Applications of machine-learning algorithms for infrared colour selection of Galactic Wolf–Rayet stars

Giuseppe Morello,^{1,2★} P. W. Morris,² S. D. Van Dyk,² A. P. Marston³
and J. C. Mauerhan⁴

¹Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

²IPAC, Caltech, 1200 E. California Blvd, Pasadena, CA 91125, USA

³ESA/JWST, Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

⁴Department of Astronomy, University of California, Berkeley, CA 94720-3411, USA

Accepted 2017 September 21. Received 2017 August 3; in original form 2016 November 12

ABSTRACT

We have investigated and applied machine-learning algorithms for infrared colour selection of Galactic Wolf–Rayet (WR) candidates. Objects taken from the *Spitzer* Galactic Legacy Infrared Midplane Survey Extraordinaire (GLIMPSE) catalogue of the infrared objects in the Galactic plane can be classified into different stellar populations based on the colours inferred from their broad-band photometric magnitudes [J , H and K_s from 2 Micron All Sky Survey (2MASS), and the four *Spitzer*/IRAC bands]. The algorithms tested in this pilot study are variants of the k -nearest neighbours approach, which is ideal for exploratory studies of classification problems where interrelations between variables and classes are complicated. The aims of this study are (1) to provide an automated tool to select reliable WR candidates and potentially other classes of objects, (2) to measure the efficiency of infrared colour selection at performing these tasks and (3) to lay the groundwork for statistically inferring the total number of WR stars in our Galaxy. We report the performance results obtained over a set of known objects and selected candidates for which we have carried out follow-up spectroscopic observations, and confirm the discovery of four new WR stars.

Key words: methods: observational – methods: statistical – stars: evolution – stars: massive – stars: Wolf–Rayet – infrared: stars.

1 INTRODUCTION

Wolf–Rayet (WR) stars are known for their spectacular spectral energy distributions (SEDs), exhibiting broad emission lines that are prominent over most observable wavelengths and strong infrared (IR) excesses due to Bremsstrahlung (free–free) emission. The emission line spectra are broadly classified according to the presence of nitrogen (the WN sequence), carbon (WC) or oxygen (WO), reflecting the abundances of CNO-cycled material from the stellar cores which have reached the surface via rotational mixing and the evaporation of external layers by strong, radiatively driven stellar winds; see van der Hucht (2001), Crowther (2007) and references therein. WR stars are believed to be descended from post-main-sequence (M–S) O-type stars with initial masses $M_{\text{init}} \gtrsim 25 M_{\odot}$ and effective temperatures $T_{\text{eff}} \gtrsim 40\,000$ K, possibly involving an intermediate luminous blue variable phase or a red supergiant phase, depending on M_{init} (Crowther 2007; Langer 2012). Massive stars in binary systems can develop into WR stars due to strip-

ping by a companion rather than inherent mass-loss due to a stellar wind (Eldridge, Izzard & Tout 2008). Such processes significantly alter the chemical composition and distribution of the interstellar medium. Further significant impact follows their explosions as Type Ib/c supernovae (Smartt 2009).

Because of their high luminosity and short lifetimes, i.e. 10^5 – 10^6 yr (Ekström et al. 2012; Georgy et al. 2012), WR stars are excellent tracers of recent star formation. A map of Galactic star-forming regions is key to understanding the history of the Milky Way, in context with the stellar demographics and the high-mass end of the stellar initial mass function. The relative abundances of giant stars in different evolutionary phases can be used to constrain the relevant time-scales in evolutionary models. This task is made difficult by dust obscuration along the Galactic plane. Estimates of the number of WR stars in the Galaxy are in the range ~ 1200 – 6500 (van der Hucht 2001; Shara et al. 2009; Rosslowe & Crowther 2015a,b). The current census of 634 WR stars¹ is much

* E-mail: giuseppe.morello.11@ucl.ac.uk

¹ This number is taken from the online catalogue by Paul Crowther (<http://pacrowther.staff.shef.ac.uk/WRcat>) on 2016 October 7.

smaller than the lowest predicted limit. While the census can never be totally complete due to practical observing limitations, Shara et al. (1999) have estimated the local population to be complete to $B < 14$. In the last decade, IR searches have doubled the number of confirmed WR stars (Hadfield et al. 2007; Mauerhan, Van Dyk & Morris 2009, 2011; Shara et al. 2009, 2012), thanks to the lower obscuration by dust at IR wavelengths compared to the visible. The current deficit in the number of confirmed WR stars is not yet settled as an observational constraint, and further IR searches should prove fruitful so that Galactic WR population completeness and evolutionary lifetimes for very high mass stars may be better understood.

In this paper, we describe our investigations into supervised machine-learning methods (Ball & Brunner 2010) to help locate the WR stars in the Galaxy that have yet to be detected and classified. The goal is to quantitatively improve on the reliability of the candidate selection. These methods use the statistical information contained in the IR colours for a set of known objects, including non-WR stellar populations that are frequently confused as WR candidates due to similar colours, to providing an automated classification of the unknown objects. The use of supervised machine-learning methods in astronomy has rapidly increased over the last decade, e.g. for automated classification of celestial objects in large catalogues and all-sky surveys (Malek et al. 2013; Kurcz et al. 2016; Lochner et al. 2016), photometric redshift estimation of galaxies (Tagliaferri et al. 2003; Lima et al. 2008; Sheldon et al. 2012; Heinis et al. 2016), morphological galaxy classification (Banerji et al. 2010; Shamir, Holincheck & Wallin 2013; Kuminiski et al. 2014; Pasquato & Chung 2016) and candidate type of object selection (Bailey et al. 2007; Yèche et al. 2010; Hsieh & Lai 2013; Marton et al. 2016). To our knowledge, this is the first time that machine-learning methods are used to classify objects in this colour space defined by J , H , K_s , $[3.6]$, $[4.5]$, $[5.8]$ and $[8.0]$ photometric bands. The tested algorithms are not limited to the search for WR stars, as they simultaneously disentangle the other stellar populations within the surveyed fields, which we regard as ‘contaminants’ and are typically younger and lower mass objects with thermal emission from dusty circumstellar shells or discs, degenerate in some colour spaces with the free-free emission from WR winds.

