Publications of the Astronomical Society of Australia (2025), **42**, e013, 21 pages
doi:10.1017/pasa.2024.133

CAMBRIDGE UNIVERSITY PRESS

Research Article

A catalogue of complex radio sources in the Rapid ASKAP Continuum Survey created using a self-organising map

Afrida Alam[1], Kevin Pimbblet[1,2] and Yjan Gordon[3]

[1]E.A. Milne Centre for Astrophysics, University of Hull, Kingston-upon-Hull, UK, [2]Centre of Excellence for Data Science, AI, and Modelling (DAIM), University of Hull, Kingston-upon-Hull, UK and [3]Department of Physics, University of Wisconsin-Madison, Madison, WI, USA

Abstract

Next generations of radio surveys are expected to identify tens of millions of new sources and identifying and classifying their morphologies will require novel and more efficient methods. Self-organising maps (SOMs), a type of unsupervised machine learning, can be used to address this problem. We map 251 259 multi-Gaussian sources from Rapid ASKAP Continuum Survey (RACS) onto a SOM with discrete neurons. Similarity metrics, such as Euclidean distances, can be used to identify the best-matching neuron or unit (BMU) for each input image. We establish a reliability threshold by visually inspecting a subset of input images and their corresponding BMU. We label the individual neurons based on observed morphologies, and these labels are included in our value-added catalogue of RACS sources. Sources for which the Euclidean distance to their BMU is $\lesssim$5 (accounting for approximately 79% of sources) have an estimated >90% reliability for their SOM-derived morphological labels. This reliability falls to less than 70% at Euclidean distances $\gtrsim$7. Beyond this threshold it is unlikely that the morphological label will accurately describe a given source. Our catalogue of complex radio sources from RACS with their SOM-derived morphological labels from this work will be made publicly available.

Keywords: Radio continuum: galaxies; methods: data analysis; catalogues

(Received 21 March 2024; revised 7 November 2024; accepted 9 December 2024)

1. Introduction

Astronomical radio emission is dominated by synchrotron emission resulting from charged particles moving at relativistic velocities through magnetic fields. Such synchrotron radiation in galaxies generally arise due to the supermassive black hole at its centre or remnants from supernovae. Extragalactic radio continuum surveys therefore generally detect two groups of galaxies: star-forming galaxies (Condon 1992) and active galactic nucleus (AGN; Kormendy & Ho 2013). While AGN emit emission across the entire electromagnetic spectrum (Padovani et al. 2017), a fraction of them (around 15-20%) are considered radio-loud and produce strong radio emission as a result of synchrotron emission from the AGN's relativistic jets (Sadler, Jenkins, & Kotanyi 1989, Kellermann et al. 1989, Urry & Padovani 1995). Early radio astronomers faced difficulties in detecting radio galaxies at optical wavelengths since they are very faint at such wavelengths and this meant that there was little overlap between optical and radio surveys (Savage & Wall 1976; Windhorst, Kron, & Koo 1984; Windhorst et al. 1985).

The advent of new technology led to a transformational period in radio astronomy between 1990 and 2004 where surveys such as Westerbork Northern Sky Survey (WENSS; Rengelink et al. 1997), NRAO VLA Sky Survey (NVSS; Condon et al. 1998), Faint Images of the Radio Sky at Twenty-Centimeters (FIRST; Becker, White, & Helfand 1995), and Sydney University Molonglo Sky Survey (SUMSS; Bock, Large, & Sadler 1999) enabled an increase of the number of known radio sources to around 2.5 million sources. This is a hundredfold increase from earlier surveys such as 3C (Edge et al. 1959) and Parkes Catalogue of Radio Sources (Ekers 1969) among others (see summary in Norris 2017). However, even with improved sensitivities at radio-wavelengths, radio catalogues from these surveys are still ∼1% in size relative to those at optical (Norris 2017).

Next-generation radio continuum surveys are expected to go further and observe tens of millions of new objects (Norris 2017). This includes 20% of galaxies which were detected by infrared surveys such as Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) and multi-spectral imaging and spectroscopic redshift surveys such as Sloan Digital Sky Survey (SDSS; York et al. 2000, Abolfathi et al. 2018). Telescopes such as the Square Kilometre Array (SKA) and its precursor and pathfinder instruments such as Australian Square Kilometre Array Pathfinder (ASKAP; Johnston et al. 2008, Hotan et al. 2021), Low Frequency Array (LOFAR; van Haarlem et al. 2013), and Murchison Widefield Array (MWA; Tingay et al. 2013, Wayth et al. 2018), in addition to the Karl G. Jansky Very Large Array (JVLA; Perley et al. 2011) are able to conduct deep continuum surveys that can detect millions of radio sources. Moreover, they would be able to carry out these surveys on much shorter timescales than earlier surveys such as the LOFAR Two-metre Sky Survey (LoTSS; Shimwell et al. 2017), the Evolutionary Map of the Universe Pilot Survey (EMU; Norris et al. 2011, Norris et al. 2021), and the Very Large Array Sky Survey (VLASS; Lacy et al. 2020).

Corresponding author: Afrida Alam; Email: a.alam-2019@hull.ac.uk

Cite this article: Alam A, Pimbblet K and Gordon Y. (2025) A catalogue of complex radio sources in the Rapid ASKAP Continuum Survey created using a self-organising map. *Publications of the Astronomical Society of Australia* **42**, e013, 1–21. https://doi.org/10.1017/pasa.2024.133

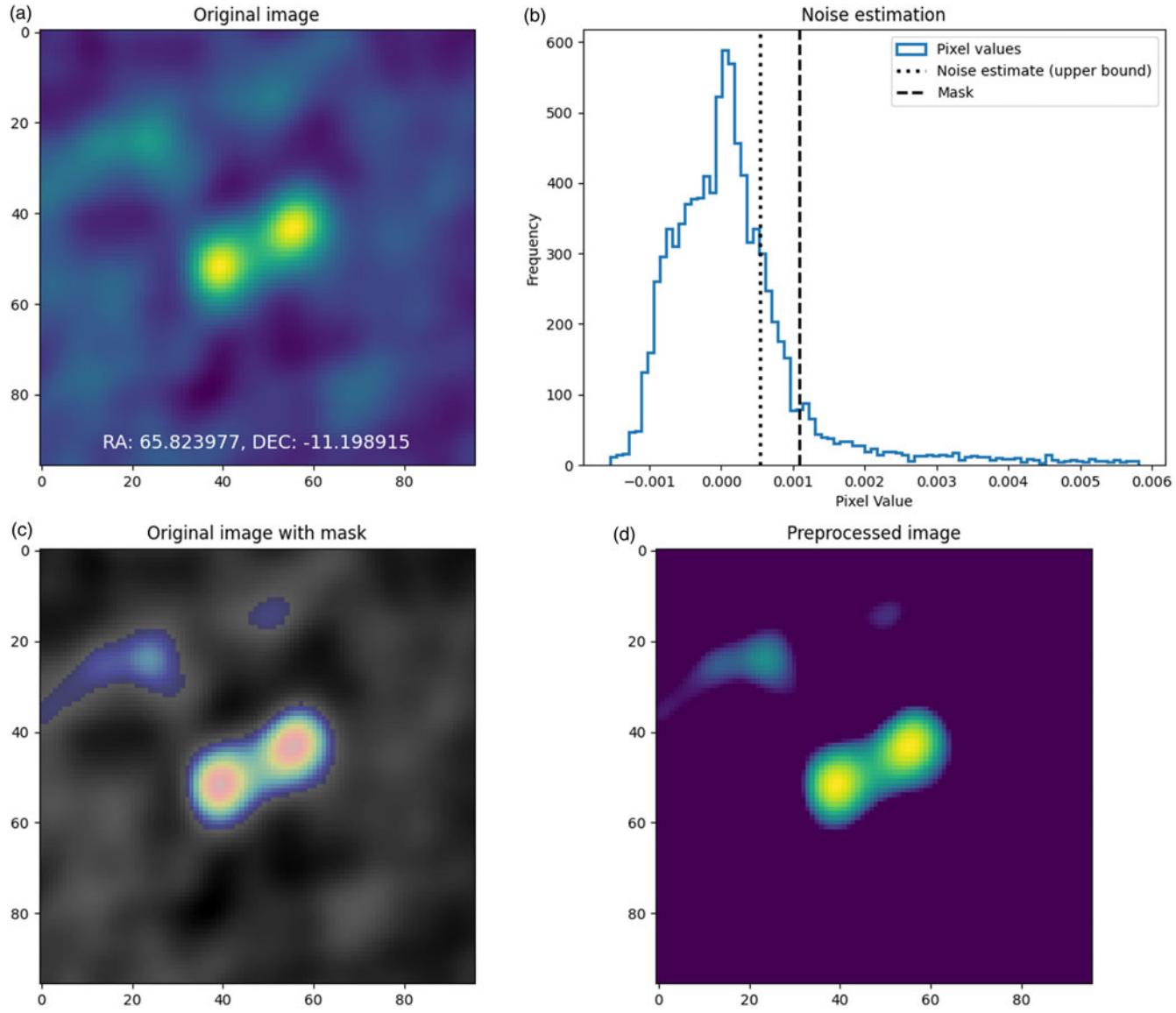

Figure 1. The preprocessing stages for each of the RACS image cutouts for a randomly chosen image cutout. On the first panel (a) we have the original image with its RA and DEC coordinates. On the second panel (b) we show the distribution of the pixel values along with the upper bound of the noise estimate and the mask applied (given by noise estimate multiplied by a minimum signal-to-noise value of 2). On the third panel (c) we show the original image with the mask overlaid on top. Once a mask is applied, we log scale the remained pixels and normalise them from 0 to 1 which gives us the final preprocessed image in the fourth panel (d).

Radio sources with a single component, where component refers to an output (generally a 2D Gaussian) from a source finding algorithm, are termed simple sources and they tend to make up around the majority of radio sources (Norris 2017). They can be easily resolved and cross-matched with catalogues in different wavelengths such as optical and infrared using techniques such as the Likelihood Ratio method (Sutherland & Saunders 1992). Complex sources, which are expected to make up a significantly smaller fraction of radio sources, are those which have multiple radio components and cannot be as easily identified as simple sources (Williams et al. 2019, Gürkan et al. 2022, Gordon et al. 2023). For example, two unresolved radio components that are in close proximity to each other might be radio emission from separate galaxies or they might be the lobes of a radio source (Gordon et al. 2023).

Given the large number of sources that are expected to be observed by next-generation surveys, classifying the morphology of the detected radio sources will be a challenging undertaking that will require novel methods of cross-identification. In this paper, we explore the use of self-organising maps (SOM; Kohonen 1990; Kohonen 2001), an unsupervised machine learning algorithm, in order to address the problem of finding complex radio sources and classifying their morphologies in the large dataset provided by SKA pathfinders, such as the Rapid ASKAP Continuum Survey (RACS; McConnell et al. 2020). Galvin et al. (2020) used SOMs to identify related radio components and the corresponding infrared host galaxy. The SOM was used on 894 415 images from FIRST and infrared data from WISE (Wright et al. 2010) centred at positions described by the FIRST catalogue. Using a SOM, they were able to identify potentially resolved radio components which

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

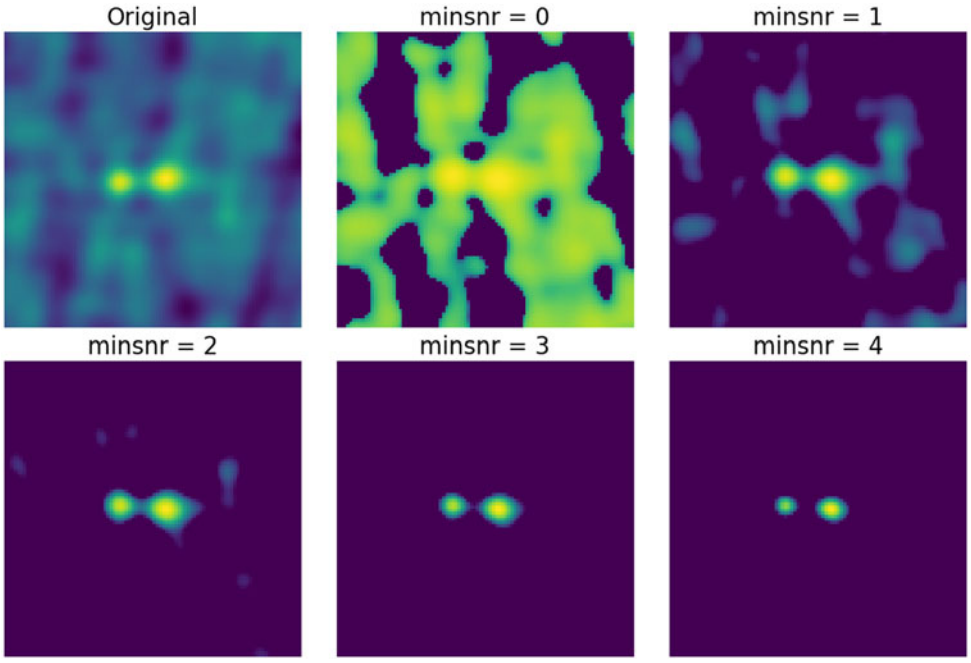

Figure 2. The original cutout and the preprocessed images of the cutouts done with different values of the minimum signal-to-noise ratio (0, 1, 2, 3, and 4) for the mask. The value was set to 2 since it is enough to mask the majority of the noise without losing much information as we can see in the images.

correspond to a single infrared host. Moreover, their approach was able to detect radio objects with interesting and rare morphologies such as 'X'-shaped galaxies and introduce a statistic that will enable the search of bent and disturbed radio morphologies. By using their method, they were able to identify 17 giant radio galaxies between 700–1 100 kpc.

