A machine-learning classifier for LOFAR radio galaxy cross-matching techniques

Lara Alegre [ID],[1]★ Jose Sabater,[1,2] Philip Best,[1] Rafaël I. J. Mostert,[3,4] Wendy L. Williams,[4] Gülay Gürkan [ID],[5] Martin J. Hardcastle [ID],[6] Rohit Kondapally [ID],[1] Tim W. Shimwell[3,4] and Daniel J. B. Smith [ID][6]

[1]*SUPA, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[2]*UK Astronomy Technology Centre, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[3]*Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands*
[4]*ASTRON, the Netherlands Institute for Radio Astronomy, Oude Hoogeveensedijk 4, NL-7991 PD Dwingeloo, the Netherlands*
[5]*Thüringer Landessternwarte Tautenburg (TLS), Sternwarte 5, D-07778 Tautenburg, Germany*
[6]*Centre for Astrophysics Research, Department of Physics, Astronomy and Mathematics, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK*

ABSTRACT

New-generation radio telescopes like LOFAR are conducting extensive sky surveys, detecting millions of sources. To maximize the scientific value of these surveys, radio source components must be properly associated into physical sources before being cross-matched with their optical/infrared counterparts. In this paper, we use machine learning to identify those radio sources for which either source association is required or statistical cross-matching to optical/infrared catalogues is unreliable. We train a binary classifier using manual annotations from the LOFAR Two-metre Sky Survey (LoTSS). We find that, compared to a classification model based on just the radio source parameters, the addition of features of the nearest-neighbour radio sources, the potential optical host galaxy, and the radio source composition in terms of Gaussian components, all improve model performance. Our best model, a gradient boosting classifier, achieves an accuracy of 95 per cent on a balanced data set and 96 per cent on the whole (unbalanced) sample after optimizing the classification threshold. Unsurprisingly, the classifier performs best on small, unresolved radio sources, reaching almost 99 per cent accuracy for sources smaller than 15 arcsec, but still achieves 70 per cent accuracy on resolved sources. It flags 68 per cent more sources than required as needing visual inspection, but this is still fewer than the manually developed decision tree used in LoTSS, while also having a lower rate of wrongly accepted sources for statistical analysis. The results have an immediate practical application for cross-matching the next LoTSS data releases and can be generalized to other radio surveys.

Key words: methods: statistical – galaxies: active – radio continuum: galaxies.

1 INTRODUCTION

The number of detected sources and the complexity of the structures in astronomical images has increased dramatically in recent years, with high-sensitivity telescopes surveying deeper but also wider areas of the sky. Radio astronomy has been at the forefront of this big data revolution, with telescopes like the LOw Frequency ARray (LOFAR; van Haarlem et al. 2013), the Very Large Array, and the Australian Square Kilometre Array Pathfinder Telescope (ASKAP; Hotan et al. 2021). These have been conducting wide radio continuum surveys, such as the LOFAR Two-meter Sky Survey (LoTSS; Shimwell et al. 2017, 2019, 2022), the VLA Sky Survey (VLASS; Lacy et al. 2020), and the Rapid ASKAP Continuum Survey (RACS; Hale et al. 2021), and the Evolutionary Map of the Universe (EMU; Norris et al. 2011), respectively. When completed, these surveys will have covered both

hemispheres and discovered tens of millions of radio sources. This brings radio astronomy into a revolutionary new era: large samples enable detailed statistical studies whilst probing the unexplored Universe at these wavelengths (see Norris 2017 for a review). In addition to producing scientific results, these surveys are also developing technology in preparation for the upcoming Square Kilometer Array (SKA; Dewdney et al. 2009), which will be the world's most powerful radio telescope. The SKA will generate massive amounts of data and is expected to detect billions of radio sources.

In order to extract the full scientific return from these surveys, it is essential to cross-match the objects detected at radio wavelengths to their counterparts at other wavelengths, particularly optical and near-infrared. This allows us to identify the host galaxies, classify the radio sources according to their morphology, black hole activity, and other characteristics, and derive basic physical properties such as redshifts, luminosities and stellar masses (e.g. Best et al. 2005; Smolčić et al. 2017; Duncan et al. 2019; Gürkan et al. 2022). The cross-identification of radio galaxies with their optical (or

★ E-mail: alegre@roe.ac.uk

infrared) counterparts is a complex process due to the extended and multicomponent nature of many radio sources, as well as the mismatch in the angular resolution between the radio and optical surveys. Traditionally, it has relied mostly on statistical methods, visual analysis, or a combination of the two (see Williams et al. 2019, hereafter referred as W19, for a discussion).

In early continuum radio surveys, the sources detected were mainly bright active galactic nuclei (AGNs); only a small proportion of these had counterparts in the all-sky optical imaging data available at that time, but the samples were small enough that dedicated deep optical imaging of individual sources could be coupled with visual analysis (e.g. Laing, Riley & Longair 1983). By the turn of the century, a statistical comparison of the Faint Images of the Radio Sky at Twenty centimetres survey (FIRST; Becker, White & Helfand 1995) with the large-area optical imaging from the Sloan Digital Sky Survey (SDSS; York et al. 2000) provided optical identifications for around 30 per cent of the $\sim 10^5$ radio source host galaxies (Ivezić et al. 2002). Recent radio surveys have been revealing still fainter sources, including higher fractions of star-forming galaxies (SFGs) that begin to dominate over AGNs at low flux densities . At the same time, deeper optical and near-infrared observations are now available over large sky areas, such as imaging from the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS-1) survey (Chambers et al. 2016) or the Dark Energy Spectroscopic Instrument (DESI) Legacy survey (Dey et al. 2019), with even deeper and wider imaging expected in the coming years from the Large Survey of Space and Time (LSST; Ivezić et al. 2019) and the Euclid Space Telescope surveys (Laureijs et al. 2011). These surveys increase both the fraction of radio sources with optical counterparts and the number of potentially confusing foreground or background sources. The simultaneous increase of possible matches and data volumes requires improvement in the current cross-matching techniques.

In LoTSS, the source density is already more than a factor of 10 times higher than in the existing widely used large-area radio continuum surveys such as the National Radio Astronomy Observatory (NRAO) VLA Sky Survey (NVSS; Condon et al. 1998), the FIRST survey, the Sydney University Molonglo Sky Survey (SUMSS; Bock, Large & Sadler 1999), and the Westerbork Northern Sky Survey (WENSS; Rengelink et al. 1997). LoTSS detected more than 300 000 sources in its first data release, containing just the first 2 per cent of the survey (LoTSS DR1; Shimwell et al. 2019), and a second data release with almost 4.4 million sources covering 5634 $\deg^2$, 27 per cent of the northern sky, has just been published (LoTSS DR2; Shimwell et al. 2022).

In LoTSS DR1, the radio sources were cross-matched with optical and near-infrared surveys, Pan-STARRS1 DR1 (Chambers et al. 2016) and the AllWISE catalogue (Cutri et al. 2013), respectively, and an optical and/or near-infrared counterpart was identifiable for 73 per cent of the LoTSS sources (W19). Compact sources, such as SFGs or compact AGNs, were cross-matched using the Likelihood Ratio technique (LR; e.g. Richter 1975; Willis & de Ruiter 1977; Sutherland & Saunders 1992; Ciliegi et al. 2003) which assesses the relative probability of a given optical source being a true counterpart against a randomly aligned optical object, based on source properties (for LoTSS DR1, the LR assessment considered both the magnitude and colour of the potential host galaxy; see Nisbet 2018, W19). This statistical method is reliable when the flux-weighted mean position of the radio emission is an accurate estimate of the location at which the radio source originates, and is therefore coincident with the optical emission. However, more extended sources cannot yet be reliably handled through these statistical methods. Furthermore, for radio sources with emission that is extended and/or split into different radio

components (e.g. double-lobed sources), source detection algorithms often fail to correctly group together the multiple radio components into a single source, generating independent entries in the radio catalogues. In other cases, the source finder can incorrectly group individual physical radio sources together into a single-blended detection. Thus, radio catalogues are not always a true description of the physical sources, leading to further inaccuracies if statistical techniques are naively applied. In LoTSS DR1, these complex-structured, multicomponent and blended sources were therefore visually cross-matched alongside manual component association or dissociation.

In order to discriminate between sources that require visual analysis and those that can be reliably cross-matched using the LR technique, W19 designed a decision tree based on the properties of the radio sources and their cross-ID LR values. This decision tree selected nearly 30 000 sources for visual inspection, corresponding to around 10 per cent of the total LoTSS DR1 sample. This was a conservative selection process, and indeed post-analysis (i.e. the examination of which ones actually required visual inspection, explained in Section 2) shows that only just over half of these sources actually required to be inspected. LoTSS DR2 covers an area almost 15 times larger than DR1, with a higher fraction of counterparts expected due to the use of the (deeper) Legacy data set for cross-matching. The large number of sources makes visual inspection very challenging for more than a small fraction of the sources, while the ultimate goal is to replace all visual analysis with automated techniques, a more practical and immediate step is to minimize the amount of unnecessary inspection.

Some progress has been made to improve the current statistical methods, for example by modifying the LR technique to tackle the blending problem (Weston et al. 2018), by replacing the LR by Bayesian approaches (Fan et al. 2015, 2020; Mallinar, Budávári & Lemson 2017), or by applying machine-learning (ML) techniques (e.g. Alger et al. 2018). Various efforts have also been made to improve the cross-matching process for the extended/multicomponent radio sources, using a ridgeline approach (Barkus et al. 2022) and deep learning techniques mainly based on Convolutional Neural Networks (CNNs), for example to group radio source components (Mostert et al., in preparation) or to find the host galaxy in previously-selected sources with multiple radio components (Alger et al. 2018). CNNs have also been used to improve the source finding and identification (e.g. Vafaei Sadr et al. 2019), or for automatic source extraction and further morphology classification (e.g. Wu et al. 2019). Deep learning has been particularly successful in automating radio galaxy morphology classification of (previously associated) multicomponent sources using CNNs (e.g. Aniyan & Thorat 2017; Lukic et al. 2018, 2019; Alhassan, Taylor & Vaccari 2018), using transfer learning (Tang, Scaife & Leahy 2019) and using clustering methods (Galvin et al. 2020; Mostert et al. 2021) combined, for example, with Haralick features (Ntwaetsile & Geach 2021). However, deep learning models, which perform feature extraction from the images before classification, require a higher number of annotated examples to train, and are also more difficult to interpret and to adapt than simpler ML models. In addition, a variety of unforeseen limitations due to limited experimentation in radio astronomy can further introduce different biases. Some examples include issues related to the use of fixed-size data images (Mostert et al. 2021) or even the image input file format (Tang et al. 2019). Furthermore, none of these methods can yet perform reliable source association and fully cross-match extended and multicomponent sources. To date, the full cross-matching of modern large radio surveys has been only achieved through citizen science projects [e.g. Radio Galaxy Zoo (RGZ), Banfield et al. 2015] and extensive science team

efforts [e.g. LOFAR Galaxy Zoo (LGZ), W19; Kondapally et al. 2021].

In this work, we propose a gradient boosting classifier (GBC) to identify which radio sources can be reliably cross-matched using the LR technique, or instead require visual inspection. We use supervised ML algorithms, which offer greater intuitive interpretation and are simpler to adjust and analyse than deep learning models. The model adopted is an ensemble of decision trees, and it was selected and optimized using Automated Machine Learning (AutoML; see Appendix A and He, Zhao & Chu 2021 for a review). While individual decision trees have been used in radio galaxy classification in the past (e.g. Proctor 2016), ensembles of decision trees have been proven to achieve better performance (Dietterich 2000). Examples of the use of ensembles of decision trees in radio astronomy include the classification of blazars using multiwavelength data (Arsioli & Dedin 2020) and the estimation of physical properties of radio sources such as redshifts (Luken et al. 2022).

We build a data set based on LoTSS DR1, which provides more than 300 000 annotated examples, and select a set of relevant features, allowing the model to successfully classify unseen sources with an accuracy of 94.6 per cent and select the ones that can be cross-matched by LR with a precision of 96.3 per cent. This helps to limit the manual analysis to the most complex sources (extended sources, sources with multiple components or blended detections), which are those for which the LR method is not successful. The results of this study are already being incorporated, by helping to identify unrelated radio components, into the automatic component association of sources larger than 15 arcsec from LoTSS DR1 (Mostert et al., in preparation). Furthermore, the methods applied in LoTSS DR1 are directly transferable to other parts of the LoTSS survey since the techniques used for processing and cross-matching the next data releases are broadly similar. Therefore, our work has immediate practical benefit for deciding which sources require visual analysis in LoTSS DR2 (Hardcastle et al. in preparation).

The paper is organized as follows. In Section 2, we describe the LoTSS DR1 data and in Section 3 we explain how these data were used to create a data set suitable for our ML classification problem. Section 4 refers to the experiments performed to select and optimize the model, including the specifications of the model adopted. The model performance and interpretation are explained in Section 5. In Section 6, we interpret the results of the model applied to the full LoTSS data sets, discussing the implications and comparing them against the methods currently used. The conclusions and a discussion of their significance for the next LoTSS data releases can be found in Section 7.

2 DATA

The data used in this work consist of LoTSS DR1 (Shimwell et al. 2019)[1] radio catalogues that were derived from the 58 mosaic images of DR1, which cover 424 deg^2 over the Hobby–Eberly Telescope Dark Energy Experiment (HETDEX; Hill et al. 2008) Spring Field (right ascension $10^h45^m00^s$ – $15^h30^m00^s$ and declination $45°00'00''$ – $57°00'00''$). LoTSS has a frequency coverage from 120 to 168 MHz, and achieves a typical rms noise level of 70 μJy beam^{-1} over the DR1 region, with estimated point source completeness of 90 per cent at a flux density of 0.45 mJy. LOFAR's low frequencies combined with high sensitivity on short baselines gives it high efficiency at detecting extended radio emission. LoTSS DR1 has an angular resolution of

6 arcsec and an astrometric precision of 0.2 arcsec, making it robust for host-galaxy identification.

In LoTSS DR1, the source detection was performed using the Python Blob Detector and Source Finder (PYBDSF; Mohan & Rafferty 2015), where a total of 325 694 PYBDSF sources were extracted with a peak detection above 5σ. PYBDSF fits Gaussians to pixel islands assigning one or multiple Gaussians to each PYBDSF source. The radio catalogues with the PYBDSF properties for both the sources and the Gaussians include positions, angular sizes and orientations, peak and integrated flux density as well as statistical errors.

PYBDSF sources do not always represent true radio sources (i.e. physically connected sources). Some of the radio components of extended sources may appear as separated and unrelated PYBDSF sources, which need to be associated together into the same source in post-processing. We refer to these as *multicomponent* sources in the rest of the paper and they account for 2.8 per cent of LoTSS DR1. In other cases, Gaussians may be incorrectly grouped into one PYBDSF source when they are actually distinct physical sources. In this case, we refer to them as *blended* sources and they make up only 0.3 per cent of LoTSS DR1. In the vast majority of cases (96.9 per cent in LoTSS DR1), however, PYBDSF correctly associates the radio emission into true physical sources. We refer to these hereafter as *single* sources. These are, in most cases, compact sources composed of only one Gaussian, but can also be extended sources composed by various Gaussians (hence our definition of singles is not the same as the 'S' code from the PYBDSF software used in W19). Even for these correctly associated sources, however, cross-matching with other surveys using statistical means alone can fail due to an incorrect (or missed) counterpart identification, especially if the source is extended and/or asymmetric. This is the case for 1.8 per cent of the sources of LoTSS DR1.

In order to enhance science quality, as part of LoTSS DR1, considerable effort was undertaken to properly associate the radio source components (or dissociate blended sources) and get the correct optical/near-infrared counterparts (W19). For the majority of LoTSS DR1 sources, PYBDSF correctly associates source components and outputs an accurate estimate of the position and radio source properties, and therefore such sources were cross-matched statistically using LR. However, complex sources with multiple components or extended emission, and incorrectly blended sources, were sent to visual inspection. This was carried out on a private LOFAR Galaxy Zoo (LGZ) project, hosted on the Zooniverse platform,[2] in which each source was inspected by at least 5 collaborators of the LOFAR consortium. The selection of the sources to be analysed in LGZ was done using a decision tree (also referred to as *flowchart*) built using the characteristics of the PYBDSF sources and Gaussians, the neighbouring sources, and the LR of any optical/IR cross-matching (see W19).

