Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **474,** 5232–5258 (2018) doi:10.1093/mnras/stx2976
Advance Access publication 2017 November 20

Galaxy And Mass Assembly: automatic morphological classification of galaxies using statistical learning

Sreevarsha Sreejith,[1]★ Sergiy Pereverzyev Jr.,[2]★ Lee S. Kelvin,[1,3] Francine R. Marleau,[1] Markus Haltmeier,[2] Judith Ebner,[2] Joss Bland-Hawthorn,[4] Simon P. Driver,[5,6] Alister W. Graham,[7] Benne W. Holwerda,[8] Andrew M. Hopkins,[9] Jochen Liske,[10] Jon Loveday,[11] Amanda J. Moffett,[12] Kevin A. Pimbblet,[13,14] Edward N. Taylor,[7] Lingyu Wang[15,16] and Angus. H. Wright[17]

[1]*Institute for Astro and Particle Physics, University of Innsbruck, A-6020 Innsbruck, Austria*
[2]*Department of Mathematics, University of Innsbruck, A-6020 Innsbruck, Austria*
[3]*Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool, L3 5RF, UK*
[4]*Sydney Institute for Astronomy, School of Physics A28, University of Sydney, NSW 2006, Australia*
[5]*International Centre for Radio Astronomy Research (ICRAR), University of Western Australia, Crawley, WA 6009, Australia*
[6]*Scottish Universities' Physics Alliance (SUPA), School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK*
[7]*Centre for Astrophysics and Supercomputing, Swinburne University of Technology, VIC 3122, Australia*
[8]*Department of Physics and Astronomy, University of Louisville, Louisville, KY 40292, USA*
[9]*Australian Astronomical Observatory, PO Box 915, North Ryde, NSW 1670, Australia*
[10]*Hamburger Sternwarte, Universität Hamburg, Gojenbergsweg 112, D-21029 Hamburg, Germany*
[11]*Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[12]*Department of Physics and Astronomy, Vanderbilt University, Nashville TN 37240, USA*
[13]*E.A. Milne Centre for Astrophysics, University of Hull, Cottingham Road, Kingston-upon-Hull, HU6 7RX, UK*
[14]*School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia*
[15]*SRON Netherlands Institute for Space Research, Landleven 12, NL-9747 AD, Groningen, The Netherlands*
[16]*Kapteyn Astronomical Institute, University of Groningen, Postbus 800, NL-9700 AV, Groningen, The Netherlands*
[17]*Argelander Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, D-53121 Bonn, Germany*

Accepted 2017 November 15. Received 2017 November 15; in original form 2017 April 19

ABSTRACT

We apply four statistical learning methods to a sample of 7941 galaxies ($z < 0.06$) from the Galaxy And Mass Assembly survey to test the feasibility of using automated algorithms to classify galaxies. Using 10 features measured for each galaxy (sizes, colours, shape parameters, and stellar mass), we apply the techniques of Support Vector Machines, Classification Trees, Classification Trees with Random Forest (CTRF) and Neural Networks, and returning True Prediction Ratios (TPRs) of 75.8 per cent, 69.0 per cent, 76.2 per cent, and 76.0 per cent, respectively. Those occasions whereby all four algorithms agree with each other yet disagree with the visual classification ('unanimous disagreement') serves as a potential indicator of human error in classification, occurring in ∼9 per cent of ellipticals, ∼9 per cent of little blue spheroids, ∼14 per cent of early-type spirals, ∼21 per cent of intermediate-type spirals, and ∼4 per cent of late-type spirals and irregulars. We observe that the choice of parameters rather than that of algorithms is more crucial in determining classification accuracy. Due to its simplicity in formulation and implementation, we recommend the CTRF algorithm for classifying future galaxy data sets. Adopting the CTRF algorithm, the TPRs of the five galaxy types are : E, 70.1 per cent; LBS, 75.6 per cent; S0–Sa, 63.6 per cent; Sab–Scd, 56.4 per cent, and Sd–Irr, 88.9 per cent. Further, we train a binary classifier using this CTRF algorithm that divides galaxies into spheroid-dominated (E, LBS, and S0–Sa) and disc-dominated (Sab–Scd and Sd–Irr), achieving an overall accuracy of 89.8 per cent. This translates into an accuracy of 84.9 per cent for spheroid-dominated systems and 92.5 per cent for disc-dominated systems.

Key words: methods: statistical – galaxies: fundamental parameters – galaxies: general – galaxies: structure.

★E-mail: sreevarsha.sreejith@student.uibk.ac.at (SS); sergiy.pereverzyev@uibk.ac.at (SPJ)

1 INTRODUCTION

Galaxies are observed to have a wide variety of forms, from bright massive ellipticals to extended late-type spirals and faint compact dwarfs. One of the first attempts in categorizing galaxies by their visual appearance was proposed by Wolf (1908). These so-called galactic nebulae were arranged according to their shape, size, and distinguishing features. No continuity or transition between these groupings was suggested. As imaging technology improved over the course of the next decade and available data sets grew, new systems for galaxy classification were proposed by many authors (e.g. Jeans 1919; Reynolds 1920). This culminated in the development of the Hubble (1936) sequence or tuning fork. The Hubble tuning fork divides galaxies into early type:[1] typically red and smooth ellipticals; late type: typically blue extended disc-like spirals, both barred and unbarred, and; a bridging population of lenticulars: systems with both a smooth bulge component and an extended yet smooth disc component. Subsequent extensions to the Hubble tuning fork have addressed a number of shortcomings in the initial classification methodology. These include the inclusion of bulgeless spirals (Shapley & Paraskevopoulos 1940), transition lenticulars (Holmberg 1958), rings (de Vaucouleurs 1959), barred lenticulars (Sandage 1961; Sandage, Sandage & Kristian 1975), and dwarfs/irregulars (Sandage & Binggeli 1984). The success of this relatively simple and extensible schema for morphological classification of galaxies has ensured that the Hubble tuning fork remains relevant almost a century later.

Hubble-type (HT) classifications have been used to explore a number of astrophysical phenomena. It was initially noted by Hubble & Humason (1931) that elliptical and lenticular galaxies preferentially favour galaxy cluster environments, indicating a potential environmental dependence on galaxy morphology. Oemler (1974) built upon this work some decades later, showing that the early-type galaxy fraction increases in dense regions. Dressler (1980) conclusively showed how the fractions of elliptical, lenticular, and spiral+irregular galaxies varied as a function of projected galaxy density: the morphology–density relation. He found that dense regions such as galaxy groups and clusters preferentially harbour elliptical galaxies, whilst less dense 'field' regions host lenticular, spiral, and irregular galaxies (See also Smith et al. 2005). This apparent relation between morphology and environment has been further explored in recent years to encompass, amongst others, galaxy mass (van der Wel 2008), star formation (Welikala et al. 2008, 2009), colour (Bamford et al. 2009), the galaxy luminosity function (Kelvin et al. 2014a, see also Baldry et al. 2006), the galaxy stellar mass function (Kelvin et al. 2014b), and galaxy structure (Hiemer et al. 2014).

Precisely how galaxies form and evolve into their various morphological configurations, and the dependence of this on environment, has been the subject of much investigation. Spitzer & Baade (1951) first suggested that merging events between galaxies, more common in dense cluster environments, may be responsible for their transition from a spiral to a lenticular morphology. Toomre (1977) went further, suggesting that elliptical galaxies may also be formed via this merging mechanism (see also White & Rees 1978). In addition to merging, a number of supplementary processes which act to modify the morphology of a galaxy have been proposed, including ram pressure stripping of spiral gas as a galaxy travels through a hot dense intracluster medium (Gunn & Gott 1972), the rapid decline of star formation due to a loss of its hot gas reservoir (strangulation: Larson, Tinsley & Caldwell 1980; Kauffmann, White & Guiderdoni 1993; Balogh, Navarro & Morris 2000; Diaferio et al. 2001), heating of the galaxy caused by rapid encounters with other nearby systems (harassment: Moore et al. 1996), and tidal interactions (Moss & Whittle 2000; Gnedin 2003b, 2003a; Park, Gott & Choi 2008). Obtaining an accurate estimate of galaxy morphology is therefore essential in order to facilitate exploration of the formation and evolution of galaxies.

Contemporary catalogues of galaxy morphology vary in size and classification methodology. Kelvin et al. (2014a, also Moffett et al. 2016) morphologically classify a local volume-limited sample of galaxies taken from the Galaxy And Mass Assembly (GAMA,[2] Driver et al. 2009) survey. Classification is performed via majority observer consensus based on visual inspection of a composite three-colour optical–near-infrared (NIR) image. Three independent expert classifiers are asked a series of questions for each galaxy: is the galaxy spheroid or disc dominated, is the galaxy a single- or multicomponent system, and is the galaxy barred or unbarred. This allows for the galaxy sample to be principally divided into elliptical (E), early-type spiral (S0–Sa), intermediate-type spiral (Sab–Scd), and late-type spiral/irregular (Sd–Irr). Additional barred classes for early- and intermediate-type spirals (SB0–SBa and SBab–SBcd, respectively) are also present. A small subset of 'little blue spheroid' (LBS) galaxies, blue compact systems (∼7.4 per cent), did not fit into this classification hierarchy and were excluded at the top level. This methodology produces accurate classifications yet remains a time consuming exercise, a problem which will only become more acute as future data sets increase in size.

A novel alternative is to enlist the support of the wider astronomy community. The Galaxy Zoo project (Lintott et al. 2008) allows for volunteer 'citizen scientists' to visually classify galaxies via a web interface. The simple and effective design of the website allows for a large number of classifiers to visit each galaxy (typically of the order ∼60), enabling rapid classification of large data sets. However, future facilities such as the Euclid space telescope and Large Synoptic Survey Telescope will probe much larger volumes, providing data sets for several billion galaxies. For these future facilities, morphological classification via visual inspection becomes increasingly prohibitive.

The concept of using automated techniques to quantify galaxy morphologies stem from this 'big data overload' scenario. Moore, Pimbblet & Drinkwater (2006) demonstrated the use of an automated Mathematical Morphology algorithm to achieve classification into ellipticals and late-type spirals using the images from Smail et al. (1997). Their approach was unique in that it had fewer free parameters and that it did not require a classifier to be trained with a machine learning algorithm. Another widely used approach to classify galaxies is by the application of statistical machine learning algorithms. Those that have been used previously used include artificial neural networks (NN), Support Vector Machines (SVM), decision trees, and random forests (RF). They are applied to either galaxy images or to parameters extracted from imaging and spectroscopic data. As part of the Kaggle challenge conducted by the Galaxy Zoo team, Dieleman, Willett & Dambre (2015) presented a convolutional neural network approach (ConvNets) to classify galaxy images. Their algorithm was designed to operate with a training set of 55 420 galaxy images, real-time evaluation set of 6158 images, and a test set of 79 975 images. Huertas-Company

[1] The naming conventions 'early type' and 'late type' refer to the complexity of visual appearance, and do not imply (nor was it meant to imply) an evolutionary sequence (Baldry 2008).

[2] http://www.gama-survey.org

et al. (2015) applied this algorithm to 58 000 (47 700 training, 5300 validation, and 5000 testing) high-redshift galaxy images[3] (median redshift $z \sim 1.25$) from five Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey fields with a result of <1 per cent misclassifications.

Abraham et al. (1996)[4] introduced a new method of discerning between early-, late-, and irregular-type galaxies, the C–A plane, where C stands for the central concentration and A for the rotational asymmetry of the galaxy. This was based on Okamura, Kodaira & Watanabe (1984) and Doi, Fukugita & Okamura (1993), both of whom proposed a strong correlation between the mean concentration index and galaxy morphology. The logged values of these two parameters are plotted in a 2D plane and the separation between the different galaxy populations are obtained by applying linear boundaries. Conselice (2003) expanded upon this method by adding a third dimension, smoothness or clumpiness of the galaxy (represented by S). He was also among the first groups to consider additional morphological types such as dwarf ellipticals, dwarf irregulars, and mergers. For more than three dimensions,[5] this method becomes difficult. Also, it presents some problems when it comes to ground-based, high-redshift data. Graham, Trujillo & Caon (2001) revealed that the concentration parameter, C was unstable in nature due to its high sensitivity to the image exposure depth. Conselice (2003) explains that while it is possible to obtain average values for CAS parameters for data from space-based telescopes (deep *Hubble Space Telescope* data being the example in the paper) up to a redshift $z \sim 3$, the same values for single galaxies will have such high uncertainties that their usage will be quite limited until such a time when deeper and high-resolution imaging can be taken.

Huertas-Company et al. (2007) offered a generalization of the CAS method using SVM. Other examples from literature where a statistical learning technique was used to classify galaxies include Banerji et al. (2010, artificial NN), Owens, Griffiths & Ratnatunga (1996, oblique decision trees), and Gauci, Zarb Adami & Abela (2010, three decision tree algorithms including an RF approach). All these methods use measured parameters as inputs to the classifying algorithms.

The goal of this paper is to explore the viability in using statistical learning methods to produce robust automated HT morphology catalogues for data sets with a greater variety in galaxy types. We have attempted to formulate a general method that will be applicable to small data sets and surveys that do not have access to such a wide variety of parameters as we do. Section 2 details the GAMA (Driver et al. 2009) data set used in this study. Section 3 describes the various statistical learning algorithms under consideration and the application of these algorithms to the data set. Results are shown in Section 4 and the conclusions and future prospects are presented in Section 5. Unless otherwise stated, a standard cosmology of (H_0, Ω_m, Ω_Λ) = (70 km s^{-1} Mpc^{-1}, 0.3, 0.7) is assumed throughout this paper.

