Astronomy
&
Astrophysics

Galaxies in the zone of avoidance: Misclassifications using machine learning tools

P. Marchant Cortés[1], J. L. Nilo Castellón[1], M. V. Alonso[2,3], L. Baravalle[2,3], C. Villalon[2], M. A. Sgró[2,3], I. V. Daza-Perilla[2,3], M. Soto[4], F. Milla Castro[1], D. Minniti[5,6,7], N. Masetti[8,5], C. Valotto[2,3], and M. Lares[2,3]

[1] Departamento Astronomía, Facultad de Ciencias, Universidad de La Serena, Av. Juan Cisternas 1200, La Serena, Chile
e-mail: p.marchantcortes.9@gmail.com
[2] Instituto de Astronomía Teórica y Experimental (IATE-CONICET), Laprida 854, X5000BGR Córdoba, Argentina
[3] Observatorio Astronómico de Córdoba, Universidad Nacional de Córdoba, Laprida 854, X5000BGR Córdoba, Argentina
[4] Instituto de Investigación en Astronomía y Ciencias Planetarias, Universidad de Atacama, Copayapu 485, Copiapó, Chile
[5] Instituto de Astrofísica, Facultad de Ciencias Exactas, Universidad Andrés Bello, Av. Fernandez Concha 700, Las Condes, Santiago, Chile
[6] Vatican Observatory, 00120 Vatican City State, Italy
[7] Departamento de Física, Universidade Federal de Santa Catarina, Trinidade 88040-900, Florianopolis, Brazil
[8] INAF – Osservatorio di Astrofisica e Scienza dello Spazio, Via Piero Gobetti 101, 40129 Bologna, Italy

Received 16 November 2023 / Accepted 28 February 2024

ABSTRACT

Context. Automated methods for classifying extragalactic objects in large surveys offer significant advantages compared to manual approaches in terms of efficiency and consistency. However, the existence of the Galactic disk raises additional concerns. These regions are known for high levels of interstellar extinction, star crowding, and limited data sets and studies.
Aims. In this study, we explore the identification and classification of galaxies in the zone of avoidance (ZoA). In particular, we compare our results in the near-infrared (NIR) with X-ray data.
Methods. We analyzed the appearance of objects in the Galactic disk classified as galaxies using a published machine-learning (ML) algorithm and make a comparison with the visually confirmed galaxies from the VVV NIRGC catalog.
Results. Our analysis, which includes the visual inspection of all sources cataloged as galaxies throughout the Galactic disk using ML techniques reveals significant differences. Only four galaxies were found in both the NIR and X-ray data sets. Several specific regions of interest within the ZoA exhibit a high probability of being galaxies in X-ray data but closely resemble extended Galactic objects. Our results indicate the difficulty in using ML methods for galaxy classification in the ZoA, which is mainly due to the scarcity of information on galaxies behind the Galactic plane in the training set. They also highlight the importance of considering specific factors that are present to improve the reliability and accuracy of future studies in this challenging region.

Key words. catalogs – surveys – infrared: galaxies – X-rays: galaxies

1. Introduction

Astronomy is moving forward as never before, with much of the progress being driven by the unprecedented amount of data produced by large surveys. New tools have begun to play an essential role in analyzing data, such as machine-learning (ML) algorithms, which have increased the efficiency with which we can identify commonalities across large databases and detect faint and complex patterns. These algorithms have become a common tool in astronomy because of the large amount of data coming from all-sky surveys, and have been well tested as classifiers for galaxy morphology (Spindler et al. 2021), young stellar object (YSO) finders (Marton et al. 2019), classifiers of variable stars using light curves (Aguirre et al. 2019), and estimators of photometric redshift (Dainotti et al. 2021), among many others.

All-sky surveys are also increasing in number at different wavelengths. This improves our understanding of the large-scale evolution and structure of the Universe, the formation of stars and galaxies, and the history of the Milky Way (MW). At lower Galactic latitudes, the data from these surveys are scarce. This region is known as the zone of avoidance (ZoA; Shapley 1961), and became more prominent as complete sky surveys increased

in number. It is more critical at $|b| < 10°$, where extragalactic sources and large-scale structure (LSS) behind the MW are obscured by dust and stellar crowding, dimming the sources by more than 25% in the optical and by about 10% in the infrared (IR) wavelengths (Henning et al. 1998). Classifying extragalactic objects within the ZoA is of critical importance. We have the opportunity to minimize this gap and explore the Local Universe in increasing detail. This endeavor is pivotal in defining the cosmography of the nearby Universe, which sheds light on the dynamics of the Local Group, giving us insights into the Universe at larger scales. It will also allow us to decipher various cosmological parameters, including the peculiar velocity of the Local Group, which exhibits a profound discrepancy relative to the cosmic microwave background dipole as seen in Loeb & Narayan (2008). In order to investigate the LSS, it is crucial to obtain a more complete redshift catalog to fill the ZoA gap in the distribution of the largest mass concentrations in the Local Universe, such as the Great Attractor (Kraan-Korteweg et al. 1996), the Perseus-Pisces Supercluster (Ramatsoku et al. 2016), and the Vela Supercluster (Kraan-Korteweg et al. 2017).

The use of near-infrared (NIR) wavelengths together with radio and X-rays has led to a new wave of extragalactic studies

in this region. With improvements in NIR cameras, it has been possible to enlarge the number of discoveries of extragalactic sources. The first photometric galaxy catalog in the ZoA was provided by the Two Micron All-Sky Survey (2MASS Extended survey, Skrutskie et al. 2006), which was carried out in order to collect the radial velocities of these galaxies (Macri et al. 2019).

More recently, the VISTA Variables in the Vía Láctea (VVV; Minniti et al. 2010) survey also covered these regions. VVV is an ESO public photometric variability survey designed to study the stellar population of the MW bulge and disk in the Z (0.87 μm), Y (1.02 μm), J (1.25 μm), H (1.64 μm), and K_s (2.14 μm) NIR passbands. The survey was carried out using the Visible and Infrared Survey Telescope for Astronomy (VISTA; Emerson et al. 2004, 2006; Emerson & Sutherland 2010) 4 m telescope at ESO, which is equipped with the VISTA InfraRed CAMera (VIRCAM), a wide-field NIR camera with a pixel scale of 0.34 arcsec pix^{-1}. The survey covers 300 deg^2 in the Galactic bulge ($-10° \leq \ell \leq 10°$; $-10° \leq b \leq 5°$) and 220 deg^2 of the Galactic disk ($295° \leq \ell \leq 350°$; $-2° \leq b \leq 2°$). With a typical limiting K_s magnitude of $17-18$ mag and exceptional data quality, it is the deepest existing data set to explore the LSS in the ZoA. Using the VVV data (Baravalle et al. 2021 and references therein), the galaxies behind the Galactic disk were selected by color cuts and visually inspection. The volume of data from the ZoA notably increased with the extension of the VVV survey, known as the VVVX (see Table 1 in Daza-Perilla et al. 2023), and the need to apply ML techniques became evident. In this sense, Daza-Perilla et al. (2023) applied ML algorithms to the northern part of the disk from the VVVX ($10° < \ell < 20°$, $-4.5° < b < +4.5°$) survey using the visual classification already performed in the southern part (VVV survey) as a training set. These authors used two samples of data, one based on the NIR images and the other based on the photometric-morphological information obtained mainly from SEXTRACTOR (Bertin & Arnouts 1996). These samples were used to separate galaxies from nongalaxies. This method was chosen because of the difficulty in obtaining both types of data, the images and photometry, and the need for double confirmation of the classification to estimate the quality of the results with each sample. To deal with the data in the ZoA, class balancing methods were applied to account for the 1:13 number imbalance in the galaxy and nongalaxy data.