The decision tree generates three main outcomes: the source association and/or identification requires *LGZ*; the source has been correctly catalogued by PYBDSF and the cross-identification (or lack of) can be made by LR; and the source is sent to a quick visual sorting (*prefiltering*), where one expert inspects the source and redirects it to one of the other two categories or identifies it as an artefact. A summary of the number of sources in each of these categories is given in Table 1, where we include in the prefiltering category 223 sources with large optical IDs that were automatically matched to a nearby (large angular size) SDSS or 2MASX galaxy since they were afterwards visually confirmed. We further exclude 2591 PYBDSF

---

[1]https://lofar-surveys.org

[2]https://www.zooniverse.org

Table 1. For each of the main categories (LR, LGZ, and prefiltering) classified by the W19 decision tree, the table gives the number of sources that were suitable for LR and the number that required visual analysis, as determined using the final outcomes after visual inspection. The final column indicates the percentage of the time that the flowchart decision was correct (i.e. the proportion of sources that were assigned correctly to each of the categories).

W19 decision	Total number	No. suitable for LR	No. requiring visual analysis	Percentage correct
LR	295 225	294 129	1096[a]	99.63
LGZ	8195	3144	5051	61.64
Prefiltering	19 683	10 079	9604	48.79
Total	323 103	307 352	15 751	95.57

[a]The 1096 sources selected by the decision tree for LR, but identified as requiring visual analysis, represent a lower limit to the true number as these were only identified when they were part of multicomponent sources for which other components were sent to LGZ (see Section 6.2 for further discussion of this).

sources identified by W19 as artefacts, except for one source which was automatically marked by the decision tree as an artefact but was instead noted during the LGZ process to be a genuine source. In LoTSS DR1, the artefacts were either removed in an initial stage of the selection process (the majority by being in the proximity of bright sources; 31 per cent) or by visual inspection (mainly during the prefiltering step; 55 per cent). In the next, LoTSS releases the improved calibration and imaging pipeline for the radio data (Tasse et al. 2021) means that we expect a lower proportion of artefacts, most of which will be clearly identifiable and removed at early stages. Furthermore, the properties of any remaining artefacts may be different due to calibration changes. For these reasons, we exclude the artefacts when constructing the ML classifier and analysing the results; our final data catalogue therefore contains 323 103 PYBDSF sources. The values quoted in Table 1 refer to PYBDSF sources and are different to the ones presented on table 5 of W19 that summarizes the total number of sources after component association or dissociation.

Using the decision tree, W19 initially classified 91.37 per cent of the sources (295 225) as being suitable for LR analysis (see Table 1) and 8.63 per cent (27 878 sources) as requiring visual inspection (either prefiltering or LGZ). These numbers correspond to sources after removal of artefacts. After visual analysis and processing of the final DR1 data, in hindsight, the conclusion is (see Section 3.1) that 95.13 per cent (307 352) could be cross-matched using LR and 4.87 per cent (15 751) required visual inspection. For the sources that were sent directly to LGZ (8195 PYBDSF sources), an examination of the final LGZ decision indicates that 5051 of them (61.64 per cent) were not correctly associated by PYBDSF and therefore, could not have had their optical identification assigned statistically by LR (or lack of identification in case of no LR match). Similarly, the prefiltering step corresponds to 19 683 PYBDSF sources for which 9604 PYBDSF sources (48.79 per cent) could not have been processed using LR. In contrast, from the 295 225 PYBDSF sources selected as suitable for cross-matching with LR, 294 129 of them (99.63 per cent) retain the LR cross-match in the final catalogue. In reality, the number of these that are correct will actually be marginally lower since these sources were not subjected to visual examination unless they were part of a multicomponent source (usually the core of a radio source) for which one of the source components was sent to visual analysis. This was the method through which the 1096 sources, sent by the decision tree to LR but which required visual analysis, were discovered. We discuss this in more detail in Section 6.2.
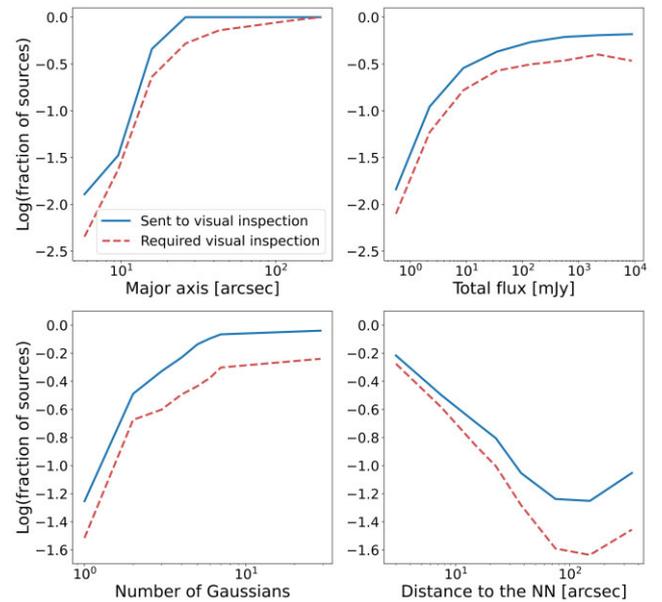
Figure 1. Fraction of PYBDSF sources sent to visual inspection by the W19 decision tree (blue lines) and the ones that actually required to be inspected (as determined from the final visual inspection outcomes; red dashed lines) as a function of different source parameters: major axis length, total flux density, total number of Gaussians that compose each PYBDSF source, and distance of each PYBDSF to its nearest neighbour (NN) PYBDSF source.

It is evident from Table 1 that overall the W19 decision tree has a high accuracy (95.57 per cent). This is mainly because most of the sources are compact and can be cross-matched by LR (where the application of statistical methods results in very high precision). However, the decision tree places about twice as many sources in to the LGZ and prefiltering categories as required, increasing the burden on visual analysis. Fig. 1 illustrates the dependence of the decision tree outcomes on some key PYBDSF source properties: the major axis length, the total radio flux density, the number of Gaussians that compose a PYBDSF source, and the distance of each PYBDSF source to its nearest neighbour (NN). In each panel, the blue line shows the fraction of sources that were sent to visual inspection, and the red dashed line shows the fraction of sources that actually needed to be inspected, as determined from the final cross-matched catalogues incorporating the LGZ outcomes. The plots show that the fraction sent for visual analysis increases with increasing source size (note that 15 arcsec was the limit used by W19 to distinguish between 'small' and 'large' sources, with all the large sources being visually inspected, either directly in LGZ or during the prefiltering stage), increasing flux density, increasing number of Gaussian components, and decreasing distances to the NN. These are in line with expectations, as they are all indications that a given source is more likely to be extended and complex. Interestingly, in all cases the red lines are broadly scaled down from the fractions sent to LGZ by about a factor of 2 with no strong parameter dependencies (fluctuations range only from around 1.5–2.5 across the parameter space). This indicates that it would not be straightforward to improve the decision tree outcomes simply by adjusting these parameter values.

3 DATA SET

In supervised ML, models are learned from a set of labelled examples drawn from the *data set*. The goal is to predict to which *class* a

previously unseen example belongs based on the value of its *features*. The data set is a key input for training the ML model and relies upon an adequate and well-profiled number of examples. We create our data set by evaluating all 323 103 PYBDSF sources from LoTSS DR1 based on their individual characteristics and assigning them to different classes (Section 3.1). We create different sets of features by using radio source parameters and optical information (Section 3.2), and we address the class imbalance problem by exploring different ways of balancing the data set (Section 3.3). The impact of these last two factors on the classification is investigated further in Section 4.

3.1 Classes

To create the classes, we first evaluate each PYBDSF source (after the results of any deblending or LGZ source association) and assigned them an 'association flag' according to different outcomes: the ones that were neither deblended nor associated with other PYBDSF sources (*singles*, flag 1); sources that were deblended (*blends*, flag 2); and PYBDSF sources that were grouped with other PYBDSF sources (*multicomponents*, flag 4). Note that a small number of sources have a combination of flags, since they were first deblended and afterwards one or more of the deblended components was grouped with another PYBDSF source (leading to flag 6).

To create these outcomes, the correspondence between each PYBDSF source and the final radio source association (or lack of association) was assessed using the PYBDSF radio source catalogue from Shimwell et al. (2019) and the final value-added catalogue (source associations and optical IDs) from W19. PYBDSF sources that were grouped with other PYBDSF sources appear as components of a radio source in the corresponding component catalogue, and PYBDSF sources that were deblended appear as two or more radio sources.

To create a final diagnosis, we also inspected the 'single' sources (i.e. the ones with the association flag 1) in order to evaluate whether the LR was a suitable method to identify the host galaxy. This is the case for those sources where the final ID in the value-added catalogue is the same as would have been drawn through LR analysis, or where there was no ID in the final catalogue and the LR analysis also predicted no ID. In contrast, if visual analysis resulted in a change in optical ID (or a change from having no LR ID to having an ID, or vice versa) then these sources are not suitable for cross-matching using the LR method. As a result of this evaluation, the sources were assigned into two classes (denoted by the flag 'accept_lr' throughout this work):

(i) **Class 1:** PYBDSF sources that were not associated with other PYBDSF sources, and were not deblended, and for which LR gave the same outcome as was finally accepted in the value-added catalogue (i.e. same host galaxy ID, or correctly gave no ID). These sources would be suitable for LR analysis.

(ii) **Class 0:** PYBDSF sources that were either associated with other PYBDSF sources in LGZ, or deblended into more than one source, or LR would obtain an incorrect ID. These sources are all unsuitable for analysis by LR alone.

The classes comprise 307 352 sources suitable for LR (class 1) and 15 751 that require visual analysis (class 0); from the latter 9072 are multiple component PYBDSF sources, 857 are blended PYBDSF sources, and 5822 are single sources for which a simple application of LR would produce an incorrect ID. Artefacts (which we exclude from the analysis) correspond to PYBDSF sources that are not in the final DR1 value-added catalogue.

Table 2. List of features used in the analysis. These were selected or calculated using different LoTSS DR1 catalogues*. The LR threshold value adopted in LoTSS DR1 ($L_{thr} = 0.639$) was used to scale LR value features (these have the suffix tlv). Features in logarithmic scale appear with the log prefix. Sources refer to PYBDSF sources.

Features	Definition and origin
Baseline (BL)	
Maj	Source major axis (arcsec)[a]
Min	Source minor axis (arcsec)[a]
Total_flux	Source integrated flux density (mJy)[a]
Peak_flux	Source peak flux density (mJy bm^{-1})[a]
log_n_gauss	No. Gaussians that compose a source[b]
Likelihood ratio (LR)	
log_lr_tlv	Log_{10}(source LR value match/L_{thr})[c]
lr_dist	Distance to the LR ID match (arcsec)[c]
Gaussians (GAUS)	
gauss_maj	Gaussian major axis (arcsec)[b]
gauss_min	Gaussian minor axis (arcsec)[b]
gauss_flux_ratio	Gaussian/source flux ratio[a, b]
log_gauss_lr_tlv	Log_{10}(Gaussian LR/L_{thr})[c]
gauss_lr_dist	Distance to the LR ID match (arcsec)[c]
log_highest_lr_tlv	Log_{10}(source or Gaussian LR/L_{thr})[c]
Nearest neighbour (NN)	
NN_45	No. of sources within $45''$ [a]
NN_dist	Distance to the NN (arcsec)[a]
NN_flux_ratio	NN flux/source flux density ratio[a]
log_NN_lr_tlv	Log_{10}(LR value of the NN/L_{thr})[c]
NN_lr_dist	Distance to the LR ID match (arcsec)[c]
Closest prototype (SOM)	
10x10_closest_prototype_x1	$\cos(2\pi$ Closest_prototype_x/10)[d]
10x10_closest_prototype_x2	$\sin(2\pi$ Closest_prototype_x/10)[d]
10x10_closest_prototype_y1	$\cos(2\pi$ Closest_prototype_y/10)[d]
10x10_closest_prototype_y2	$\sin(2\pi$ Closest_prototype_y/10)[d]

Note. * [a]PYBDSF radio source catalogue (Shimwell et al. 2019).
[b]Gaussian component catalogue (Shimwell et al. 2019).
[c]Gaussian and PYBDSF source LR catalogues (W19).
[d]Self-Organizing Map for LoTSS DR1 (SOM; Mostert et al. 2021).

3.2 Features

As input features for the ML classifier we used radio source parameters along with properties of the LR matches for both the PYBDSF source being considered and its nearest neighbour (NN). We discuss these below and list them in Table 2.

The radio features were built from the PYBDSF catalogue from Shimwell et al. (2019), where each PYBDSF source has an identifier (source name) with the corresponding radio properties; here, we use the major and minor axis sizes and the peak and total flux densities. In addition to these basic radio properties, we used the LR value of the best match and the distance to this match. We computed the LR values for the PYBDSF sources and for each of the Gaussians that comprise a PYBDSF source in the same way as described in W19, with minor modifications that resulted from improvements of the original code (Kondapally et al. 2021).

We also used the Gaussian component catalogue (described in Shimwell et al. 2019), which contains the radio information for all the Gaussians that compose each PYBDSF source. We use the number of Gaussian components comprising a source (indicative of the morphological complexity of the source), and also use the properties (major and minor axis size, fractional source flux density, and LR match properties) of the Gaussian with the highest LR value,

or of the brightest Gaussian if the LR of all Gaussian components is below the LoTSS DR1 LR threshold adopted in W19.

We also used the radio and LR properties of the NN source. In addition, we computed the number of radio sources within 45 arcsec (used as an estimate of the local source density, which might be indicative of the presence of multicomponent sources) and the flux ratio between the source and its NN.

Finally, we investigated using the positions of the LoTSS DR1 sources on a cyclic Self-Organizing Map (SOM; Mostert et al. 2021) as input features. The SOM provides information of the different LoTSS DR1 morphological source 'prototypes' on a two-dimensional grid.

In ML, the quality of the features affects the ability of the model to learn. In order to feed useful features that can be more easily interpreted by the algorithm, we made the following transformations to the data:

(i) We searched the catalogues for missing values (e.g. LR values where there was no potential host within the 15 arcsec search radius) which we assigned extreme values (e.g. a very low arbitrary value of 10^{-135} in the case of LR), even though the tree models adopted in Section 4 can in general handle missing data well.

(ii) We used the log value of the number of Gaussians, since complex sources can be made of dozens of Gaussians (up to 51 in LoTSS DR1).

(iii) We encode the values of the SOM morphological prototypes into cyclical features. The prototypes are located on a square grid with (x, y) coordinates. Each radio source is mapped to the prototype of the SOM that it most resembles. We transformed the corresponding (x, y) coordinates by using a sine and a cosine transform: this creates 2 new features from each of the original ones, but ensures the cyclical nature of the SOM is retained. We set the values of the prototypes to an arbitrary high value of 10^{20} when the source is not available.

(iv) We set the value of the LR to a log scale, although this choice has no effect on our results (decision tree models, which we adopt in Section 4, are not sensitive to feature transformations). Using a log scale allows this feature to be used interchangeably with different classifiers (e.g. neural networks).

(v) We create a feature which uses either the LR of the source or the LR of the Gaussian component with the highest LR value if the LR of the source is lower than the LR of one of the Gaussians that make up the source. This is more indicative of a LR match when the source is composed by multiple Gaussians, one of which traces the radio core (and it is the same if the source is only composed by one Gaussian). This can also be indicative of a blended source, especially if the source LR value is below the LR threshold.

(vi) We further scaled the LR values by dividing them by the LR threshold value used to process the sources in the HETDEX field (only sources for which the match had a LR value higher than 0.639 got an ID or no ID via this method). This has the advantage of making the model appropriate for future LoTSS fields that might use different optical/near-IR data sets with a correspondingly different LR threshold.