The layout of the rest of this paper is as follows: in Section 2 we provide a brief description of the RACS data products used in this paper, and in Section 3 we outline the preprocessing steps for the cutout images from RACS as well as the SOM training process. In Section 4,we analyse the results from the SOM training and describe the inspection and mapping done to create a catalogue of complex sources, and we summarise our conclusions in section 5. This is followed by an Appendix A containing additional figures exploring the properties of individual neurons in the trained SOM grid.

2. Data

The survey used in this paper is from the first epoch of RACS, which is the first all-sky survey conducted with the full ASKAP telescope (McConnell et al. 2020, Hale et al. 2021). ASKAP is an array of 36 antennas, each with a 12-meter diameter. Each antenna is equipped with a phased array feed (PAF) and is capable of dual polarisation. At 800 MHz, each ASKAP pointing has a field of view of $\approx$31 deg^2 (Hotan et al. 2021). RACS is the deepest radio survey covering the entire southern sky to date. It is able to connect low-frequency surveys such as TIFR GMRT Sky Survey (TGSS; Intema et al. 2017) and Galactic and Extragalactic Allsky Murchison Widefield Array survey (GLEAM; Hurley-Walker et al. 2017) to surveys such as NVSS (Condon et al. 1998) at 1.4 GHz and VLASS at 3 GHz (Lacy et al. 2020).

We specifically use the data products from the first public data release from RACS, RACS-Low, which is made up of 903 tiles south of declinations of $+41°$ and covered a total survey area of 34 240 deg^2. It is centred at 887.5 MHz, with 15-min integrations and 288 MHz of bandwidth with 1 MHz wide channels, and they achieved a nominal sensitivity between 0.25 and 0.3 mJy/beam (McConnell et al. 2020; Hale et al. 2021). The first Stokes I catalogue from Hale et al. (2021) is derived from 799 tiles that have been convolved to a common resolution of 25 arcsec. It covered most of the sky in the declination region $\delta = -80°$ to $+30°$, excluding the region $|\,b\,| < 5°$ in the Galactic plane. The catalogue uses Python Blob Detection and Source Finder (PyBDSF; Mohan & Rafferty 2015) to detect regions of radio emission which are fitted with 2D Gaussian components. As such, the catalogue contains both single source components as well as sources with multiple components which are defined as single sources or islands of pixels fitted with multiple Gaussians. The catalogue contains 2 123 638 sources, of which 1 872 361 (88.17%) are simple sources with a single Gaussian component and 251 277 sources (11.83%) are complex with multiple components. Since the objective of our work is to identify complex sources, the focus will be on this latter group of sources with multiple components.

3. Methodology

3.1 Self-Organising Map (SOM)

Given that the next-generation surveys are expected to become more data intensive and detect vast number of radio sources, machine learning methods can help us identity and classify these detected sources in a more efficient and labour-saving manner. They can be divided into two approaches: supervised and unsupervised machine learning. Supervised machine learning broadly describe algorithms that learn to represent an unknown and possibly complex function by training a mapping function between input data and their assigned training labels. For example, Aniyan

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

& Thorat (2017) used convolution neural network (CNN), a type of supervised machine learning method, to classify radio sources from FIRST into Fanaroff-Riley (Fanaroff & Riley 1974) and bent-tailed morphology classes, and had a success rate of approximately 95% depending on the morphology presented, with bent-tailed radio galaxies being the most identifiable. Lukic et al. (2018) used CNNs trained on radio sources from Radio Galaxy Zoo (RGZ; Banfield et al. 2015) Data Release 1 to classify sources into compact and different classes of extended sources and achieved a 94.8% accuracy rate.

Conversely, unsupervised machine learning methods are algorithms that do not need labelled data or any prior knowledge about the dataset. Instead these algorithms focus on identifying any existing structures within a dataset. Examples of unsupervised machine learning methods include k-means clustering (Ikotun et al. 2022), Gaussian mixture models (Everitt et al. 2011), principal component analysis (Jolliffe & Cadima 2016), and SOMs (Kohonen 1990; Baron 2019). SOMs (also known as Kohonen maps) are a type of neural network which output a low-dimensional, usually two-dimensional, representation of the input dataset. SOMs use a competitive learning process in order to map the input dataset onto a grid of discrete neurons. The neurons are each assigned a unique position, i, onto a regular lattice and initialised with prototype weights, w, which are usually zeros or small random values. A single iteration of training involved randomly selecting an input data sample, d, from the reference training dataset, D, and comparing it to the current state of each of the neurons. The neuron with the best similarity score is referred to as the best-matching unit (BMU). The position of the BMU, j, is then used to update the weights of the other neurons in the grid.

Euclidean distance is one of the similarity metrics that the SOM algorithm can use to quantify the similarity between the input data and the neurons in the SOM grid. This can be done by calculating the straight-line distances between them. However, this can result in problems with some data types, including astronomical images, which do not maintain invariance between certain types of transformations such as flipping and rotation. As a result, Polsterer et al. (2016) developed the software Parallelised rotation and flipping INvariant Kohonen-maps (`PINK`)[1] which builds upon the SOM algorithm and introduces a minimisation procedure to best align a source of random orientation onto the neurons. This ensures that sources which are similar are grouped despite any differences in rotation. A SOM training using `PINK` starts with the weights of all the neurons being initialised with randomly generated numbers or zeros. `PINK` then rotates and flips all input images a set number of times (specified by the rotations parameter which will be discussed in more detail in Section 3.3). The similarity between the input images, including the rotated and flipped copies, and all the neurons are calculated so that the BMU, i.e. the neuron with the shortest Euclidean distance to a given input image and is therefore the best representation of the input, can be identified (see Section 2 of Polsterer et al. 2016 for more details on the method used in the `PINK` framework, and Polsterer, Gieseke, & Igel 2015 for details on an earlier version of the framework). The weighting function implemented in `PINK` is described by Equation (1) (Polsterer et al. 2016):

$$w'_i = w_i + \alpha(t) \cdot G_{ij} \cdot (\phi(d) - w_i) \tag{1}$$

[1]Parallelised rotation and flipping INvariant Kohonen maps (`PINK`): https://github.com/HITS-AIN/PINK.

Table 1. The hyperparameters used in the four training stages: the width of the neighbourhood function G_{ij} given by σ, the learning rate α, the number of rotations and iterations.

Stage	σ	α	Rotations	Iterations
1	1.5	0.1	92	5
2	1.0	0.05	180	5
3	0.7	0.05	360	5
4	0.5	0.005	360	10

where:

- w_i is the initial weight vector of neuron i.
- w'_i is the updated weight vector of neuron i.
- α is the learning rate and this parameter controls how much the weights are updated as training progresses.
- G_{ij} is the neighbourhood function which controls the extent to which the BMU neuron j influences the weight update of neuron i. `PINK` currently supports three possible neighbourhood functions of which the Gaussian distribution function is the most common and is used for this training. The width of the Gaussian neighbourhood function establishes the BMU's region of influence such that the weights of the neurons which are closer to the BMU are updated more than neurons which are further away.
- d is the current input data, and the term $(\phi(d) - w_i)$ aligns d onto w_i.

`PINK` uses a modified Euclidean distance metric (Polsterer et al. 2016; Galvin et al. 2019) to measure similarity:

$$\Delta(A, B) = \underset{\forall \phi \in \Phi}{\text{minimise}}(\phi) \sqrt{\sum_{c=0}^{C} \sum_{x=0}^{X} \sum_{y=0}^{Y} \left(A_{c,x,y} - \phi(B_{c,x,y})\right)^2} \tag{2}$$

where A and B are a given neuron and input image and c is their corresponding channel, x and y are the coordinates of the pixels, ϕ corresponds to an affine image transformation which has been drawn from a set of image transformations Φ, i.e. the set of all rotated and flipped copies of the input, and is optimised to best align the input image with the features of the neuron by finding the ϕ with the shortest Euclidean distance from all the possible rotations and flips. `PINK` can also impose either a circular or quadratic region over which the Euclidean distances are calculated. A quadratic region can cause variations in these calculations due to the impact of other sources, especially bright sources, potentially moving into the masked region as the images are rotated (Vantyghem et al. 2024). We use a circular mask in order to minimise the effects of sources near the edges of the images. Once the Euclidean distances have been calculated and a BMU has been determined, the neuron positions on the SOM grid are evaluated against the neighbourhood function and their weights are modified accordingly so that they can be a better representation of the input. The previous steps are then iterated over all input images in the training dataset an X amount of times, keeping in consideration that X must be large enough to allow the SOM to converge. After enough training iterations have taken place such that stable SOM can be produced, all input images in the dataset

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

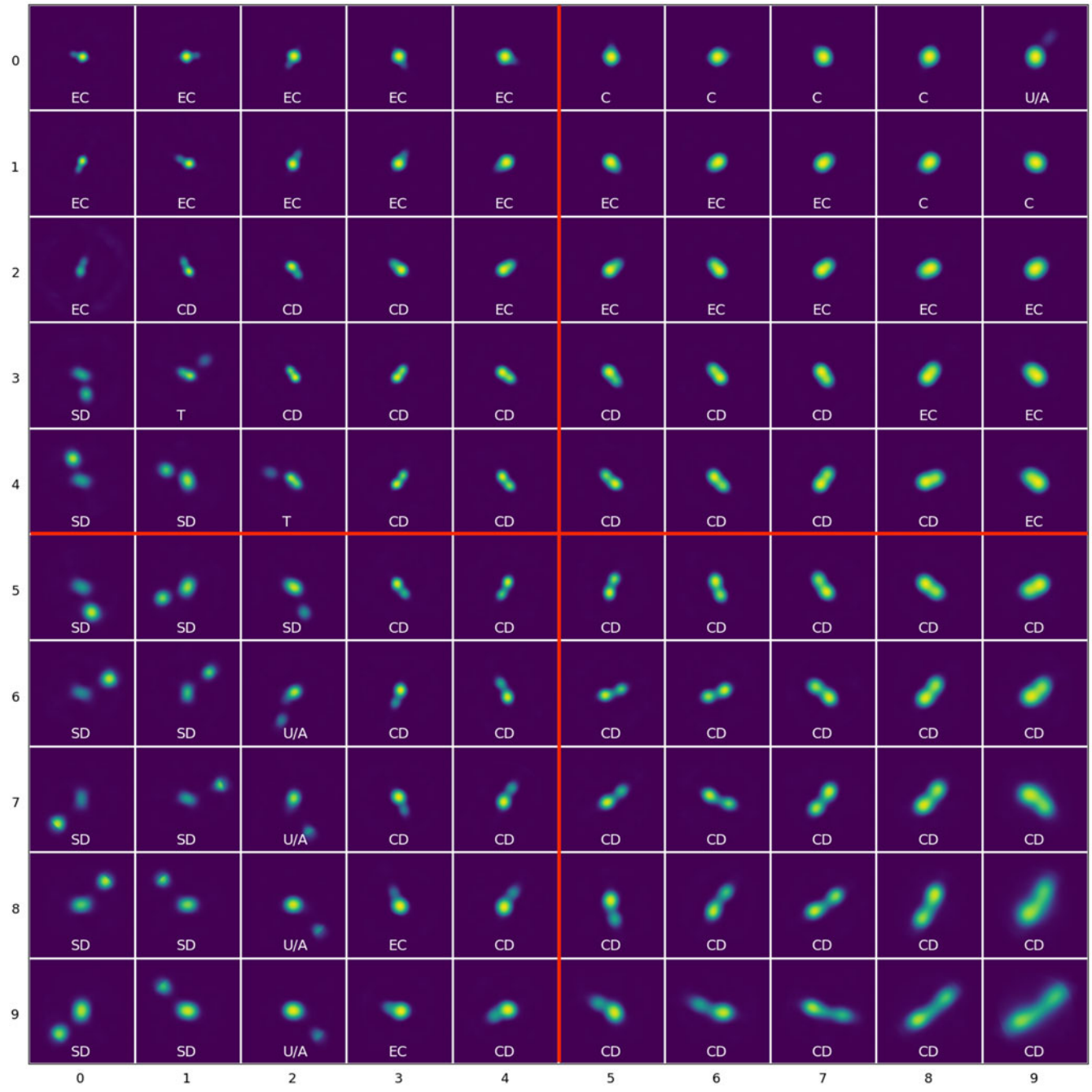

Figure 3. The trained 10x10 SOM with manual labels of their morphological labels: **C** (Compact) sources, **EC** (Extended Compact), **CD** (Connected Double) sources, **SD** (Split Double) sources, **T** (Triple) sources, **U/A** (Uncertain/Ambiguous) sources. The labels on the axis indicate the neuron coordinate in the SOM grid such that the top left neuron is (0, 0) with morphological label EC. The SOM can also divided into four quadrants: top left, top right, bottom left, and bottom right (marked in red) for additional analysis.

are mapped to the derived neurons in order to determine the distances to the neurons and find the best match regions.