Previous to Daza-Perilla et al. (2023), only a few works had used ML techniques over the ZoA regions. Vavilova et al. (2018) generated galaxy distributions and properties to compare the artificial survey with the real data in the region. Jones et al. (2019) applied convolutional neural network (CNN) and evolutionary algorithms with VISTA and UKIDSS data to study the behavior of these tools in this region. The results in this latter case are promising, with a good percentage of accuracy but the authors nevertheless suggest that results should always be visually checked for this particular area.

In addition to NIR, X-rays offer an excellent window onto extended sources in the ZoA thanks to the transparency of the MW for hard X-ray emission. Zhang et al. (2021, hereafter, ZZW21) performed an automated classification of sources using ML techniques in the entire 4XMM-DR9 survey, which covers a large part of the sky, including the thin disk of the Galaxy. These authors also used data from AllWISE (Wright et al. 2010), SDSS (York et al. 2000), and LAMOST (Cui et al. 2012) surveys at different wavelengths. Despite the increasing number and quality of data, the identification of extragalactic sources at lower galactic latitudes is challenging due to high levels of interstellar extinction and contaminating light from stars and galactic

objects. Furthermore, it is worth noting that the surveys used in ZZW21 are all-sky surveys and do not concentrate on mapping at low latitudes, in contrast to the VVV survey. Our main goal is to compare the classifications of ZZW21 with the NIR data from the VVV survey, in order to provide quantitative validation of the ZZW21 classifications in these regions.

The paper is organized as follows: in Sect. 2 we describe the data, tools, and procedures used in this work to select the sources to be compared with the results of ZZW21. In Sect. 3, special attention is given to the visual inspection of the sources and in Sect. 4 we provide specific examples and define "interesting zones" that can be used to understand the nature of the objects. Finally, in Sect. 5, we present a discussion about the application of ML algorithms in the ZoA and outline our main conclusions.

2. The data

The VVV survey provides NIR data in the southern Galactic disk. In recent years, we have used these data to identify and study galaxies behind the disk. We ran SEXTRACTOR+PSFEX (Bertin 2011) on all the images, finding 177 838 607 sources, mainly Galactic objects, such as stars, star associations, groups, and forming star regions. The extragalactic sources have colors that are different from those of Galactic objects and we used a methodology that selects galaxy candidates based on color cuts (Baravalle et al. 2018, 2019). All the candidates were visually inspected using processed images available at the ESO Science Archive and the VISTA Science Archive[1] (VSA; Cross et al. 2012). The VVV near-IR Galaxy Catalog (VVV NIRGC; Baravalle et al. 2021) is the result of this procedure, and is the largest catalog of galaxies in the southern Galactic disk. It consists of 5563 visually confirmed galaxies behind the MW.

The European Space Agency's X-ray Multi-Mirror Mission (*XMM-Newton*; Jansen et al. 2001) was launched in 1999, making observations in the X-ray, ultraviolet, and optical passbands. The 4XMM survey (Webb et al. 2020) has further improved our understanding of the X-ray Universe, providing a deep and detailed look at the X-ray sources in the sky. With extensive X-ray source catalogs, 4XMM has allowed astronomers to study a wide range of phenomena in the $0.2-12$ keV energy range, including X-ray binaries, supernova remnants, and X-ray-emitting stars. The typical positional accuracy is about 2 arcsec.

Zhang et al. (2021) employed ML techniques to classify galaxies in X-rays and included information from optical and IR in the training set. The final sample from ZZW21 consists of 550 124 objects, each with a classification parameter indicating the probability that the object is a star, a galaxy or a quasi-stellar object (QSO). The probability for galaxy classification is determined by the ML algorithm employed in combination with the bands used. The settings with best accuracy for each sample are X-ray only (Rotation Forest, 77.80%), X-ray/optical (LogitBoost, 92.82%), X-ray/IR (Random Forest, 89.42%), and X-ray/optical/IR (LogitBoost, 94.26%). The minimum probability to be classified as a galaxy using only X-ray information with the Rotation Forest algorithm is 0.333, which is one-third of the sample. From the sample of ZZW21, there are 15 423 objects in the VVV disk region (between $295° < \ell < 350°$ and $-2° < b < 2°$), which the authors classified as stars, galaxies, and QSOs. As there are no galaxies in these regions obtained using optical data and only a few with IR in these regions, we decided to use the classification and probabilities P_X from only X-ray
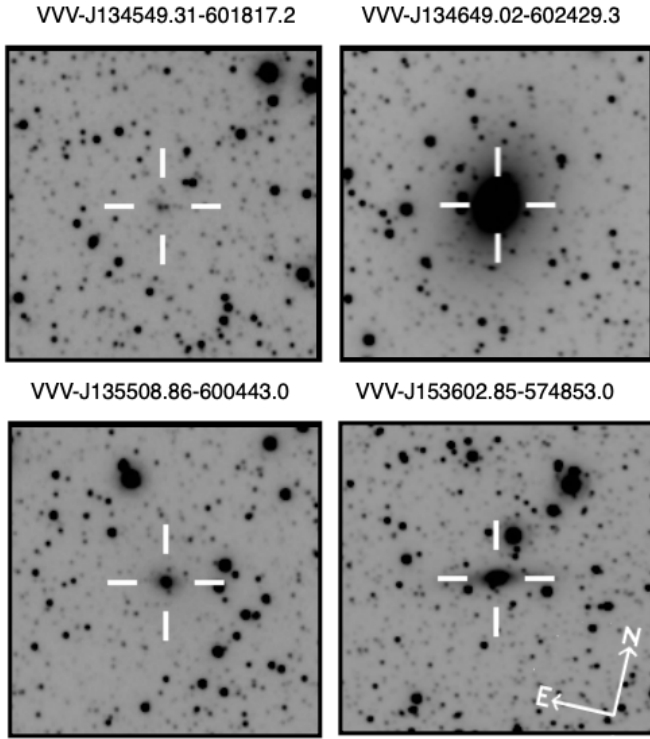
---

[1] http://horus.roe.ac.uk/vsa/

VVV-J134549.31-601817.2   VVV-J134649.02-602429.3

VVV-J135508.86-600443.0   VVV-J153602.85-574853.0

Fig. 1. Galaxies in common between VVV NIRGC and the galXray sample. The galaxies are shown in the K_s VVV passband with $1' \times 1'$ size. The orientation of all images is shown in the bottom-right panel.

data. There are 1666 stars (10.80%), 9726 galaxies (63.06%), and 4031 QSOs (20.14%), with an important imbalance, mainly between stars and galaxies. Hereafter, we refer to the subsample of 9726 ZZW21 galaxies in the VVV regions of the galactic disk as the galXray sample. The median of their probabilities P_X to be galaxies is 0.664. There are 4829 sources with higher-than-median probabilities of being considered galaxies, which represent 49.65% of the sample. Our main goal in this work is to detect these X-ray galaxies in the NIR passbands of the VVV survey.

The galXray sample was cross-matched with the VVV NIRGC catalog, which revealed only four galaxies in common with differences in positions of smaller than 1.3 arcsec. The VVV NIRGC has only 45 galaxies in common with other authors (namely Schröder et al. 2007, 2019; Williams et al. 2014; Said et al. 2016; Baravalle et al. 2021, Sect. 2.2) and the galXray sample has only one galaxy in common with Schröder et al. (2007): DZOA 4653–11 (J134736.00–603703.8) with a CMB radial velocity of $4041 \pm 86\,\mathrm{km\,s^{-1}}$ (Radburn-Smith et al. 2006) and a probability of $P_X = 0.552$ of being a galaxy from ZZW21. Figure 1 shows the VVV K_s images of the four galaxies in common between VVV NIRGC and the galXray sample, each of $1' \times 1'$ in size. These are clearly early-type galaxies with probabilities P_X of 0.575 and 0.561 (upper panels) and bulges with P_X of 0.765 and 0.778 (lower panels). These objects have K_s magnitudes brighter than 15.06 mag, as reported by Baravalle et al. (2021).