3.3 Balancing the data set

The number of objects in the two classes created previously is heavily imbalanced: class 1 has 307 352 sources while class 0 comprises 15 751 objects. The major problem with imbalanced data sets is the tendency of the model to get specialized in the class with more examples (i.e. to overfit to class 1). For that reason, we explore different ways of creating a balanced data set by under-

and oversampling (cf. Collell, Prelec & Patil 2018, and references therein).

We performed undersampling of the majority class by extracting a random sample of 15 751 objects from class 1 (which is the number of sources available in class 0). Undersampling is the standard method adopted throughout the experiments (Section 4); we use 31 502 sources, comprising the same number of examples in both classes.[3] In these experiments, we used a training set (used to train the model) of 75 per cent of the data set, and a test set (used to evaluate the model) of 25 per cent. When performing model selection and optimization (see Section 4.4) we use a 10-fold cross-validation (CV), otherwise we test and train the models on 10 different randomly sampled data sets and use the mean value as the model performance.

Since both under- and oversampling have the potential to affect performance, we conducted experiments to determine which method was the best. We created a synthetic training data set with ADASYN (He et al. 2008), an adaptive sampling technique that is used to generate synthetic examples of the minority class (class 0) by using the original density distribution of the sources in this class. To avoid data leakage, we re-sampled only 75 per cent of the minority class (11 841 sources) and tested on a test set comprising the remaining 25 per cent of these sources (which is balanced as well). The number of sources in the training set before and after re-sampling is 303 386:303 386 for class 1 and 11 841:301 738 for class 0, respectively. We compare the performance using the model trained using under- and oversampling in Section 4.4.2.

Finally, it should be emphasized that although both under- and oversampling techniques aim to create a balanced data set that can generalize well for the two different classes, the distribution of the sources is inherently highly imbalanced and objects that need to be visually inspected are relatively rare in LoTSS (and other deep radio surveys). For that reason, when applying the model trained on a balanced data set to the real (imbalanced) data, which we do in Section 6, other factors require consideration; we discuss these in detail in Section 6.1.

4 EXPERIMENTS

We start by defining in Section 4.1 the metrics that will be used to evaluate the performance of the classifier. In Section 4.2, we create a baseline model for the experiments. This is a less complex yet still effective model that produces acceptable results but has room for improvement. The baseline model was selected using the Tree-Based Pipeline Optimization Tool (TPOT; Olson et al. 2016b), and consists of a GBC; see Appendix A for an overview of the ML models and AutoML tools used. In order to improve model performance, we examine the impact of adding different sets of features in Section 4.3, and optimizing the model hyperparameters in Section 4.4.

4.1 Performance metrics

Accuracy is the most common metric to evaluate the performance of an ML classifier. Accuracy can be given as the percentage of the correctly classified inputs relative to the overall classifications: accuracy = (TP+TN)/(TP+FP+TN+FN), where in our case the numbers of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) correspond to

---

[3]Note that when the SOM features are included, this data set is reduced to 31 320 objects because a small fraction of the LoTSS DR1 sources (on the borders of the mosaics) do not have SOM information.

(i) **TP:** sources correctly classified as suitable for LR methods;

(ii) **FP:** sources that should be visually inspected, but which the classifier deems suitable for LR;

(iii) **TN:** sources correctly classified as requiring visual inspection;

(iv) **FN:** sources that could be done by LR techniques but are being sent by the classifier to visual analysis.

When training and testing with a balanced number of examples in each category, accuracy shows the robustness of the classifier. For our binary classification on a balanced data set, the classifier returns a probability of the source being able to be accepted by LR (class 1), with the probability of being class 0 (requiring LGZ) being 1 minus this probability. 0.50 is the normal threshold value used to discriminate between the two, and it is the value we adopted when evaluating the results in this section and in Section 5. We do, however, investigate other thresholds in order to evaluate the model applied to an imbalanced data set in Section 6. When evaluating the results we are mainly concerned with minimizing the number of sources wrongly accepted through LR (FP), while keeping a low number of sources that need to be sent to visual analysis (FN and TN sources). That is another reason why we further investigate 'threshold moving' in order to establish more suitable cut-off probabilities.

We also analyse the values of recall [also known as sensitivity or true positive rate (TPR)] and precision for our two classes. Precision can be defined as the fraction of sources predicted as being from a certain class that are actually from that class [e.g. TP/(TP+FP)], and recall as the fraction of sources from a certain class that are predicted correctly [e.g. TP/(TP+FN)]. The overall balance between precision and recall for the different classes is given by the F1-score [2×(precision×recall)/(precision + recall)]. In Section 5, we also use the false positive rate (FPR) to illustrate the performance of the classifier. The FPR corresponds to the fraction of sources from class 0 that are incorrectly classified [FP/(FP+TN)].

In our analysis, we further define the 'LGZ scale-up factor' that corresponds to the total number of sources that we would have to visually inspect scaled by the ones we should really inspect [(FN+TN)/(TN+FP)]. In other words, it represents the multiplicative factor of additional galaxies we would have to send to LGZ besides the ones that should be sent. We compare it with the false discovery rate [FDR = FP/(FP+TP)], which corresponds to the fraction of sources deemed to be suitable for cross-matching by LR that are classified incorrectly.

4.2 Baseline

In order to create a baseline model, we used TPOT and a set of baseline features (BL) that contain only basic radio source information: PYBDSF peak and total fluxes, major and minor axis sizes, and the logarithm of the number of Gaussians that compose each PYBDSF source.

We ran TPOT using a set of conservative parameters: three generations (number of iterations of the optimization process; see Appendix A for more details), a population size of 20 (number of candidate solutions TPOT retains in each generation), and a 10-fold CV (number of data splits where each pipeline is trained and evaluated). This allows TPOT to search for 600 different models in each run. The choice of the values for these parameters is subjective, and higher values would enable the search for more model combinations. However, running TPOT for a larger number of generations and population size would drive TPOT towards more complex ML pipelines with stacked models that could cause the

Table 3. Accuracy on the test sets of the GBC model before and after optimization of the hyperparameters, and using cumulative sets of features: baseline features (BL), PYBDSF source LR features (LR), PYBDSF Gaussian features (GAUS), nearest neighbour (NN) and SOM features (SOM) as described in Table 2. In each case the GBC was run on 10 different undersamplings with random sampling of the dataset into training and test sets, and the mean of these 10 is quoted. The standard deviation between the 10 data sets is typically around 0.2 per cent.

Set of features	Accuracy achieved (per cent)	
	Baseline hyperparameters	Optimised GBC
(0) BL	88.7	88.7
(1) BL and LR	90.2	90.2
(2) 1 and GAUS	90.3	90.3
(3) 2 and NN	94.4	94.6
(4) 3 and SOM	94.7	94.8

model to overfit; this is a current challenge of the method (see Olson et al. 2016b for a discussion). Therefore, we define low values for the TPOT parameters, and use it to get recommended pipelines. In that way, we select a simple model that provides interpretability for our experiments and we perform model optimization at a later stage.

We performed different TPOT runs and we found a consistent selection of tree-based models as the favoured choice: using different balanced random samples of the full data set, TPOT would select a GBC or occasionally an XGBoost (XGBoost is an optimized version of a GBC that can include regularization and allows further optimization due to the amount of parameters that can be tuned); when using subsets of the data (half-size data set) a Random Forest or an Extra Trees classifier was favoured. For all the models, we achieved an internal CV accuracy of around 89 per cent and a test accuracy within ±0.5 per cent of the CV value. The GBC achieved higher performance on the CV tests but the Random Forest models showed a higher generalization ability when training with only 50 per cent of the data set. This suggests that the smaller data set does not contain enough examples for TPOT to detect strong patterns among the features and therefore it fits a model that performs well with higher variance data. This also indicates that the classification could benefit from adding more relevant features and could be improved using a GBC model (with optimized hyperparameters, such as a bigger ensemble size and/or a different learning rate). For our baseline model, we therefore select a GBC with 100 estimators and a learning rate of 0.01, which are also the hyperparameters suggested by TPOT. The complete specifications of the baseline model can be seen in Table 4.

4.3 Feature selection

We started with the baseline model and investigated the impact of adding different sets of features (as described in Table 2) to the classifier; their impact on classification is illustrated in Table 3. These comprise four sets of features in addition to the (0) baseline features: (1) LR information of the PYBDSF source; (2) properties of the Gaussian component with the highest LR value (or the brightest Gaussian if none have a LR match); (3) the nearest PYBDSF neighbour information; and (4) the positions of the PYBDSF sources on the SOM.

Source LR features: The addition of the LR features (LR value and LR distance) of the PYBDSF source increases the performance accuracy of the baseline model by about 1.5 per cent (from 88.7 per cent to 90.2 per cent; see Table 3). This improvement is expected, as the presence or absence of a potential host galaxy at the expected position is a strong indicator of whether the source has been correctly associated.
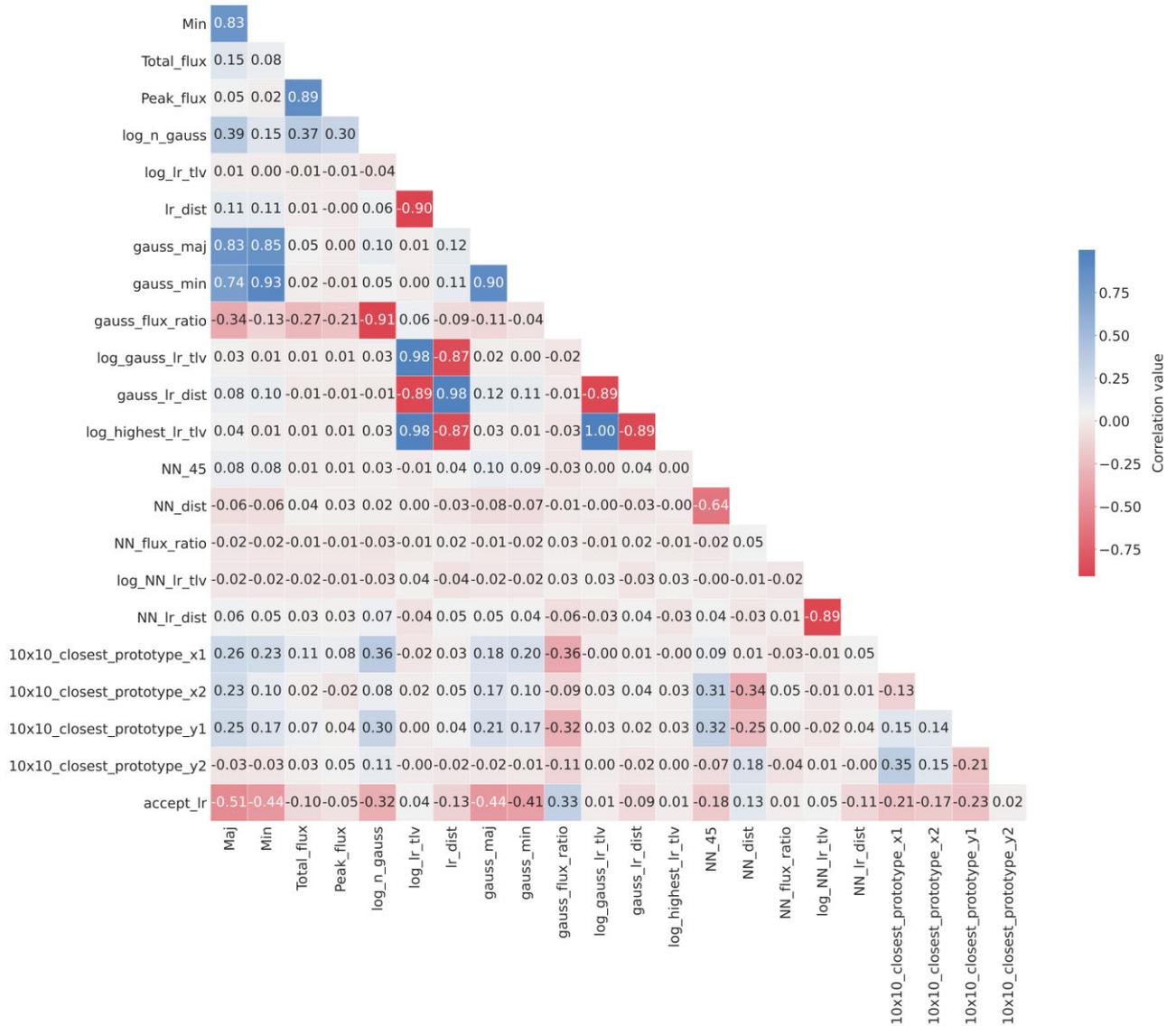
Figure 2. Correlation matrix using Pearson correlation. This shows the correlation coefficients between each of the different input features considered for the modelling (blue for positive linear correlation, red for negative linear correlation, scaling from 1 to -1). The bottom row provides the correlation of each parameter with the final 'accept_lr' outcome, indicating the strength of any linear relation between the features and the target class.

Gaussian (GAUS) features: The addition of the Gaussian features has a small impact on the model with only minor improvements for the classification. When adding these features to the Baseline features and the LR features, the improvement is 0.1 per cent. Fig. 2 shows the correlation between different input features (and with the resulting classification). It is evident from this plot that the flux ratio relative to the source and the size of the Gaussian (gaus_min and gaus_max) do show, respectively, a strong positive and strong negative correlation with the 'accept_lr' output, and thus contain useful information. However, the sizes of the Gaussians show a very strong correlation with the sizes of the sources, the flux ratio between the Gaussian and the source is highly correlated (inversely) with the number of Gaussians that composed each PYBDSF source, and the Gaussian LR features (log_gauss_lr_tlv and gauss_lr_dist) are also highly correlated with the source LR parameters (not least because most sources are composed by single Gaussian components). Thus, the inclusion of the Gaussian features does not introduce much new information. Nevertheless, we include these features in our final model, as they are easily available and offer marginal improvement.

Nearest neighbour (NN) features: Adding the NN information has the greatest impact on the model performance, improving the classification by more than 4 per cent. Even though there is not a strong linear correlation with the 'accept_lr' output in Fig. 2, the NN_dist, NN_lr_dist and log_NN_lr_tlv, and the NN_45 parameters provide valuable additional information for the classification, as does the flux ratio of the NN source relative to the source under consideration.

Self-organizing Map (SOM) features: Experiments using solely the baseline and the SOM features improves the classifier by about 2.5 per cent compared to the baseline only. Impressively, if using only the SOM as input features (not shown in Table 3), the model achieves a classification of almost 80 per cent, which demonstrates the power of the morphological representation for the classification. However, it also demonstrates that some essential information contained within the baseline features is not retrievable from the SOM alone.

The addition of the SOM features on top of all of the other different experiments improved the model accuracy by 0.3 per cent, to 94.7 per cent on the baseline model and 0.2 per cent on the

final model. This indicates that the information encoded in the SOM, through a visual representation of the source (compact versus extended emission, single versus blended versus multiple radio component source, etc.) does provide some additional information over the other features. However, this is limited, due to the correlations between the SOM and other features as seen in Fig. 2. Due to the relatively small improvement, and because the SOM features come from an external source, we have decided to exclude the SOM from our final model.

Deconvolved features: We also investigated using the deconvolved (DC) major and minor axis instead of the measured values, and we found the same results. We ran the model using the DC and non-DC major and minor axis for both the PYBDSF sources and the Gaussians and the differences were negligible. Baseline experiments replacing the measured sizes by the deconvolved sizes of the sources pointed to a small improvement on the classifier, but well within the range of the variance of the model. In our final model, we opted to use the non-deconvolved sizes as these are potentially more robust against inaccurately measured beam sizes; however, this choice is arbitrary and is not expected to have a significant effect on the classifier for LoTSS DR1.

4.4 Model optimization

4.4.1 Selection of model and model hyperparameters

After feature selection, we performed further experiments using TPOT to optimize the model hyperparameters using a single data set. The hyperparameters are used to adjust the learning process (e.g. learning rate) and the model specifications (e.g. number of estimators, i.e. trees, on a tree-based model). We ran TPOT for three generations with a population size of 5, and a cross-validation (CV) of 10 and the sets of features from Table 2 excluding the SOM. The range of values we defined for TPOT to perform the search, and the optimized set of model hyperparameters for the GBC model finally selected, can be seen in Table 4.