The key hyperparameters to consider during the training process are: the number of neurons in the SOM grid, width of the neighbourhood function, learning rate, and the number of rotations and iterations. The number of neurons specifies the size of the SOM, and should be large enough to represent the dominant structures in the training data but not so large that it becomes time-consuming to compute. The width of the neighbourhood function used cannot be too wide as that would impact all neurons in the grid, however it also cannot be too narrow as that could potentially lead to individual neurons being decoupled and only creating smaller clusters that do not accurately capture the similarities between neighbouring neurons. The learning rate, if

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

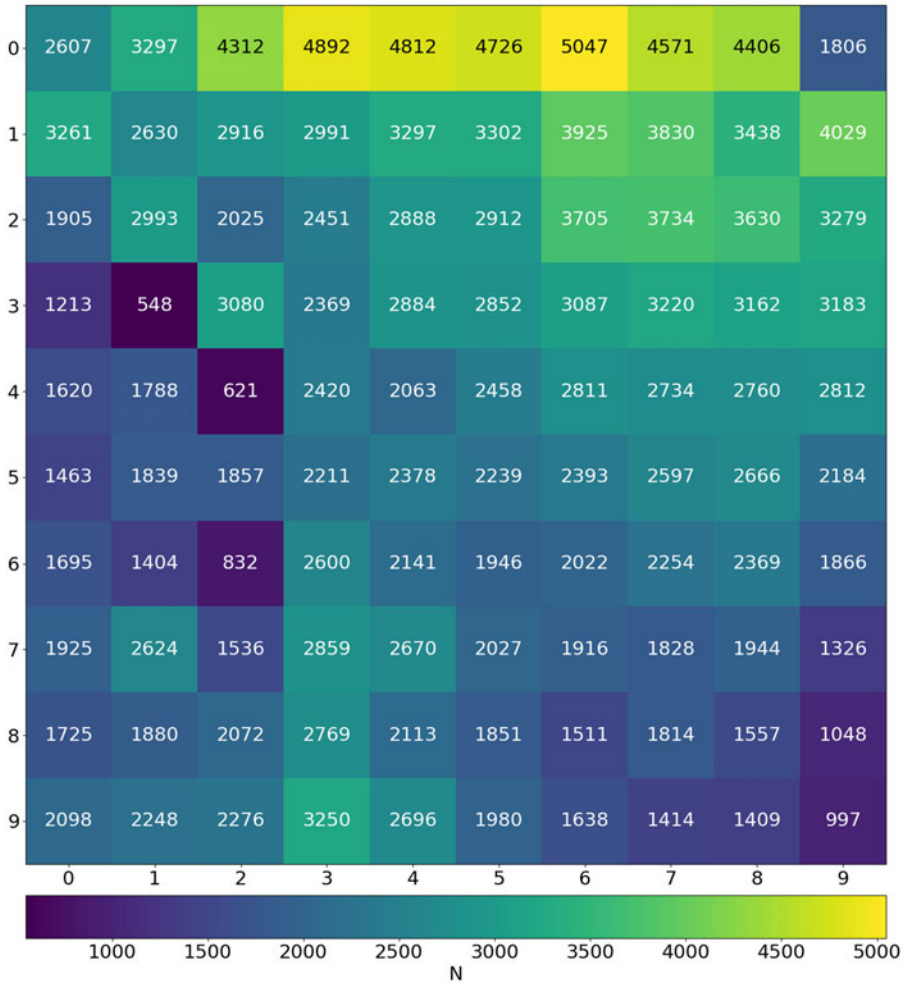

Figure 4. Density map showing the Best-Matching Unit (BMU) count across the trained SOM.

too small, will result in long computational time, but if it is too large then the changes in the weight updates would be too abrupt (Mostert et al. 2021).

3.2 Preprocessing RACS image cutouts

Prior to training a SOM, the data is preprocessed to ensure that the training set emphasises the morphological structures and features present and interference from noise is minimised. The first step is to filter the RACS-Low catalogue to select complex sources with multiple components which are classed as 'multi-Gaussian' (given by 'M' in column 'S Code' in the catalogue) by PyBDSF (Mohan & Rafferty 2015, Hale et al. 2021). This results in 251 277 individual radio sources. Next, we take 96×96 pixel cutouts from the RACS image tiles centred around each coordinate which corresponds to a $4'$ field-of-view. The vast majority of radio sources are expected to have sub-arcmin sizes, with only the largest sources having angular extents greater than $4'$ (Lara et al. 2001; Proctor 2016). A cutout of angular size $4'$ will be large enough to capture most radio sources, including possible extended structures, while still being small enough to avoid any interference from unrelated nearby sources in the tile. There may exist more than one coordinate for a given object due to decomposition, hence the same object can be in more than one cutout. As such, if there are multiple cutouts for a given 'target' RA and Dec coordinates, we choose the cutout that is closest to this target. This is done by calculating the angular distance between the pixel coordinates of the target and the reference CRPIX1 and CRPIX2 pixel coordinates of the cutouts. The most central cutout, i.e. the cutout with the smallest angular distance to the reference pixel coordinates, is selected.

For our preprocessing, we use the python package `PYINK`,[2] which has a set of useful tools to aid in training and analysing a SOM using `PINK` and can be used to create the bespoke binary file it requires. The image cutouts for each coordinate are preprocessed with `PYINK`, and Fig. 1 shows the preprocessing stages for an arbitrary cutout taken from our sample. The first stage is estimating the noise in each image cutout. We measure the outlying pixels that deviate from the median by more than three times the median absolute deviation (MAD) of the pixels. We clip and remove these flagged outlier pixels and this process is repeated twice. A Gaussian was fitted to the pixel intensity distribution of the remaining unflagged pixels. The standard deviation of this fitted distribution gives the estimate of the background noise, which

[2] `PYINK`: https://github.com/tjgalvin/pyink (commit `176177b`).

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

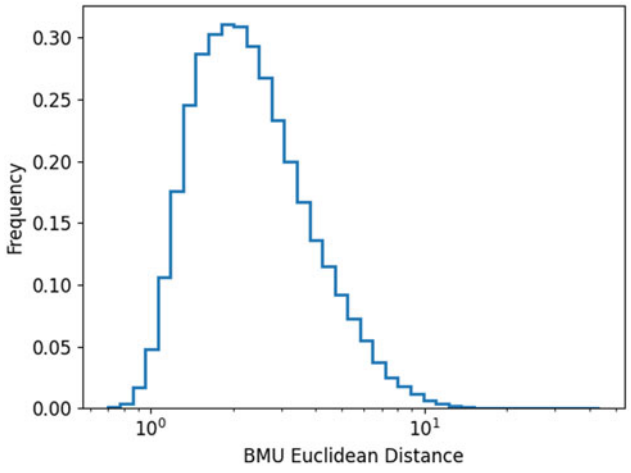

Figure 5. The distribution of the Euclidean distance between input images and their corresponding BMU neuron.

we can consider to be the upper bound of the noise. A mask is then created by multiplying the noise estimation and a minimum signal-to-noise ratio (the dotted and dashed lines in the second panel (b) in Fig. 1 corresponds to the noise estimation and mask, respectively). Tests were run with different values of the minimum signal-to-noise (Fig. 2), and based on these we determine that a value of 2 is sufficient in masking the majority of the noise in the cutout whilst still capturing the important structures present. Once a mask is applied, it filters the pixels to only retain those that satisfy the mask conditions and pixels outside this mask were filled in to have values of 0. We then apply a log scaling to the pixels and normalise them from 0 to 1 to give the final preprocessed image. Out of the 251 277 image cutouts that underwent the preprocessing method 18 cutouts failed. Visual inspections of these cutouts show that they tend to have diffuse or large scale structure which cover a larger fraction of the cutouts, and as such `PYINK` finds a higher noise level estimation than `PyBDSF`. This results in edge cases where the signal-to-noise within the cutout is not representative of the local signal-to-noise for large sources. Due to the higher noise level, the data array is empty once the mask (the product of the `PYINK` noise estimation and a minimum signal-to-noise value of 2) is applied during preprocessing. We use the remaining 251 259 sources to train the SOM.

3.3 Training

Following the preprocessing, we train a 10×10 SOM using `PINK` with four training stages. In order to do so, we have to establish certain hyperparameters for each stage: the width of the neighbourhood function (σ), the learning rate (α), and the number of rotations and iterations. As stated in Section 3.1, the width of the σ function sets how much of the neighbourhood should be updated in each iteration, the learning rate controls how much the weights are updated during each iteration, rotations gives the number of rotations and flips `PINK` performs for each input image during the training, and iterations is the number of times each individual item in the training dataset is used to update the SOM.

During the training process, we want to first establish the broad morphologies and subsequently fine-tune the SOM and identify smaller structures and details present (Galvin et al. 2019; Mostert et al. 2021). For the first training stage, the neurons are initialised with random noise. We keep the rotations and iterations to a lower number initially, and this also has the added advantage of decreasing the computational time. The number of rotations are subsequently increased at each stage, and in the final training stage we also increase the number of iterations in order to capture the finer details. Kohonen (2001) states that a larger neighbourhood function is able to capture the broad or global structures in the dataset, and so decreasing the size of the neighbourhood function as training progresses ensures that the smaller or localised details are also represented in the SOM. We follow the same principle for the learning rate which controls the magnitude of the weight updates during each iteration. Hence, the width of the neighbourhood function (σ) and the learning rate (α) are set to 1.5 and 0.1, respectively, and they are decreased in subsequent training stages.

The SOM grid is trained on the hyperparameters established in Table 1 and is shown in Fig. 3. Here the neuron positions are given by (y, x) and the origin in this coordinate scheme is the left-most column of the top row, such that the right-most column of the top row has a coordinate of (0, 9). The training is done on a single GPU node with a NVIDIA A40 48GB GPU and took approximately 14 days, however we note that CUDA was disabled and this may have affected the time taken for the training.

Table 2. Intervals based on Euclidean distance between randomly chosen input images and their BMU.

Intervals	Euclidean distance	Total sources	Sample sources
1	$0.69 \leq \mathrm{ED} < 2.74$	114 392	338
2	$2.74 \leq \mathrm{ED} < 4.78$	83 327	288
3	$4.78 \leq \mathrm{ED} < 6.83$	33 501	183
4	$6.83 \leq \mathrm{ED} < 8.87$	12 139	110
5	$8.87 \leq \mathrm{ED} < 10.91$	4 648	68
6	$10.91 \leq \mathrm{ED} < 12.96$	1 845	42
7	$12.96 \leq \mathrm{ED} < 15.00$	739	27
8	$15.00 \leq \mathrm{ED} \leq 43.82$	668	25

4. SOM inspection and mapping

Once the SOM is trained, we map all images onto the final SOM and plot a density map in order to see the neurons which were most frequently selected as the BMU (Fig. 4). The neuron (0, 6) is selected as the BMU for the highest number of input images, 5 047 images which accounts for approximately 2% of the total input images. Moreover, the neurons in the first row were more likely to be selected as BMU, showing that even for 'multi-Gaussian' sources relatively simple morphologies are dominant. Whereas the neurons (3, 1) and (4, 2) which we label as Triple sources (see Section 4.1 for a description of the labels) are selected as BMU the least number of times, for 548 and 621 input images, respectively. Generally, we see that the neurons towards the edges, especially in the bottom half of the grid, were chosen to be the BMU much less frequently than those near the top half. We also see that the impact of the circular region over which the Euclidean distances are calculated are more visible for some neurons than others and can be seen in the SOM grid, for example neuron (9, 9) at the bottom right. We also note that structures outside this region are essentially noise and potentially do not carry any real meaning.

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

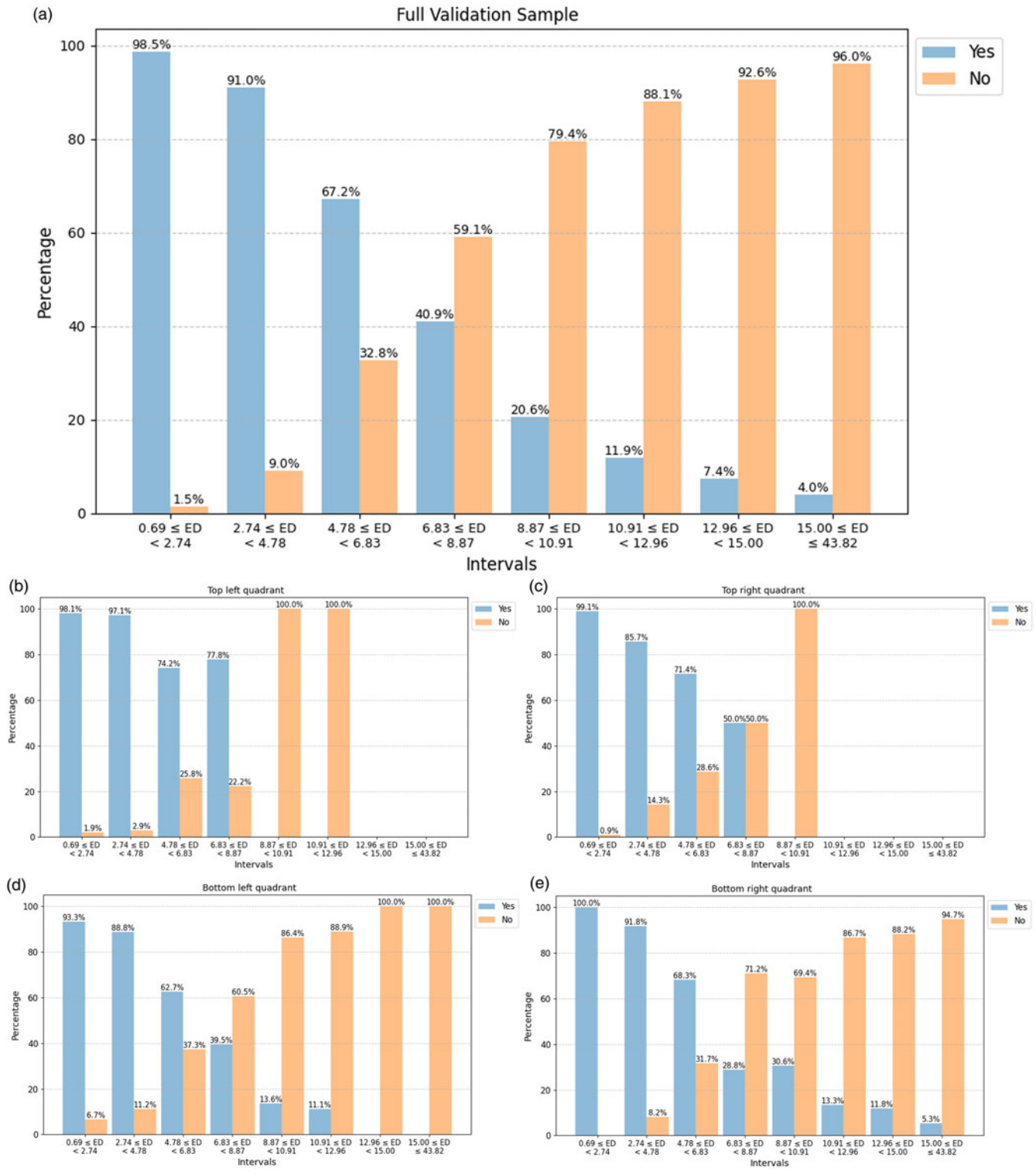

Figure 6. Upper panel (a): A manual validation of the match between original input images in the full validation sample and their corresponding BMU, where the sample is divided into smaller intervals on the Euclidean distance (Table 2). Lower panels (b–e): Distribution of the 'Yes' and 'No' matches from the validation scheme above split into the SOM quadrants: top left quadrant (b), top right quadrant (c), bottom left quadrant (d), and bottom right quadrant (e).

