VEXAS: VISTA EXtension to Auxiliary Surveys

Data Release 2: Machine-learning based classification of sources in the Southern Hemisphere^{★,★★}

V. Khramtsov^{1,2}, C. Spiniello^{3,4}, A. Agnello⁵, and A. Sergeyev^{1,6}

¹ Institute of Astronomy, V. N. Karazin Kharkiv National University, 35 Sumska Str., Kharkiv, Ukraine
e-mail: vld.khramtsov@gmail.com

² Department of Data Science, Quantum, 20, Otakara Yarosha lane, Kharkiv, Ukraine

³ Sub-Dep. of Astrophysics, Dep. of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

⁴ INAF, Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Naples, Italy

⁵ DARK, Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen Ø, Denmark

⁶ Institute of Radio Astronomy of the National Academy of Sciences of Ukraine, 4, Mystetstv St., Kharkiv 61002, Ukraine

Received 14 December 2020 / Accepted 4 March 2021

ABSTRACT

Context. We present the second public data release of the VISTA EXtension to Auxiliary Surveys (VEXAS), where we classify objects into stars, galaxies, and quasars based on an ensemble of machine learning algorithms.

Aims. The aim of VEXAS is to build the widest multi-wavelength catalogue, providing reference magnitudes, colours, and morphological information for a large number of scientific uses.

Methods. We applied an ensemble of thirty-two different machine learning models, based on three different algorithms and on different magnitude sets, training samples, and classification problems (two or three classes) on the three VEXAS Data Release 1 (DR1) optical and infrared (IR) tables. The tables were created in DR1 cross-matching VISTA near-infrared data with Wide-field Infrared Survey Explorer far-infrared data and with optical magnitudes from the Dark Energy Survey (VEXAS-DESW), the Sky Mapper Survey (VEXAS-SMW), and the Panoramic Survey Telescope and Rapid Response System Survey (VEXAS-PSW). We assembled a large table of spectroscopically confirmed objects (VEXAS-SPEC-GOOD, 415 628 unique objects), based on the combination of six different spectroscopic surveys that we used for training. We developed feature imputation to also classify objects for which magnitudes in one or more bands are missing.

Results. We classify in total $\approx 90 \times 10^6$ objects in the Southern Hemisphere. Among these, $\approx 62.9 \times 10^6$ ($\approx 52.6 \times 10^6$) are classified as ‘high confidence’ (‘secure’) stars, $\approx 920\,000$ ($\approx 750\,000$) as ‘high confidence’ (‘secure’) quasars, and ≈ 34.8 (≈ 34.1) million as ‘high confidence’ (‘secure’) galaxies, with $p_{\text{class}} \geq 0.7$ ($p_{\text{class}} \geq 0.9$). The DR2 tables update the DR1 with the addition of imputed magnitudes and membership probabilities to each of the three classes.

Conclusions. The density of high-confidence extragalactic objects varies strongly with the survey depth: at $p_{\text{class}} > 0.7$, there are 11 deg^{-2} quasars in the VEXAS-DESW footprint and 103 deg^{-2} in the VEXAS-PSW footprint, while only 10.7 deg^{-2} in the VEXAS-SM footprint. Improved depth in the mid-infrared and coverage in the optical and near-infrared are needed for the SM footprint that is not already covered by DESW and PSW.

Key words. astronomical databases: miscellaneous – catalogs – surveys – methods: data analysis – virtual observatory tools

1. Introduction

Wide-field, digital, astronomical surveys have the potential to advance our knowledge in observational astrophysics. The footprint, depth, and above all accessibility are paramount in the search for rare objects (e.g., extremely-metal-poor or ultra-cool-dwarf stars, high-redshift quasars, strong gravitational lenses), and for the statistical properties of astronomical objects at large, such as the clustering properties of galaxies or the assembly history of our own Milky Way (MW, see e.g., the review of Helmi 2020 and references therein).

* The DR2 catalogues are also available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via <http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/651/A69>

** VEXAS is publicly available through the ESO Phase 3, https://archive.eso.org/scienceportal/home?data_collection=VEXAS

Over the last decade, different facilities have surveyed wide areas of the sky, over various frequencies in the optical and near-infrared (NIR). They have extended the pioneering work of the Sloan Digital Sky Survey (SDSS, DR1, Abazajian et al. 2003; DR14, Abolfathi et al. 2018) to wider and homogeneous footprints, and in some cases with better depth and image quality. The Panoramic Survey Telescope & Rapid Response System DR1 (PanSTARRS1, Chambers et al. 2016) has covered $30\,000 \text{ deg}^2$ contiguously to the north of declinations $\delta > -30 \text{ deg}$, in *grizy* magnitudes, and down to $i \approx 23.1$ (5σ); the Dark Energy Survey (DES, Dark Energy Survey Collaboration 2016) has reached deeper magnitudes ($i \approx 23.50$, 10σ) over a 5000 deg^2 contiguous footprint in the south in *grizY*; and the SkyMapper Southern Sky Survey (SM, Wolf et al. 2018) has covered the whole Southern Hemisphere in *uvgriz*, albeit to shallower limits and with a coarser image quality. To complement those optical surveys, different initiatives have mapped

wide areas in the NIR, mostly in the south (e.g., with the Visible and Infrared Survey Telescope for Astronomy, VISTA, Emerson et al. 2006).

In principle, these existing surveys can already provide a wealth of information on multiple fields of observational astrophysics, especially in preparation for more ambitious endeavours such as the *Vera Rubin* Observatory LSST (Ivezić et al. 2019) and the ESA-*Euclid* mission. In practice, matters are complicated by their heterogeneous and separate coverage. The SDSS was transformational because it provided homogenised information and additional tables with higher-level data, such as object classifications, photometric redshifts, and stellar masses and sizes of galaxies. All in all, calibrated and cross-matched multi-wavelength catalogues are crucial.

This is indeed the purpose of the VISTA EXtension to Auxiliary Surveys (VEXAS) project (Spiniello & Agnello 2019, hereafter Paper I), which extends infrared data of VISTA from the X-ray (ROSAT All Sky Survey, Boller et al. 2014, 2016; The *XMM-Newton* Serendipitous Survey, Watson et al. 2001) up to the radio domain (SUMSS, Bock et al. 1999) in a fully homogenised database.

The VEXAS DR1 catalogues, which are publicly available via the ESO Phase 3¹, include multi-bands photometric data as well as objects with a spectroscopic follow up (from the Sloan Digital Sky Survey, SDSS, DR14, Abolfathi et al. 2018 and/or from the 6dF Galaxy Survey, 6dFGS, Jones et al. 2004). The core requirement of VEXAS is a reliable photometry in more than one band. This condition, together with the detection in at least two surveys (via cross-match), should minimise, if not completely eliminate, the number of spurious detections in the final catalogues.

For the VEXAS Data Release 2 (DR2), which we describe in this paper and release to ESO Phase3, we continue the effort of providing the scientific community with useful multi-wavelength photometric catalogues, but we also add the probability of it being a galaxy, a star, or a quasar for each catalogue source. This macro-classification is obtained through a machine learning (ML) ensemble of algorithms initially developed for a similar effort on a smaller footprint (Khrantsov et al. 2019), but improved and re-trained on purpose to match the VEXAS catalogues' composition and varying depth and coverage.

The paper is organised as follows. In Sects. 2 and 3, we provide details on the input tables and training tables that we used for the classification. In Sect. 4 we present an exhaustive description of our classification pipeline, including details on the feature imputation technique and on the ensemble learning. This latter is based on three different classifiers which were used to build 32 single models solving different problems, with different features and training datasets. In Sect. 5 we assess the performance and quality of our pipeline on a test sample. In Sect. 6 we present the main results of the classification process, including the validation with external data. Finally, in Sect. 7, we draw our conclusions and discuss possible future developments of the VEXAS Project.

2. The input data: the VEXAS optical and infrared tables

In Paper I, we covered the Southern Galactic Hemisphere (SGH) below the Galactic plane (i.e., to $b < -20^\circ$). As

¹ The VEXAS tables are available both from the [Archive Science Portal](#) or through the [Catalog Facility](#) (active links in this footnote in the online version).

as a first step, we assembled a table of 198 608 360 detections from the VISTA-Kilo-Degree Infrared Galaxy Survey (VIKING, Sutherland 2012; Edge et al. 2013) and the VISTA Hemisphere Survey (VHS, McMahon et al. 2013), with reliable Petrosian magnitudes and small uncertainties in at least one band (see Eq. (1) in Paper I). We then cross-matched the table with the AllWISE Source Catalogue (Cutri 2013), which includes ≈ 747 million objects over the whole sky, with a median angular resolution of 6.1, 6.4, 6.5, and 12.0 arcsec for the four bands ($W1$, $W2$, $W3$, and $W4$), respectively. Setting a matching radius of 3 arcsec, we built the VEXAS-AllWISE table of 126 372 293 objects with near- and far-infrared photometry. Subsequently, we matched the VEXAS-AllWISE table with the following three different multi-band optical wide-sky photometric surveys: DES, PanSTARRS1, and SkyMapper, resulting in the tables VEXAS-DESW, VEXAS-PSW, and VEXAS-SMW. In each table, each entry has reliable photometry measured in at least one of the VISTA infrared bands (Z , Y , J , H , K_s), two of the WISE bands ($W1$, $W2$, $W3$, $W4$), and three optical bands (u , g , r , i , z , y)². These magnitudes are always provided in their native system of reference (AB for optical and Vega for infrared) and corrected for extinction.

In this VEXAS-DR2, we used a filtered version of the DR1 tables: we removed all sources fainter than 25^m in each considered band since below this value the extrapolation of the training sample cannot be tested or trusted as a result (see Sect. 6.1 and Appendix A). In the case of VEXAS-PSW, we also applied a more severe cut on the optical magnitudes and associated uncertainties: we restricted ourselves to magnitudes brighter than the mean 5σ point-source limiting sensitivity values given in Chambers et al. (2016) and we filtered out all sources with uncertainties $> 1^m$. For the VEXAS-SMW table, in DR1 we limited ourselves to $\delta < -30^\circ$, since above this declination the PSW coverage is uniform and at least two magnitudes deeper. Here in DR2 we consider the whole coverage of Sky Mapper in the SGH, extending the VEXAS-SMW table to ~ 32 M unique objects. This allowed us to compare the results obtained from training the pipeline on SMW with those obtained by training it on DESW and PSW (see Sect. 6.3).

These three tables constitute the photometric datasets used for this VEXAS-DR2. Their sky coverage is shown in Fig. 1 and the numbers of the objects are listed in Table 1.

Throughout the paper, these three tables are treated separately. As we describe in Sect. 6, we used our classification pipeline on each of them and obtained three tables of classified objects in output, which update and replace the three optical+NIR VEXAS DR1 tables.

3. The training samples

In this section we describe the different training datasets that we used for the ML process. We combined data from six different spectroscopic surveys (SDSS DR16, Ahumada et al. 2020; WiggleZ, Drinkwater et al. 2018; GAMA DR3 Baldry et al. 2018; OzDES, Childress et al. 2017, 2QZ Croom et al. 2001 and 6dFFS DR3, Jones et al. 2009) in order to build a training sample as large as possible and as complete as possible in all the three classes of objects (STAR, GALAXY, and QSO).

In the following sub-sections, we describe how we selected and combined these different sets into the training, validation,

² Throughout the paper, we always indicate the infrared band from VISTA with an uppercase Y and the optical one from DES or PS with a lowercase y .

Fig. 1. Sky coverage view of the three input VEXAS optical+IR tables. The colours indicate the number of objects per deg^2 , as shown by the side bar, obtained using a Hierarchical Equal Area isoLatitude Pixelation of a sphere (HEALPIX) with resolution equal to 9.

Table 1. Number of objects and sky coverage of each of the three optical cross-matched VEXAS tables we gave as input to our classification pipeline.