Since there is some discussion in literature about boosting methods overfitting under certain circumstances (see Appendix A for references) we give special attention to check that the model we use does not overfit. Therefore, and for verification purposes, we tested different possible combinations of hyperparameters. Increasing the learning rate and increasing the number of estimators both make the model increase its accuracy; for example, for 1000 estimators the accuracy is able to reach values higher than 99 per cent on the training set and 94.8 per cent on the test set. However, a training set performance close to 100 per cent is a strong indication that the model is overfitting, especially with the significant difference in performance between the training and test sets (although the high accuracy on the test set shows the model is still able to generalize). TPOT favours the use of 500 estimators, which offers good results and minimizes the risk of overfitting. Our optimized GBC model achieves an internal TPOT CV score of 94.6 per cent and an average accuracy of 94.6 per cent on the test and 95.9 per cent on the training set.[4] These are also the values obtained for the model trained and optimized using a single data set which we further use to present the results in the next section. This is within 0.2 per cent of the

---

[4]Note that this accuracy cannot be fairly compared against the accuracy of the decision tree of W19 quoted in Table 1, since the latter is for a very unbalanced data set and is optimized for performance on the majority population of class 1 sources. We compare the ML performance against that of the W19 decision tree in Section 6.3.

performance with 1000 estimators, but by using a smaller number of estimators we reduce the complexity of the model as well as training time, and can have higher confidence that the model is not overfitting.

We also investigated an XGBoost model, as this was also favoured by TPOT. The best XGBoost model achieves an internal TPOT CV score of 94.6 per cent and an average accuracy of 94.7 per cent on the test and 96.6 per cent on the training set. This is a marginally superior performance on the test set to the GBC model, but within the scatter of different data set selections, and also has a higher difference between test and training set performance. Given this, we opt to retain the less complex GBC model for our final analysis.

Overall, as can be seen from Tables 3 and 4, the hyperparameters and performance for the optimized GBC model are not dissimilar from those of the baseline model.

4.4.2 Training with re-sampling

To test whether under- or oversampling is a better approach, we applied the optimized classifier on the re-sampled data (see Section 3.3). Not surprisingly, we found that training the model with more examples of class 0 (even if they are synthetic) results in a higher precision for this class. Additionally, when compared to training without resampling, it results in a more proportional model performance across the two classes. This model reduces the number of sources that need to be visually inspected (the value of recall for class 1 increases), but this comes at the cost of accepting more sources for LR than should be (precision on class 0 decreases). This increase in the number of false positives is not in alignment with our science goals, as these sources will all remain incorrectly classified in the final analysis. The overall performance for the re-sampled data sets decreases by 0.7 per cent in accuracy on the test set, compared to the undersampling method, while the accuracy for the training set increases by 1.16 per cent. This difference is particularly evident for sources in class 1, for which the model got too specialized: it achieves 98.41 per cent precision on the training set, which does not allow it to generalize well on the test set for this class. This is the most probable reason why the model accepts too many false positive sources as suitable for LR analysis. We conclude that training with re-sampling leads to overfitting the classifier, and hence we opt for training the final classifier with undersampling instead.

5 MODEL PERFORMANCE AND INTERPRETATION

5.1 Final model performance

The model that we adopt in the rest of the paper is the GBC model with the optimized hyperparameters described in Table 4 and the 18 features (which exclude the SOM features) from Table 2, trained and tested on a balanced data set created with undersampling. In Table 5, we present the suite of metrics defined in Section 4.1 to assess the performance a binary classifier, in order to illustrate the overall performance of the model, as well as the performance on the different classes. The results presented here are run on an independent test set and adopt a standard cut-off probability of 50 per cent between the two classes.

Our best model achieves an overall accuracy of 94.6 per cent on the test set, and just 1.3 per cent higher on the training set. The model can be seen to favour precision for class 1 (sources that can be cross-matched using LR) and recall for class 0 (sources that require visual inspection). These are the values we intend to optimize: while we

Table 4. GBC model hyperparameters: baseline, tuning values, and finally adopted optimized hyperparameters obtained by TPOT optimization. The learning rate controls how quickly the loss is corrected at each iteration; no. of estimators corresponds to the number of sequential trees create by the model; max depth represents the maximum tree extension; subsample is the proportion of data used in each tree; min samples split corresponds to the minimum number of examples necessary to split a tree into different branches; min samples leaf is the minimum number of examples required in a terminal leaf; and max features is the maximum number of features to take into consideration while searching for the optimal split.

Hyperparameters	Baseline GBC	Search values	Optimized GBC
Learning rate	0.01	0.001, 0.01, 0.05, 0.1, 0.5, 1	0.01
No. of estimators	100	100, 250, 500, 1000	500
Max depth	10	Range (1, 11, steps = 1)	8
Subsample	0.75	Range (0.05, 1.01, steps = 0.05)	0.15
Min samples split	6	Range (2, 21, steps = 1)	12
Min samples leaf	10	Range (1, 21, steps = 1)	5
Max features	0.35	Range (0.05, 1.01, steps = 0.05)	0.6

Table 5. Performance on the test and training sets: the results give the overall accuracy, and the F1-score, precision and recall for each class (where 1 = suitable for LR; 0 = requires LGZ), for a decision tree threshold of 0.50 or 50 per cent. The results quoted are for a single undersampled balanced data set.

	Test set	Training set
Accuracy	0.946	0.959
F1-score 1	0.945	0.958
F1-score 0	0.947	0.960
Precision 1	0.963	0.975
Precision 0	0.930	0.944
Recall 1	0.928	0.942
Recall 0	0.964	0.976

want to avoid a high number of visual inspections it is more important to reduce the number of sources accepted as class 1 when they do not belong to that class. From the total number of sources accepted as being suitable for LR analysis, 96.3 per cent are actually from that class; similarly, 96.4 per cent of the sources that need to be visually inspected are sent to visual inspection. While this means that there is already a low percentage of sources wrongly predicted to be class 1, in practice the number that will end up being mis-classified is even smaller as some of these sources will be corrected during the LGZ process (see corrections applied in Section 6.2). The model yields slightly lower values of precision for class 0 and recall for class 1, meaning that the model sends more sources to visual inspection than needed. Overall, the classification predictions send around 7 per cent more sources (in a balanced data set) to visual inspection than needed to be inspected; this percentage will be significantly higher when applying the model to a highly imbalanced data set with many more sources in class 1.

For illustration, we show in Fig. 3 the ROC curve of the model. This shows the true positive rate (TPR) against the false positive rate (FPR) plotted for different thresholds. The plot illustrates the performance of the model on detecting a source that can be processed by LR (i.e. a positive test) as we achieve values close to a TPR of 1 and FPR of 0; and an AUC (Area Under the Curve) for the test set of 0.98 (where an AUC of unity would correspond to the perfect classifier). Instead of using the default 0.50 threshold for balanced data sets, we can further explore a more suitable cut-off threshold closer to the top left corner of the curve, which is particularly important when dealing with imbalance data sets. We therefore explore the effect of varying the cut-off threshold in Section 6.1 in order to optimize the trade-off between the number of sources wrongly accepted as suitable for LR and the number of sources sent to visual inspection.
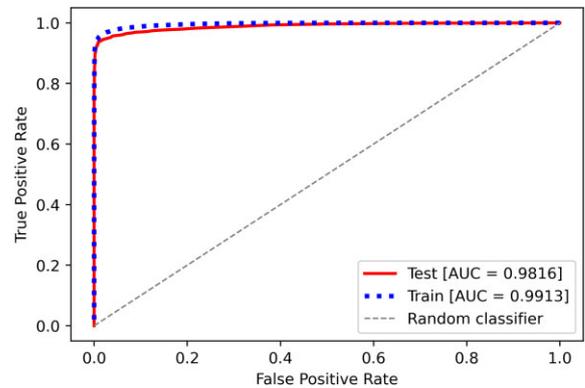
Figure 3. Receiver Operating Characteristic (ROC) curve of the optimized model for a training and test balanced data set, showing that this has an Area Under the Curve (AUC) close to unity, where 1 would be the value for a perfect classifier classifier. The true positive rate [TP/(TP + FN)] is the rate at which a source suitable to cross-match with LR is correctly identified as such out of all the ones that can be done using this method, while the false positive rate [FP/(TN + FP)] is the proportion of sources that are incorrectly predicted to be suitable to LR out of all the ones that require visual inspection.

5.2 Feature importance in the model

To interpret the importance of the different features for the classification, we use SHAP (SHapley Additive exPlanations; Lundberg & Lee 2017), through the use of a PYTHON package explicitly applied to tree-based ML models (Lundberg et al. 2020). The method measures the impact of different features on the model classification by averaging the contribution of a particular feature compared to when that feature is absent for the prediction.

The SHAP values are computed individually for each source in the training set, and the left-hand panel of Fig. 4 shows how the values of each feature contribute to the classification. SHAP values are given in units of log of odds, with positive SHAP values implying that the value of the feature favours class 1 sources and negative SHAP values implying that the feature value favours class 0 sources. The colour-coding on the plot indicates the value of the input feature compared to the range of values of that feature for all sources. Thus, for example, higher values of the major axis are associated with sources that have highly negative SHAP values (class 0), while lower major axis values favour class 1.

The right-hand panel of Fig. 4 shows the global contribution of the different features to the model predictions, in descending order. These correspond to the mean of the absolute SHAP values per feature across all the data on the training set. The features at
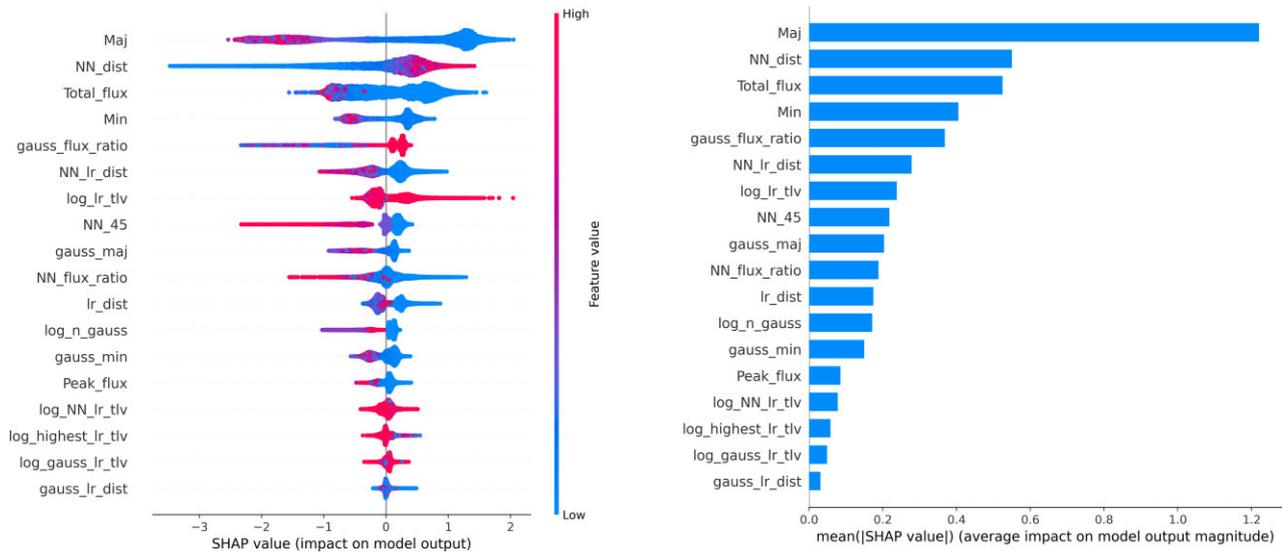
Figure 4. *Left.* SHAP values for each feature and for each source within the training set. The colour coding indicates the value of the feature for that source compared to the range of values for that feature across all sources, as indicated by the colour bar, and the thickness of the plot indicates the density of sources at a given SHAP value. Larger absolute SHAP values indicate higher impact in the prediction. *Right:* SHAP feature importance computed as the mean of the absolute SHAP values. These are ordered such that the features with the highest predictive power are at the top.
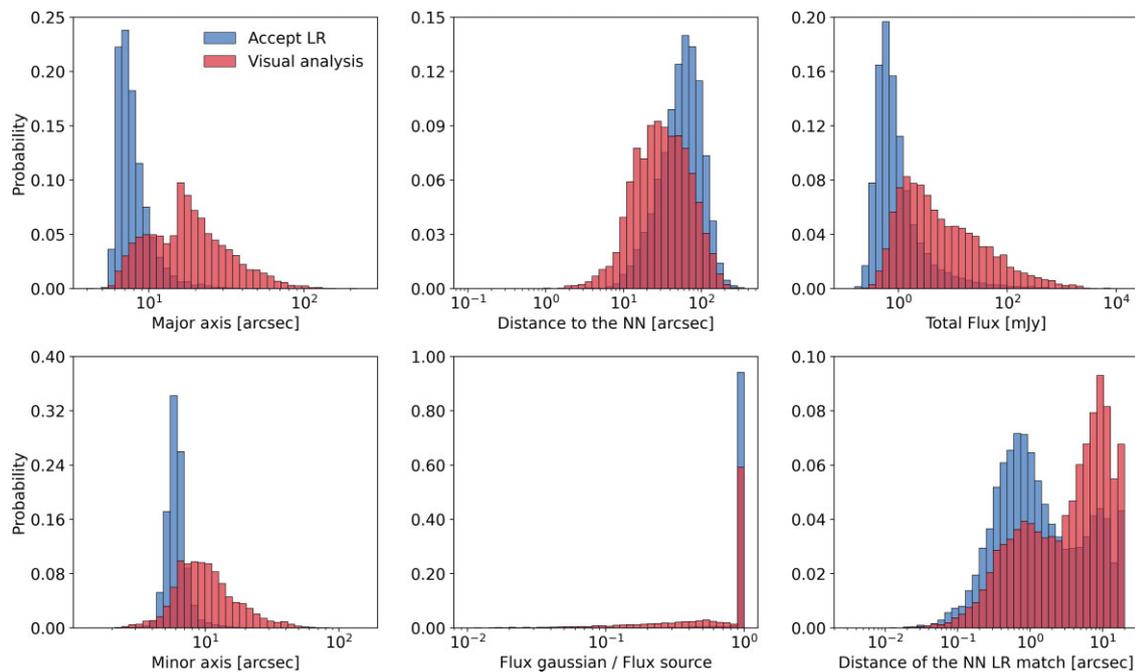
Figure 5. Probability distributions of the most distinctive features, as identified by the SHAP analysis. In each case, blue corresponds to sources that are suitable for statistical match by LR (class 1) and red represents sources that require visual analysis (class 0). For all of these features, a systematic offset in feature values between class 1 and class 0 sources is apparent, in the direction that would be expected from the radio source properties (see the text for details).

the top of the plot are those with the highest predictive power: these are the major axis of the source followed by the distance to the source's NN. The features towards the bottom of the plot provide the least predictive power of those considered in the model.

Fig. 5 shows the distribution of feature values for the six features picked out to have the highest predictive power. Specifically, it shows histograms of the distributions of feature values for the two classes

of objects (class 0, class 1), each normalized to the total number of sources of that class. In each case a distinction between the two classes is apparent, and is in the direction which would be expected. Smaller sources (both major axis in the upper left and minor axis in the lower left) have a higher probability of having a correct cross-match by statistical means, as opposed to more extended sources, which are more likely to be resolved and possibly complex. Brighter sources (upper right) are also more likely to require visual analysis,

due to the predominance of more extended AGNs at higher flux densities compared to more compact AGNs and SFGs at fainter flux densities (see discussion in W19). Sources for which the Gaussian component contains only a fraction of the total flux density and hence other Gaussian components must also be present, indicating an extended source, are also more likely to need visual analysis (lower middle panel), as compared to compact sources with all of their flux in a single Gaussian. Finally, those sources with a close near neighbour (upper middle panel), especially when that near neighbour does not have a close LR match (lower right-hand panel) are also indicative of multicomponent sources which require visual analysis.