Another important property of a SOM to consider is coherence which can be a useful tool for morphological studies such as this. The coherence gives us the total number of times where the neuron which was the next best match for an input image (i.e. had the second lowest value of Euclidean distance between it and the input image) was neighbouring (both next to or diagonally) the neuron chosen as BMU for the same image. A high coherence value indicates that neighbouring neurons are similar to each other and the

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

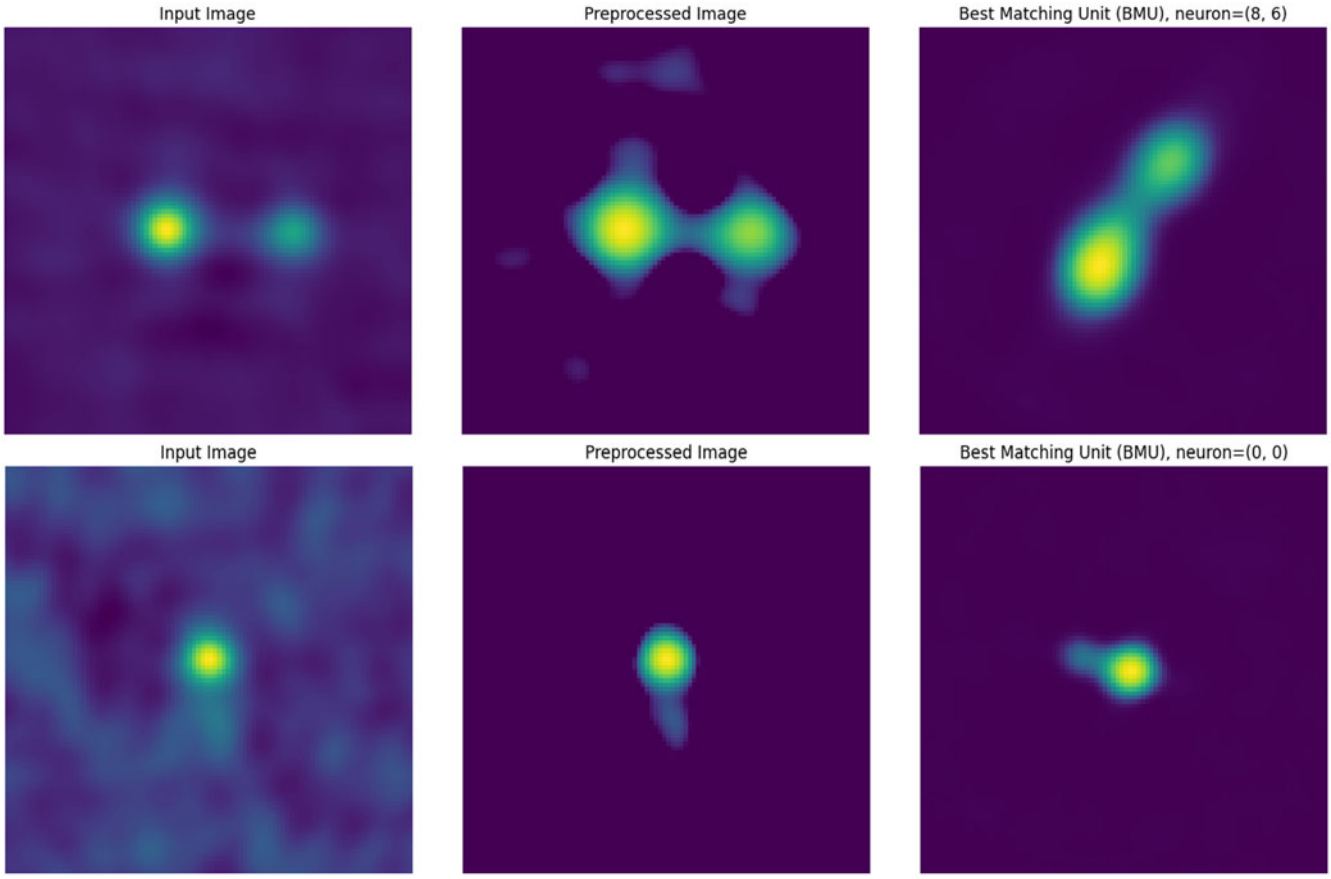

Figure 7. Examples of a 'Yes' match for an Input Image, and its corresponding Preprocessed Image and Best-Matching Unit (BMU).

underlying structures of the SOM are well-organised and represented in the SOM grid. The coherence value in our SOM is 229 678 which means that for 91.4% of the input images, their second closest best-matching neuron is adjacent to the best-matching neuron in the grid. This shows that the different neighbourhoods in our SOM grid are well-established and so the regions in and of themselves are also a useful metric of morphology and not just the precise coordinate. We see this in our trained SOM (Fig. 3) where similar representative images or morphologies are closer to each other and can be grouped into similar morphological classes.

We plot the distribution of the Euclidean distance between all the input images and their corresponding BMU (Fig. 5; distributions of the Euclidean distances for each individual neurons in the SOM can be found in the Appendix A Figs. A1, A2, A3, and A4). The distance in this paper ranged from 0.69 to 43.82, where larger Euclidean distance values indicate there are larger differences between an image and the neuron. It should be noted that the Euclidean distance values generally depend on the input images used to train the SOM as well as the specific training configurations of the system and so is therefore unique to each SOM. A visual inspection can be done to establish an approximate distance threshold at which point the input image starts to no longer resemble its associated BMU.

To perform a quantitative validation of the similarity between input image cutout and the BMU we create a smaller validation sample by dividing the range of Euclidean distances into 8 intervals and selecting random sources from each (Table 2). The first 7 intervals comprised of distances in the range 0.69–15.0 as 99.7% of our dataset have distances in that range, and the 8th interval consisted of those with distances 15.0 to 43.82. Next, we take the $\sqrt{n}$ of the number of sources in each interval to create the validation sample so a manual validation of the matches can be performed in a more efficient and less time-consuming manner. Prior to the validation, we inspect multiple random input images and their BMU to review the criteria of a match so as to avoid bias and ensure consistent labelling. A visual inspection is subsequently done on the validation sample by looking at each of the original input images in the sample and their BMU to see if they matched or mostly matched (rotating and flipping the input images to match the BMUs if needed), such that the input image can be reasonably believed to have contributed to the BMU neuron which is an aggregate of all input images for which it is chosen as the BMU. These matches were designated as 'Yes' with the others being assigned as 'No' (the upper panel a in Fig. 6). While this is a subjective match based upon our visual inspection, it provides a good baseline for reliability in the similarity matches (see Figs. 7 and 8 for examples of how the similarity was judged).

As expected, the number of 'Yes' matches between the input and the BMU decrease with Euclidean distances. In Intervals 1 and 2 (which cover the distance 0.69–4.78 overall and comprise of 78.7% of our dataset), the number of 'Yes' matches are 98.5% and 91.0%, respectively, and can be considered to be the most

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

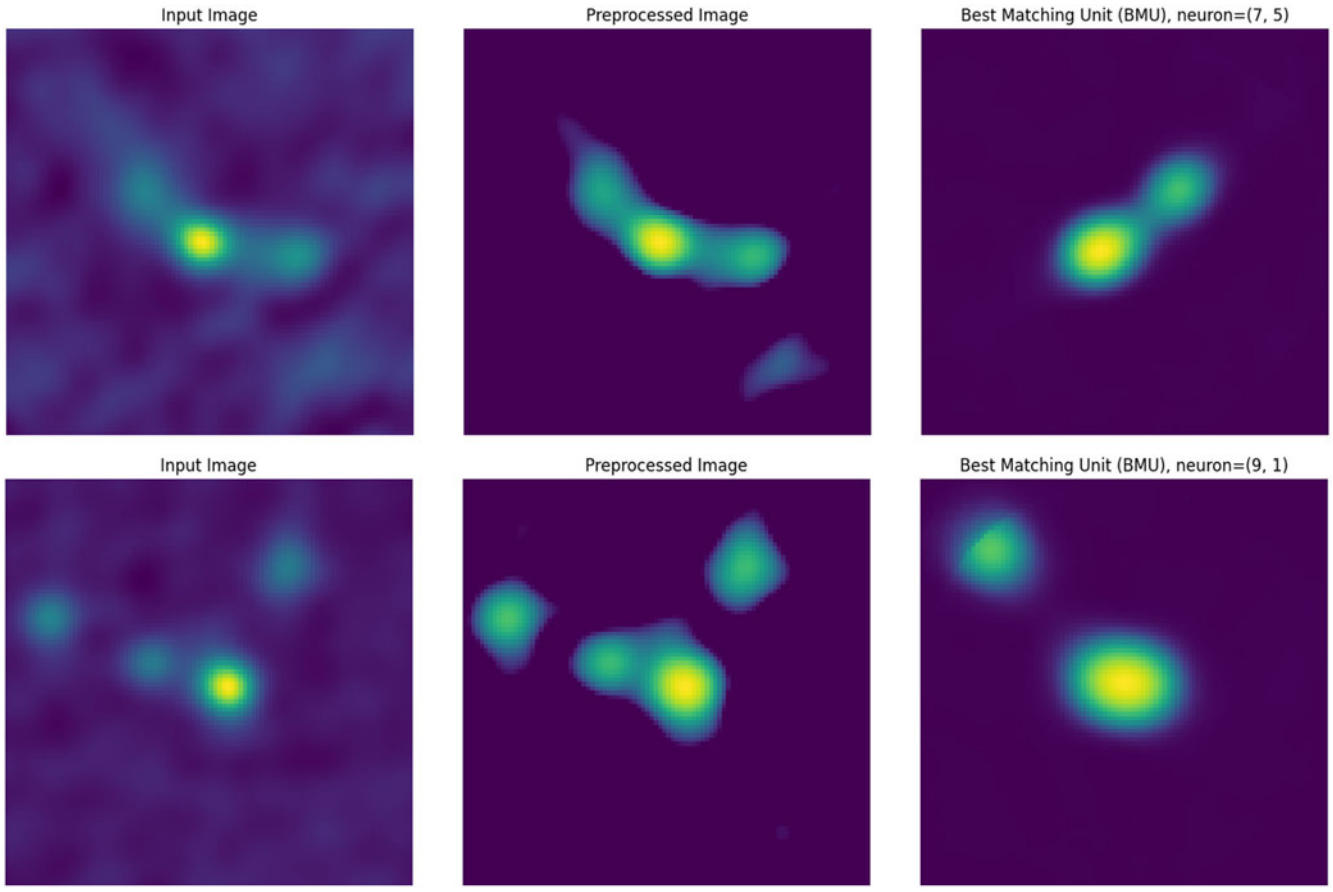

Figure 8. Examples of a 'No' match for an Input Image, and its corresponding Preprocessed Image and Best-Matching Unit (BMU).