We also performed a cross-match, with a radius of 1.3 arcsec, between the galXray sample with the original output of SEXTRACTOR+PSFEX in the Galactic disk to see the morphology of the sources. We found 3229 sources in common, of which 3225 are Galactic sources. Of these, 3183 are point sources and 64 extended ones, mainly consisting of gas clouds. Figure 2

shows the distribution of galaxies within the VVV Southern disk area obtained from the galXray sample, the VVV NIRGC, and Schröder et al. (2007). In general, the galXray sample is distributed evenly throughout the inner regions of the disk and exhibits high extinctions at very low Galactic latitudes. On the other hand, the VVV NIRGC is located at higher latitudes. There is no overlap between the surveys due to the high interstellar extinctions and the differing wavelengths of each survey. Also shown are the overdensities present in the area introduced by Soto et al. (2022) and the A_V isocontours derived from the extinction maps of Schlafly & Finkbeiner (2011).

The remaining 6497 sources with no counterpart in the SEXTRACTOR+PSFEX catalog, which constitute 66.8% of the galXray sample, underwent cross-matching with all surveys available in VizieR. This cross-match yielded 732 sources in common with the VVV-DR2 survey, which are predominantly bright objects, specifically stars. Additionally, the remaining 5765 sources did not yield any results. Furthermore, these sources lack counterparts in all wavelengths, including the *Gaia*-DR3 survey (Gaia Collaboration 2016). This sample is referred to as the NOmatch sample throughout this paper. Figure 3 summarizes the procedure adopted to select the final sample to identify the different sources that are part of the galXray sample of galaxies in the ZZW21 study.

3. Visual inspection

To improve our understanding of the sources in the NOmatch sample, we visually inspected the 5765 sources using $30'' \times 30''$ Z, Y, H, J, and K_s stamps centered at the position of the sources in the VVV-DR5 images. The inspection involved observing differences in surface brightness of the sources in the five VVV NIR passbands as we did for the first time in Baravalle et al. (2018) for galaxy identification and classification. Stellar objects exhibit comparable surface brightness in all five VVV passbands, whereas extended sources possess larger surface brightness at longer wavelengths (J, H, and K_s), but are faint or barely detectable at shorter wavelengths (Z and Y). This inspection considered both the central source and its surroundings in order to identify and characterize common features. These features are not necessarily mutually exclusive, which means a single source can have multiple associated features. To improve the overall study, we kept the primary classification. The purpose of this inspection was to validate the galaxy classification of ZZW21. The sources detected by *XMM-Newton* might not be observable in the NIR of the VVV survey because of the severe conditions of the Galactic disk with high interstellar extinction and stellar contamination. Our inspection allowed us to divide the sample into ten distinct groups or categories.

Here, we provide a brief description of the various features found in the NIR images of the southern Galactic disk region in order of the number of cases found. The first and most common are normal "crowded regions" and "stars". At lower latitudes, the presence of the Galactic disk causes the fields to be heavily contaminated by stars. Together they account for 73.7% of the features found in the NOmatch sample. As noted above, the stars have approximately the same surface brightness regardless of the passband. No extended structures are observed in the frames. "Empty central regions" refer to areas in the center that lack detectable sources in the NIR passbands.

"Star associations", "Saturated bright stars", and "spikes" represent less than 5% of the cases. The star associations feature comprises up to four stars located in close proximity to each other in or near the stamps' central regions. Bright stars
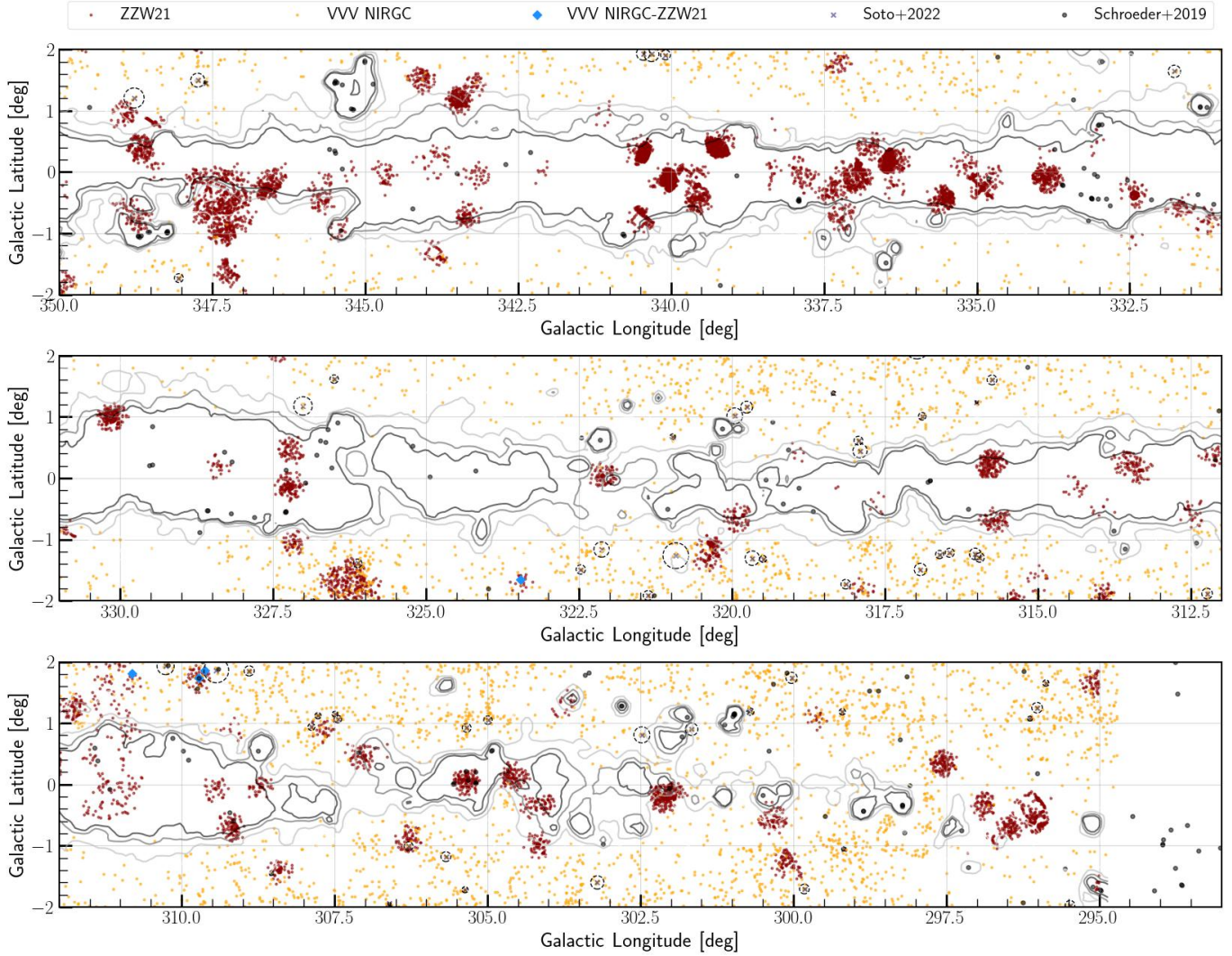
Fig. 2. Distribution of galaxies from the ZZW21 in the VVV Southern disk region. The galaxies from galXray sample are represented by red dots, the confirmed galaxies from the VVV NIRGC are in orange, the four galaxies in common between them are shown as blue diamonds, and galaxies from Schröder et al. (2019) as black dots. The center of the overdensities reported by Soto et al. (2022) are represented with a black "x" centered on a black dotted circle, which denotes the radius of each overdensity amplified by a factor of four. The A_V isocontours derived from the extinction maps of Schlafly & Finkbeiner (2011) are superimposed in a gray gradient with levels of 11, 15, 20, and 25 mag.