We show how the methods perform in re-classifying known objects, yielding a list with confidence metrics included. We also present some new WR discoveries with follow-up spectroscopic observations of candidates from this list. More observations will be useful to assess the empirical success rate with this method, hence allowing us to estimate the completeness of the known populations from the candidate lists obtained over the different regions of the sky. Finally, we discuss possible future improvements with additional information, higher statistics and more sophisticated methods.

2 METHODS AND SIMULATIONS

2.1 Heuristic IR colour selection

WR stars are a significant source of free-free emission due to the scattering of electrons in the neighbourhood of H^+ and He^+ ions, and as shown by Morris et al. (1993), the intrinsic (unreddened) shapes of the continuum energy distributions can be represented as a power law from optical to far-IR and radio wavelengths for the set of then-known Galactic and Magellanic Cloud WR stars. The flat continuum energy distributions yield broad-band IR colours that may be distinguished, at least partially, from other populations, as

shown by Hadfield et al. (2007) and Mauerhan et al. (2009), who report the discoveries of WR stars from the spectroscopic follow-up of candidates obtained with broad-band IR colour selection.

The colours used by Hadfield et al. (2007) were derived from the 2 Micron All Sky Survey or 2MASS (Skrutskie et al. 2006) and the *Spitzer* Galactic Legacy Infrared Midplane Survey Extraordinaire or GLIMPSE (Benjamin et al. 2003; Churchwell et al. 2009). After visual inspection of the colour–colour diagrams of their sample of GLIMPSE + 2MASS objects, including the known Galactic WR stars, they identified two regions, in the $[3.6] - [4.5]$ versus $[3.6] - [8.0]$ and $J - K_s$ versus $K_s - [8.0]$ diagrams, with the highest concentration of WR stars, essentially by eye. WR candidates were selected among the unclassified objects that fall in both regions. Additional selection criteria, such as photometric measurement quality and low confusion, were adopted to reduce uncertainties and shorten the candidate list from $\sim 100\,000$ to ~ 5000 objects. The spectroscopic follow-up of 261 WR candidates revealed 25 new WR stars, implying a WR recovery rate of ~ 10 per cent, as reported in Mauerhan et al. (2009). An overall success rate that includes other emission line OB-type stars that are spectroscopically similar and are probably related in evolutionary ‘transition’ has been around 25 per cent. Mauerhan et al. (2011) further refined the colour selection criterion by adding boundaries from the $J - H$ versus $H - K_s$ and K_s versus $J - K_s$ colour–magnitude diagrams. They reported 60 new WR stars, discovered at a higher WR success rate of ~ 20 per cent in what remains to this point to be a primarily ad hoc approach.

2.2 Supervised machine-learning classification

Selecting the WR candidates in a list of objects based on their multiband photometry is a conceptually straightforward classification problem: identify to which of the set of ‘classes’ an object belongs, when only limited information, the so-called features (e.g. the photometric colours), is available. Supervised machine-learning algorithms can be used to infer a classification function from a labelled training set. The training set contains examples of known objects, for which both the features and the class are reported.

The goodness of a classification algorithm, whether it is heuristic or machine-learning based, can be measured from the results obtained over a test set of objects for which both the features and the classes are known. Note that the objects in the test set can also be in the training set. The algorithm performance can be visualized by a specific table or ‘confusion matrix’ in which the columns represent the instances in a predicted class while each row represents the instances in an actual class (or vice versa). In other words, the element a_{ij} in the confusion matrix is the number of elements of the true class i , for which the predicted class is j . The confusion matrix for a perfect classification algorithm would have all non-diagonal elements equal to 0. In particular, we will refer to the following standard estimators (Perry, Kent & Berry 1955; Olson & Delen 2008):

- (i) the WR *hit rate* or *precision*, i.e. the percentage of true (known) WR stars in the inferred candidate list;
- (ii) the WR *completeness*² or *recall*, i.e. the percentage of WR stars correctly identified among those ones in the training set;

² This statistical estimator does not directly yield population completeness, but rather a theoretical upper limit, if spectroscopic follow-up of *all* candidates, with some constraints on the photometric quality, is performed (and the training set is representative of the true stellar populations).

Table 1. Classes and sizes in the adopted training set: the first column reports the total number of objects per class in the training set; the second column the numbers of objects with available photometric measurements in all bands (J , H , K_s , [3.6], [4.5], [5.8], [8.0]); the third column reports the numbers of objects with available photometric measurements in the K_s , and the fourth column IRAC bands.

	Max. size	7 bands	K_s + IRAC
WR stars	279	225	260
YSOs	1087	985	985
Be stars	79	79	79
AGB cand.	1722	713	1379
M-S stars	125	125	125
Total	3 292	2127	2828

(iii) the *total precision*, i.e. the percentage of objects correctly classified.

Because WR is the first class in the confusion matrices represented in this paper, the sum over the first row is the number of known WR stars (in the test set), while the sum over the first column is the number of WR candidates. The WR hit rate and completeness are obtained by dividing the first element of the matrix (top left) by the sum over the first column and over the first row, respectively. The total precision is the sum of the diagonal terms divided by the sum of all the elements in the confusion matrix, which is the number of test objects.