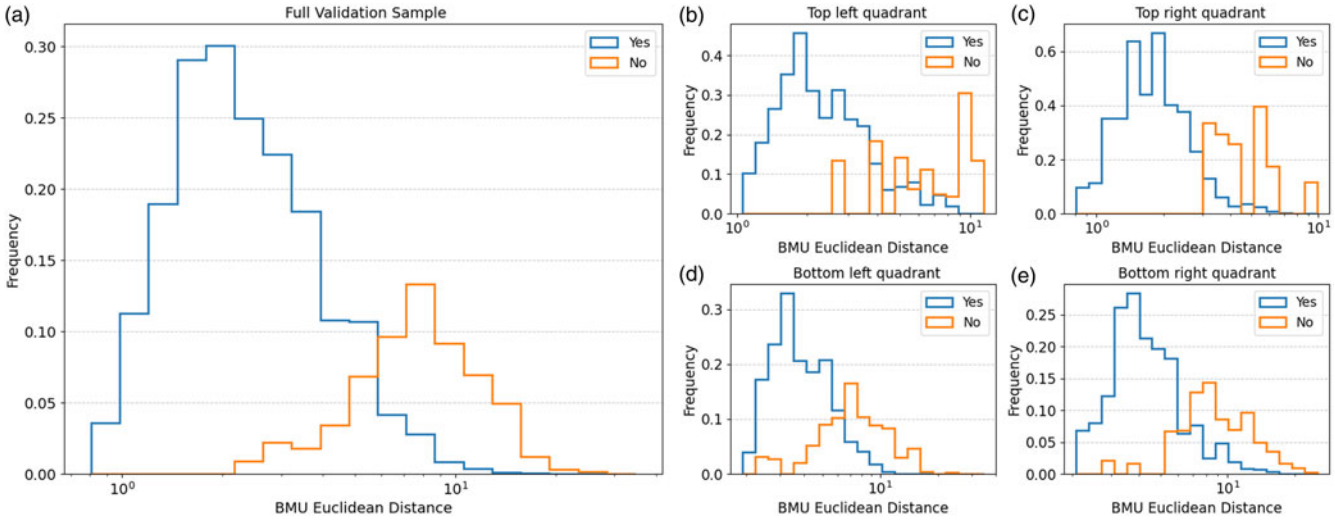

Figure 9. The distribution of the Euclidean distance between input images and their corresponding BMU for the 'Yes' and 'No' matches in the validation sample (a). The distance distributions for the sources in the validation sample grouped into SOM regions: Top left quadrant (b), top right quadrant (c), bottom left quadrant (d), and bottom right quadrant (e).

reliable. In Interval 3, the percentage of 'Yes' matches (67.2%) have decreased from the previous intervals, but is still considerably higher than the 'No' matches. In Interval 4 and beyond (Euclidean distance larger than 6.83), the number of 'Yes' matches continue to decrease and are significantly less than their 'No' counterparts. As a result, their morphological labels become less reliable, with Intervals 7 and 8 being the least reliable given that more than 90% of their input images do not visually match their

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

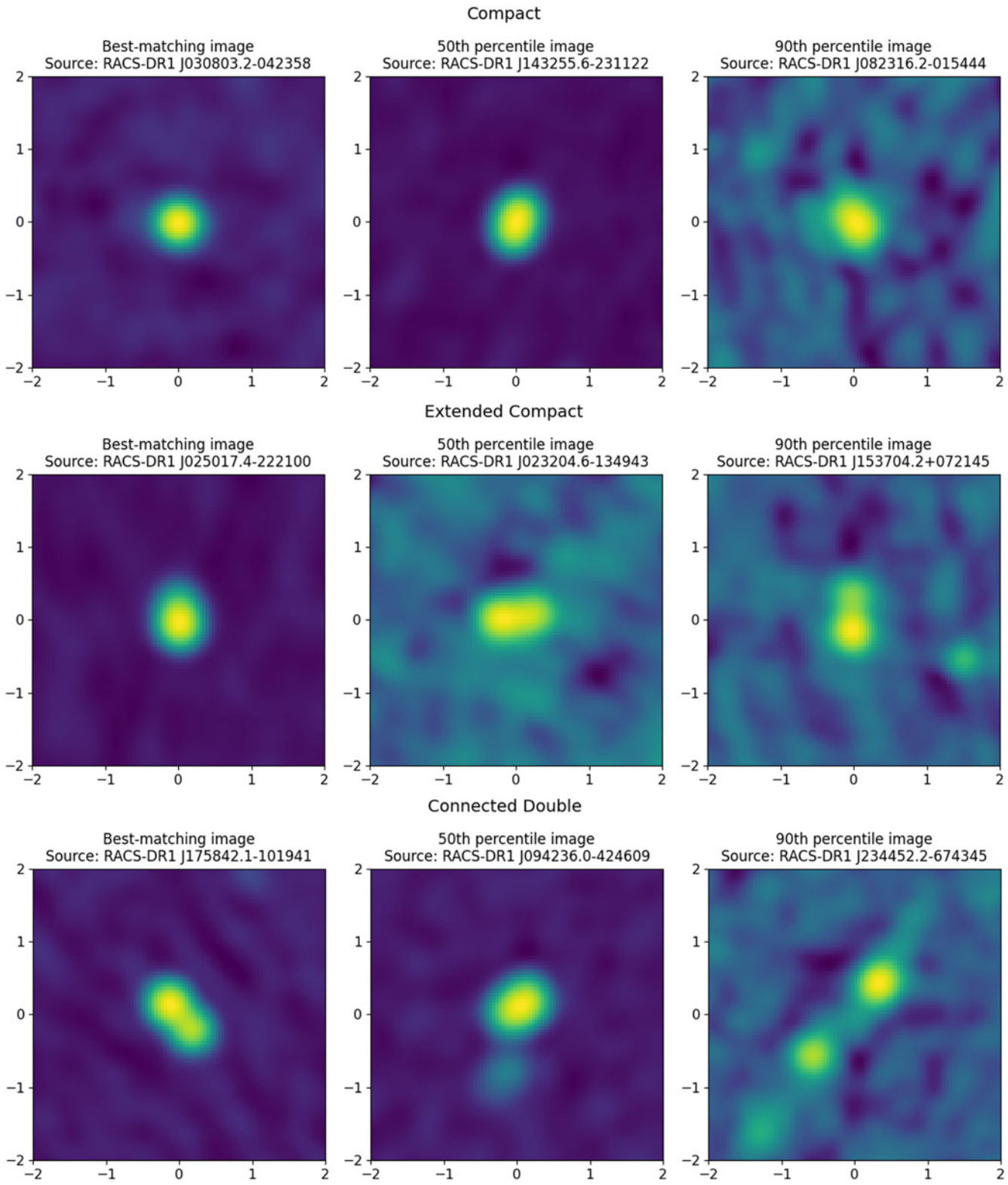

Figure 10. The best-matching, 50th percentile and 90th percentile images for the labels Compact (C), Extended Compact (EC) and Connected Double (CD).

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

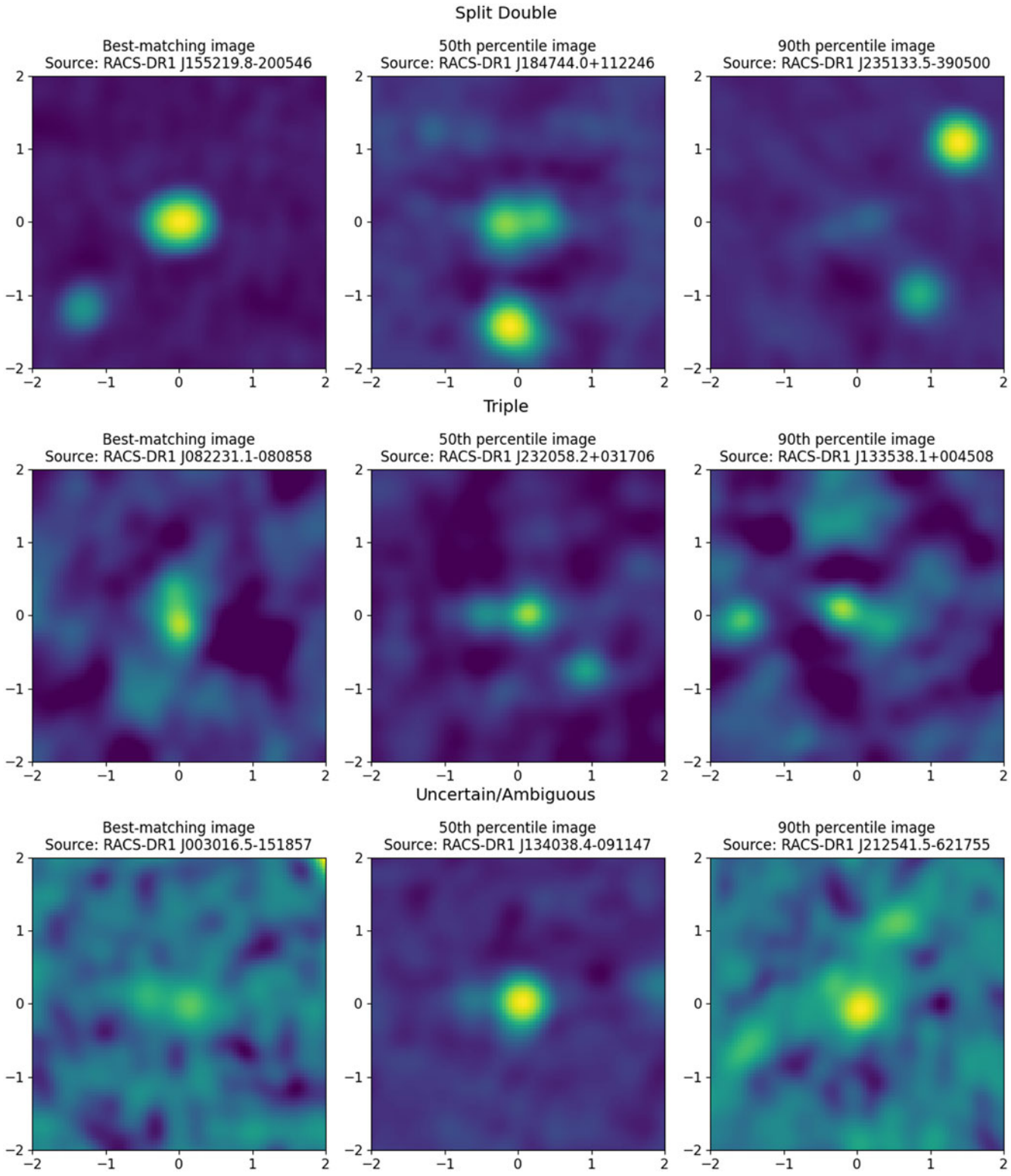

Figure 11. The best-matching, 50th percentile and 90th percentile images for the labels Split Double (SD), Triple (T) and Uncertain/Ambiguous (U/A).

BMU. This provides us with a potential second subset of unique and complex galaxies for further study and can enable us to potentially find rare and unexpected objects within the RACS catalogue. Therefore, for Euclidean distances to the BMU $\lesssim 5$, which account for approximately 79% of our sources, the reliability of the morphological labels in our catalogue is estimated to be >90%, and this reliability drops down to less than 70% at Euclidean distances $\gtrsim 7$. We also qualitatively group the validation sample into SOM

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

Table 3. Summary of the classification of the morphological labels. From left to right we give the morphological label, the number of neurons which were assigned said label, the total number of sources in the RACS catalogue once the labels were transferred, and the split of the sources into each reliability percentage from the validation process based on Euclidean distances.

Morphological label	Neurons	Total sources	98.5%	91.0%	67.2%	40.9%	20.6%	11.9%	7.4%	4.0%
Compact	6	26 217	23 964	2 225	28	0	0	0	0	0
Extended compact	25	81 396	59 920	18 964	2 250	238	23	1	0	0
Connected double	48	106 671	29 181	48 612	16 915	6 745	2 866	1 273	529	550
Split double	14	25 379	18	6 644	11 345	4 753	1 724	567	210	118
Triple	2	1 169	2	317	669	165	13	3	0	0
Uncertain/Ambiguous	5	10 427	1 307	6 565	2 294	238	22	1	0	0

sub-regions and plot the distribution of the 'Yes' and 'No' matches to see if it varies depending on the region (the lower panels b-e in Fig. 6). This is done by splitting the full validation sample according to which SOM quadrant (see Fig. 3 for the four quadrants marked in red) the BMU of the validation sources are located in: top left (235 sources), top right (285 sources), bottom left (264 sources), and bottom right (297 sources). It should be noted that since the sampling for the validation subset was done randomly we do not have equal numbers of each quadrant present. However, the general trend of a higher number of 'Yes' matches than 'No' at lower Euclidean distances, particularly for the first three intervals, is still present. In addition, for the top left (panel a) and top right (panel b) quadrants we see that there are fewer sources with high Euclidean distances that fall within the intervals on the right hand side unlike the bottom quadrants (panels c and d). This indicates that the sources in the top quadrants skew towards lower Euclidean distances.

Subsequently, we plot the distribution of the Euclidean distances between all the input images and their corresponding BMU in the validation sample for both 'Yes' and 'No' matches (panel a in Fig. 9), as well as the split into the SOM quadrants (panels b to e). For the full validation sample we see that the distribution for 'Yes' matches peaks at ∼2, and the skews heavily towards the left-hand side of the graph at lower Euclidean distances with the majority of the values falling within distances of ∼5. Whereas, the distribution for 'No' matches span a wider range of Euclidean distances, and when compared to the 'Yes' matches it skews more towards the higher end of Euclidean distances. This indicates that at higher Euclidean distances we expect the fraction of input images which are similar to their BMU to start decreasing. This trend generally holds in the four SOM quadrants with 'Yes' matches skewing mostly towards lower Euclidean distances, and the 'No' matches leaning more towards higher distances in comparison (panels b-e in Fig. 9). We can also see that for the top left and top right quadrants there are fewer 'No' matches on the left-hand side which is in contrast to the bottom left and bottom right quadrants where the distributions vary over a wider range of distances. These results can be attributed to the neurons in the top SOM quadrants generally being dominated by relatively simple and smaller structures, whereas the neurons at the bottom quadrants have larger or more extended sources. These larger and more extended sources are more likely to have higher levels of background and noise when compared to the simpler sources even following image preprocessing. In addition, their extended sizes could prevent them from being fully captured by the current cutout size of 4′. As a result, they might not be modelled as well as simpler sources during the SOM training. As such, we would expect the Euclidean distances between the top quadrant neurons and their input images to be comparatively lower than for the bottom quadrants (see Figs. A1, A2, A3, A4 in Appendix A for the distributions of the Euclidean distances split by individual neurons in each quadrant for the whole dataset).

