Identifying galaxies, quasars and stars with machine learning: a new catalogue of classifications for 111 million SDSS sources without spectra

A. O. Clarke¹, A. M. M. Scaife¹, R. Greenhalgh¹ and V. Griguta¹

Jodrell Bank Centre For Astrophysics
e-mail: alex.clarke-3@manchester.ac.uk

Received 19 September 2019 / Accepted –

ABSTRACT

We use 2.4 million spectroscopically labelled sources from the SDSS catalogue to train an optimized random forest classifier using photometry from the Sloan Digital Sky Survey (SDSS) and Widefield Infrared Survey Explorer (WISE). We apply this machine learning model to 111 million previously unlabelled sources from the SDSS photometric catalogue without existing spectroscopic observations. Our new catalogue contains 49.7 million galaxies, 2.4 million quasars, and 59.2 million stars. We provide individual classification probabilities for each source, with 6.4 million galaxies (13%), 0.35 million quasars (14%) and 44.3 million stars (75%) having classification probabilities greater than 0.99, and 34.8 million galaxies (70%), 0.77 million quasars (32%) and 55.3 million stars (93%) having classification probabilities greater than 0.9. We determine Precision, Recall and F₁ score as a function of feature selection, including scenarios with additional external variables. We investigate the effect of class imbalance on our machine learning model and discuss the implications of transfer learning for populations of sources at fainter magnitudes than the training set. We use a non-linear dimension reduction technique (Uniform Manifold Approximation and Projection: UMAP) in unsupervised, semi-supervised and fully-supervised schemes to visualise the separation of galaxies, quasars and stars in a two dimensional space using the 2.4 million spectroscopically labelled training sources. We confirm that when the same algorithm is applied to the 111 million sources without spectra it is in strong agreement with the class labels applied by our random forest model.

Key words. galaxies – galaxies:quasars – stars – surveys – catalogs – methods:data analysis – methods:statistical

1. Introduction

The classification scheme of galaxies, quasars and stars is one of the most fundamental in astronomy. The early cataloguing of stars and their distribution in the sky led to the understanding that they make up our own galaxy (Herschel 1789) and, following the distinction that Andromeda was a separate galaxy to our own (Opik 1922; Hubble 1929), numerous galaxies began to be surveyed as more powerful telescopes were built. The designation of quasar arose after radio emission was detected from unresolved star-like sources with high redshifts (e.g: 3C48 and 3C273; Smith & Hoffleit 1961; Greenstein & Matthews 1963; Matthews & Sandage 1963; Schmidt 1963; Greenstein & Schmidt 1964). This emission was later demonstrated to be produced by accretion disks surrounding super-massive black holes at the centre of some galaxies (Burbidge et al. 1963; Rees 1984; Begelman et al. 1984). For quasars, the emission from this central region on scales less than a light year, known as an Active Galactic Nucleus (AGN; Urry & Padovani 1995), is expected to dominate over the light from the host galaxy. Large samples of quasars (e.g: Pâris et al. 2018) are now routinely selected through the identification of characteristic high-ionisation emission lines in their optical spectrum (e.g: C IV, Mg II; Francis et al. 1991; Vanden Berk et al. 2001), as well as via spectroscopic follow-up of optical sources where a radio counterpart is also present (Gürkan et al. 2019). Whilst quasars are typically located at redshifts high enough to be unresolved with optical telescopes, some nearby resolved galaxies (Seyfert galaxies; Weedman 1977; Antonucci 2012) are also labelled as quasars, having bright and compact cores with associated emission lines from AGN, although they are comparatively less bright than high-redshift quasars.

Providing classification labels for astronomical catalogues containing large numbers of sources has a wide range of benefits, both for studies of individual systems and for statistical population analyses. In particular, a significant range of science goals are dependent on large samples of quasars, which are still the minority class. This consideration has been important in motivating the construction of new facilities such as the Square Kilometre Array (SKA; Jarvis et al. 2015) and the Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al. 2009; LSST Dark Energy Science Collaboration 2012; Ivezić et al. 2019). Such science objectives include Lyman-α forest surveys (Rauch 1998; McDonald & Eisenstein 2007), cosmic magnetism studies (Scranton et al. 2005), general cosmology (Leistedt & Peiris 2014; Hutsemékers et al. 2005) and the evolution of galaxies (Schmidt & Green 1983; Sanders et al. 1988; Kauffmann & Haehnelt 2000), amongst others.

Millions of sources have already been catalogued from telescopes such as (e.g.) the Sloan Digital Sky Survey (SDSS; Aguado et al. 2019), the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) and the LOw Frequency ARray (LOFAR; van Haarlem et al. 2013; Shimwell et al. 2019). The next generation of telescopes are predicted to detect many orders of magnitude more sources, requiring more sophisticated techniques to create and manage these vast data sets.
of magnitude more sources. The LSST is expected to catalogue approximately 20 billion galaxies and a similar number of stars (Ivezić et al. 2019). Source count predictions for the SKA indicate a source density of around ten galaxies per square arcminute for phase one of the array, and up to 75 galaxies per square arcminute for phase two (assuming a detection threshold of $S_{1.4GHz} \geq 100\mu Jy$). This results in totals of 1 and 8 billion sources, respectively, for a survey area of 3r steradians (Jarvis et al. 2015). Consequently, it is becoming unfeasible for astronomers to manually verify and label individual sources and, whilst efforts such as Galaxy Zoo (Lintott et al. 2008, 2011) bring in many more people to help sift through data, this effort alone is not expected to be able to keep up with the source counts anticipated for the next generation of telescopes. For such large datasets, machine learning algorithms are becoming an increasingly valuable tool for analysis and data exploration. The development of such algorithms in computer science fields has accelerated rapidly in the last decade, focusing on processing large datasets in high performance computing workflows and cloud computing systems (Jones 2014; Wu et al. 2016).

Although distinguishing astronomical source type is normally straightforward where detailed data are available, such as spectroscopy and multi-wavelength observations, the complexity of obtaining detailed observations for millions of individual sources is time consuming and generally impractical for the largest samples of sources given the survey speeds of the current generation of telescopes. In contrast, classifying sources using only photometry in multiple wavebands, and labelling them based on their colours (differences between pairs of photometry measurements), is comparatively fast. With three orders of magnitude less wavelength coverage than spectroscopy, photometry cannot capture the same detail as spectra; however, it can capture the overall shape of the spectrum that distinguishes different types of sources. In the optical, stars show a black-body spectrum, galaxies show the superposition of many black-body spectra, and quasars show a comparatively flatter spectrum due to AGN emission. The most widely used colours for source classification are $u-g$ and $g-r$ in SDSS data, and $w1-w2$ and $w2-w3$ in WISE data (Nikutta et al. 2014; Peters et al. 2015). Consequently, photometry data have been demonstrated to be useful as machine learning features in source-type classification by a number of studies (e.g: Carrasco et al. 2015; Schindler et al. 2019; Kang et al. 2019; Nakoneczny et al. 2019; Bai et al. 2019). Furthermore, testing whether a source is resolved or unresolved can help distinguish the extended profiles of galaxies from stars and quasars (Aguado et al. 2019), and be a useful machine learning feature.

In this work, we apply supervised, unsupervised and semi-supervised machine learning algorithms to classify galaxies, quasars and stars using SDSS and WISE data. In Section 2 we introduce the SDSS and WISE data and describe our feature set. In Section 3 we introduce the machine learning algorithms, describe how the models were optimised using a spectroscopically selected training dataset and give an in-depth evaluation of the model performance as a function of source magnitude and classification. In Section 4 we apply our optimised model to an SDSS photometric catalogue of 111 million previously unclassified sources. In section 5 we discuss the results, and in Section 6 we draw our conclusions. All of the code used in this paper to gather, process and analyse the data is available on our Github repository.

1 https://www.github.com/informationcake/SDSS-ML/

2. Data

Two datasets were constructed from the SDSS Data Release 15 (DR15 Aguado et al. 2019) catalogue. These comprise (i) a labelled dataset of spectroscopically observed sources, and (ii) an unlabelled dataset of photometrically observed sources with no associated spectroscopic observation. In both cases we select sources that have WISE counterparts.