2.3 The training set

Our training set includes published catalogues of known WR stars, young stellar objects (YSOs, Evans et al. 2003), emission line B-type or ‘Be’ stars (Zhang, Chen & Yang 2005), asymptotic giant branch (AGB; Robitaille et al. 2008) candidates, and a selection of M-S stars obtained by cross-correlating the GLIMPSE objects from a small sky area with the CDS SIMBAD. The sizes of each class are reported in Table 1. Fig. 1 shows the [3.6] – [4.5] versus [3.6] – [8.0] diagram for the training objects and unknown objects from the GLIMPSE catalogue. Note that the training classes tend to form clusters in different regions of the [3.6] – [4.5] versus [3.6] – [8.0] colour space, but they are partially overlapping. Also, there are large areas of the colour space which do not contain any of the training objects, but they do contain many unknown ones. In future studies, it will be desirable to extend the training set, to make it more representative of the population diversity which is present in the GLIMPSE catalogue, then allowing more realistic estimates of the performances of the candidate selection criteria, for WR stars as well as for other classes of objects.

2.4 K -nearest neighbours (k -NN)

k -nearest neighbours (k -NN) is one of the simplest machine-learning algorithms (Altman 1992) because of its non-parametric, instance-based nature. Non-parametric means it makes no assumptions about the functional form to use for classification (e.g. polynomials in the colour space). Instance-based means that it does not learn a model, but computes the classification directly from the training objects. For these reasons, it is ideal for exploratory studies of complex classification problems where interrelations between variables and classes are complicated. It has been successfully used in a variety of contexts in different fields, such as the classification of cosmic rays (Borione et al. 1995), medical imaging

(Ramteke & Khachane Monali 2012; Thamilselan & Sathiaselan 2016), measuring photometric redshift of galaxies (Lima et al. 2008; Cunha et al. 2009; Sheldon et al. 2012), creating mock catalogues (Xu et al. 2013) and morphological galaxy classification (Shamir et al. 2013; Kuminski et al. 2014; Pasquato & Chung 2016). The k -NN algorithm classifies an unknown object based on the classes of the k -NNs in the training set, where k is a user-defined constant. A metric needs to be defined in order to measure the distances between objects and to find the k -NN to the query. Then, each neighbour will give a ‘vote’ for its own class, and the class with the highest sum of votes is attributed to the query. In this paper, we adopt the weighted k -NN approach, in which the votes are inversely proportional to the distance.

2.5 (Multi)dual-colour k -NN

As a first approach, we performed a k -NN classification using all possible colour–colour (or dual-colour) combinations. We adopted a standard Euclidean metric, i.e. the distance between two objects $o_i = (c_{1,i}, c_{2,i})$ and $o_j = (c_{1,j}, c_{2,j})$, being two colours c_1 and c_2 , is

$$d_{c_1, c_2}(o_i, o_j) = \sqrt{(c_{1,i} - c_{1,j})^2 + (c_{2,i} - c_{2,j})^2}. \quad (1)$$

Across all seven photometric bands used in our study, there are $(7 \times 6)/2 = 21$ distinct colours, defined as differences in magnitude between two bands. Symmetric colours, e.g. $K_s - [8.0]$ and $[8.0] - K_s$, are counted as one, because the change of sign has no effect on the distances. The possible colour–colour combinations are $(21 \times 20)/2 = 210$. We also made some tests without using the J and H bands, as discussed in Sections 3.1 and 3.2. In this case, there are $(5 \times 4)/2 = 10$ distinct colours, leading to $(10 \times 9)/2 = 45$ colour–colour combinations.

2.6 Weighted-colour k -NN

We chose to combine the information from multiple colours into a unique metric, rather than considering multiple dual-colour classification schemes. Our largest colour space is six-dimensional, as a maximum of six independent colours can be obtained from seven photometric bands, e.g. $J - H$, $H - K_s$, $K_s - [3.6]$, $[3.6] - [4.5]$, $[4.5] - [5.8]$ and $[5.8] - [8.0]$ (by differentiating consecutive bands). All the other colours are sums of those six, e.g. $[3.6] - [8.0] = ([3.6] - [4.5]) + ([4.5] - [5.8]) + ([5.8] - [8.0])$; hence, they would not add information. We tested using a weighted-Minkowski metric:

$$d_{w-M}(o_i, o_j) = \sqrt{\sum_l w_l^2 (c_{l,i} - c_{l,j})^2}, \quad (2)$$

where c_l are the independent colours. We tested the following colour bases:

- (i) $J - H$, $H - K_s$, $K_s - [3.6]$, $[3.6] - [4.5]$, $[4.5] - [5.8]$ and $[5.8] - [8.0]$;
- (ii) $K_s - [3.6]$, $[3.6] - [4.5]$, $[4.5] - [5.8]$ and $[5.8] - [8.0]$.

The weights, w_l , can be optimized for a specific goal, e.g. detecting WR stars, by finding the weights that maximize the WR hit rate in the test set. We generated 10 000 random weight vectors for each colour basis, computed the classification with the corresponding metrics, and stored and compared the outcomes. The results are discussed in Section 3.2.

Figure 1. $[3.6] - [4.5]$ versus $[3.6] - [8.0]$ diagram with unknown objects from the GLIMPSE catalogue (grey) and the training objects, i.e. WR stars (green), young stellar objects (blue), asymptotic giant branch candidates (magenta), Be stars (red) and M-S stars (cyan). The ‘+’ symbols under the legend represent the median error bars for points in the training set. The reddening vector (Indebetouw et al. 2005) is represented by the arrow near the y-axis.

2.7 Dealing with an incomplete training set

Our current training set covers a relatively small region of the colour space built up from the totality of the GLIMPSE objects (see Fig. 1), and it was not possible to extend the training set by class as most of the objects are not identified in CDS SIMBAD. It is this unidentified group that constitutes the candidate sample, but a criterion can be set to decide whether an object can or cannot be confidently classified, based on its ‘similarity’ with a number of objects in the training set. A possibility is to define some arbitrary colour cuts, as proposed in the previous heuristic studies, and then applying the k -NN classifier only to those objects within the selected colour regions. A more clever approach is to put a constraint on the neighbours’ distances, e.g. setting a threshold on the sum of the inverse distances of k neighbours, hereafter referred to as ‘score’. A low score is associated with test objects at large distances from the training objects, and the classification is then considered less reliable. Different kinds of constraints and threshold values can be more or less effective, depending on the properties of the training set (e.g. Dubuisson & Mylène 1993; Guttormsson et al. 1999; Arlandis, Perez-Cortes & Cano 2002; Markou & Singh 2003).