[3] The training set actually consists of 8000 galaxies from the Great Observatories Origins Deep Survey-South field, which are rotated randomly three times and over three filters to obtain 58 000 galaxy images (Huertas-Company et al. 2015).

[4] The use of concentration index parameter for galaxy classification can be traced as far back to Shapley & Sawyer (1927) and Morgan (1958).

[5] Please note that dimensions refer to the number of parameters used for the classification process. This terminology is used increasingly when referring to SVM methods where a kernel function (Gaussian in most cases) is applied to non-linearly separable data to project the parameter space into a higher dimension where the data are linearly separable.

2 DATA

In this section, we briefly describe the GAMA survey from which our data sample is taken, the parameters that we have chosen and the justifications for choosing these specific parameters.

2.1 Galaxy And Mass Assembly

GAMA is a project designed to study the low-redshift galaxy population, combining data from eight ground-based and four space-based facilities. It involves both spectroscopic and multiwavelength imaging programmes which are designed to study structures along the scales from 1 kiloparsec (kpc) to 1 megaparsec (Mpc) in the nearby Universe ($z \lesssim 0.25$). The main goal of the GAMA survey is to test and verify the hierarchical structure formation scenario that emerges from the Λcold dark matter cosmological model by measuring the structure growth rate, halo mass function, and star-forming efficiency of galaxies in groups.

The GAMA spectroscopic survey was carried out on the AAOmega multi-object spectrograph on the Anglo-Australian Telesecope (AAT). It includes ~300 000 galaxies with magnitudes down to $r \sim$ 19.8 mag [r being the Galactic extinction corrected Petrosian magnitude in the r band from Sloan Digital Sky Survey Data Release (SDSS DR6); Adelman-McCarthy et al. 2008] spanning an area of ~286 deg^2. The GAMA imaging programme compiles and reprocesses data from a number of other contemporary imaging surveys (see Driver et al. 2009 for details). The reprocessed optical and NIR imaging has a pixel-scale resolution of 0.339 arcsec pixel^{-1} . The master GAMA input catalogue, InputCatAv07, is primarily based on SDSS DR7 (Abazajian et al. 2009) photometry. The majority of the redshifts have been attained as part of the GAMA spectroscopic campaign on the AAT (Hopkins et al. 2013). Additional redshifts are obtained from a number of surveys including the SDSS (Smee et al. 2013), Two-degree-Field Galaxy Redshift Survey (2dFGRS) (Colless et al. 2001), Millennium Galaxy Catalogue (Driver et al. 2005) and others. Full details may be found in Driver et al. (2009) and Baldry et al. (2014).

2.2 Galaxy sample

The galaxy sample used in this paper is from DR2 of the GAMA survey (Liske et al. 2015) which gives spectra, redshifts, and supplementary information regarding 72 225 objects from GAMA DR1 (Driver et al. 2011). Our primary sample consists of 7941 galaxies which have been visually classified into 11 HTs [Kelvin et al. 2014a; Moffett et al. 2016; see Table 1; refer to the VisualMorphologyv02 catalogue in the VisualMorphology Data Management Unit (DMU) for further details], spanning a redshift range of $0.002 \leq z \leq 0.06$.

From our initial sample of 7941 galaxies, we have excluded those objects that are classified as a 'star' or 'artefact' (GAMA HT codes 50 and 60; 374 in number) in the VisualMorphology02 catalogue. We have also excluded an additional 39 objects for which the values were missing for one or more of our chosen parameters. Therefore, the final sample that we apply our statistical learning methods to consists of 7528 objects. Of these, the number of objects of each morphological type are: ellipticals – 856 (11.4 per cent ± 3.3), LBS – 869 (11.5 per cent ± 2.0), early-type spirals – 833 (11.1 per cent ± 0.7), intermediate-type spirals – 1432 (19.0 per cent ± 6.0), and late-type spirals and irregulars – 3538 (47.0 per cent ± 5.9). We computed uncertainties in the sample based on standard deviations of the classifications by the three human classifiers.

Table 1. HTclassifications in the GAMA catalogue and their distribution in our data set. The complete data set consists of 7941 objects from which we remove 374 objects that are visually classified as a 'star' or 'artefact' (GAMA HTs 50 and 60) and 39 objects that do not have valid values for the parameters we have chosen. Of the remaining 7528 objects, we combine the unbarred (11) and barred (12) early-type spirals as well as the unbarred (13) and barred (14) intermediate-type spirals to form two new composite data types 1112 and 1314 (henceforth combinedly referred to as S0–Sa and Sab–Scd, respectively).

GAMA Hubble type code	Galaxy type	Abbreviation	Number of objects (per cent in final 7528 sample)	Of which in training set	Of which in test set
1	Elliptical	E	856 (11.4 per cent ± 3.3)	682 (11.3 per cent)	174 (11.6 per cent)
2	Little blue spheroid	LBS	869 (11.5 per cent ± 2.0)	689 (11.4 per cent)	180 (12.0 per cent)
11	Early-type spirals	S0–Sa	833 (11.1 per cent ± 0.7)	657 (10.9 per cent)	176 (11.7 per cent)
12	Early-type spirals (barred)	SB0–SBa			
13	Intermediate-type spirals	Sab–Scd	1432 (19.0 per cent ± 6.2)	1152 (19.1 per cent)	280 (18.6 per cent)
14	Intermediate-type spirals (barred)	SBab–SBcd			
15	Late-type spirals and irregulars	Sd–Irr	3538 (47.0 per cent ± 5.9)	2842 (47.2 per cent)	696 (46.2 per cent)
50	Artefact	Artefact	374	–	–
60	Star	Star			
–	Incomplete features	–	39	–	–

Notes: Additional HTs of Not Elliptical (10) and Uncertain (70) Morphologies are available in the GAMA VisualMorphology DMU, though these were derived for a different sample via a different method and as such are not used in this study (see Driver et al. 2012 for further details).

Table 2. Parameters chosen from the GAMA catalogues and the derived parameters used for training and testing our algorithms. The parameters in the top panel are those given to the machine learning algorithms as input. Those in the bottom panel are used to derive those in the top panel (with the exception of visual HT), but were not used directly.

Parameter Name	Catalogue column name	Notes	Units	Table	Reference
Stellar mass	logmstar	Logged in catalogue	$\log_{10}(M_{\odot})$	StellarMassesv18	Taylor et al. (2011)
Mass-to-light ratio	logmoverl_i	Logged in catalogue	$\log_{10}(M_{\odot}/L_{\odot,i})$	StellarMassesv18	Taylor et al. (2011)
$g-i$ colour	gminusi	Not logged	mag	StellarMassesv18	Taylor et al. (2011)
$u-r$ colour	uminusr	Not logged	mag	StellarMassesv18	Taylor et al. (2011)
Absolute magnitude	absmag_r	Not logged	mag	StellarMassesv18	Taylor et al. (2011)
Ellipticity	GALELLIP_r	Not logged	no unit	SersicCatSDSSv09	Kelvin et al. (2012)
Sérsic index	GALINDEX_r	Logged	no unit	SersicCatSDSSv09	Kelvin et al. (2012)
Half-light radius in kpc	–	Logged	$\log_{10}$(kpc)	–	–
Kron radius in kpc (semimajor axis)	–	Logged	$\log_{10}$(kpc)	–	–
Kron radius in kpc (semiminor axis)	–	Logged	$\log_{10}$(kpc)	–	–
Half-light radius	GALRE_r	–	arcsec	SersicCatSDSSv09	Kelvin et al. (2012)
Kron radius	KRON_RADIUS	–	units of A_IMAGE or B_IMAGE	ApMatchedCatv06	Hill et al. (2011)
Angular size (semimajor axis)	A_IMAGE	Used to calculate Kron radius in kpc	pixels	ApMatchedCatv06	Hill et al. (2011)
Angular size (semiminor axis)	B_IMAGE	Used to calculate Kron radius in kpc	pixels	ApMatchedCatv06	Hill et al. (2011)
Redshift	Z_TONRY	Used to calculate Kron and half-light radii in kpc	no unit	DistancesFramesv14	Baldry et al. (2012)
Hubble type	HUBBLE_TYPE_CODE	Barred and unbarred counterparts merged for training the algorithms	no unit	VisualMorphologyv02	Kelvin et al. (2014a) Moffett et al. (2016)

2.3 Chosen parameters

The choice of input parameters is crucial for the effectiveness of statistical learning algorithms. We want to recreate the classification process that the human eye would perform upon seeing an image, using parameters extracted from such an image. Ideally we would choose parameters that clearly demarcate the different classes of galaxies. Table 2 lists the parameters that we have chosen from the GAMA data base for each galaxy, the tables they have been taken from and the relevant references.

It is decidedly non-trivial to differentiate between galaxies using only parameters that give similar information, for example, galaxy colour. In Lange et al. (2015), the separation between early- and late-type galaxies in the GAMA catalogue are defined as $u - r = 1.5$ mag and $g - i = 0.65$ mag. Values greater than these would represent the redder (early-type) galaxies, while values less than these would represent bluer (late-type) galaxies. Using only colour to ascribe morphology of a galaxy gives a good general picture of the apparent bimodality of the local galaxy population, but neglects the fact that colour traces star formation, while morphology reflects the dynamic evolution of the galaxy. While they are related, they are not the same. The colour information alone may bias against certain morphological types such as blue ellipticals and red spirals (see fig. 20 of Kelvin et al. 2012). The addition of extra features such as Sérsic index undoubtedly helps provide a more accurate separation of early- and late-type galaxies (Driver et al. 2006; Cameron et al. 2009).

Our objective has been to choose a broad range of parameters that will allow us to successfully morphologically classify galaxies with minimal failures. We have been careful to select astrophysically meaningful parameters that denote different aspects of the physicality of a galaxy. As listed in Table 2, we have parameters that are known to directly trace galaxy morphology (Sérsic index, stellar mass, and colour), parameters that trace galaxy morphology indirectly (mass-to-light ratio) and parameters that are based on galaxy structure (Kron radius, ellipticity, half-light radius, and absolute magnitude). We have attempted to remove the effects of redshift on all the chosen parameters. We also note that in this work, we have not accounted for the errors in the chosen set of parameters.

The total stellar mass, mass-to-light ratio, absolute magnitude, and $g - i$ and $u - r$ colours are taken from the table StellarMassesv18 in the GAMA DMU Stellar Masses (Taylor et al. 2011). Total stellar masses have been derived using stellar population synthesis (SPS) modelling using Bruzual and Charlot models (Bruzual & Charlot 2003) assuming a Chabrier initial mass function (Chabrier 2005). SDSS and VISTA-VIKING photometry have been used for this calculation (roughly equivalent to rest-frame $u - Y$). The mass-to-light ratio has been calculated using the SDSS rest-frame i band. The $g - i$ and $u - r$ colours are rest-frame colours using AB photometry that has been k-corrected to redshift $z = 0$ calculated from the spectral energy distribution (SED) fit. Together, these colours provide a wide wavelength baseline. Absolute magnitude has been calculated using the rest-frame r band from the best SPS SED fit.

Ellipticity, Sérsic index, and half-light radius have been taken from the table SersicCatSDSSv09 in the DMU Sérsic Photometry (Kelvin et al. 2012). These are based on 2D single Sérsic function fits to SDSS r-band images.

We obtained Kron radii in arcseconds by multiplying the Kron radius with the angular sizes in semimajor and minor axes and the angular resolution of the main GAMA imaging data set (0.339arcsec pixel^{-1}). These values were converted into kpcs using flow-corrected spectroscopic redshifts from the catalogue DistancesFramesv14 (Baldry et al. 2012).

We use morphology for training purposes and to test the robustness of our algorithms. We also note that our parent sample (Kelvin et al. 2014a; Moffett et al. 2016) is magnitude limited ($M_r < -17.4$ mag) and we do not expect it to be overly sensitive to dwarf galaxy populations. The complete list of parameters that we have used for training and testing are given in Table 2.

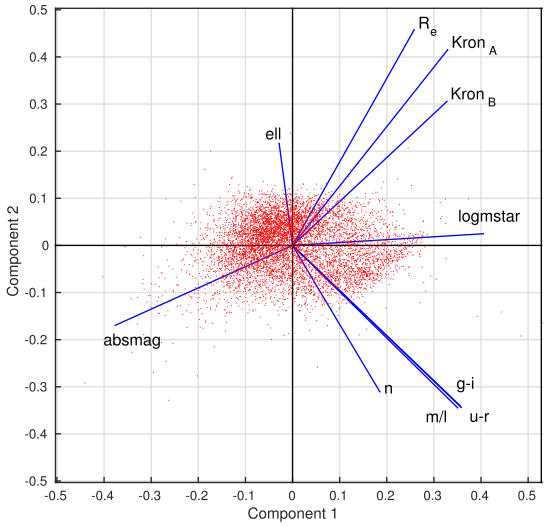

Figure 1. Results of PCA performed on the selected parameters to determine their impacts on the classification process. The component labels correspond to the parameters given in Table 2 in the following manner: ell = ellipticity; R_e = half-light radius in kpc; $\mathrm{Kron_A}$ = Kron radius in kpc (major axis); $\mathrm{Kron_B}$ = Kron radius in kpc (minor axis); logmstar = stellar mass; g-i = $g - i$ colour; u-r = $u - r$ colour; m/l = mass-to-light ratio; n = Sérsic index; absmag = absolute magnitude. Please see Table 2 for more details. The analysis was performed using the MATLAB function `pca`.