Table	Sky area (deg^2)	Declination	Num (#) of objects
VEXAS-DESW	4900	$-70^\circ < \delta < 0^\circ$	35 381 482
VEXAS-PSW	3800	$-30^\circ < \delta < 0^\circ$	21 891 609
VEXAS-SMW	9300	$\delta < 0^\circ$	31 999 995

Notes. These tables were assembled in Paper I, but have been slightly changed here (see text for more details). Both VEXAS DR1 and DR2 include only objects with Galactic latitude $b < -20^\circ$.

and testing samples. The data always include spectroscopic classification of the sources and redshift information for extragalactic objects. We applied some selection criteria on the spectra to select only sources passing the quality level thresholds recommended directly by the data providers. These criteria are described in the following for each of the six spectroscopic surveys separately.

3.1. SDSS DR16

The largest spectroscopic sample we used for training is the sixteenth data release of Sloan Digital Sky Survey (SDSS DR16, Ahumada et al. 2020). SDSS DR16 is the forth release of the current SDSS phase IV (Blanton et al. 2017), and it collects the most recent imaging, photometric, and spectroscopic data. In particular, it comprises optical spectra from the extended Baryon Oscillation Spectroscopic Survey (eBOSS, Dawson et al. 2016), including data from the eBOSS sub-projects – the SPectroscopic IDentification of ERosita Sources (SPIDERS, Clerc et al. 2016; Dwelly et al. 2017) survey and the Time-Domain Spectroscopic Survey (TDSS, Morganson et al. 2015) and infrared stellar spectra from the Apache Point Observatory Galaxy Evolution Experiment 2 (APOGEE-2, Majewski et al. 2016). Finally, integral field spectroscopic observations of nearby galaxies from the Mapping Nearby Galaxies at Apache Point Observatory (MaNGA, Bundy et al. 2015) are included too.

SDSS DR16 covers $\sim 15\,000\,\text{deg}^2$ and includes 5 789 200 spectra in total, among which 5 107 045 are unique³. From these,

we selected only sources with $\text{zWarning} = 0$, which resulted in a sample of 890 374 stars, 751 741 quasars, and 2 638 083 galaxies.

3.2. WiggleZ final DR

The WiggleZ Dark Energy Survey (Drinkwater et al. 2010) is a survey of $\sim 10^5$ objects with ultraviolet photometry from the Galaxy Evolution Explorer (GALEX, Martin et al. 2005) survey, specifically selected in order to limit the sample to $z \sim 0.5$ emission-line galaxies ($z_{\text{median}} \approx 0.6$). The final data release of the WiggleZ survey (Drinkwater et al. 2018) contains a spectroscopic classification for 225 415 sources (mostly blue galaxies) across a $1000\,\text{deg}^2$ area split into seven fields. Here, we only used galaxies with well-defined redshifts (with the quality parameter $Q > 3$), and we limited the sample to sources with $z > 0.0024$, as recommended by Drinkwater et al. (2010), to remove possible stellar contamination from the catalogue. This results in a set of 144 040 galaxies.

3.3. GAMA

The Galaxy And Mass Assembly (GAMA⁴, Driver et al. 2011) survey is mainly aimed at redshift measurements of galaxies. Here, we employed the GAMA Data Release 3 (Baldry et al. 2018), which consists of $\sim 214\,000$ galaxies, some of which were also observed with other surveys (SDSS, WiggleZ, etc.) and added to the catalogue. This survey is split into five sky fields (with an area of $\approx 60\,\text{deg}^2$ each), three of which are near the celestial equator and two are in the southern celestial hemisphere.

We cleaned the GAMA DR3 sample of galaxies with the following criteria: $0.05 < z < 0.9$, $nQ > 1$, where nQ is the ‘normalised quality scale’ and measures the probability that the redshift estimate for a given spectrum is correct⁵. This allowed us to collect a sample of 192 422 galaxies with precisely measured redshifts.

3.4. OzDES DR1

The Australian Dark Energy Survey (OzDES, Childress et al. 2017; Yuan et al. 2015) is a spectroscopic survey to measure

⁴ <http://www.gama-survey.org/>

⁵ For spectra with $nQ < 1$, it is not possible to measure a redshift.

³ Selected from the ‘SpecObj’ table via the CasJobs platform.

redshifts of ~ 2500 Type Ia supernovae host-galaxies over the redshift range $0.1 < z < 1.2$ and derive reverberation-mapped black hole masses for ~ 500 quasars over $0.3 < z < 4.5$. The OzDES First Data Release (OzDES DR1⁶) contains the redshifts of $\sim 15\,000$ sources that were observed during the first three years of observations. We used only the 14 693 sources with redshift-quality flag > 3 .

3.5. 2QZ

The 2dF QSO Redshift Survey (2QZ, Croom et al. 2001) is a spectroscopic survey of $\sim 40\,000$ quasars observed over two $75^\circ \times 5^\circ$ declination strips. We retrieved the final catalogue, which combines 2QZ with the 6dF QSO Redshift Survey (Croom et al. 2001) and comprises 49 424 sources. Since 2QZ includes observations over two different epochs, we filtered out sources that have been classified in different classes or that have a different redshift in one epoch with respect to the other. Moreover, we also excluded all sources for which redshift estimation is missing in both epochs (see Table 2 in Croom et al. 2003). For sources with more than one redshift estimate, we simply computed the average.

The labels STAR, QSO, or GALAXY for the selected objects were retrieved directly from the catalogue. The final sample, after cleaning, consisted of 39 639 sources.

3.6. 6dFGS

The 6dF Galaxy Survey (6dFGS, Jones et al. 2009) aims at studying galaxies spectroscopically in the southern sky. The final data release of 6dFGS contains 136 304 spectra of mostly extragalactic objects. We selected only objects with $q_z \geq 3$ and labelled those with $q_z = 6$ as stars, as described on the 6dFGS description pages.

3.7. The final VEXAS spectroscopic table

Combining the six ‘cleaned’ spectroscopic tables described above, and cross-matching with the three input tables (using a matching radius of $1.5''$ since the resolution in the optical is better than that in the infrared), we assembled a final spectroscopic table for each VEXAS table. In particular, for VEXAS-DESW, VEXAS-PSW, and VEXAS-SMW, we find 293 584, 328 821, and 211 092 unique spectroscopic sources, respectively. More detailed information is available in Table 2 on how many objects per class were found in each survey.

Despite the fact that we used these three separate spectroscopic samples for the three VEXAS input tables, we also release the VEXAS-SPEC-GOOD, comprising 415 628 unique VEXAS objects with photometry in the optical and infrared and a secure and clean spectroscopic classification. In total, there are 89 222 unique STAR, 35 179 unique QSO, and 291 227 unique GALAXY in the released spectroscopic table. The redshift distribution of the spectroscopic objects is plotted in Fig. 2, where the sources are colour coded by their class, as indicated by the caption. The footprint of the table, which is colour-coded by the object density (number of objects per deg^2), is instead plotted in Fig. 3.

In this paper, the three final VEXAS spectroscopic tables are used in Sect. 4 to train our classification models and validate their performance.

To this purpose, each spectroscopic table is further split into a main and an auxiliary sub-table for the training phase. The

Table 2. Number of objects with a spectroscopic match from one or more spectroscopic surveys used in this paper to train the machine learning pipeline.

Survey	Class	VEXAS DESW	VEXAS PSW	VEXAS SMW
SDSS	STAR	56 187	74 794	77 100
	QSO	23 820	26 872	4660
	GALAXY	144 495	162 717	65 937
2QZ	STAR	537	151	701
	QSO	3499	657	699
	GALAXY	368	114	46
OzDES	STAR	1614	1216	1569
	QSO	1191	998	220
	GALAXY	7246	4438	1778
6dFGS	STAR	506	649	1544
	GALAXY	25 274	21 792	54 093
GAMA	GALAXY	33 050	34 551	11 578
WiggleZ	GALAXY	7851	12 759	723

Fig. 2. Redshift distribution of the sources in the VEXAS-SPEC-GOOD table, colour coded by the object class.

main sub-table is then divided into training, testing, and validating samples. More details are given below.

3.8. Splitting spectroscopic datasets

Each final spectroscopic table described above is separated into main and auxiliary samples for the workflow described in Sect. 4. These two parts are used at different stages during the training of the algorithms to mitigate the effect of the presence of wrong classification labels. In fact, despite the selection criteria we applied on the spectroscopic tables, the final labelling and classification of the spectroscopic datasets is not ideal (as shown in Appendix B). Many sources have spectra that are too noisy to provide a trustworthy classification and some objects have also been misclassified by the survey pipelines. Therefore, we undertook an iterative process, alternating the training and sample composition. At training, the classifier(s) learns from the main training sample, and then it predicts the classes of all the sources in the auxiliary sample. At this point, all of the sources for which the predicted class are consistent with the original

⁶ <http://www.mso.anu.edu.au/ozdes/DR1>

Fig. 3. Sky coverage view of the VEXAS-SPEC-GOOD final table. The colour indicates the number of objects per deg^2 , in logarithmic scale, as shown by the side bar, obtained as in Fig. 1.

spectroscopic-based class are added to the main training sample, and the classifier is run again, learning each time from a larger data set. As an initial main sample, we adopted the SDSS DR16 catalogue as it comprises the largest number of objects over the three macro-classes. We used the remaining five spectroscopic datasets as auxiliary instead.

The main (SDSS) sample was divided into training (60%), validation (20%), and testing (20%). The training sample was then used to train the classification algorithm, the validation sample to control and tune the classifier during its learning process, and, finally, the testing sample was used to check the quality of the classifier, which was run on previously unseen data, as described in Sect. 5. The auxiliary sample was added only to the main training sample.

4. Classification pipeline

In this section, we highlight the main characteristics of our classification pipeline in more detail. We make the code available via [Github](#)⁷.

To classify the sources of the VEXAS input tables into STAR, QSO, and GALAXY, we set up a pipeline able to resolve, or at least to take into account, the major limitation of the VEXAS DR1 dataset, in particular regarding the fact that some magnitudes are missing for a non-negligible number of sources in one or more bands. This can be an issue because not all machine learning algorithms can handle missing data. We used the technique of feature imputation to fill the missing magnitudes, and we trained a separate ML algorithm on purpose, which we describe below (Sect. 4.1). We finally give details on the feature sets (Sect. 4.2) and algorithms, both on single classifiers and ensemble learning (Sect. 4.3), which we used in our classification pipeline.

4.1. Dealing with missing magnitudes: imputation

As already highlighted in Paper I, the VEXAS coverage is not homogeneous across the different bands. This is particularly true for SM and in general for the infrared, or at least for the Y and H bands, given the strategy of the VHS survey, and for $W3$ and $W4$ from WISE. This means that for many entries in the catalogue, one or more magnitudes in one or more bands can be missing

Table 3. Percentage of objects with a measured magnitude in each of the listed bands for the VEXAS input tables.

Band	VEXAS-DESW	VEXAS-PSW	VEXAS-SMW
u	–	–	19.0%
g	98.5%	89.3%	51.9%
r	99.6%	97.3%	65.0%
i	99.6%	100.0%	84.9%
z	99.8%	98.0%	94.1%
y	99.4%	95.7%	–
Y	6.4%	51.4%	29.1%
J	98.0%	98.8%	99.7%
H	31.6%	24.1%	16.7%
K_S	91.5%	94.2%	99.4%
$W1$	100.0%	100.0%	100.0%
$W2$	88.6%	87.6%	96.0%
$W3$	20.0%	18.4%	17.9%
$W4$	4.7%	4.5%	4.3%

simply because this particular object has not been observed by the corresponding survey. Another reason is also that an object is detected in some bands, but it is too faint in others (e.g., $z \sim 3$ objects lack flux in the u -band, [Steidel et al. 1996](#)). Clearly, disentangling between these two cases would help in the classification of the sources with missing magnitudes.

