ideal PRF, $p_{\rm th} = 0$, where all objects propagate along all branches to all terminal nodes, unlike $p_{\rm th} = 1$ which denotes a classical RF, where each object propagates to only one terminal node. The former requires a higher amount of computation time, and for any given object there may be nodes with small propagation probability. Stopping the propagation at these nodes reduces the run time without decreasing overall performance. Reis et al. (2018) found that reducing the probability threshold below a value of $p_{\rm th} = 0.05$ does not significantly improve the prediction accuracy, and only increases computation time, so we used this value in our work.

3.3.3 Number of trees parameter

To determine the optimum number of trees for the PRF, for each of the data sets, the data set was split into training and testing and then the classifier was trained on this at $n_{\rm trees} = 1, 5, 10, 25, 50, 100, 200, 500$, and then the score (fraction of correct classifications when the classifier is used on the test set) was calculated. It should be noted that since the extragalactic sources dominate the test set, the score therefore is dominated by the extragalactic accuracy, and the score is not representative of all classes. Next, the whole data set was split randomly again and then the classifier was trained again on the different number of trees. This is done for five iterations for each $n_{\rm trees}$ and then the score is averaged for each value of $n_{\rm trees}$.

From this it was seen that for all data set configurations after $n_{\rm trees} > 100$ the score for each classifier plateaus. Therefore the value of the number of trees was set to $n_{\rm trees} = 100$. This was done for training sets with and without colour information and it was found that the overall accuracy/score is greater when colour features are included. It was also found that the classifier trained and tested on the 'MC' training data set was found to have a lower accuracy than the classifiers trained and tested on the individual 'LMC' and 'SMC' training data sets.

3.3.4 Balanced versus imbalanced data sets

A balanced training set would have all classes roughly equal in size. An imbalanced training set would have large differences between class sample sizes. This imbalance can cause a poor predictive performance for the minority classes (e.g. Khoshgoftaar, Golawala & Hulse 2007; More & Rana 2017), as most machine learning algorithms operate under the assumption of an equal sample size for each class.

Ensemble methods such as random forests can mitigate the effects of imbalanced data sets by training each tree on an independently randomly selected subset of the training set and then combining the results of all the trees together. However, for extremely imbalanced data sets, when randomly selecting a subset to train a tree on, if a minority class is too small then only a few or even none at all of the minority class may be selected for a particular tree, meaning there will be trees that have not seen the minority class at all, so will not know how to classify them. This effect can be mitigated by balancing the data set.

Balancing the data set can either be done by downsampling, which reduces all the class sizes to the smallest class size, or upsampling, which increases/augments the minority class with synthetic data so that all the class sizes are the same as the largest class size. Downsampling works well if spread in parameter space is preserved (such as in Kinson et al. 2021, 2022), however, comes with the caveat of potentially losing important information if the training sets are heavily imbalanced and therefore will not be used here. Upsampling maintains the same amount of information (though with the possible caveat of overfitting due to replication of non-relevant features) so this strategy was adopted. See online Appendix Section B for the comparison between not balancing, downsampling and upsampling the training set effects the classifier's precision and recall.

Upsampling can be done in one of two ways. Either by using machine learning on the minority class to generate synthetic data points based on the real data of the minority class sample; or by randomly copying objects from the minority class sample to increase the sample size. The latter method was used as it maintains that only real data is used whilst balancing the data set so that each tree will randomly sample sources from each class. To do this the 'resample' function of PYTHON's SCIKIT-LEARN module was used to upsample all the class samples to the same size as the majority class, so that all classes have an equally sized training set.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

The SMC and LMC training samples were upsampled to the size of the galaxy/Unknown class. This was only done after the training sample was split into training and test sets, and only on the training sets. This process was so that the same objects did not end up in both the training and test set. This was trained and tested three times and the results averaged. From this we found that the addition of the upsampling overall increased the number of confident ($P_{\rm class}$ > 80 per cent) correct classifications, especially for those with the smaller training sets.

Note that an imbalance in the data set can also be caused by bias within the classes themselves. In this case, it would be selection bias of the sources being bright enough in the optical to be spectroscopically observed. We do not address this issue in this paper as it is beyond the scope of our work and is due to limiting to the spectroscopic data that are already in hand. It is, however, somewhat mitigated by the inclusion of the Unknown class, which was found to decrease the number of confident wrong predictions. One way this could be improved upon is by finding suitable, brighter targets in the Milky Way for some classes, for example, RGB stars of SMC metallicity in the globular cluster 47 Tucanae. Another is by spectroscopically observing more sources that are not necessarily bright enough in the optical, but perhaps instead in the IR, with a telescope such as the recent *JWST* (Gardner et al. 2006).

3.3.5 Final data configuration

The overall accuracy does not tell one how the classifier performs on individual classes, this can be shown through the use of confusion matrices. A confusion matrix shows the comparison between the true labels (y-axis) versus the predicted labels (x-axis). A perfect classifier would show a value of 1 (100 per cent) in a diagonal line from top left to bottom right of the confusion matrix, which represents the recall (the ratio of $\frac{tp}{tp+fn}$, where tp is the number of true positives and fn is the number of false negatives) of each class, whilst all the other values would be 0 (0 per cent). This would show that all classes have been predicted correctly. The values for an entire row should sum to 1, showing the distribution of class predictions for each class.

The final data set configuration to be trained upon was for separate classifiers for the LMC and SMC, where both data sets will share extragalactic sources from both regions, whilst keeping stellar sources specific to the Cloud they are from. The PRF has 100 trees. As the PRF randomly selects a subset of sources from the training set to train from for each decision tree, each run of the classifier creates different trees. How it randomizes the selection can be locked in by setting a 'random seed state'. The classifier was trained and tested on the data set ten times, using ten different random seed states, to create ten confusion matrices. The values were then averaged to create an 'average' confusion matrix. The confusion matrix for the SMC-trained classifier tested on SMC data can be found in Fig. 2 (right panels), and the confusion matrix for the LMC-trained classifier tested on LMC data can be found in Fig. 2 (left panels).

The overall accuracy (score) of the SMC classifier is found to be 0.79 ± 0.01, and the overall accuracy of the LMC classifier is found to be 0.87 ± 0.01. For both classifiers the AGN class has one of the highest recall, $\sim$ 90 per cent of all AGN in the test set are classified correctly for the SMC and LMC. For the AGN misclassified as other sources, they are most often misclassified as galaxies, which is not unexpected. The precision (the ratio of $\frac{tp}{tp+fp}$, where fp is the number of false positives) of the AGN class is not as great, as other sources are misclassified as AGN. For the LMC sightlines $\sim$3 per cent of galaxies and $\sim$9 per cent of PNe are misclassified as AGN. However, for the SMC sightlines, $\sim$6 per cent of PNe, $\sim$8 per cent of H II/YSOs, $\sim$ 5 per cent of pAGB/pRGB, $\sim$4 per cent of galaxies, $\sim$2 per cent of AGB and $\sim$1 per cent of RGB and OB are misclassified as AGN. Overall, most AGN will be classified correctly, with some expected confusion between AGN and galaxies, which is not unexpected as AGNs are hosted in galaxies with varying levels of obscuration and luminosity of the AGN emission, making it hard to discern a heavily obscured or low luminosity AGN from a galaxy with no AGN. There will, however, be some stellar interlopers, most often PNe. Though, when limited to only the high confidence sources (>80 per cent probability of class being correct), as seen in the bottom panels of Fig. 2, we see that the precision is improved as only $\sim$1 per cent of galaxies are misclassified as AGN for both SMC and LMC classifiers, and only $\sim$5 per cent of PNe are misclassified as AGN for the SMC classifier.

The recall of the post-AGB/RGB class is the worst (39 per cent and 44 per cent for the SMC and LMC, respectively), though when restricting to high confidence sources it is then PNe that have the worst recall ($\sim$88 per cent) for the LMC classifier. For the SMC and LMC, post-AGB/RGB stars are mostly misclassified as other stellar sources, with only $\sim$5 per cent (<1 per cent for high confidence sources) misclassified as AGN for the SMC, and <1 per cent for the LMC. It is not surprising that post-AGB/RGB have the lowest recall, since they are intrinsically one of the rarest source populations in our fields and thus our spectroscopic sample is also limited to a small number of examples (46 and 33 sources in the SMC and LMC respectively, with only 75 per cent of these used for training) and is likely a biased sample that does not accurately probe the full range of source properties for this class. With the upsampling, it is possible that the classifier has been overfit to this class.

Simplifying the confusion matrix by combining the classes into extragalactic, Magellanic, PM and Unknown, as seen in Fig. 3, shows that, when restricting to probabilities >80 per cent, the classifier is working well. 99.8 per cent and 99.9 per cent of all extragalactic sources are predicted as extragalactic, for the LMC and SMC fields, respectively. These confusion matrices show that most of the misclassification occurring is not between extragalactic and Magellanic classes, but within them, showing that in instances where the classifiers do not obtain the correct class, it will likely classify it as either within or outside the Clouds correctly.

The misclassification of stellar sources as AGN, and AGN as stellar sources, reflects what has been found anecdotally in the Magellanic Clouds. Classifications based on photometry have led to stars masquerading as AGN and vice versa in the Magellanic Clouds, such as SAGE0536AGN (Hony et al. 2011; van Loon & Sansom 2015) and SAGE0534AGN (Pennock et al. 2022), two AGN which were first thought to be evolved stars in the LMC, and Source 5 and Source 8 from the study of a small sample of AGN in Pennock et al. (2022), which were revealed to be stars in the SMC instead. These sources are within the data sets to be trained upon, which increases the likelihood the PRF would classify similar sources correctly, but it is possible their small number might not be enough.

Overfitting of a machine learning model can generally be spotted by using the classifier on the training set, and if the performance is much better than on the test set, then the model is overfitting. For the SMC classifier used on the SMC training set, the average accuracy was 0.92 ± 0.01, and for the LMC classifier used on the LMC training set the average accuracy was 0.88 ± 0.01. Both classifiers only performed marginally better on the training sets compared to the test sets (0.90 ± 0.01 for the SMC and 0.87 ± 0.01 for the LMC), meaning the machine learning model is not overfitting.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

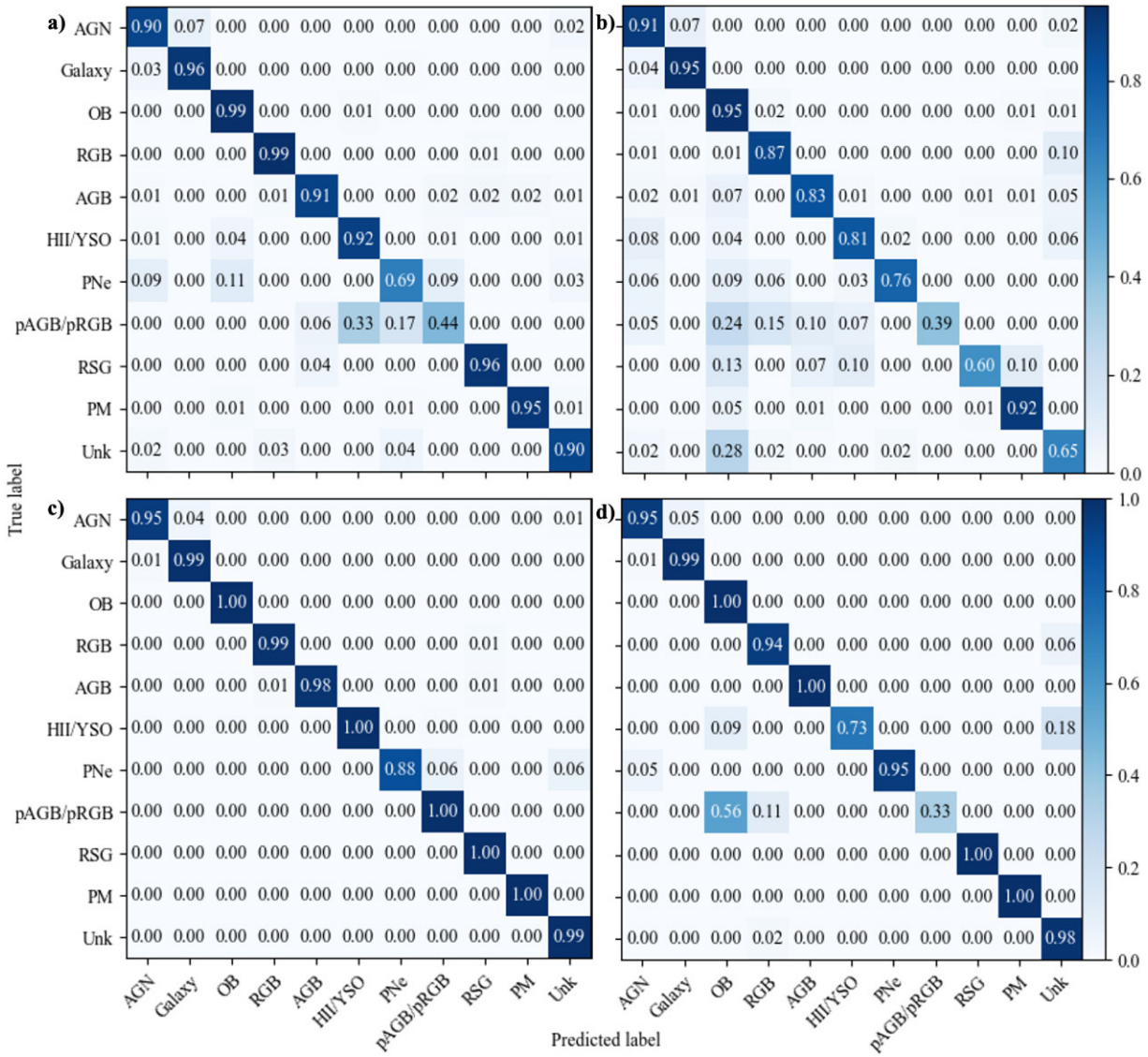

Figure 2. Confusion matrix of final classifier trained and tested on the LMC (left) and SMC (right) data sets. The top panels, (a) and (b), show all sources, whilst the bottom panels, (c) and (d), are restricted to sources that were classified with probabilities >80 per cent. This data set configuration includes extragalactic and foreground sources from both the LMC and SMC, but Magellanic stellar sources only from the respective Clouds.

3.3.6 Feature importance

The PRF algorithm can calculate overall feature importance for the entire classifier. This level of importance is calculated as 'mean decrease impurity', which is defined as the total decrease in node impurity (weighted by the probability of reaching that node, which is approximated by the proportion of samples reaching that node), averaged over all trees over the ensemble (Breiman 2001). In other words, how well each feature separates the sample into the expected classes (the decrease in class impurity). The values of the feature importances are then normalized, such that they all sum to one.

The classifiers are trained on the full data sets for SMC and LMC, from which the feature importances were calculated. This was then repeated ten times and the feature importances were then averaged for each feature. The top 15 ranked feature importances for both the SMC and LMC classifiers can be seen in Fig. 4, the full list of feature importances are available as a data product alongside the paper. For both the SMC and LMC classifiers the top 15, whilst in a different order, only have three features not in common. The feature importance plateaus with a slight decline after this towards the least important features. The full list of features and importances is available as online supplementary material. Note that feature importances are only calculated for all classes, and not individual classes.

VMC photometry and colours rank high amongst the feature importance, most likely due to all sources having at least one observation in the YJK_s bands. However, it is unlikely that is the sole reason for their high importance, therefore the colours and photometry are providing good distinction between the sources as well, showing the near-IR is a powerful resource in separating different classes.