2.4 Principal component analysis

We perform principal component analysis (PCA, Pearson 1901) on the parameters that we have chosen from the GAMA catalogues (see Section 2.3, Table 2). PCA is one of the methods by which parameters are generally chosen for functions such as classification. In our case, we had already defined the criterion for choice of parameters as their distance independence or the possibility of removal of their distance dependence. Therefore, our PCA is a secondary method, to see statistically, the impact each parameter has on the classification process. It was done using the MATLAB function `pca`. Approximately 86 per cent of the variability in our parameters is contained in Components 1–3 of PCA. For visualization convenience, we have plotted the first two components in Fig. 1.

Of the two plotted components, Component 1 contains ∼57 per cent of the variance of the parameters and Component 2 contains ∼17 per cent. Both stellar mass (logmstar) and absolute magnitude (absmag) have a significant impact on Component 1, but a smaller contribution towards Component 2. The parameters $g - i$ (g–i) and $u - r$ (u–r) colours and mass-to-light ratio (m/l) have very similar contributions to both the components, and are therefore redundant to a great extent.

Of the other parameters that we have chosen, Sérsic index (n), Kron radii ($\mathrm{Kron_A}$ and $\mathrm{Kron_B}$), and half-light radius (R_e) seem to have significant contributions towards both Components 1 and 2, thereby representing sizeable variability in the data set. Ellipticity (ell) seems to be the one with the least variance among our parameters. A detailed analysis of how much each parameter affects the classification process is given in Section 4.

2.5 Data preprocessing

Classes 12 and 14 are the barred counterparts of classes 11 and 13. Their numbers are low in our sample, at 80 and 195, respectively. A potential reason for this, as noted in Kelvin et al. (2014a) is that

Figure 2. A sample of galaxies classified as elliptical (type 1, E) in the GAMA visual morphology catalogue. Postage stamps are log scaled, span an area of 3 × Kron radius of each galaxy, and are ordered from top-left to bottom-right by increasing stellar mass. Overlaid on each galaxy image are: (top left) the GAMA CATAID of the galaxy; (top right) the numeric HT codes indicating the predicted classification as determined by the SVM, CT, CTRF, and NN classifiers, respectively; (bottom left) the total stellar mass in units of $\log_{10}(M_{\odot})$, and; (bottom right) the flow-corrected spectroscopic redshift of the galaxy. The row-wise median physical scales for these galaxies in kiloparsecs are 5.5, 5.4, 7.4, 7.5, and 2.9.

there were noticeable disagreements among the classifiers about the presence of bars in these systems. Another reason could be that, for edge-on systems, it is impossible to verify the presence of bars and therefore they would be classified as unbarred. Due to the relatively low numbers of galaxy systems hosting bars in our sample, we opt to merge the barred classes with their unbarred counterparts. We merge the classes 11 and 12 (S0–Sa and SB0–SBa) to form a new class 1112. Likewise, we merge classes 13 and 14 (Sab–Scd and SBab–SBcd) to form a new class 1314. This simplifies the classification problem, albeit marginally. The machine learning classifier that we formulate concentrates on predicting the GAMA Hubble types 1, 2, 1112, 1314, and 15. Figs 2–6 show examples of each galaxy type from our final sample. They are created using SDSS *g*-, *r*-, and *i*-band imaging by the GAMA Panchromatic Swarp Imager tool.[6]

[6] http://gama-psi.icrar.org/psi.php

Figure 3. As Fig. 2, but for LBS (type 2) galaxies. The row-wise median physical scales for these galaxies in kiloparsecs are 4.6, 5.1, 4.0, 5.0, and 19.0.

Each image spans a diameter equivalent to 3 × Kron radius of the galaxy in arcseconds, and is log scaled.

To construct and evaluate classifiers using statistical learning methods, the data sample is randomly split into training and test sets. The training set is used for constructing classifiers, containing 80 per cent of the data sample. The test set is used for the evaluation of the classifiers' prediction abilities, containing the remaining 20 per cent of galaxies. In our case, the training and test sets contain 6022 and 1506 galaxies, respectively. We consistently use the same training and test sets for all considered statistical learning methods described in Section 3. The data are normalized before training, i.e. we centre each parameter at its mean value, and scale it to have unit standard deviation. The distribution of HTs for the full data sample, training, and test subsets are presented in Table 1.

3 METHODS

In this section, we outline the galaxy classification problem in the context of statistical learning. We also describe the methods that we apply to solve this classification problem.

3.1 The classification problem

We consider the parameters of a galaxy to be components of a multidimensional vector $\boldsymbol{x} = \left(x_1, x_2, \ldots, x_p\right)^\top \in \mathbb{R}^p$, where $(\cdot)^\top$ denotes the transpose of a vector or matrix. Thus, $\boldsymbol{x}$ is a $p \times 1$ column vector. In our case $p = 10$, and we use the parameters described in Table 2.

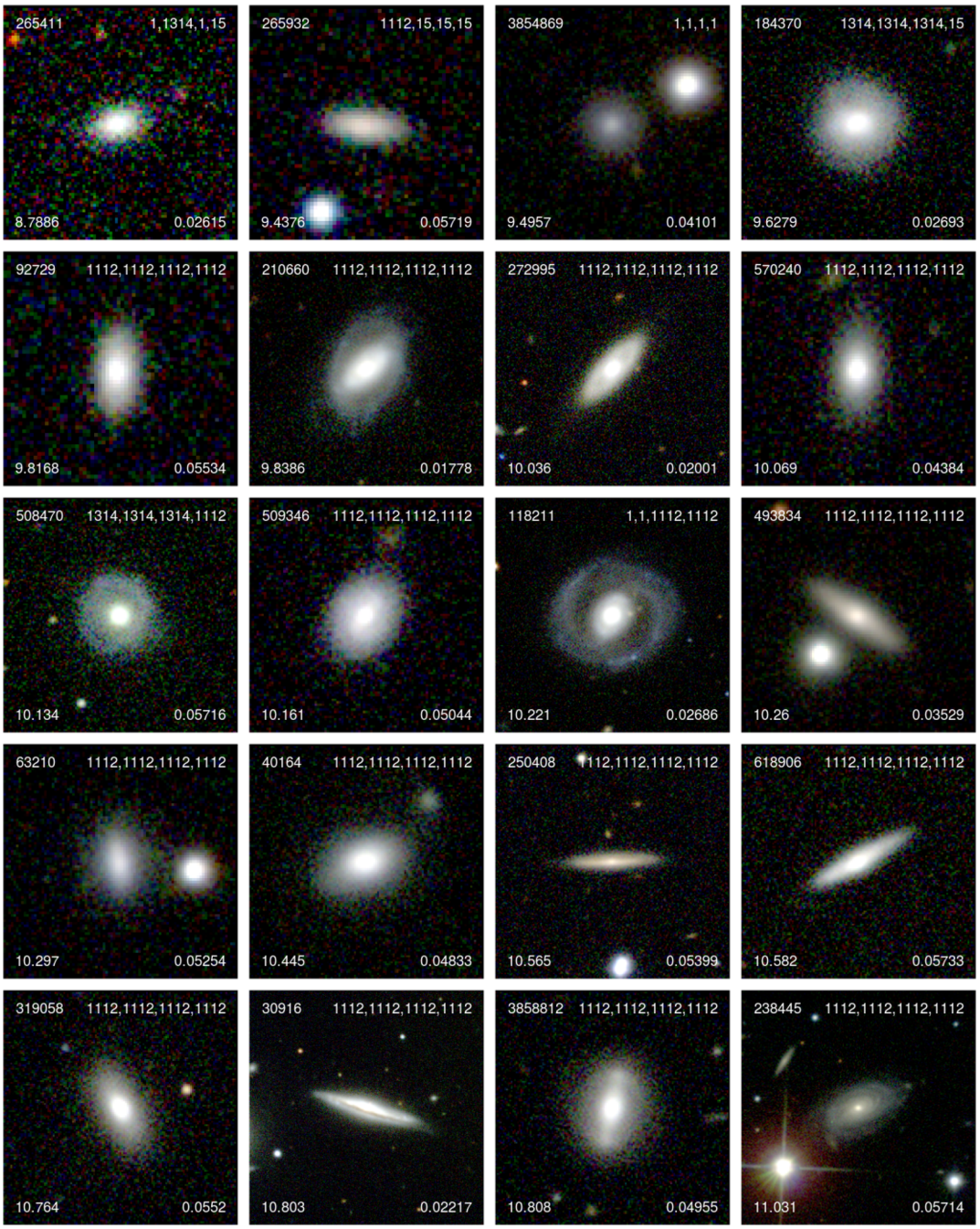

Figure 4. As Fig. 2, but for early-type spiral (type 1112, S0–Sa) galaxies. The row-wise median physical scales for these galaxies in kiloparsecs are 15.9, 24.3, 17.4, 13.5, and 11.8.

In the context of statistical learning, the vector space $\mathbb{R}^p$ is often called *feature space*, the elements $\boldsymbol{x} \in \mathbb{R}^p$ are called *feature vectors*, and the components x_i of the feature vectors are called *features*. The feature vector $\boldsymbol{x}$ belongs to one of the T classes. For convenience, we label the classes as $1, 2, \ldots, T$. In our case $T = 5$, and the classes correspond to the considered HTs as $\{1, 2, 1112, 1314, 15\} \mathrel{\widehat{=}} \{1, 2, 3, 4, 5\}$. Let $y \in \{1, 2, \ldots, T\}$ denote the class label of $\boldsymbol{x}$.

Suppose that there is an *ideal classifier* $f^* : \boldsymbol{x} \mapsto y$ that for each feature vector $\boldsymbol{x}$ assigns its true classification y. A statistical learning method aims to construct a classifier $f : \boldsymbol{x} \mapsto y$ that approximates f^*. For this purpose, statistical learning methods use observational data of the pairs $(\boldsymbol{x}_i, y_i)$ that contain feature vectors $\boldsymbol{x}_i$ for which the corresponding class y_i is known. A set made up of such pairs $(\boldsymbol{x}_i, y_i)$ is called the *training set*, and we denote it as $\mathcal{Z} = \{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, N\}$.

Every statistical learning method consists of a family of classifiers f that depends on certain parameters. Using a learning procedure, a particular classifier is chosen from this family based on the classifier's behaviour on the training data set. The selection is typically done such that the classification is well predicted on the training set, i.e. $f(\boldsymbol{x}_i) \approx y_i$, so as to give low training errors. The quality of the

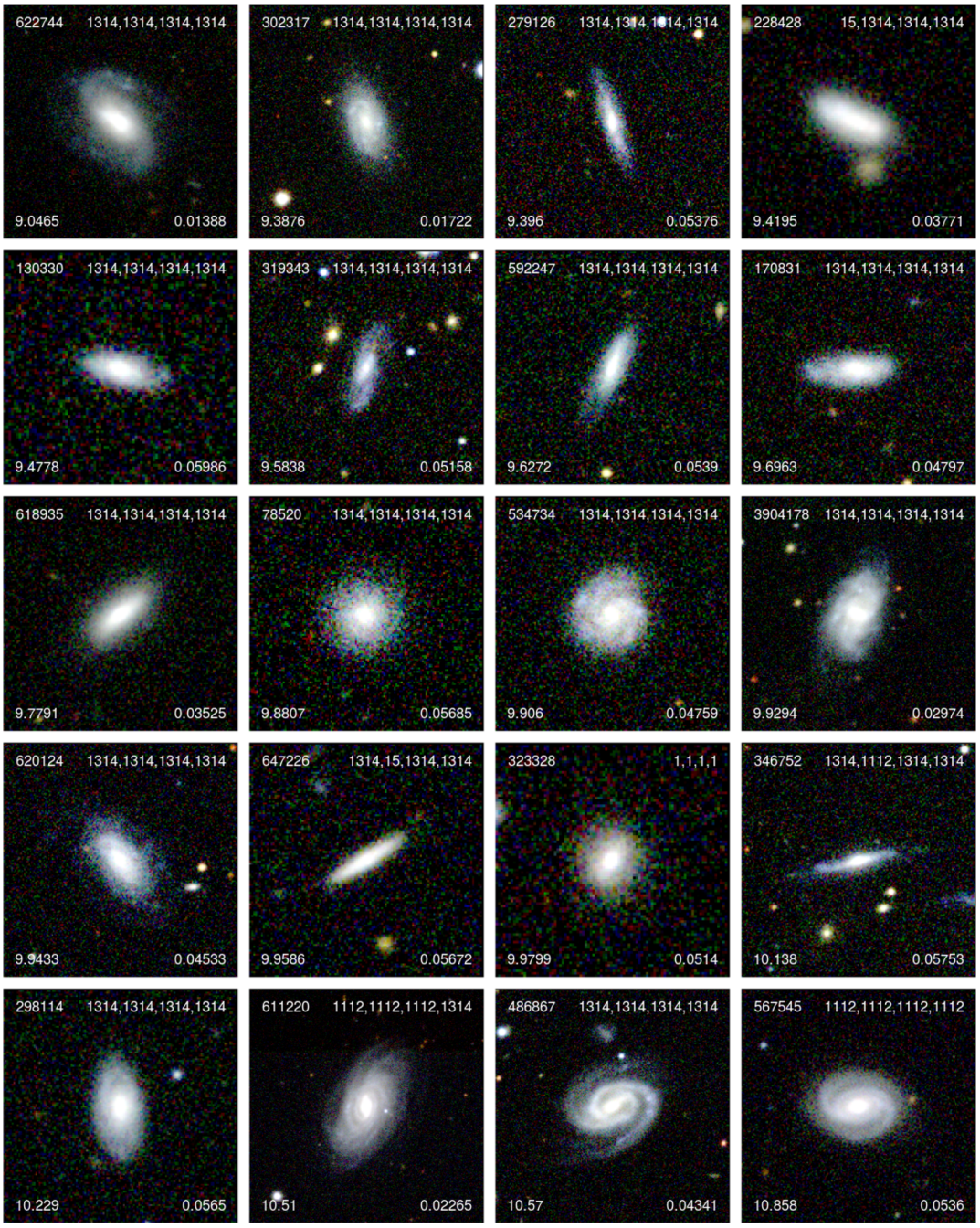

Figure 5. As Fig. 2, but for intermediate-type spiral (type 1314, Sab–Scd) galaxies. The row-wise median physical scales for these galaxies in kiloparsecs are 12.5, 21.8, 12.4, 17.5, and 18.1.

classifier is then evaluated on the test set, where the classification is known. The data of the test set are not used for constructing the classifier. Thus, the performance of the classifier on the test set can be seen as an estimation of its performance on sets with unknown classification.