6 APPLICATION TO FULL LOTSS DATA SETS

In this section, we apply our model to the full LoTSS DR1 data set, and also make a preliminary evaluation of its performance on a subset of LoTSS DR2. When applying the trained ML model to the full LoTSS DR1 data set there are two main points that require consideration. Firstly, unlike the data set used to train and test the model, LoTSS DR1 is highly imbalanced. In Section 6.1, we investigate varying the cut-off probability to select a value that is more suitable for this class distribution problem rather than using the default 0.50 threshold. We also define the parameters by which we will assess the performance of the model in order to select the appropriate threshold. Secondly, it should be noted that some sources wrongly classified by the algorithm as being suitable for LR (false positives) may be recovered (corrected) if additional components of the same (multicomponent) source are sent to LGZ. This may particularly be the case for the cores of extended radio sources: the core itself is compact and aligns with the optical host galaxy so may have a higher LR match, pushing towards a class 1 prediction, but the surrounding extended lobes are far more likely to be predicted to need LGZ. We examine and correct for this issue in Section 6.2.

To investigate the overall performance of the classifier in different regions of parameter space, we compare our results with those of the W19 decision tree in Section 6.3 and investigate the success of the classifications for different source properties (as defined from the SOM) in Section 6.4. Finally, we conclude the evaluation of the model on LoTSS DR1 by examining the nature of those sources that deliver false positive outcomes (i.e. are sent to LR but should require LGZ) in Section 6.5. In Section 6.6, we further apply our model directly to LoTSS DR2 as a first step to evaluate how the model performs in a completely unseen data set.

6.1 Threshold moving for an imbalanced data set

The distribution of the two classes in the LoTSS DR1 data set is severely skewed towards class 1, and the default 0.50 threshold value does not represent an optimum cut-off probability between the two classes. The model prediction threshold reflects the proportion of examples in the two classes that were used to train the classifier; as a result, when the model is applied to the entire, imbalanced LoTSS DR1 data set, the majority of sources are classified as belonging to class 1, which is the most frequent class. Therefore, we tune the decision threshold, often known as 'threshold moving', which is a common approach used to optimize the predictions for imbalanced data sets (e.g. Collell et al. 2018). The effect of changing the threshold is demonstrated on the ROC curve in Fig. 6.

Instead of evaluating the whole performance of the model solely with the typical metrics (accuracy, precision, etc.), we seek in
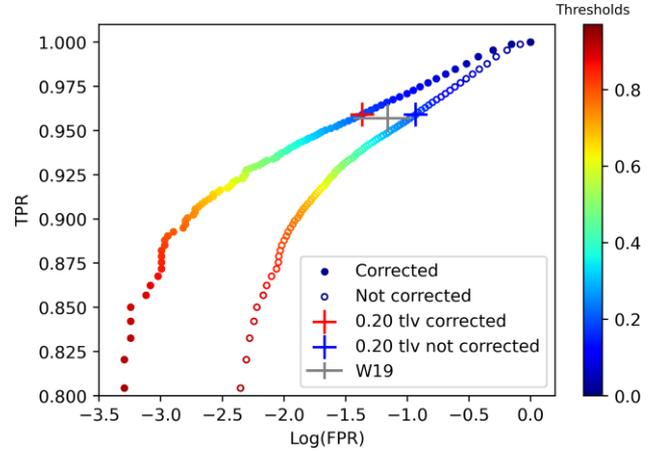
Figure 6. Zoom in of the ROC for the full LoTSS DR1 data set, showing the different threshold levels. Note that to better visualize the results, the *x*-axis is on a log scale, and only the upper values of the *y*-axis are shown (cf. Fig. 3). The open (lower) symbols represent the raw results from the model fitting, and the filled (upper) symbols demonstrate the improvement which results from the corrections for recovered false positives (see Section 6.2). The threshold value adopted is indicated by the red and blue crosses which corresponds to a false positive rate (FPR) of 11.7 per cent for not corrected values and 4.3 per cent for the corrected values, and a true positive rate (TPR) of 95.8 per cent. The grey point indicates the results of the W19 decision tree using the raw values from Table 1, with the horizontal error bar representing the potential spread from uncorrected to corrected values if the false positive recovery rate for W19 would be the same as for the classifier.
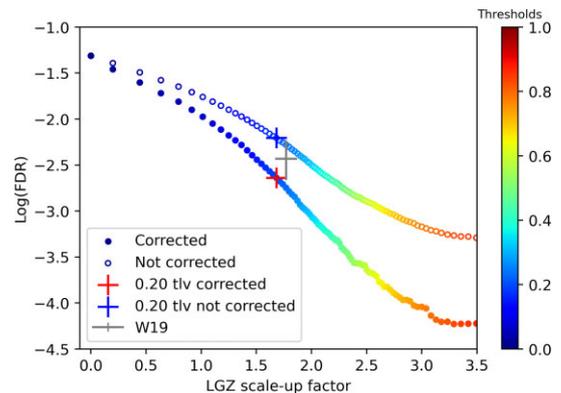
Figure 7. A comparison of the performance metrics adopted for analysis of our model for different values of the cut-off threshold between the two classes. The *y*-axis is the false discovery rate (FDR), which measures the fraction of sources accepted for LR that were incorrectly selected. The *x*-axis is the LOFAR Galaxy Zoo (LGZ) scale-up factor, which measures the total number of sources that the model selects for visual inspection divided by the number that we should really inspect. This is the combination of parameters that we aim to minimize. Symbols are as in Fig. 6.

particular to minimize the number of sources wrongly predicted as suitable to process with LR while keeping the number of sources sent to visual inspection low. These two requirements can be captured by (i) the false discovery rate, FDR = FP/(FP+TP), which quantifies the fraction of sources sent to LR which are incorrect; and (ii) a parameter we refer to as the LOFAR Galaxy Zoo scale-up factor, given by (TN+FN)/(TN + FP), which expresses the factor by which the number of sources selected for visual analysis in LGZ is higher than the number actually required to be sent (cf. Table 1). In Fig. 7, we show how the comparison between these two metrics changes

as we change the cut-off threshold (open symbols, colour-coded by threshold level).

Although Fig. 7 does not dictate which threshold value to use, the practical requirement to keep the LGZ scale-up factor to below ∼2 pushes for a lower value of the threshold than the nominal 0.50 value, while the threshold values should not be so low to allow a false discovery rate above about 1 per cent. In practice, we adopt a threshold value based on comparison with the W19 decision tree results. After correction for recovered components (Section 6.2), the classifier outperforms the W19 decision tree in *both* false discovery rate and LGZ scale-up factor for thresholds in the range of 0.18–0.25. We select a threshold level of 0.20, as a round number towards the centre of this range. This threshold value corresponds to an LGZ scale-up factor of 1.68 and a false discovery rate of 0.006 for the raw model outputs.

6.2 Corrections adopted

Corrections were determined to account for the multicomponent sources wrongly classified as suitable to cross-match by LR (FP) that would subsequently be recovered by LGZ. Specifically, we analyse the prediction for each PYBDSF component that makes up a multicomponent radio source and if at least one of the components is sent to visual analysis by the model, the source is removed from the FP group. The sources recovered in this way are discussed in Section 6.5; in many cases these are the cores of radio sources (which on their own resemble a compact radio source) for which the more extended lobes are sent to LGZ.

We calculated the number of recovered sources for each different threshold value. The filled symbols on Fig. 6 demonstrate the improvement that these corrections make to the ROC curve analysis, and those on Fig. 7 demonstrate the impact on our metric plot (FDR versus LGZ scale-up factor) after applying these corrections to the FDR. Except at the very lowest thresholds, the improvement that the corrections make to the FDR is very significant; the fraction of recovered sources increases for higher threshold values, leading to very low FDRs at high thresholds, but with the cost of a higher number of visual inspections. For our adopted threshold of 0.20, 63 per cent of the FP sources are recovered, resulting in a corrected false discovery rate of 0.002. This is also shown in the confusion matrix for that cut-off level, presented in Fig. 8. In the analysis that follows, these corrections are applied unless stated otherwise.

6.3 Performance relative to W19 decision tree

In this section, we compare the performance of our model against that of the W19 decision tree for the same data set. First, in Fig. 9 we present the confusion matrix for the final model, split by the three main decision tree outcomes of W19: suitable for LR, send to LGZ, or requires prefiltering.

It can be seen that the performance of the model on the 'LR group' is excellent with nearly 99 per cent of the sources being deemed by the classifier to be suitable for LR. Furthermore, of the 1096 sources that were incorrectly selected by the W19 decision tree as 'LR' but which were subsequently re-classified during the LGZ process (e.g. by being examined in parallel with another LGZ source) the classifier correctly sends the majority (over 600 sources) to LGZ, and of the rest all but 75 are recovered by having an alternate component of the source sent to LGZ. The classifier does send 3710 sources to LGZ that W19 sent directly to LR and which have a label of being suitable for LR. However, it is important to note that none of these sources has been visually examined to confirm that the W19 label is

Figure 8. Confusion matrix for all the sources in LoTSS DR1 using the optimized model and a threshold value of 0.20. The confusion matrix shows how examples belonging to each class are assigned correctly and incorrectly to the 2 possible classes. A perfect classifier would produce a confusion matrix filled diagonally with only TP (top left) and TN (bottom right) values, where the FP (bottom left) and FN (top right) would have values of zero, as defined in Section 4.1. The background colours illustrate the proportion of sources in the matrix (given also by the percentage values in brackets) with darker colours representing a greater number of sources. The numbers presented correspond to the corrected values (see Section 6.2).

Figure 9. The model confusion matrix (for a threshold level of 0.20), split by the three main decision tree outcomes of W19: LR, LGZ, and prefiltering. The FP values quoted are after corrections, with the numbers in brackets showing the values prior to corrections. As may be expected, the highest classification accuracy is for the LR sources, and the lowest accuracy is for the population of sources with intermediate parameter values deemed by W19 to require prefiltering.

correct: where the W19 decision tree provided a LR classification, that was simply adopted by W19 (unless LGZ examination of a different PYBDSF component overrode that). There may, therefore, be (many) examples amongst these 3710 sources that, like the 1096 sources discussed above, would have been re-labelled had they been visually examined and for which the classifier is therefore correct. We explore this further below, and in Section 6.5.

For the sources selected by W19 to go directly to the LGZ process, the classifier provides an overall accuracy of 73.5 per cent, with the lower value mostly driven by nearly 2000 sources being sent to LGZ despite being suitable for LR. Nevertheless, amongst the 3000 sources in the W19 LGZ sub-sample that were found (after visual examination) to be suitable for LR, the classifier is able to send over one-third of these directly to LR, thus reducing the LGZ scale-up factor.

The classifier performance is poorest on the sources sent by W19 for prefiltering. This is not surprising, since these are generally sources with intermediate parameter values, between the compact LR sources and the extended LGZ examples. Again the classifier is able to send around one-third of the true LR sources directly to LR, but still assigns nearly 7000 sources incorrectly to the LGZ class, providing the largest contribution to the LGZ scale-up factor.

The prefiltering category also contains the largest number of false positives (339 after corrections).

We also compare the performance of our model in our metrics of FDR versus LGZ scale-up factor, against those of the W19 decision tree. The LGZ scale-up factor of the W19 decision tree is easily calculated from the numbers in Table 1 and corresponds to a value of 1.77, while the 1096 PYBDSF sources identified as false positives implies a W19 FDR of 0.004. The FDR and LGZ scale-up factor thus determined for the W19 decision tree are shown in Fig. 7. Compared to these, the ML model with a threshold of 0.20 achieves both a lower false discovery rate and a lower LGZ scale-up factor. Furthermore, as discussed above, the value of 0.004 represents a lower limit to the FDR of W19 because the objects selected as being suitable for LR analysis were, in general, not visually examined, and thus false positives were not identified. We can estimate the total number of FPs by assuming that the fraction of sources rescued in this way is broadly the same as the ML model (the 'corrections' calculated in Section 6.2). This value depends weakly on the threshold value adopted (for similar theshold values). For thresholds around 0.20 we calculated above that 63 per cent of the sources are rescued. If 1096 sources correspond to 63 per cent then we can estimate the total number of false positives in the W19 decision tree will be approximately 1730 sources.[5] This would correspond to a higher FDR of 0.006.

To gain a better understanding of which types of sources the model performs well on, and on which it performs badly, in Fig. 10 we reproduce a simplified version of the W19 decision tree, and examine the model confusion matrix at different locations of the decision tree. In the W19 decision tree, sources are first classified as 'Large' (major axis larger than 15 arcsec) or 'Small' (under 15 arcsec), with a small number being associated with nearby large optical galaxies (above 60 arcsec of radius in the 2 µm all sky-survey extended source catalogue, Jarrett et al. 2000). The 'Large' sources are then separated by W19 in flux density (above or below 10 mJy total flux densities), where W19 send the brighter large sources all to LGZ and the fainter large sources all to prefiltering. The performance of the classifier on these two sub-categories is comparable to that on the general 'LGZ' and 'prefiltering' classes discussed above: these large sources produce more than half of the false negatives that lead to the above-unity LGZ scale-up factor. We examine the nature of these extended sources in more detail in Section 6.4.

For the 'Small' sources, W19 next examined whether the source is relatively 'Isolated' (no NN within 45 arcsec) or not. Isolated sources were examined to see if they were composed of single or multiple Gaussians. 'Single Gaussians' were sent by W19 to LR and it can be seen that the classifier achieves a remarkable accuracy of 99.98 per cent on these sources, which comprise nearly 58 per cent of the full sample. This subset of sources probably explains why the addition of LR features was found to only offer a small improvement in model performance in Section 4: these small, isolated, single Gaussian sources can almost entirely be sent for LR analysis based on their radio properties alone, and the LR provides no extra information. This does imply, however, that the addition of the LR features has much more impact in the other branches of the decision tree than the raw statistics of Table 3 suggest – indeed, for the 'Large' and for multiple Gaussian sources, the addition of the LR information provides around 5 per cent increase in accuracy compared to the baseline.

For sources with multiple Gaussians, the W19 decision tree was complicated, but can be simplified to consider those sources for which the PYBDSF source has an LR match above the LR threshold, those for which the PYBDSF source does not but one of the Gaussian components does, and those for which neither source nor any of the Gaussians has an LR match above the threshold. The classifier performs fairly well (accuracy ≈90 per cent) on the first and third of these classes, but less well (accuracy ≈60 per cent) on the Gaussian LR matches, which are only 0.5 per cent of the complete sample but contribute nearly 30 per cent of the (corrected) false positives in the whole sample. This implies that it may be possible to improve the classifier through better consideration of which Gaussian features to include (e.g. a second Gaussian to assist in identifying blended sources; see Section 6.5) but such an investigation is beyond the scope of this paper.

If the NN is within 45 arcsec ('Not Isolated') then the number of other PYBDSF sources within 45 arcsec is counted: sources with at least 4 others within that distance ('Clustered') were sent by W19 to LGZ, and the classifier similarly sent most of these to LGZ. The 'Unclustered' sources were then examined as for the isolated sources, into single or multiple Gaussian components and looking at the LR matches for the latter. In this case, the performance on the single Gaussians (30 per cent of the overall sample) is less strong than for the isolated single-Gaussian sources, both in terms of false positives and LGZ scale-up factor, but still achieves 97.8 per cent accuracy. This illustrates that the near-neighbour components are impacting the classifier. Similarly, the performance on the multiple Gaussians is poorer than for the isolated sources (overall 71.6 per cent accuracy), in the sense of having a higher LGZ scale-up factor (more false negatives), albeit with a lower false positive rate.

6.4 Performance as a function of source properties

We also investigate the model performance as a function of source morphology and different source characteristics. For this, we consider the SOM, and separate the locations of the sources within this into six different morphological categories following Mostert et al. (2021). These six categories (described in more detail below) are 'extended singles'; 'compact doubles'; 'core-dominated doubles'; 'large diffuse lobes'; 'extended doubles'; and 'single lobe/near neighbour'. Added on to these are the sources classified by Shimwell et al. (2019) as 'unresolved', which were not considered on the SOM.