In general, simply removing the sources which have partial magnitude information would severely reduce the VEXAS dataset. This can be seen from Table 3 where we list the percentage of objects with measured magnitude in each band for each input VEXAS table. Even without considering the Y and H bands from VISTA and $W3$ and $W4$ from WISE, which we excluded because of their shallower coverage, we would lose roughly 10–15% of the sources for VEXAS-DESW and VEXAS-PSW and almost 80% of the sources in VEXAS-SMW (or 50% discarding the u -band). Rather than limiting the analysis to only sources with photometry measured in each band, a strategy for handling missing magnitudes has to be adopted within our pipeline if we aim to maximise the multi-band coverage.

Furthermore, we note that a simple imputation with mean, median, or constant values would allow us to preserve the total number of sources, but it would not guarantee the reliability of the classification results for sources with missing magnitudes. A better way to handle this is to use an intelligent imputer, which is able to find a relationship between missed and measured magnitudes for each source, which should reflect its true underlying properties.

We follow this approach here, using an autoencoder (AE) neural network. In general, an AE is trained to reproduce the input features, and as a consequence the hidden layers end up selecting combinations of input features that carry most of the underlying physical information. For this reason, AEs are typically used for dimensionality reduction, compressing the information through a hidden layer with a lower dimension than the input feature space. In our case, we are interested in reconstructing missing magnitudes, so the input to our AE is a vector of magnitudes of which some of them are masked out at random, and the output target is the full (un-masked) magnitude vector. We give further detail on our AE architecture in Appendix C.1. In order to test the reliability of the imputation procedure, we first performed a test on a sample of objects with known magnitudes to quantify the goodness of imputed

⁷ [VEXAS Github repository](#) (active link in the online version)

magnitudes. Appendix C.2 gives further details about it, including the trustworthiness of the output magnitudes.

To test the effect of feature imputation on the classification, instead, we ran our pipeline sometimes with and sometimes without imputation (see Sect. 4.3), and we compared the results obtained in both cases. However, we caution the reader that, in the case of VEXAS-SMW, the imputation, especially for quasars, might not work properly for all the objects. This is possibly due to the fact that the number of quasars with magnitudes measured in all bands present in this table is very low (<8000)⁸. The imputer therefore does not have enough knowledge on the spectral energy distribution of this class of objects and thus it does not predict good values for the missing magnitudes by looking at the measured ones. More details are provided in Appendix C.3.

In addition, for all the classes, the percentage of objects with missing magnitudes in the optical bands is much larger for VEXAS-SMW, as visible from Table 3, and this makes imputation harder. For this reason, in the case of VEXAS-SMW, we obtained and released the following two different classifications: the fiducial one with imputation and an independent one without imputation and based on the decision-tree based CatBoost algorithm (Dorogush et al. 2018; Prokhorenkova et al. 2018), which is able to deal with missing magnitudes but it is not as complete and generalisable as the ensemble learning META_MODEL described in Sect. 4.3. We anticipate that the two algorithms result in an object classification which is the same for 99.5% of the sources in the VEXAS-SMW table (see Sect. 6 and Fig. 8).

4.2. Feature set

The basic principle of our feature set construction is that STAR, QSO, and GALAXY are different in their morpho-photometric properties. Thanks to the broad wavelength range covered by the VEXAS tables, we have a large number of magnitudes ($u, g, r, i, z, y, J, K_s, W1, W2$)⁹ at our disposal, and at least one morphological parameter highlighting the ‘stellarity’ for each entry in the catalogue.

In the following, we indicate the source index as i and the magnitude indices as j, k , and we describe the input feature space that we used for the system classification, which is based on the following:

Colour indices. Different types of sources have different spectral energy distributions (SEDs) and thus they lie in different regions in colour-colour diagrams. Having a large set of magnitudes, we can form all the possible colours as pairwise differences between two magnitudes: $m_{i,j} - m_{i,k}$. For the models where feature imputation is not used, if one of the two considered magnitudes is not measured for a given source, the colour obtained from that magnitude will also be missing for that source (see below).

Scaled magnitudes. The physical characteristics of the sources cannot be retrieved directly from the raw magnitudes. However, magnitude information could be important for some of the classifiers, which could create ‘their own features’ space (e.g., artificial neural networks), or the learned relationship between magnitudes and classes with a distance-based approach

(e.g., k nearest neighbours). In both cases, the shape of the SEDs is expressed with a set of magnitudes. Rather than using standard magnitudes, we used ‘scaled magnitudes’ (m'), which are better suited to capture the relative ‘intensity’ in each band for each source, independently of the intrinsic differences in brightness between the different objects. In particular, for each source, we identified the maximum magnitude across all bands and we re-scaled all the other bands to that value: $m'_{i,j} = m_{i,j}/\max(m_i)$. This scaling helps increase the classification reliability on very faint sources.

Stellarity index. As additional morphological information, we used the VISTA P_{STAR} parameter which is an index of the ‘stellarity’ of the sources, and it has been proved to be a very important feature in the classification of objects (e.g., Khramtsov et al. 2019). This parameter has values between 0 and 1, where 1 corresponds to a point-like source and 0 corresponds to an extended source. It is more generally available than other indicators, for example psf versus model magnitudes, as it is obtained directly at the level of image segmentation.

4.3. Ensemble learning

In the presence of noisy datasets, as ours are (see Appendix B), an ensemble of different classifiers performs better than a single trained and optimised classifier (Kuncheva 2004; Myoung-Jong et al. 2006; Lessmann et al. 2015). For this reason, we used an ensemble of classifiers here based on different algorithms.

In general, in ensemble learning, the final classification is obtained by joining the predictions of the individual classifiers. In particular, we adopted a stacking procedure, in which a meta-classifier uses the predictions of a number of individual classifiers as input features, and it learns to predict the classes using the input probabilities derived from single models.

We defined 32 different models based on the following three algorithms:

1. Artificial neural network (ANN). ANN learns the function in an iterative way, connecting input features with the target values, producing the weighted sum of neurons’ output, which are connected with each other by layers. Our ANN is made up of seven layers; the first one accepts the raw features as input, and the following six transform them into $5 \times k$ features, where k ranges from $k = 5$ to $k = 1$. The last layer consists of neurons, the number of which corresponds to the number of output classes. To prevent overfitting and to speed up the learning, we used the batch-normalisation and dropout techniques; all layers are connected via a rectified linear unit activation function. The ANN minimises the binary cross-entropy with the *Adam* optimiser (Kingma & Ba 2014; with a learning rate equal to 10^{-2}) over ten epochs, and the weights are saved for the epoch with the lowest loss function value on the validation dataset.

2. k -nearest neighbours (k NN). The k NN classifies a source by selecting the most prevalent class within the k nearest neighbours. Here, we selected $k = 15$ as the number of neighbours to be used in the k NN algorithm and used the Euclidean distance as the distance metric between sources in the feature space.

3. CatBoost (Dorogush et al. 2018; Prokhorenkova et al. 2018). CatBoost is a gradient boosting (Friedman 2000) decision-tree ensemble algorithm. To train CatBoost, we used all default parameters, except for the number of iterations, which in our case is equal to 8000. We adopted 500 early-stopping steps, meaning that we stopped the growth of trees as soon as the validation score did not increase after 500 steps.

⁸ We have obtained this number a posteriori from the classified table presented in Sect. 6

⁹ We do not consider the Y and H bands from VISTA, nor the $W3$ and $W4$ from WISE, as their coverage is sparse and very limited, see Table 3.

Table 4. Description of the 32 individual classifiers that we combined into the ensemble learning, varying the algorithms, classification problem, input bands, feature sets, the use of imputation, and the auxiliary training sample.

Mod. #	Method	Classes	Input bands	Features	Imputation	Aux. sample
Complete models (one for each algorithm)						
1	ANN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
2	kNN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
3	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
Different band-sets						
4	ANN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}$	sc.mags, colours, P_{STAR}	True	True
5	kNN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}$	sc.mags, colours, P_{STAR}	True	True
6	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}$	sc.mags, colours, P_{STAR}	True	True
7	ANN	STAR, QSO, GAL	$g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
8	kNN	STAR, QSO, GAL	$g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
9	CB	STAR, QSO, GAL	$g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
10	ANN	STAR, QSO, GAL	$J, K_S, g, r, i, z, W1, W2$	sc.mags, colours, P_{STAR}	True	True
11	kNN	STAR, QSO, GAL	$J, K_S, g, r, i, z, W1, W2$	sc.mags, colours, P_{STAR}	True	True
12	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, W1, W2$	sc.mags, colours, P_{STAR}	True	True
Different classification problems						
13	ANN	STAR, QSO+GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
14	kNN	STAR, QSO+GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
15	CB	STAR, QSO+GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
16	ANN	STAR+GAL, QSO	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
17	kNN	STAR+GAL, QSO	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
18	CB	STAR+GAL, QSO	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
19	ANN	STAR+QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
20	kNN	STAR+QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
21	CB	STAR+QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	True
The effect of imputation and auxiliary sample						
22	ANN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	False
23	kNN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	False
24	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	True	False
25	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	False	True
26	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours, P_{STAR}	False	False
Classification without stellerity						
27	ANN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours	True	True
28	kNN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours	True	True
29	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	sc.mags, colours	True	True
Classification without magnitudes						
30	ANN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	colours, P_{STAR}	True	True
31	kNN	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	colours, P_{STAR}	True	True
32	CB	STAR, QSO, GAL	$J, K_S, g, r, i, z, y/u^{(a)}, W1, W2$	colours, P_{STAR}	True	True

Notes. ^(a)We use the y band for PS and DES, and the u band for SM.

Based on these three algorithms, we built a total of 32 different models which were trained with different choices of input features. Moreover, some models were built to solve slightly different classification problems: in few cases we changed the number of output classes, merging together stars and quasars (i.e., classifying point-like versus extended sources), or merging quasars and galaxies together (i.e., extra-galactic versus galactic sources). Finally, in some cases, we employed feature imputation (see Sect. 4.1) and/or the auxiliary training sample (see Sect. 3), while in some other cases we did not. The characteristics of these models are listed in Table 4, where we also highlight the effect we aimed to study with each group of models.

In principle, one could then combine the models using a simple averaging. However, we preferred to use a slightly more sophisticated weighted averaging. We used logistic regression

(LR) as meta-classifier, trained to predict STAR, QSO, and GALAXY classes based on the probability scores derived from all the single models we set up. LR is closely related to a linear regression, but it is constructed using the Sigmoid function to limit the result in the range [0; 1]. LR is described by the following equation:

$$\hat{y}(\mathbf{C}, \mathbf{P}) = (1 + \exp^{-\mathbf{C}^T \cdot \mathbf{P}})^{-1}, \quad (1)$$

where $\mathbf{P}$ is the vector of input probabilities, $\mathbf{C}$ is the corresponding coefficient, and $\mathbf{C}^T \cdot \mathbf{P} = \sum_i C_i P_i$ is a scalar product of both vectors, with P_i being the probability to belong to a certain class given by a single model, and C_i the corresponding regression coefficient.

The regression model uses L_1 -regularisation:

$$L = \sum_j (y - \hat{y}(\mathbf{C}, \mathbf{P}))^2 + \lambda \sum_i |C_i|, \quad (2)$$

where L is a loss function and λ is a regularisation parameter (which we set to $\lambda = 1$). In the above equation, summing over j , one proceeds through all training objects; whereas summing over i , one goes through all the coefficients. This type of regularisation pushes the coefficients corresponding to useless input features to be zero, or very close to zero, hence helping us to select the most informative input features (in our case, probabilities of individual models). This is not the case for the L_2 -regularisation ($\lambda \sum_i C_i^2$), for example, which would make the selection of important input features more difficult.