The methods that we consider here for classifying galaxies are: SVM, Classification Trees (CT), Classification Trees with Random Forest (CTRF), and NN. We have used the realization of these methods in MATLAB R2014b. The outputs provided by the algorithms that we have formulated are multi-class labels, denoting which galaxy type the algorithms deem the galaxy to be of. They are described in detail in the following subsections.

3.2 Support Vector Machines

The SVM method was originally designed for binary classification (Cristianini & Shawe-Taylor 2000; Hastie, Tibshirani & Friedman 2009, chapter 12). In this method, for each feature vector $\boldsymbol{x}$, there is a class label $z \in \{-1, 1\}$. Therefore for each $\boldsymbol{x}_i$ in the training set, the corresponding class is z_i. The details of the structure

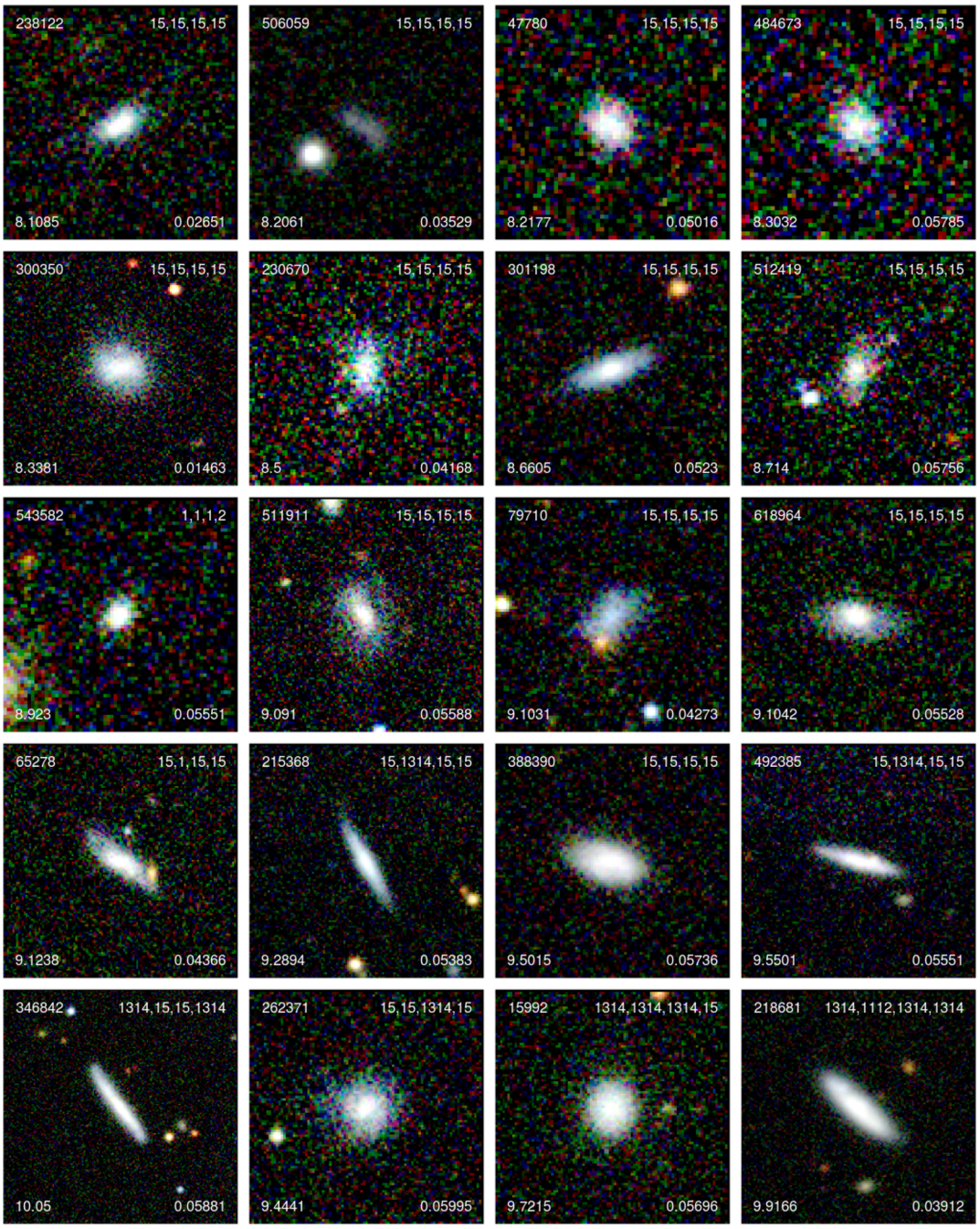

Figure 6. As Fig. 2, but for late-type spiral and irregular (type 15, Sd–Irr) galaxies. The row-wise median physical scales for these galaxies in kiloparsecs are 28.9, 9.6, 8.1, 12.4, and 17.0.

and definitions of the SVM classifier that we employ are given in Appendix A1.

We use the MATLAB function `svmtrain` for constructing SVM classifiers. For computing the result $f(\boldsymbol{x})$ of the SVM classifier f, function `svmclassify` has been used.

In order to use SVM for multiclass classification, the multiclass classification problem is reduced into a series of binary classification problems. For this purpose, we consider a tree structure approach (Campbell 2001). We propose a tree formed by the binary classifiers C_{15}, $C_{\rm sp}$, $C_{\rm E}$, and $C_{\rm a}$ as depicted in Fig. 7. This tree structure is inspired by the distribution of HTs in our data set represented in Table 1. Here, C_{15} is the binary classifier that classifies a galaxy as HT 15 or not. $C_{\rm sp}$ then classifies into spirals and not spirals. Further classification is done by $C_{\rm E}$ into HT 1 (E) or HT 2 (LBS). $C_{\rm a}$ splits the output of the $C_{\rm sp}$ binary classifier into HTs 1112 and 1314. All the binary classifiers in this tree structure are constructed with the SVM method. At each binary classifier, the data are split by roughly 50 per cent.

3.3 Classification Trees with hyper-rectangular partitions

In the CT method, the feature space is partitioned into a set of hyper-rectangular regions $R_{\rm m}$ (Breiman et al. 1984; Hastie et al. 2009, chapter 9). An example of such a partition is presented in Fig. 8.

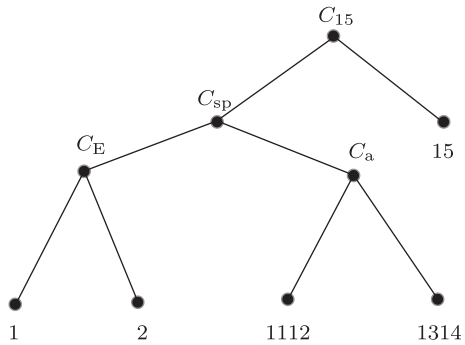

Figure 7. The binary CT determined for the SVM method. The classifier C_{15} classifies a galaxy as HT 15 or not. Then, $C_{\rm sp}$ classifies into spirals and not spirals. Further classification is done by $C_{\rm E}$ into HT 1 (E) or HT 2 (LBS). $C_{\rm a}$ splits the output of the $C_{\rm sp}$ classifier into HTs 1112 and 1314. All the binary classifiers in this tree structure are constructed with the SVM method.

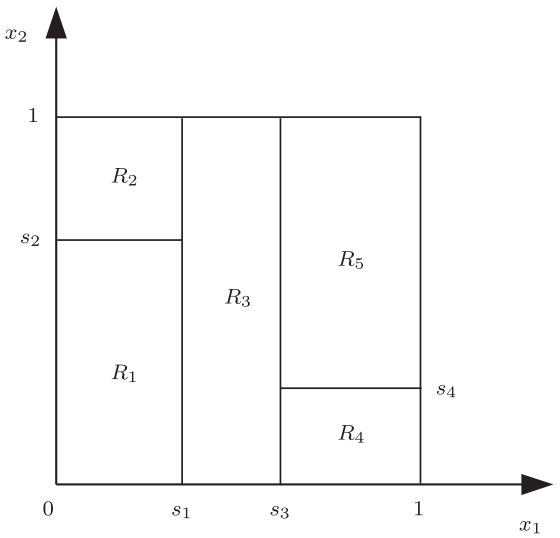

Figure 8. Illustrative example CT method using hyper-rectangular partitions. This unit square is successively split ($s_1 - s_4$) into five nodes R using the two features x_1 and x_2.

The goal of this method is to make the partitions such that each region $R_{\rm m}$ contains training feature vectors that belong only to one class, say $k_{\rm m} \in \{1, 2, \ldots, T\}$, or at least the majority of the training feature vectors in $R_{\rm m}$ is from one class $k_{\rm m}$. Then, for each feature vector $\boldsymbol{x}$, the CT classifier identifies a region $R_{\rm m}$ that contains $\boldsymbol{x}$, and then assigns $k_{\rm m}$ as the predicted class for $\boldsymbol{x}$. The method is discussed in detail in Appendix A2.

The CT partitioning can also be represented by a binary tree, i.e. the partition presented in Fig. 8 can be represented by the tree in Fig. 9. The top node of the tree, which is called root, represents the complete feature space. Feature vectors that satisfy the condition $x_1 < s_1$ are assigned to the next lower node on the left, while the other feature vectors are assigned to the next lower node on the right, and so on. The nodes at the bottom of the tree, which are called terminal nodes or leaves, correspond to the regions of the final partition of the feature space: $R_1, R_2, \ldots, R_5$.

The node splitting is recursively repeated for the new nodes. The node is not split if any of the following conditions is satisfied:

(i) The node is pure.

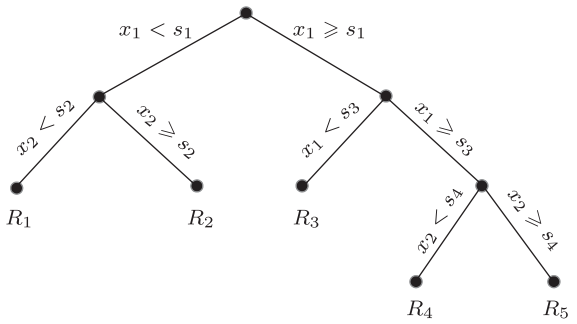

Figure 9. A binary classification tree determined for the CT method as applied to the example unit square shown in Fig. 8.

(ii) The node contains less than a certain number (standard value adopted here is 10) of training feature vectors.

(iii) Any node splitting gives new nodes that contain less or equal to a certain number (standard value adopted here is 0) of training feature vectors.

(iv) If a certain number of nodes (the default value for the MATLAB function that generates the node splitting is $N - 1$) are created.

For our work, we constructed the CT classifier using the MATLAB function `fitctree` and the function `predict` was used for computing the result of the CT classifier. In the constructed CT classifier for our data set, a full description of the derived nodal splits becomes increasingly complex beyond the first leaf. Therefore, we describe the splits which were determined up to and including the first leaf only. The splitting feature in the top node (i.e. at the root of the constructed tree) is x_1 which corresponds to the stellar mass of a galaxy. The split point for this feature was determined to be $\log M = 9.276$. The next leaf node (in the regime $x_1 < 9.276$) has the splitting feature x_6, which is the half-light radius, with the split point determined to be $\log R_{\rm e} = 0.0514$. The alternative node (i.e. the galaxies in the regime $\log M \geq 9.276$) has the splitting feature x_8, which is $u - r$ colour, with the split point $u - r = 1.842$.

The structure of the classifier in the CT method is quite simple. Notably, no arithmetic operation is used for estimating the class of the feature vector $\boldsymbol{x}$. Only a comparison between numbers is used. Therefore, the evaluation of the result of the CT classifiers is very fast, which is a distinct advantage of this method.