For SDSS, photometric measurements are provided in five optical bands: $u(\lambda = 0.355 \mu m)$, $g(\lambda = 0.477 \mu m)$, $r(\lambda = 0.623 \mu m)$, $i(\lambda = 0.762 \mu m)$, $z(\lambda = 0.913 \mu m)$, with associated errors. A correction for Galactic extinction is also provided per band (Schlegel et al. 1998). WISE provides photometric measurements in four infrared bands: $w1(\lambda = 3.4 \mu m)$, $w2(\lambda = 4.6 \mu m)$, $w3(\lambda = 12 \mu m)$, and $w4(\lambda = 22 \mu m)$ with their associated errors. SDSS photometric measurements, as described in (Stoughton et al. 2002), are optimised for different types of source. For unresolved point sources (e.g. stars and quasars) that are well-described by the point spread function (PSF), the best measure of the total flux is determined by fitting a PSF model to the source, referred to as $psfMag$. However, for resolved sources such as galaxies this is not the case, and a better measure of total flux comes from a model fitted to a source’s radial profile. There are several of these model magnitudes associated with each catalog source. The $devMag$ and $expMag$ magnitudes are associated with de Vaucouleurs and exponential model fits, respectively. These magnitudes are calculated from independent models in each of the 5 bands. In addition, there is the $modelMag$, which uses the better of the two fits in the $r$-band as a matched aperture to calculate the flux in all bands. For extended sources, this option provides the best measurement of colours due to the flux being measured over equivalent apertures across all bands. A composite model magnitude $cmdMag$ is also defined, taking the best fit from either the de Vaucouleurs or the exponential model in each band and obtaining a linear combination of the two that best fits the image. In practice, $cmdMag$ is the optimum total flux indicator, agreeing well with $psfMag$ and $modelMag$. However, it does not result in as high a signal-to-noise measurement for colours compared to $modelMag$.

Fig. 1. Histogram of 2.4 million spectroscopically observed galaxies, quasars and stars from both the SDSS and BOSS instruments (after removing unclean data). There is a double peaked distribution (particularly with galaxies) due to each instrument having different target magnitude selection criteria.
For each source we determine if it is resolved or point-like by calculating the difference between the \( cmodel \) magnitude and the PSF magnitude in the \( r \)-band:

\[
\text{resolved} = |psf_r - cmod_r|.
\]

### 2.1. Spectroscopic data

SDSS data release 15 includes 2,538,781 unique sources, spectroscopically observed and cross-matched with detections in the WISE catalogue. Each source in the catalogue is labelled as "STAR", "GALAXY" or "QSO" (quasar) depending on the outcome of fitting various pre-defined models to the spectrum of each source (Bolton et al. 2012).

We obtain the DR15 data by submitting Structured Query Language (SQL) jobs to the CasJobs component of SkyServer API. We then apply various conditions upon retrieving the data to ensure the dataset is clean, with minimal contamination from sources which could have incorrect classifications or problematic photometry.

Firstly, 18,426 sources have more than one WISE match within five arc-seconds. We remove the duplicate entries, only keeping the entry with the closest WISE match. Next, we remove 117 sources that have -9999 for the \( cmodel \) magnitudes where the fit has failed. Sources with spectra that are known to have problems are indicated in the DR15 dataset by the \( z\text{warning} \) flag. We only select sources where this flag has a value of either 0 or 16. If the flag is zero, this indicates the spectra has no problems. If the flag is 16, this indicates the "MANY_OUTLIERS" warning which is only present for data taken with the SDSS spectrograph and not with the BOSS spectrograph, and usually indicates a high signal-to-noise spectrum or broad emission lines in a galaxy. Consequently, it rarely signifies a true error. These \( z\text{warning} \) conditions remove 81,239 sources, 64,962 of which are due to the "SMALL_DELTA_CHI2" warning \( (z\text{warning} = 4) \), which indicates that the chi-squared value of the best fit spectrum is too close to that of the second best. Finally, we remove 511 sources where the value of any of the WISE band magnitudes are set to 9999, where extracting the magnitude has failed.

In total the cleaning process described above removes 100,293 sources: 66,701 galaxies, 28,354 quasars and 5238 stars. This leaves a total of 2,438,488 sources, with 1,777,307 galaxy, 295,576 quasars and 365,605 stars. The class ratio is approximately 6.0:1.0:1.2. Within the resulting dataset, 1,293,359 sources were observed using the BOSS spectrograph (53%) and 1,145,129 were observed using the SDSS spectrograph (47%). Due to the different survey goals for each of these spectrographs, there is a difference in the magnitude distribution for the sources measured in each case. This is illustrated in Figure 1, where a double peaked distribution for both galaxies and quasars is evident.

### 2.2. Photometric data

The complete SDSS photometric catalogue contains 1,231,051,050 entries, including repeat observations. To create the catalogue of previously unclassified photometrically observed sources, we use sources from the SDSS \textit{PhotoPrimary} table with WISE matches that have no associated spectroscopic observation and clean photometry (where the flag \( \text{clean}=1 \)).

We limit the dataset to have SDSS magnitudes between 0 and 35 after the correction for Galactic extinction, and WISE magnitudes between 0 and 30. This results in a catalogue of 111,344,899 sources. A small number of SDSS sources (51,866) were found to be associated with more than one WISE source. We removed duplicate WISE associations by selecting the WISE source closest in angular separation to the SDSS position. Following these selection steps, the remaining dataset contains 111,293,033 unclassified photometrically observed sources.

### 2.3. Feature Set

We use SDSS and WISE photometry, and the \( \text{resolved} \), parameter, as features in our machine learning models. We do not explicitly calculate colours, and instead give all five SDSS bands, and all four WISE bands as features. We use the dataset of spectroscopically observed sources to train, validate and test a model to predict the class labels which are spectroscopically confirmed (we do not use the spectra itself as a feature or in the analysis). We use this model to predict the class labels on the second dataset of unlabelled photometrically observed sources that do not have spectra in SDSS.

### 3. The machine learning model

In this work we use a random forest (thoroughly reviewed in Louppe 2014) as a supervised learning algorithm to classify galaxies, quasars and stars. A random forest is an ensemble of independent decision trees where each individual tree is trained on a random subset of both features and data samples. The predicted classification for a new data sample comes from a majority consensus classification across the full set of approximately uncorrelated machine learning models from all the trees in the forest. These principles make a random forest robust to over-fitting, and minimise the variance and bias in its predictions. In addition, random forests are a commonly used algorithm for supervised learning problems due to a number of other strengths: they can deal with numerical and categorical features over different scales, are effective at multi-class problems, and naturally return classification probabilities and feature importance rankings.

We divide our spectroscopically-classified dataset into a training dataset and a testing dataset. The fraction of the complete training set used for each category is discussed in Section 3.1. The training set is used to train the random forest classifier and fit the machine learning model; it must be large enough for the random forest to extract a sufficient amount of information on the expected types of galaxies, quasars and stars and their relation to the features we are using. The test dataset is used to derive a variety of performance metrics which assess how the machine learning model would perform on unseen data. These metrics will be used to assess the confidence of the classifications when applying the model to unlabelled sources without spectra. When creating these two subsets of data, we ensure that each contains the same ratio \( (6.0:1.0:1.2) \) of the three different classes in order not to bias training, validation or testing.

Instead of splitting out a fixed validation dataset we implement a cross-validation scheme during training to tune the hyper-parameters, described in more detail in Section 3.2. Validation of this type is used to ensure that the fitted machine learning model does not over- or under-fit the training data.
We use precision, recall and F₁ score as metrics to assess the performance of the model.

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},
\]

indicates how good the classifier is at identifying true positives (TP), which are the correctly identified sources. A low precision for an individual class would indicate a low fraction of positive identifications.

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},
\]

indicates how good the classifier is at minimising false negatives. A low recall for an individual class would indicate it is often misclassified as another class.

\[
F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}},
\]

is the harmonic mean of precision and recall and is used as an overall performance metric. In this multi-class scheme these metrics are calculated per class to show the relative performance of each class.

### 3.1. Training data volume

Given the large number of sources available to us in the spectroscopic catalogue, we initially split the full dataset into two halves, using one half as a preliminary training set, and the other as a preliminary test set. By taking increasing fractions of the full training set to train a random forest model, we can use these models to predict the class labels of the sources in a consistent test dataset and assess how the volume of data in the training set affects the model performance.

During training we set a weight per class which is inversely proportional to the frequency of that class in the dataset. This is used to prevent the imbalance between the different classes, primarily caused by an excess of galaxies, influencing the fit. Figure 2 shows how the F₁ score per class varies as a function of the fraction of the full training set used to fit the model. The performance increase becomes linear for all classes when using more than 60% of the of the full training dataset, equivalent to 30% of all sources with spectra. When using fewer than 60% of the samples in the full training set the classes show a non-linear increase in performance that is more significant for the minority classes (quasars and stars). This effect is due to the training data subset not capturing a broad enough range of examples for all classes and indicates that a minimum of 60% of the training dataset, equivalent to 30% of the full spectroscopic catalogue, should be used to fit the machine learning model in order to avoid biased results.

Given the class imbalance in our dataset that arises from there being approximately 5 times more galaxies than stars, and 6 times more than quasars, we repeat the procedure described above including only 20% of the galaxies in the training set. Balancing the three classes in the training dataset whilst maintaining the same unbalanced test dataset gives the result shown by a dashed line in Figure 2. The resulting F₁ scores for the model are lower for all classes, decreasing by the same fraction for all fractions of the training dataset. There is a small increase in precision for galaxies, but a larger decrease in precision for quasars and stars. There is a decrease in recall for galaxies, but an increase in recall for quasars and stars. By limiting the number of galaxies available to train on, the model is hindered on all classes, with a particularly large drop in the precision of quasars (a large increase in false positives in this class from galaxies misclassified as quasars). Overall our approach of including weights in the fitting of the random forest is robust to the galaxy class imbalance without requiring us to sub-sample this class. Reducing the number of galaxies in the training set results in a poorer model for all classes.