Our META_MODEL, trained to separate the classes from each other¹⁰, solves a three-class problem in a ‘one-versus-rest’ scheme. This means that, in output, we have three LR models, one for each class, and thus three ‘meta probabilities’ for each object to belong to each of the three classes (one for STAR versus QSO+GALAXY, one for QSO versus STAR+GALAXY and one for GALAXY versus STAR+QSO). We note that strictly speaking, it would have been more accurate to solve the multi-classes problem. However, for computational simplicity, we restricted ourselves to the ‘one-versus-rest’ scheme, but we highlight that this is a very good approximation as the three output probabilities always sum to 1 within better than 10^{-5} . We also emphasise that the META_MODEL is more general and comprehensive than a single model and it allows us to better extrapolate the classification results to regimes that may not be covered by the current spectroscopic training set.

5. Testing the quality of the classification pipeline

Before running the pipeline on the full VEXAS input tables, we performed a number of quality checks on the ‘unseen’ test sample to understand the strengths and limitations of our classification. An extensive description of these tests and their results both on the META_MODEL and on the single ones, based on five different classification metrics and on regression coefficients, are given on the Github VEXAS repository. There, we also provide a visualisation of the relative importance of each classification feature for each single model, confirming that the VISTA P_{STAR} parameter has a high classification power. However, we stress that even though one feature is relatively important in a single model, this does not necessarily imply that the same feature is driving the ensemble classification. The main conclusion we can draw analysing the performance on each single model on the three different VEXAS tables, in terms of classification quality metrics, is that joining different classification methods, based on different algorithms and input parameters, significantly improved the classification results. In addition, we note that for some models, the ANN algorithm performs worse than the other two, especially in classifying stars and quasars. We also tried to change the configuration of the ANN to obtain a better score, but we were not successful. We nevertheless kept this algorithm in the final ensemble learning because this does not happen for all the models and for all the input tables in the same way and it makes the derived uncertainties on the classification more conservative.

Here, for simplicity, we only provide a visualisation of the distribution of predictions across the classes in the form of confusion matrices. This is one of the most commonly used classification quality metrics for ML based algorithms in astrophysics.

In general terms, a confusion matrix shows the fraction of sources belonging to a certain ‘true’ class and classified within a predicted class. For an m -class problem, the confusion matrix is defined as the $m \times m$ table, where each row represents the number of instances in a predicted class, and each column represents the number of the instances in a ground-truth class¹¹. Each cell of the table represents the number of sources from a class i , classified as class j , and the diagonal shows how many objects (often given in percentage) have been correctly classified. The confusion matrices for each of the three different VEXAS test tables, computed from the META_MODEL, are plotted in Fig. 4.

The scores obtained for the objects classified correctly (on the diagonal) are very high in all cases. We always obtained $>99\%$ of correct identifications on the STAR and GALAXY classes, and $>94\%$ on the QSO class, which is the least populated class. We also note that for all tables, the largest fraction of misclassified objects occurred for QSO classified as GALAXY, reflecting the fact that we can have both central-engine-dominated and host-dominated emission in these classes.

An interesting fact we can deduce from the confusion matrices is that the model trained on the VEXAS-DESW dataset (top panel) performs slightly better in splitting the sources into galactic and extragalactic. In fact, the contamination from stars in galaxies and quasars is the lowest: only $\approx 0.4\%$ STAR are misclassified as QSO or GALAXY, versus $\geq 0.6\%$ for the other two tables.

Unfortunately, the presence of the u -band in VEXAS-SMW does not seem to help to classify quasars; however, we stress again that only 19% of the full table has u -band detections and the imputation procedure is sub-optimal in SMW. We will tackle this problem in forthcoming releases of VEXAS, extending the cross-match with the second data release of the NOAO Source Catalog (NSC, Nidever et al. 2021), comprising 3.9 billion objects over $\sim 35\,000$ square degrees with $ugrizY$ precise photometry. Finally, we also observe that the fraction of STAR classified as QSO in VEXAS-SMW (bottom panel) is equal to zero and the fraction of QSO classified as STAR is also minimal (0.05%). However, we believe that this is mainly a limitation of the AE imputation on quasars, as we show in the next section and in Appendix C.

6. Classification results and validation

For each object in the VEXAS input catalogues, our META_MODEL returned three numbers, representing the probability of belonging to each of the three classes of objects: p_{STAR} , p_{GALAXY} , and p_{QSO} , obtained solving the ‘one-versus-rest’ problem. In general, a source belongs to a given class when the corresponding probability for that class is the highest. However, we considered as trustable classifications those for which the probability is $p_{\text{class}} \geq 0.5$. With this simple assumption, starting from the input ≈ 33 million objects in VEXAS-DESW, ≈ 22 in VEXAS-PSW, and ≈ 32 in VEXAS-SMW, we classify 1.6% of the sources as quasars, 37.4% as stars, and 61.0% as galaxies for DESW, and 1.9% quasars, 48.4% stars, and 49.7% galaxies for PSW. For VEXAS-SMW instead, 91.4% of the sources are classified as stars, 8.35% as galaxies, and only 0.25% are classified as quasars. The percentages relative to the classes are thus comparable for VEXAS-DESW and VEXAS-PSW, while they are very different for VEXAS-SM, which is roughly two magnitudes shallower than the other two surveys.

Depending on the scientific purpose one might have, it could be more appropriate to use an even more severe threshold, with

¹⁰ We note that, while each individual classifier was trained on the training sample, the meta-classifier was trained on the predictions on the validation sample. This helps in preventing overfitting.

¹¹ Transposing the table does not affect the results.

Fig. 4. Confusion matrices obtained for the META_MODEL on the test samples (20% of the SDSS spectroscopic dataset) for the three tables.

the purpose of maximising the purity of the obtained sample as much as possible, but consequently penalising the total number of selected objects. On the other hand, with the aim of assembling the largest possible sample of a given class of objects, one might use slightly lower probability thresholds, paying the price of a larger number of contaminants. This is why we list the number of objects classified in each class as a function of the

Table 5. Number of objects classified in each class as a function of the probability threshold for the DESW, PSW, and SMW final tables.

VEXAS-DESW			
p	STAR	QSO	GALAXY
0.5	13 225 467	552 847	21 593 339
0.6	13 207 001	524 388	21 543 204
0.7	13 185 692	496 514	21 479 584
0.8	13 156 283	465 135	21 381 884
0.9	13 105 319	422 868	21 145 300

VEXAS-PSW			
p	STAR	QSO	GALAXY
0.5	10 586 167	408 598	10 886 179
0.6	10 558 641	380 388	10 832 329
0.7	10 526 795	351 648	10 769 276
0.8	10 485 364	318 899	10 680 432
0.9	10 410 731	271 353	10 505 175

VEXAS-SMW			
p	STAR	QSO	GALAXY
0.5	29 242 620	80 943	2 673 869
0.6	29 214 063	76 406	2 645 062
0.7	29 181 114	72 004	2 612 243
0.8	29 135 737	67 144	2 569 647
0.9	29 050 417	59 434	2 494 954

probability threshold for each VEXAS catalogue in Table 5. In the released DR2 tables, we provide the probabilities to belong to each class, so that the users can select their preferred threshold.

A visualisation of the class distribution of the objects in the output catalogues is plotted in Fig. 5 in the form of a triangle density plot. The colours indicate the $\log_{10}$ number of objects contained in a probability cell with a size of 0.01×0.01 . Each corner represents the maximum probability of belonging to a given class ($p_{\text{class}} = 1$) and the dashed horizontal black lines show the threshold level of $p_{\text{class}} = 0.7$, which we define as threshold for the ‘high-confidence objects’ as it corresponds to objects which have a probability with a confidence of more or less one σ (0.67). This is, in our opinion, the natural compromise between purity and completeness and is used throughout the remainder of the paper for all the plots and tables.

We show the density of each class of object on the sky in Fig. 6. Also, we report the mean values obtained for high-confidence objects in Table 6.

Once again, we note that for VEXAS-DESW and VEXAS-PSW, the numbers are similar, especially for stars ($\approx 2850\text{--}3050 \text{ deg}^{-2}$) and quasars ($\approx 110\text{--}115 \text{ deg}^{-2}$). For galaxies, the density in DESW is higher with $\approx 4700 \text{ deg}^{-2}$ (versus $\approx 3100 \text{ deg}^{-2}$ in PSW). The situation is instead very different for VEXAS-SMW, where the density of stellar objects is comparable, but that of extragalactic sources is more than ten times smaller. As already previously stated, we attribute this difference to the shallower depth of the Sky Mapper Survey.

Since quasars and galaxies are more or less uniformly distributed on the sphere, the fact that the spatial distribution of these objects is not always uniform is due to spatial changes in the surveys’ depth caused by extinction, different exposure times, or a different number of bands observed in that region. For stars, instead, towards the borders we approach the MW disk, and this, of course, increases the density of galactic objects by a factor of ~ 10 .

Fig. 5. Density plot of the number of objects as a function of probability. The colour-bar indicates the $\log_{10}(N)$ per each probability cell of size 0.01×0.01 . The black horizontal dashed lines indicate where the threshold $p_{\text{class}} \geq 0.7$.

Fig. 6. Spatial density of each class of objects in the three VEXAS tables. *Top row*: results obtained for VEXAS-DESW, the *middle row* refers to VEXAS-PSW, and the *bottom row* shows the results obtained for VEXAS-SMW. Towards the borders, the contribution from MW stars increases the density of STAR. For QSO and GALAXY, the non-uniform spatial distribution is mainly due to different depths reached by the surveys in that region.

We remind the reader that our metrics for the classification accuracy are averaged over the sky. This means that there may be more cross-contamination towards the low galactic-latitude edges of the VEXAS footprint where the MW stellar density is higher. In the future, and with more demanding iterations, one may also fold in some spatial information to rebalance the scores accounting for how many objects for each class are expected at a given position. In this DR2, we preferred

to simply give the mean output scores because any further iteration depends on accurate coverage maps and on a proper model for the MW stellar density.

To further assess the reliability of the classification results, in Fig. 7, we plotted two among the most common infrared and optical magnitude-colour and colour-colour diagrams used in the literature to separate the different classes of objects (e.g., Stern et al. 2005, 2012; Assef et al. 2013; Chehade et al. 2018).

Table 6. Density of high-confidence ($p_{\text{class}} \geq 0.7$) STAR, QSO, and GALAXY in the three output tables.

Table	Class	Density (#/deg ²)
VEXAS-DESW	STAR	2889.3
	QSO	111.0
	GALAXY	4697.1
VEXAS-PSW	STAR	3037.8
	QSO	103.1
	GALAXY	3106.5
VEXAS-SMW	STAR	3324.8
	QSO	10.7
	GALAXY	300.0

The contours show object density levels, with the different colours indicating the three families of objects, as specified in the caption. The stellar locus (with low $J - K_s$) and the galaxy and quasar regions (with higher $W2$ and $W1 - W2$) are clearly visible. These plots also show that the three macro-classes cannot be separated with a very high accuracy through simple colour cuts, emphasising the need for full ML-based classification approaches.

For VEXAS-SMW, the three classes of objects do not show the same separation as in the other two tables (bottom-right panel, $J - K_s$ versus $g - i$). In Appendix C.3, we show that this is caused by the imputation, which suffers from the under-representation of quasars in the input sample. However, Fig. 8 demonstrates that the classification is not affected by this problem. This figure shows the difference between the class of each object predicted by the META_MODEL and that predicted by a single CatBoost model (#25, without imputation). For 99.5% of the objects, the difference is null, and the three p_{class} obtained from the two models are almost identical. The classification changes only for $\sim 14.6 \times 10^4$ sources ($\approx 0.5\%$ of the total), mainly from STAR to GALAXY or viceversa ($\text{Class}_{\text{meta}} - \text{Class}_{25} = \pm 2$). The classification for QSO is almost completely unaffected, which demonstrates that the META_MODEL is not affected by the sub-optimal imputation for this class.