Far-IR background emission is either the top feature or close to it. As discussed in Section 3.2, this is most likely due to far-IR being used as an analogue of extinction, which would affect the photometry in bluer bands for where far-IR flux density is higher. Also, far-IR is an indicator of how close to the centre of each of the Clouds a source

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

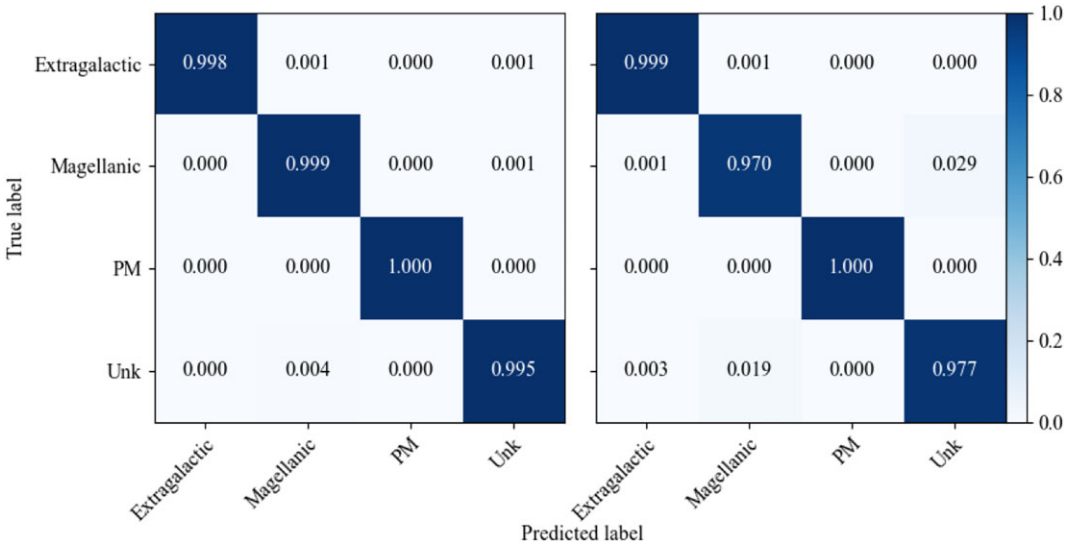

Figure 3. Simplified confusion matrix of final classifier trained and tested on the LMC (left) and SMC (right) data sets, restricted to sources that were classified with probabilities >80 per cent. The AGN and galaxy classes have been combined into the 'Extragalactic' label. The Magellanic stellar classifications have been combined into the 'Magellanic' label.

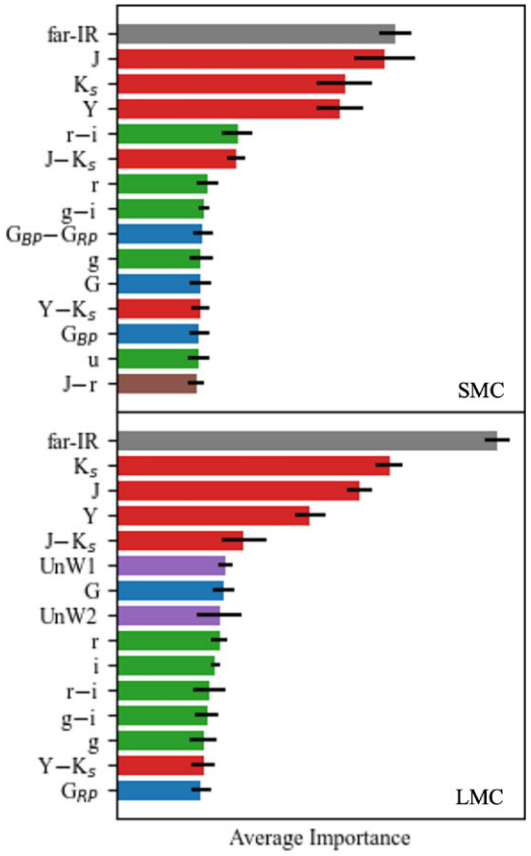

Figure 4. Top 15 important features for the SMC (left) and LMC (right) classifier. Note that, whilst the top 15 features are in a different order, they are almost all the same features for both the SMC and LMC classifier. The bars are colour-coded to represent the survey origins, where green is SMASH, blue is *Gaia*, red is VMC, purple is unWISE, grey is HERITAGE and brown represents a colour calculated from bands in different surveys.

is (higher far-IR nearer the centre), where stellar density is higher closer to the centre and thus a source is more likely to be a star rather than a background extragalactic object. Away from the Clouds, at the edges of the survey area, there is little to no far-IR emission from the Clouds and a source is more likely to be a background extragalactic object rather than a star.

Because all of the sources used for training have classifications based on spectroscopy, this means that they tend to be bright enough to be observed with *Gaia*, hence why the *Gaia* colours and photometry rank quite highly. However, despite this high dependence on *Gaia* photometry and colours, the proper motions in RA and DEC do not rank nearly as high, at ranks 85 (0.0027 ± 0.0002) and 67 (0.0036 ± 0.0004) for proper motion in RA for the SMC and LMC, respectively, and at ranks 48 (0.0046 ± 0.0004) and 110 (0.0017 ± 0.0001) for proper motion in DEC for the SMC and LMC, respectively. This is unexpected since proper motions would be the most obvious way of separating the high proper-motion stars and extragalactic sources from the Magellanic stellar sources. This could be due to the Clouds lying at the very limits of the usability of the *Gaia* data (e.g. Vasiliev 2018; Gaia Collaboration 2021b) making proper motions have significant uncertainties at the distances of the Clouds, especially for the fainter sources ($G > 18$ mag). It should be noted that an increased uncertainty in a feature tends to lead to a decrease in class probability, especially for the more important features.

The worst features are most likely due to an abundance of missing values for these features, brought upon by either a lack of coverage in certain areas and/or a lack of depth, such as for colours based on SMASH and AllWISE photometry (e.g. any SMASH photometry − any AllWISE photometry), which are the least important features for both classifiers. Leaving these features in should not affect the accuracy of the classifiers as they have been deemed unimportant, and therefore unlikely to be relied upon to make a classification.

4 RESULTS

The full VMC data set for the SMC and LMC consists of 29 514 739 and 103 172 194 sources, respectively, where, of the sources not classed as Unknown, ∼9 per cent (SMC) and ∼6 per cent (LMC) are

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Table 3. Distribution of the classifications of sources in the SMC and LMC fields for all sources, sources with $P_{\rm class}$ > 60 per cent and < 80 per cent and sources with $P_{\rm class}$ > 80 per cent.

Class	All	60 per cent < $P_{\rm class}$ < 80 per cent	$P_{\rm class}$ > 80 per cent
SMC	29 514 739	4012 812	10 478 568
SMC (Known)	7953 558	350 287	707 939
AGN	680 721	112 070	7902
Galaxy	67 522	17 158	3167
OB	4735 098	5006	8739
RGB	1119 492	203 795	682 502
PNe	1112 906	442	48
Post-AGB/RGB	29 385	2907	89
AGB	137 618	703	2382
HII/YSO	38 365	324	89
PM	30 777	7676	2777
RSG	1674	206	244
Unknown	21 561 181	3662 525	9770 629
LMC	103 172 194	7176 530	46 218 151
LMC (Known)	30 889 945	580 880	397 899
AGN	1593 270	403 382	42 605
Galaxy	230 515	49 503	23 979
OB	336 1 81	27 303	64 178
RGB	628 388	61 794	237 841
PNe	44 118	545	77
Post-AGB/RGB	4411	20	33
AGB	25 376 797	5218	15 892
HII/YSO	214 791	19 601	3268
PM	2455 355	12 721	8898
RSG	6119	793	1128
Unknown	72 282 249	6595 650	45 820 252

classed as extragalactic. Table 3 shows the distribution of classes for the entire SMC and LMC fields, as well as for sources with class probabilities ($P_{\rm class}$) >60 per cent and >80 per cent. This shows that the majority of sources (that are not classed as Unknown) in the SMC and LMC fields are classified as stars, as expected. OB and RGB stars have the highest number, which could be caused by other stellar sources that were not trained upon being classed as these classes, such as bluer stars for the OB class and redder stars for the RGB class. The majority of extragalactic sources are expected to be galaxies not hosting an AGN, however, galaxy counts tend to be lower than AGN. This could be explained by the host galaxies used to train the PRF being all low redshift sources, which could mean the higher redshift galaxies are being predicted to be other classes. It could also be because AGN can be detected out to higher redshifts than galaxies as the bright AGN continuum can be seen when the fainter continuum of a galaxy cannot, therefore we can find more AGN. High-z galaxies could potentially be classed as RGB stars if the galaxy is particularly red and dusty, but could also be classified as Unknown, which is where most of the fainter sources are expected to end up due to lack of faint sources in the training set.

For the known (sources not classed as Unknown) sources with $P_{\rm class}$ > 80 per cent in Table 3, it can be seen that the SMC sources outnumber the LMC sources. This can be attributed to the RGB class, as most of the classes are larger in number in the LMC, except the RGB, which is less than half the SMC RGB number. This may be due to the training set, as the SMC RGB training set had fainter examples of RGB than the LMC training set. Therefore, these fainter RGBs are not being picked up by the LMC classifier, and the number of the RGBs is less for the LMC.

A layout of the results table for the classification of all the sources can be seen in the online Appendices in Table C1. The catalogues of sources are separated into high-confidence sources ($P_{\rm class}$ > 80 per cent), mid-confidence sources (60 per cent < $P_{\rm class}$ < 80 per cent) and low-confidence sources ($P_{\rm class}$ < 60 per cent). Some low/mid-confidence AGNs have the possibility of being moved up to the high-confidence catalogue if they are found to be associated with an X-ray and/or radio detection (see Sections 5.4 and 5.5), or if the combined AGN and galaxy probabilities put them into a higher threshold (see Section 4.2).

For the rest of this work, unless stated otherwise, we will be referring to the high-confidence sources when exploring their distributions and properties.

4.1 PRF classification spatial distributions across the SMC and LMC fields

The spatial distributions of each of the classes across the SMC and LMC fields can be seen in Figs 5–8. Note that the most confident class predictions tend to be in the areas where all the photometric surveys overlap (see Fig. 1 for comparison), and therefore the PRF had access to the most complete data set to classify with.

The spatial distributions of the extragalactic sources are expected to be homogeneous when not looking through a nearby galaxy. In the presence of the SMC and LMC, the spatial distribution is expected to be mostly homogeneous, but highest away from the centres of the SMC and LMC, and decrease as the stellar density increases towards the centres of the SMC and LMC, as stellar sources are more likely to be in the way of the background extragalactic sources and extinction becomes more prominent. The spatial distribution of the sources classed as AGN and galaxies, seen in Fig. 5, is as expected, the number of sources slightly decreases towards the centres of the Magellanic Clouds. The highest density areas are where there is overlap between SMASH and VMC data sets (see Fig. 1), the combination of which would therefore allow for more confident classifications.

The spatial distribution of the foreground Milky Way stars can be seen in Fig. 6. This is expected to be homogeneous across the sky and this is the spatial distribution that we see.

The spatial distribution of the combined Magellanic stellar sources can be seen in Fig. 7. The spatial distribution for the LMC is as expected with the Magellanic sources concentrating in the centre. The spatial distribution of sources for the SMC is not as expected. Though the numbers do become fewer towards the edge of the survey region, the sources extend to the edges of the VMC survey area where extragalactic sources are expected to dominate. The majority of the sources causing this unexpected behaviour are classified as RGB stars.

After removing the dominating RGB class, we see the spatial distribution in Fig. 8, in which the sources concentrate in the centre of the Magellanic Clouds as expected. We can see the stellar structures of the Magellanic Clouds. The SMC is known to have a bar structure with an extension towards the East, which is what we are seeing here. We can also see the bar structure of the LMC (El Youssoufi et al. 2019) clearer after removing the RGBs. Since RGBs tend to be older, this could suggest that the bars are not an old structure.

The spatial distribution maps of the individual classes (except PM stars) can be seen in online Appendix Section D.

The spatial distribution of galaxies is mostly homogeneous as expected, with a decrease to the highest stellar densities in the centres of the two Magellanic Clouds, which is a more obvious effect for the LMC catalogue. AGNs, on the other hand, are mostly homogeneous across the entire survey footprint, with no decrease

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

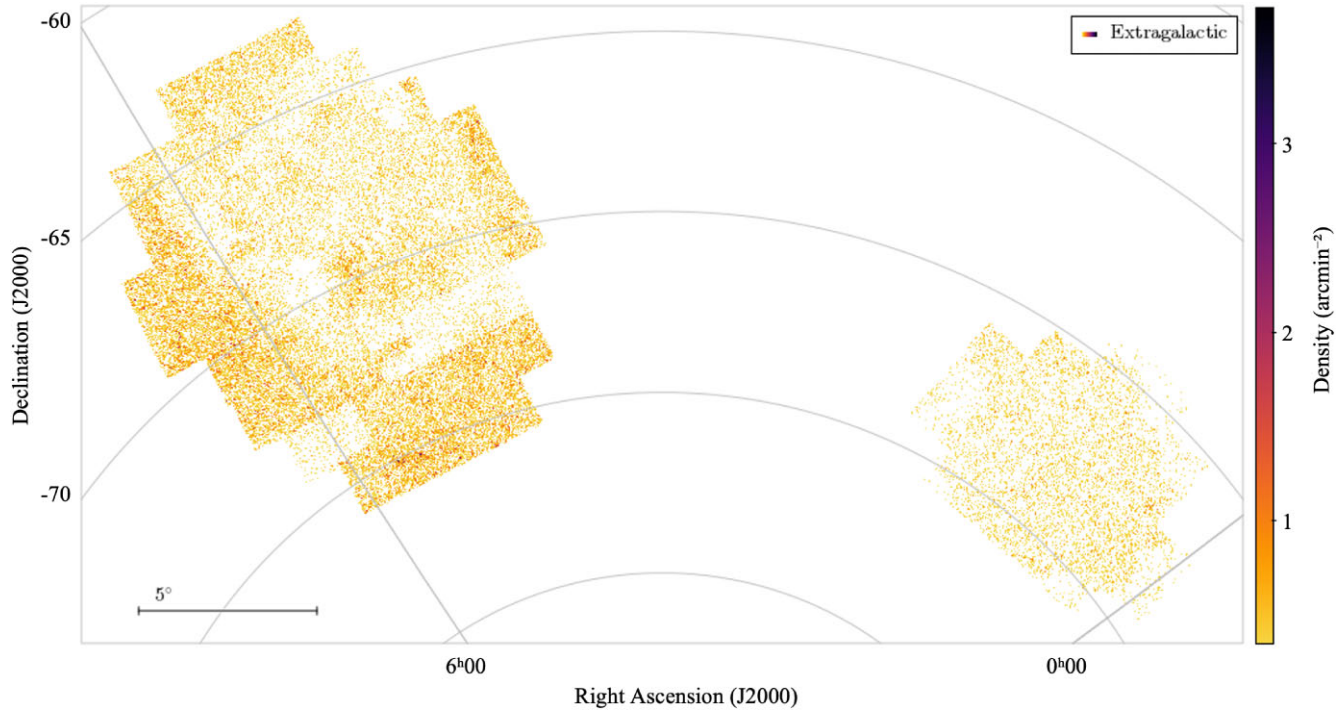

Figure 5. The sky density of the combined AGN and galaxy sources with $P_{\rm class} > 80$ per cent for the LMC (left) and SMC (right). The density of sources identified as extragalactic is fairly uniform over the survey areas, with a slight increase away from the centres of the Magellanic Clouds (as expected due to extinction and source confusion) and with some visible structure due to the differing footprints of some of the data sets that enable robust identification of extragalactic sources.