However, CT classifiers are known to have the following drawback. $f(\boldsymbol{x}_{\rm i})$ can be in a good agreement with $y_{\rm i}$, but outside the training set, the predictive performance of the CT classifier may be rather poor. This phenomenon is called *overfitting*. To overcome this drawback, the idea of *RF* has been proposed (Hastie et al. 2009, chapter 15; Breiman 2001). This leads to the CTRF method that we explore in the next subsection.

3.4 Classification Trees with Random Forest

The essential idea of the CTRF method is to improve the performance of a single CT by averaging over several differently trained CTs. In order to achieve this, a certain number of samples are created by random sampling *with replacement* from the training set. The sampling is done using uniform distribution, where each sample is of the same size as the original training set. By using sampling with replacement, any element of the training set can be selected more than once for the same random sample. More details on this process are given in Appendix A3.

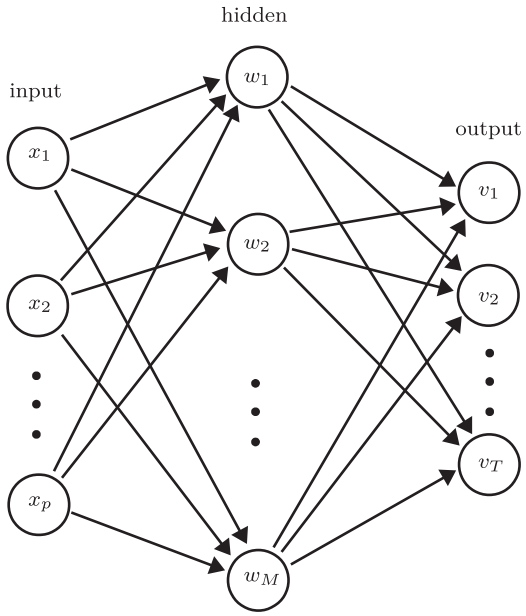

Figure 10. A network diagram for the single hidden layer feed-forward NN.

Each CT classifier in a RF is trained on a different sample of the training data. Moreover, the use of the modified CT learning algorithm, namely the use of random subsets of the features, ensures the *decorrelation* between the constructed CT classifiers. This means that the tree structure of the involved CT classifiers differ from one CT to another. These two properties allow the combination via majority vote of the CTs in the RF to correct the overfitting of each CT classifier. For building our CTRF classifier, we used the MATLAB class `TreeBagger`, and the function `predict` was used for calculating the outcome of the CTRF classifier.

The choice of the number of samples B in RF can be done by observing the *out-of-bag error*. This error is the mean prediction error on each training example using only the CT classifiers that did not have this example in their training sample (Hastie et al. 2009, p. 593). In our case, we observed that this error stabilizes for $B = 100$, and therefore, we used this number for our CTRF classifier.

3.5 Single hidden layer feed-forward Neural Networks

The last statistical learning method that we consider is NN (Hastie et al. 2009, chapter 11). This is a classification method inspired by the central nervous system or biological NN of animals. In comparison to the other mentioned methods, NN constructs classifiers with a more complicated mathematical structure, and the algorithms for constructing NN classifiers are more complex. However, a typically good performance of the NN classifiers outside the training sets makes them very popular.

An NN consists of units that are organized in layers. Typically, a network diagram, such as in Fig. 10, is used to represent an NN. In this work, we implement the most widely used NN ensemble called the *single hidden layer feed-forward NN*. It consists of three layers: the input layer, hidden layer, and output layer.

The units in the input layer correspond to the features x_i. The kth unit v_k in the output layer models the probability for the feature vector to belong to class k. The units in the hidden layer w_m, $m = 1, 2, \ldots, M$, can be seen as additional features that are derived from the features x_i. The structure of the NN that we have considered is explained in more detail in Appendix A4.

For defining our NN classifier, we used the MATLAB function `patternnet`. Then, the weights of the NN classifier were determined using the function `train`, and the evaluation of the result of the classifier was performed. We consider values for the number of units in the hidden layer M in the interval [10,500] and examine the performance of the corresponding NN classifiers on the so-called validation set. For this set, we randomly sample 15 per cent of the elements in the training set. These elements were not used for training the NN classifiers. We find that the True Prediction Ratio (TPR) for the validation set increases as a function of M; however, the relative increase in TPR significantly diminishes as we tend towards larger values of M. We therefore adopt $M = 500$ as the optimal trade-off between classification accuracy and computational complexity of the NN classifier.

4 RESULTS

The CT, CTRF, SVM, and NN codes are run using the parameters shown in Table 2. Fig. 11 shows the classification success rate for each morphological type considered in addition to the total sample ('all'). Galaxy populations are arranged along the x-axis, as indicated. Classification success rate is characterized by the parameter TPR shown on the y-axis. TPR (y-axis)[7] represents the quality measure of the classifiers. It is defined as the ratio of the number of correctly classified galaxies to the total number of galaxies considered. The TPR for the machine learning algorithms CT, CTRF, SVM, and NN are represented by the colours yellow, green, pink, and blue, respectively, for each morphological type. As can be seen, the morphological-type Sd-Irr (Type 15) typically returns the highest success ratio at ∼90 per cent. The morphological-type Sab–Scd (Type 1314) returns the lowest average success ratio, typically in the range ∼55 per cent. Potential reasons for this are discussed in detail in Section 5, but principally revolve around the idea that our algorithms in their current configuration may be more suited to classify single component rather than more complex multicomponent systems. The overall average success rate across all morphological types is found to be ∼76 per cent, with the notable exception of the CT method (see Table 3).

Classification errors can be also characterized using a *confusion matrix*, $\left(a_{ij}\right)_{i,j=1}^{T}$. The entry of this matrix a_{ij} in the ith row and jth column is the number of galaxies from the class j that are classified as the class i by the classifier.

Note that the above considered quality measure TPR of a classifier for the class j can be calculated using the confusion matrix $\left(a_{ij}\right)_{i,j=1}^{T}$ of this classifier:

$$\mathrm{TPR_j} = \frac{a_{jj}}{\sum_{i=1}^{T} a_{ij}}.$$

This quality measure is also known under the names *true positive rate* or *recall*.

The TPR of a classifier for all classes is calculated as

$$\mathrm{TPR_{all}} = \frac{\sum_{j=1}^{T} a_{jj}}{\sum_{i,j=1}^{T} a_{ij}}.$$

[7] Here onward, this parameter is used interchangeably with accuracy of classification (Sokolova & Lapalme 2009).

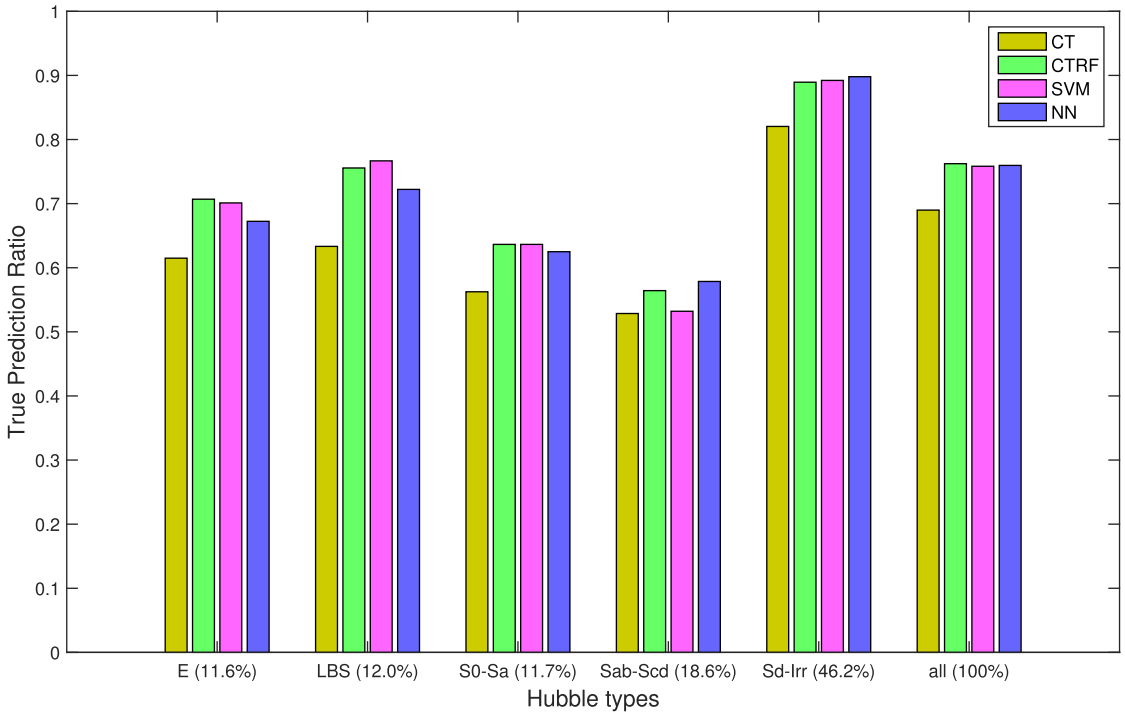

Figure 11. Histograms showing the TPRs from panel 1 of Table 3. The different HTs in our sample are represented on the *x*-axis and the TPR values for each type as obtained by the four statistical learning algorithms are shown on the *y*-axis. The percentage of galaxies of a certain type are shown in brackets next to the HT codes.

Table 3. TPRs in percentages for the classifiers obtained by the methods considered in Section 3 on the test set are given in panel 1. Panel 2 represents the results of binary classification using CTRF method. The galaxy types E, LBS, and S0–Sa are collectively considered as spheroid-dominated systems and Sab–Scd and Sd–Irr as disc-dominated systems.

HT	E	LBS	S0–Sa	Sab–Scd	Sd–Irr	All
	1	2	1112	1314	15	
CT	$61.5^{+3.5}_{-3.8}$	$63.3^{+3.4}_{-3.7}$	$56.3^{+3.7}_{-3.8}$	$52.9^{+3.0}_{-3.0}$	$82.0^{+1.4}_{-1.6}$	$69.0^{+1.2}_{-1.2}$
CTRF	$70.7^{+3.2}_{-3.7}$	$75.6^{+2.9}_{-3.5}$	$63.6^{+3.5}_{-3.8}$	$56.4^{+2.9}_{-3.0}$	$88.9^{+1.1}_{-1.3}$	$76.2^{+1.1}_{-1.1}$
SVM	$70.1^{+3.2}_{-3.7}$	$76.7^{+2.9}_{-3.4}$	$63.6^{+3.5}_{-3.8}$	$53.2^{+3.0}_{-3.0}$	$89.2^{+1.1}_{-1.3}$	$75.8^{+1.1}_{-1.1}$
NN	$67.2^{+3.4}_{-3.7}$	$72.2^{+3.1}_{-3.6}$	$62.5^{+3.5}_{-3.8}$	$57.9^{+2.9}_{-3.0}$	$89.8^{+1.0}_{-1.3}$	$76.0^{+1.1}_{-1.1}$
	Spheroid-dominated			Disc-dominated		All
CTRF	$84.9^{+1.4}_{-1.7}$			$92.5^{+0.8}_{-0.9}$		$89.8^{+0.7}_{-0.8}$

In addition to the TPR, another useful characteristic of the classifier performance is the *Positive Predictive Value* (PPV) or *precision*. It is calculated for the class j using the confusion matrix $\left(a_{ij}\right)_{i,j=1}^{T}$:

$$\mathrm{PPV_j} = \frac{a_{jj}}{\sum_{i=1}^{T} a_{ji}}.$$

Another important characteristic is the *F-score* of the classifier. For the class j, it is defined as the harmonic mean of $\mathrm{TPR_j}$ and $\mathrm{PPV_j}$:

$$\mathrm{F_j} = \frac{2 \cdot \mathrm{TPR_j} \cdot \mathrm{PPV_j}}{\mathrm{TPR_j} + \mathrm{PPV_j}}.$$

The confusion matrices and the mentioned performance characteristics of the considered classifiers are presented in Tables 4–8. The actual classification is given in the columns and the

Table 4. Confusion matrix and performance characteristics for five galaxy classes for the SVM classifier.

SVM classification	Visual classification				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
E	122	12	35	9	7
LBS	13	138	3	12	30
S0–Sa	22	0	112	30	2
Sab–Scd	10	3	24	149	36
Sd–Irr	7	27	2	80	621
	Performance characteristics				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
TPR	70.1	76.7	63.6	53.2	89.2
PPV	66.0	70.4	67.5	67.1	84.3
F	68.0	73.4	65.5	59.4	86.7

Table 5. As for Table 4, but for the CT classifier.

CT classification	Visual classification				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
E	107	21	38	11	17
LBS	4	114	3	8	41
S0–Sa	37	3	99	34	9
Sab–Scd	15	9	31	148	58
Sd–Irr	11	33	5	79	571
	Performance characteristics				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
TPR	61.5	63.3	56.3	52.9	82.0
PPV	55.2	67.1	54.4	56.7	81.7
F	58.2	65.1	55.3	54.7	81.9

Table 6. As for Table 4, but for the CTRF classifier.

CTRF classification	Visual classification				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
E	123	15	31	5	11
LBS	8	136	4	10	31
S0–Sa	24	1	112	25	2
Sab–Scd	8	2	26	158	33
Sd–Irr	11	26	3	82	619
	Performance characteristics				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
TPR	70.7	75.6	63.6	56.4	88.9
PPV	66.5	72.0	68.3	69.6	83.5
F	68.5	73.7	65.9	62.3	86.2

Table 7. As for Table 4, but for the NN classifier.