We also tested a probability estimator based on the colour distribution of the known WR stars, as an alternative selection criterion to the k -NN scores. For a given object, we considered independent³ colour intervals centred on the relevant colour values and with a predetermined length, and then we calculated the fraction of known WR stars that fell within those intervals simultaneously. In a two-

colour space, this process would be equivalent to counting the fraction of WR stars within a rectangle centred on the query point. An advantage of the probability estimator is that it provides an ordered priority list of candidates for spectroscopic follow-up. Of course, it may bias towards selecting the new candidates near the bulk of the observed sample and rejecting those near the tails of the distribution, but this potential issue is common to any other selection criterion.

The probability estimator approximates the probability of a WR star to have the relevant colour values within the chosen intervals, based on the observed sample. In other words, it is an index of similarity between the query object and the WR cluster in the training set. It does not correspond with the probability of the object being a WR star because of the observed colour values, which also depends on the number of objects of other classes falling in the same colour intervals, and is largely uncertain due to inherent incompleteness of the training set (known classes and relevant sizes are driven by selection effects due to targeted search programmes).

3 RESULTS

3.1 (Multi)dual-colour k -NN classifications

We tested the performances of all the 210 dual-colour combinations in reclassifying the training objects, with different choices of the number of neighbours considered ($k = 1, 5, 10, 15, 20$ and 50). We found that the top three WR hit rates (precision) and completeness (recall) are realized by the same dual-colour combinations, independent of k , though not always with the same order. These combinations are $[3.6] - [4.5]$ versus $[3.6] - [8.0]$, $[3.6] - [4.5]$

³ That is, such that each colour cannot be obtained as a combination of the others.

Figure 2. Left-hand panel: WR hit rate for the top three dual-colour combinations, reported for $k = 1, 5, 10, 15, 20$ and 50 . Right-hand panel: the same for the WR completeness.

Figure 3. Total precision for the top two dual-colour combinations, reported for $k = 1, 5, 10, 15, 20$ and 50 .

versus $[4.5] - [8.0]$ and $[3.6] - [8.0]$ versus $[4.5] - [8.0]$. It is worth noting that they have the same information content, as the colours in one combination can be calculated from the colours in another, but the relevant Euclidean metrics (equation 1) are not mathematically equivalent. Fig. 2 reports the WR hit rate and completeness as a function of k , for the top three dual-colour combinations and their average values. Both statistics increase asymptotically with k . In particular, when varying k from 1 to 10, the WR hit rate increases from ~ 80 per cent to ~ 83 per cent, while the WR completeness increases from ~ 80 per cent to ~ 86 per cent. Higher k values up to 50 only improve the WR statistics by less than 1 per cent.

The total precision has a similar dependence on k , but it is optimized for different dual-colour combinations, typically $[4.5] - [8.0]$ versus $K_s - [5.8]$ and $[4.5] - [8.0]$ versus $H - [4.5]$. Fig. 3 reports the total precision as a function of k .

We focus on the results obtained with $k = 10$, as a compromise between the best classification performances and the computing time (complexity scales as $O[k]$). Tables 2 and 3 report the confusion matrices for the $[3.6] - [4.5]$ versus $[3.6] - [8.0]$ and $K_s - [5.8]$ versus $[4.5] - [8.0]$, respectively. In both cases, the WR stars are very well distinguished from the Be stars, M-S stars and AGB candidates, and the greatest source of confusion is with the YSO population. The Be stars are not well recovered, most likely because the relevant training set is the smallest one. YSOs and AGB candidates are quite degenerate, but the relevant precision and recall are $\gtrsim 70$ per cent for

Table 2. Confusion matrix for $[3.6] - [4.5]$ versus $[3.6] - [8.0]$ with $k = 10$. The numerosity of each class is denoted in parenthesis in the first column.

True \ Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (225)	192	27	3	1	2	85.3
YSOs (985)	37	599	19	9	321	60.8
Be (79)	2	30	18	29	0	22.8
M-S (125)	0	9	7	109	0	87.2
AGBc (713)	0	257	0	0	456	64.0
Precision (per cent)	83.1	65.0	38.3	73.6	58.5	64.6

Table 3. Confusion matrix for $K_s - [5.8]$ versus $[4.5] - [8.0]$ with $k = 10$. The numerosity of each class is denoted in parenthesis in the first column.

True \ Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (225)	182	40	2	0	1	80.9
YSOs (985)	49	708	6	7	215	71.9
Be (79)	6	30	22	21	0	27.8
M-S (125)	1	5	5	114	0	91.2
AGBc (713)	0	195	0	0	518	72.7
Precision (per cent)	76.5	72.4	62.9	80.3	70.6	70.0

the $K_s - [5.8]$ versus $[4.5] - [8.0]$. Note that the statistics inferred from the reclassification of the training set do not apply, in general, to the classification of all the GLIMPSE objects, as the training set is incomplete (see Section 2.7), but they are indicative of the level of degeneracy between these classes of objects in the colour space analysed.

We investigated whether the k -NN classification would improve, if the results from the multiple dual-colour classifications are taken as votes, and the final classes attributed to the test objects are the ones with the majority of votes. We refer to this approach as multi dual-colour k -NN. Table 4 reports the corresponding confusion matrix. The total precision is higher than that for any dual-colour classification, and, in particular, the statistics relative to the Be stars are greatly improved. Table 5 reports the analogous confusion matrix for the multi dual-colour k -NN obtained by using only the 45 dual-colour combinations with the K_s and four IRAC bands, showing, overall, similar performances. The information contained in the J and H bands appears to be important to distinguish the Be stars from the YSOs and M-S stars. This test was motivated by the low number of reliable WR candidates that we found in the GLIMPSE

Table 4. Confusion matrix for the multi dual-colour k -NN classification with $k = 10$ (210 dual-colour combinations and seven GLIMPSE photometric bands). The numerosity of each class is denoted in parenthesis in the first column.