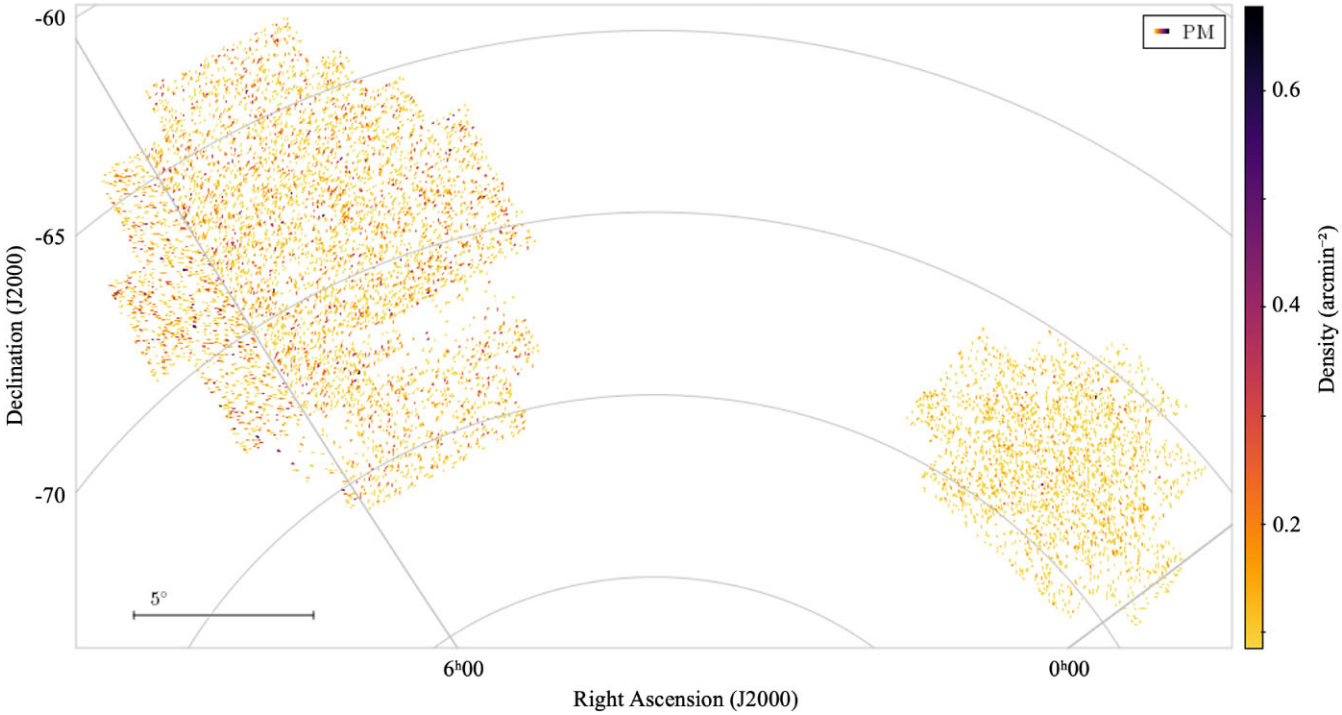

Figure 6. The sky density of the foreground high proper-motion (PM) sources with $P_{\rm class} > 80$ per cent for the LMC (left) and SMC (right). The density of sources identified as foreground stars is fairly uniform over the survey areas, as expected.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

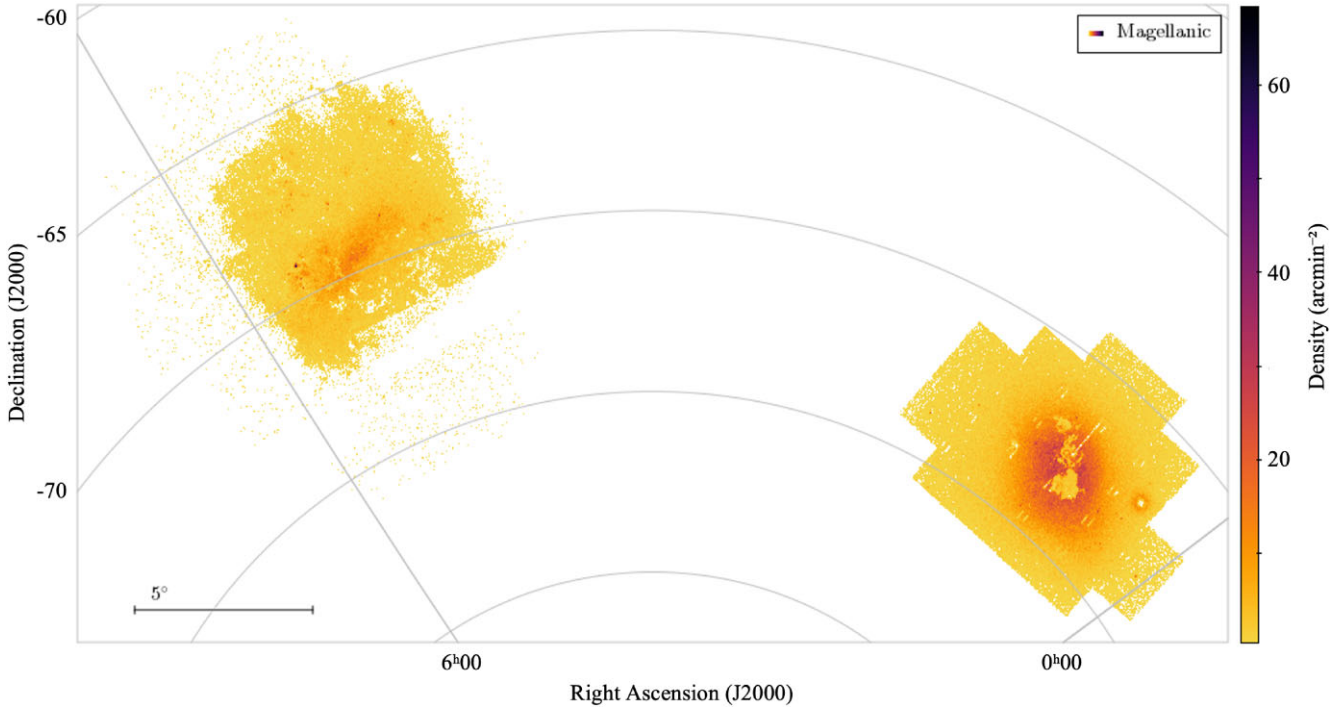

Figure 7. The sky density of the combined stellar Magellanic sources with $P_{\rm class}$ > 80 per cent for the LMC (left) and SMC (right). The distribution is dominated by intermediate-age/old RGB stars.

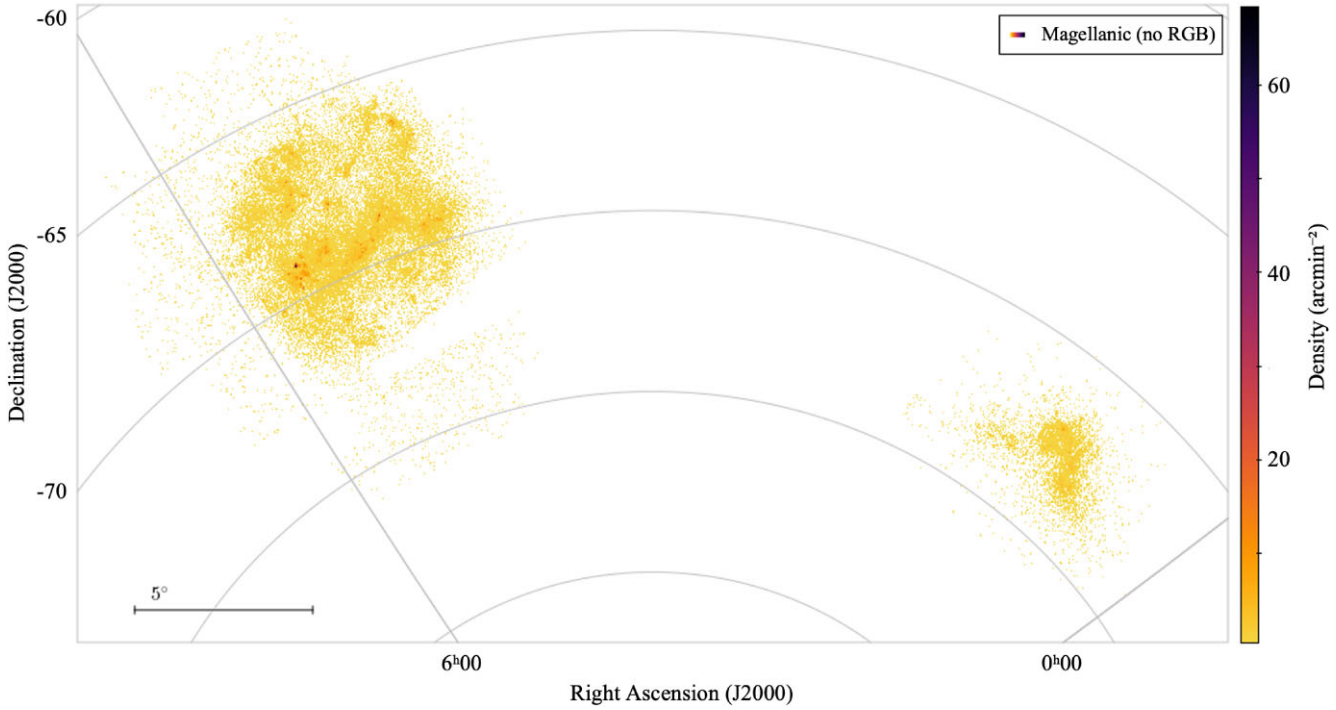

Figure 8. The sky density of the combined stellar Magellanic sources with RGB sources removed and $\boldsymbol{P}_{\rm class}$ > 80 per cent for the LMC (left) and SMC (right).The sources identified as Magellanic concentrate in the centres of the Clouds with fewer sources on the outskirts of the survey area. The sources concentrate more in the bar structure of the Clouds and the distribution is dominated by young stars.

towards the galactic centres. The lack of decrease towards the centres of the Clouds for the AGN spatial distribution could be due to brighter AGNs being easier to see through the Clouds than the fainter background galaxies.

The two most noticeable globular clusters, 47 Tucanæ and NGC 362, show noticeably fewer sources with probability >80 per cent. This is most likely because no sources from these clusters, which are closer to us than the SMC, were used in training. They can be traced

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

mainly by the spatial distributions of RGB, Unknown and PM stars, as expected.

Overall, the different classes are distributed across the VMC fields of the LMC and SMC mostly as expected.

4.2 AGN and galaxy classifications

Some sources have a combined AGN and Galaxy probability that makes them a high confidence extragalactic source, if not necessarily a high confidence AGN or Galaxy. These sources tend to be obviously extended in VISTA images, proving their most likely extragalactic nature.

For this reason sources with a combination of AGN and Galaxy probabilities $\geq$80 per cent (56 134 and 140 212 for the SMC and LMC, respectively) are moved to the high confidence catalogue, and those that have $\geq$ 60 per cent and $<$80 per cent (126 514 and 302 530 for the SMC and LMC, respectively) are moved to the mid-confidence catalogue.

5 DISCUSSION

5.1 Classification of unusual dust-dominated AGN

As part of the work done in Pennock et al. (2022), unusual AGN that appeared to showcase dust emission almost entirely from the AGN, that are often misidentified as stellar objects, were found using unsupervised machine learning and spectroscopically observed. This work was then continued into the LMC. Some of these sources were used in training. The PRF classifications that we obtain here for the full sample of candidate dust-dominated AGN from Pennock et al. (2022) and the extended sample in the LMC field can be seen in Table 4.

For the SMC sources, most are confident AGN, except the two known stars. SMCtSNE5 is confidently classed as an AGB star, which is not unexpected as it is a carbon star. SMCtSNE8, which is a long-period variable AGB star with a Hα emission line from shock dissipation, is classed as a low confidence AGN ($\sim$54 per cent probability) with a $\sim$15 per cent probability of being a pAGB/RGB and $\sim$12 per cent probability of being an AGB, as the next highest possibilities.

For the LMC sources, 15 sources are classed as AGN with varying confidences. LMCtSNE14 (AGN, $z \sim 0.4$) is classed as a galaxy with a $\sim$ 30 per cent chance of being an AGN, which, considering its visible host galaxy in VMC images, is not that surprising. We find that, although not predicted confidently as AGN, dusty AGB stars and YSOs are the main contaminants of AGN-dust dominated samples.

This shows that unusual dust-dominated AGN that have often been mistaken for dusty Magellanic objects are being classified correctly by the PRF, most likely helped by the inclusion of similar sources in the training set. This does, however, show that emission line stars have a chance of being classed as AGN, though with possibly low probabilities.

5.2 Unseen stellar classes

One way to ascertain the performance of the classifier in separating extragalactic from stellar sources is to test it on classes it has not seen before. The classifications from SAGE-spec (Ruffle et al. 2015; Jones et al. 2017) were used as part of the training set for the classifier. Not all the classes in this data set were added to the training set as they were deemed too few in number for training purposes and/or not a well-defined class (e.g. emission-line stars). The SAGE-spec data sets for the SMC and LMC has 18 and 47 sources, respectively, that were not used. The predicted probabilities and classes of these 18 sources were extracted from the full SMC and LMC data sets and can be seen in Tables 5 and 6, respectively.

Overall, for the SMC classifier, all of the sources were classed as stellar sources. All the sources are predicted to have a $<$10 per cent probability of being an AGN and $<$5 per cent probability of being a galaxy. This implies that similar stellar sources of these natures would most likely not be predicted to be extragalactic. RCrB are often associated with post-AGB stars so it is not surprising that two out of three were classed as AGB stars. WR can be similar to O type stars so would be expected to be predicted as an OB star, which all but one are. The WR star predicted as an RGB star could possibly be a case of a dusty WR star.

For the LMC classifier, two sources were classified as UNK in the SAGE-spec catalogue, but were classed by the PRF as AGN. The rest of the sources were predicted to be one of the stellar classes. Just as for the SMC, the RCrB were mostly classed as AGB, WR were mostly predicted to be OB stars and the one BSG was predicted to be an OB star. The two SNRs were predicted to be HII/YSOs, the two LBVs were predicted as PM stars, the seven RVTau were predicted as post-AGB/RGB and AGB stars and the YSG was predicted as an RSG. These are all unsurprising as these stellar objects share properties with the stellar classes they have been classed as.

These results are promising for AGN and galaxy classifications, as this shows that Magellanic classes that have not been trained upon are classified as one of the other Magellanic classes.

5.3 Colour–magnitude selections of PRF classified sources

In this section, we explore the class distributions across colour–colour and colour–magnitude diagrams in the optical, near-IR, and mid-IR. Here, we focus on the sources with $P_{\rm class} > 80$ per cent.

5.3.1 *Optical*

From the optical colour–magnitude diagram seen in Fig. 9, we can see in the Unknown sources the structure of stellar sequences (Nidever et al. 2017). The main sequence stars are expected to start from the bottom of the diagram and fork to the left, which is what we see in the Unknowns. It is not surprising that main sequence stars would be picked up as Unknowns as they were not a class that was trained on. The right fork of the Unknowns is expected to be supergiants and RGB stars. The sources classed as RGB stars and RSG stars do follow this fork, so it is likely that not all the RGB and RSG stars were classified, possibly due to mismatches between the different photometry and/or missing data. AGN can be mostly found in the middle of the expected stellar sequence, meaning they would be hard to classify in just optical alone. Galaxies tend to concentrate just to the right of the stellar sequences, and it is possible the Unknowns in this region are also galaxies that were not accounted for in the training set.

5.3.2 *Near-IR*

The near-IR VISTA colour–magnitude diagram ($J - K_{\rm s}$ versus $K_{\rm s}$) can be seen in Fig. 10. Most sources that have $J - K_{\rm s} > 1$ mag and $K_{\rm s} > 12$ mag are expected to be background galaxies and quasars (see region L in Cioni et al. 2014, 2016), and from Fig. 10 (right panel) we see that the AGN and galaxy populations follow this.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Table 4. Classifications of sources from Pennock et al. (2022) and Pennock et al. (in preparation) that were candidate AGN with little to no dust from the host galaxy and that also had a tendency to be misclassified as dusty stellar sources in the Magellanic Clouds. C star refers to carbon stars and Em∗ refers to emission-line stars. Note that some of these sources were used in the training set, indicated by the 'T?' column, by either 'Y' (yes) or 'N' (no).