For the sake of completeness, in this DR2, we release both sets of three probabilities (with and without imputation) in the classified VEXAS-SMW table. Moreover, in all three released VEXAS tables, we added a flag column to identify the imputed magnitudes.

6.1. Safe ranges, outliers, and saturation

One of the most common issue in machine learning based classifications is that the ‘depth’ of the data is larger than that of the training set. The most common solution is to cut out the input tables, limiting the inference to bright objects only hence avoiding any extrapolation to unseen regions in the space of features (e.g., Khramtsov et al. 2019, 2020).

However, this approach is not desirable in our case since it would violate the main purpose of the VEXAS project, being to collect as much information as possible on the multi-wavelength sky and thus classify the largest possible number of sources. Therefore, here, we do not restrict the classification to the brightest sources only, but nevertheless caution the readers that at the very bright and faint ends, the classification might be less secure.

We plotted the completeness of the training samples in the r -band for each VEXAS input table in the upper panel of Fig. 9.

While for the VEXAS-PSW table, the distribution of training and input are very similar, the training sample for the VEXAS-SMW (VEXAS-DESW) table is clearly incomplete at the bright (faint) end. To better quantify this, we define as ‘safe ranges’ the r -band regions where the ratio between the training set and the corresponding input table is larger than 0.1%. These safe ranges, which are highlighted with shaded green regions in the plots, are $r < 22^{\text{m}}$ for VEXAS-DESW, $14^{\text{m}} < r < 23.2^{\text{m}}$ for VEXAS-PSW, and $14.5^{\text{m}} < r < 19.5^{\text{m}}$ for VEXAS-SMW. For magnitudes outside of these intervals, we caution the readers that the classification might be slightly bias due to an under-sampled training set.

A possibility to improve the coverage of an under-sampled training set has been provided in the literature based on a re-weighting procedure of the training sample (e.g., Sánchez et al. 2014; Bonnett et al. 2016). Unfortunately, a similar algorithm only marginally helps in our case since there are objects outside the safe range region with one or more colours which are completely ‘off-models’ (i.e., they have an absolute value larger than the maximum covered by the training set). In a sense, this region is not under-sampled by the training, but not sampled at all. We therefore define an ‘outlier’ as a source for which at least one colour has a value smaller (larger) than the minimum (maximum) value allowed by the training set. The number of outliers as a function of the r -band magnitude is plotted in the bottom panels of Fig. 9. For those objects, the classification is likely to be biased. We note, however, that the number of outliers is much smaller than the number of objects and it represents only $\sim 0.5\%$ of the whole dataset for every VEXAS table. Finally, for the VEXAS-DESW table, where the input is much deeper than the training sample, we report in Appendix A a test which demonstrates that despite an under-sampled training set our classification reaches fair performances.

In conclusion, in the classified VEXAS tables released in this DR2, we inserted a column containing a ‘warning flag’, which takes a value of 0 if the source is within the safe ranges, 1 if it is outside of them but not an outlier, according to the distribution of its colours, and 2 if it is an outlier.

At the bright end, we note that in addition to a poorly sampled training set, there is the bigger problem of saturation, as we demonstrate below. Figure 10 shows the distribution of the r -band magnitude in the three VEXAS tables for objects classified as high-confidence STAR (red), GALAXY (green), or QSO (blue). A further confirmation that the sub-optimal imputation for the VEXAS-SMW does not bias the classification results comes from the fact that the distribution of sources in the r -band is almost identical using the $p_{\text{class}} = 0.7$ obtained from the META_MODEL or those obtained from model #25.

All plots show similar behaviours of the luminosity functions: at the faint end, they all have the same shape except for the depth cut-off, which is dominated by the WISE depth in DESW and PSW and by the optical in SMW; at the bright end, there are secondary bumps and cutoffs which are an artefact of saturation, which occurs at $r \approx 15$ in DESW, $r \approx 14$ in PSW, and $r \approx 11$ in SMW. The saturated objects amount to $\approx 1\%$ in DESW, and an overall correction is

$$\begin{aligned} \text{mag}_{\text{corr}} &= 8.0 + 1.25(\text{mag}_{\text{DES}} - 12.0) \\ &\text{if } \text{mag}_{\text{SM}} < \text{mag}_{\text{DES}} + 1 \text{ and } \text{mag}_{\text{SM}} < 16.5 \end{aligned} \quad (3)$$

for $\text{mag} = (g, r, i)$, and

$$\begin{aligned} z_{\text{corr}} &= 8.0 + 1.1(z_{\text{DES}} - 11.5) \\ &\text{if } z_{\text{SM}} < z_{\text{DES}} + 1 \text{ and } z_{\text{SM}} < 16.5. \end{aligned} \quad (4)$$

Fig. 7. Selected colour-colour and magnitude-colour diagrams of VEXAS high-confidence ($p_{\text{class}} = 0.7$) STAR (red contours), QSO (green contours), and GALAXY (blue contours), over the three optical footprints (VEXAS-DESW, *left column*, VEXAS-PSW, *middle column*, and VEXAS-SMW *right column*), split according to the predicted class. *Bottom right panel:* issue with the imputation for VEXAS-SMW, which is described in detail in Sect. 6 and in Appendix C.3.

Fig. 8. Comparison between the classification obtained with our fiducial META_MODEL and with a single CatBoost Model (#25). We note that 99.5% of the sources are classified in the same class.

Comparing SMW and PSW, there is an overall offset $gri_{\text{PS1}} - gri_{\text{SM}} = 0.2$ and $z_{\text{PS1}} - z_{\text{SM}} = 0.7$ on all magnitudes.

The issue of saturation on the performance of the classification at the bright end can also be seen in the astrometric properties of the objects, for example the proper motions of bright objects classified as quasars, which are discussed in the next section.

6.2. Astrometric validation with Gaia

The latest data release of ESA-*Gaia*, Early Data Release 3 (EDR3, [Gaia Collaboration 2021](#)) provides five astrometric parameters (positions α, δ , proper motions μ_α, μ_δ , and parallaxes ϖ) for ≈ 1.468 billion sources, covering the whole celestial sphere up to $G \lesssim 21^{\text{m}}$ ¹². This makes the *Gaia*-EDR3 an excellent means to test the purity of our catalogue, especially for quasars and stars and, indirectly, also for galaxies (see below).

We cross-matched each VEXAS output table with the *Gaia* EDR3 catalogue using a $1''.5$ matching radius, and we considered sources with defined astrometric parameters. Table 7 lists the number of sources with a cross-match in *Gaia* EDR3, which

¹² This limit corresponds to $r \approx 21^{\text{m}}$ for quasars at $z \leq 3$ ([Proft & Wambsganss 2015](#)).

Fig. 9. *Top:* histogram of the r magnitudes for the input catalogues (red) and the training sample (blue) for each of the VEXAS tables. *Bottom:* fraction of outliers (see text for more details) as a function of the r magnitudes for each of the VEXAS tables.

were split according to our classification pipeline (using the threshold $p_{\text{class}} \geq 0.7$). In all cases, more than 90% of the matched objects were classified as STAR.

To assess the purity of the GALAXY sample, we used the simple argument that, by construction, the *Gaia* catalogue should contain very few galaxies (Robin et al. 2012), so very few of the objects with high p_{GALAXY} should have a match in *Gaia* EDR3. This is, of course, only a rough approximation since there might be a number of galaxies that *Gaia* still detects, but without accurate proper motions and parallaxes, as the *Gaia* astrometric solution fits only for point-like sources (Lindgren et al. 2021). Indeed, resolved objects can be detected in *Gaia* because their G -band magnitudes are >0.1 mag higher than the synthetic magnitude $G_{\text{RB}} = G_{\text{RP}} - 2.5 \log_{10}[1 + 10^{0.4(G_{\text{RP}} - G_{\text{BP}})}]$ from the blue and red passbands, whenever these are available (Agnello & Spiniello 2019). From the cross-match with *Gaia*, we found $\approx 1\,440\,000$, $\approx 1\,100\,000$, and $\approx 1\,910\,000$ objects classified by our algorithm as GALAXY in VEXAS-DESW, VEXAS-PSW, and VEXAS-SMW, respectively. Among these, only $\leq 20\%$ sources have measured proper motions and parallaxes, and the majority are motionless, suggesting that they could indeed be extended sources at $z > 0$ (galaxies). The remaining objects can instead be stars that are misclassified by our algorithm, or very compact galaxies.

To assess the purity of the QSO class, we analysed the proper motions and parallaxes of the objects classified as high confidence quasars as a function of their r -band magnitude. As quasars are very distant sources, they have proper motions of only a few micro-arcseconds, due to different cosmological effects (Bachchan et al. 2016), within the current accuracy of the ESA-*Gaia* relativistic solution (Lindgren et al. 2021). Hence, we checked the proper motions and parallaxes of all the objects classified as quasars and with a match in *Gaia* to

test the assumption that they are indeed zero-proper motion and zero-parallax sources within the systematic errors. In Fig. 11, we plotted the two proper motion components ($\mu_{\alpha*}$, μ_{δ}) and the parallax as a function of the r -band magnitude for high confidence quasars ($p_{\text{qso}} \geq 0.7$) in each of the three cross-matches between VEXAS-DR2 and *Gaia*-EDR3. Clearly, for magnitudes fainter than the saturation limits (~ 15 for VEXAS-DESW, top row; ~ 14 for VEXAS-PSW, middle row; and ~ 11 for VEXAS-SMW, top row; see also Fig. 10), QSO are perfectly consistent to be motionless. This is not true for STAR and GALAXY, as we show in Table 7 where we report the root-mean-square error (RMSE) on the parallax and on each component of the proper motion for all the objects with a match in *Gaia*, split in the three classes.

6.3. Internal validation

In this section, we present a qualitative validation of the classification results obtained comparing the probabilities derived from the different output tables on common objects. We cross-matched the three output tables using a radius of $1.5''$. There are 7 817 243 objects in common between VEXAS-DESW and VEXAS-PSW, 7 743 844 objects in common between VEXAS-PSW and VEXAS-SMW, and, finally, 9 912 142 objects in common between VEXAS-SMW and VEXAS-DESW.

In Fig. 12 we provide a visualisation of this internal validation. For each class of objects (different columns) and for each cross-match between pair of surveys (different lines), we plotted the histograms of the difference between the probabilities computed by either.

The agreement among classification results is excellent as the histograms always peak around 0. More quantitatively, in all cases and for all classes, $<3\%$ of the common objects have probabilities that disagree by more than 0.1 and only between 0.5 and

Fig. 10. Magnitude distribution in r -band of the sources classified in each class for each of the three tables. For the VEXAS-SMW, we show both the classification obtained from the META_MODEL (solid lines) and from the single CatBoost model (#25, dashed), which does not use imputation.

1.2% of the sources have probability that differ by more than 0.5. Thus, for more than 95% of the common sources, the independent classifications obtained from the three tables are in perfect agreement. This simple test also provides further evidence that the sub-optimal imputation for the VEXAS-SMW table does not play a role during the classification step for the great majority of the sources (99.5%).

7. Conclusions and outlook

The increase in the depth, footprint, and sharpness of wide-field imaging surveys posits as many opportunities as challenges. Multiple endeavours, both current and upcoming, aim to collect large spectroscopic samples to map the Milky Way through its stellar content, and the large-scale structure of the Universe through spatial correlations of galaxies or quasars (Kollmeier et al. 2017; de Jong et al. 2019). An outstanding issue in these fibre-fed spectroscopic surveys is the target pre-selection, which was already acknowledged in the preparation of earlier quasar samples (Croom et al. 2003; Dawson et al. 2013), and it is even more pressing in the Southern Hemisphere, where u -band coverage is much shallower. More generally, the use of simple colour cuts to select objects of interest can significantly affect the overall completeness of spectroscopic follow-up samples, with direct consequences on the science. All of this is complicated by the patchiness and partial overlap of different surveys. Here, we have deployed a collection of different machine-learning techniques to circumvent these issues, and we provide membership probabilities in macro-classes.