### 3.2. Hyper-parameter optimisation

Random forest algorithms are inherently designed to prevent over-fitting through their selection of random subsets of data and features during training. We use a validation dataset in order to tune the hyper-parameters of the random forest to optimise the model, which could lead to over- or under-fitting if set incorrectly. To do this we use cross-validation with a five-fold split. This means that 80% of our full training dataset (40% of the full spectroscopic catalogue) is used to fit the machine learning model in each fold, and 20% of our training set (10% of the full spectroscopic catalogue) is used for validating each fold. Given the results described in Section 3.1, this reduction in the full training dataset will not have an adverse effect on the performance of our model applied to each fold.
We optimize over three main hyper-parameters that influence the fitting of the random forest model: the number of trees in the forest \( n_{\text{estimators}} \), the maximum number of features per decision tree \( \text{max} \_ \text{features} \), and the minimum number of samples required for a decision tree to split \( \text{min} \_ \text{samples} \_ \text{leaf} \). All other hyper-parameters are left at their default values whilst tuning these parameters one at a time.

We found that using 200 estimators in the random forest gave the best performance whilst minimising run-time. Using 1000 trees did not provide a notable increase in performance over 200, but required five times longer to train the model. This is consistent with the expected computational complexity for random forests, which scales linearly with the number of estimators (Louppe 2014). Increasing the \( \text{min} \_ \text{samples} \_ \text{leaf} \) parameter in the range from one to 500 always results in a large drop in the \( F_1 \) score, and so we keep it set to one. Adjusting the \( \text{max} \_ \text{features} \) parameter in the range from two to six results in a very small increase in \( F_1 \) score. However, given that we only have 10 features, substantially increasing this parameter will lead to over-fitting. We therefore use the the default value which is the square root of the number of features, which is rounded down to three in our case.

After optimising the hyper-parameters individually, we verify the optimisation by re-fitting the random forest using all three optimised hyper-parameters together, whilst also implementing five-fold cross-validation. This demonstrated that the model had not over- or under-fitted the training data, and would generalise to an unseen dataset.

Finally, to maximise the volume of data available for training, we re-fit the machine learning model without implementing a cross-validation scheme. This uses the complete training dataset (50% of the full spectroscopic catalogue) and the random forest with optimised hyper-parameters. The resulting machine learning model is our production model.

### 3.3. Feature optimisation

A number of other measurements are available for each source in the SDSS catalogue, which might alternatively be used as machine learning features. In order to establish that there was no advantage in using these alternative features we evaluated the model performance for a variety of feature iterations using the test dataset. For each of these feature iterations we also repeated the procedures in Section 3.2 to establish that the changes to the feature set were not significant enough to affect the optimal hyper-parameter values. The precision, recall and \( F_1 \) scores per class are shown in Table 1 for each different set of features.

Including the \( \text{resolved} \_ \text{parameter} \) (Equation 1) as a feature gives an improved classification performance for all classes, mostly for the recall of quasars (Table 1). This is illustrated in Figure A.1 where we see a large improvement in the classification of unresolved stars and quasars, and resolved galaxies. We use the \( r \)-band magnitude for this measure as it has the highest signal to noise. We see the performance decrease when using either \( u \)- or \( z \)-bands for this purpose due to the poorer signal to noise in these bands. Using the three high signal to noise bands, \( g \), \( r \), \( i \), to calculate the resolved nature of a source as three separate features does not increase performance as this does not provide the classifier with any additional information, consequently we use only the \( r \)-band value in order to minimise the number of features.

Excluding WISE magnitudes from the features gives worse results for all classes, affecting quasars the most with an \( F_1 \) score dropping from 0.952 to 0.936. This is equally due to an increase in false negatives (quasars classified as galaxies or stars) with recall being lower by 0.018, and an increase in false positives (stars and galaxies classified as quasars), with precision lower by 0.015. Only using WISE magnitudes as features significantly lowers the \( F_1 \) scores for all classes, particularly so for stars.

Performing a correction for extinction due to our own Galaxy does not alter the performance of the model for any class. We choose to include SDSS extinction-corrected magnitudes in our feature set. The extinction correction for the \( u \), \( g \), \( r \), \( i \), \( z \) bands has a mean and standard deviation of 0.13 ± 0.08, 0.10 ± 0.06, 0.07 ± 0.04, 0.05 ± 0.03, 0.04 ± 0.02 respectively for spectroscopically observed sources.

Using \( \text{cmode} \) magnitudes in place of PSF magnitudes for the SDSS bands gives a poorer performance across all target classes, with the most significant effect evident for stars. Using a combination of \( \text{cmode} \) and PSF magnitudes provides no improvement over using PSF magnitudes alone.

Overall, the optimal combination of features includes the SDSS PSF magnitudes, which we choose correct for Galactic extinction, the WISE magnitudes and the \( \text{resolved} \_ \text{parameter} \). The relative weight assigned to each feature from the resulting model is shown in Figure 3.

### 3.4. Performance evaluation

The random forest model derived in Section 3.2 was applied to our test dataset of 1.2 million spectroscopically confirmed sources (50% of our full spectroscopic catalogue) using the optimal combination of features described in Section 3.3. Figure 3 shows how the features vary for correct and missed sources per class. There are 884 211 correct galaxies, 139 572 correct quasars and 176 403 correct stars. 3405 galaxies were missed as quasars, 7605 quasars were missed as galaxies, 4071 stars were missed as galaxies, 1038 galaxies missed as stars, 2328 stars missed as quasars and 611 quasars missed as stars. Figure 5 shows the distribution of correctly identified and missed sources as a function of feature space. The average magnitude and one standard deviation range is shown for the SDSS and WISE bands for the left side in Figure 5, and a histogram of the \( \text{resolved} \_ \text{parameter} \) is shown on the right side. This is shown pairwise for all classes (top: quasar/star, middle: galaxy/star, bottom: galaxy/quasar) indicating in which feature ranges particular sources are misclassified as others. Furthermore, a histogram of each of the SDSS and WISE magnitude features is shown in Figure A.2, per class, for correct and missed sources.
In summary, it can be seen that correctly identified galaxies and stars show a similarly shaped spectrum across the SDSS bands, with galaxies being brighter in the WISE bands. Correctly identified quasars show a flatter spectrum across the SDSS bands compared to correctly identified galaxies and stars. Correctly identified quasars also have a much larger \( w1 - w2 \) difference than galaxies and stars. Correctly identified quasars and stars are mostly unresolved, with similar distributions, whilst correctly identified galaxies are mostly resolved.

Quasars missed as galaxies (7605 sources) appear similar to galaxies across both the SDSS and WISE bands, and are more often resolved (bottom row of Figures 5 and A.2). Most of these sources have redshifts less than 1 (see Figure 4). They can appear as resolved galaxies where the light from the galaxy dominates over the AGN, giving a galaxy-like spectrum across the SDSS and WISE bands whilst there is still an AGN present (e.g. the top row of Figure 6). Alternatively, they can be more unresolved galaxies that are much redder (e.g. the bottom row of Figure 6), although their resolved parameters are still much larger on average than point source quasars and stars (shown by the dashed pink line in the bottom right panel of Figure 5). Their distribution per magnitude feature in SDSS and WISE is shown in Figure A.2. In particular, their distribution has a shape that follows that of the correctly identified galaxies (shown by the dashed pink line in the bottom row of Figure A.2). Overall, quasars missed as galaxies tend to be resolved objects (at low redshifts) which are redder in colour.

Galaxies missed as quasars (3405 sources) appear similar to quasars in the WISE bands, and have a spectrum across the SDSS of a similar shape to quasars but are fainter than quasars on average. Compared to correctly identified galaxies, they tend to be unresolved (shown by the dashed grey line in the bottom right panel of Figure 5), but have a similar distribution in redshift (Figure 4). In the SDSS bands they appearing bluer on average than correctly identified galaxies, with a flatter spectrum. Figure A.4 shows a selection of examples of galaxies missed as quasars. Their distribution per magnitude feature in SDSS and WISE is shown in the bottom row of Figure A.2.

Galaxies missed as stars (1038 sources) look similar to correctly identified galaxies in the SDSS bands, but are fainter in the WISE bands (middle row of Figures 5 and A.2). Figure A.5 shows a selection of examples of these sources. The top-left image in Figure A.5 is a case where the SDSS classification pipeline is incorrect, and our model has correctly identified it as a star. There may be a small number other cases like this, although without manually inspecting all of them (which we have not done), it is not possible to find these cases automatically with this dataset. Their distribution in redshift and the resolved parameter is similar to correctly identified galaxies (middle row of Figures 4 and 3).