Name	RA	DEC	PRF Class	$P_{\rm class}$	*T*?	Spec. Class
SMCtSNE1	00:36:16.99	−74:31:31.3	AGN	0.99	Y	AGN
SMCtSNE2	01:13:37.08	−74:27:55.3	AGN	0.97	Y	AGN
SMCtSNE3	00:31:56.89	−73:31:13.6	AGN	0.96	Y	AGN
SMCtSNE4	00:26:02.54	−72:47:18.0	AGN	0.98	Y	AGN
SMCtSNE5	00:48:25.70	−72:44:03.0	AGB	0.99	Y	C star
SMCtSNE6	01:14:08.00	−72:32:43.3	AGN	0.98	Y	AGN
SMCtSNE7	00:55:51.51	−73:31:10.0	AGN	0.97	Y	AGN
SMCtSNE8	01:22:36.90	−73:10:16.7	AGN	0.45	N	Em∗
SMCtSNE9	01:21:08.40	−73:07:13.1	AGN	0.99	Y	AGN
SMCtSNE10	01:15:34.09	−72:50:49.3	AGN	0.94	Y	AGN
SMCtSNE11	00:39:10.78	−71:34:09.9	AGN	0.98	Y	AGN
SMCtSNE12	00:51:16.95	−72:16:51.5	AGN	0.98	Y	AGN
SMCtSNE13	00:57:32.80	−72:13:02.0	AGN	0.99	Y	AGN
SMCtSNE15	00:34:05.30	−70:25:52.3	AGN	0.82	Y	AGN
SMCtSNE16	00:49:52.50	−69:29:56.0	AGN	0.34	Y	AGN
LMCtSNE2	06:15:04.01	−66:17:16.4	AGN	0.69	Y	AGN
LMCtSNE3	05:33:57.69	−64:20:24.9	AGN	0.87	Y	Galaxy
LMCtSNE4	05:01:10.84	−73:36:35.0	AGN	0.92	Y	AGN
LMCtSNE5	05:41:12.99	−64:11:53.7	AGN	0.69	N	?
LMCtSNE6	05:45:05.73	−64:11:19.3	AGN	0.60	N	AGN
LMCtSNE7	05:20:19.84	−73:55:37.3	AGN	0.31	N	AGN
LMCtSNE8	05:32:10.38	−73:57:22.3	AGN	0.36	N	AGN
LMCtSNE9	04:38:50.67	−72:17:12.6	AGN	0.50	N	AGN
LMCtSNE10	05:14:17.90	−72:20:19.2	AGN	0.56	Y	AGN
LMCtSNE11	04:51:38.41	−71:02:06.1	AGN	0.95	Y	AGN
LMCtSNE12	05:40:55.08	−70:34:46.9	OB	0.84	N	Em∗
LMCtSNE13	05:22:52.28	−69:50:42.6	HII/YSO	0.21	N	Em∗
LMCtSNE14	05:51:43.28	−68:45:43.0	Galaxy	0.58	Y	AGN
LMCtSNE15	05:22:30.52	−67:54:43.6	OB	0.71	N	Em∗
LMCtSNE16	05:31:48.96	−67:21:33.8	HII/YSO	0.41	N	Star
LMCtSNE17	05:31:54.44	−68:26:40.4	HII/YSO	0.99	Y	Em∗
LMCtSNE18	05:48:22.29	−67:58:53.3	AGB	0.44	N	Em∗
LMCtSNE19	05:04:47.16	−66:40:30.7	HII/YSO	0.69	N	Em∗
LMCtSNE20	05:53:57.48	−66:50:01.6	AGN	0.90	N	AGN
LMCtSNE21	06:10:52.23	−66:30:11.5	AGN	0.62	N	AGN
LMCtSNE22	05:19:42.45	−65:02:16.8	AGN	0.88	Y	AGN
LMCtSNE23	05:49:13.47	−64:29:29.2	AGN	0.60	N	AGN
LMCtSNE24	05:43:34.33	−64:22:58.2	AGN	0.83	N	AGN

Cioni et al. (2014) also states that a minority of RGB stars could also be scattered to this region due to larger extinctions. Fainter Unknown sources in this region are most likely extragalactic sources at higher redshift than in the training set.

The stellar classes in Fig. 10 mostly avoid the extragalactic classes. The ones that do not, YSOs, PNe, post-AGB/RGB and AGB stars, are most likely reddened due to dust. The RGB stars that are $K_{\rm s} > 16$ are in the region expected for RGB stars, and the RGB stars brighter than this are in the area for dusty AGB stars.

5.3.3 *Mid-IR*

The distributions of extragalactic and stellar sources are plotted in AllWISE colour–colour diagrams (for expected distributions see, e.g. Stern et al. 2012; Assef et al. 2013; Nikutta et al. 2014). The extragalactic sources can be seen in Fig. 11 (right panel) where both AGN and galaxies occupy expected regions in colour–colour space. The galaxy class occupies the region expected for both star-forming and elliptical galaxies, whilst AGN occupy the region of QSOs and Seyferts. The foreground stars (PM) are also plotted here and are nicely centred on (0,0) in Vega colours, as expected. The 'Unknown', sources are shown to overlap with several of the classes, but spread further to the bottom right than the other classes. Note that this is where the WISE sources with low signal to noise (S/N < 3) tend to end up.

The distributions of the stellar Magellanic sources across the AllWISE colour–colour diagram can be seen in Fig. 11 (left panel). The AGB sequence can be clearly seen. The populations of RGB, OB, and RSG stars tend to concentrate below $W1-W2 \sim 0$, unlike the extragalactic sources which tend to concentrate above $W1-W2 \sim 0$. The PNe and YSO and post-AGB/RGB are the classes that show the most cross-over with the extragalactic classes, which is not unexpected as they are known to be hard to differentiate from extragalactic sources in colour–colour diagrams. It is noted that sources below $W1-W2 \sim -1$ mag are mostly found within the higher density regions (centre of the SMC). This could be due to *WISE*

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Table 5. Sources from Ruffle et al. (2015) that were not used in the training of the PRF classifier. The SAGE classes are Wolf-Rayet stars (WR), R Coronæ Borealis variable stars (RCrB), Blue supergiants (BSG), S-type stars (S star), Symbiotic stars (Sym. star), and stars of indiscernible type (star).

Source name	SAGE Class	RA (J2000)	DEC (J2000)	PRF Class	P_{class}
SMC-WR9	WR	00:54:32.2	−72:44:36	OB	0.99
SMC-WR12	WR	01:02:52.2	−72:06:52	OB	0.96
GSC 09141−05631	WR	00:43:42.2	−73:28:54	OB	0.99
SMC-WR2	WR	00:48:31.0	−73:15:45	OB	0.99
SMC-WR3	WR	00:49:59.3	−73:22:14	OB	0.99
SMC-WR4	WR	00:50:43.4	−73:27:05	OB	0.98
RMC 31	WR	01:03:25.2	−72:06:44	OB	0.80
SMC-WR11	WR	00:52:07.5	−72:35:38	OB	0.99
MSX SMC 014	RCrB	00:46:16.4	−74:11:13	AGB	0.74
MSX SMC 155	RCrB	00:57:18.2	−72:42:35	AGB	0.56
AzV 404	Star	01:06:29.4	−72:22:09	OB	0.52
BFM 1	S star	00:47:19.3	−72:40:04	AGB	0.97
AzV 456	Star	01:10:55.8	−72:42:57	OB	0.84
AzV 23	Star	00:47:38.9	−73:22:54	OB	0.50
OGLE SMC-SC10 107856	RCrB	01:04:53.0	−72:04:04	AGB	0.43
MSX SMC 185	Sym. star	00:54:20.0	−72:29:09	PNe	0.53
HD 5980	WR	00:59:26.7	−72:09:54	OB	0.62
HD 6884	BSG	01:07:18.1	−72:28:04	AGB	0.41

photometry being affected by blends, making the longer wavelength, poorer angular resolution data appear brighter.

The OB stars appear to have two populations in the AllWISE colour–colour diagram. One population that concentrates below $W1-W2 \sim 0$ as expected, and a smaller one that concentrates just above this, a redder population, which also happens to be the expected area for galaxies. Some possibilities are that it is either the star lighting up surrounding ISM ('Pleiades effect', e.g. Ivanov et al. 2024; Sheets et al. 2013), a nascent star (B[e] star), or a mature, Be star with an excretion disc, in which case the red $W1$–$W2$ colour is caused by free-free emission from the circumstellar ionized gas (rather than dust). Spectroscopically observed stars undergoing the Pleiades effect (Sheets et al. 2013), as well as B[e] and Be stars (Reid & Parker 2012) have been plotted on the AllWISE colour–colour diagram in Fig. 12. From this we can see that redder sources are most likely Be or B[e] stars, with near-IR excess most likely due to free−free emission.

Overall, the colour–colour and colour–magnitude diagrams for the different wavelength regimes show that the majority of sources are being separated where expected and that even in areas where multiple classes can be found the PRF is still capable of separating the sources. For comparisons between the distributions of the training set versus the classed sources, see online Appendix Section E.

5.4 Classifications of the radio population

Cross-matching with the radio ASKAP SMC (Joseph et al. 2019) and LMC (Pennock et al. 2021) catalogues with a 2 arcsec search radius, gives 1047/7736 and 8120/54612 sources for SMC and LMC, respectively, which have a P_{class} >80 per cent. The numbers of radio sources per class can be seen in Table 7. A search radius of 2 arcsec was used, as when cross-matching with a larger search radius of 5 arcsec it was seen that the AGN class, the sources that are the most likely true counterparts, peaked at a separation radius of $\sim$ 1 arcsec. Doubling this to a search radius of 2 arcsec was used to include the majority of AGN matches whilst reducing the number of mismatches.

The majority (78 per cent for the SMC and 70 per cent for the LMC) of these sources are classed as Unknown. From the other classes, the class with the highest number is AGN, followed by galaxies, as expected as such sources are often radio bright. However, for both the SMC and LMC, RGBs number >50. RGBs are not expected to be radio sources, so this implies that there were some misclassification, or that these RGB are not the true counterparts to the radio sources. Due to the significant differences in resolution between both data sets, mismatching is not unexpected. The separation between the VMC coordinates and ASKAP coordinates for the RGB class is in general larger than AGN ($\sim$ 1.35 arcsec for RGB compared to $\sim$ 1 arcsec for AGN), which implies that these RGB sources could be merely mismatches.

Of the other stellar sources, H II/YSO and PNe are expected to be associated with a radio detection, and foreground stars are close enough that a radio detection is possible. One of the brightest and well known radio sources in the LMC is supernova SN 1987A, which matched with 2 sources in the PRF catalogue within 1 arcsec. Both of which had a classification of H II/YSO with probability 31–35 per cent, with the next highest class as AGN with probability 22–27 per cent. So, the classifier did not know what to class it as and did not put it in the Unknown class, proving that it is quite a unique source.

In relation to the full catalogue of sources with P_{class} > 80 per cent, the fraction of AGN with a radio detection is $\sim$1.24 per cent and $\sim$3.89 per cent and for galaxies is $\sim$1.58 per cent and $\sim$2.54 per cent, for the SMC and LMC, respectively. The expected radio loud population is about 10 per cent, but ours is limited to the likeliest AGN, i.e. those that are well sampled in the training data, so its possible that the missing fraction is in the lower confident AGN population and/or the Unknown class.

It has been seen that there is an upturn in the number of sources towards fainter flux densities, representing the beginning of the faint galaxy population, as well as the radio quiet AGN population (e.g. Pennock et al. 2021). Therefore, it is expected that the number of radio detected galaxies will have increased towards the fainter fluxes. Looking at the radio flux density distribution at 888 MHz (LMC) and 1320 MHz (SMC), and the ratio of galaxy to AGN counts in Fig. 13, it can be seen at the brightest fluxes, there are less galaxies compared to AGN, as expected, and towards lower flux densities the

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Table 6. Sources from Jones et al. (2017) that were not used in the training of the PRF classifier. The SAGE classes are Wolf-Rayet stars (WR), R Coronæ Borealis variable stars (RCrB), Blue supergiants (BSG), Yellow supergiants (YSG), S-type stars (S star), Symbiotic stars (Sym. star), RV Tauri stars (RVTau), SNR, unknown (UNK), luminous blue variable stars (LBV), and stars of indiscernible type (star).

Source name	SAGE class	RA (J2000)	DEC. (J2000)	PRF class	$P_{\rm class}$
LHA 120-N 82	WR	04:53:30.30	−69:17:49.2	HII/YSO	0.39
HD 268 813	STAR	04:54:23.23	−70:26:56.8	RSG	0.39
RP 1631	RCrB	05:00:35.35	−70:52:00.5	AGB	0.60
HV 2281	RVTau	05:03:05.05	−68:40:25.0	pA/RGB	0.82
LMC-BM 11–19	STAR	05:03:43.43	−67:59:19.0	AGB	0.86
RP 1878	UNK	05:04:34.34	−67:52:21.4	AGN	0.39
	BSG	05:06:39.39	−68:22:09.5	OB	0.99
HV 915	RVTau	05:14:18.18	−69:12:35.3	pA/RGB	0.41
	STAR	05:15:26.26	−67:51:27.0	AGB	0.99
KDM 3196	STAR	05:18:08.08	−71:51:53.6	AGB	0.65
HV 2444	RVTau	05:18:45.45	−69:03:22.0	AGB	0.58
	STAR	05:19:45.45	−69:30:00.0	AGB	0.88
HV 942	RCrB	05:21:48.48	−70:09:57.2	AGB	0.46
HV 5829	RVTau	05:25:19.19	−70:54:10.1	AGB	0.56
MACHO 82.8405.15	RVTau	05:31:51.51	−69:11:46.3	pA/RGB	0.57
	STAR	05:32:07.07	−70:10:25.0	AGB	0.94
SHP LMC 256	UNK	05:34:44.44	−67:37:50.5	AGN	0.78
KDM 5345	UNK	05:38:24.24	−66:09:00.4	AGB	0.99
MACHO 81.9728.14	RVTau	05:40:01.01	−69:42:14.8	OB	0.31
	UNK	05:45:46.46	−67:32:39.1	AGB	0.40
KDM 6247	STAR	05:47:57.57	−68:14:57.1	AGB	0.91
HV 2862	RVTau	05:51:23.23	−69:53:51.4	pA/RGB	0.84
PMP 133	STAR	05:52:53.53	−69:30:35.3	PM	0.96
HD 270 754	STAR	04:47:05.05	−67:06:53.3	OB	0.97
HD 32 402	WR	04:57:24.24	−68:23:56.8	OB	0.87
HD 269 187	STAR	05:14:04.04	−67:15:50.8	PM	0.78
S Dor	LBV	05:18:14.14	−69:15:01.4	PM	0.35
OGLE LMC-RCB-10	RCrB	05:20:48.48	−70:12:13.0	AGB	0.91
HD 36 402	WR	05:26:04.04	−67:29:57.1	OB	0.72
W Men	RCrB	05:26:25.25	−71:11:11.8	AGB	0.40
HD 269 662	LBV	05:30:52.52	−69:02:58.9	PM	0.53
HV 12 620	STAR	05:33:00.00	−70:41:23.6	AGB	0.82
HV 2671	RCrB	05:33:49.49	−70:13:23.5	HII/YSO	0.40
SN 1987A	SNR	05:35:28.28	−69:16:11.3	HII/YSO	0.36
W61 27–27	STAR	05:36:04.04	−69:01:30.4	OB	0.91
	WR	05:36:44.44	−69:29:46.0	OB	0.98
SNR B0540-69.3	SNR	05:40:11.11	−69:19:54.5	HII/YSO	0.76
HD 269 953	YSG	05:40:12.12	−69:40:04.8	RSG	0.39
IRAS 05413–6934	UNK	05:40:54.54	−69:33:18.7	HII/YSO	0.99
MSX LMC 1795	RCrB	05:42:22.22	−69:02:59.6	AGB	0.90
LHA 120-S 61	WR	05:45:52.52	−67:14:25.8	OB	0.80
HD 270 422	STAR	05:56:48.48	−66:39:05.0	RSG	0.35
HD 270 467	STAR	05:58:12.12	−66:20:23.6	PM	0.81
WOH G 642	STAR	05:59:21.21	−66:31:56.6	PM	0.99
HD 41 466	STAR	06:00:19.19	−66:13:27.5	PM	0.34
HD 270 485	STAR	06:00:53.53	−66:55:48.0	PM	0.99
HD 271 776	STAR	06:01:38.38	−66:35:20.0	PM	0.92

number of galaxies compared to AGN increases, also as expected. However, at about $F_{888\rm MHz}, F_{1320\rm MHz} < 7$ mJy, the number of galaxies compared to AGN starts decreasing towards lower flux densities, which is unexpected, as at lower flux densities we would expect an increase in faint galaxies, but it could be that the faint AGN population is easier to classify than the faint galaxy population. The trend is less clear for the SMC where source statistics are poorer.