We have trained a total of 32 different classifiers, with different techniques (ANN, kNN, and CB) and input features, using magnitudes from optical surveys (DES, PS1, and SkyMapper), VISTA (J and K_s), and WISE magnitudes ($W1$ and $W2$). To deal with missing entries, we used a machine-learning based feature imputation. The final classification is an aggregation of all the different classifiers. Each of them has undergone extensive vetting to identify the physical properties that drive its performance, which in turn is quantified in multiple ways to find robust thresholds in the classification scores. While a clear separation can be made between most stars and extragalactic objects, the separation between galaxies and quasars is less abrupt, reflecting the range of central-engine-dominated and host-dominated emission in these objects. As simple colour-magnitude diagnostics show, our classification generalises simple colour-magnitude cuts that have been proposed in the literature (Stern et al. 2005, 2012; Assef et al. 2013; Chehade et al. 2018), but it also deals with the overlap of different classes, especially towards the faint end in WISE magnitudes and towards higher-redshift quasars.

The three VEXAS optical+IR classified tables, with object IDs, coordinates, optical and infrared magnitudes, including the imputed ones (which are flagged), and the probability to belong to each class, are publicly released and available for the scientific community as part of the VEXAS Data Collection (DR2) via the ESO Phase 3¹³. The machine-learning classification pipeline and the code are available on the VEXAS Github repository¹⁴.

The density of extragalactic objects varies with the survey depth, quite expectedly. Considering only high-confidence objects for which the probability of belonging to the class is $p_{\text{class}} \geq 0.7$, our classifiers yielded 111 qso deg⁻² for the VEXAS-DESW footprint (≈ 4900 deg²) and 103 qso deg⁻² for the VEXAS-PSW footprint (≈ 3800 deg²). These numbers roughly meet the requirements of the 4MOST selection for baryon acoustic oscillations measurements. The

¹³ The active link to the VEXAS collection through the [Science Portal](#) and the [Catalog Interface](#) are given in this footnote in the online version. While we are processing the Phase 3 documentation and format, we release the table via a temporary repository ([here](#)).

¹⁴ <https://github.com/VEXAS-team/VEXAS-DR2>

Fig. 11. Dependencies of astrometric parameters on r -band magnitudes for objects classified as quasars ($p_{\text{QSO}} > 0.7$) from each survey. From left to right we plotted the two proper motion components (μ_{α^*} , μ_{δ}) and the parallax. Solid lines and filled areas represent the median values and standard deviations of the parameters in each magnitude bin.

Table 7. Cross-match with *Gaia* EDR3 for the three VEXAS classified tables and for each class of objects.

Table	Class	Total matches	N_{Good}	N_{Good} ($r > 15^{\text{m}}$)	μ_{α^*} (mas yr $^{-1}$)	μ_{δ} (mas yr $^{-1}$)	ϖ (mas)
VEXAS-DESW	STAR	12 397 335	11 815 841	10 740 398	5.566 ± 11.464	-3.722 ± 11.891	1.012 ± 1.119
	QSO	281 681	237 191	235 311	0.041 ± 1.117	-0.051 ± 1.208	-0.020 ± 0.751
	GALAXY	1 445 987	169 198	158 700	1.367 ± 5.782	-1.354 ± 5.907	0.067 ± 1.343
VEXAS-PSW	STAR	10 108 741	9 580 815	8 885 919	3.810 ± 11.345	-6.014 ± 10.819	0.986 ± 1.142
	QSO	205 789	172 019	171 940	0.155 ± 2.035	-0.544 ± 2.464	0.007 ± 0.842
	GALAXY	1 098 769	171 344	170 264	1.635 ± 6.326	-3.428 ± 7.059	0.304 ± 1.308
VEXAS-SMW	STAR	29 551 196	29 182 681	26 150 059	2.750 ± 10.299	-4.633 ± 10.748	0.872 ± 0.995
	QSO	68 857	65 442	65 364	0.322 ± 3.544	-0.633 ± 3.774	0.074 ± 0.561
	GALAXY	1 911 503	408 894	407 107	1.766 ± 8.58	-3.146 ± 9.498	0.681 ± 1.380

Notes. From left to right we list the following: the total number of identified sources, the number of sources with measured astrometric parameters (N_{Good}), the same number but only for sources that are not saturated ($r > 15^{\text{m}}$), the two proper motion components, and the parallax, with the corresponding standard deviations.

density drops to $\approx 10 \text{ qso deg}^{-2}$ for the VEXAS-SMW footprint ($\approx 9300 \text{ deg}^2$).

All in all, the combined survey depth is the limiting factor. The VEXAS-SMW footprint is limited by the shallow optical coverage of SkyMapper, which also has less uniform coverage than DES and PanSTARRS. A solution would be to consider only NIR and mid-IR magnitudes in the classification, which in turn requires more uniform coverage in the NIR and smaller uncertainties in the WISE magnitudes. This may be obtained thanks to unWISE, a re-processing of WISE, and NeoWISE imaging (Lang 2014; Schlafly et al. 2019), or with forced photometry (on SkyMapper and unWISE cutouts) based on VISTA detections.

We caution that these samples are only the very first step for cosmological measurements, which also require spectroscopic redshifts and well characterised coverage maps. The spectroscopic follow-up is already the aim of southern surveys (4MOST, de Jong et al. 2012; SDSS-V, Kollmeier et al. 2017). Since this paper is focused on the classification techniques and their application to VEXAS, the coverage maps are outside the scope of this work and will be part of a subsequent release.

Acknowledgements. The authors wish to thank the anonymous referee for a very constructive and useful report, which improved the quality of the final manuscript. CS is supported by a Hintze Fellowship at the Oxford Centre

Fig. 12. Distribution of the probability differences on common objects in pairs of tables. *Top row:* VEXAS-DESW $\times$ VEXAS-PSW. *Middle row:* VEXAS-PSW $\times$ VEXAS-SMW. *Bottom row:* VEXAS-SMW $\times$ VEXAS-DESW. Each column shows one class of objects (left STAR, middle QSO, right GALAXY).

for Astrophysical Surveys, which is funded through generous support from the Hintze Family Charitable Foundation. AA is supported by a grant from VILLUM FONDEN (project number 16599). This project is funded by the Danish council for independent research under the project ‘Fundamentals of Dark Matter Structures’, DFF – 6108-00470. This research has made use of the services of the ESO Science Archive Facility and of the cross-match service provided by CDS, Strasbourg. The authors are thankful to Laura Mascetti and the ESO Archive Science Group Team, led by Magda Arnaboldi, for the precious help in making the VEXAS tables Phase-3 compliant and releasing them through the Phase-3 Science Archive. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP),

Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University. Wigglez acknowledges financial support from The Australian Research Council (grants DP0772084, LX0881951 and DP1093738 directly for the WiggleZ project, and grant LE0668442 for programming support), Swinburne University of Technology, The University of Queensland, the Anglo-Australian Observatory, and The Gregg Thompson Dark Energy Travel Fund at UQ. GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC

(Australia), the AAO, and the participating institutions. Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G. This paper uses data from the VIMOS Public Extragalactic Redshift Survey (VIPERS). VIPERS has been performed using the ESO Very Large Telescope, under the “Large Programme” 182.A-0886. The participating institutions and funding agencies are listed at <http://vipers.inaf.it>. This research uses data from the VIMOS VLT Deep Survey, obtained from the VVDS database operated by Cesam, Laboratoire d’Astrophysique de Marseille, France. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. Based on observations made with ESO Telescopes at the La Silla or Paranal Observatories under programme ID(s) 179.A-2005(A), 179.A-2005(B), 179.A-2005(C), 179.A-2005(D), 179.A-2005(E), 179.A-2005(F), 179.A-2005(G), 179.A-2005(H), 179.A-2005(I), 179.A-2005(J), 179.A-2005(K), 179.A-2005(L), 179.A-2005(M), 179.A-2005(N), 179.A-2005(O).