might be present either at the center of the stamps or in other areas, and could significantly contaminate the objects. The saturated stars with larger radii and spikes from nearby stars could strongly affect the surrounding area. The spikes display a diffraction pattern of a massive nearby star on the stamps. Less than 1% of the cases correspond to star-forming regions, which are typically found in the disks of spiral galaxies. We include the "star-forming regions" observed in most passbands with no variations in the star brightness. We also define "photometrically ultravariable stars" (PUVS), which refers to the presence of a star near the center that varies consistently in brightness between passbands. Although sometimes corresponding to variable stars, these objects are occasionally misidentified as extended sources. We also designate "photometrically ultravariable star-forming regions" (PUVSFR), which show a significant high surface brightness levels at longer wavelengths (J, H, and K_s), often spanning a large portion of the frame and even with the presence of saturated stars. At shorter wavelengths, these appear to be made up of gas clouds surrounding the stars. In 13 cases, we also found bright star associations characterized by several mas-

sive stars, mostly saturated and displaying spikes. These massive stars are close in the sky suggesting that they belong to the same highly star-forming region with high X-ray emission.

Table 1 summarizes the features found in the stamps of the NOmatch sample together with the number of occurrences and the percentage relative to the total number of sources. Figure 4 shows some examples in the VVV K_s passband of the different features. The ZZW21 galaxies with probabilities P_X higher than 0.95 in the galXray sample were visually inspected and all the objects are bright stars. This represents the 9% of the sample. Upon request, the samples of objects with different features can be provided.

4. Detailed inspection of interesting zones

To study the distribution of the X-ray sources in the galXray sample at lower Galactic latitudes, we defined some interesting zones inspired by the patterns visible in Fig. 2. We identified the 9726 sources classified as galaxies from the galXray sample, the VVV NIRGC, and Schröder et al. (2019).
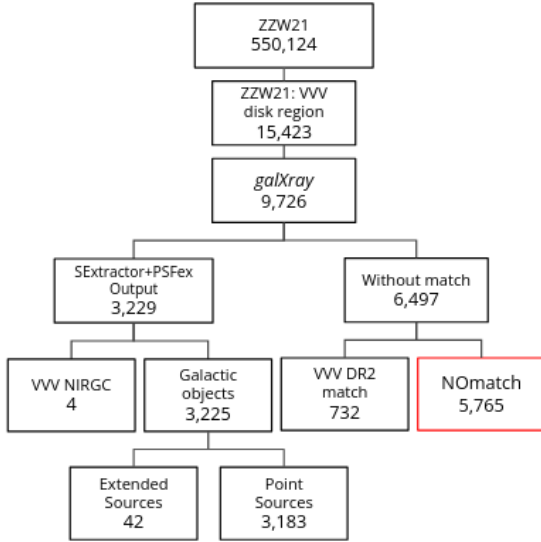
Fig. 3. Flow-chart showing the selection of sources from the ZZW21 sample making up the NOmatch sample, which is our main concern in this work.

Table 1. Features found in the NIR images through visual inspection.

Features in the NIR	Total number	%
Crowded region	3243	56.25%
Central star	1008	17.48%
Empty central region	971	16.84%
Star association	266	4.65%
Saturated bright star	153	2.61%
Spikes	48	0.83%
Normal SFR	29	0.50%
PUVS	22	0.38%
Massive star association	13	0.23%
PUVSFR	12	0.21%

We now focus our interest on the distribution of the sources from the galXray sample represented in red. We selected five interesting zones, Z1 to Z5. These are distributed across the Galactic plane, each one with distinct characteristics and showing a high concentration of X-ray sources. Figure 5 shows the distribution of sources classified as galaxies color-coded according to the probability P_X of being a galaxy, as defined by ZZW21. Most of the regions exhibit high interstellar extinction, particularly Z3, whilst Z5 has the lowest extinction in comparison. Furthermore, Z1 to Z4 are located in the Norma Supercluster region (Woudt & Kraan-Korteweg 2001), which is of major importance for the LSS and the Great Attractor.

Figure 6 shows the X-ray images in the eb3 channel (1.0−2.0 keV) from the *XMM-Newton* Observatory of the interesting zones Z1 to Z5 with the galXray sources highlighted as red points in each specific region. This passband helps us to verify the existence of extended Galactic structures because it is more affected by the presence of strong interstellar absorption and is less contaminated by hard X-ray emission. A peculiar Galactic extended structure resembling a bubble can be identified in some of these images. In the Z1, Z2, and Z5 zones, the symmetrical shape suggests a supernova remnant (SNR) with the additional presence in X-rays of a central point source, as in the case of Z1. Conversely, Z4 seems more indicative of a
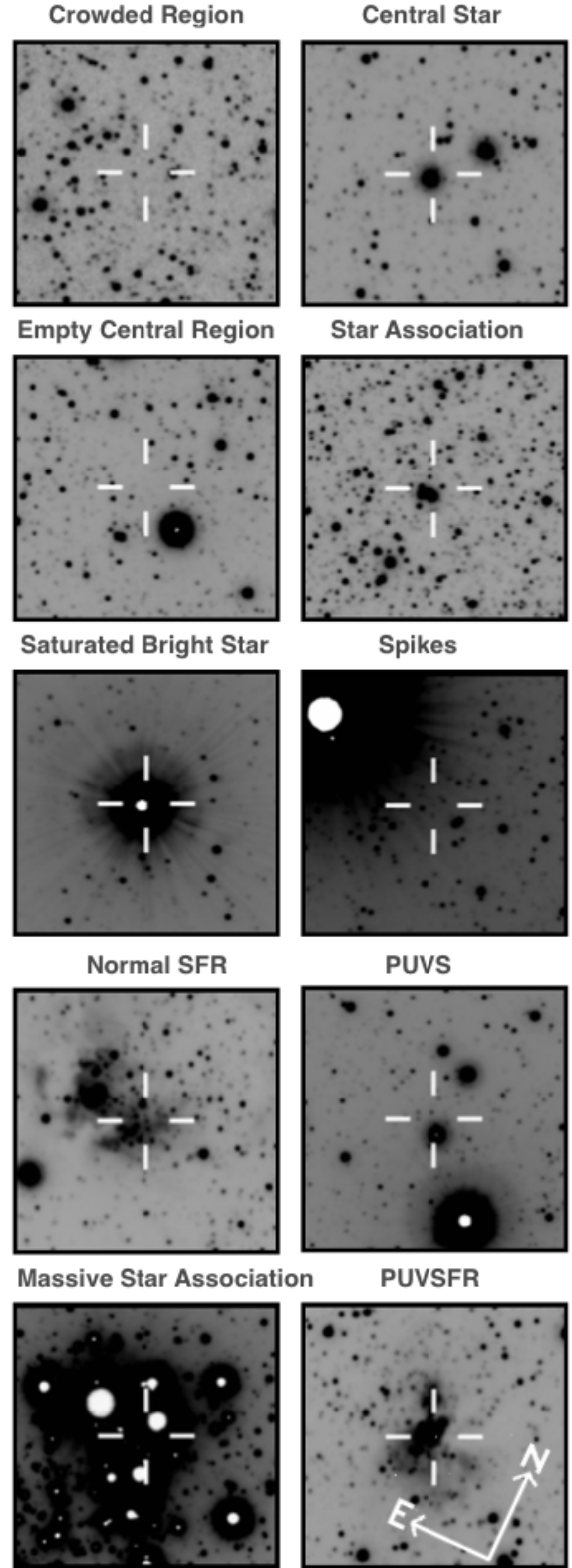