Stars missed as galaxies (4071 sources) have a similarly shaped spectrum to stars and galaxies in the SDSS and WISE bands, but are much fainter and mostly resolved (middle row of Figures 5 and A.2). The top row of Figure A.6 shows examples of these resolved sources with resolved, > 1. The first image shows a galaxy with a foreground star at its centre. The second image shows a galaxy with a foreground star just off-centre, which is close and bright enough to dominate the spectrograph, hence why the SDSS pipeline has incorrectly labelled it as a star whilst our algorithm has correctly labelled it as a galaxy. The fifth image shows a foreground star on top of a galaxy where the SDSS pipeline has correctly labelled both the star and the background galaxy (which was in our training set) from their spectra. However our algorithm has missed this star as a galaxy, likely due to contamination from the background.
galaxy, which is 0.35 magnitudes brighter in the PSF r-band, and the large value of the resolved, parameter. The bottom row of Figure ?? shows the rarer examples of more unresolved sources (resolved, < 0.3), where the reason for the misclassification is predominantly due to these stars having a spectrum across the SDSS and WISE bands that is very similar to galaxies.

Quasars missed as stars (611 sources) tend to be fainter in the w2-band than correctly identified quasars, which results in a smaller w1 − w2 difference, and have a spectrum in the SDSS bands similar to stars (top row of Figures 5 and A.2). They are much fainter in the u-band than correct quasars making them appear redder on average. They have a similar distribution in the resolve, parameter to correct quasars, correct stars and missed stars (top right panel of Figure 5). They tend to be at higher redshifts, as seen in the top panel of Figure 4, since lower redshift quasars are more likely to be resolved and misclassified.

Fig. 5. Left: Average magnitude and one standard deviation error bar for each feature (waveband) for correct and missed sources, per class. Error bars and lines are offset in the x-axis per feature for clarity. The resolved feature was also used in the model and extinction corrected PSF magnitudes were used for SDSS. The F1 score is 0.991 for galaxies, 0.952 for quasars and 0.978 for stars. Right: Histogram of the resolved, parameter per class, for correct and missed sources.
as galaxies (bottom panel in Figure 4). The main reason for their misclassification is due to the shape of their spectrum across the SDSS and WISE bands, and being very faint in the $u$-band. Some examples of these are shown in Figure A.7.

Stars missed as quasars (2328 sources) tend to be much fainter across all bands than correctly identified stars and have a larger $w_1 - w_2$ difference (top row of Figures 5 and A.2). They are significantly fainter in the $z$-band than correct stars, giving them a flatter spectrum across the SDSS bands. They have a similar distribution in the resolved parameter to correct stars, correct quasars and missed quasars. The main reason for their misclassification is due to the shape of their spectrum across the SDSS and WISE bands, and being very faint sources where there are much fewer stars in the training set. Some examples of these are shown in Figure A.7.

3.4.1. Precision, recall and $F_1$ score

The average precision, recall and $F_1$ scores per class are shown in Table 1. These metrics represent the overall performance of the model on the test dataset of 1.2 million spectroscopically confirmed sources (50% of our full spectroscopic catalogue). Furthermore, we calculate precision, recall and $F_1$ score for sources binned along various variables to assess the model performance throughout the 10-dimensional feature space.

Figure 7 shows histograms of precision, recall and $F_1$ score over PSF $r$-band magnitude, PSF $r$-band error, a one dimensional feature and the resolved parameter. The shaded area shows a one standard deviation confidence interval calculated per bin from the Wilson score interval (Wilson 1927). The Wilson interval score behaves well with probability distributions close to zero or one (unlike a Normal distribution) or with small sample sizes. A histogram per class is also shown to guide where the source counts drop, normalised relative to the galaxy class which peaks at a half. Furthermore, Figure A.3 shows histograms of precision, recall and $F_1$ score over each of the individual SDSS and WISE magnitude bands.

Stars are classified correctly with very high precision as a function of the PSF $r$-band magnitude. The precision has values of $0.99 - 1$ between PSF $r$-band magnitudes of $14 - 18$, dropping to $0.95$ at a magnitude of $21$, before rising again and then tailing off as the source density of stars decreases. The recall for stars is higher than $0.95$ in the magnitude range $14 - 19$, but drops to $0.73$ at magnitude $22$ before rising again. This drop in recall is due to an increase in false negatives. These stars missed as galaxies (4071 sources) are mostly stars super-posed along the line of sight towards background galaxies, increasing their resolved parameter. The precision, recall and $F_1$ score as a function of each magnitude feature in SDSS and WISE is shown in Figure A.3. The other source of false negatives, the stars missed as quasars (2328 sources), are mostly faint unresolved sources with a distribution per magnitude feature in SDSS and WISE shown in Figure A.3. Overall, the $F_1$ score for stars is greater than $0.8$ for the majority of sources, excluding the very high and low magnitude limits where the number of sources is low. The $F_1$ score drops at magnitudes $20 - 22.5$ due to an increase in false negatives. As a function of the resolved parameter (bottom row in Figure 7) the precision stays above 0.9, dropping when resolved, is greater than 4. The recall drops significantly for resolved stars.

Galaxies are classified correctly with high precision as a function of the PSF $r$-band magnitude. This is higher than $0.95$ from magnitude $15.5$ to $23$, and reduces to $0.8$ from magnitude $23$ to $25$ as the source density falls. The recall for galaxies is also very high, being above $0.98$ from magnitude $16$ to $23.5$, only dropping either side of this as the source density significantly decreases. Overall the $F_1$ score is very high over the entire magnitude range, only dropping when the source density significantly falls. As a function of the resolved parameter, the precision and recall for galaxies drops below $0.95$ when resolved, is less than $0.2$ or greater than $6$, as the source density significantly drops.

Quasars are classified correctly with high precision as a function of the PSF $r$-band magnitude, and is greater than $0.95$ between magnitude $15.7$ and $20.7$. Precision falls either side of this as the source density drops. The recall for quasars is above $0.95$ between magnitude $18.6$ and $20.7$. Outside of this range the recall drops due to an increased number of false negatives. These are mostly due to quasars missed as galaxies (7605 sources) while only a small number (611 sources) are due to quasars missed as stars. The bottom row in Figure 7 shows how the precision and recall drop below 0.95 when the
resolvedr, parameter is greater than 0.8 and 1.1, respectively. Whilst the precision stays high over the majority of the source population, the recall drops significantly for resolved sources due to resolved galaxies being misclassified as quasars.

Whilst fainter sources generally have larger PSF r magnitude errors we do not see evidence that this affects the performance of the classifier (second row of Figure 7). For example, the recall for stars falls as the PSF magnitude error increases, however the precision remains high. This drop in recall for stars (due to an increase in misclassified stars) is seen at fainter magnitudes from 20 to 22, where the classifier confuses them with galaxies and quasars. In general for all three classes, the F1 scores only drop as a function of PSF r error when the source density falls significantly (sources with significantly big or small errors).

3.4.2. Classification probabilities

The random forest algorithm naturally returns the probabilities per class of a classification, which sum to one over the number of possible classes. The class with the largest probability is the predicted class assigned by the algorithm. The left of Figure 8 shows a histogram per classification possibility over the probabilities for the assigned class. For example, the dashed pink line shows quasars that the random forest predicted were stars, and is a histogram of the probability that the random forest assigned to it being a quasar. Incorrect classes must have probabilities less than one half, whilst correct classes must have probabilities greater than one third. The right side of Figure 8 shows the same but as a cumulative normalised histogram.

The 0.99-1 bin in Figure 8 contains 73%, 53% and 83% of the correctly classified galaxies, quasars and stars respectively. Furthermore, 97% of galaxies, 84% of quasars and 95% of stars have classification probabilities above 0.9. Quasars are the weakest performing class in this regard, where in general the correctly classified quasars have lower classification probabilities than correctly classified stars or galaxies. This indicates that this class is the most difficult to classify.

The last column in Figure 7 has the classification probabilities (dashed lines) and one standard deviation (dotted lines) plotted along with the F1 scores. Overall they follow the trend in the F1 scores for all classes. They only start to decrease in areas with very low numbers of sources (for example at the edges of the plots).

Figure 9 shows a two-dimensional histogram of the PSF r magnitude and classification probability. Misclassified sources are also plotted individually and a normalised one-dimensional histogram of the PSF r magnitude is overlaid. Overall the random forest probabilities show a similar distribution over the whole PSF r magnitude range. Misclassified sources are mostly found with r magnitudes from 20 to 22.

3.4.3. Limiting the training set

One of the primary intended uses for machine learning algorithms is to classify sources detected in new astronomical surveys based on models trained from existing data, a form of transfer learning (e.g: Pratt et al. 1991; Tang et al. 2019). As telescopes become more powerful, new surveys typically become progressively deeper in sensitivity and recover fainter populations of sources. One possible limitation for this form of transfer learning is therefore the introduction of biases when deploying a model on a population of sources that are fainter than the population used for training. Here we consider how the performance of our machine learning model changes as a function of source brightness.