It should be noted that the AGN and galaxies being compared here only represent the more confident classifications of the PRF, a fraction of the true number in this field that will have been detected with ASKAP. More AGN than galaxies are also found, most likely due to them being easier to detect at higher redshifts, which could account for the increase towards lower radio flux densities, as we are identifying the radio quiet AGN, but not the fainter star-forming galaxies. It should also be noted that a source classed as a galaxy with a radio association could indicate AGN activity that is not visible at other wavelengths.

The 1320 MHz band was used over the 960 MHz for the SMC, as the observation at 960 MHz did not use the full ASKAP array, but the 1320 MHz observation did, and would therefore have similar depths to the LMC 888 MHz observation.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

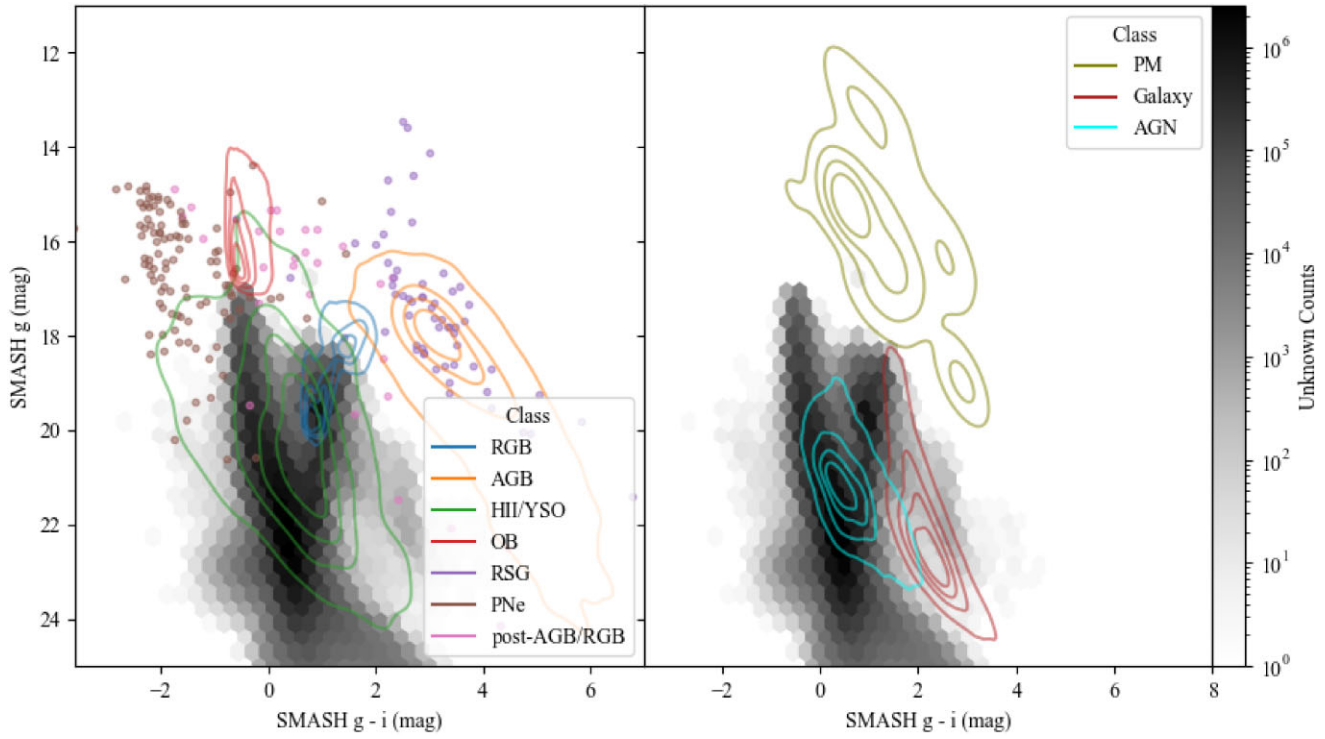

Figure 9. SMASH colour–magnitude diagrams of the sources classified as Magellanic sources (left), foreground stars and extragalactic (right) in the SMC and LMC fields. The contours represent a probability distribution in intervals of 0.2. The sources identified as Unknown with $P_{\rm class} > 80$ per cent are represented as a 2D histogram in the background of both plots.

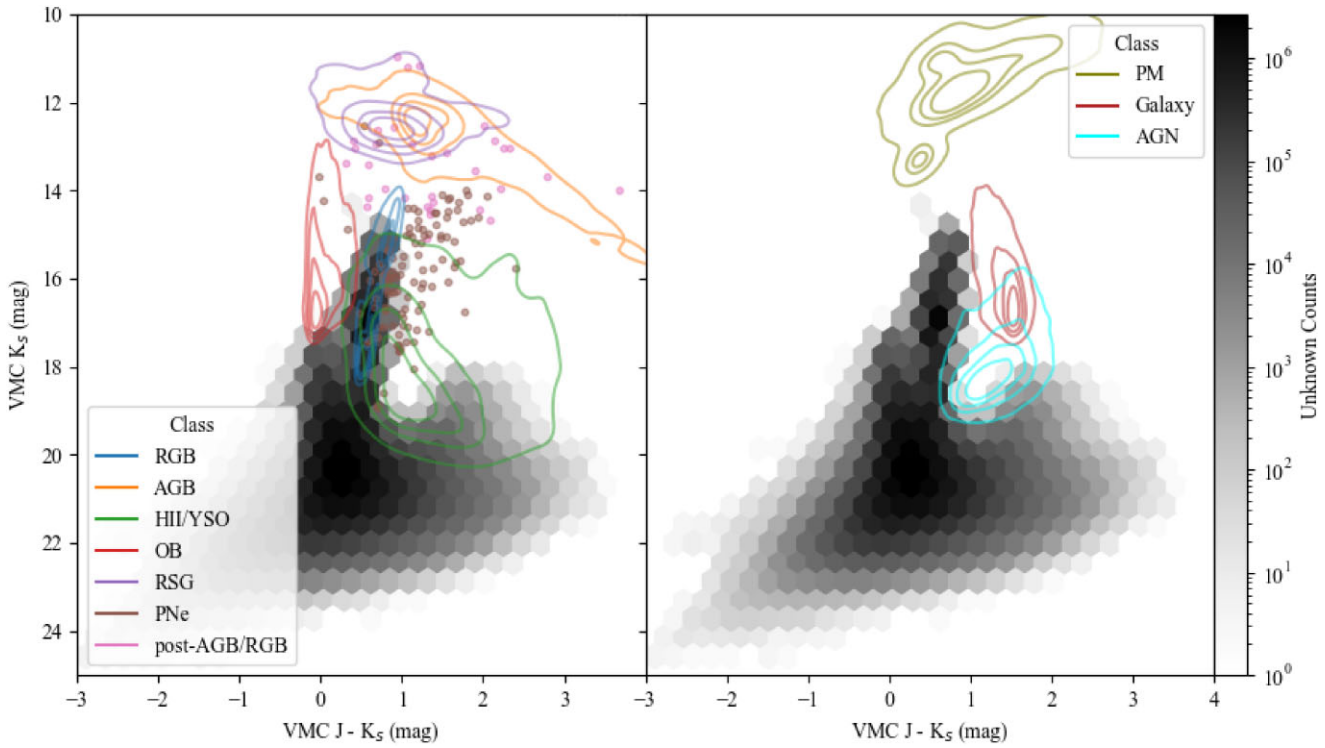

Figure 10. VISTA colour–magnitude diagrams of the sources classified as Magellanic sources (left), foreground stars and extragalactic (right) in the SMC and LMC fields. The contours represent a probability distribution in intervals of 0.2. The sources identified as Unknown with $P_{\rm class} > 80$ per cent are represented as a 2D histogram in the background of both plots.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

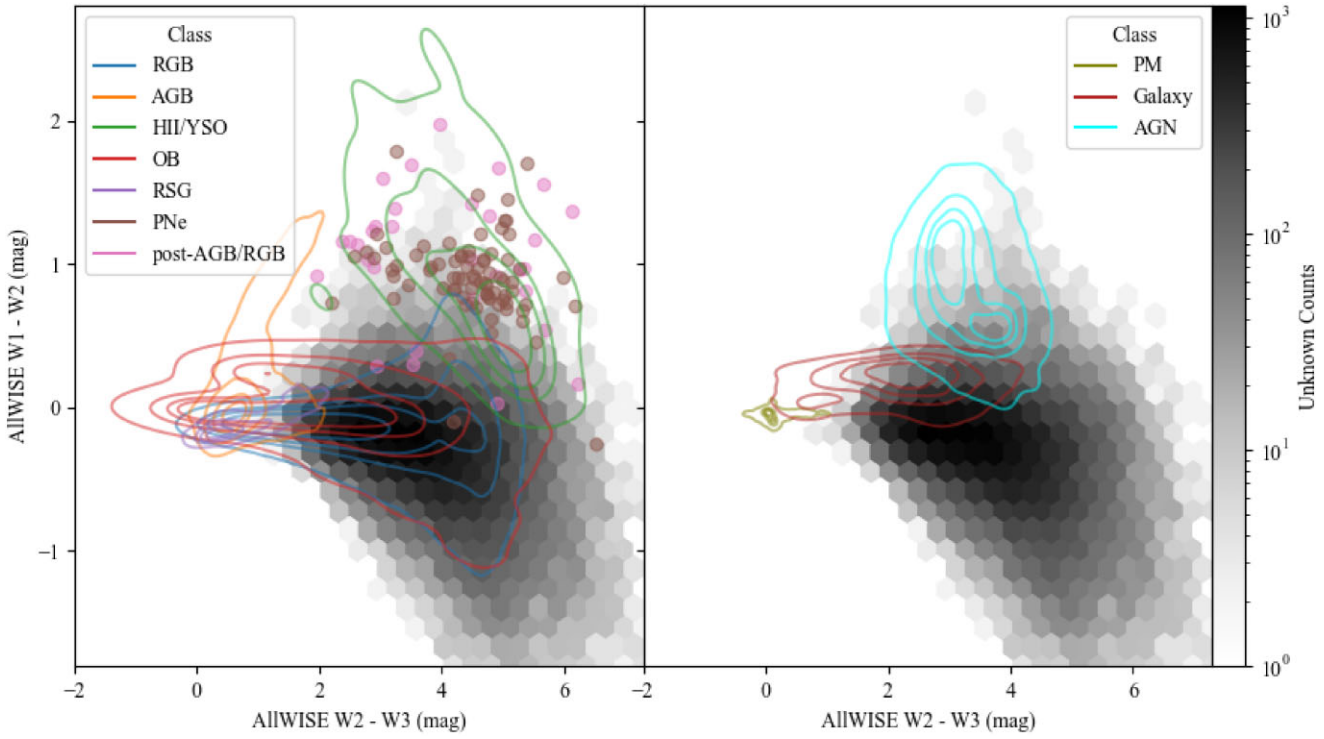

Figure 11. AllWISE colour–colour diagrams of the sources classified as Magellanic sources (left), foreground stars and extragalactic (right) in the SMC and LMC fields. The contours represent a probability distribution in intervals of 0.2. The sources identified as Unknown with $P_{\rm class} > 80$ per cent are represented as a 2D histogram in the background of both plots.

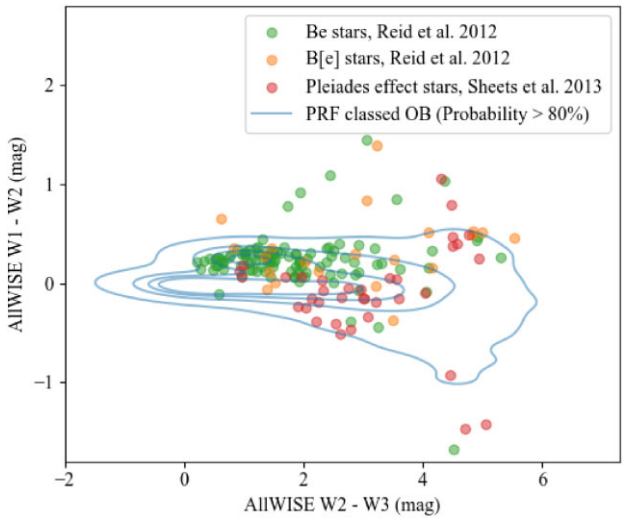

Figure 12. AllWISE colour–colour diagram of the sources classified as OB stars (blue contours). The contours represent a probability distribution in intervals of 0.2. Overplotted are stars that exhibit the Pleiades effect (red circles), Be stars (green circles) and B[e] stars (orange circles).

5.5 Classifications of the X-ray population

Cross-matching with the *XMM–Newton* catalogues for the SMC (Sturm et al. 2013) and all-sky (Webb et al. 2020) catalogues at a cross-matching radius less than the positional error in the X-ray coordinates for each source yields a total of 627 and 4794 X-ray sources with reliable PRF classifications ($P_{\rm class}$ >80 per cent) for

Table 7. The number of sources per class that were cross-matched with ASKAP within a 2 arcsec search radius.

Class	All	60 per cent $<P_{\rm class}$ <80 per cent	$P_{\rm class}$ >80 per cent
SMC	4469	965	1047
Unk	2208	544	813
AGN	790	192	98
Galaxy	431	189	50
RGB	153	33	76
OB	540	1	4
HII/YSO	36	4	1
AGB	18	0	3
PNe	285	1	2
RSG	0	0	0
Post-AGB/RGB	5	1	0
PM	3	0	0
LMC	37 375	6450	8120
Unk	15 752	2002	5660
AGN	10 399	3012	1658
Galaxy	4224	1361	609
RGB	197	25	57
OB	1324	16	63
HII/YSO	206	23	28
AGB	3880	1	1
PNe	292	7	38
RSG	1	1	0
Post-AGB/RGB	6	0	2
PM	1154	2	4

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

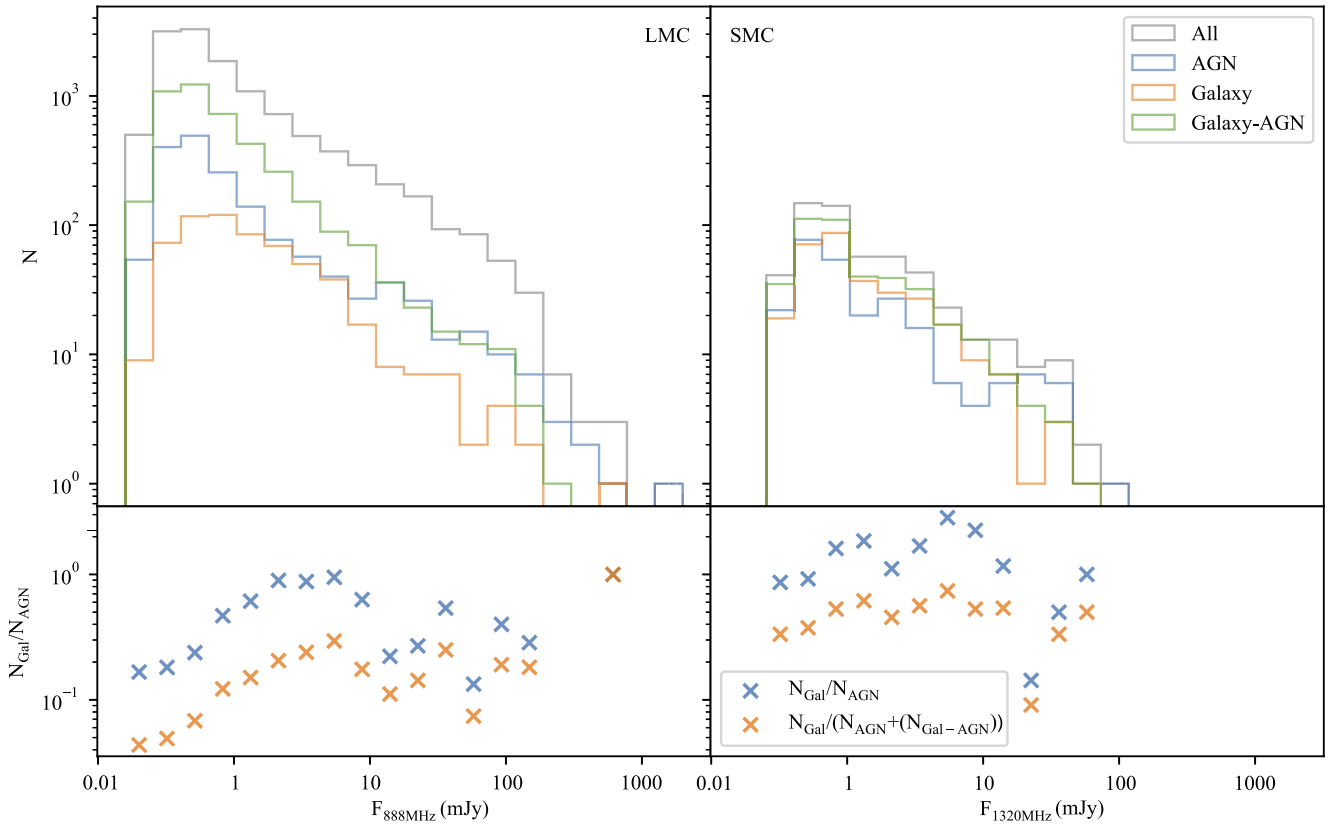

Figure 13. Radio LMC ASKAP 888 MHz (left) and SMC ASKAP 1320 MHz (right) flux density distributions and the ratio of galaxy to AGN counts (bottom panels) of the predicted AGN and galaxy sources with P_{class} > 80 per cent. A galaxy-AGN subset has also been included where the individual AGN and galaxy probabilities are <80 per cent, but the combined probabilities are >80 per cent.

the SMC and LMC, respectively. The number of sources per class can be seen in Table 8. Most of the sources are Unknown. Of the other classifications, the highest count is AGN, as expected.