Fig. 4. Examples of objects found in the NOmatch sample are shown in the VVV K_s passband with the size of $1' \times 1'$. The orientation of all images is shown in the bottom-right panel.

star-forming region, with the gas distribution concentrated in one specific area. It is also observed that the sources of the galXray sample present a nonuniform distribution, favoring the structure of the hot gas in each zone. An exception is Z3, where the
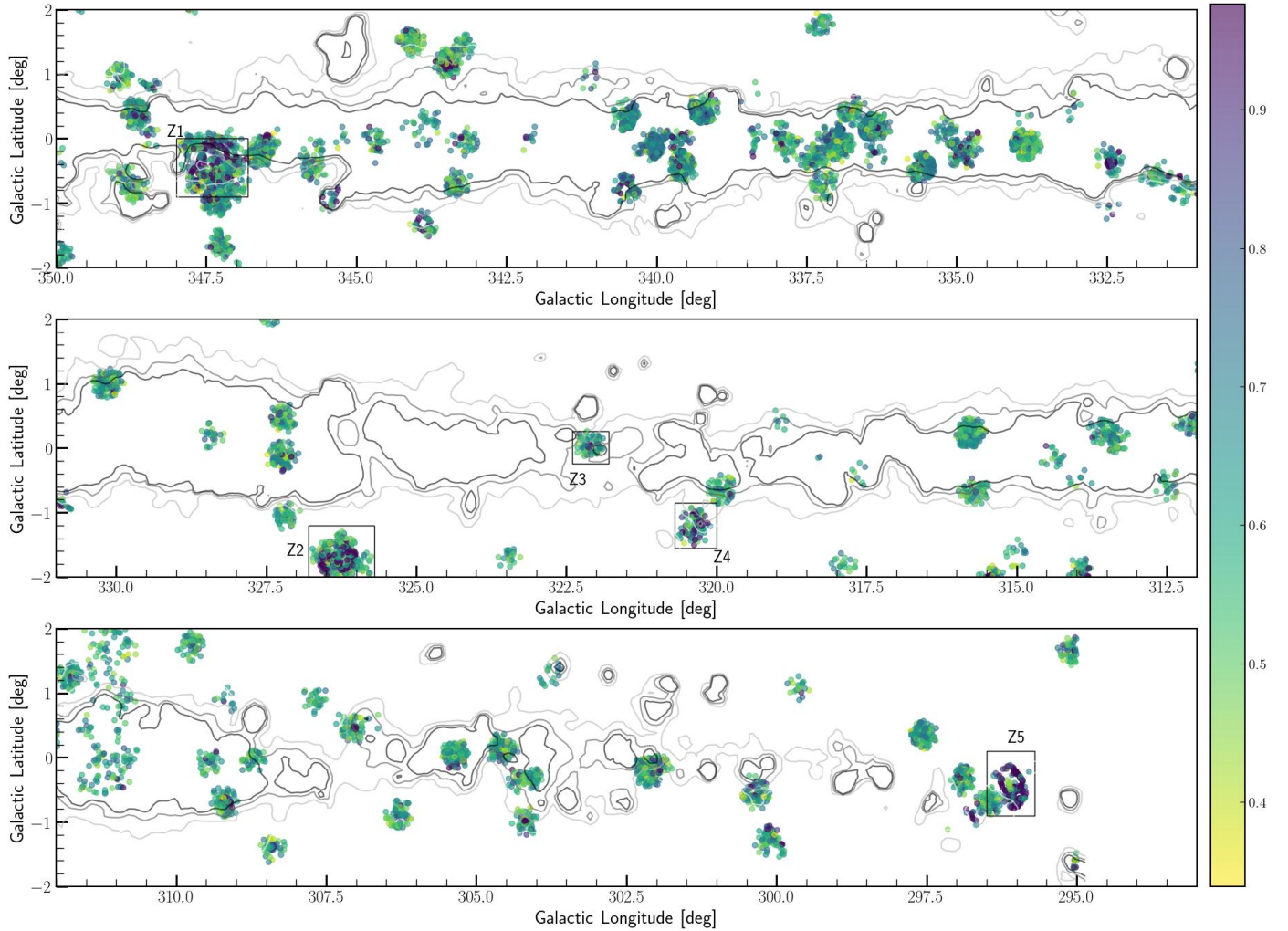
Fig. 5. Distribution of sources classified as galaxies in the southern Galactic disk of the VVV survey. The galaxies from Zhang et al. (2021) are color-coded according to their probability P_X of being a galaxy. The black squares shows the "interesting zones" studied. The A_V isocontours derived from the extinction maps of Schlafly & Finkbeiner (2011) are superimposed in gray scale with levels of 11, 15, 20, and 25 mag.

distribution of the sources shows a nearly homogeneous distribution throughout the area.

Zones Z1 to Z5 were also observed at different wavelengths, including radio, mid-infrared (MIR), NIR, and optical, using images provided by the Sydney University Molonglo Sky Survey (SUMSS; Bock et al. 1999), AllWISE (Wright et al. 2010; Mainzer et al. 2011), VVV, and the Supercosmos H-alpha Survey (SHS; Parker et al. 2005), respectively. Figures 7–11 show the interesting zones at the most relevant wavelength of each survey.

4.1. Supernova remnants

In the Z1 region, centered at α: $17^h\,13^m\,21.31$, δ: $-39^d\,41^m\,32.05$, it is possible to observe a symmetrical circular structure in the XMM images that resembles a sphere surrounding a central point source. We note that 562 galXray objects in this region (red points; see upper left panel of Fig. 6) follow the X-ray emission structure, concentrating over the darker regions. According to different studies (Slane et al. 1999; Vasquez et al. 2005; Cassam-Chenaï et al. 2004; Tateishi et al. 2021), this region corresponds to the shell-type SNR RX J1713.7–3946, which is characterized by nonthermal emission. Its central

X-ray point source is 1WGA J1713.4–3949, which has an X-ray-to-optical flux ratio that is consistent with that of a neutron star. Another possibility is that the source is an extragalactic background source (Slane et al. 1999), and therefore it is highly probable that the central point source seen in X-rays is the compact relic of the supernova progenitor of the remnant in the category of type II supernovae (Cassam-Chenaï et al. 2004). The other point source 12 arcmin away from the central X-ray source corresponds to the red super giant star HD 155603.

In the Z3 region, centered at (α: $15^h\,20^m\,34\overset{s}{.}605$, δ: $-57^d\,07^m\,56\overset{s}{.}599$), the central source shown in the upper right panel of Fig. 6 corresponds to the neutron star X-ray binary Circinus X-1 within a SNR, known as one of the brightest X-ray sources in the sky. Heinz et al. (2013) studied the natal SNR of the accreting neutron star Circinus X-1, comparing the emission in X-ray with radio. In the SUMSS stamp (radio) in Fig. 9, the SNR and the jet of the binary source (Sadeh et al. 1979; Phillips et al. 2007; Johnston et al. 2016; Coriat et al. 2019) can be seen at the center.