Figure 10 shows the precision, recall and F1 score when upper limits were set on the r magnitude in the training set. For a given limit, 50% of all sources above the limit were used to train the model, and all remaining sources were used to test the model. Figure 11 shows the relationship between the r magnitude training limit and the r magnitude at which the F1 score drops below a certain fraction of the original value. These figures show that the model maintains its performance up to the training magnitude limit, as expected; however, beyond this limit the model weakens at various rates for the three classes of source. For galaxies the precision is not strongly affected by imposing a magnitude cut in the training, whilst the recall is significantly affected. This indicates that the number of false negatives, in this case galaxies misclassified as stars and quasars, are significant when imposing a magnitude limit on the training set. For quasars the recall is most affected when the r magnitude training limit is below 19. For stars the recall is not significantly affected by imposing a training magnitude limit, whilst the precision is significantly affected.

When imposing the r magnitude training limit at its lowest value of 18, the overall fraction of each class used to fit the model is 7.3%, 3.7% and 33.9% for galaxies, quasars and stars, respectively. Comparing Figure 2 and Figure 10 indicates that on average the F1 scores should still be much higher despite this reduction in the number of training samples. Furthermore, the F1 score as a function of the PSF r magnitude when training on 1% of sources shows the same distribution as that shown in the first row of Figure 7, which used 50% of the data for training. In other words, training on 1% of each class gives a distribution in r magnitude of F1 scores that is similar to training on 50% of each class.

Overall this demonstrates that when limiting the training set to sources below magnitude 18 or 19, the training set does not contain enough examples of each class that are representative out to fainter magnitudes. This degraded performance is alleviated significantly once the training set goes up to and beyond magnitude 19.5, but overall the models cannot be applied to sources at much fainter magnitudes than were included in the training set without resulting in a significantly poorer performance. Furthermore, this decrease in performance will not be quantifiable for a distribution of sources at magnitudes fainter than those in the test dataset. Consequently, we consider that the metrics used to quantify the performance of our machine learning model are only relevant within the magnitude range explored by our training dataset, as shown in Figure 7.

3.4.4. Confidence of classification for individual sources

In the previous section we made approximations in order to specify how the model performs as a function of individual features, reducing the full 10-dimensional space used by the performance metrics. However these feature-specific metrics are still only correct on average within a specific range of feature space, and cannot be used to quantify performance for individual sources. In order to narrow further the range of feature space that a performance metric represents, we implement a nearest neighbour approach for a subset of sources in one and ten dimensions, and we investigate the relation between these metrics and the classification probabilities returned by the random forest for individual sources.

For each source the full set of ten features are transformed into one single feature by using PCA, see Figure 7. 75% of the variance from the full feature set is retained in this...
Fig. 7. Precision, recall and F1 score as a function of: PSF r magnitude (first row), PSF r magnitude error (second row), features transformed into one dimension (third row), and the resolved parameter (fourth row). The shaded area shows a one standard deviation confidence interval calculated per bin from the Wilson score interval. A histogram per class is shown normalised relative to galaxies which peaks at a half. In the last column the classification probabilities from the random forest model for the assigned classes are also shown by dashed lines. Where there are no data to calculate a value, it defaults to one third. Galaxies are in grey, quasars in pink and stars in blue.
Fig. 8. Histogram (left) and cumulative normalised histogram (right) of the random forest classification probabilities using a bin size of 0.005. The 0.99-1 bin contains 73%, 53% and 83% of the correctly classified galaxies, quasars and stars respectively. Correct classifications must have a probability greater than one third and incorrect classifications must have a probability less than one half. In general correctly classified quasars have lower classification probabilities than correctly classified stars or galaxies.

Fig. 9. Random forest classification probabilities plotted as a function of PSF r magnitude. A normalised histogram of the PSF r magnitude is also overlaid, though note that the missed galaxies histogram is multiplied by a factor of 10 to be visible. The 0.99-1 bin contains 73%, 53% and 83% of the correctly classified galaxies, quasars and stars respectively (as detailed in Figure 8).
Fig. 10. Precision, recall and F$_1$ score plotted as a function of PSF $r$ magnitude for galaxies (left), quasars (middle), and stars (right). Each of the coloured lines depicts the upper limit of the $r$ magnitude for the random forest training. Below each of these limits 50% of the dataset was used for training.

Fig. 11. The $r$ magnitude training upper limit plotted against the $r$ magnitude at which the F$_1$ score drops to a fraction of the F$_1$ score (shown in the legend) without a $r$ magnitude training limit. Below this limit 50% of the data were used for training. A linear relationship is maintained for stars (right) since a significant fraction of the training set is still included below the $r$ magnitude limit of 18.5. Galaxies (left) and quasars (middle) show a non-linear jump between $r$ magnitude of 19 and 19.5 for F$_1$ score fractions of 0.9 (galaxies) and less than 0.7 (quasars). This is also seen in Figure 10, likely due to a large change in the fraction of sources trained on.
is high, and only begin to become visible at the edges of the figure where source density drops. The error in the $F_1$ score is calculated using the Wilson score interval, and is only significant when the source density drops at the edges of the plot. In the top right panel of Figure 12 we show the classification probabilities for the same sources. As shown in Figure 8, quasars generally have lower classification probabilities, whilst galaxies and stars have higher values. The probabilities for all classes are positively correlated with the $F_1$ score calculated from the 10,000 nearest neighbours.

Whilst this nearest neighbour method does not fully explore the 1-D feature space (only using 8% of the test dataset), the $F_1$ scores and probabilities are in agreement with those shown in Figure 7 (the final panel of the third row).

We repeat this process using all of the features, finding the 10,000 nearest neighbours for each of 100,000 sources in the original 10-dimensional feature space (bottom row of Figure 12). The error bars on the 1-D feature are now much larger, as a sample of 10,000 nearest neighbours selected in 10-D will have a wider distribution than in 1-D. A broadly similar trend in $F_1$ scores is seen when compared with that for the 1-D case, see the top left panel of Figure 12 and Figure 7. However, we now see a much larger scatter in the $F_1$ scores for each class as a result of exploring more sparsely populated areas of the 10-D feature space. In particular, there are regions in the 10-D feature space where the $F_1$ scores are significantly below the average, most notably for stars and quasars. This occurs because the source density in the 10-D space can drop significantly, even when the corresponding 1-D feature is densely populated with a high $F_1$ score. In the bottom right panel of Figure 12 we show the mean classification probability per source calculated from its nearest neighbours in 10-D. There are significantly more galaxies, quasars and stars with lower probabilities due to the lower source density in some regions of the 10-D feature space. Despite this, the classification probabilities are still positively correlated with the $F_1$ score. In other words, areas of the 10-D feature space which had low $F_1$ scores calculated from their nearest neighbours in 10-D also have low classification probabilities.

Overall, Figure 12 shows that the random forest probabilities are positively correlated with the calculated $F_1$ scores across the full 10 dimensional feature space. This means that even in sparsely populated regions of the 10 dimensional feature space, the random forest probabilities reflect this uncertainty in the classification for individual sources.

Calculating $F_1$ scores in the 10 dimensional space like this is highly computationally intensive. For each of our 111 million photometrically observed sources that we classify in Section 4, finding 10,000 nearest neighbours from the 1.2 million sources in our spectroscopic test dataset, and calculating $F_1$ scores localised to each newly classified source would require significant computational resources. Whilst this would be beneficial when using classification algorithms in higher dimensions that do not return a classification confidence per source, the classification probabilities returned by the random forest provide a more efficient solution. Furthermore, through testing a subset of 100,000 sources we have demonstrated that the random forest probabilities are in agreement with the localised $F_1$ scores. Overall, the random forest probabilities are an effective method to estimate the reliability of a classification, representing the confidence of classifications in that particular area of 10-D feature space.
3.5. Unsupervised clustering with UMAP

In order to interrogate the performance of our machine learning model further, we use a form of unsupervised machine learning. Unsupervised machine learning algorithms have no prior knowledge of target class labels, and attempt to group sources by inferring their similarities in feature space. Here we use the UMAP (Uniform Manifold Approximation and Projection: McInnes et al. 2018) non-linear dimensionality reduction algorithm to explore potential patterns or clustering within our dataset. UMAP is a non-linear dimension reduction algorithm based on manifold learning techniques and topological data analysis. It is designed to preserve structure information on local scales, but is also effective at preserving structure on global scales, particularly as compared to alternative non-linear algorithms such as t-distributed Stochastic Neighbour Embedding (t-SNE: van der Maaten & Hinton 2008). It also has significantly superior scaling performance to t-SNE (McInnes et al. 2018, Figures 5, 6 and 7), allowing for much larger samples sizes and number of dimensions. UMAP can also be semi-supervised or fully-supervised, and allows for metric learning where a model can be applied to unseen data. Furthermore, unlike t-SNE, it returns meaningful distances between clusters of points; however, the dimensions returned in embedded space have no specific meaning, unlike, for example, principal component analysis where the returned dimensions are the direction of greatest variance in the original data.