True \ Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (225)	114	83	2	0	26	50.7
YSOs (985)	12	737	1	9	226	74.8
Be (79)	0	7	49	23	0	62.0
M-S (125)	1	3	5	116	0	92.8
AGBc (713)	0	134	0	0	579	81.2
Precision (per cent)	89.8	76.5	86.0	78.4	69.7	75.0

Table 5. Confusion matrix for the multi dual-colour k -NN classification with $k = 10$ (45 dual-colour combinations, K_s and four IRAC bands). The numerosity of each class is denoted in parenthesis in the first column.

True \ Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (260)	212	37	1	1	9	81.5
YSOs (985)	33	632	5	7	308	64.2
Be (79)	6	34	16	23	0	20.2
M-S (125)	0	6	3	116	0	92.8
AGBc (1,379)	0	129	0	0	1 250	90.6
Precision (per cent)	84.5	75.4	64.0	78.9	79.8	78.7

catalogue with photometric measurements in all the seven bands, and the better statistics obtained with the dual-colour k -NN classifications using the bands at wavelengths longer than J and H .

The k -NN automatically provides some criteria to measure the level of confidence for the class attributed to a test object, e.g. in a single dual-colour classification, the relative weighted sums of votes obtained for that class relative to the total sum of votes from the k neighbours. For the multi dual-colour approach, it is the relative number of votes. Our tests on the training set confirm that the objects with the highest values of such confidence estimators are correctly classified, while most of the misclassified objects have similar number of votes (or similar weights) for at least two classes. For example, if accepting only those objects for which ≥ 75 per cent of the votes (i.e. ≥ 159 out of 210) are for the same class, the number of misclassified objects decreases from 25 per cent to $\lesssim 8$ per cent, mostly attributable to the degeneracy between YSOs and AGB candidates. No training objects are mistakenly classified as WR stars with more than 134 out of 210 votes (≈ 64 per cent), or more than 36 out of 45 votes (80 per cent, using K_s and four IRAC bands only). Spectroscopic follow-up of new candidates is needed to assess the validity of such k -NN confidence metrics.

3.2 Weighted-colour k -NN classification

Fig. 4 shows the WR completeness versus hit rate obtained for all random colour weights when reclassifying the training set (see Section 2.6). In most cases, both statistics are above 70 per cent, but an appropriate choice of colour weights can lead to ~ 90 per cent hit rate and ~ 80 per cent completeness, or vice versa, even when using the K_s and four IRAC bands only. Table 6 reports the sets of colour weights leading to the best classification results, in terms of WR hit

Figure 4. WR completeness versus hit rate for the weighted-colour k -NN with random weighting vectors.

rate, WR completeness⁴ and total precision. Tables 7–12 report the relevant confusion matrices.

In most cases, the statistics obtained using the K_s and four IRAC bands are comparable, within a few per cent, to the ones obtained using all seven bands. This test confirms that the information contained in the J and H bands is mostly important to distinguish the Be stars from the YSOs and M-S stars.

3.3 Uniform comparison with the previous heuristic approaches

The heuristic IR colour selection method adopted in the previous searches is targeted to select an optimal list of WR candidates, without attempting a classification into multiple classes. Hence, we can only compare the statistics relative to the WR stars obtained with the heuristic selection criteria and the k -NN classification. We found that by applying the criteria imposed by Mauerhan et al. (2011) on our training set, the resulting WR candidates list includes 61 per cent of the WR sample (completeness) with 43 per cent of true positives (hit rate). The empirical hit rate quoted in Mauerhan et al. (2011) is 20 per cent of their followed-up candidates from the GLIMPSE catalogue. Based on the discussion in Section 2.7, it is not surprising that the hit rate inferred from our training set is overestimated. However, it is significant that the majority of dual-colour k -NN classifications report much better statistics over the same sample of objects despite using less information, i.e. two colours against the six colours and one magnitude used to define the heuristic selection criteria. The multi dual-colour and weighted-colour approaches further improve the classification results, as discussed in Sections 3.1 and 3.2.

As an additional test, we consider the list of false WR candidates that have been found during the previous searches by Hadfield et al. (2007) and Mauerhan et al. (2009, 2011). The list consists of 216 non-WR objects for which a robust classification is not available. Therefore, we only check how many of those objects are mistakenly classified as WR stars with the different approaches. As the heuristic criteria have been progressively refined, we note that only 92 out of the 216 false candidates meet the latest selection criteria implemented by Mauerhan et al. (2011), and, in particular, solely due to the constraint added on $J - H$ versus $H - K_s$. We should

⁴ Among multiple sets of colour weights leading to the same WR completeness, we selected the one with the highest hit rate.

Table 6. Colour weights leading to the best classification results in terms of WR hit rate, WR completeness and total precision, with and without using the J and H bands.

Optimized statistics	$J - H$	$H - K_s$	$K_s - [3.6]$	$[3.6] - [4.5]$	$[4.5] - [5.8]$	$[5.8] - [8.0]$
WR hit rate	0.486	0.234	0.481	0.649	0.133	0.198
WR completeness	0.369	0.473	0.000	0.420	0.501	0.460
Total precision	0.238	0.387	0.360	0.607	0.325	0.436
WR hit rate	–	–	0.628	0.749	0.152	0.144
WR completeness	–	–	0.258	0.578	0.628	0.454
Total precision	–	–	0.156	0.749	0.492	0.417

Table 7. Confusion matrix for the weighted-colour k -NN with $k = 10$, colours using all the seven bands and weights optimizing the WR hit rate. The numerosity of each class is denoted in parenthesis in the first column.