The galXray sample in the Z1 and Z3 regions provides a clear delineation of the gas structure in the SNRs. The sources in these two regions have median probabilities of being galaxies P_X of 0.712 and 0.674, respectively.
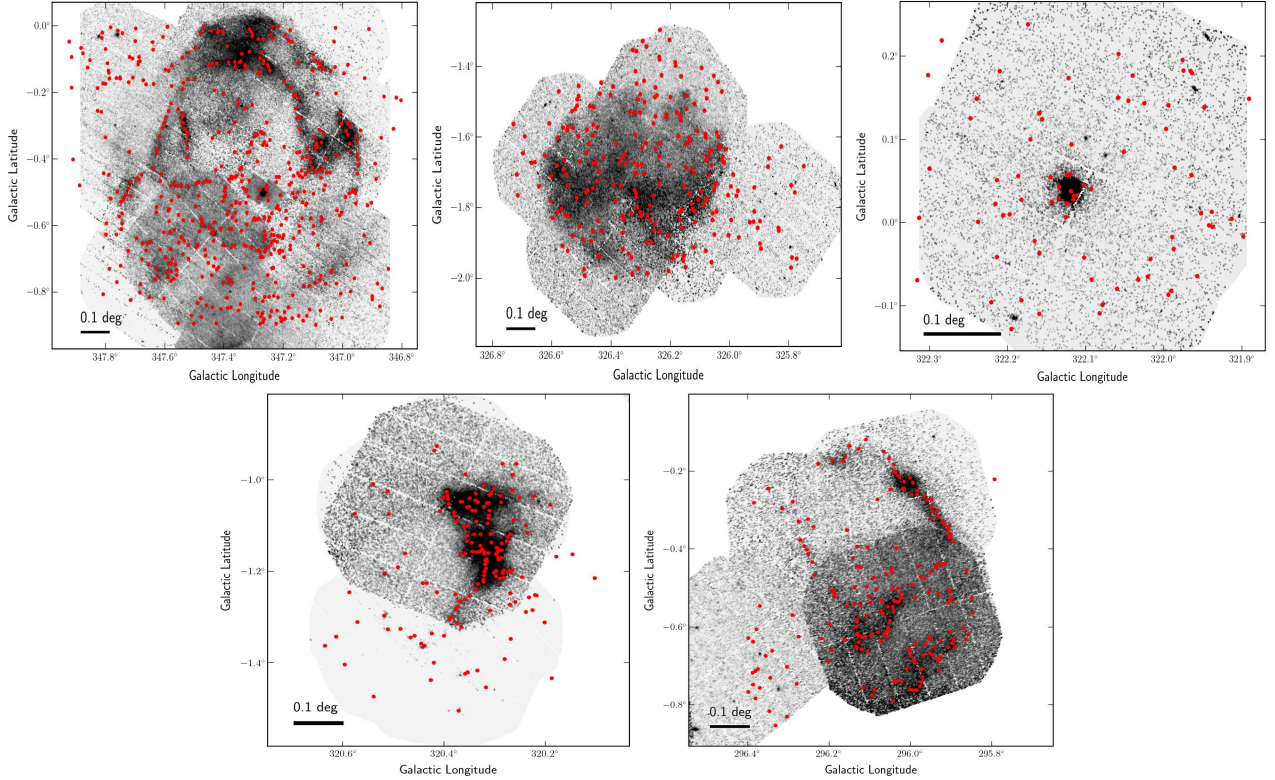
Fig. 6. XMM eb3 images of the Z1 to Z5 interesting zones, from top-left to bottom-right, respectively, including the galXray objects of ZZW21 in each region in red dots.
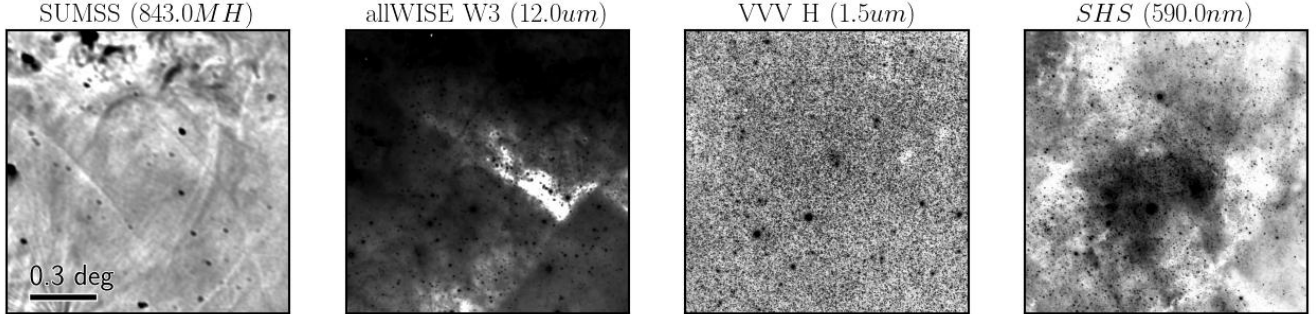

Fig. 7. Stamps at different wavelengths for the Z1 interesting zone with available surveys.

4.2. Other regions

In the Z2 (α: $15^\text{h}\,52^\text{m}\,26\overset{s}{.}793$, δ: $-56^\text{d}\,11^\text{m}\,29\overset{s}{.}343$), we observe that the galXray emission data (red points, Fig. 6) follow the distribution of the hot gas circular structure for XMM and SUMSS images. For Z4 (α: $15^\text{h}\,14^\text{m}\,43\overset{s}{.}7601$, δ: $-59^\text{d}\,09^\text{m}\,40\overset{s}{.}265$), nonuniform structure is found, both in the eb3 channel of XMM and other passbands (Figs. 6 and 10). In Z5 (α: $11^\text{h}\,51^\text{m}\,50\overset{s}{.}437$, δ: $-62^\text{d}\,35^\text{m}\,19\overset{s}{.}304$), and the red points of the galXray sample in Fig. 11 follow part of a double-shell structure, which is also observed in the SUMSS passbands. At present, no additional data are available in the literature for these areas.

Table 2 summarizes the statistics of the probabilities that the sources are galaxies from the work of ZZW21 for the interesting zones, including those from all-sky XMM data and galXray sample for comparison. The table shows the identification and the number of objects in each sample in Cols. (1) and (2) and the quartiles in the Cols. (3)–(5). The sources in Z2, Z4, and Z5 have probabilities P_X of being galaxies with median values of 0.649, 0.744,

and 0.966, respectively. The median of Z5 is the highest and is a very different value compared with the other regions. Also, it is somewhat puzzling that the quartile distribution of the classification probability in the ZZW21 and galXray samples does not show significant changes between the all-sky and the Galactic disk results, despite the different extinction levels.

5. Discussion and final remarks

Classifying extragalactic objects, especially through automated methods, is very challenging at lower Galactic latitudes due to high interstellar extinction and the small number of galaxies compared with stars. Zhang et al. (2021) used machine learning algorithms and defined the training set using X-ray data. These authors presented an all-sky catalog of galaxies with associated probabilities. The ML algorithms might be deficient for the detection and classification of galaxies that are outside the range of characteristics used in the training models. In this work, we
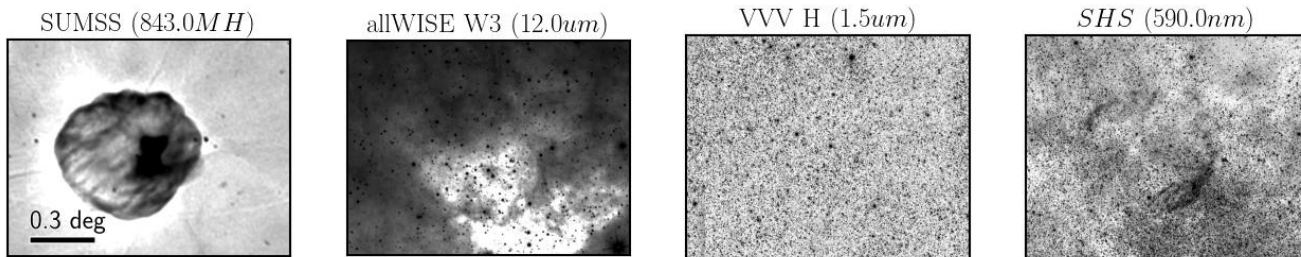
SUMSS (843.0MH) allWISE W3 (12.0um) VVV H (1.5um) SHS (590.0nm)

0.3 deg

Fig. 8. Same as Fig. 7, but for Z2 region.