References

- Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2003, *AJ*, **126**, 2081
- Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. 2018, *ApJS*, **239**, 18
- Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, *ApJS*, **235**, 42
- Ahumada, R., Prieto, A., Almeida, C., et al. 2020, *ApJS*, **249**, 3
- Agnello, A., & Spiniello, C. 2019, *MNRAS*, **489**, 2525
- Agnello, A., Treu, T., Ostrovski, F., et al. 2015, *MNRAS*, **454**, 1260
- Agnello, A., Schechter, P. L., Morgan, N. D., et al. 2018a, *MNRAS*, **475**, 2086
- Agnello, A., Lin, H., Kuropatkin, N., et al. 2018b, *MNRAS*, **479**, 4345
- Andreani, P., Cimatti, A., Loinard, L., et al. 2000, *A&A*, **354**, L1
- Anguita, T., Schechter, P. L., Kuropatkin, N., et al. 2018, *MNRAS*, **480**, 5017
- Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, *ApJ*, **772**, 26
- Bachchan, R. K., Hobbs, D., & Lindegren, L. 2016, *A&A*, **589**, A71
- Bañados, E., Venemans, B. P., Decarli, R., et al. 2016, *ApJS*, **227**, 11
- Baldry, I. K., Liske, J., Brown, M. J. I., et al. 2018, *MNRAS*, **474**, 3875
- Blanton, M., Bershad, M., Abolfathi, B., et al. 2017, *AJ*, **154**, 28
- Bochanski, J. J., Munn, J. A., Hawley, S. L., et al. 2007, *AJ*, **134**, 2418
- Bock, D. C.-J., Large, M. I., & Sadler, E. M. 1999, *AJ*, **117**, 1578
- Boller, T., Freyberg, M., & Truemper, J. 2014, *The X-ray Universe*, **2014**, 40
- Boller, T., Freyberg, M. J., Trümper, J., et al. 2016, *A&A*, **588**, A103
- Bonifacio, P., Monai, S., & Beers, T. C. 2000, *AJ*, **120**, 2065
- Bonnert, C., Troxel, M. A., Hartley, W., et al. 2016, *Phys. Rev. D*, **94**, 042005
- Brown, M. J. I., Jannuzi, B. T., Dey, A., & Tiede, G. P. 2005, *ApJ*, **621**, 41
- Bundy, K., Bershad, M. A., Law, D. R., et al. 2015, *ApJ*, **798**, 7
- Carnall, A. C., Shanks, T., Chehade, B., et al. 2015, *MNRAS*, **451**, L16
- Carnero Rosell, A., Santiago, B., dal Ponte, M., et al. 2019, *MNRAS*, **489**, 5301
- Cassata, P., Cimatti, A., Kurk, J., et al. 2008, *A&A*, **483**, L39
- Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]
- Chehade, B., Carnall, A. C., Shanks, T., et al. 2018, *MNRAS*, **478**, 1649
- Childress, M. J., Lidman, C., Davis, T. M., et al. 2017, *MNRAS*, **472**, 273
- Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. 2020, *A&A*, **639**, A84
- Clerc, N., Merloni, A., Zhang, Y.-Y., et al. 2016, *MNRAS*, **463**, 4490
- Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2001, *MNRAS*, **322**, L29
- Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2003, *MNRAS*, **349**, 1397
- Cross, N. J. G., Collins, R. S., Mann, R. G., et al. 2012, *A&A*, **548**, A119
- Cutri, R. M., et al. 2013, VizieR Online Data Catalog:II/328
- Daddi, E., Cimatti, A., & Renzini, A. 2000, *A&A*, **362**, L45
- Dark Energy Survey Collaboration (Abbott, T., et al.) 2016, *MNRAS*, **460**, 1270
- Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, *AJ*, **145**, 10
- Dawson, K., Kneib, J.-P., & Percival, W. 2016, *AJ*, **151**, 44
- de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, *Proc. SPIE*, **8446**, 84460T
- de Jong, J. T. A., Kuijken, K., Applegate, D., et al. 2013, *The Messenger*, **154**, 44
- de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, *The Messenger*, **175**, 3
- Diehl, H. T., & Dark Energy Survey Collaboration 2012, *Phys. Proc.*, **37**, 1332
- Dorogush, A. V., Ershov, V., & Gulin, A. 2018, ArXiv eprints [arXiv:1810.11363]
- Drinkwater, M. J., Jurek, R. J., Blakeet, C., et al. 2010, *MNRAS*, **401**, 1429
- Drinkwater, M. J., Byrne, Z. J., Blake, C., et al. 2018, *MNRAS*, **474**, A451
- Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, *MNRAS*, **413**, 971
- Drlica-Wagner, A., Bechtol, K., Rykoff, E. S., et al. 2015, *ApJ*, **813**, 109
- Dwelly, T., Salvato, M., Merloni, A., et al. 2017, *MNRAS*, **469**, 1065
- Edge, A., Sutherland, W., Kuijken, K., et al. 2013, *The Messenger*, **154**, 32
- Elston, R., Rieke, G. H., & Rieke, M. J. 1988, *ApJ*, **331**, L77
- Elston, R., Rieke, M. J., & Rieke, G. H. 1989, *ApJ*, **341**, 80
- Emerson, J., McPherson, A., & Sutherland, W. 2006, *The Messenger*, **126**, 41
- Friedman, J. H. 2000, *Ann. Stat.*, **29**, 1189
- Gaia Collaboration (Brown, A. G. A., et al.) 2021, *A&A*, **649**, A1
- Gilmore, G., Randich, S., Asplund, M., et al. 2012, *The Messenger*, **147**, 25
- González-Fernández, C., Hodgkin, S. T., Irwin, M. J., et al. 2018, *MNRAS*, **474**, 5459
- Gonzalez-Perez, V., Baugh, C. M., Lacey, C. G., & Kim, J.-W. 2011, *MNRAS*, **417**, 517
- Helmi, A. 2020, *ARA&A*, **58**, 205
- Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, *MNRAS*, **465**, 1454
- Hodgkin, S. T., Irwin, M. J., Hewett, P. C., et al. 2009, *MNRAS*, **394**, 675
- Inada, N., Oguri, M., Shin, M.-S., et al. 2012, *AJ*, **143**, 119
- Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, *ApJ*, **873**, 111
- Jones, D. H., Saunders, W., Colless, M., et al. 2004, *MNRAS*, **355**, 747
- Jones, D. H., Read, M. A., Saunders, W., et al. 2009, *MNRAS*, **399**, 683
- Kaiser, N., Aussel, H., Burke, B. E., et al. 2002, *Proc. SPIE*, **154**
- Khramtsov, V., Akhmetov, V., & Fedorov, P. 2020, *A&A*, **644**, A69
- Khramtsov, V., Sergeyev, A., Spiniello, C., et al. 2019, *A&A*, **632**, A56
- Kingma, D. P., & Ba, J. 2014, ArXiv eprints [arXiv:1412.6980]
- Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. 2017, *Advances in neural information processing systems (NIPS)*, **30**, 971
- Kollmeier, J. A., Zasowski, G., Rix, H. W., et al. 2017, ArXiv eprints [arXiv:1711.03234]
- Kong, X., Daddi, E., Arimoto, N., et al. 2006, *ApJ*, **638**, 72
- Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, *MNRAS*, **454**, 3500
- Kuncheva, L. I. 2004, *Combining Pattern Classifiers: Methods and Algorithms* (John Wiley & Sons)
- Lang, D. 2014, *AJ*, **147**, 108
- Lemon, C. A., Auger, M. W., McMahon, R. G., et al. 2018, *MNRAS*, **479**, 5060
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. 2015, *Eur. J. Oper. Res.*, **247**, 124
- Le Fevre, O., Cassata, P., Cucciati, O., et al. 2013, *A&A*, **559**, A14
- Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021, *A&A*, **649**, A2
- Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, *MNRAS*, **452**, 2087
- Mainzer, A., Bauer, J., Grav, T., et al. 2011, *ApJ*, **731**, 53
- Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., et al. 2003, *ApJ*, **599**, 1082
- Majewski, S., APOGEE Team, & APOGEE-2 Team 2016, *Astron. Nachr.*, **337**, 863
- Martin, D. C., Fanston, J., Schiminovich, D., et al. 2005, *ApJ*, **619**, L1
- Maturi, M., Bellagamba, F., Radovich, M., et al. 2019, *MNRAS*, **485**, 498
- Mauch, T., Murphy, T., Buttery, H. J., et al. 2003, *MNRAS*, **342**, 1117
- McCarthy, P. J., Persson, S. E., & West, S. C. 1992, *ApJ*, **386**, 52
- McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, *A&A*, **544**, A156
- McInnes, L., & Healy, J. 2018, ArXiv e-prints [arXiv:1802.03426]
- McMahon, R. G., Banerji, M., Gonzalez, E., et al. 2013, *The Messenger*, **154**, 35
- Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, *New Astron.*, **15**, 433
- Morganson, E., Green, P. J., Anderson, S. F., et al. 2015, *ApJ*, **806**, 244
- Muñoz, J. A., Falco, E. E., Kochanek, C. S., et al. 1998, *Ap&SS*, **263**, 51
- Myoung-Jong, K., Sung-Hwan, M., & Ingoo, H. 2006, *Expert Syst. Appl.*, **31**, 241
- Nakoneczny, S., Bilicki, M., Solarz, A., et al. 2019, *A&A*, **624**, A13
- Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, *ApJS*, **208**, 5
- Nidever, D. L., Dey, A., Fasbender, K., et al. 2021, *AJ*, **161**, 192
- Oguri, M., & Marshall, P. J. 2010, *MNRAS*, **405**, 2579
- Ostrovski, F., Lemon, C. A., Auger, M. W., et al. 2018, *MNRAS*, **473**, L116
- Petrillo, C. E., Tortora, C., Vernardos, G., et al. 2019, *MNRAS*, **484**, 3879
- Pozzetti, L., Hoekstra, H., Röttgering, H. J. A., et al. 2000, *A&A*, **361**, 535
- Proft, S., & Wambsganss, J. 2015, *A&A*, **574**, A46
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A., 2018, *Advances in Neural Information Processing Systems*, **31**, 6638
- Reed, S. L., McMahon, R. G., Banerji, M., et al. 2015, *MNRAS*, **454**, 3952
- Robin, A., Luri, X., Reylé, C., et al. 2012, *A&A*, **543**, A100
- Roche, N. D., Almaini, O., Dunlop, J., Ivison, R. J., & Willott, C. J. 2002, *MNRAS*, **337**, 1282
- Roy, N., Napolitano, N. R., La Barbera, F., et al. 2018, *MNRAS*, **480**, 1057
- Salvato, M., Buchner, J., Budavari, T., et al. 2018, VizieR Online Data Catalog:JMNRA/473/4937
- Sánchez, C., Carrasco Kind, M., Lin, H., et al. 2014, *MNRAS*, **445**, 1482
- Saracco, P., Longhetti, M., Severgnini, P., et al. 2005, *MNRAS*, **357**, L40
- Schlaflly, E. F., Meisner, A. M., & Green, G. M. 2019, *ApJS*, **240**, 30
- Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, *ApJ*, **500**, 525
- Scodiegio, M., Guzzo, L., Garilli, B., et al. 2018, *A&A*, **609**, A84
- Scoville, N., Aussel, H., Benson, A., et al. 2007, *ApJS*, **172**, 150
- Shanks, T., Metcalfe, N., Chehade, B., et al. 2015, *MNRAS*, **451**, 4238

- Shin, J., Shim, H., Hwang, H. S., et al. 2017, *J. Korean Astron. Soc.*, **50**, 61
- Shipp, N., Drlica-Wagner, A., Balbinot, E., et al. 2018, *ApJ*, **862**, 114
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, **131**, 1163
- Spiniello, C., & Agnello, A. 2019, *A&A*, **630**, A146
- Spiniello, C., Agnello, A., Napolitano, N. R., et al. 2018, *MNRAS*, **480**, 1163
- Spiniello, C., Sergeyev, A. V., Marchetti, L., et al. 2019a, *MNRAS*, **485**, 5086
- Spiniello, C., Agnello, A., Sergeyev, A. V., et al. 2019b, *MNRAS*, **483**, 3888
- Steidel, C. C., Giavalisco, M., Pettini, M., et al. 1996, *ApJ*, **462**, L17
- Stekhoven, D. J., & Bühlmann, P. 2011, *Bioinformatics*, **28**.1, 112
- Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, *ApJ*, **631**, 163
- Stern, D., Assef, R. J., Benford, D. J., et al. 2012, *ApJ*, **753**, 30
- Stiavelli, M., Treu, T., Carollo, C. M., et al. 1999, *A&A*, **343**, L25
- Sutherland, W. 2012, *Science from the Next Generation Imaging and Spectroscopic Surveys*, 40
- Taylor, M. B. 2005, *Astronomical Data Analysis Software and Systems XIV*, **347**, 29
- Thompson, D., Beckwith, S. V. W., Fockenbrock, R., et al. 1999, *ApJ*, **523**, 100
- Tonry, J. L., Stubbs, C. W., Lykke, K. R., et al. 2012a, *ApJ*, **750**, 99
- Tonry, J. L., Stubbs, C. W., Kilic, M., et al. 2012b, *ApJ*, **745**, 42
- Treu, T., Agnello, A., & Strides Team 2015, *Am. Astron. Soc. Meeting Abstr.*, **225**, 318.04
- Truemper, J. 1982, *Adv. Space Res.*, **2**, 241
- Venemans, B. P., Findlay, J. R., Sutherland, W. J., et al. 2013, *ApJ*, **779**, 24
- Venemans, B. P., Verdoes Kleijn, G. A., Mwebaze, J., et al. 2015, *MNRAS*, **453**, 2259
- Vikram, V., Chang, C., Jain, B., et al. 2015, *Phys. Rev. D*, **92**
- Voges, W., Aschenbach, B., Boller, T., et al. 1999, *A&A*, **349**, 389
- Voges, W., Aschenbach, B., Boller, T., et al. 2000, *IAU Circ.*, **7432**, 3
- Yuan, F., Lidman, C., Davis, T. M., et al. 2015, *MNRAS*, **452**, 3047
- Watson, M. G., Auguères, J.-L., Ballet, J., et al. 2001, *A&A*, **365**, L51
- Whitmore, B. C., Allam, S. S., Budavári, T., et al. 2016, *AJ*, **151**, 134
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868
- Wolf, C., Bian, F., Onken, C. A., et al. 2018, *PASA*, **35**, e024

Appendix A: Testing the classification at the faint end of VEXAS-DESW

As described in detail in Sect. 6.1, the r -band magnitude ranges covered by the input tables are larger than those covered by the corresponding tables. This is particularly true for the VEXAS-DESW table, which reaches magnitudes as faint as $r \sim 25^m$ (see Fig. 9). Therefore, in this appendix, we carry out a test on this table to assess the performance of our ML classification at such faint magnitudes.

Unfortunately, the number of faint objects with a footprint within VEXAS-DESW and with a secure spectroscopic classification is very limited and mainly made of galaxies. We considered the SDSS test sub-sample (see Sect. 3.8) and three other different spectroscopic surveys: the DEEP2 redshift survey DR 4 (Newman et al. 2013), the VIMOS Public Extragalactic Redshift Survey (VIPERS) DR2 (Scodeggio et al. 2018), and the VIMOS VLT Deep Survey (VVDS) final data release (Le Fevre et al. 2013). We selected all objects with magnitudes fainter than 22^m as they are outside the safe ranges defined in Sect. 6.1.