NN classification	Visual classification				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
E	117	13	28	7	4
LBS	9	130	3	11	27
S0–Sa	27	0	110	23	2
Sab–Scd	12	3	27	162	38
Sd–Irr	9	34	8	77	625
	Performance characteristics				
	E	LBS	S0–Sa	Sab–Scd	Sd–Irr
TPR	67.2	72.2	62.5	57.9	89.8
PPV	69.2	72.2	67.9	66.9	83.0
F	68.2	72.2	65.1	62.1	86.3

Table 8. As for Table 4, but for the binary CTRF classifier.

Binary CTRF classification	Visual classification	
	Spheroid	Disc
Spheroid	450	73
Disc	80	903
	Performance characteristics	
	Spheroid	Disc
TPR	84.9	92.5
PPV	86.0	91.9
F	85.5	92.2

classification predicted by the classifiers in rows. The rows and columns represent the five galaxy types.

For Tables 4–7, the left diagonal represents the objects that are correctly classified by the respective classifiers. For e.g. in Table 4, 122, 138, 112, 149, and 621 objects which were visually classified as E, LBS, S0–Sa, Sab–Scd, and Sd–Irr were correctly classified by the SVM classifier. The other columns show how many of the objects were classified into which other galaxy types. The same format is followed in all the confusion matrices.

A general trend that is observed for all classifiers is that the 'misclassifications' by the classifiers are mostly from neighbouring classes. For e.g. in Table 4, most of the misclassifications by the SVM classifier of the visual E galaxies are as type S0–Sa. Another interesting inference is that galaxies visually classified as classes LBS and Sd–Irr are frequently confused with each other by all four classifiers. This hints at a possible similarity in properties between these galaxy types.

The confusion matrix of the binary CTRF classifier shown in Table 8 is similar to that of the multiclass classifiers. The actual and predicted classifications are represented by the columns and rows, respectively. 450 spheroid-dominated and 903 disc-dominated objects are classified correctly by the binary classifier, while the misclassifications are for 80 and 73 objects, respectively.

The PPV for the corresponding classes gives a measure of classification error by showing how exact the classifier is. For e.g. in Table 4, in the case of type Sab–Scd, while the SVM classifier only positively classifies 53.2 per cent of the time, there is a probability that when it does, it is 67.1 per cent correct. This measure depends heavily on how balanced the data set is, i.e. if there are more objects of a certain galaxy class in the data sample, that particular galaxy type will have a higher value of PPV. This can be seen clearly in the case of galaxy-type Sd–Irr for all the classifiers. It can also be observed in the case of the binary CTRF classifier, for which the data set is more balanced than for multiclass classification, there is a subsequent increase in the PPV of spheroid-dominated objects (which is still the minority class).

The F-score represents the balance between the precision and recall for the classifier. For an unbalanced data set such as ours, the classifier could, in theory, get a higher accuracy rate just by choosing a majority class. In such cases, an F-score is often used to choose an optimum classifier, by choosing one that has consistently high F-scores for all the classes. In case of the four algorithms considered in this study, that classifier is CTRF as can be seen for both the binary and multiclass classifications.

The CT algorithm is observed to be the lowest grossing method over the entire sample, with an average accuracy of 69.0 per cent. The other three methods, CTRF, SVM, and NN have comparable values for classification accuracy at 76.2 per cent, 75.8 per cent, and 76.0 per cent, respectively. This leads us to conclude that perhaps the choice of parameters is a more important factor in classification accuracy rather than the choice of algorithms. Fig. 12 represents the classification efficiencies of these three methods by GAMA HT and for the entire test set. Here, CTRF, SVM, and NN algorithms are represented by green, pink, and blue, respectively. The number of objects that are classified 'correctly' by each method is shown in brackets next to the algorithm labels. The number of objects not classified 'correctly' by any of the three algorithms is given in the top left corner, while the total number of visual HTs is given in the top right corner. As can be seen in the case of each individual visual HT and in the total test set (panel 6), the overall performance of the CTRF classifier is slightly better than the other two. Based on these results, we recommend the CTRF classifier for

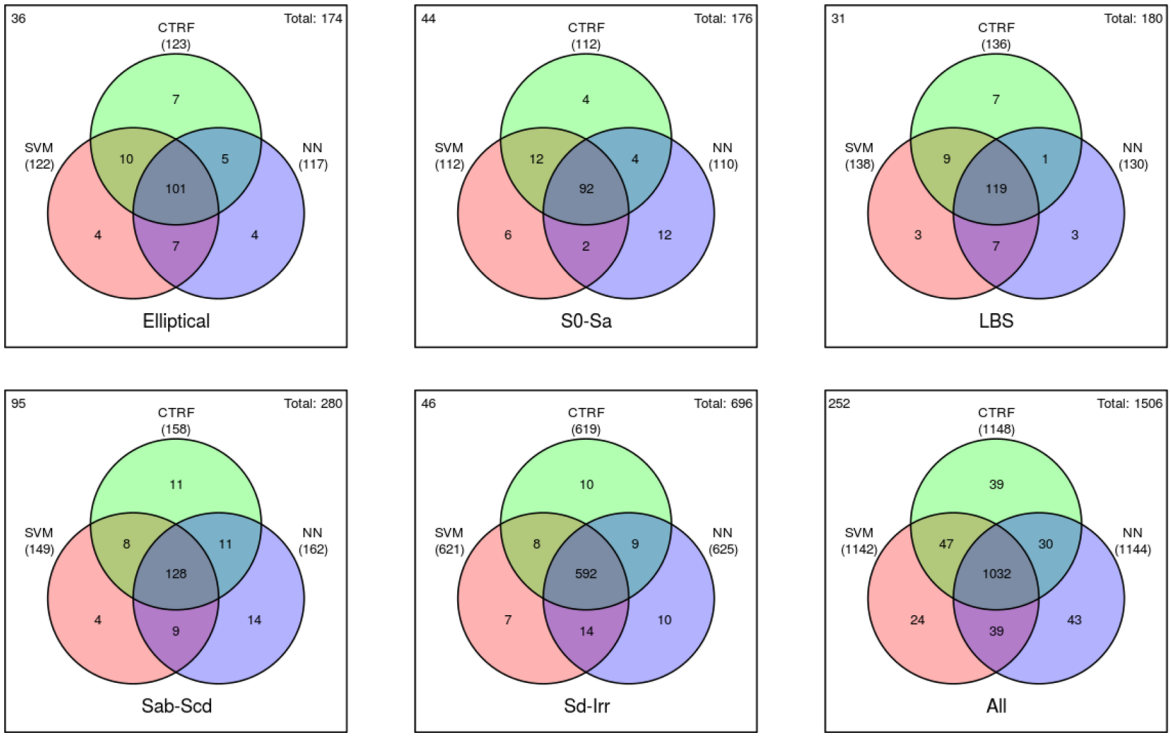

Figure 12. Venn diagrams representing the effectiveness of classification by CTRF, SVM, and NN methods for each GAMA HT and over all types. The number of objects 'correctly' classified by each method is shown in brackets next to the algorithm labels. The number of objects which were not classified 'correctly' by any method is shown in the top left corner, while the total number of objects is given in the top right corner.

further use in astrophysical practice. Even though the improvement in classification accuracy is marginal, CTRF has a simpler mathematical structure. The CTRF machine learnt classifications will be our primary automatic classifications used for further analysis below.

Figs 2–6 show several example postage stamp images of different galaxy types from our test set. The postage stamps span an area of 3× Kron radius of each galaxy and are ordered according to their stellar masses (low-mass galaxies at the top and high-mass galaxies at the bottom). Classifications for different statistical learning algorithms are overlaid on the top right corner of these images in the order SVM, CT, CTRF, and NN. As can be seen, the majority of machine learnt classifications agree well with their visual HT, however, there are instances where one or more algorithms classify a galaxy as something different from its visual classification. All four algorithms are in agreement with each other in 1040 out of the 1506 galaxies in our test set. And out of these 1040 objects, 143 (i.e. ∼10 per cent of the total test set) differ from the respective visual classification. This 'unanimous disagreement' occurs with varying frequency for the different morphological types:[8] ∼9 per cent for type E, ∼9 per cent for type LBS, ∼14 per cent for type S0–Sa, ∼21 per cent for type Sab–Scd, and ∼4 per cent for type Sd–Irr. This phenomenon could be due to two reasons, (1) the visual classification might be inaccurate and, based on the parameters that were used for training, the galaxy belongs to a different class, or, (2) some vital information to classify this galaxy is missing, i.e. the given parameters are not sufficient. Fig. 13 shows a few examples of galaxies that exhibit this phenomenon. Further analysis of this interesting occurrence is required to explore why a host of machine learning algorithms may consistently agree with one another yet disagree with the human eye.

[8] All the numbers quoted here (and henceforth in the same context) are percentages on the total test set.

4.1 Analysis : CTRF classifier

Figs 14 and 15 represent the TPRs obtained by the CTRF classifier as a function of the total stellar mass and redshift, respectively, for the galaxies in our test set. In both cases, the errors are calculated using the `aqbeta` function from the astro library in R (Cameron 2011). This estimates the confidence intervals from quantiles of a beta distribution fit to the data, and is especially suited for small to intermediate data samples.

In Fig. 14, the TPRs obtained by the CTRF classifier are plotted against the total stellar masses of the galaxies from our test set. The first panel represents all galaxies, while the distributions of distinct GAMA HTs are plotted in the subsequent panels (see the legend). We find that the accuracy in classification decreases as the total stellar mass increases. This becomes evident in the extreme mass trends observed for HTs S0–Sa and Sab–Scd. In case of elliptical galaxies (type 1, E), the TPR values seem to be increasing after a dip at $\log_{10} M_{\odot} \sim 10.5$. This seems to be a real rather than a statistical effect, as the bin centred at $\log_{10} M_{\odot} = 10.5$ has more objects in it than the one centred at $\log_{10} M_{\odot} = 11$. For type Sd–Irr, the success rate drops significantly from ∼90 per cent at low mass to ∼30 per cent at $\log_{10} M_{\odot} > 10$. It seems that the algorithm finds it increasingly difficult to classify type Sd–Irr at higher masses, however, we note that the very low number statistics for this population in this mass regime (both in training and test sets), as evidenced by the relatively large error bars could also

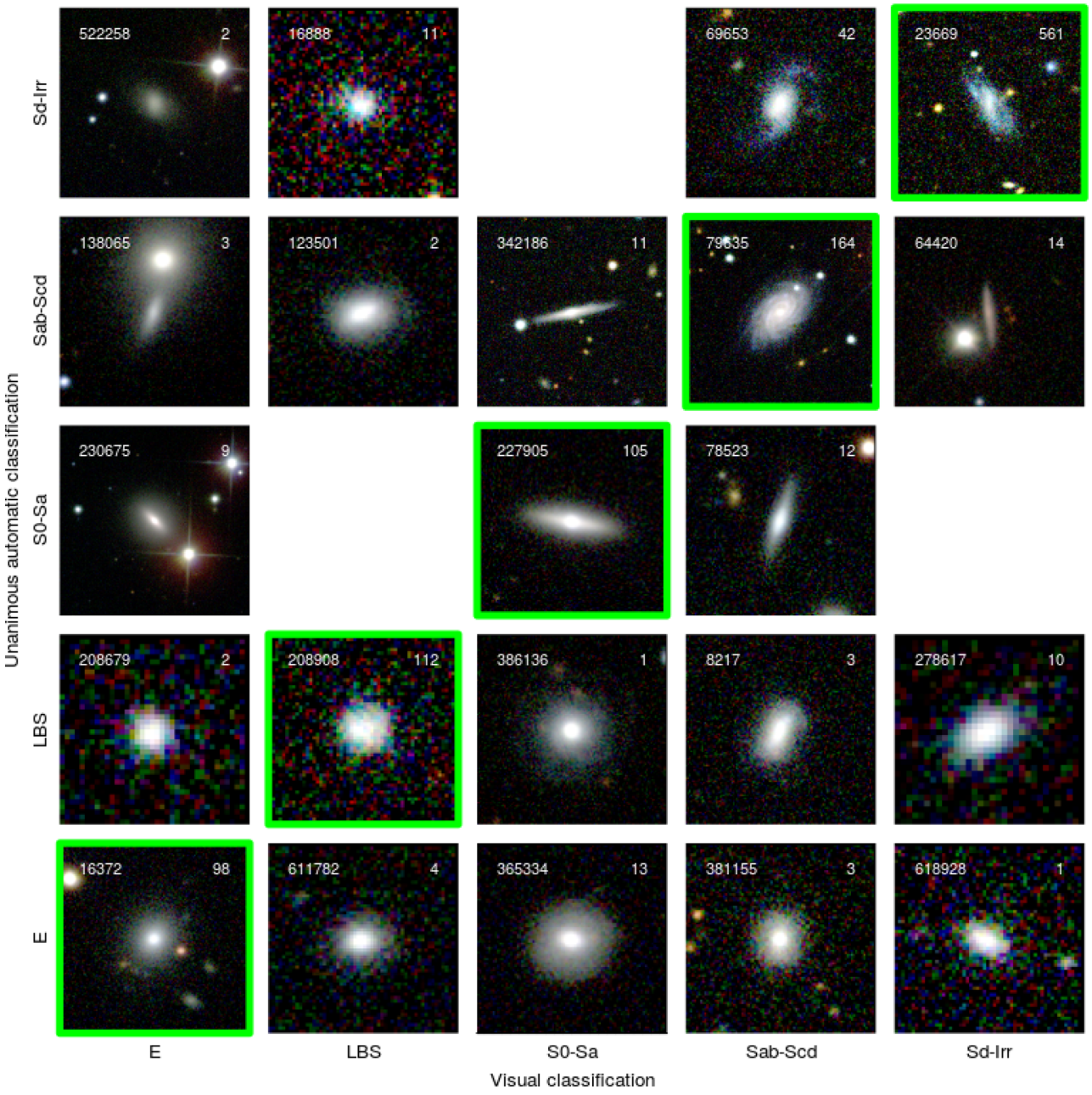

Figure 13. Figure illustrating unanimous disagreement. The *x*-axis represents the visual classification of the objects, while the *y*-axis shows the unanimous automatic classifications. For example, the galaxy in the bottom most row with ID 611782 has been visually classified as LBS while all four algorithms used in this study classify it as type E. The prime diagonal represents objects for which the visual classification and the four algorithms are in agreement (highlighted in green). The number of objects in each bin is noted in the top right corner of each postage stamp. The other blank spaces denote the absence of objects of *x*-axis type unanimously classified by the four algorithms as the *y*-axis type.

be a contributing factor. This trend holds true for type LBS as well. Moffett et al. (2016) note that types LBS and Sd–Irr together account for only about 10 per cent of the total stellar mass density of the parent sample, and that their frequencies drop to nearly zero above the mass range $\log_{10} M_{\odot} = 10.0$. The reason for the decrease in TPR values in case of early- and intermediate-type spirals is not clear at this time, but may be related to the increasingly apparent complexity of structure in galaxies of these types at higher mass regimes.