SUMSS (843.0MH) allWISE W3 (12.0um) VVV H (1.5um) SHS (590.0nm)

0.1 deg
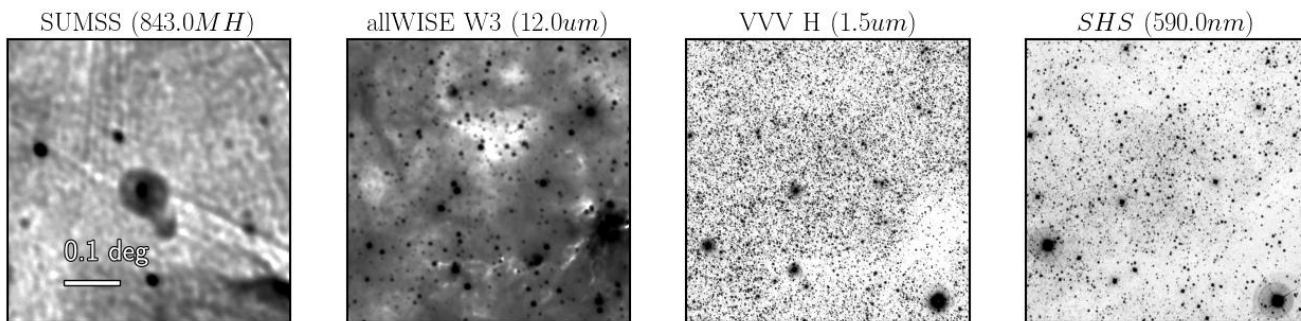
Fig. 9. Same as Fig. 7, but for Z3 region.

SUMSS (843.0MH) allWISE W3 (12.0um) VVV H (1.5um) SHS (590.0nm)

0.3 deg
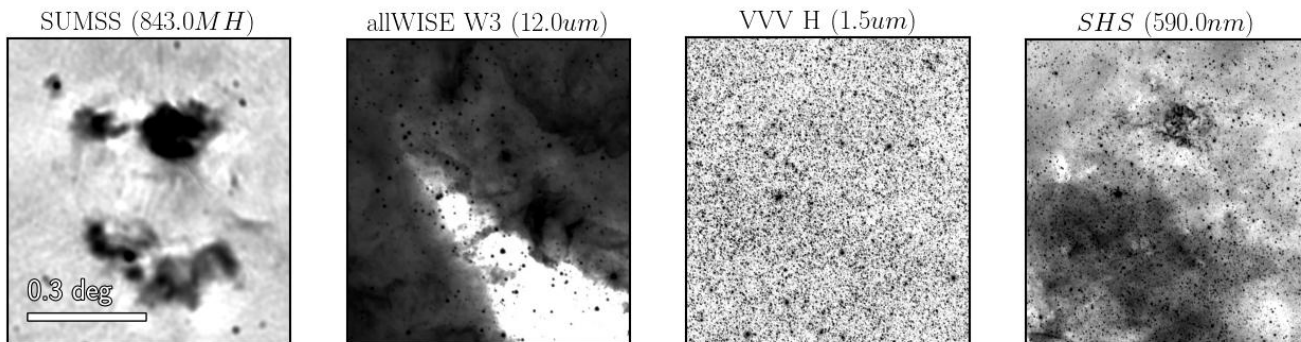
Fig. 10. Same as Fig. 7, but for Z4 region.

compared the galaxies found by these latter authors at energetic wavelengths with the objects obtained by SEXTRACTOR+PSFEX in the NIR regime.

Zhang et al. (2021) classified galaxies using the Rotation Forest method, relying solely on X-ray data. However, data from other wavelength ranges were incorporated during the training phase, which introduced emissions with potentially distinct signals. It is important to double check with a different data set, such as the VVV survey in the NIR, because the use of diverse data sources may lead to misclassifications, especially at lower Galactic latitudes. In such cases, it is crucial to consider data transfer or transfer learning when performing classifications. This strategic approach allows knowledge gained from one data set to enhance the algorithm's performance on another, potentially mitigating challenges associated with varied signal and noise characteristics. In the case of ZZW21, the algorithm was trained using SDSS data that mainly consist of bright and large galaxies. No optical galaxies were found in the studied region. It is possible that the algorithm may classify objects with X-ray emission as galaxies, but this may only be valid when they are bright and not obscured by Galactic dust. Hence, even objects with X-ray emission resembling that of SDSS galaxies should

be treated with caution. In the samples through the Galactic disk, certain objects may be erroneously classified as galaxies, when in reality they are blended stars or other extended Galactic structures.

From the all-sky 4XMM-DR9 survey, there are 15 423 objects from ZZW21 within the VVV southern Galactic disk area. In this sample, ZZW21 classified 1666 stars, 4031 quasars, and 9726 galaxies. In this region, Baravalle et al. (2021) obtained the VVV NIRGC catalog with 5563 visually confirmed galaxies. The cross-match between these two samples results in only four galaxies in common. There is also one galaxy reported by Schröder et al. (2007) in this region.

The 5765 objects of the NOmatch sample were visually inspected in this work. This was the most critical part, which included regions of high interstellar extinction and stellar crowding. It is evident that the majority of the ten different features summarized in Table 1 correspond to Galactic regions normally crowded with several stars distributed over the whole area. Around 25% of the cases exhibit a bright or saturated star at the center of the stamp. The cases of Normal Crowded Region and Empty Central Region account for 73% of all cases. The absence of galaxies in the NIR regime does not necessarily imply
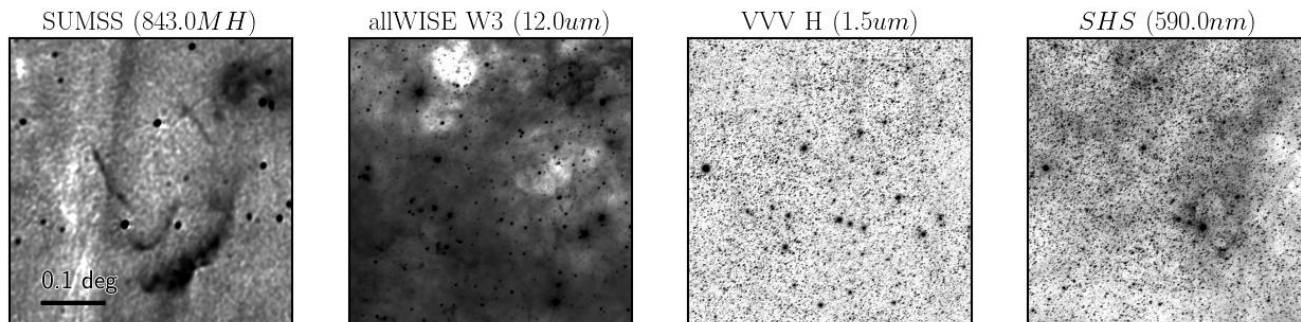
SUMSS (843.0MH)   allWISE W3 (12.0um)   VVV H (1.5um) SHS (590.0nm)

Fig. 11. Same as Fig. 7, but for the Z5 region.

Table 2. Statistics of the probabilities P_X for different samples and zones.

Galaxy sample	# objects	Q1	Q2	Q3
ZZW21	191 528	0.538	0.657	0.804
galXray	9726	0.552	0.664	0.743
Z1	562	0.591	0.712	0.953
Z2	244	0.539	0.649	0.958
Z3	78	0.552	0.674	0.712
Z4	158	0.566	0.744	0.963
Z5	203	0.704	0.966	0.972

a lack of galaxies in these regions, but rather suggests that it may be challenging to detect them, especially in regions with higher interstellar extinctions. The galaxies can be observed with the XMM but remain undetectable in the NIR passbands. Hence, the most significant result is the lack of galaxies in these regions.