4.1 Annotating the SOM

Once a reliability threshold for the similarity between an input image and its BMU had been established, the next stage is to manually label or tag each of the 100 neurons based on their shown morphology. We have decided on 6 tags which broadly encompasses the morphologies seen in the SOM grid:

- **C** (Compact) sources without any significant features other than the central core and are circular or nearly circular.
- **EC** (Extended Compact) sources which are compact sources with either a bright central core and some extended structures, such as an elongated compact core, a tail or additional neighbouring components.
- **CD** (Connected Double) sources which comprise of two distinguishable lobes of comparable sizes or brightness which are either connected or the angular distance between them is relatively minimal.
- **SD** (Split Double) sources which comprise of two distinguishable lobes of comparable sizes or brightness with a clear angular separation where the separation is relatively large.
- **T** (Triple) sources which comprise of three distinguishable lobes of relatively comparable sizes or brightness.
- **U/A** (Uncertain/Ambiguous) sources which contains characteristics that might be present in more than one of the previous labels and it is not clear which label would be the best fit.

We tag the individual neurons with the aforementioned labels and transfer these labels to each of the input images from their BMU (see Figs. 10 and 11 to see examples of the best-matching, 50th percentile and 90th percentile input images for each of the labels). Once the labels have been transferred, we can combine them with the reliability percentage from the validation scheme.

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

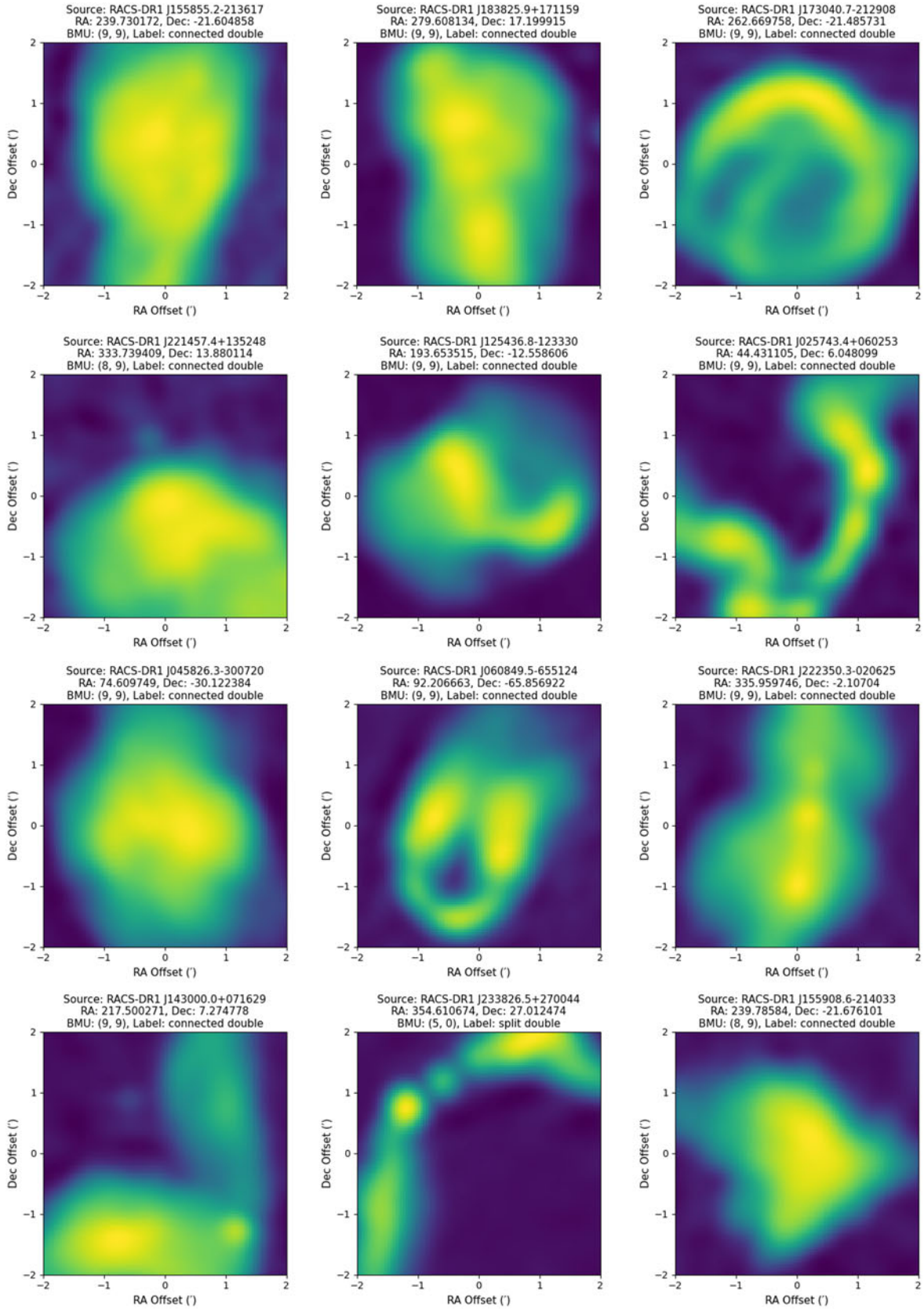

Figure 12. The 12 sources with the largest Euclidean distances between their input images and BMU. For each source we give its source name, RA, Dec, BMU in the SOM grid and its morphological label after label transfer. The distance for these sources range from 34.36 to 43.82 and they all have a reliability percentage of around 4.0% based on the validation scheme.

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

Table 3 gives a summary of the classification of the morphological labels along with the number of neurons which have been assigned these labels, the total number of sources in the RACS catalogue after label transfer, and the split of these sources into the reliability percentages from the validation process: 98.5%, 91.0%, 67.2%, 40.9%, 20.6%, 11.9%, 7.4% and 4.0%. The labels 'Compact' and 'Extended Compact' can be considered the most reliable since a greater fraction of their total sources (91.4% and 73.6%, respectively) have a reliability of $\gtrsim$98.5%. This is to be expected given their relatively simple morphologies. We also see that the majority of neurons assigned these labels are located in the top left and top right quadrants of the SOM which generally has lower Euclidean distances both in the validation sample (Fig. 9) and in the overall SOM (Figs. A1, A2, A3, A4 in Appendix A). Approximately 73% of 'Connected Double' sources fall within the reliability threshold of 91.0% or higher which demonstrates an overall good match. For 'Split Double' and 'Triple' sources, the biggest fraction of sources have around 67.2% reliability match and so these labels can be considered moderately reliable. It should be noted that the label 'Triple' has the fewest number of sources, approximately 0.45% of the total sources, and the two neurons assigned this label are also least commonly chosen as BMU (Fig. 4). As such, while the reliability percentage for this label is relatively moderate, the limited number of sources does limit the extent to which the label can be utilised. Majority of the 'Uncertain/Ambiguous' sources have a high reliability match of $\gtrsim$91.0% which indicates that the associated neurons capture the ambiguous nature of their morphologies well. We can further assess the reliability of the transferred morphological label for each source by also considering the morphological labels of its next best-matching neurons to see if they are consistent. We find that for 75.55% of sources, the labels of its second best-matching neuron matches the label from its BMU. For third and fourth best-matching neurons the percentage of matches with the BMU label decreases to 71.83% and 68.98%, respectively. This aligns with our expectations as the similarity between the source and the neuron will diverge with increasing Euclidean distance. Therefore, for the majority of sources the transferred morphological labels from the BMU are consistent with the labels of its next best-matching neurons, especially the second best-matching neuron.

It is important to note that these labels are not universally agreed upon labels but are instead subjective and based upon our visual inspections of the neurons on this specific trained SOM. As such, they are not transferable to other SOMs even if they are trained on the same data. However, the labels once transferred to our value-added catalogue of RACS sources will help us quickly distinguish the broad morphological features present within the dataset (Rudnick 2021; Bowles et al. 2023). Moreover, it can help us quickly identify atypical and rare sources. In Fig. 12 we show the 12 sources with the largest Euclidean distances between their input images and their BMU. These sources have relatively unusual and interesting morphologies, and their high BMU Euclidean distances is due to them not being very well represented by the neurons. These sources have a reliability percentage of around 4.0% from our validation scheme, which further indicates that their SOM-derived morphological labels are not very dependable. A more thorough study of these sources will be done in the next paper.

Table 4. Description of the columns in the catalogue created in this paper.

Catalogue column	Description
`source_name`	Name of the source as given in the RACS catalogue which follows the IAU convention JHHMMSS.S±DDMMSS with the prefix RACS-DR1
`source_id`	ID of the source as given in the RACS catalogue which is the RACS tile ID along with the Src_ID generated by PyBDSF
`ra`	Right Ascension coordinate of the source (in degrees)
`dec`	Declination coordinate of the source (in degrees)
`bmu`	The position of the Best-Matching Unit (BMU), i.e. the neuron which best-matched the input image, in the SOM grid (Fig. 3). The neuron coordinates are in the form (y, x).
`euclidean_distance`	The Euclidean distance between the source and BMU in the SOM grid
`morphological_label`	The SOM-derived morphological label (see 4.1 for more information on the labels used)
`match_percent`	The reliability percentage which gives the percentage of input images which matched with its BMU based on visual inspection of a smaller validation sample

4.2 Catalogue of complex sources

For each source in our dataset, we add the position of its BMU in the SOM grid, the Euclidean distance between it and the BMU, the transferred morphological label based on the visual inspection of the individual neurons, and the reliability percentage based on the Euclidean distance to create our final catalogue of complex sources. The catalogue contains the following columns: `source_name`, `source_id`, `ra`, and `dec` from the RACS catalogue, `bmu` which gives the position of the best-matching neuron in the SOM grid, `euclidean_distance` which gives the Euclidean distance between the BMU and the input image, the `morphological_label` based on observed morphological features present based on the visual inspections, `match_percent` which gives the reliability percentage, i.e. the percentage of sample input images which matched with its BMU based on the Euclidean distance (see Table 4 for more details on the columns). The first 30 rows of the catalogue are shown in Table 5. The full catalogue produced in this paper will be made available in CDS VizieR (Ochsenbein, Bauer, & Marcout 2000) and other key databases after publication.

5. Conclusions

Next-generation surveys are expected to identify vast number of sources, and as a result will require novel methods of cross-identification. Machine learning methods, especially SOMs, can be used to address the problem of finding complex radio sources in the large dataset provided by SKA pathfinders, such as RACS. In order to do so, we build and train a SOM on sources with multi-Gaussian components from the RACS-Low catalogue. Once

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

Table 5. The first 30 rows from the final catalogue of complex sources created using the SOM.

source_name	source_id	ra	dec	bmu	euclidean_distance	morphological_label	match_percent
RACS-DR1 J001232.8+135445	RACS_0000+12A_1102	3.136777	13.912659	(1, 3)	1.613091	extended compact	98.5%
RACS-DR1 J001218.2+120733	RACS_0000+12A_1115	3.076231	12.125925	(9, 6)	5.108846	connected double	67.2%
RACS-DR1 J001217.3+140104	RACS_0000+12A_1122	3.072128	14.017922	(4, 9)	3.465037	extended compact	91.0%
RACS-DR1 J001213.7+134434	RACS_0000+12A_1137	3.057311	13.742932	(2, 1)	4.860659	connected double	67.2%
RACS-DR1 J001159.7+111637	RACS_0000+12A_1141	2.998874	11.277090	(9, 6)	4.630018	connected double	91.0%
RACS-DR1 J001201.7+120116	RACS_0000+12A_1149	3.007460	12.021187	(0, 3)	1.390036	extended compact	98.5%
RACS-DR1 J001155.1+101847	RACS_0000+12A_1155	2.979664	10.313124	(9, 7)	7.687910	connected double	40.9%
RACS-DR1 J001152.5+125156	RACS_0000+12A_1174	2.968830	12.865726	(5, 8)	3.627663	connected double	91.0%
RACS-DR1 J001156.9+135007	RACS_0000+12A_1178	2.987112	13.835327	(7, 1)	4.852356	split double	67.2%
RACS-DR1 J001140.2+134548	RACS_0000+12A_1210	2.917669	13.763403	(7, 3)	3.632242	connected double	91.0%
RACS-DR1 J001135.7+124553	RACS_0000+12A_1217	2.898931	12.764777	(6, 5)	4.224261	connected double	91.0%
RACS-DR1 J001129.2+104835	RACS_0000+12A_1221	2.871919	10.809794	(0, 5)	2.472950	compact	98.5%
RACS-DR1 J001122.6+101542	RACS_0000+12A_1229	2.844409	10.261744	(5, 0)	5.126052	split double	67.2%
RACS-DR1 J001131.1+134736	RACS_0000+12A_1230	2.879728	13.793518	(1, 8)	1.701950	compact	98.5%
RACS-DR1 J001119.8+100738	RACS_0000+12A_1236	2.832730	10.127255	(0, 2)	2.427596	extended compact	98.5%
RACS-DR1 J001115.9+111800	RACS_0000+12A_1246	2.816345	11.300248	(4, 4)	2.145172	connected double	98.5%
RACS-DR1 J001115.4+144607	RACS_0000+12A_1265	2.814549	14.768678	(0, 6)	2.099214	compact	98.5%
RACS-DR1 J001109.8+122838	RACS_0000+12A_1267	2.791108	12.477418	(8, 5)	3.741234	connected double	91.0%
RACS-DR1 J001108.2+123532	RACS_0000+12A_1272	2.784499	12.592249	(0, 6)	1.701353	compact	98.5%
RACS-DR1 J001106.3+125027	RACS_0000+12A_1273	2.776347	12.840854	(6, 7)	2.314408	connected double	98.5%
RACS-DR1 J001030.7+105827	RACS_0000+12A_1336	2.628291	10.974313	(1, 8)	1.747133	compact	98.5%
RACS-DR1 J001034.6+133848	RACS_0000+12A_1338	2.644539	13.646904	(5, 9)	2.476797	connected double	98.5%
RACS-DR1 J001023.8+121937	RACS_0000+12A_1363	2.599256	12.327161	(2, 3)	3.860162	connected double	91.0%
RACS-DR1 J001022.2+134639	RACS_0000+12A_1368	2.592654	13.777504	(4, 7)	3.052129	connected double	91.0%
RACS-DR1 J001018.2+143337	RACS_0000+12A_1375	2.576152	14.560517	(0, 8)	1.829436	compact	98.5%
RACS-DR1 J000952.3+124426	RACS_0000+12A_1411	2.468288	12.740782	(9, 9)	15.941885	connected double	4.0%
RACS-DR1 J000951.3+141738	RACS_0000+12A_1427	2.463940	14.293970	(4, 0)	5.859904	split double	67.2%
RACS-DR1 J000945.2+141442	RACS_0000+12A_1444	2.438447	14.245052	(1, 1)	2.470556	extended compact	98.5%
RACS-DR1 J000927.4+095842	RACS_0000+12A_1456	2.364251	9.978556	(3, 3)	1.716938	connected double	98.5%
RACS-DR1 J000933.5+144146	RACS_0000+12A_1465	2.389622	14.696120	(1, 2)	1.801990	extended compact	98.5%