From the cross-matching it can be found that for sources with P_{class} > 80 per cent, ∼1.49 per cent, and ∼1.02 per cent of all AGNs are X-ray detected, for the SMC and LMC, respectively. However, the *XMM–Newton* surveys of the SMC and LMC do not cover the full VMC survey areas, and concentrate mainly in the centres of the Clouds, so the true percentages are most likely much higher.

In Fig. 14, we plot the unWISE *W*1 and SMASH *g*-band magnitudes as a function of X-ray flux for Magellanic (left) and extragalactic and foreground stars (right). In these plots, the extragalactic and stellar sources tend to separate. The extragalactic tend to concentrate at the fainter optical/IR magnitudes for a given X-ray flux (e.g. Hornschemeier et al. 2001; Civano et al. 2015; Nandra et al. 2015; Salvato et al. 2018).

Sources classified as Unknown which have X-ray detections are mainly concentrated on the extragalactic side of the plots. This side of the plots is mainly occupied by AGN, so a tentative AGN label can be assigned to these Unknowns. These sources tend to be the fainter sources, where there are fewer spectroscopic observations for the classifier to learn from. When plotted on an AllWISE colour–colour diagram, such as the one seen in Fig. 11, the Unknown sources (P_{class} > 80 per cent) that are bright enough to be detected in the *W*3 band (12 sources) tend to concentrate where galaxies are expected, with a few falling into the region where the AGN area overlaps the galaxy area. It is possible that these are obscured AGNs, which the training set for AGNs would have been biased against. Ultraluminous

Table 8. The number of sources per class that were cross-matched with *XMM–Newton* within the positional error of each source's coordinates.

Class	All	60 per cent <P_{class} <80 per cent	P_{class} >80 per cent
SMC	1607	381	627
Unk	823	235	411
AGN	364	109	118
Galaxy	25	6	4
RGB	103	23	39
OB	173	3	36
H II/YSO	9	0	0
AGB	17	0	1
PNe	44	0	1
RSG	12	1	1
Post-RGB/AGB	4	2	0
PM	33	2	16
LMC	10 139	1499	4794
Unk	7157	1066	4127
AGN	1277	324	436
Galaxy	84	17	7
RGB	192	30	105
OB	278	11	65
H II/YSO	213	21	4
AGB	727	4	7
PNe	14	0	0
RSG	13	0	2
Post-RGB/AGB	0	0	0
PM	184	26	41

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

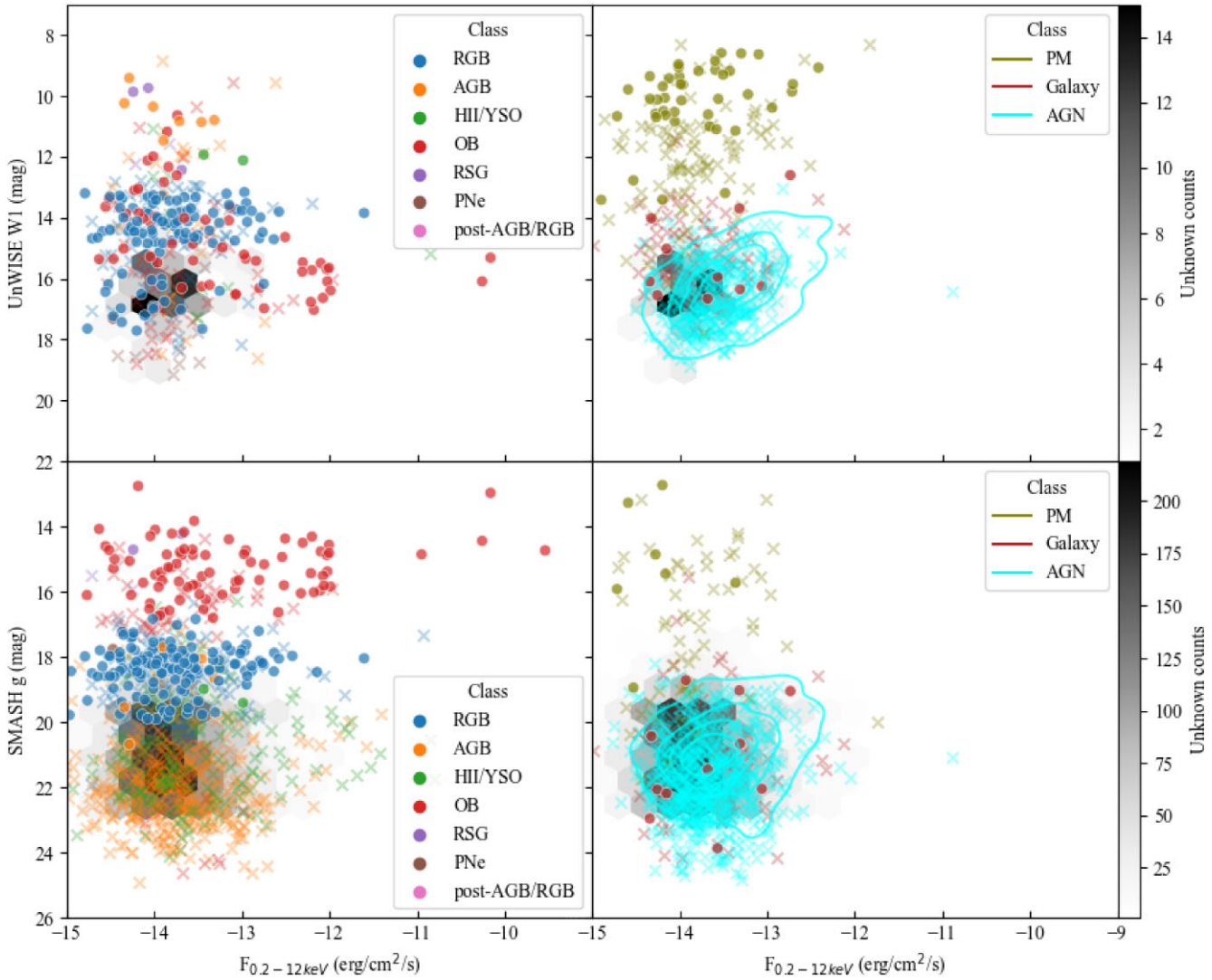

Figure 14. X-ray flux versus unWISE *W*1 (top) and SMASH *g* (bottom) bands for the Magellanic (left) and extragalactic and foreground sources (right). The high confidence ($P_{\rm class}$ > 80 per cent) Unknowns are plotted in the background of all plots. Contours and filled circles refer to sources with $P_{\rm class}$ > 80 per cent, whilst crosses refer to sources with $P_{\rm class}$ < 80 per cent. The contours represent a probability distribution in intervals of 0.2.

infrared galaxies (ULIRGs) tend to be found in the lower right corner where AGNs and galaxies overlap, which supports that these sources are more obscured.

The X-ray catalogue for the SMC from Sturm et al. (2013) also provides classifications within their catalogue. PRF classed OB stars that have cross-matches within the SMC X-ray catalogue indicate that sources classed as OB by the PRF are a mix of foreground stars and high-mass X-ray binaries (HMXB), which appear to separate in the unWISE *W*1 band, with foreground stars tending to be brighter, and HMXBs fainter and closer to extragalactic magnitudes. The OB stars that are classed as HMXB in Sturm et al. (2013), amongst which is SMC X-1, are grouped with the other stellar sources in optical, however. This could indicate that X-ray sources that are grouped with the extragalactic at IR magnitudes but grouped with the stellar sources at optical magnitudes could be given a tentative HMXB classification.

Furthermore, 53 out of 54 HMXBs found in the LMC (Maitra et al. 2019, 2021a, b; Haberl et al. 2022, 2023) were also classified as OB by the PRF. These inlcude the well-known LMC X-1 and LMC X-4 with $P_{\rm OB}$ > 96 per cent. Of these sources, those with the lowest probabilities (<70 per cent) for being an OB star were found to have a 10–20 per cent probability of being a YSO or PNe (LMC X-3, 4XMM J052546.5-694451 and 4XMM J052417.1−692533). Of these three sources, LMC X-3 is atypical as it is a Roche Lobe filling Black hole binary in the LMC with a B3V/B5V companion (e.g. Cowley et al. 1983; Soria et al. 2001). 4XMM J052546.5-694451 and 4XMM J052417.1−692533 are questionable HMXB candidates due to lack of expected Hα emission from a decretion disc (van Jaarsveld et al. 2018), which makes their lower confidence classifications not unexpected. The source that could not be classified, RX J0512.6−6717, was a ROSAT selected candidate and had a large position error circle of 7 arcsec, within which the *Gaia* counterpart could not be identified. However, cross-matching with the PRF catalogue with a matching radius 7 arcsec gave 11 sources, one of which (at 78.17215, −67.28993) was classed as an OB star at $P_{\rm Class}$ ∼ 88 per cent. If confirmed to be the true counterpart, this shows the PRF has the potential to match X-ray detections to their optical/near-IR counterparts.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

The true identity of the sources classed as RGB stars remains in contention. RGB stars should not have observable X-ray emission at Magellanic distances, so their inclusion in the X-ray–detected sources is questionable. As with the radio detected sources, it is possible that if the RGB class is the correct classification, then they are not the true counterpart. In Fig. 14, the RGB-classified sources are concentrated between the stellar and extragalactic, which also happens to be the location of some of the few galaxies with X-ray emission. It is possible that these RGB-classified sources are actually red dusty star-forming galaxies. However, when looking at the distribution at separation between VMC and *XMM–Newton* coordinates it can be seen that the AGNs have a median of 0.59 and 0.71 arcsec for the SMC and LMC, respectively, whilst RGBs display a median of 0.97 and 1.22 arcsec, which might indicate that these RGBs could be merely spurious alignments.

Overall, the X-ray–detected sources are classified as expected, and that the sources classified as Unknown which have a corresponding X-ray detection can be a tentative extragalactic/AGN classification. We have also shown that sources classified by the PRF with an associated X-ray detection are HMXBs, and that the PRF classifier has great potential to find the optical/near-IR counterparts to X-ray detections.

5.6 Quaia comparison

The Quaia survey (Storey-Fisher et al. 2023) is an all-sky spectroscopic quasar catalogue that used low-resolution BP/RP spectra from *Gaia* to identify AGN candidates, and has had cuts applied based on *Gaia* brightness ($G < 20.5$ mag), proper motions and unWISE colours to obtain a purer sample. Cross-matching Quaia with the VMC survey with a cross-matching radius of 1 arcsec yields 4325 and 1906 sources in the LMC and SMC, respectively. This leaves 34 and 5 sources in Quaia that are inside the LMC and SMC VMC survey footprints, respectively, that were not matched with a source in the VMC. It is surprising that sources that are $G < 20.5$ mag are not picked up in the deeper VMC catalogue.

Of the sources in the VMC survey, 427 were in the training set. 422 of these were known AGN and two were known galaxies. However, there was one known pAGB/RGB from the LMC, SMP LMC 11, and one known AGB from the SMC, OGLE SMC-SC5 255936. These classes are known to be variable sources, which could have played a part in their misclassification in Quaia.

After removing the sources in the training sets there were 4072 and 1733 sources in the LMC and SMC left, respectively. The highest fraction of sources is, as expected, classed as AGN (3335/4072 and 1622/1694 for the LMC and SMC, respectively). Five sources from the LMC footprint are classed as galaxies, with the second likeliest classification of an AGN, so these are most likely galaxies with an AGN component.

There are four sources classed as AGB from the SMC and seven from the LMC. Seven of these have $P_{\rm class} > 80$ per cent of being an AGB. With the one known AGB being classed as an AGN by Quaia, it is possible these sources are actually AGB. The *Gaia* proper motions would not be as capable of separating the stars in the Magellanic Clouds as in the Milky Way, unWISE colours ($W1 - W2$, see Fig. 11) could have been in the AGN regime and the low-resolution spectra might not have provided good enough S/N to make a good classification. Furthermore, AGB are known to vary regularly and Simbad contains matches for 8/11 of the AGB, and they are all classed as variable stars.

There are also 18 sources classed as HII/YSOs in the Quaia sample. Only two have mid-confidence of being this class ($P_{\rm class}$ of $\sim$ 79 per cent and 75 per cent), the rest have $P_{\rm class} < 60$ per cent. The majority of these sources also have AGN as the next likeliest class.

Five sources in the SMC were classed as RGB stars with $P_{\rm class} <$ 80 per cent, with five sources with <60 per cent. One source in the LMC was classed as an OB star but this had a $P_{\rm class} < 60$ per cent, so is unlikely to be the true class.

Lastly, 99 and 709 sources from the SMC and LMC, respectively, were classed as Unknown, suggesting that there is possibly a population of AGN that are being missed by the classifier. The sources the PRF classed as AGN had a magnitude range of $16.3 < G < 20.5$ mag, whilst the Unknown classed sources had a magnitude range of $18.8 < G < 20.5$ mag. This shows that the Unknown *Gaia* QSOs are fainter examples, for which there are less spectroscopically observed sources to train the classifier upon.

Looking at the Quaia sources that had no match in the PRF catalogues, VISTA images revealed that at the majority of the Quaia coordinates there are multiple sources in close proximity at the expected coordinates. If there is a galaxy underneath an AGN a larger separation is possible, as centres for galaxies are more difficult to establish. The majority (>90 per cent) of Quaia sources that matched with the VMC had a match within <0.2 arcsec. We increased the cross-matching radius to 5 arcsec and matched the Quaia sources with no match in the PRF catalogue again to find that all but one of the sources had matches between 1 and 1.8 arcsec. Amongst the matches are 15 AGN, 4 galaxies, 14 Unknown and 5 low confidence stellar classes. The majority of these classes being extragalactic implies that these are the true counterparts, though may not be as reliable due to the distance between counterparts.