True class	STAR	77 (90.59%)	1 (1.18%)	7 (8.24%)
	QSO	0 (0.00%)	82 (71.30%)	33 (28.70%)
	GALAXY	34 (0.47%)	44 (0.60%)	7201 (98.93%)
		Predicted class		
		STAR	QSO	GALAXY

Fig. A.1. Confusion matrix obtained for the META_MODEL on the deep external spectroscopic samples for the VEXAS-DESW.

We then applied some criteria to filter out unreliable spectroscopic measurements. In particular, for DEEP2, we required that the spectroscopic class is not empty; for VIPERS, we required that the integer part of z_{flg} flag is $>2^{15}$. For VVDS, we accepted sources with $ZFLAGS$ equal to 3, 4, or 9 as galaxies, sources with $ZFLAGS$ equal to 13, 14, or 19 as quasars, and sources with a redshift of $z < 0.001$ as stars.

Finally, we matched the coordinates of all sources with secure spectroscopic classification for each survey with the

VEXAS-DESW classified table, using a matching radius of $1.5''$. In this way, we obtained a set of 7279 GALAXY, 115 QSO, and 85 STAR with $r > 22^m$.

Figure A.1 shows the comparison between the ML predicted and spectroscopically measured classes in the form of a confusion matrix. Almost all galaxies were classified correctly from our pipeline ($\sim 99\%$ of the total number of objects classified as such). For stars and quasars the performance of our ensemble learning is less optimal, but still acceptable, with $\sim 10\%$ of STAR and $\sim 30\%$ of QSO misclassified (as galaxies mostly). We note that the lower performance obtained in the QSO class might due to the fact that the spectroscopic surveys we used here are targeting galaxies. This means that probably the majority of these quasars have bright host galaxies, which contaminate their colours, making the classification harder.

In conclusion, we can state that, with respect to the classification obtained within the safe regions (see Fig. 4, top panel), the performance of our ML classification is worse, as expected. However, it is still highly reliable, at least for the GALAXY class.

Appendix B: Noisy spectroscopic datasets

In the main text of this manuscript, we have argued that the spectroscopic classification sample that was used to train the ensemble learning was not perfect, despite the manual cleaning we applied. Here, we show an experiment to confirm this statement. For this experiment, we only used the VEXAS-PSW table, as an example, but this result could be easily replicated for the other cases, assuming the uniformity of the label noise in spectroscopic surveys across the sky.

We used a dimensionality reduction algorithm called Uniform Manifold Approximation and Projection (UMAP, McInnes & Healy 2018). The utility of these types of algorithms has already been shown in many publications, for example Nakoneczny et al. (2019) and Clarke et al. (2020). We applied the UMAP algorithm to the imputed optical+IR magnitudes from the VEXAS-PSW table, reducing the input parameter space of 9 magnitudes ($grizy$ from PS, K_S and J from VISTA, and $W1$ and $W2$ from WISE) to only two values. This makes it possible to produce 2D figures where the objects belonging to a different class lie on different regions, as shown in Fig. B.1 for VEXAS-PSW (upper, bigger panel) and for the six spectroscopic tables used in Sect. 3. We colour-coded the three classes consistently with the main body of the paper: STAR are shown in red, QSO in green, and GALAXY in blue. In the upper panel, all spectroscopic matches are plotted and, visually, each family is clustered in a different region in the UMAP reduced parameter space. However, when we split the two dimensional projection into three different panels, one for each of the different classes, as we do in the smaller panels, survey by survey, we can clearly see that there are objects belonging to a given class but laying in a region corresponding to another one. This is particularly true for stars (red points), and partially for galaxies (blue points).

¹⁵ See [the survey documentation](#) provided as a link in the online version of this manuscript.

Fig. B.1. Two-dimensional projection of the VEXAS-PSW sources with a spectroscopic match created with UMAP. Each point represents a source within the two-dimensional reduced space, and it is colour coded according to its spectroscopic classification. *Bottom lower panel:* each line of three-plots is relative to a different spectroscopic survey, as indicated by the titles. Each panel within a line highlights a different class of objects. Grey points represent the sources belonging to the other two classes instead.

Appendix C: Magnitude imputation

In the main body of the paper, we have imputed magnitudes in $grizy(u)JK_S W1W2$ using an AE ANN. The need for an intelligent feature imputation is an essential step to prevent reducing the size of the VEXAS DR2 classified tables dramatically, which prevents us from fulfilling the final VEXAS purpose of building a homogeneous multi-band photometric catalogue with a sky coverage that is as large as possible. However, we have also shown that a possible issue with the imputation of the VEXAS-SMW table, for which the percentage of missing magnitudes in the optical is much larger than that in the other two tables, might exist. In this appendix, we first describe in more detail the AE architecture, then we provide results of the magnitude imputation carried out on a training sample for which the ‘true’ magnitudes were known a priori. Finally, we focus on the VEXAS-SMW case, showing evidence supporting the fact that the object classification based on the ensemble learning is not affected by the non-optimal imputation.

C.1. AE architecture

With respect to the traditional AE architecture, our imputer presents a few modifications. First, we added a skip-connection¹⁶, which concatenates the output AE layer (with a size equal to the number of input magnitudes) and the input magnitudes. The skip-connection simply creates a layer that is $2\times$ the size of the number of magnitudes. We also added one neuron to this layer, which contains the stellarity parameter P_{STAR} . This layer was then connected with the final layer, where each output represents an imputed magnitude. The activation function of each layer is a scaled exponential unit, (SELU, Klambauer et al. 2017). Finally, we also added the additional information on the stellarity of the objects (P_{STAR}) at the input layer to further help the AE learn useful representations and obtain a more accurate magnitude imputation as a result. The AE architecture is summarised in Table C.1.

Table C.1. AE imputer architecture construction.

Description	Size	Activation
Input	10	–
Encoder	7	SELU
	5	SELU
Decoder	7	SELU
	9	SELU
Skip-connection	19	–
Output	9	Linear

Notes. Each row represents a layer with its size (number of neurons) and activation function. The input layer refers to the input set of magnitudes with missed values ($grizy(u)JK_S W1W2$) and additional features (P_{STAR} in our case), and the output layer refers to the imputed magnitudes. The skip-connection concatenates the input and output of the decoder and has a size equal to $2\times$ the number of magnitudes plus an additional input value (P_{STAR}).

The input magnitude entries that were masked at random were replaced with zeros. For computational ease, we divided all magnitudes by 25, so that they all belonged to the $0 <$

¹⁶ The idea of skip-connection is introduced in the U-Net network for the segmentation of images.

(mag/25) < 1 interval. We trained the AE with a logcosh loss function and the NADAM optimiser¹⁷.

C.2. Training and testing the performance of the AE imputation

To train the AE imputer, we created three tables, one for each survey, comprising only objects with the entire set of measured magnitudes ($g, r, i, z, y, J, K_S, W1, W2$ for DESW and PSW, and $u, g, r, i, z, J, K_S, W1, W2$ for SMW). Then we masked out these magnitudes for a number of objects proportional to the complementary fraction given in Table 3, with the exception of the W1-band, for which we assumed a rate of measurements equal to 99.9% and thus masked out this magnitude for only 0.1% of the tables. This was done to keep the training of the AE as general as possible. In this way, we retrieved 28 537 891 (81% of the full table), 15 592 387 (71%), and 5 923 761 (19%) sources for DESW, PSW, and SMW, respectively. As already noted above, the portion of sources with the entire set of measured magnitudes is much smaller for VEXAS-SMW, compared to the other two tables. This is probably causing a sub-optimal performance for the imputation in SMW, as we discuss in the next section.

The sample was then split into training and validation samples in a proportion of 85%–15%. At this point, the AE was trained over 150 epochs, with a batch size equal to 256 sources. We scheduled the learning rate value to converge deeper into the loss minimum: if our loss on the validation sample did not change during seven epochs, we decreased the value of the learning rate to a factor of 0.1.

Finally, we ran the imputation pipeline on the validation objects (the remaining 15%) with ‘hidden’ magnitudes. We therefore obtained imputed values which we then compared to the ‘true’ ones. The results of this test are presented in Figs. C.1–C.3 where we plotted the ‘true’ magnitudes versus the imputed ones for the bands on which imputation was computed for the VEXAS-DESW, VEXAS-PSW, and VEXAS-SMW sources, respectively.

The magnitudes are plotted in their original units, which are AB for the optical and Vega for the infrared. We note that we replicated, in scale, the percentage of objects with missing magnitudes to be imputed in each band as presented in Table 3.

Qualitatively, in all the plots, the imputed magnitudes agree well with the ‘true’ ones, which already demonstrate the validity of the imputation process. For each band and in each table,

we computed the coefficient of determination R^2 between true and imputed magnitudes. For DESW and PSW, we obtained $0.88 < R^2 < 0.98$ for all magnitudes. For the SMW, almost all imputed magnitudes follow the one-to-one line with $R^2 > 0.95$. In particular g, r, i, z magnitudes were recovered with $R^2 > 0.99$, while the worst scores were obtained for the u and $W1$ bands. The imputation for the $W1$ magnitude shows the worst results in all tables, possibly due to the relatively small fraction of sources with a missing $W1$ magnitude in the training sample¹⁸. Anyways, this would not affect the classification results at all, since in reality we did not perform any imputation on $W1$ because by definition 100% of the VEXAS objects have measured $W1$ magnitudes. We also note that the VEXAS-SMW table consists mostly of STAR (see e.g., Table 5), so the obtained scores could not be representative of the other two classes. This holds particularly for quasars: among the whole SMW training sample for the AE, quasars constitute only $\approx 0.1\%$ (≈ 8000 objects), which is possibly too small of a fraction to obtain a satisfactory imputation of some magnitudes for this subclass of sources. Hence, in the next section we focus in particular on VEXAS-SMW, showing that indeed the imputation is sub-optimal, but that this does not bias the classification through the ensemble learning described in the main body.

C.3. A sub-optimal imputation for VEXAS-SMW

In Fig. 7 of the main body, we show that, while for VEXAS-DESW and VEXAS-PSW a clear (albeit partial) separation between objects belonging to different classes is visible, for the VEXAS-SMW case (right panel) this is not the case. To demonstrate that this is indeed caused by the sub-optimal performance of the AE imputation, in Fig. C.4, we plotted high confidence objects ($p_{\text{class}} = 0.7$) for which none of the magnitudes were missing (thus where the imputation was not necessary). It is clear that for VEXAS-DESW (left) and VEXAS-PSW (middle) nothing changes; whereas for VEXAS-SMW (right), the confusion between objects classified in different classes is substantially reduced and three different regions are clearly visible.

Importantly, this sub-optimal imputation does not affect the classification of the objects in the three classes obtained through the ensemble learning (see Fig. 8). This is, in fact, flexible enough and based on different models, of which only some use imputation.

¹⁷ http://cs229.stanford.edu/proj2015/054_report.pdf

¹⁸ We remind the readers that the frequency of masking magnitudes to create the training sample was taken from Table 3, and a very small percentage was also hidden for $W1$ as a further test.

Fig. C.1. Results of the magnitude imputation on the training sample for the VEXAS-DESW table. The red line shows the one-to-one correlation.

Fig. C.2. Results of the magnitude imputation on the training sample for the VEXAS-PSW table. The red line shows the one-to-one correlation.

Fig. C.3. Results of the magnitude imputation on the training sample for the VEXAS-SMW table. The red line shows the one-to-one correlation.

Fig. C.4. Selected colour-colour and magnitude-colour diagrams of VEXAS high-confidence ($p_{\text{class}} = 0.7$) STAR (red contours), QSO (green contours), and GALAXY (blue contours), for only objects without missing magnitudes, over the three optical footprints (VEXAS-DESW, *left column*; VEXAS-PSW, *middle column*; and VEXAS-SMW, *right column*), split according to the predicted class. While, for the first two tables, the situation is unchanged with respect to Fig. 7; for the VEXAS-SMW, the separation obtained without imputation is much more net between the three classes.