True\Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (225)	174	27	2	1	21	77.3
YSOs (985)	14	742	4	11	214	75.3
Be (79)	1	0	59	19	0	74.7
M-S (125)	0	2	7	115	1	92.0
AGBc (713)	0	112	1	0	600	84.2
Precision (per cent)	92.1	84.0	80.8	78.8	71.8	79.5

Table 8. Confusion matrix for the weighted-colour k -NN with $k = 10$, colours using all the seven bands and weights optimizing the WR completeness. The numerosity of each class is denoted in parenthesis in the first column.

True\Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (225)	211	12	1	0	1	93.8
YSOs (985)	36	747	7	8	187	75.8
Be (79)	1	0	50	28	0	63.3
M-S (125)	1	2	10	112	0	89.6
AGBc (713)	0	116	0	0	597	83.7
Precision (per cent)	84.7	85.2	73.5	75.7	76.1	80.7

Table 9. Confusion matrix for the weighted-colour k -NN with $k = 10$, colours using all the seven bands and weights optimizing the total precision. The numerosity of each class is denoted in parenthesis in the first column.

True\Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (225)	206	13	2	1	3	91.6
YSOs (985)	25	775	5	9	171	78.7
Be (79)	1	0	58	20	0	73.4
M-S (125)	0	2	8	114	1	91.2
AGBc (713)	0	101	0	0	612	85.8
Precision (per cent)	88.8	87.0	79.5	79.2	77.8	83.0

also note that while robust stellar classifications were not carried out for the non-WR objects in the previous searches, a number of emission line objects with luminous OB star characteristics in the near-IR were identified, probably linked to WR stars in their evolution. For the purposes of testing and comparing machine-learning algorithms, the performance estimators adhere strictly to the WR stars.

The multi dual-colour k -NN approach would mistakenly find again 56 false candidates from the list of 216 if the information from all bands is used, which is 40 per cent less false candidates than that obtained with the state-of-the-art heuristic criteria. If considering

Table 10. Confusion matrix for the weighted-colour k -NN with $k = 10$, colours using the K_s and four IRAC bands, and weights optimizing the WR hit rate. The numerosity of each class is denoted in parenthesis in the first column.

True\Pred	WR	YSOs	Be	M-S	AGBc	recall (per cent)
WR (260)	210	26	5	1	18	80.8
YSOs (985)	21	619	24	11	310	62.8
Be (79)	3	32	18	26	0	22.8
M-S (125)	0	5	5	115	0	92.0
AGBc (1 379)	0	164	0	0	1 215	88.1
precision (per cent)	89.7	73.2	34.6	75.2	78.7	77.0

Table 11. Confusion matrix for the weighted-colour k -NN with $k = 10$, colours using the K_s and four IRAC bands, and weights optimizing the WR completeness. The numerosity of each class is denoted in parenthesis in the first column.

True\Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (260)	241	13	3	1	2	92.7
YSOs (985)	39	656	21	7	262	66.6
Be (79)	5	14	36	24	0	45.6
M-S (125)	0	4	3	118	0	94.4
AGBc (1 379)	0	138	0	0	1 241	90.0
Precision (per cent)	84.6	79.5	57.1	78.7	82.5	81.0

Table 12. Confusion matrix for the weighted-colour k -NN with $k = 10$, colours using the K_s and four IRAC bands, and weights optimizing the total precision. The numerosity of each class is denoted in parenthesis in the first column.

True\Pred	WR	YSOs	Be	M-S	AGBc	Recall (per cent)
WR (260)	238	17	2	1	2	91.5
YSOs (985)	40	661	18	8	258	67.1
Be (79)	4	15	32	28	0	40.5
M-S (125)	1	3	2	119	0	95.2
AGBc (1 379)	0	114	0	0	1 265	91.7
Precision (per cent)	84.1	81.6	59.3	76.3	83.0	81.9

only the K_s and IRAC bands, in an effort to include fainter or more extinct objects which are not observed in the J and H bands, the number of false candidates goes up to $\gtrsim 120$, either using our default multi dual-colour or weighted-colour approaches, which is again ~ 40 per cent less than that obtained with the heuristic criteria (using the same photometric bands). Furthermore, the number of false candidates obtained with machine learning can be reduced, in all cases, by accepting only those candidates with a higher confidence. For example, even when limited to the information in the K_s and

IRAC bands, the number of false candidates from the same list can be reduced to less than 92 by adding a threshold in the number of votes (27 out of 45) for the multi dual-colour k -NN, or in the weighted sum of votes for the WR class (~ 0.6) for the weighted-colour approach.

Mauerhan et al. (2011) suggested more restrictive criteria that, a posteriori, might have increased their empirical hit rate up to ~ 50 per cent. We found that, with those criteria, the hit rate inferred from our training set increases up to 84 per cent, close to the best values estimated for the machine-learning approaches, but the trade-off is a much lower completeness down to 39 per cent.

4 INITIAL CONFIRMATION OF CANDIDATES

While a description of our investigations into the use of the k -NN classifier to select candidate WR stars is the focus of this paper, the efficacy of the approach can be proven only by observational follow-up with the resulting lists of candidates. Near-IR spectroscopy of candidate WR stars was obtained on two half-nights on 2016 June 17 and 18 (UT), using the SpeX medium-resolution spectrograph on the 3-meter Infrared Telescope Facility (IRTF) telescope (Rayner et al. 2003), located on the summit of Mauna Kea in Hawaii. Sky conditions were cloudy for a substantial fraction of the two nights. Spectra were acquired during intermittent breaks in cloud cover; high thin cirrus clouds was always a factor that limited us to observing candidates with $K_s < 11$ mag, which is in the part of colour space where stellar population degeneracies increase. We were able to obtain useful spectra of 20 candidate WR stars, equally distributed over the two nights.