Considering first the unresolved sources, the top panel of Fig. 11 shows the confusion matrix for these sources. Perhaps surprisingly, more than 9000 of these sources have 'LGZ' labels, and the classifier also sends a further 7556 sources to LGZ, corresponding to a significant proportion of the LGZ scale-up factor. To investigate the reason for this, in the bottom panel of Fig. 11 we show how the different classifier outcomes vary with the size of the source major axis. Despite these sources being identified to be 'unresolved' by Shimwell et al. (2019), the major axis sizes can extend to more than 20 arcsec; this is because the Shimwell et al. classification adopts a signal-to-noise dependent size envelope for separating unresolved from extended sources based on their integrated flux density to peak brightness ratios, and so at low signal-to-noise ratio where there is substantial scatter in the flux ratio it is possible to have quite large 'unresolved' sources. It is not surprising that LR is not appropriate for these, as the radio position is poorly defined. Fig. 11 indeed shows that both the true negative and false negative percentages increase

---

[5]Note that these extra false positives will be mis-labelled in the input data set, most likely comprising some of the false negatives in the LR subset of Fig. 9 as discussed above, and thus the performance of the ML model may therefore be fractionally higher than quoted.
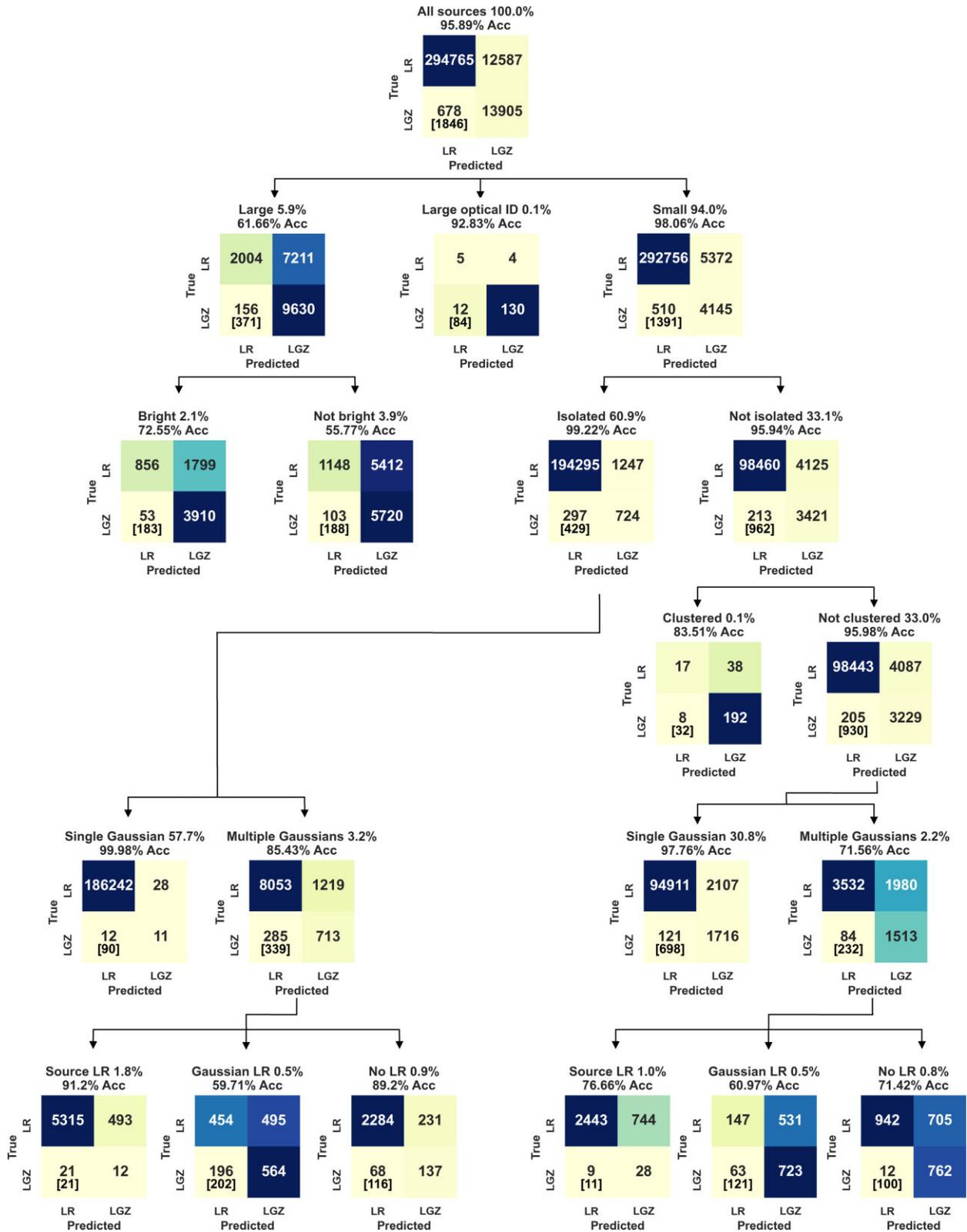
Figure 10. A simplified version of the W19 decision tree, showing the performance of the classifier (in the form of the confusion matrix) at different locations on the decision tree.

with increasing major axis size, each reaching ≈10 per cent at a major axis size of 15 arcsec.

Fig. 11 also illustrates that beyond 15 arcsec in size, the true negative fractions suddenly jump to 40 per cent. This is due to a feature of the training sample: all sources larger than 15 arcsec in size were visually examined by W19, and thus we expect them to be all correctly labelled, but at smaller sizes those sources for which the W19 decision tree predicted 'LR' were not visually examined;
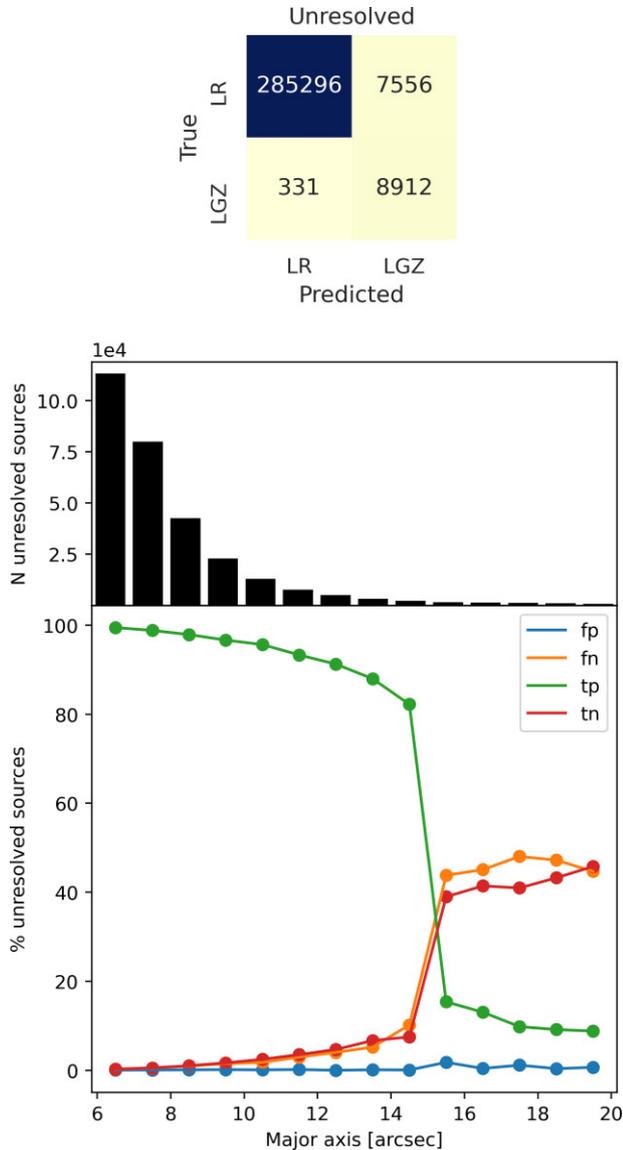
Figure 11. *Top:* The confusion matrix for the sources classified as 'unresolved' by Shimwell et al. (2019). *Bottom:* The distribution of major axis sizes of these unresolved sources, and the variation of the different classifier outcomes as a function of the major axis size. The predicted LGZ outcomes are primarily associated with those sources with larger major axis sizes. The jump at a major axis of 15 arcsec is associated with the training sample characteristics (see the text for more details). Note that the ≈6000 sources larger than 20 arcsec are not included on this plot.

as discussed in Section 6.1 some of these may be wrongly labelled. This suggests that the LGZ-fraction at sizes just below 15 arcsec may be somewhat higher than the labels suggest.

Note that although the jump appears pronounced, only a small fraction of the 'unresolved' sources have these large sizes, as can be seen in the histogram in Fig. 11. Specifically, 10 516 (3.5 per cent of the unresolved sources) have sizes between 12 and 15 arcsec, and 12 281 (4.1 per cent of the unresolved sources) are larger than 15 arcsec; these small numbers will not have a large effect on the classification outcomes. It is interesting to note that the false negative fraction shows a large jump at 15 arcsec size as well: due to the issues of the training sample, the classifier is learning that 15 arcsec is a critical size above which sources are more likely to require LGZ.

This suggests that with an improved training sample that did not contain this issue, the performance of the classifier could potentially be improved even further than that presented here.

Considering the extended sources, Fig. 12 displays the confusion matrices for each of the six categories of extended sources, along with three example thumbnails of each category, drawn from the SOM representative sources. For both the 'extended single' sources [those fitted by PYBDSF as a single Gaussian, but classified by Shimwell et al. (2019) as resolved] and the 'compact doubles' (typically two Gaussian components in the PYBDSF source, but small angular size), the performance of the classifier is similar: over 75 per cent of both categories are classifiable by LR, and the classifier performs reasonably well (accuracy ≈77 per cent) but sends about twice as many sources to LGZ as required. For the 'core-dominated doubles', which show a bright central component but extended emission, the classifier sends about 70 per cent of the sources to LGZ, presumably due to the extended emission, although in reality 60 per cent would be classifiable by LR due to the central component (the other 40 per cent are not, as most are split into multiple PYBDSF sources). Similarly for the more 'extended doubles', the classifier sends the majority to LGZ even though around half are symmetric enough that LR could be used. For the sources called 'large diffuse lobes' by Mostert et al. (2021) (which typically comprise either one or two extended lobes), the classifier achieves an accuracy of over 75 per cent by correctly sending the majority of the sources to LGZ, and again erring on the side of caution with an above-unity LGZ scale-up factor but few false positives. Finally Mostert et al. (2021) define a category of 'single lobes', but we re-define this as 'single lobe/near neighbour' because investigation reveals that while some of these are indeed one lobe of a double, two-thirds are single-component sources (classifiable by LR) for which there just happens to be a near neighbour. The classifier achieves a good accuracy (69 per cent) on these sources but again sends nearly twice as many as necessary to LGZ in order to minimize the number of false positives. Overall, it is clear that the performance on the extended sources is poorer than that on the 'unresolved sources', but still relatively strong: the total LGZ scale-up factor for these extended sources is only ≈1.8, not much higher than that of the unresolved sources, and the extended sources provide less than 300 false positives after corrections, with a false discovery rate below 4 per cent.

6.5 Examination of false positives

Finally, in Fig. 13 we provide a montage of examples of false positive sources: these are the most critical failures, because of the lack of visual inspection. The false positives can be categorized into four main categories, illustrated in the first four rows of the figure. The top row of the figure shows examples of multicomponent sources that get recovered (corrected) because one of the other PYBDSF components that makes up the source is sent to LGZ. These sources account for 63 per cent of all false positives. They are dominated by cases of the cores of radio sources for which the more extended lobes are sent to LGZ (e.g. in the first and second columns), but also include sources showing small extensions selected as a separate PYBDSF source (third column; in some cases these may be noise and in other cases they may be genuine extensions), and even a small number of radio source lobes rescued by other components of the source (fourth column).

The second row shows additional multicomponent sources, which are not recovered. In these cases, which amount to about 10 per cent of all false positives, it is essential to examine the sources with LGZ in order to properly associate the different PYBDSF sources into the same physical source and to identify the host galaxy, but the classifier
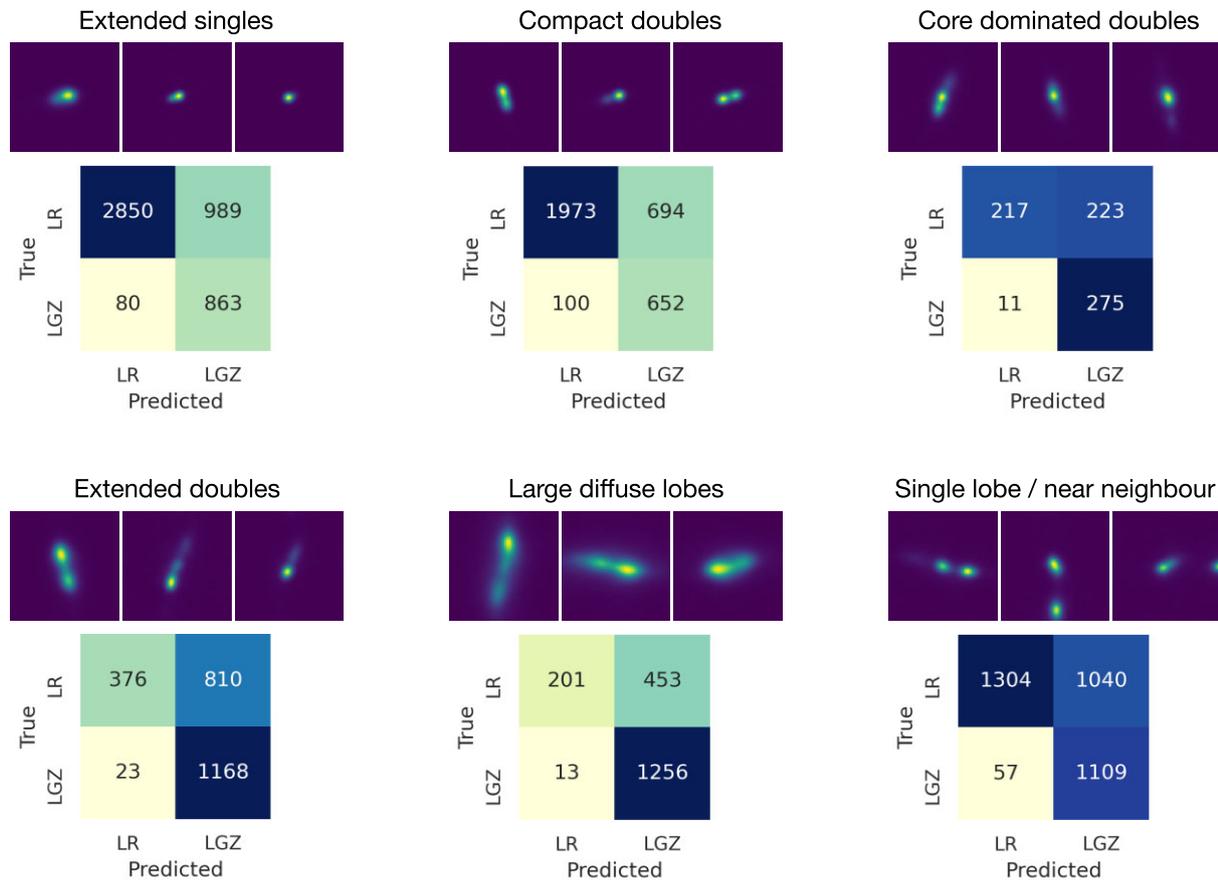
Figure 12. For six different broad morphological classes of extended sources defined by Mostert et al. (2021), the figure shows the confusion matrix, along with three example thumbnails drawn from the SOM representative sources.

predicts that all of the PYBDSF components are suitable for LR. These sources are typically relatively compact, two-component sources; sometimes, it is clear from the radio structure that these form a single source (e.g. the example in the first column), whereas in many cases this is only apparent when examining the optical and infrared data and noting the presence of a host galaxy between the two lobes (examples in second and fourth columns). Finally, a proportion of these multicomponent sources represent sources with weak extensions, some of which may be calibration artefacts (see example in the third column). In future work, it would be worth investigating whether the performance of the classifier on these multicomponent sources could be improved by including an additional feature related to the LR at the flux-weighted position between a source and its NN (corresponding to roughly where a host galaxy would be expected if the two sources form part of a double source).