Overall, the majority ($\sim$86 per cent) of AGN found in Quaia are found by the classifier. This is not unexpected, as these are most likely the more obvious type I (broad-line) AGN, which make up the most of the AGN class in the training set. The type II (narrow-line) AGNs, however, are harder to test against as spectroscopic samples tend to be biased towards type I, so the sources classed as unknown are more likely the type II AGN.

5.7 YSOs in the LMC

YSOs are dusty sources that can exhibit emission lines, making them easily confused with AGN, therefore ascertaining that YSOs are not being misclassified as AGN or vice versa by the PRF is necessary. A recent study by Kokusho et al. (2023) has compiled all the YSOs in the area of the LMC, numbering 4097, the majority of which are candidates located using photometry and SED fitting. The number of H II/YSOs the PRF detects is greater in the LMC than the SMC, and shows signs of structure [Fig. 15; see also Section D (Fig. D8) in the online Appendix for the SMC)].

Fig.15 shows an expected distribution, with easily identifiable structures such as 30 Dor and the Southern molecular ridge below it. N11 is seen to the right and most of the Henize H II regions scattered across the face of the LMC.

Cross-matching with the PRF results with a matching radius of 1 arcsec yields 2715 matches, and restricting to those not in the training set, numbers 2274. Further restricting to only those with $P_{\rm class} > 80$ per cent gives 630 sources.

Of the 630 high-confidence sources, 226 are classified as H II/YSOs and 117 are classed as Unknown, where the Unknown sources tend to be the fainter sources. 105 and 6 are classed as AGN and galaxies, respectively. The rest are classified as OB (58), AGB (49), PNe (8), RGB (5), post-AGB/RGB (1), and PM (1).

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

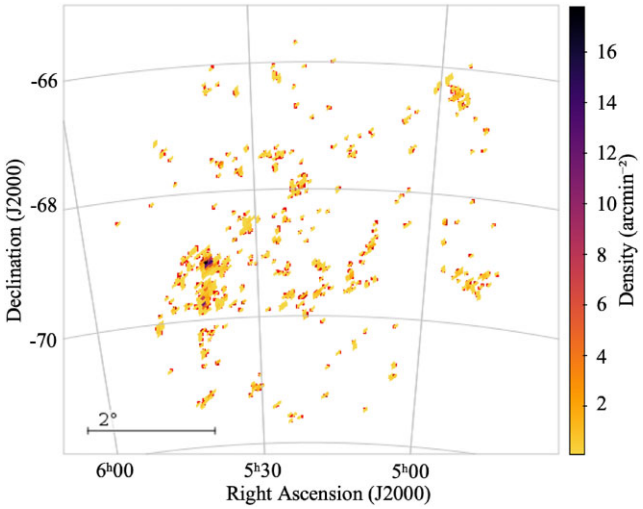

Figure 15. The sky density distribution of the H II/YSOs with $P_{\rm class} > 80$ per cent in the LMC.

About a sixth of the YSOs have been classed as an extragalactic class. It should, however, be noted that of the sources from this study that are also in the PRF training set (441 sources), some are spectroscopically confirmed to be other classes. 297 of them are confirmed as H II/YSOs. 47 and 11 are spectroscopically confirmed as AGNs and galaxies, respectively. The rest are classed as AGB (24), post-AGB/RGB (22), OB (20), PNe (18), RGB (1), and RSG (1). So, it is not unexpected that not all the sources classed as YSOs in Kokusho et al. (2023) are classed as such here.

5.8 The unknown class

The Unknown class represents all the sources that were not given a similar enough representative in the training set. The majority of these sources are the fainter sources that the photometry for is mostly missing and that we have little spectroscopy for due to the limits of ground based spectroscopy instruments. There are, however, still brighter sources amongst the Unknowns, that are made up of classes and/or subclasses that were not accounted for in the training sets due to lack of available spectroscopy.

The distribution of the Unknown class can be seen in the online Appendix D in Fig. C10. From this it can be seen that the Unknowns cover the entirety of the VMC LMC field, and the area covered by the SMASH survey for the VMC SMC field. For the SMC field this implies that the confident Unknown sources are those with coverage from most if not all the photometric surveys, indicating that these sources, specifically the brighter sources with fewer missing data, do not have a match in the training set, and that they are classes that are not trained on, rather than not having enough information.

Selection criteria for selecting stars of different ages using colour–magnitude diagrams have been created for the SMC and LMC (Cioni et al. 2014, 2016; El Youssoufi et al. 2019) for sources down to $K_s > 19.8$ mag and only those with a 70 per cent probability of being stars and with photometric uncertainties <0.1 mag. We apply the selection criteria from El Youssoufi et al. (2019) to the confident Unknowns to provide a tentative classification and to discover the most underrepresented classes in the training set for the Magellanic sources.

After applying the selection criteria to the Unknowns, 91 per cent and 71 per cent of sources for the LMC and SMC, respectively, were too faint for the selection cut. The division of the sources that were bright enough can be seen in Table 9. See online Appendix Section F for a figure showing the selection criteria plotted on top of the Unknown distribution in near-IR CMD space.

Table 9. The number of Unknowns with $K_s < 19.8$ mag (Vega) and $P_{\rm class} > 80$ per cent that are categorized as different stellar populations based on the near-IR CMD selection from El Youssoufi et al. (2019), for only those sources with a 70 per cent probability of being stars and with photometric uncertainties <0.1 mag. Note that 'All' represents all sources with $K_s > 19.8$ mag. The table of sources with the VMC colour–magnitude classification will be made public with the final VMC public release.

Region	Dominant stellar Population	SMC	LMC
All		875 233	13 077 735
A	Main sequence	34 032 (4 per cent)	200 596 (2 per cent)
B	Main sequence	136 738 (16 per cent)	886 994 (7 per cent)
C	Main sequence	150 780 (17 per cent)	2216 757 (17 per cent)
D	Main sequence and subgiants	132 100 (15 per cent)	3063 970 (23 per cent)
E	RGB	306 060 (35 per cent)	4459 540 (34 per cent)
F	Milky Way	43 686 (5 per cent)	532 015 (4 per cent)
G	Supergiants and giant stars	1 (<1 per cent)	0
H	Supergiants and giant stars	0 (<1 per cent)	16 (<1 per cent)
I	Supergiants and giant stars	21 (<1 per cent)	32 360 (<1 per cent)
J	Red clump stars	420 (<1 per cent)	1325 249 (10 per cent)
K	RGB	7 (<1 per cent)	296 100 (2 per cent)
L	Extragalactic	71 378 (8 per cent)	64 138 (<1 per cent)

From Table 9, we can see that, as expected, the main-sequence stars (populations A, B, C, and D) make up the majority of the Unknowns, most likely due to their lack of corresponding class in the training sets, where only O and B main sequence stars are accounted for. RGBs (populations E and K) are the next largest population in the Unknowns, despite the majority of sources already being classified as RGB by the PRF, implying that we are not capturing the full scope of the RGB star class with our training sets for the PRF, despite it already being the largest class.

Population L is where extragalactic sources can be found. From Table 9, it can be seen that ∼135 000 sources across the Clouds can be given a tentative extragalactic classification. Additional information would be required to further separate the sources into AGN and galaxies.

Another way of giving a tentative classification to sources is by using wavelengths that were not used in the PRF. A source being associated with an X-ray or radio detection tends to imply an extragalactic source rather than a stellar source, as seen in Sections 5.5 and 5.4. This means we can give a tentative extragalactic classification, though with the caveat that cross-matching can lead to contamination from spurious alignments with Magellanic or foreground stars and that there are some stellar objects that do emit in the X-ray and radio, though they tend to be easy to pick out in the X-ray.

To improve the success of this machine learning technique, more spectroscopy is needed to bolster the training set (for both stellar and extragalactic populations), especially for fainter sources and

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

those detected in IR but not in the optical, amongst which would be higher redshift galaxies/AGN, as well as heavily dust reddened sources. Therefore, spectroscopy from telescopes such as the *JWST* (Gardner et al. 2006), 4MOST (de Jong et al. 2012), and WEAVE (Dalton et al. 2012) would provide greater potential in identifying the fainter extragalactic population. The training set could potentially be augmented to reduce selection bias by using simulations/models, or by taking observational data of high-redshift sources from a deep field survey, modelling the spectral energy distribution, and then estimating what the surveys used in the area of the VMC would measure in each waveband. Another method would be to take the known sources and model how they would look at fainter magnitudes to regain some of the fainter sources from the Unknown class.

In terms of features, we have selected the best current survey data available with unique contributions. The VISTA data is deep and is complemented by the optical SMASH data, though the SMASH survey does not cover the entirety of the VMC fields (see Fig. 1). The future Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) would be an improvement on the SMASH data, as it would cover all of the VMC field and would reach comparable depths in its first data release. Furthermore, there is no current complementary mid-IR photometry that reaches the same depths. Mid-IR is a wavelength range that is particularly powerful in identifying AGN due to being sensitive to the emission from the dust surrounding the accretion disc. Therefore, photometry from an IR telescope that reaches greater depths than *WISE* and *Spitzer* would provide greater potential in identifying the extragalactic.

6 CONCLUSIONS

In summary, we trained a probabilistic random forest on the UV–IR photometry of spectroscopically observed sources in the field of the VMC survey, augmented with AGN and galaxies from the SDSS observations of the GAMA09 field. This yielded overall accuracies of 0.79 ± 0.01 for the SMC classifier, and 0.87 ± 0.01 for the LMC classifier. For the extragalactic classifications the classifiers yielded accuracies of 0.93 ± 0.01 for both classifiers. When restricted to $P_{\rm class} > 80$ per cent the accuracy of the classifiers were 0.98 ± 0.01 and 0.90 ± 0.01 for the LMC and SMC, respectively.

The classifiers were used on the entirety of the LMC and SMC PSF catalogues and the sources were separated into three catalogues with different ranges of probabilities of the classification being correct (low-confidence – $P_{\rm class} < 60$ per cent, mid-confidence – 60 per cent $< P_{\rm class} < 80$ per cent, high-confidence – $P_{\rm class} > 80$ per cent). At high confidence, we classify a total of 707 939 and 397 899 sources in the SMC and LMC, respectively, with a total of 50 507 AGNs and 27 146 galaxies (>49 500 and >26 500 of which, respectively, are new candidates) across the two Clouds. Looking at the high-confidence classifications, we find

(i) The spatial distributions of the different classes across the VMC fields of the SMC and LMC are as expected. The extragalactic and foreground sources being mostly homogeneous across the field and the Magellanic sources concentrating in the centres of the Clouds with fewer sources towards the edges of the fields.

(ii) We tested the classifiers on stellar and extragalactic sources that are known to be confused with each other from Pennock et al. (2022). The results showed that all the AGNs were classified correctly. We showed that even sources that are often confused with another class are well classified by the classifiers. However, emission-line stars have the possibility of being classed as AGN, though not necessarily with high confidence.

(iii) We tested the behaviour of the classifiers on classes which they have not trained upon (65 sources from the SAGE-spec catalogues; Ruffle et al. 2015; Jones et al. 2017). From this, we found that for all stellar sources the classifiers classified them as another stellar class. For the two sources that were classified as an AGN their spectral classification was not known so they may have been AGN. This means that stellar classes that we have not trained upon are unlikely to be misclassified as extragalactic sources.

(iv) Plotting the sources across optical, near-IR and mid-IR colour–colour and colour–magnitude diagrams showed that the classes separated as expected, and that where the classes do overlap the classifiers are still able to discern between the different classes. This shows that the large array of features from optical to far-IR is helping to separate sources that would have been otherwise hard to untangle in single colour–colour/magnitude diagrams.

(v) Investigating the sources that had a corresponding ASKAP 888/960/1320 MHz radio or *XMM–Newton* X-ray detection showed that, as expected, the majority of the radio/X-ray detected sources were (when restricting to sources not classed as Unknown) predominantly classed as extragalactic (∼89 per cent and ∼64 per cent, respectively).

(vi) The proportions of radio AGNs and galaxies were found to vary with radio flux density. The brightest flux densities are dominated by AGN, then towards lower flux densities the fraction of galaxies increases as we start to pick up fainter emission from star-formation from galaxies, as expected. Unexpectedly, at about $F_{\rm 888MHz}$, $F_{\rm 1320MHz} < 7$ mJy, the number of galaxies compared to AGN starts decreasing again. We expect this could be accounted for by selection bias, where the faint AGN population is easier to classify than the faint galaxy population.

(vii) Quaia survey (Storey-Fisher et al. 2023) AGN candidates were predominantly classified as AGN (∼ 85 per cent), as expected. Only ∼14 per cent of the Quaia AGN candidates were classed as Unknown. This implies that the sample of AGN in the Quaia catalogue are well represented by the AGN training sample for the PRF classifiers, which are mostly made up of the bright broad-line AGN. Those classed as Unknown are possibly the underrepresented narrow-line region AGN.

(viii) VMC near-IR colour–colour magnitude diagrams of the brightest Unknown sources ($K_s < 19.8$ mag) revealed that the main classes missing from the classifier are main-sequence stars and fainter examples (than are currently in the training set) of Milky Way stars (PM) and RGB stars.

(ix) It is also possible to give a tentative extragalactic classification to Unknowns that have X-ray or radio counterparts. Plotting unWISE *W*1 and SMASH *g* bands against *XMM–Newton* flux showed that the majority of the Unknowns lie in the regions in these plots occupied by AGNs and galaxies. However, the possibility of spurious alignments does lower the reliability of this.

The majority (∼71 per cent for all sources, ∼98 per cent for sources with $P_{\rm class} > 80$ per cent) of sources are classed as Unknown. Whilst some of this is due to some missing classes such as main-sequence stars (other than O and B types) and fainter Milky Way stars, this is mostly due to these sources being fainter than the spectroscopically observed examples we provided the classifier to be trained upon. Therefore more spectroscopy of fainter sources is required. For the sources that are within the brightness range of the training set, but were still classed as Unknown, more classes are required, such as main sequence stars that are not of type O or B, as well as fainter examples of Milky Way stars. Deeper photometry from surveys such as optical LSST, as well as a complimentary mid-IR

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

survey, would be preferable to reduce the amount of faint sources in the VMC catalogue with the majority of features missing. However, without the spectroscopy for sources at fainter magnitudes to train upon, the majority will remain Unknown.

ACKNOWLEDGEMENTS

We thank the anonymous referee for their feedback, which helped improve the paper. CMP and acknowledges funding from an STFC (Science and Technology Facility Council) studentship and from a UKRI (UK Research and Innovation) Future Leaders Fellowship (grant: MR/T020989/1). JEMC acknowledges funding from an STFC studentship. This research was supported in part by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 682115). We thank the Cambridge Astronomy Survey Unit (CASU) and the Wide Field Astronomy Unit (WFAU) in Edinburgh for providing the necessary data products under the support of the Science and Technology Facility Council (STFC) in the UK.

This paper uses observations made at the SAAO 1.9-m under programme Pennock-2019-05-74-inch-257, and with the Southern African Large Telescope (SALT) under programmes 2021-1-SCI-018 (PI: van Loon), 2021-1-SCI-029 (PI: van Loon), 2021-1-SCI-032 (PI: van Loon) and 2021-2-SCI-017 (PI: Anih). We would like to thank Francois van Wyk for his assistance in acquiring the observations at the 1.9-m telescope during late-COVID lockdown.

This research made use of ASTROPY,[8] a community-developed core PYTHON package for Astronomy (Astropy Collaboration 2013, 2018). We have made extensive use of the SIMBAD Database at CDS (Centre de Données astronomiques) Strasbourg, the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, CalTech, under contract with NASA, and of the VizieR catalog access tool, CDS, Strasbourg, France.

Whilst we here refer to the 'Magellanic' Clouds as their common names, we acknowledge and disapprove of the associated colonial heritage of which the inequities and disparities still endure today.