We first used our training dataset (half of our 2.2 million spectroscopically observed sources), to train a UMAP model embedding the 10 dimensional data into 2 dimensions. In this way, UMAP sees the labels of these sources whilst fitting the embedding which maps them into the 2 dimensional space. Once trained, we applied the model to the training data, the unseen test dataset, and also all of the 2.2 million spectroscopically observed sources together (train and test datasets). In all three cases, the resulting embedding in 2 dimensions is exactly the same. Figure 13 shows all 2.2 million sources embedding into 2 dimensions. Each point has been colour coded by its class label, and where multiple classes are present in a single pixel bin those colours are combined proportionally. The brightness of the pixel represents the number of sources within that bin, and is on a logarithmic scale.

In an unsupervised scheme, we applied UMAP to our test dataset (1.2 million spectroscopically observed sources) without supplying the class labels to the algorithm, and colouring each point by the class label afterwards. The resulting embedding in 2 dimensions is indistinguishable from that shown in Figure 13, showing that UMAP achieves the same result in both supervised and unsupervised scenarios. In a semi-supervised scheme we give UMAP half of the labels for the sources in our test dataset, and the same result is achieved.

In both supervised and unsupervised scenarios, UMAP has effectively separated the data into distinct classes in this 2-dimensional space. Furthermore, it has distributed the classes in a way that represent more than just their class label. The structures present in Figure 13 represent the shape of the spectrum across the SDSS and WISE bands, plus how resolved the sources are. Galaxies (shown in grey) that are predominantly star forming are in the lower left of the plot (termed the blue cloud in colour-magnitude diagrams). From here the galaxies transition into a long thin structure, referred to as the green valley (Salim 2014; Angthopo et al. 2019), which contains galaxies where star-formation has been quenched. There is then a break in the structure where source density drops considerably, before a semi-separate structure is formed at the top of the plot containing old passive galaxies (termed red sequence galaxies). Stars (shown in blue) are arranged much like a Hertzsprung Russell diagram. Red main sequence stars are in the lower right, following multiple tracks up to larger blue main sequence stars, with giants and super-giants further to the left. White dwarfs are shown as a separate cluster of blue beneath the these giants. Quasars (shown in pink) represent a class of source that are a intermediate between galaxies and stars, being mostly unresolved and in an area of colour space between galaxies and stars. Bluer quasars are towards the lower left of the structure, whilst redder quasars are towards the upper right.

When taking a closer view of Figure 13 it is noticeable that each of the dominantly coloured regions is contaminated at a low level by colours from other classes. For example the region dominated by galaxies is contaminated at a low level by stars and quasars (shown by pink and blue colours within the grey). This is most noticeably so for the galaxy region, less so for the quasar region, and barely noticeably so in the star region. This qualitatively reflects the precision and recall for each class presented by the random forest. Stars have the highest precision (few false positives in the blue region of Figure 13), followed by galaxies then quasars. Galaxies have the best recall (few false negatives in the pink and blue region of Figure 13), however stars and quasars have a poorer recall, shown by blue and pink appearing outside of their respective classes.

Whilst we have used UMAP as a qualitative assessment of the class labels, it affirms that using UMAP in an unsupervised, semi-supervised, or fully-supervised scheme is an effective tool in understanding the distribution of sources where labels are not available. In section 4.2 we explore using UMAP on the 111 million un-labelled photometrically observed sources.

4. Classifying new sources

4.1. Applying the random forest model

Using the machine learning model described in Section 3, we classify the 111 293 033 previously unlabelled SDSS photometric sources described in Section 2.2 as either a galaxy, quasar or star. This returns 49 714 482 galaxies, 2 419 610 quasars and 59 158 941 stars.

Figure 15 shows that the average shape of a spectrum across the SDSS and WISE bands for these newly labelled photometrically observed sources is similar to that of the spectroscopically observed sources. Newly labelled galaxies and quasars are fainter in all bands, whilst newly labelled stars show similar mean magnitude in all bands. The resolved, feature effectively distinguishes stars from galaxies for these new sources, see Figure 15, although there are significantly more newly labelled galaxies that are unresolved point sources than were in the spectroscopically observed dataset.

Figure 16 shows the distribution of classification probabilities for the newly labelled sources. The 0.99–1 probability bin is the most populated bin, containing 6 369 632 galaxies (13%), 349 597 quasars (7%) and 44 336 520 stars (75%). There are 34 812 404 galaxies (70%), 766 107 quasars (32%) and 55 290 532 stars (93%) with probabilities greater than 0.9. This is lower than for spectroscopically observed sources (97%, 84%, 95%). The cumulative histogram in the right of Figure 16 shows that newly labelled stars have a similar distribution to those with spectra. However, newly labelled galaxies and quasars have significantly more sources with lower classification probabilities than those with spectra. This indicates that stars are easier to
Fig. 13. 2.4 million spectroscopically observed sources processed with UMAP to reduce from 10 features to two. The resulting two dimensions are plotted, with sources binned per pixel and colours combined proportional to how many of each class are in that pixel bin. Galaxies are shown in grey, quasars in pink and stars in blue. The brightness corresponds to the total source count in that pixel on a logarithmic scale. Whilst the axis have no physical meaning in this two dimensional space, UMAP returns proportionate distances between points and clusters, effectively displaying local and global structures. In this way, red galaxies appear in the top of the plot, connected to green valley galaxies which lead to bluer star-forming galaxies in the lower left. Stars are arranged much like a Hertzsprung Russell diagram, with red main sequence stars in the lower right, following the main sequence up to larger blue main sequence stars, and giants further to the left. White dwarfs are shown as a separate cluster beneath the giants. Quasars represent a class of source that are an intermediate between galaxies and stars, being mostly unresolved in an area of colour space between galaxies and stars. Redder quasars are in the top right, whilst bluer quasars appears in the lower left of the structure.

classify than the other two classes, and quasars are the most difficult, having lower classification probabilities on average.

Figure 17 shows how the classification probabilities vary as a function of the PSF r-magnitude. For stars fainter than PSF r-magnitude 18 the classification probabilities are very high. However, above magnitudes of 18 there are many more stars with lower classification probabilities. This is likely due to the fact that below magnitudes of 18 there are far fewer examples of the two other classes, leading to correctly classified stars showing stronger performance than galaxies and quasars. The middle plot in Figure 17 shows newly labelled quasars, which show a significant drop in classification probabilities below PSF.
With quasars and galaxies having distributions peaked above significantly in favor of quasars at these low magnitudes.

We note that the classification probabilities are always closely most likely stars being misclassified as quasars. In this scenario it is rare to have quasars at magnitudes less than 14, they are second highest probability assigning them as stars. Given that $r$-magnitude 16. These sources tend to be unresolved, with the second highest probability assigning them as stars. Given that it is rare to have quasars at magnitudes less than 14, they are most likely stars being misclassified as quasars. In this scenario we note that the classification probabilities are always closely split between the two classes (stars and quasars), and are never significantly in favor of quasars at these low magnitudes.

There are photometric redshifts available for 47 million sources in our catalogue. Figure 18 shows their distribution, with quasars and galaxies having distributions peaked above $z = 0.1$, and stars having a broader distribution cover lower redshifts. When looking at the classification probabilities for these stars with photometric redshifts, we see no difference to that of the overall distribution of probabilities. In other words, it is unlikely that these stars are all misclassified by our model. It is likely that the SDSS photometric redshift pipeline has fitted these stars with spectra.

4.2. Clustering with UMAP

We applied UMAP to 10% of the 111,344,899 million photometrically observed sources in an unsupervised scheme, shown in the left plot of Figure 19, where sources are coloured by the classification labels assigned by the random forest model. UMAP effectively separates the galaxies and stars as labelled by our random forest model, but does not pick out quasars as a separate cluster. However, quasars are one continuous structure mainly mixed in with the cluster of galaxies. Stars are clustered into various groups with complex structure. Galaxies are clustered as a single group which broadly resembles the global structures displayed in the spectroscopic sources, see Figure 13. The middle plot in Figure 19 includes the 2.2 million spectroscopically observed sources, included without labels, combined with 10% of the unlabelled photometric sources. This unsupervised approach, guided by the inclusion of unlabelled spectroscopically observed sources, improves the clustering, particularly for quasars which are now clearly grouped together and have less overlap with other classes. The right hand plot in Figure 19 includes the same sources as the middle plot, but in this case the labels for the spectroscopically observed sources were explicitly used by the UMAP clustering algorithm. This semi-supervised approach helps UMAP achieve the optimum clustering as defined by the galaxy, quasar, star classification scheme. The galaxy-star separation is quite distinct, although quasars are not as well separated as previously, see Figure 13. Global structures resembling those shown in the spectroscopically observed sources are broadly preserved in this semi-supervised approach, although not as clearly. The majority of the stars can be found in a single continuous cluster, with white dwarf stars shown as an overlapping but distinct structure.