We also used the results of several surveys at different wavelengths to perform a visual panchromatic inspection. We defined five "interesting zones" based on the data distribution from ZZW21. Images of each area were compared at different wavelengths, from radio to X-rays (Figs. 7–11). They also show high probabilities of being a galaxy, as seen in Fig. 5 and Table 2. Based on these images and previous results, we might conclude that these interesting zones correspond to the emission from extended Galactic events rather than individual galaxies. The sources classified as galaxies by ZZW21 belonging to the interesting zones are part of extended Galactic structures, such as SNRs or star-forming regions. Further studies in these regions are needed.

Imbalanced datasets are frequently encountered in ML and pattern recognition (Lemaître et al. 2017), compromising the learning process. Most of the standard ML algorithms expect balanced class distribution or an equal misclassification cost (He & Garcia 2009). This problem can be critical when trying to distinguish galaxies from nongalaxies at lower Galactic latitudes where the numbers of stars and associations are extremely important. Conventional classifiers often prioritize the minimization of the overall error rate, which can lead to a bias towards the majority class, resulting in an inaccurate classification. The most common metric used in classification is accuracy, which measures the ratio of correct predictions to the total number of input samples (Blagec et al. 2020). In the zone of avoidance, when dealing with imbalanced data sets, which are often found when classifying galaxies and nongalaxies, accuracy alone may not be the optimal metric. In Daza-Perilla et al.

(2023), we used the F1-score as a better option, as it considers both the quantity and quality of the classification. Our classification algorithms were centered on the 5509 VVV NIRGC galaxies and 74 238 nongalaxies in the southern part of the Galactic disk using the VVV survey and the VVV NIRGC. We considered regions with varying Galactic extinction levels, employing both the CNN method with NIR images and the XGBoost method with photometric and morphological VVV NIRGC data. These two samples were used as training sets to separate galaxies from nongalaxies in the northern Galactic disk using the VVVX survey, taking into account the number imbalance present in the dataset.

Our work highlights the importance of having a representative training set when working on the ZoA using ML. An appropriate training set ensures accurate and reliable classification, improving the purity of the positive class in particular, and the results in general, which is of paramount importance when mapping the LSS at lower galactic latitudes.

References

Aguirre, C., Pichara, K., & Becker, I. 2019, MNRAS, 482, 5078
Baravalle, L. D., Alonso, M. V., Nilo Castellón, J. L., Beamín, J. C., & Minniti, D. 2018, AJ, 155, 46
Baravalle, L. D., Nilo Castellón, J. L., Alonso, M. V., et al. 2019, ApJ, 874, 46
Baravalle, L. D., Alonso, M. V., Minniti, D., et al. 2021, MNRAS, 502, 601
Bertin, E. 2011, in Astronomical Data Analysis Software and Systems XX, eds. I. N. Evans, A. Accomazzi, D. J. Mink, & A. H. Rots, ASP Conf. Ser., 442, 435
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Blagec, K., Dorffner, G., Moradi, M., & Samwald, M. 2020, arXiv e-prints [arXiv:2008.02577]
Bock, D. C. J., Large, M. I., & Sadler, E. M. 1999, AJ, 117, 1578
Cassam-Chenaï, G., Decourchelle, A., Ballet, J., et al. 2004, A&A, 427, 199
Coriat, M., Fender, R. P., Tasse, C., et al. 2019, MNRAS, 484, 1672
Cross, N. J. G., Collins, R. S., Mann, R. G., et al. 2012, A&A, 548, A119
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Res. Astron. Astrophys., 12, 1197

Dainotti, M. G., Bogdan, M., Narendra, A., et al. 2021, ApJ, 920, 118
Daza-Perilla, I. V., Sgró, M. A., Baravalle, L. D., et al. 2023, MNRAS, 524, 678
Emerson, J., & Sutherland, W. 2010, The Messenger, 139, 2
Emerson, J. P., Sutherland, W. J., McPherson, A. M., et al. 2004, The Messenger, 117, 27
Emerson, J., McPherson, A., & Sutherland, W. 2006, The Messenger, 126, 41
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
He, H., & Garcia, E. A. 2009, IEEE Trans. Knowl. Data Eng., 21, 1263
Heinz, S., Sell, P., Fender, R. P., et al. 2013, ApJ, 779, 171
Henning, P. A., Kraan-Korteweg, R. C., Rivers, A. J., et al. 1998, AJ, 115, 584
Jansen, F., Lumb, D., Altieri, B., et al. 2001, A&A, 365, L1
Johnston, H. M., Soria, R., & Gibson, J. 2016, MNRAS, 456, 347
Jones, D., Schroeder, A., & Nitschke, G. 2019, arXiv e-prints [arXiv:1903.07461]
Kraan-Korteweg, R. C., Woudt, P. A., Cayatte, V., et al. 1996, Nature, 379, 519
Kraan-Korteweg, R. C., Cluver, M. E., Bilicki, M., et al. 2017, MNRAS, 466, L29
Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, J. Mach. Learn. Res., 18, 1
Loeb, A., & Narayan, R. 2008, MNRAS, 386, 2221
Macri, L. M., Kraan-Korteweg, R. C., Lambert, T., et al. 2019, ApJS, 245, 6
Mainzer, A., Bauer, J., Grav, T., et al. 2011, ApJ, 731, 53
Marton, G., Ábrahám, P., Szegedi-Elek, E., et al. 2019, MNRAS, 487, 2522
Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, New Astron., 15, 433
Parker, Q. A., Phillipps, S., Pierce, M. J., et al. 2005, MNRAS, 362, 689
Phillips, C. J., Deller, A., Amy, S. W., et al. 2007, MNRAS, 380, L11
Radburn-Smith, D. J., Lucey, J. R., Woudt, P. A., Kraan-Korteweg, R. C., & Watson, F. G. 2006, MNRAS, 369, 1131
Ramatsoku, M., Verheijen, M. A. W., Kraan-Korteweg, R. C., et al. 2016, MNRAS, 460, 923
Sadeh, D., Meidav, M., Wood, K., et al. 1979, Nature, 278, 436
Said, K., Kraan-Korteweg, R. C., Jarrett, T. H., Staveley-Smith, L., & Williams, W. L. 2016, MNRAS, 462, 3386
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schröder, A. C., Mamon, G. A., Kraan-Korteweg, R. C., & Woudt, P. A. 2007, A&A, 466, 481
Schröder, A. C., van Driel, W., & Kraan-Korteweg, R. C. 2019, MNRAS, 482, 5167
Shapley, H. 1961, J. R. Astron. Soc. Can., 55, 273
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Slane, P., Gaensler, B. M., Dame, T. M., et al. 1999, ApJ, 525, 357
Soto, M., Sgró, M. A., Baravalle, L. D., et al. 2022, MNRAS, 513, 2747
Spindler, A., Geach, J. E., & Smith, M. J. 2021, MNRAS, 502, 985
Tateishi, D., Katsuda, S., Terada, Y., et al. 2021, ApJ, 923, 187
Vasquez, J., Cappa, C., & McClure-Griffiths, N. M. 2005, MNRAS, 362, 681
Vavilova, I. B., Elyiv, A. A., & Vasylenko, M. Y. 2018, Russ. Radio Phys. Radio Astron., 23, 244
Webb, N. A., Coriat, M., Traulsen, I., et al. 2020, A&A, 641, A136
Williams, W. L., Kraan-Korteweg, R. C., & Woudt, P. A. 2014, MNRAS, 443, 41
Woudt, P. A., & Kraan-Korteweg, R. C. 2001, A&A, 380, 441
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Zhang, Y., Zhao, Y., & Wu, X.-B. 2021, MNRAS, 503, 5263