DATA AVAILABILITY

The results table of source coordinates and corresponding classifications, as well as the VMC colour–magnitude classifications of the Unknowns, will be made available on the VSA[9] (VISTA Science Archive) ESO archive[10] as part of VMC DR7, as well as on CDS[11] (Centre de Données astronomiques) when the paper is published. The training data for the SMC and LMC classifiers and the table of feature importances for the two classifiers created in this work are available as supplementary material to this article.

REFERENCES

Ahumada R. et al., 2020, ApJS, 249, 3
Assef R. J. et al., 2013, ApJ, 772, 26
Assef R. J., Stern D., Noirot G., Jun H. D., Cutri R. M., Eisenhardt P. R. M., 2018, ApJS, 234, 23
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Bell C. P. M. et al., 2019, MNRAS, 489, 3200
Bell C. P. M. et al., 2020, MNRAS, 499, 993
Bell C. P. M. et al., 2022, MNRAS, 516, 824
Blanton M. R. et al., 2017, AJ, 154, 28
Breiman L., 2001, Machine Learning, 45 , 5
Buckley D. A. H., Swart G. P., Meiring J. G., 2006, in Stepp L. M., ed., Proc. SPIE Conf. Ser. Vol. 6267, Ground-based and Airborne Telescopes. SPIE, Bellingham, p. 62670Z
Burgh E. B., Nordsieck K. H., Kobulnicky H. A., Williams T. B., O'Donoghue D., Smith M. P., Percival J. W., 2003, in Iye M., Moorwood A. F. M., eds, Proc. SPIE Conf. Ser. Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes. SPIE, Bellingham, p. 1463
Carrera R., Conn B. C., Noël N. E. D., Read J. I., López Sánchez Á. R., 2017, MNRAS, 471, 4571
Choudhury S. et al., 2021, MNRAS, 507, 4752
Cioni M. R. L. et al., 2011, A&A, 527, A116
Cioni M. R. L. et al., 2013, A&A, 549, A29
Cioni M.-R. L. et al., 2014, A&A, 562, A32
Cioni M.-R. L. et al., 2016, A&A, 586, A77
Civano F. et al., 2015, ApJ, 808, 185
Cole A. A., Tolstoy E., Gallagher John S. I., Smecker-Hane T. A., 2005, AJ, 129, 1465
Cowley A. P., Crampton D., Hutchings J. B., Remillard R., Penfold J. E., 1983, ApJ, 272, 118
Crause L. A. et al., 2019, J. Astron. Telesc. Instrum. Syst., 5, 024007
Crawford S. M. et al., 2010, in Silva D. R., Peck A. B., Soifer B. T., eds, Proc. SPIE Conf. Ser. Vol. 7737, Observatory Operations: Strategies, Processes, and Systems III. SPIE, Bellingham, p. 773725
Cusano F. et al., 2021, MNRAS, 504, 1
Cutri R. M. et al., 2013, Explanatory Supplement to the AllWISE Data Release Products. Available at: http://wise2.ipac.caltech.edu/docs/release/allwise/expsup/
Dalton G. B. et al., 2006, in McLean I. S., Iye M., eds, Proc. SPIE Conf. Ser. Vol. 6269, Ground-based and Airborne Intrumentation for Astronomy. SPIE, Bellingham, p. 62690X
Dalton G. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. SPIE, Bellingham, p. 84460P
De Bortoli B. J., Parisi M. C., Bassino L. P., Geisler D., Dias B., Gimeno G., Angelo M. S., Mauro F., 2022, A&A, 664, A168
de Jong J. T. A. et al., 2017, A&A, 604, A134
de Jong R. S. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. SPIE, Bellingham, p. 84460T
DeBoer D. R. et al., 2009, IEEE Proc., 97, 1507
Dickey J. M. et al., 2013, PASA, 30, e003
Dorigo Jones J., Oey M. S., Paggeot K., Castro N., Moe M., 2020, ApJ, 903, 43
Driver S. P. et al., 2011, MNRAS, 413, 971
El Youssoufi D. et al., 2019, MNRAS, 490, 1076
Emerson J., McPherson A., Sutherland W., 2006, Messenger, 126, 41
Esquej P. et al., 2013, A&A, 557, A123
Evans C. J. et al., 2015a, A&A, 574, A13
Evans C. J., van Loon J. T., Hainich R., Bailey M., 2015b, A&A, 584, A5
Finkbeiner D. P. et al., 2004, AJ, 128, 2577
Flesch E. W., 2015, PASA, 32, e010
Flesch E. W., 2019, preprint (arXiv:1912.05614)
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2021a, A&A, 649, A1
Gaia Collaboration, 2021b, A&A, 649, A7
Gaia Collaboration, 2023, A&A, 674, A41
Gardner J. P. et al., 2006, Space Sci. Rev., 123, 485
Geha M. et al., 2003, AJ, 125, 1
Gordon K. D., Meixner M., Meade M., Whitney B. A., Engelbracht C. W., Bot C., 2011, AJ, 142, 102
Griffin M. J. et al., 2010, A&A, 518, L3

[8] http://www.astropy.org
[9] http://vsa.roe.ac.uk
[10] http://archive.eso.org
[11] https://cds.unistra.fr

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

Grin N. J. et al., 2017, A&A, 600, A82
Groenewegen M. A. T., Blommaert J. A. D. L., 1998, A&A, 332, 25
Groenewegen M. A. T. et al., 2019, A&A, 622, A63
Groenewegen M. A. T. et al., 2020, A&A, 636, A48
Gullieuszik M. et al., 2012, A&A, 537, A105
Haberl F., Maitra C., Vasilopoulos G., Maggi P., Udalski A., Monageng I. M., Buckley D. A. H., 2022, A&A, 662, A22
Haberl F. et al., 2023, A&A, 671, A90
Hamuy M., Suntzeff N. B., Heathcote S. R., Walker A. R., Gigoux P., Phillips M. M., 1994, PASP, 106, 566
Hickox R. C., Alexander D. M., 2018, ARA&A, 56, 625
Hony S. et al., 2011, A&A, 531, A137
Hopkins A. M. et al., 2013, MNRAS, 430, 2047
Hornschemeier A. E. et al., 2001, ApJ, 554, 742
Hotan A. W. et al., 2014, PASA, 31, e041
Hotan A. W. et al., 2021, PASA, 38, e009
Ivanov V. D. et al., 2016, A&A, 588, A93
Ivanov V. D. et al., 2024, A&A, 687, A16
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jarvis M. J. et al., 2013, MNRAS, 428, 1281
Johnston S. et al., 2008, Exp. Astron., 22, 151
Jones D. H. et al., 2009, MNRAS, 399, 683
Jones H., Saunders W., Colless M., Read M., Parker Q., Watson F., Campbell L., 2005, in Fairall A. P., Woudt P. A., eds, ASP Conf. Ser. Vol. 329, Nearby Large-Scale Structures and the Zone of Avoidance. Astron. Soc. Pac., San Francisco, p. 11
Jones O. C. et al., 2017, MNRAS, 470, 3250
Joseph T. D. et al., 2019, MNRAS, 490, 1202
Kamath D., Wood P. R., Van Winckel H., 2014, MNRAS, 439, 2211
Khoshgoftaar T. M., Golawala M., Hulse J. V., 2007, in 19th IEEE International Conference on Tools with Artificial Intelligence, (ICTAI 2007). IEEE, Patras, Greece, p. 310
Kinson D. A., Oliveira J. M., van Loon J. T., 2021, MNRAS, 507, 5106
Kinson D. A., Oliveira J. M., van Loon J. T., 2022, MNRAS, 517, 140
Kobulnicky H. A., Nordsieck K. H., Burgh E. B., Smith M. P., Percival J. W., Williams T. B., O'Donoghue D., 2003, in Iye M., Moorwood A. F. M., eds, Proc. SPIE Conf. Ser. Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes. SPIE, Bellingham, p. 1634
Kokusho T., Torii H., Kaneda H., Fukui Y., Tachihara K., 2023, ApJ, 953, 104
Kozłowski S., Kochanek C. S., Udalski A., 2011, ApJS, 194, 22
Kozłowski S. et al., 2012, ApJ, 746, 27
Kozłowski S. et al., 2013, ApJ, 775, 92
Kumar A. et al., 2012, Proc. SPIE Conv. Ser. Vol. 8443, Space Telescopes and Instrumentation 2012: Ultraviolet to Gamma Ray. SPIE, Bellingham, p. 84431N
Lacy M. et al., 2004, ApJS, 154, 166
Lamb J. B., Oey M. S., Segura-Cox D. M., Graus A. S., Kiminki D. C., Golden-Marx J. B., Parker J. W., 2016, ApJ, 817, 113
Maitra C. et al., 2019, MNRAS, 490, 5494
Maitra C., Haberl F., Maggi P., Kavanagh P. J., Vasilopoulos G., Sasaki M., Filipović M. D., Udalski A., 2021a, MNRAS, 504, 326
Maitra C., Haberl F., Vasilopoulos G., Ducci L., Dennerl K., Carpano S., 2021b, A&A, 647, A8
Mauch T., Murphy T., Buttery H. J., Curran J., Hunstead R. W., Piestrzynski B., Robertson J. G., Sadler E. M., 2003, MNRAS, 342, 1117
McConnell D. et al., 2016, PASA, 33, e042
McConnell D. et al., 2020, PASA, 37, e048
Meixner M. et al., 2006, AJ, 132, 2268
Meixner M. et al., 2010, A&A, 518, L71
More A. S., Rana D. P., 2017, in 1st International Conference on Intelligent Systems and Information Management (ICISIM). IEEE, India, p. 72
Murphy T. et al., 2010, MNRAS, 402, 2403
Nandra K. et al., 2015, ApJS, 220, 10
Netzer H., 2015, ARA&A, 53, 365
Neugent K. F., Levesque E. M., Massey P., Morrell N. I., Drout M. R., 2020, ApJ, 900, 118
Nidever D. L. et al., 2017, AJ, 154, 199
Nikutta R., Hunt-Walker N., Nenkova M., Ivezić v., Elitzur M., 2014, MNRAS, 442, 3361
Noël N. E. D., Conn B. C., Carrera R., Read J. I., Rix H. W., Dolphin A., 2013, ApJ, 768, 109
Noël N. E. D., Conn B. C., Read J. I., Carrera R., Dolphin A., Rix H. W., 2015, MNRAS, 452, 4222
Norris R. P., Hopkins A. M., Afonso J., Brown S., Condon J. J., 2011, PASA, 28, 215
Oliveira J. M. et al., 2011, MNRAS, 411, L36
Oliveira J. M. et al., 2013, MNRAS, 428, 3001
Oliveira J. M. et al., 2019, MNRAS, 490, 3909
Padovani P. et al., 2017, A&AR, 25, 2
Parisi M. C., Grocholski A. J., Geisler D., Sarajedini A., Clariá J. J., 2009, AJ, 138, 517
Parisi M. C., Geisler D., Grocholski A. J., Clariá J. J., Sarajedini A., 2010, AJ, 139, 1168
Parisi M. C., Gramajo L. V., Geisler D., Dias B., Clariá J. J., Da Costa G., Grebel E. K., 2022, A&A, 662, A75
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pennock C. M. et al., 2021, MNRAS, 506, 3540
Pennock C. M. et al., 2022, MNRAS, 515, 6046
Pilbratt G. L. et al., 2010, A&A, 518, L1
Poglitsch A. et al., 2010, A&A, 518, L2
Reid W. A., Parker Q. A., 2012, MNRAS, 425, 355
Reis I., Baron D., Shahaf S., 2018, AJ, 157, 1
Rezaeikh S., Javadi A., Khosroshahi H., van Loon J. T., 2014, MNRAS, 445, 2214
Ripepi V. et al., 2015, MNRAS, 446, 3034
Roman-Duval J. et al., 2019, ApJ, 871, 151
Rubele S. et al., 2012, A&A, 537, A106
Rubele S. et al., 2015, MNRAS, 449, 639
Rubele S. et al., 2018, MNRAS, 478, 5017
Ruffle P. M. E. et al., 2015, MNRAS, 451, 3504
Salvato M. et al., 2018, MNRAS, 473, 4937
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Schlafly E. F., Meisner A. M., Green G. M., 2019, ApJS, 240, 30
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Schumacher G., Mondaca E., Warner M., Martinez M., Estay O., Abbott T. M. C., 2010, in Radziwill N. M., Bridger A., eds, Proc. SPIE Conf. Ser. Vol. 7740, Software and Cyberinfrastructure for Astronomy. SPIE, Bellingham, p. 77402H
Seale J. P., Looney L. W., Chu Y.-H., Gruendl R. A., Brandl B., Chen C. H. R., Brandner W., Blake G. A., 2009, ApJ, 699, 150
Secrest N. J., Dudik R. P., Dorland B. N., Zacharias N., Makarov V., Fey A., Frouard J., Finch C., 2015, ApJS, 221, 12
Shaw R. A., Stanghellini L., Mutchler M., Balick B., Blades J. C., 2001, ApJ, 548, 727
Sheets H. A., Bolatto A. D., van Loon J. T., Sandstrom K., Simon J. D., Oliveira J. M., Barbá R. H., 2013, ApJ, 771, 111
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Soria R., Wu K., Page M. J., Sakelliou I., 2001, A&A, 365, L273
Stern D., Eisenhardt P., Gorjian V., Kochanek C., 2005, ApJ, 631, 163
Stern D. et al., 2012, ApJ, 753, 30
Storey-Fisher K., Hogg D. W., Rix H.-W., Eilers A.-C., Fabbian G., Blanton M., Alonso D., 2023, ApJ, 964, 69
Sturm R. et al., 2013, A&A, 558, A3
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29
Thilker D. A., Bianchi L., Simons R., 2014, American Astronomical Society, AAS Meeting #223, Washington, DC, p. 355.11
Tody D., 1986, in Crawford D. L., ed., Proc. SPIE Conf. Ser. Vol. 627, Instrumentation in Astronomy VI. SPIE, Bellingham, p. 733
Tody D., 1993, in Hanisch R. J., Brissenden R. J. V., Barnes J., eds, ASP Conf. Ser. Vol. 52, Astronomical Data Analysis Software and Systems II. Astron. Soc. Pac., San Francisco, p. 173
van Gelder M. L. et al., 2020, A&A, 636, A54
van Loon J. T., Sansom A., 2015, MNRAS, 453, 2342

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

van Jaarsveld N., Buckley D. A. H., McBride V. A., Haberl F., Vasilopoulos G., Maitra C., Udalski A., Miszalski B., 2018, MNRAS, 475, 3253
van Loon J. T. et al., 1998, A&A, 329, 169
van Loon J. T., Zijlstra A. A., Groenewegen M. A. T., 1999a, A&A, 346, 805
van Loon J. T., Groenewegen M. A. T., de Koter A., Trams N. R., Waters L. B. F. M., Zijlstra A. A., Whitelock P. A., Loup C., 1999b, A&A, 351, 559
van Loon J. T., Cioni M. R. L., Zijlstra A. A., Loup C., 2005, A&A, 438, 273
van Loon J. T., Marshall J. R., Cohen M., Matsuura M., Wood P. R., Yamamura I., Zijlstra A. A., 2006, A&A, 447, 971
van Loon J. T., Cohen M., Oliveira J. M., Matsuura M., McDonald I., Sloan G. C., Wood P. R., Zijlstra A. A., 2008, A&A, 487, 1055
Vasiliev E., 2018, MNRAS, 481, L100
Walborn N. R. et al., 2014, A&A, 564, A40
Webb N. A. et al., 2020, A&A, 641, A136
Wenger M. et al., 2000, A&AS, 143, 9
Whiting M. T., 2020, in Ballester P., Ibsen J., Solar M., Shortridge K., eds, ASP Conf. Ser. Vol. 522, Astronomical Data Analysis Software and Systems XXVII. SPIE, Bellingham, p. 469
Wright E. L. et al., 2010, AJ, 140, 1868
Zivkov V. et al., 2018, A&A, 620, A143
Zivkov V. et al., 2020, MNRAS, 494, 458

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

suppl_data

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article/537/2/1028/7954762 by Macquarie University user on 17 March 2025

© 2025 The Author(s).
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.