Fig. 15 is a similar representation of the TPRs with the redshifts of all the galaxies in the test set along the *x*-axis. The first panel represents all the galaxies in our test set, while the succeeding panels represent the different HTs (see the legend). For the total sample, the trend is to be expected, considering that we have attempted to choose redshift independent parameters. However, we observe varying trends along the subpopulations. The trend for each HT subpopulation is similarly consistent with a flat relation with redshift, with the notable exception of type Sab–Scd, for which the TPR is lower at low redshifts and goes on to increase at higher redshifts. This may be due to the fact that local galaxies are better resolved than distant galaxies, and therefore the automated algorithms may be having a harder time processing the extra structural data. The

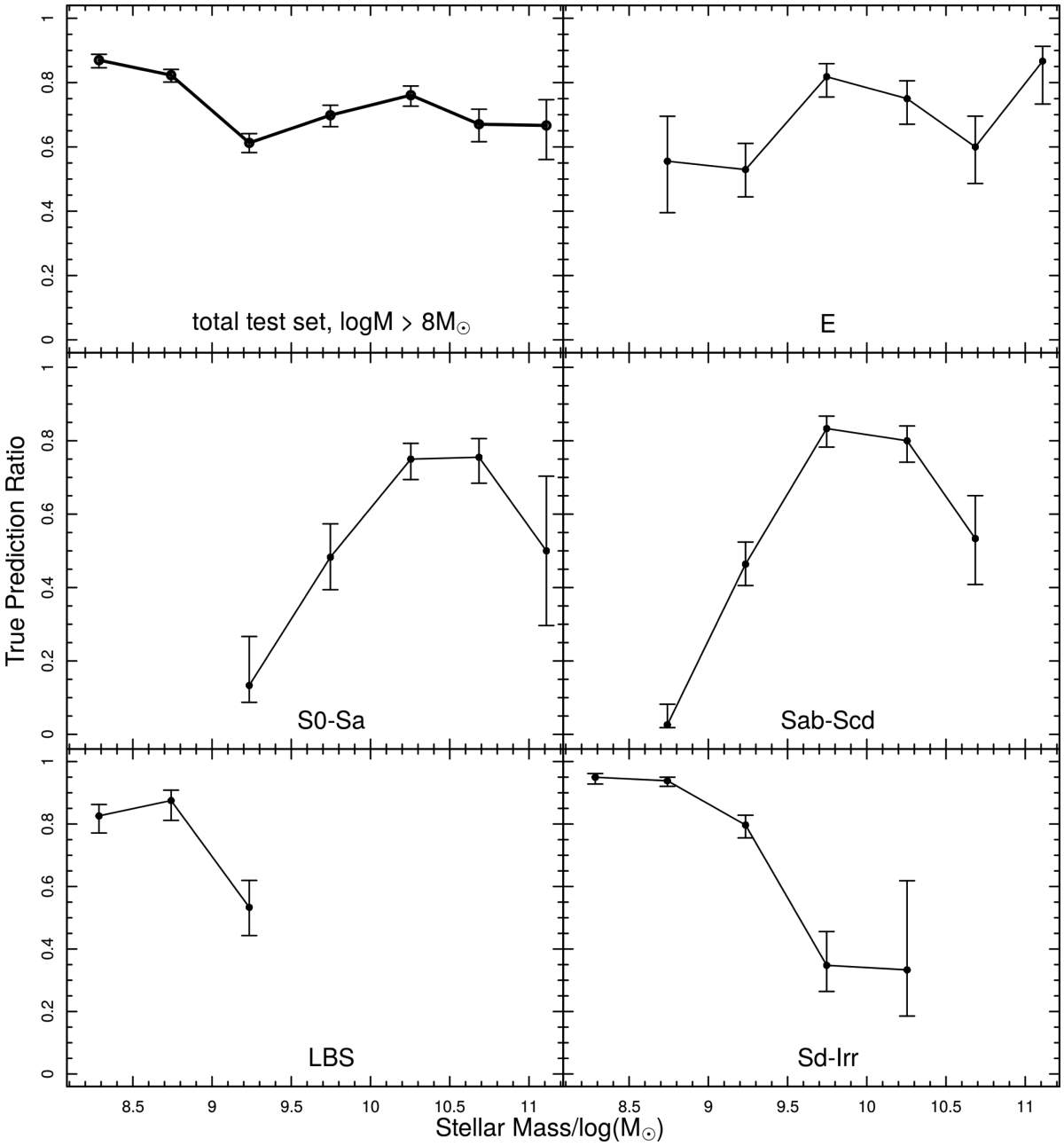

Figure 14. Representation of the TPR as a function of total stellar mass (log) for the method that we recommend, CTRF. The distribution over the total test set is represented in the first panel. The individual contributions of the different GAMA HTs are plotted in the subsequent panels as indicated. The lower and upper boundary fractional errors for the data set are calculated by using the `aqbeta` function from the astro library in R (Cameron 2011).

apparent angular scale from $z = 0.02$ to 0.06 decreases by a factor of $\sim$3, which has the effect of blurring stellar populations within the galaxies.

Figs 16–20 show the location of galaxies in the Sérsic index – $g - i$ colour plane with each figure representing a different visual HT morphology. Data point types and colours represent the morphological types assigned to each galaxy by the CTRF classifier. The marginal histograms represent the distributions of $g - i$ colour (top) and Sérsic index (right) for the visual and CTRF classifications. The efficiency of classification by the CTRF classifier for different HTs can be visually inspected from these histograms.

Fig. 16 shows all visually classified elliptical galaxies in the Sérsic index versus $g - i$ colour plane. Most of the objects for which the classifier is unable reproduce the visual classification are determined to be early-type spirals (S0–Sa). The objects that have been classified by the CTRF classifier as S0–Sa are all redward of the main population, whilst other types are scattered in the blue low Sérsic index tail of the E distribution. One reason for this could be the potential systematic misclassification of face-on red S0 galaxies as ellipticals. If true, our machine learning algorithm may provide a robust automated means by which we could apply corrections to currently existing visual morphological data sets to address the issue of E/S0 confusion. Another reason for this 'spheroid-disc tension' between the human eye and the automated algorithms could be the presence of discy elliptical 'ES' (Liller 1966; Graham, Ciambur & Savorgnan 2016; Savorgnan & Graham 2016) class with

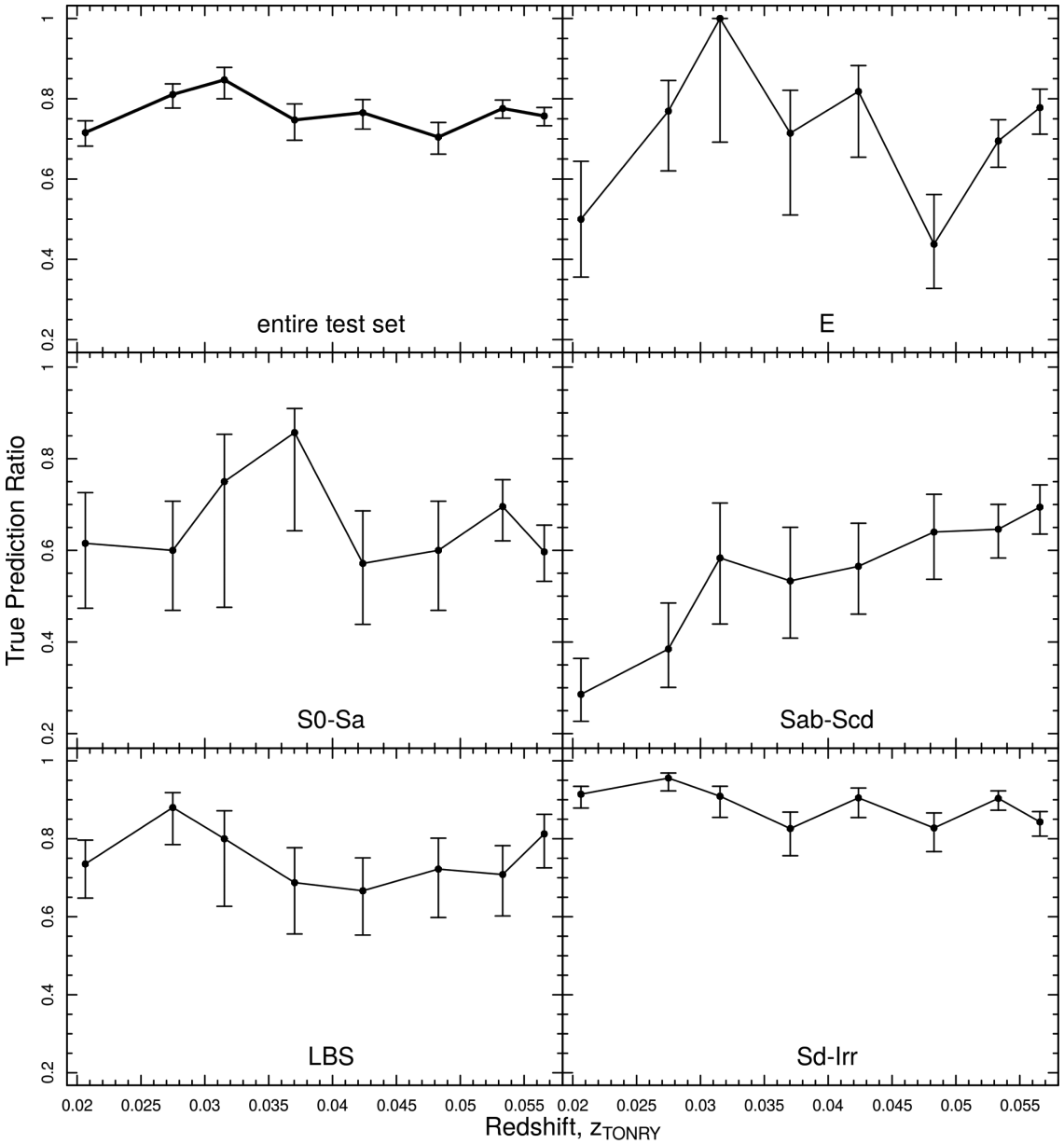

Figure 15. As Fig. 14, but as a function of redshift.

intermediate discs in our sample. It could also be a wider 'red disc detection' issue, however, we note that the Sérsic indices for many of these objects are of the order of $n \sim 4$ which indicates spheroid-dominated systems.

Fig. 17 shows objects that are visually classified as LBS (type 2, represented as green squares). The instances where the CTRF classifier is not in agreement with the visual classifications are represented by the other colours and points in the scatter plot. In general, most of the objects which were not found to be LBS by the CTRF method have been classified as late-type spirals and irregulars, except towards the redder end of the scatter plot, where they have been classified as elliptical galaxies. We note that in the visual classification of this particular type, the 'blue colour' was a secondary characteristic, the objects were primarily classified on the basis of their shape and size.

Fig. 18 shows objects visually classified as early-type spiral galaxies (type 1112, S0–Sa, barred and unbarred, represented as black diamonds). The CTRF classifier's classifications that do not agree with the visual morphology are almost equally divided between ellipticals (red circles) and intermediate-type spirals (purple triangles). They seem to be uniformly distributed in Sérsic index space, while there appears to be some dependence in $g - i$ colour, with the objects classified as ellipticals clustered in an area redder than the objects that are classified as intermediate-type spirals. Classification as intermediate-type spiral follows a trend observed by Owens et al. (1996), in that differentiating between neighbouring classes of galaxies such as these is more difficult than differentiating between non-neighbouring classes. The population of elliptical galaxies we find might be an indicator that the human eye is fallible when classifying this type of galaxy. Very few objects are classified as late-type spirals and irregulars or LBS (mostly at the bluer end).