the SOM is trained, each input image has a neuron which has been assigned as its best representative or BMU. We label the neurons based on observable morphological structures and then transfer these labels back to the sources from their BMU. This yields a catalogue of complex radio sources, which can be used for further studies. We visually inspect a smaller subset of input images and their BMU to determine a reliability threshold for the similarity metric, which in this case is a modified Euclidean distance. We find that for Euclidean distances of less than 2.74 there is around a 98.5% chance that a randomly chosen input image will match its BMU, but this percentage decreases with Euclidean distance as expected. This, however, gives us the opportunity to study the most unusual and rare objects present in the data by filtering the catalogue to identify sources with high BMU Euclidean distances, and this will be the topic of our next paper.

The catalogue created consists of 251 259 objects from RACS-Low and has additional columns added which include: the best-matched neuron or BMU, the Euclidean distance between the input image and its BMU, the morphology label based on its BMU as well as a general confidence level calculated through visual inspections.

Acknowledgement. Y.A.G. was supported by the US National Science Foundation (NSF) Grant AST 22-06053.

Our research made use of parallelised rotation and flipping INvariant Kohonen-maps or `PINK`[3] as well as the python package `PYINK`.[4] We would like to thank Dr Tim Galvin for his help and advice on the installation and usage of `PINK`.

This work made use of Astropy:[5] a community-developed core Python package and an ecosystem of tools and resources for astronomy (Astropy Collaboration et al. 2013; Astropy Collaboration et al. 2018; Astropy Collaboration et al. 2022) as well as the Astropy affiliated package Astroquery

[3] `PINK`: https://github.com/HITS-AIN/PINK; Polsterer et al. 2016.
[4] `PYINK`: https://github.com/tjgalvin/pyink.
[5] http://www.astropy.org.

(Ginsburg et al. 2019). Other packages used in this paper were Matplotlib (Hunter 2007) and NumPy (Harris et al. 2020).

This scientific work uses data obtained from Inyarrimanha Ilgari Bundara/the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (https://ror.org/05qajvd42). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund. This paper includes archived data obtained through the CSIRO ASKAP Science Data Archive, CASDA (https://data.csiro.au).

Data availability. The catalogue of SOM-derived complex sources will be uploaded to CDS VizieR and other key databases after publication, and the links will be made available in due course.

References

Abolfathi, B., et al. 2018, ApJS, 235, 42
Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20
Astropy Collaboration, et al. 2013, A&A, 558, A33
Astropy Collaboration, et al. 2018, AJ, 156, 123
Astropy Collaboration, et al. 2022, ApJ, 935, 167
Banfield, J. K., et al. 2015, MNRAS, 453, 2327
Baron, D. 2019, Machine Learning in Astronomy: A Practical Overview, doi: 10.48550/ARXIV.1904.07248
Becker, R. H., White, R. L., & Helfand, D. J. 1995, ApJ, 450, 559
Bock, D. C.-J., Large, M. I., & Sadler, E. M. 1999, AJ, 117, 1578
Bowles, M., et al. 2023, MNRAS, 522, 2584
Condon, J. J. 1992, ARA&A, 30, 575
Condon, J. J., et al. 1998, AJ, 115, 1693
Edge, D. O., Shakeshaft, J. R., McAdam, W. B., Baldwin, J. E., & Archer, S. 1959, MmRAS, 68, 37
Ekers, J. A. 1969, AuJPA, 7, 3
Everitt, B., Landau, S., Leese, M., & Stahl, D. 2011, Cluster Analysis (5th edn.; Wiley)
Fanaroff, B. L., & Riley, J. M. 1974, MNRAS, 167, 31P
Galvin, T. J., et al. 2019, PASP, 131, 108009
Galvin, T. J., et al. 2020, MNRAS, 497, 2730
Ginsburg, A., et al. 2019, AJ, 157, 98
Gordon, Y. A., et al. 2023, ApJS, 267, 37
Gürkan, G., et al. 2022, MNRAS, 512, 6104
Hale, C. L., et al. 2021, PASA, 38, doi: 10.1017/pasa.2021.47
Harris, C. R., et al. 2020, Natur, 585, 357
Hotan, A. W., et al. 2021, PASA, 38, e009
Hunter, J. D. 2007, CSE, 9, 90
Hurley-Walker, N., et al. 2017, MNRAS, 464, 1146
Ikotun, A., Ezugwu, A., Abualigah, L., Abuhaija, B., & Heming, J. 2022, InS, 622, doi: 10.1016/j.ins.2022.11.139
Intema, H. T., Jagannathan, P., Mooley, K. P., & Frail, D. A. 2017, A&A, 598, A78
Johnston, S., et al. 2008, ExA, 22, 151
Jolliffe, I., & Cadima, J. 2016, PTRSA, 374, 20150202
Kellermann, K. I., Sramek, R., Schmidt, M., Shaffer, D. B., & Green, R. 1989, AJ, 98, 1195
Kohonen, T. 1990, Proc. IEEE, 78, 1464
Kohonen, T. 2001, Self-Organizing Maps (3rd edn.; Springer)
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Lacy, M., et al. 2020, PASP, 132, 035001
Lara, L., et al. 2001, A&A, 370, 409
Lukic, V., et al. 2018, MNRAS, 476, 246
McConnell, D., et al. 2020, PASA, 37, doi: 10.1017/pasa.2020.41
Mohan, N., & Rafferty, D. 2015, PyBDSF: Python Blob Detection and Source Finder, Astrophysics Source Code Library, record ascl:1502.007
Mostert, R. I. J., et al. 2021, A&A, 645, A89
Norris, R. P. 2017, NatAs, 1, 671–678
Norris, R. P., et al. 2011, PASA, 28, 215
Norris, R. P., et al. 2021, PASA, 38, e046
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Padovani, P., et al. 2017, A&ARv, 25, doi: 10.1007/s00159-017-0102-9
Perley, R. A., Chandler, C. J., Butler, B. J., & Wrobel, J. M. 2011, ApJ, 739, L1
Polsterer, K., Gieseke, F., Igel, C., Doser, B., & Gianniotis, N. 2016, in 24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27–29, 2016, Bruges, Belgium, publication status: Published
Polsterer, K. L., Gieseke, F., & Igel, C. 2015, in Astronomical Society of the Pacific Conference Series, Vol. 495, Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV), ed. A. R. Taylor, & E. Rosolowsky, 81
Proctor, D. D. 2016, ApJS, 224, 18
Rengelink, R. B., et al. 1997, A&AS, 124, 259
Rudnick, L. 2021, Galaxies, 9, 85
Sadler, E. M., Jenkins, C. R., & Kotanyi, C. G. 1989, MNRAS, 240, 591
Savage, A., & Wall, J. V. 1976, AuJPA, 39, 39
Shimwell, T. W., et al. 2017, A&A, 598, A104
Sutherland, W., & Saunders, W. 1992, MNRAS, 259, 413
Tingay, S. J., et al. 2013, PASA, 30, doi: 10.1017/pasa.2012.007
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
van Haarlem, M. P., et al. 2013, A&A, 556, A2
Vantyghem, A. N., et al. 2024, A&C, 47, 100824
Wayth, R. B., et al. 2018, PASA, 35, doi: 10.1017/pasa.2018.37
Williams, W. L., et al. 2019, A&A, 622, A2
Windhorst, R. A., Kron, R. G., & Koo, D. C. 1984, A&AS, 58, 39
Windhorst, R. A., Miley, G. K., Owen, F. N., Kron, R. G., & Koo, D. C. 1985, ApJ, 289, 494
Wright, E. L., et al. 2010, AJ, 140, 1868
York, D. G., et al. 2000, AJ, 120, 1579–1587

Appendix A. BMU Euclidean Distances of individual neurons in SOM grid

In Section 4, we have Fig. 5 which gives the distributions of the BMU Euclidean distance for the overall dataset. Here, we have included additional Figs. A1, A2, A3, and A4 which explore the distribution of the BMU Euclidean distance for individual neurons in the trained SOM grid divided into the four quadrants in the grid, i.e. the top left, top right, bottom left and bottom right (see the quadrants as indicated in red in Fig. 3). On the top quadrants, we mostly have smaller or simpler structures for which the Euclidean distances skews more towards the lower end. However, there is a trend for higher Euclidean distances in the bottom quadrants, especially the bottom right quadrant, and this can be attributed to there being larger or more extended sources in this SOM regions compared to the top left quadrant of the SOM. For these sources, we would expect a higher level of background and noise pixels, and their cutout size might also not be large enough to capture their full extent. As such, the higher Euclidean distances could be due to them not being modelled as well during SOM training.

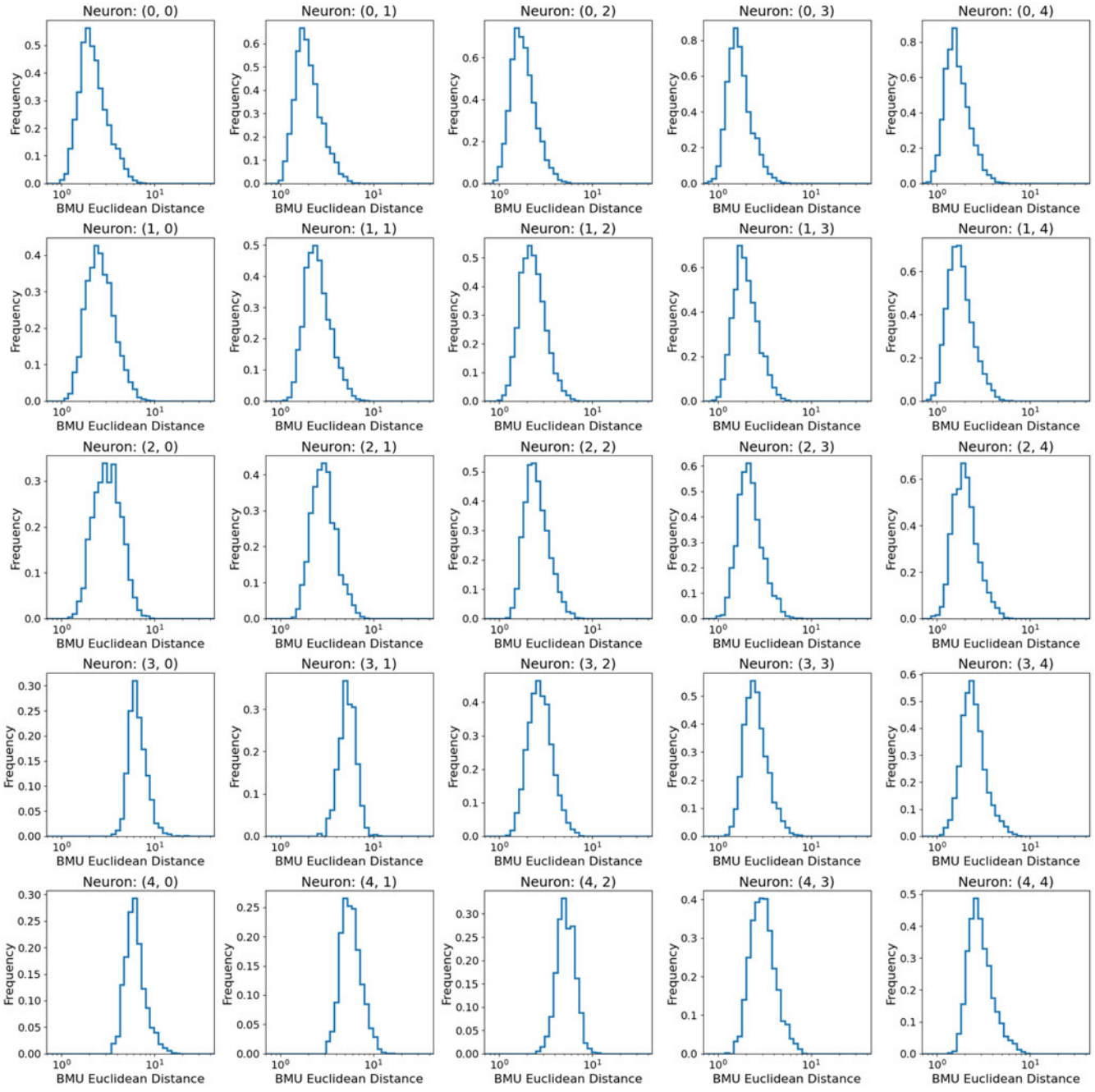

Figure A1. The distributions of the BMU Euclidean distance between an individual neuron in the SOM grid and all the input images for which it was chosen as the BMU for all the neurons in the top left quadrant of the SOM (Fig. 3).