The third row of Fig. 13 shows examples of blended sources (about 10 per cent of the false positives). These are cases where two physical sources have been merged into the same PYBDSF component, and these need to be examined and separated, but the classifier predicts that LR is appropriate. The optical images make the deblending requirement obvious, but it is understandable that this is difficult for the classifier to identify where the central component is substantially brighter in the radio and has a strong LR match. It is possible that if the LR of the second brightest Gaussian component was included as an additional feature of the classifier the performance on these objects could be improved.

The last two bottom rows represent sources that amount for about the remaining 20 per cent of the false positives. The fourth row presents examples of single sources (i.e. sources where PYBDSF has correctly identified the physical radio source) which the classifier predicts can be done by LR, but where the LR outcome disagrees with the final W19 identification outcome. There can be many different reasons for this. The first column shows a source where the LR selects the more northerly galaxy, closer to the radio centroid, but examination of the radio contours led the LGZ participants to conclude that the southern galaxy is the true host. The second and third columns both give cases where the galaxy close to the radio centroid has an LR value above the threshold level, but the LGZ participants concluded that this was not sufficiently robust to accept, and found no ID. The fourth column shows an example where the LR finds no identification, but in LGZ it was concluded that this was an asymmetric source with the galaxy on the right-hand component being the host. It should also be noted that the LGZ process is not perfect and some of these single components may be mis-labelled, and should actually be true positives rather than false positives. The fifth row of Fig. 13 demonstrates this: these are all examples of single sources deemed by the classifier to be suitable for LR (and to have an identification) but judged by W19 decision tree not to be. In all of these cases, the LR identification does appear to be robust. This suggests that these sources may be wrongly labelled by W19 and that the classifier is consequently performing even better than quoted.
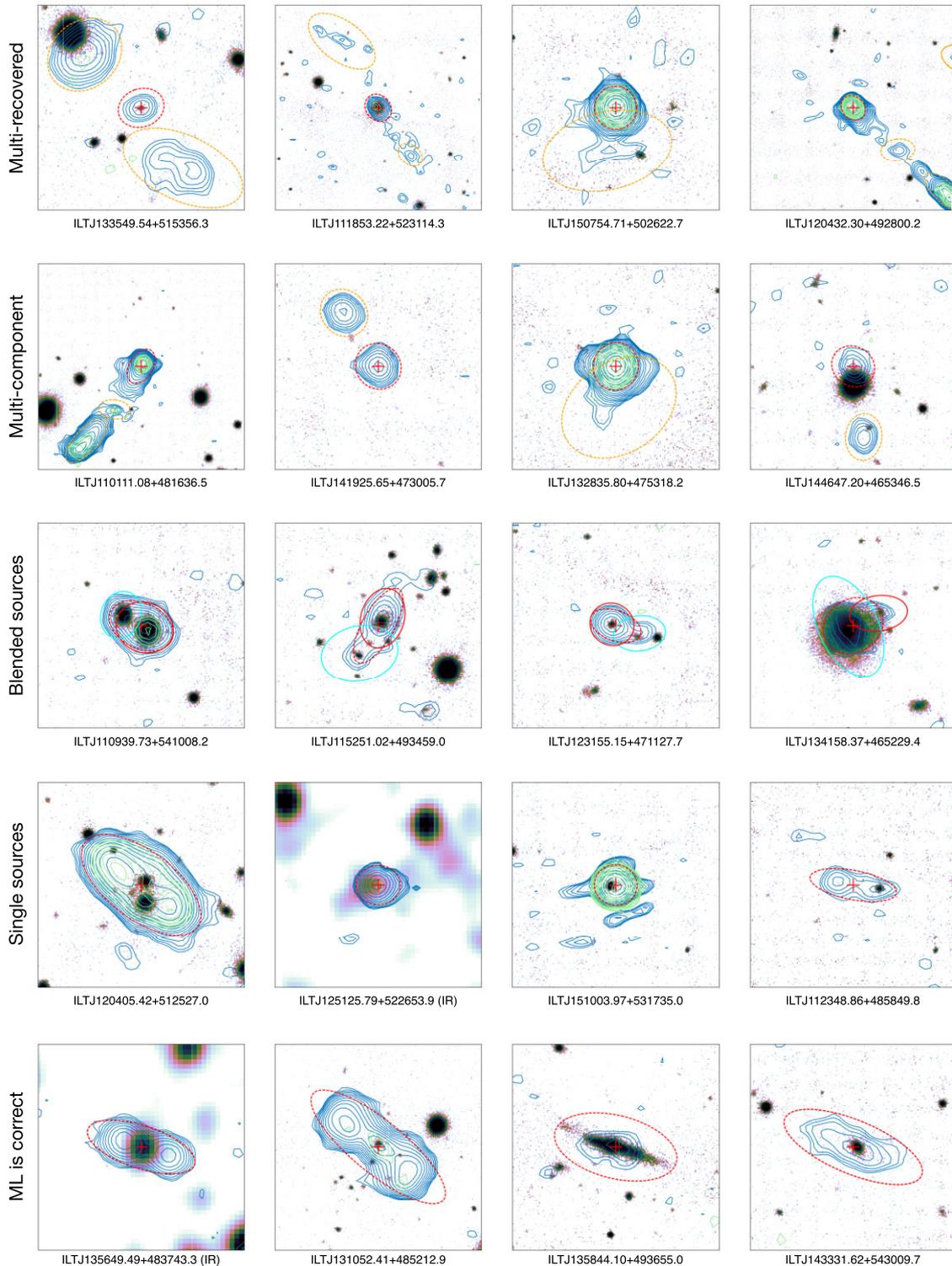
Figure 13. Examples of 'false positive' classifications, where the model predicts that a LR approach is suitable, but in reality LR gives the wrong outcome and examination by LGZ is required. In all figures, the red cross and red dashed ellipse indicate the PYBDSF source being examined, the dark blue contours indicate the LOFAR radio emission, and the green contours indicate the higher frequency 1.4 GHz radio emission from the FIRST survey. Yellow dashed ellipses indicate other PYBDSF sources that need to be combined to form a multicomponent source; solid Yellow ellipses indicated unrelated sources. For the blended sources (row 3) the blue and red solid ellipses indicate the deblended components. The top row shows examples of multicomponent sources where the false positive PYBDSF source is recovered (corrected) because a different component of the same source is sent to LGZ. The second row shows multicomponent sources where none of the components is sent to LGZ. The third row shows blended sources, where LGZ is required to separate the PYBDSF source into two physical sources. The fourth row shows single components (correctly associated) but for which the LR prediction does not match the final W19 ID outcome. For some of these, as indicated in the final row, the W19 label appears to be incorrect and the machine learning (ML) correct. See the text for further discussion.

6.6 Application to LoTSS DR2 subset

We have applied the model trained on LoTSS DR1 data directly to a subset of LoTSS DR2 in a small region in which LGZ source association and cross-matching process has already been completed for sources with total flux density higher than 4 mJy (these bright sources were examined first in LGZ in order to prepare targets for the WEAVE-LOFAR spectroscopic survey; Smith et al. 2016). Since LoTSS DR2 contains almost 14 times more sources LoTSS DR1, the application of ML methods is crucial to help managing these large data sets. We find that for the same threshold of 0.20, the classifier recommends that 11.8 per cent of LoTSS DR2 sources require visual analysis, compared to 8.2 per cent for LoTSS DR1. Investigation reveals that this difference is largely due to the source declinations: at declinations above about 50 deg the classifier sends 9.7 per cent of LoTSS DR2 sources for visual analysis, which is not much higher than the DR1 statistics, but that fraction increases as we move to lower declinations. This declination dependence is likely to be largely due to the lower sensitivity of the LoTSS survey at lower declinations (Shimwell et al. 2022), which raises the median image rms. This means that a larger fraction of the detected sources are at higher flux densities, where they are more likely to be multicomponent and require LGZ (see the upper right-hand panel of Fig. 5). Adjusting the prediction threshold to a higher value would therefore help to increase the correct classifications at lower declinations. An additional factor may be the increasing size of the LOFAR beam at lower declinations (the use of deconvolved sizes in our features might have mitigated this).

To further test the performance of the model on DR2, we examine and compare the output predictions within this DR2 region. From a sample of 59 122 sources brighter than 4 mJy, the classifier achieves an accuracy of 76 per cent; this compares with an accuracy of 82.7 per cent for sources brighter than 4 mJy in DR1, without taking into consideration recovered source components in both cases. The lower accuracy for the DR2 data is mostly associated with the classifier sending more sources to LGZ, as discussed above. Considering that the classifier has not been trained on DR2, but simply applied with its DR1-determined hyperparameters (and the DR1 cut-off threshold) directly on the DR2 data set, this shows that it has a strong ability to generalize to an unseen data set. The optical cross-matching for LoTSS DR2 (Hardcastle et al. in preparation) will differ from that of DR1 in the use of the DESI Legacy imaging surveys (Dey et al. 2019) instead of Pan-STARRS as the primary optical survey; however, our use of 'log_lr_tlv', that is, the logarithm of the ratio of the LR relative to the threshold value, as the primary LR feature should mitigate against these differences in the cross-match survey.

7 CONCLUSIONS AND FUTURE OUTLOOK

In order to get the most science out of the survey catalogues being produced by the new generation of radio interferometers, it is necessary to properly associate radio source components into physical sources, and then cross-match those sources with multiwavelength data. This enables us to identify the host galaxies and correctly derive the physical properties of the radio sources. To address the question of which sources are suitable for simple statistical cross-matching, and which ones require a more advanced (currently visual) approach, we trained a machine-learning (ML) classifier using LoTSS DR1 and applied it to different LoTSS releases. The main conclusions of our work are as follows:

(i) Our best model is a tree-based gradient boosting classifier, and achieves an accuracy of 95 per cent on a balanced data set.

This accuracy is maximized by appropriate choice of features in the model: inclusion of information on nearest neighbour (NN) radio sources, on the properties of any LR match, and on the composition of the radio source in terms of Gaussian components all improve the model.

(ii) The full LoTSS data set is highly imbalanced, with the majority ($\approx$95 per cent) of the sources being suitable for LR analysis. Adaption of the default 0.50 probability threshold for the classifier would result in far too many of these sources being predicted to require visual analysis. An optimized threshold of 0.20 restricts the LGZ sample to only 68 per cent larger than strictly required, while keeping the false discovery rate (i.e. the fraction of those sources accepted by LR that should have required LGZ) to only 0.2 per cent. With this threshold, the classifier outperforms the manually defined decision tree used for LoTSS DR1 by W19 in both the LGZ scale-up factor and the false discovery rate.

(iii) We have investigated the performance of the classifier on sources of different radio morphologies and with different source characteristics. As expected, performance is strongest for the most compact sources, achieving an accuracy of over 98 per cent on sources with a major axis size smaller than 15 arcsec (and over 99.9 per cent on the subset of these that have no near neighbours and can be well-modelled by a single Gaussian). The accuracy drops to just above 60 per cent for sources larger than 15 arcsec in size, primarily due to sending substantially more sources to LGZ than required.

The efficiency of the ML approach means that it can be applied to other radio surveys, and in particular to future data releases of the LoTSS survey, where the radio data are almost identical in nature to the DR1 sample analysed here (although there will be small differences, associated with improvements in the calibration scheme and a changing telescope beam as we move to lower declination; see Shimwell et al. 2022 for more details). Because of these results, the classifier outcomes derived for the full DR2 sample have been used, in conjunction with the W19 decision tree, to identify the LoTSS DR2 sources that are being sent to LGZ; Hardcastle et al. (in preparation) will provide more details.

In conclusion, the ML classifier that we have developed has been shown to have a high accuracy at identifying those sources for which a statistical cross-matching process is insufficient, and to outperform a manually-defined decision tree in both the false discovery rate, and in the number of sources that are predicted to require the time-consuming visual analysis step. The classifier has been demonstrated to be able to generalize to unseen data sets; it already has immediate application in the cross-matching of the LoTSS DR2 and can be easily applied to other radio surveys.

The classifier could potentially be further improved by the inclusion of additional features, for example, the LR of a second Gaussian component to assist in identifying blended sources, an LR at the flux-weighted position between a source and its NN to help identify multicomponent sources, or additional properties such as the local noise level or the source signal-to-noise ratio. However, even if the classifier were improved still further, the number of sources that require more than statistical cross-matching will still remain large, and visual analysis of all of these will become impractical as radio surveys continue to grow in size. The crucial next step is therefore to be able to replace visual analysis as the process to handle those sources. To this end, work to automatically associate multicomponent sources (e.g. Mostert et al., in preparation) and to improve automatic source cross-matching for extended sources (e.g. ridge-line based approaches; Barkus et al. 2022) is on-going.

The automatic source association of Mostert et al. (in preparation) actually makes use of the ML classifier developed here to reject unassociated compact sources that lie within the boundary of more extended multicomponent sources. It is likely that a selection of different ML and deep learning techniques will need to be developed and combined to fully solve this problem.

DATA AVAILABILITY

The tabular data underlying this article are provided in the supplementary online material (see Table B1 for columns description). The data sets were derived from LoTSS Data Release 1 publicly available at https://lofar-surveys.org/dr1_release.html.

REFERENCES

Alger M. J. et al., 2018, MNRAS, 478, 5547
Alhassan W., Taylor A., Vaccari M., 2018, MNRAS, 480, 2085
Aniyan A. K., Thorat K., 2017, ApJS, 230, 20
Arsioli B., Dedin P., 2020, MNRAS, 498, 1750
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Banfield J. K. et al., 2015, MNRAS, 453, 2326
Banzhaf W., Francone F. D., Keller R. E., Nordin P., 1998, Genetic Programming: An Introduction., Morgan Kaufmann Publishers Inc., San Francisco
Barkus B. et al., 2022, MNRAS, 509, 1
Barsotti D., Cerino F., Tiglio M., Villanueva A., 2022, Class. Quantum Gravity, 39, 085011
Bauer E., Kohavi R., 1999, Machine Learn., 36, 105
Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559
Best P. N., Kauffmann G., Heckman T. M., Brinchmann J., Charlot S., Ivezić Ž., White S. D. M., 2005, MNRAS, 362, 25
Bock D. C.-J., Large M. I., Sadler E. M., 1999, AJ, 117, 1578
Chambers K. C. et al., 2016, preprint (arXiv:1612.05560)
Ciliegi P., Zamorani G., Hasinger G., Lehmann I., Szokoly G., Wilson G., 2003, A&A, 398, 901
Collell G., Prelec D., Patil K. R., 2018, Neurocomputing, 275, 330
Condon J. J., Cotton W., Greisen E., Yin Q., Perley R., Taylor G., Broderick J., 1998, AJ, 115, 1693
Cutri R. M. et al., 2014, VizieR Online Data Catalog, II/328
De Rainville F.-M., Fortin F.-A., Gardner M.-A., Parizeau M., Gagné C., 2012, J. Machi. Learn. Res., 13, 2171
Dewdney P., Hall P., Schillizzi R., Lazio J., 2009, Proc. Inst. Electr. Electr. Eng. IEEE, 97, 1482
Dey A. et al., 2019, AJ, 157, 168
Dietterich T. G., 2000, Machine Learn., 40, 139
Duncan K. J. et al., 2019, A&A, 622, A3
Eiben A. E., Smith J., 2015, Nature, 521, 476
Fan D., Budavári T., Norris R. P., Hopkins A. M., 2015, MNRAS, 451, 1299
Fan D., Budavári T., Norris R. P., Basu A., 2020, MNRAS, 498, 565
Feurer M., Klein A., Eggensperger K., Springenberg J., Blum M., Hutter F., 2015, in Cortes C., Lawrence N., Lee D., Sugiyama M., Garnett R., eds, Advances in Neural Information Processing Systems, 28, Curran Associates, Inc., New York, p. 2962
Friedman J. H., 2001, Ann. Stat., 29, 1189
Friedman J. H., 2002, Comput. Stat. Data Anal., 38, 367
Galvin T. J. et al., 2020, MNRAS, 497, 2730
Gürkan G. et al., 2022, MNRAS, 512, 6104
Hale C. L. et al., 2021, PASA, 38, e058
He H., Bai Y., Garcia E. A., Li S., 2008, in IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, New York, p. 1322
He X., Zhao K., Chu X., 2021, Knowledge-Based Systems, 212, 106622
Hill G. et al., 2008, in Kodama T., Yamada T., Aoki K., eds, ASP Conf. Ser. Vol. 399, Panoramic Views of Galaxy Formation and Evolution. Astron. Soc. Pac., San Francisco, p. 115
Hotan A. W. et al., 2021, PASA, 38, e009
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ivezić Ž. et al., 2002, AJ, 124, 2364
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jarrett T. H., Chester T., Cutri R., Schneider S., Skrutskie M., Huchra J. P., 2000, AJ, 119, 2498
Jin H., Song Q., Hu X., 2019, in Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, p. 1946
Kondapally R. et al., 2021, A&A, 648, A3
Kruk S. et al., 2022, A&A, 661, A85
Lacy M. et al., 2020, PASP, 132, 035001
Laing R. A., Riley J. M., Longair M. S., 1983, MNRAS, 204, 151
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Le T. T., Fu W., Moore J. H., 2020, Bioinformatics, 36, 250
Luken K. J., Norris R. P., Park L. A., Wang X. R., Filipović M., 2022, Astron. Comput., 39, 100557
Lukic V., Brüggen M., Banfield J. K., Wong O. I., Rudnick L., Norris R. P., Simmons B., 2018, MNRAS, 476, 246
Lukic V., Brüggen M., Mingo B., Croston J., Kasieczka G., Best P., 2019, MNRAS, 487, 1729
Lundberg S. M., Lee S.-I., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., eds, Advances in Neural Information Processing Systems 30. Curran Associates, Inc., New York, p. 4765
Lundberg S. M. et al., 2020, Nature Mach. Intel., 2, 2522
Mallinar N., Budavári T., Lemson G., 2017, Astron. Comput., 20, 83