SpeX was used in the short cross-dispersed mode (SXD), with a slit width of 0.8 arcsec, providing a spectral resolving power of $R \approx 1000$. All spectra were acquired in an ABBA nodding sequence in order to subtract the sky background, and to suppress the contribution of bad pixels. The spectra were reduced and extracted using the IDL-based software package SPEXTOOL, specially designed for the reduction of data obtained with SpeX on the IRTF (Cushing, Vacca & Rayner 2004). Telluric corrections were derived using spectra of A0V standard stars, and executed using the IDL package XTTELLCOR (Vacca, Cushing & Rayner 2003), which applies and removes model H I absorption lines from the A0V standard star before application to the science data.

The WR candidates were selected from the GLIMPSE I catalogue (highly reliable) adopting the multi dual-colour k -NN classification scheme to the objects with reliable measurements in the K_s and IRAC bands. The Galactic latitude range of the available targets was $l = 10^\circ - 35^\circ$. We set a magnitude cut-off of $K_s \leq 13$ mag, i.e. a practical cut-off for observational follow-up at medium class telescopes with near-IR spectrographs, such as the 3-meter IRTF. We removed the known objects in the CDS SIMBAD from the final WR candidate list.

Although several emission line stars were detected, no WR candidates were confirmed on the first night. This is not terribly surprising or disappointing due to the difficult observing conditions and the effects these had on our selection of candidates to follow-up based on their reliability. On the second night, we implemented the probabilistic estimator defined in Section 2.7 and the number of votes as preferential criteria to select the ‘best candidates’ in our list. In this way, we discovered four new WR stars (their IR spectra are shown in Fig. 5). The hit rate for the second night is 40 per cent, while the average over the two nights is 20 per cent. More details are reported in Table 13.

Table 13. Observation summary.

RA	DEC	Spectrum	Rank	Votes
First night				
279.949 57	−5.665 45	Ble ^a	26	38
282.969 06	1.053 65	CO em	35	37
271.525 27	−19.328 18	CO abs	130	29
280.610 60	−6.069 31	CO abs	150	35
272.473 90	−17.809 85	CO abs	172	30
275.703 60	−15.651 53	CO abs	175	31
279.907 95	−3.699 48	CO abs	231	36
281.211 04	−2.960 66	Ble	232	23
272.610 51	−17.839 05	CO abs	465	37
280.327 38	−3.549 36	CO abs	730	29
Second night				
275.746 07	−13.208 70	WN4-5	2	40
281.086 02	−2.614 15	WN4-5	12	43
278.348 12	−10.933 98	Ble	24	44
277.085 35	−12.276 84	WN4-5	36	41
281.212 67	−4.080 24	Ble	77	41
283.342 38	0.160 59	CO abs	94	43
274.381 01	−15.447 57	CO abs	107	43
277.137 88	−9.802 16	CO abs	254	40
276.192 80	−11.907 58	WN4-5	310	41
279.267 45	−7.259 35	Ble	392	45

Note: ^a + emission line at 2.21 μ m.

CO refers to first overtone bandheads detected in the 2.28–2.36 μ m range. Since we do not know the distances, all Ble classifications are tentative.

The first empirical results indicates that the k -NN methods perform as well as or better than the heuristic approaches adopted in the previous literature, in terms of WR hit rate, although the current estimates are clearly limited to the low number statistics of our short observing run. We stress the fact that many of our best ranked candidates could not be observed during the second night, as part of them were covered by clouds or were too faint to be observed with the given sky conditions. Also, we note that three of the four WR discoveries comply with the heuristic criteria adopted by Mauerhan et al. (2011), and only one of them is in the WR sweet spot (see also Section 3.3).

5 FUTURE DEVELOPMENTS

The work presented in this paper is a proof of concept for the use of machine-learning classification algorithms for WR candidate selection, and, more in general classification of the objects in the GLIMPSE catalogue solely based on their IR photometry. Our preliminary results are promising and encourage us to pursue further developments as ways to improve the WR detection success rate.

First, observational campaigns involving the spectroscopic follow-up of GLIMPSE objects will increase the size of the classes and the number of classes in the training set. The current training set is at least one order of magnitude smaller than what is typical in other machine-learning classification studies (e.g. Kurcz et al. 2016; Marton et al. 2016). With these low numbers, the use of complex machine-learning classifiers, such as artificial neural networks (Jeffrey & Rosner 1986), is out of the question. In addition to the k -NN algorithms, we tested several variants of kernel support vector machine (Cortes & Vapnik 1995), Naive Bayes (Zhang 2004) and Random Forests (Breiman 2001). Their classification performances strongly depend on the details of the algorithms, e.g. different choices of kernel and weighted optimizations, but all of them give similar results, at their best, when applied to the current

training set. Therefore, a detailed comparison of the different algorithms appears to be premature. Our choice to focus on the results obtained with the k -NN algorithms is mainly motivated by their relative simplicity.

After the performances of the classification algorithms will be assessed, based on tests on a statistically significant sample (either fully empirical or synthetic), the total number of Galactic WR stars can be estimated from the number of candidates, given the WR hit rate and completeness for the algorithm adopted.

WR candidates can be selected from the GLIMPSE catalogue, based on their IR colours. In this paper, we tested, for the first time, the use of machine-learning classifiers to quantitatively improve the primarily eyeball approach of defining ‘sweet spots’ in colour space for WR candidates (Hadfield et al. 2007; Mauerhan et al. 2009, 2011). The preliminary tests discussed in this paper are very promising, as the machine-learning algorithms perform better than the previous heuristic approaches, if tested over the same sample. Also,

ACKNOWLEDGEMENTS

REFERENCES

- Altman N. S., 1992, *Am. Stat.*, 46, 175
- Arlandis J., Perez-Cortes J. C., Cano J., 2002, in Kasturi R., Laurendeau D., Suen C., eds, *Proc. 16th Int. Conf. Pattern Recognit.* Quebec City, QC, Canada. p. 576
- Bailey S., Aragon C., Romano R., Thomas R. C., Weaver B. A., Wong D., 2007, *ApJ*, 665, 1246
- Ball N. M., Brunner R. J., 2010, *Int. J. Mod. Phys. D*, 19, 1049
- Banerji M. et al., 2010, *MNRAS*, 406, 342
- Benjamin R. A. et al., 2003, *PASP*, 115, 953
- Borione A. et al., 1995, *Int. Cosm. Ray Conf.*, 2, 676
- Breiman L., 2001, *Mach. Learn.*, 45, 5