We also implement supervised dimension reduction by training UMAP on 50% of the spectroscopically labelled sources and using the fitted model to transform all unlabelled photometric sources into the resulting two dimensional space. This form of metric-learning allows us to efficiently apply the UMAP transformation derived from 1.2 million spectroscopically observed sources, see Figure 13, to all 111 million photometrically observed sources. This is computationally advantageous compared to running UMAP on 111 million photometric sources in an un- or semi-supervised scheme. The results are shown in Figure 20, where sources are colour coded post-projection with the labels assigned by the random forest algorithm. As expected for this metric learning scheme, the global and local clustering patterns appear the same as those from the spectroscopically observed sources shown in Figure 13. There is a clear separation of the three classes as designated by the random forest algorithm, similar to that seen for the spectroscopically observed sources. Given the significant increase in source density in this figure, regions that were previously unpopulated in Figure 13 are now filled, and there is overlap where each of the three classes intersect. We can see that small clusters of stars are often co-located with the main clusters of galaxies and quasars, but we do not see galaxies or quasars noticeably in the main cluster of stars. Galaxies overlap the beginning of the cluster of quasars, but do not overlap the cluster of stars with any significance.

The metric-learning and un-supervised or semi-supervised schemes implemented with UMAP achieve similar results, but appear to have biases towards different classes. In a semi-supervised scheme, UMAP appears to perfectly separate stars from galaxies, see Figure 19. However, in a metric-learning scheme there is contamination of stars in the galaxy and quasar clusters, e.g. Figure 13. Conversely, the metric-learning scheme is able to more clearly separate quasars into their own cluster, whilst the semi-supervised scheme overlaps quasars significantly with galaxies. Whilst the un-supervised and semi-supervised implementations only use 10% of the photometric sources, whilst the metric-learning scheme uses all photometric sources, the same result is achieved when limiting the metric-learning method to the same subset of sources. With quasars being the minority class in the set of photometrically observed sources, metric learning helps UMAP pick out these rarer sources into their own cluster, at the cost of some contamination from stars.

Overall, the labels assigned to photometrically observed sources by the random forest model are consistent with the clusters derived from UMAP. Furthermore, even when classes overlap in the embedded two dimensional space, the structures visualised in overlapping regions are coherent, see Figure 19, and consistent with the labels from the random forest. This can be interpreted as non-overlapping clusters of sources being distinguished in the higher 10-dimensional space by the random forest, but overlapping in this two dimensional space due to the embedding derived by UMAP. UMAP’s effectiveness in separating clusters increases when including spectroscopically observed sources in an unsupervised scheme, and further
increases in a semi-supervised scheme. The metric-learning scheme has the advantage that it is computationally efficient, and more clearly separates quasars into their own cluster.

4.3. Catalogue description

The catalogue resulting from this work contains just over 111 million labelled sources. These sources all have associated WISE matches but do not have associated spectra in SDSS. We provide the catalogue under a digital object identifier: https://www.doi.org/10.5281/zenodo.3459294. The whole catalogue is in a single Pandas (McKinney et al. 2010) data frame, saved as a Pickle file (.pkl). We also provide it as Pandas data frames per class, which are sorted by classification probability in descending order. The column names are described in Table 2. An extract from the full catalogue is shown in Appendix B for illustration. There are 6 369 632 galaxies (13%), 349 597 quasars (7%) and 44 336 520 stars (75%) with classification probabilities greater than 0.99. There are 34 812 404 galaxies (70%), 766 107 quasars (32%) and 55 290 532 stars (93%) with classification probabilities greater than 0.9.

5. Discussion

5.1. Feature selection

We have used a set of 10 features that are readily available for a large number of sources in order to predict class labels. As random forests compare random sets of features in each decision tree estimator, they effectively calculate colours in different bands naturally throughout the forest. This method helps remove bias and over-fitting, which could arise when only using a few colours to identify objects. Furthermore, utilising all SDSS and WISE wavebands, and the resolved parameter together in this way provides a greater confidence in the classifications over using selected colours alone. We observe that removing absolute magnitude dependence from the model, i.e. taking the difference.
between the photometry from each band and the SDSS $r$-band photometry, resulting in a set of nine total features, did not alter the results.

The feature rankings returned from the random forest show that the $resolved_r$ parameter separates the classes very well. This is also shown by the histograms in the right panel of Figure 5. The next best features are the SDSS $z$ and $i$ magnitudes, which provide the best estimate of the shape of the spectrum across the SDSS bands. Other SDSS and WISE bands provide additional colour information, however the W4 band is provides very little additional information in this classification scheme.

Whilst higher signal to noise colours can be calculated using modelMag, we found that this made no difference to the performance of our classifier. The higher signal to noise colours calculated from modelMag are therefore likely not significant compared to the ensemble approach of biased estimators from the random forest.

SDSS provide a binary classification of star or galaxy depending on their $resolved_r$, parameter being greater or less than 0.145 (Aguado et al. 2019). Due to its simplicity, in many cases this classification is incorrect (for example the middle right panel of Figure 5). Our method, leveraging information from the photometry alongside the $resolved_r$ parameter, provides a more accurate classification label, along with a quantified estimation of its probability.
We have found that the UMAP algorithm is an effective, efficient, and flexible tool for astronomical source classification. Implementing UMAP in a metric-learning scheme where spectroscopically observed sources are used to learn an embedding that is then applied to photometrically observed sources is particularly powerful for separating our target classes. The prior training allows a minority class, in this case quasars, to be picked out by UMAP are consistent with our class labels. Furthermore, our labels assigned from the random forest model show that even when classes overlap in this two dimensional space, the structures picked out by UMAP are consistent with our class labels.

5.2. Class imbalance

In our new labelled photometric catalogue there are 28 times more galaxies, 8 times more quasars and 162 times more stars than in the previously available spectroscopic dataset. It is notable that the ratio between the three classes is different in the spectroscopic catalogue used for training compared to the resulting photometric catalogue. This bias is expected as a result of spectroscopic surveys prioritising particular sources at different magnitude depths. For example there is greater scientific demand for galaxy spectra over those of stars, and a rare occurrence of bright quasars to target. It is therefore expected that the ratio of the three classes will be different in our new photometric catalogue compared to the ratio in the spectroscopic sources training data. Figure 2 demonstrates that the model still performs well when trained on several orders of magnitude fewer sources, and that sub-sampling the galaxy class hinders the model. However, under-sampling any classes in such a way that restricts the distribution in magnitude space does affect the results, as is described in Section 3.4.3. We note that training the random forest model on only 1% or 10% of the spectroscopically observed sources still results in the same ratio of source classifications in the resulting catalogue, though the classification probabilities are slightly lower.

5.3. UMAP

We have found that the UMAP algorithm is an effective, efficient, and flexible tool for astronomical source classification. Implementing UMAP in a metric-learning scheme where spectroscopically observed sources are used to learn an embedding that is then applied to photometrically observed sources is particularly powerful for separating our target classes. The prior training allows a minority class, in this case quasars, to be picked out by UMAP are consistent with our class labels. A further advantage of the metric-learning scheme is that it can be efficiently scaled to embed any number of previously unseen sources using a parallel or distributed computing system, making it a viable algorithm to use with the next generation of surveys from telescopes such as LOFAR, SKA and LSST.

However we note that while we could use UMAP as a pre-processing tool in anticipation of a supervised clustering scheme in two dimensions, for example a nearest neighbour approach, this will always be less accurate than a supervised learning scheme in the higher 10-dimensional feature space. Hence, our use of UMAP is primarily as a diagnostic tool for this dataset, in particular for visualising the effectiveness of the classification labels assigned to photometrically observed sources by the random forest.

5.4. Spectroscopic follow-up

Our new catalogue contains 2.4 million quasars, which is approximately five times more than the 526 356 spectroscopically and visually confirmed quasars previously identified from SDSS (Pâris et al. 2018). Spectroscopic follow-up observations of these new quasars could be done most efficiently by prioritising those with high classification probabilities. In particular there are 349 597 sources which have classification probabilities greater than 0.99. As an increasing number of quasars are spectroscopically confirmed, these can be progressively included into the training set to improve the model’s performance. As discussed in Section 3.4.4, increasing the density of sources in the 10-D feature space by spectroscopically confirming more quasars will increase the random forest classification probabilities for other quasars in similar areas of the feature volume. This would be particularly effective at fainter magnitudes where the training set has lower source counts. An example of this is available in the literature already, where in one particular case a recent spectroscopic survey of hot white dwarf stars confirmed a quasar classification in our new catalogue (Fix et al. 2015).
Fig. 20. UMAP applied to all 111 million photometrically observed sources in a metric-learning scheme, reducing the number of dimensions (features) from 10 to two. UMAP was trained on half (1.2 million) of the spectroscopically observed sources, and then all 111 million photometrically observed sources were embedded into this two dimensional space. Colours were added afterwards for galaxies (grey), quasars (pink) and stars (blue), where their labels were derived from the random forest model trained on the same spectroscopically observed sources. Source are binned per pixel, and colours are combined proportional to how many of each class are in that pixel bin. The brightness corresponds to the total source count in that pixel on a logarithmic scale. UMAP effectively separates the classes and is in a strong agreement with the random forest.