Mason L., Baxter J., Bartlett P., Frean M., 1999, in Solla S., Leen T., Muller K., eds, Advances in Neural Information Processing Systems. 12, MIT Press, US, p. 512

McKinney W. et al., 2010, in van der Walt S., Millman J., eds, Proc. 9th Python in Science Conference. p. 51

Mohan N., Rafferty D., 2015, Astrophysics Source Code Library, record ascl:1502.007

Molino P., Dudin Y., Miryala S. S., 2019, preprint (arXiv:1909.07930)

Mostert R. I. et al., 2021, A&A, 645, A89

Nisbet D. M., 2018, PhD thesis, University of Edinburgh, United Kingdom

Norris R. P., 2017, Nature Astrono., 1, 671

Norris R. P. et al., 2011, PASA, 28, 215

Ntwaetsile K., Geach J. E., 2021, MNRAS, 502, 3417

Olson R. S., Bartley N., Urbanowicz R. J., Moore J. H., 2016a, in Proceedings of the Genetic and Evolutionary Computation Conference 2016, ACM, New York, p. 485

Olson R. S., Urbanowicz R. J., Andrews P. C., Lavender N. A., Creis Kidd L., Moore J. H., 2016, in Giovanni S., Paolo B., eds, Applications of Evolutionary Computation: 19th European Conference. Springer International Publishing, US, p.123

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Proctor D., 2016, ApJS, 224, 18

Rengelink R., Tang Y., De Bruyn A., Miley G., Bremer M., Roettgering H., Bremer M., 1997, A&AS, 124, 259

Richter G. A., 1975, Astron. Nachr., 296, 65

Schapire R. E., Freund Y., 2014, Boosting: Foundations and Algorithms. MIT Press, US

Shimwell T. W. et al., 2017, A&A, 598, A104

Shimwell T. W. et al., 2019, A&A, 622, A1

Shimwell T. W. et al., 2022, A&A, 659, A1

Smith D. J. B. et al., 2016, in SF2A-2016: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, p. 271

Smolčić V. et al., 2017, A&A, 602, A1

Sutherland W., Saunders W., 1992, MNRAS, 259, 413

Sutton C. D., 2005, Handbook Stat., 24, 303

Tang H., Scaife A. M., Leahy J., 2019, MNRAS, 488, 3358

Tarsitano F., Bruderer C., Schawinski K., Hartley W. G., 2022, MNRAS, 511, 3330

Tasse C. et al., 2021, A&A, 648, A1

Trevor H., Robert T., Jerome F., 2009, Elements Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York

Vafaei Sadr A., Vos E. E., Bassett B. A., Hosenie Z., Oozeer N., Lochner M., 2019, MNRAS, 484, 2793

van Haarlem M. P. et al., 2013, A&A, 556, A2

Waskom M. L., 2021, J. Open Source Softw., 6, 3021

Weston S. D., Seymour N., Gulyaev S., Norris R. P., Banfield J., Vaccari M., Hopkins A. M., Franzen T. M. O., 2018, MNRAS, 473, 4523

Williams W. L. et al., 2019, A&A, 622, A2

Willis A. G., de Ruiter H. R., 1977, A&AS, 29, 103

Wu C. et al., 2019, MNRAS, 482, 1211

York D. G. et al., 2000, AJ, 120, 1579

Zimmer L., Lindauer M., Hutter F., 2021, IEEE Trans. Pattern Anal. Mach. Intell., 43, 3079

Zuntz J. et al., 2021, Open J. Astrophys., 4, 13

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

LOFARMachineLearningClassifier_MasterTable.csv

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

APPENDIX A: MACHINE-LEARNING TOOLS AND ALGORITHMS

A1 AutoML

In Section 4, we streamline model selection and optimization using Automated Machine Learning (AutoML). AutoML generates optimal ML pipelines by identifying the best model and model hyperparameters. AutoML has already been used in astronomy with the application of open source AutoML toolkits, and the use of artificial intelligence platforms. For instance, Arsioli & Dedin (2020) investigated the LUDWIG framework (Molino, Dudin & Miryala 2019) in the classification of blazars, and Zuntz et al. (2021) used AUTO-KERAS (Jin, Song & Hu 2019) to select one of the models for the LSST-DESC 3x2pt Tomography Optimization Challenge. Tarsitano et al. (2022) used the MODULOS.AI platform to select the best CNN architecture to perform optical galaxy morphological classification, Barsotti et al. (2022) used the DATAROBOT platform to predict gravitational waveforms from compact binaries, and Kruk et al. (2022) used the GOOGLE CLOUD AUTOML VISION to train a CNN for the Hubble Asteroid Hunter project. Other AutoML framework examples include the Tree-based Pipeline Optimization Tool (TPOT, Olson et al. 2016a) and AUTO-SKLEARN (Feurer et al. 2015) for traditional ML; and AUTO-PYTORCH (Zimmer, Lindauer & Hutter 2021) for deep learning. In this work we use TPOT, which we will explain in more detail next.

A1.1 TPOT

TPOT is an open source AutoML tool that evaluates different ML pipelines using genetic programming (GP; Banzhaf et al. 1998). In the field of evolutionary computation, GP (and its variants) are the most widely used type of evolutionary algorithm (e.g. Eiben & Smith 2015). By using a function that minimizes the error in the solution, these algorithms search for an optimal candidate within a group of potential solutions. TPOT was further developed to incorporate pipeline design automation; it performs feature selection, pre-processing and engineering, besides algorithm searching and optimization. It uses the PYTHON SCIKIT-LEARN library to implement both individual and ensemble tree-based models (decision trees; random forests and gradient boosting), non-probabilistic and probabilistic linear models (support vector machines and logistic regression), k-nearest neighbours; and it uses PYTORCH for neural networks. The code can be used for both classification and regression problems, and has been adapted to work with large datasets of features (Le, Fu & Moore 2020).

TPOT is built on Distributed Evolutionary Algorithms in PYTHON (DEAP; De Rainville et al. 2012), a framework that implements evolutionary computation. TPOT implements GP by creating trees of pipeline operators and evolving those operators in order to maximize accuracy. In brief (see Olson et al. 2016a; Olson et al. 2016b, for more details), in the first iteration (i.e. generation) TPOT sets and evaluates a random number of tree-based pipelines (i.e. population). The next generation is constructed as follows. First, 10 per cent of the new pipelines are copies of the highest accuracy pipeline from the previous generation; 90 per cent of new pipelines are selected from the previous generation using a three-way tournament selection with a two-way parsimony (i.e. three random pipelines are evaluated by first eliminating the one with the poorest performance and then choosing the simplest of the remaining two). Next, a proportion of these new generation pipelines are modified; 5 per cent of the

pipelines suffer a one-point *crossover*, which consists of swapping the contents of two random pipelines at a random split in the tree of operators. For 90 per cent of the remaining new pipelines a *mutation* is applied, where random operators are inserted, removed or replaced in the pipelines. The process is repeated for the number of generations defined.

A2 Ensembles of decision trees

We provide a brief description of ensembles of decision trees, with particular focus on the gradient boosting classifier (GBC), which is the type of algorithm we chose to apply (see Section 4). Ensembles of decision trees are sets of decisions trees, typically containing between 100 and 1000 trees. On their own, individual trees have moderate performance, but when combined, the ensemble achieves strong performance. There are different ways of creating these ensembles. Two common techniques are bagging and boosting. In bagging (e.g. random forest) the trees are created in parallel using splits between the features and the final prediction is, in general, given by the average of the predictions or the majority of the votes of the trees. By contrast, in boosting (e.g. gradient boosting) each tree is constructed sequentially by minimizing a loss function from the preceding tree, and in general trees (also referred as *weak learners*) with better performance have higher weight on the final predictions. (see e.g. Bauer & Kohavi 1999; Sutton 2005, for details and comparison of the methods). Since bagging models output average predictions, they reduce the variance of the model, and are therefore more robust to outliers and defective features (since these will be mainly ignored). In boosting, the trees grow in the direction where the loss is minimized. Therefore, each additional tree reduces the bias of the model. By aggregating the predictions from all the trees, boosting also reduces the model variance (Schapire & Freund 2013). As a consequence, boosting models are more powerful than bagging models, but they can also overfit in some cases, especially when the number of trees is increased: since each iteration reduces the training error, this can be made arbitrarily small by growing trees, which can lead to overfitting to the training data (Trevor, Robert & Jerome 2009).

The model used in this work is a GBC, for which the original formulation can be found in Friedman (2001). It is a stochastic boosting model (Friedman 2002) that uses a functional gradient descent (Mason et al. 1999). Consider an input training set of n examples, where each example has a set of feature values x and an output value y (where for our binary classifier y is defined as 0 or 1). The model sequentially builds an ensemble of weak learners, whose output prediction after iteration m is F_m.

The weak learners are constructed by first initializing a very simple model (F_0) in which the output prediction is a constant for all sources; this constant may be set to zero or may be chosen to minimize the initial loss function L_0. The loss function is defined based on the difference between predicted and true values, summed across the full training population: for the binary classifier used in this work, a binary log loss function (also known as binary cross-entropy or binomial deviance) is used:

$$L_m = \frac{1}{n} \sum_{i=1}^{i=n} y_i \log F_{m_i} + (1 - y_i) \log(1 - F_{m_i}), \tag{A1}$$

where L_m is the loss function for tree m and F_{m_i} is the model prediction for source i in iteration m. For each subsequent iteration, m, the procedure is then as follows. First, the *pseudo-residuals* for each training source are calculated from the model. Pseudo-residuals r for each source i are defined as

$$r_{i,m} = \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}. \tag{A2}$$

A modified data set is then made with the input parameters x, and output values of r. A tree is then fitted to this data set, with the resultant predictions $h_m(x_i)$. Using these predictions, the model prediction F_m is defined as

$$F_m(x) = F_{m-1}(x) + \nu h_m(x), \tag{A3}$$

where ν is the shrinkage parameter, commonly referred as learning rate, which scales the contribution of each tree by a factor between 0 and 1, acting as a regularization method (Friedman 2002). This value must be such that there is a trade-off with the number of trees M in the model. The loss function for the new tree can then be calculated, and the process is repeated until a final prediction $F_M(x)$ is produced. Due to the way that the model is constructed, it can be considered to be a weighted additive combination of all of the individual weak learners from which it is comprised:

$$F_M(x_i) = \sum_{m=1}^{M} \nu h_m(x_i). \tag{A4}$$

APPENDIX B: MASTER TABLE

An electronic table provides the source identification and feature data used as input to the ML algorithm, along with the source identification flags and diagnostic flags, and the final model prediction. Table B1 describes the columns provided in that table, which also include the columns from Table 2.

Table B1. Master table columns description. These were selected or computed using different catalogues (*a-f* listed in the footnote below).

Column	Definition and origin
Source information	
Source_Name	PYBDSF source identifier (typically a combination of RA and Dec. position)[a]
RA	PYBDSF source right ascension (deg)[a]
DEC	PYBDSF source declination (deg)[a]
Source_Name_final	Final radio source name (after any source association or deblending); NULL if artefact [d]
RA_final	Final radio source right ascension (deg)[d]
DEC_final	Final radio source declination (deg)[d]
AllWISE_lr	Source identifier of near-infrared AllWISE counterpart cross-matched by likelihood ratio[c,f]
AllWISE_final	Source identifier of finally-assigned near-infrared AllWISE counterpart[d]
objID_lr	Source identifier of optical Pan-STARRS cross-matched by likelihood ratio[c,f]
objID_final	Source identifier of finally-assigned optical Pan-STARRS counterpart[d]
Mosaic_ID	HETDEX mosaic which contains the source image[d]
Gaus_id	Gaussian component identifier used as feature[b]
NN_Source_Name	PYBDSF Source_Name of the nearest neighbour[a]
Identification flags	
W19dt	W19 decision tree main outcomes [0-LGZ, 1-LR (ID or no ID), 2-prefiltering, 3-large optical IDs, −99-artefacts][d]
Diagnosis flags	
association	PYBDSF source association diagnosis [1-single, 2-blended, 4-multicomponent, −99-artefacts][f]
accept_lr	Source suitable to LR technique [0-false,1-true, −99-artefact][f]
multi_component	Multicomponent source [0-false, 1-true, −99-artefact][f]
ML features	
(Several columns)	Machine-learning features from Table 2
Additional ML features	
n_gauss	Number of Gaussians that compose the PYBDSF source[b]
gauss_total_flux	Integrated flux density of the Gaussian component used as feature (mJy)[b]
Deconvolved sizes	
DC_Maj	PYBDSF source deconvolved major axis (arcsec)[a]
DC_Min	PYBDSF source deconvolved minor axis (arcsec)[a]
gauss_dc_maj	Gaussian deconvolved major axis (arcsec)[b]
gauss_dc_min	Gaussian deconvolved minor axis (arcsec)[b]
Likelihood ratio (LR) values	
lr	LR value match for the PYBDSF source[c,f]
gauss_lr	LR value match for the Gaussian[c,f]
highest_lr	Highest LR value match between the Gaussian and the source[c,f]
NN_lr	LR value match for the PYBDSF nearest neighbour[c,f]
Self-organizing map (SOM)	
10x10_closest_prototype_x	Row position of the PYBDSF source on the LoTSS DR1 cyclic 10x10 SOM[e]
10x10_closest_prototype_y	Column position of the PYBDSF source on the LoTSS DR1 cyclic 10x10 SOM[e]
Predictions	
probability_lr	Prediction probabilities to accept the LR match [range 0-1, 0-false, and 1-true][f]
dataset	Data set splitting [0-not on the training or test sets, 1-training set, 2-test set][f]
prediction_0.20	Predictions for 20 per cent threshold [0-send to LGZ, 1-accept LR, 2-recovered PYBDSF source component][f]

Note. Artefacts are flagged with the value of −99. [a]LoTSS DR1 PYBDSF radio source catalogue (Shimwell et al. 2019); [b]LoTSS DR1 PYBDSF Gaussian component catalogue (Shimwell et al. 2019); [c]LoTSS DR1 Gaussian and PYBDSF source LR catalogues (W19); [d]Optical LoTSS DR1 source catalogue (W19); [e]LoTSS DR1 self-organized map (SOM; Mostert et al. 2021); [f]Results calculated in this work.

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