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

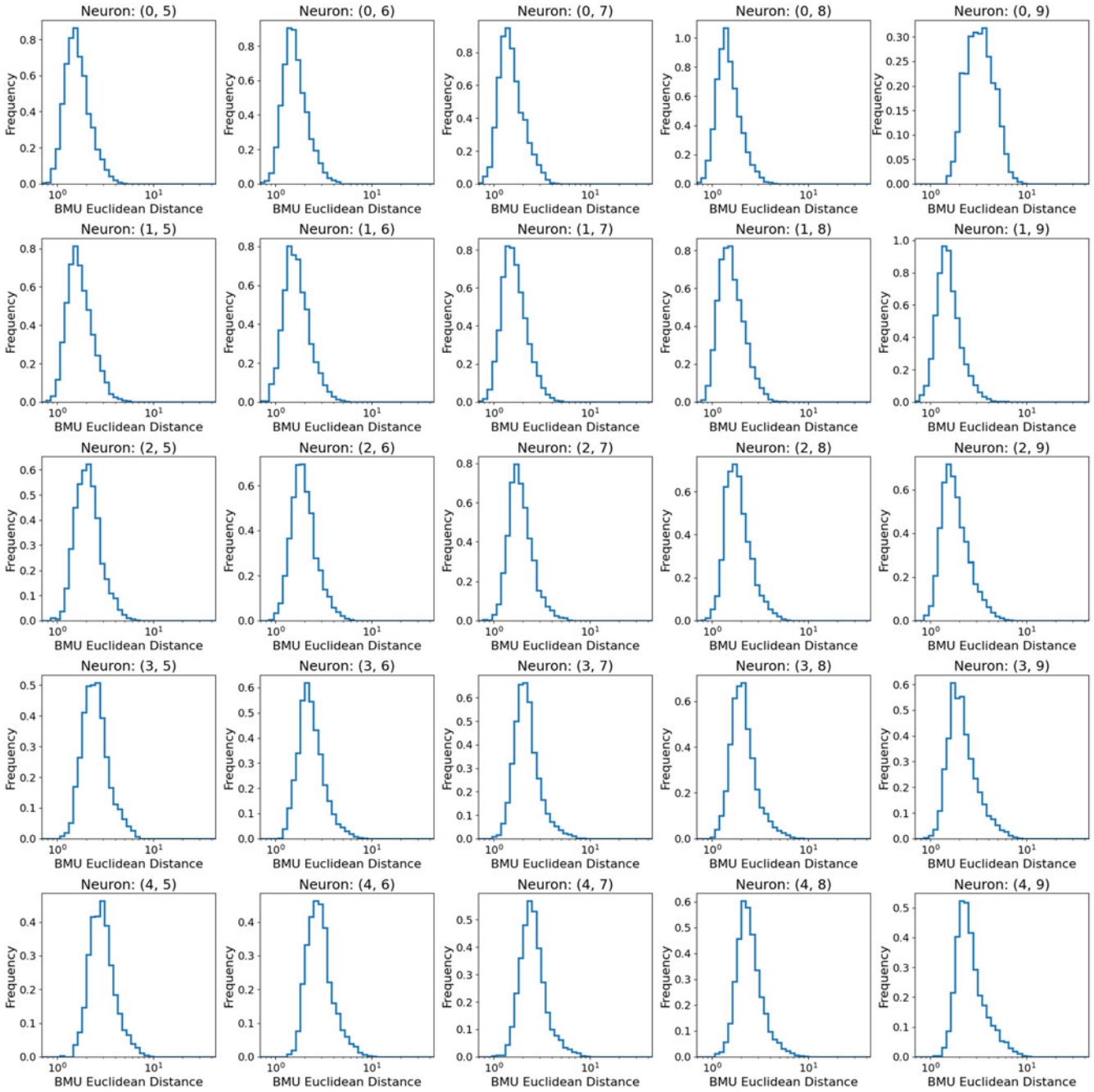

Figure A2. The distributions of the BMU Euclidean distance between an individual neuron in the SOM grid and all the input images for which it was chosen as the BMU for all the neurons in the top right quadrant of the SOM (Fig. 3).

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

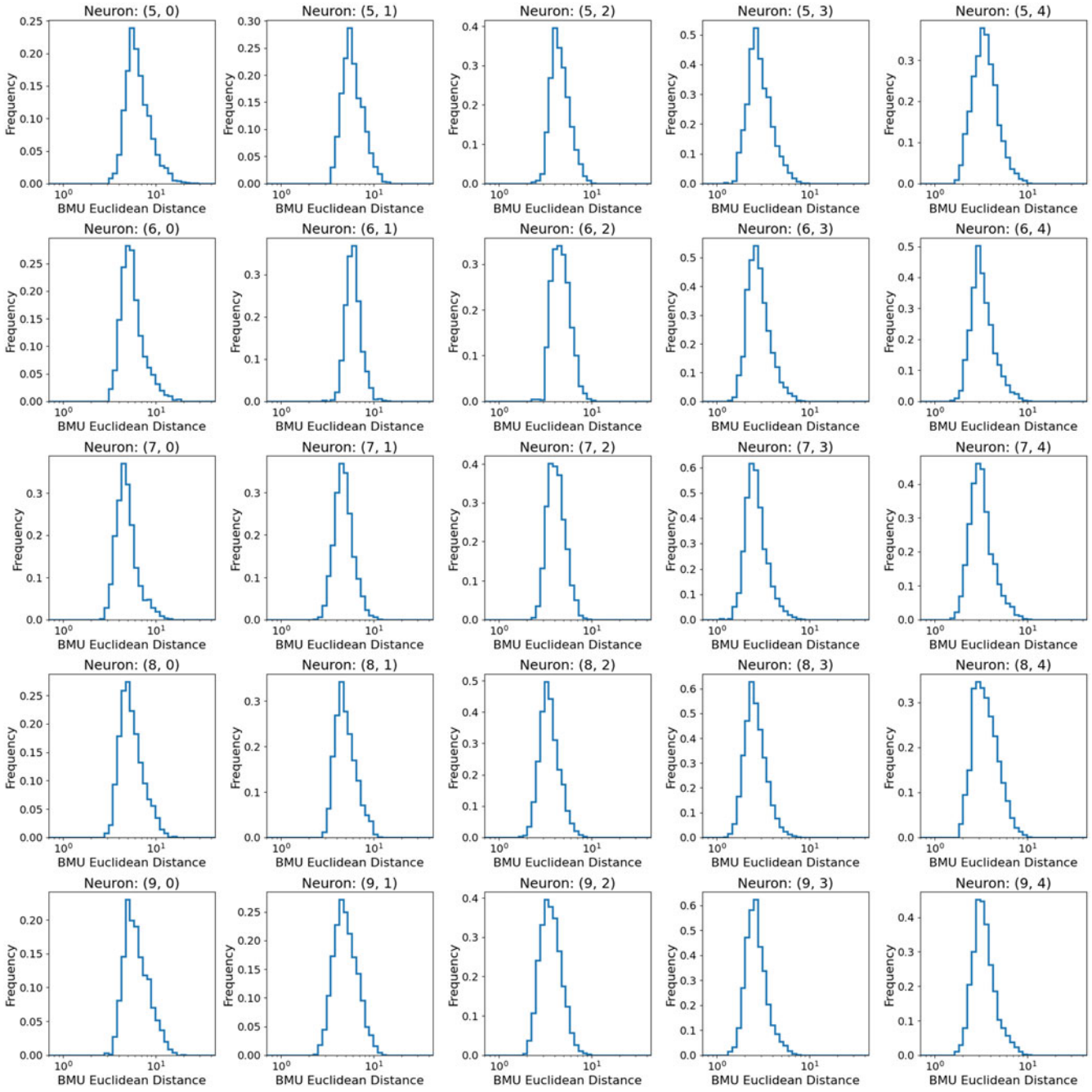

Figure A3. The distributions of the BMU Euclidean distance between an individual neuron in the SOM grid and all the input images for which it was chosen as the BMU for all the neurons in the bottom left quadrant of the SOM (Fig. 3).

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press

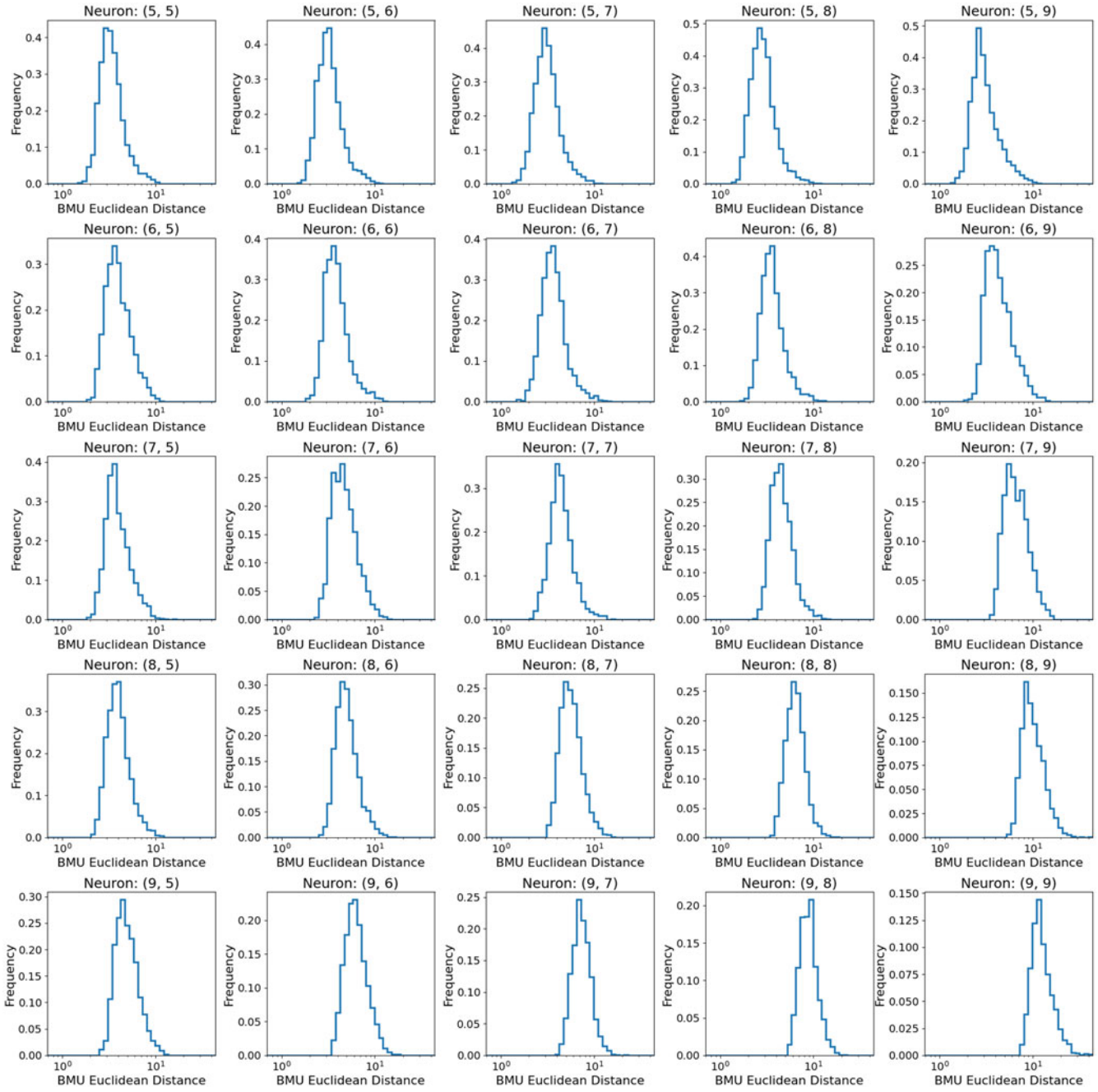

Figure A4. The distributions of the BMU Euclidean distance between an individual neuron in the SOM grid and all the input images for which it was chosen as the BMU for all the neurons in the bottom right quadrant of the SOM (Fig. 3).

https://doi.org/10.1017/pasa.2024.133 Published online by Cambridge University Press