5.5. Future surveys using machine learning

The next generation of telescopes performing large surveys are expected to take observations over many years before completion, using increasingly more efficient and sensitive instruments. Given the cadence of data acquisition and processing, waiting until each survey is complete before building a classification pipeline will not maximise the scientific output from the instrument to the community. Instead, an iterative scheme of re-training and updating an existing model as more data become available will be very important for the telescopes which aim to survey large portions of the sky over many years. Furthermore, the efficiency of such an approach would be improved were models developed specifically with the aim of increasing a training set gradually over the course of the survey. In such schemes it is very important to have metrics that can
assess the reliability of classifications as a function of multiple variables that change over the lifetime of the survey. In this paper we have demonstrated some of these aspects, providing classification metrics as a function of feature space, as well as variables not included as features, e.g: magnitude error and redshift.

6. Conclusions

We have trained a random forest machine learning model on 1.2 million spectroscopically confirmed sources from SDSS in order to classify sources as galaxies, quasars and stars. As features we have used photometry from both SDSS and WISE bands, plus a measure source extension. We have used cross-validation to tune the hyper-parameters in the model, and ensure that the class imbalance in the training data does not affect the performance of the model. Using a test dataset of a further 1.2 million spectroscopically confirmed sources we determine that the random forest achieves F1 scores of 0.991, 0.950, and 0.975 for galaxies, quasars and stars, respectively. Precision, recall and F1 score are also derived as a function of individual features as well as other astronomical parameters, in order to illustrate how the performance of the classifier varies across the three classes. We use the classification probabilities from the random forest as a measure of the likelihood of an individual classification, and we show that this value is in agreement with the F1 score derived from a nearest neighbour search around each source.

We apply the random forest model to 111,293,033 previously unlabelled photometrically observed sources from SDSS. Our model returns 49,714,482 galaxies, 2,419,610 quasars and 59,158,941 stars. Each source has an associated classification probability per class. 49,714,482 galaxies (13%), 349,597 quasars (7%) and 44,336,520 stars (75%) have classification probabilities greater than 0.99. 34,812,404 galaxies (70%), 766,107 quasars (32%) and 55,290,532 stars (93%) have classification probabilities greater than 0.9.

To validate our model we use a non-linear dimension reduction technique (UMAP) in supervised, unsupervised and semi-supervised schemes to reduce the number of features from ten to two, in order that the result can be plotted and the class labels visualised. In all three of these schemes UMAP clearly separates the galaxy, quasar and star classes in two dimensions for the spectroscopic sources.

We applied UMAP to 10% of our new photometric catalogue of labelled sources (11 million sources) in three different schemes: (i) We used UMAP in an unsupervised scheme, where it is in a strong agreement with the star and galaxy class labels assigned by the random forest, but is poor at separating quasars; (ii) we used UMAP in an unsupervised scheme, but also including 1.2 million spectroscopically observed sources (without labels) in the dataset. This gives a superior performance, separating quasars from stars and galaxies more effectively. (iii) We use UMAP in a semi-supervised scheme, including 1.2 million spectroscopically labelled sources in the dataset. This gives a further improvement in the result, clearly separating the classes. When the three classes overlap in any of these three scenarios, the labels from the random forest show coherent structures per class, demonstrating that in the original 10 dimensional space there is a clearer separation of the classes than in the two dimensional space.

We further utilised the ability of UMAP to perform metric-learning by training it on the full set of 10 features for 1.2 million spectroscopically labelled sources, reducing this feature space to two dimensions, and using the resulting model to embed all 111 million unlabelled photometric sources into this two dimensional space. This scheme allows the minority class (quasars) to be clustered more clearly than in other schemes. However, this comes at the cost of more contamination from stars in both the quasar and galaxy areas of the plot, which was not seen in the three scenarios described above. However, one significant advantage of the metric-learning scheme is that it can be efficiently scaled in a parallel or distributed computing system to embed any number of unlabelled sources using a model built from a significantly smaller, partially labelled training data set. This makes it a viable algorithm to run on the next generation of surveys (e.g. LSST, SKA, LOFAR) for both classification and data visualisation.

Acknowledgements. The authors gratefully acknowledge support from the UK Research & Innovation Science & Technology Facilities Council (UKRI-STFC). We also thank Justin Bray for numerous discussions helping to guide this work. We made use of the Python packages Scikit-learn (Pedregosa et al. 2011), Matplotlib (Hunter 2007), Pandas (McKinney et al. 2010) and Datashader: https://www.datashader.org.

Appendix A: Supplementary plots
Fig. A.1. Correct and missed spectroscopically observed sources per class when including the resolved parameter as a feature (top) and when excluding it as a feature (bottom). Both use SDSS extinction corrected PSF magnitudes and WISE magnitudes as features. The $F_1$ score per class increases from 0.988 to 0.991 (galaxy), 0.942 to 0.952 (quasar), 0.974 to 0.978 (star).
Fig. A.2. Histogram of each magnitude feature, per class, for correct and misclassified sources from the random forest model applied to the test dataset of 1.2 million spectroscopically confirmed sources. Galaxies are shown in grey, quasars in pink and stars in blue.
Fig. A.3. Precision, recall and F₁ score plotted as a function of each magnitude feature per class, as derived from the random forest model applied to the test dataset of 1.2 million spectroscopically confirmed sources. In the top row, a histogram of each magnitude feature is also shown as a dotted line per class (normalised to one third) to show the source density. In the bottom row the classification probabilities from the random forest model for the assigned classes are shown by dashed lines. Where there are no data to calculate a value, it defaults to one third. Galaxies are shown in grey, quasars in pink and stars in blue.
Fig. A.4. Examples of galaxies missed as quasars. The spectroscopically observed target is at the centre in a red box, and photometrically observed sources are circled in blue. From top left to bottom right in order of how resolved they are, their SpecObjIDs, resolved $r$ parameters and redshifts are: 527093262050158592, 1.627 ($z=0.162$); 337809312191637504, 1.298 ($z=0.186$); 3087234932888594432, 1.137 ($z=0.094$); 658795790411524096, 1.129 ($z=0.199$); 2976929452103067648, 1.064 ($z=0.303$); 1928687220470867968, 0.919 ($z=0.162$); 5492327033231876096, 0.742 ($z=0.882$); 1780082440165943296, 0.296 ($z=0.280$); 4946128968019320832, 0.090 ($z=0.640$); 7990778382286561280, 0.083 ($z=0.460$); 9339371187510157312, 0.030 ($z=1.078$) and 162140257022233600, 0.022 ($z=0.051$).

Fig. A.5. Examples of galaxies missed as stars. The spectroscopically observed target is at the centre in a red box, and photometrically observed sources are circled in blue. From top left to bottom right in order of how resolved they are, their SpecObjIDs, resolved $r$ parameters and redshifts are: 892985450690537472, 4.685 ($z=0.003$); 3383360912827195392, 2.357 ($z=0.0$); 5584568912597458944, 1.858 ($z=0.724$); 3340604199258843136, 0.386 ($z=0.038$); 2402769422277175296, 0.120 ($z=0.089$) and 686843782766152322, 0.004 ($z=0.656$). Note that the second source is actually a real star, incorrectly labelled by the SDSS classification pipeline.

Fig. A.6. Examples of stars missed as galaxies. The spectroscopically observed target is at the centre in a red box, and photometrically observed sources are circled in blue. The top row are resolved sources with SpecObjIDs and resolved $r$ parameters of: 1142953103180457984, 2.323; 1190170526966900736, 1.609; 2430897686078056448, 1.133; 2660651076583188480, 5.279; 3141329535987902464, 1.498 and 538202045280835456, 1.897. Top bottom row are unresolved sources with SpecObjIDs and resolved $r$ parameters of: 9568057781371510784, 0.268; 9929352360244518912, 0.063; 4543105995102986240, 0.294; 472548831264663808, 0.270; 4728959746027003904, 0.286 and 9197691961909157888, 0.097.
Fig. A.7. Examples of quasars missed as stars. The spectroscopically observed target is at the centre in a red box, and photometrically observed sources are circled in blue. Their SpecObjIDs and redshifts from left to right are: 1578634038246139904 (z = 4.450), 4316783906617532416 (z = 2.257), 6092406673766326272 (z = 3.087), 6434584599746027520 (z = 3.998), 7539033580141256704 (z = 2.995), 7595591358561099776 (z = 1.818).

Fig. A.8. Histogram of each magnitude band for all 111 million newly classified sources. Galaxies are plotted in grey, quasars in pink, and stars in blue.
Table B.1. A sample of 20 galaxies, quasars and stars from our new catalogue of classified sources. The full catalogue can be downloaded using our digital object identifier: https://www.doi.org/10.5281/zenodo.3459294.

| objid | ra         | dec         | psf_u_corr | psf_g_corr | psf_r_corr | psf_i_corr | psf_z_corr | w1     | w2     | w3     | w4     | resolvedr | class_pred | class_prob_galaxy | class_prob_quasar | class_prob_star |
|-------|------------|-------------|------------|------------|------------|------------|------------|--------|--------|--------|--------|-----------|------------|-------------------|-------------------|----------------|