- Cardelli J. A., Clayton G. C., Mathis J. S., 1989, *ApJ*, 345, 245
- Churchwell E. et al., 2009, *PASP*, 121, 213
- Cortes C., Vapnik V., 1995, *Mach. Learn.*, 20, 273
- Crowther P. A., 2007, *ARA&A*, 45, 177
- Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, *MNRAS*, 396, 2379
- Cushing M. C., Vacca W. D., Rayner J. T., 2004, *PASP*, 116, 362
- Dubuisson B., Mylène M., 1993, *Pattern Recognit.*, 26, 155
- Ekström S. et al., 2012, *A&A*, 537, A146
- Eldridge J. J., Izzard R. G., Tout C. A., 2008, *MNRAS*, 384, 1109
- Evans N. J., II et al., 2003, *PASP*, 115, 965
- Georgy C., Ekström S., Meynet G., Massey P., Levesque E. M., Hirschi R., Eggenberger P., Maeder A., 2012, *A&A*, 542, A29
- Guttorfsson S. E., Marks R. J., II, El-Sharkawy M. A., Kerszenbaum I., 1999, *IEEE Trans. Energy Convers.*, 14, 16
- Hadfield L. J., Van Dyk S. D., Morris P. W., Smith J. D., Marston A. P., Peterson D. E., 2007, *MNRAS*, 376, 248
- Heinis S. et al., 2016, *ApJ*, 821, 86
- Hillier D. J., 1987, *ApJS*, 63, 947
- Hillier D. J., Lanz T., 2001, in Ferland G., Savin D. W., eds, *ASP Conf. Ser. Vol. 247, Spectroscopic Challenges of Photoionized Plasmas*. *Astron. Soc. Pac.*, San Francisco, p. 343
- Hsieh T.-H., Lai S.-P., 2013, *ApJS*, 205, 5
- Indebetouw R. et al., 2005, *ApJ*, 619, 931
- Jeffrey W., Rosner R., 1986, *ApJ*, 310, 473
- Kuminski E., George J., Wallin J., Shamir L., 2014, *PASP*, 126, 959
- Kurcz A., Bilicki M., Solarz A., Krupa M., Pollo A., Malek K., 2016, *A&A*, 592, A25
- Langer N., 2012, *ARA&A*, 50, 107
- Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, *MNRAS*, 390, 118
- Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, *ApJS*, 225, 31
- Malek K. et al., 2013, *A&A*, 557, A16
- Markou M., Singh S., 2003, *Signal Process.*, 83, 2481
- Marton G., Tóth L. V., Paladini R., Kun M., Zahorecz S., McGehee P., Kiss C., 2016, *MNRAS*, 458, 3479
- Mauerhan J. C., Van Dyk S. D., Morris P. W., 2009, *PASP*, 121, 591
- Mauerhan J. C., Van Dyk S. D., Morris P. W., 2011, *AJ*, 142, 40
- Morris P. W., Brownsberger K. R., Conti P. S., Massey P., Vacca W. D., 1993, *ApJ*, 412, 324
- Olson D. L., Delen D., 2008, *Advanced Data Mining Techniques*. Springer, Berlin
- Pasquato M., Chung C., 2016, *A&A*, 589, A95
- Perry J. W., Kent A., Berry M. M., 1955, *Am. Documentation*, 6, 242
- Ramteke R. J., Khachane Monali Y., 2012, *Int. J. Adv. Comput. Res.*, 2, 190
- Rayner J. T., Toomey D. W., Onaka P. M., Denault A. J., Stahlberger W. E., Vacca W. D., Cushing M. C., Wang S., 2003, *PASP*, 115, 362
- Robitaille T. P. et al., 2008, *AJ*, 136, 2413
- Rosslowe C. K., Crowther P. A., 2015a, *MNRAS*, 447, 2322
- Rosslowe C. K., Crowther P. A., 2015b, *MNRAS*, 449, 2436
- Shamir L., Holincheck A., Wallin J., 2013, *Astron. Comput.*, 2, 67
- Shara M. M., Moffat A. F. J., Smith L. F., Niemela V. S., Potter M., Lamontagne R., 1999, *AJ*, 118, 390
- Shara M. M. et al., 2009, *AJ*, 138, 402
- Shara M. M., Faherty J. K., Zurek D., Moffat A. F. J., Gerke J., Doyon R., Artigau E., Drissen L., 2012, *AJ*, 143, 149
- Sheldon E. S., Cunha C. E., Mandelbaum R., Brinkmann J., Weaver B. A., 2012, *ApJS*, 201, 32
- Skrutskie M. F. et al., 2006, *AJ*, 131, 1163
- Smartt S. J., 2009, *ARA&A*, 47, 63
- Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, *Neural Networks for Photometric Redshifts Evaluation*. Springer, Berlin, p. 226
- Thamilselvan P., Sathiaselan J. G. R., 2016, *Int. J. Appl. Eng. Res.*, 11, 4223
- Vacca W. D., Cushing M. C., Rayner J. T., 2003, *PASP*, 115, 389
- van der Hucht K. A., 2001, *New Astron. Rev.*, 45, 135
- Xu X., Ho S., Trac H., Schneider J., Poczos B., Ntampaka M., 2013, *ApJ*, 772, 147
- Yèche C. et al., 2010, *A&A*, 523, A14
- Zhang H., 2004, in Valerie B. Zdravko M., eds, *Proc. Twenty-eight Int. Florida Artif. Intell. Res. Soc. Conf. (FLAIRS 2004)*, AAAI Press, Palo Alto
- Zhang P., Chen P. S., Yang H. T., 2005, *New Astron. Rev.*, 10, 325

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