Fig. 19 shows objects that are visually classified as intermediate-type spirals (type 1314, Sab–Scd, purple triangles). In most instances where the CTRF classifier disagrees with the visual

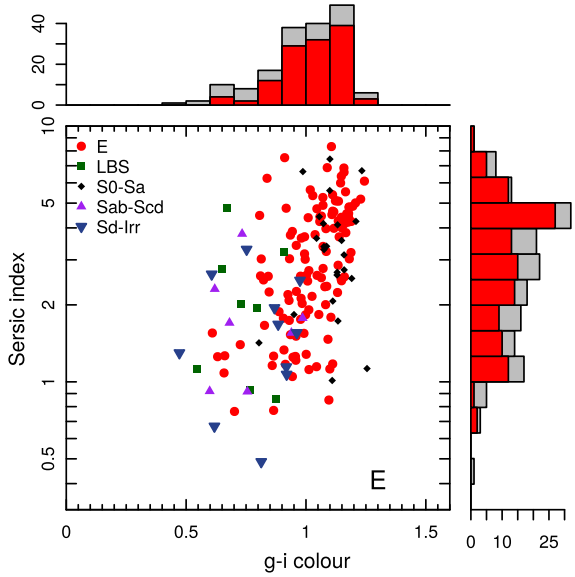

Figure 16. Scatter plot with marginal histograms showing all visually classified elliptical (type 1, E) galaxies in Sérsic index and $g - i$ colour space. Data point colours and types vary according to their CTRF classification, as indicated by the inset legend. Marginal histograms show the distribution for all (grey) and visually classified elliptical (red) galaxies.

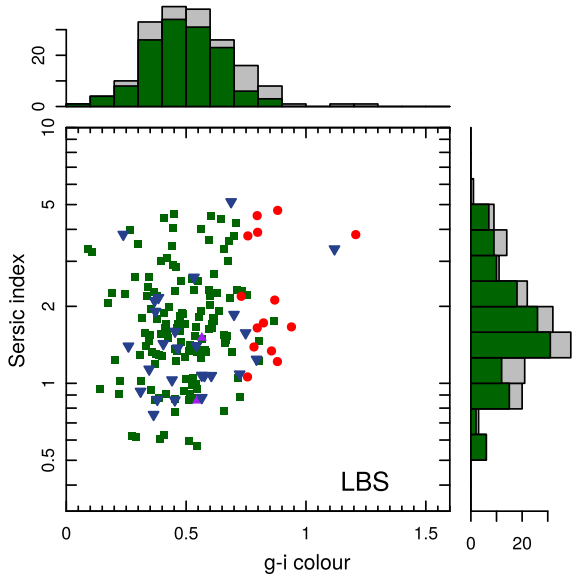

Figure 17. As Fig. 16, for LBS (type 2).

classification, it classifies objects as late-type spirals and irregulars. However, at the redder and higher Sérsic index end, some objects are classified as early-type spirals. This is also the galaxy type for which the classifiers of the machine learning algorithms that we have applied disagree the most with visual classifications.

Fig. 20 shows objects that are visually classified as late-type spirals and irregulars (type 15, Sd–Irr, represented as blue triangles pointing down). For this particular galaxy type, all four machine learning algorithms have a high agreement rate with the visual classifications (>80 per cent). As is shown, the disagreements are evenly divided between types LBS and intermediate-type spirals, while there are a few objects classified as ellipticals. The classifications as LBS and ellipticals could be an indication that these

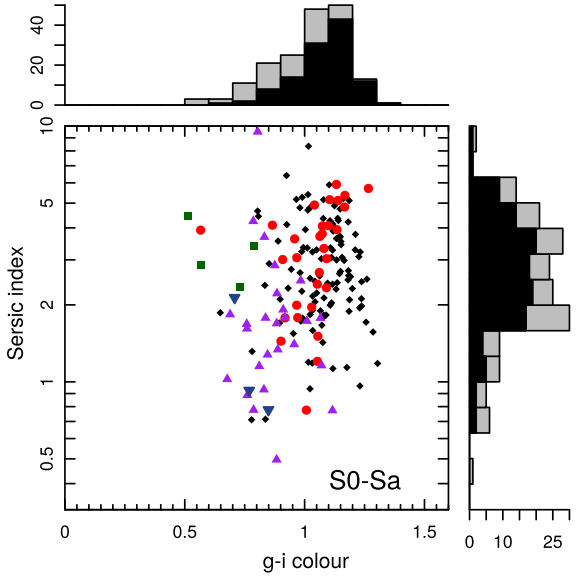

Figure 18. As Fig. 16, for early-type spirals (type 1112, S0–Sa, barred, and unbarred).

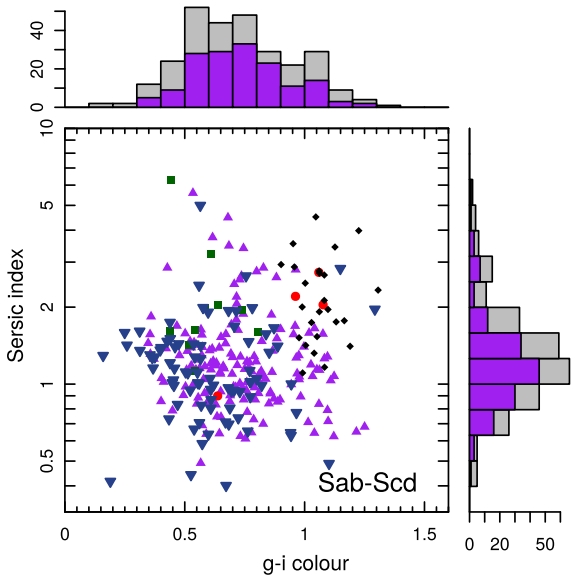

Figure 19. As Fig. 16, for intermediate-type spirals (type 1314, Sab–Scd, barred, and unbarred).

objects may have more in common with early-type galaxies than is currently conceived. The classifications as intermediate-type spirals are likely due to the Owens et al. (1996) observations mentioned previously.

4.2 Impact of chosen parameters on the CTRF classifier

We perform a sensitivity test to ascertain the impact of each parameter on the classification process of our CTRF algorithm. In order to achieve this, we remove all the parameters mentioned in the upper panel of Table 2 one by one, and obtain the TPRs, retraining the CTRF classifier in each instance. The results of this are shown in Table 9.

The removal of Sérsic index lowers the overall rate of accuracy the most, by almost 1.4 per cent. All other increases and decreases from

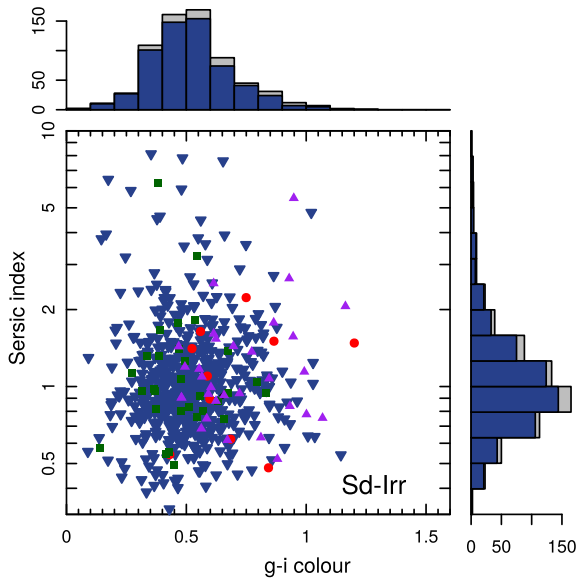

Figure 20. As Fig. 16, for late-type spirals and irregulars (type 15, Sd–Irr).

the overall TPR caused by the removal of parameters are within the error limits defined in Table 3. The only parameter whose removal causes an increase in the overall TPR is absolute magnitude, by 0.3 per cent. This indicates that for the total data sample, Sérsic index is the parameter that contributes most to the classification process by the CTRF algorithm. This, however, does not hold true for the individual HTs.

Removal of $u - r$ colour and stellar mass does not affect the classification in the case of elliptical galaxies. Absolute magnitude and mass-to-light ratio have an almost similar effect on the TPR values, albeit in different directions. When absolute magnitude is removed, the TPR value increases by 1.1 per cent and when mass-to-light ratio is removed, the value decreases by 1.2 per cent. The parameters for which the accuracy falls outside the error bars are ellipticity and half-light radius.

In case of LBS galaxies, the parameters that affect the classification process the most are half-light radius, Kron radius (semimajor and semiminor), mass-to-light ratio, and stellar mass. The parameters that have a similar effect on the classification rate are Kron radius (semiminor axis), mass-to-light ratio, and stellar mass, a decrease by ∼4 per cent. The decrease in TPR values is drastic in the case of both half-light radius and Kron radius (semimajor axis), ∼7 per cent and ∼5 per cent, respectively.

For early-type spiral galaxies, the changes in TPR are within the error bars except in the case of Sérsic index. When Sérsic index is removed prior to training the classifier, the accuracy drops by ∼5 per cent. The effects caused by the absence of Kron radius (semiminor), mass-to-light ratio and $u - r$ colour are analogous, a decrease of ∼1 per cent. Same is the case with Kron radius (semimajor) and absolute magnitude, by ∼2 per cent. When $g - i$ colour is excluded from the process, the TPR values remain the same as that from the original run.

The change in accuracy for intermediate-type spirals after removing the parameters one by one, are all within the error limits of the values from Table 3. As in case of early-type spirals, removing $g - i$ colour has no effect on the original TPR values. Sérsic index and Kron radius (semiminor) contribute to a decrease in TPR values by ∼0.7 per cent each; Kron radius (semimajor), half-light radius, and $u - r$ colour to an increase by ∼0.4 per cent each; and mass-to-light ratio and stellar mass to an increase by ∼1 per cent each. Removing absolute magnitude seems to matter the most, by increasing the accuracy by ∼2 per cent.

Table 9. Results of parameter sensitivity test on the CTRF algorithm in percentages are shown in panel 1. In panel 2, similar results for redundant parameters according to the PCA performed in Section 2.4 (Fig. 1) are shown. The results for the CTRF classifier from the original run are shown in panel 3.

Parameter removed	E 1	LBS 2	S0–Sa 1112	Sab–Scd 1314	Sd–Irr 15	All
Sérsic index	67.8	76.7	58.5	55.7	87.9	74.8
Kron radius in kpc (semiminor axis)	69.0	71.7	62.5	55.7	88.7	75.2
Half-light radius in kpc	65.5	68.9	60.8	56.8	90.5	75.3
Kron radius in kpc (semimajor axis)	67.8	70.6	61.4	56.8	89.8	75.5
Ellipticity	66.7	74.4	63.1	57.1	88.8	75.6
Mass-to-light ratio	69.5	71.7	62.5	57.5	89.1	75.8
$g - i$ colour	67.2	76.7	63.6	56.4	88.8	76.0
Stellar mass	70.7	71.7	64.2	57.5	88.8	76.0
$u - r$ colour	70.7	74.4	62.5	56.8	89.0	76.0
Absolute magnitude	71.8	73.3	61.4	58.2	90.0	76.5
Mass-to-light ratio and $g - i$ colour	68.4	75.0	63.7	60.0	89.1	76.6
Mass-to-light ratio and $u - r$ colour	67.8	75.6	60.8	57.9	89.7	76.2
$u - r$ colour and $g - i$ colour	69.5	76.1	61.4	60.4	89.4	76.8
All chosen parameters	$70.7^{+3.2}_{-3.7}$	$75.6^{+2.9}_{-3.5}$	$63.6^{+3.5}_{-3.8}$	$56.4^{+2.9}_{-3.0}$	$88.9^{+1.1}_{-1.3}$	$76.2^{+1.1}_{-1.1}$

The changes in TPR in case of late-type spirals and irregulars are mostly within the error bars of the original results, except in the case of half-light radius where it increases by ∼2 per cent, which seems to have the most impact on classification accuracy as well. Ellipticity, $g - i$ colour, and stellar mass have a similar effect on the TPR values (decrease by 0.1 per cent). $u - r$ colour seems to have a similar impact on the classifier's performance for this galaxy class, an increase of the TPR by 0.1 per cent.

In the PCA, we performed (represented in Fig. 1), ellipticity was found to be the parameter which contained the least variability. But as can be seen from Table 9, while it might not be the most important parameter overall, it has a significant impact in the classification accuracies of individual HTs, especially elliptical galaxies. The TPR of ellipticals fall by 4 per cent when this parameter is removed.

Also represented in Fig. 1 is the redundancy of the parameters, mass-to-light ratio, and $g - i$ and $u - r$ colours. We also explore here, the impact on the classification accuracies when these parameters are removed two at a time. These results are represented in the second panel of Table 9.

When mass-to-light ratio and $g - i$ colour are removed, there is a marginal increase in the overall TPR value, to 76.6 per cent. This increase is reflected in the individual HTs, S0–Sa, Sab–Scd, and Sd–Irr. The accuracies take a consequent dip in case of types E and LBS.

The removal of mass-to-light ratio and $u - r$ colour does not make a significant overall impact, with the TPR value remaining the same as that of the original run, at 76.2 per cent. Among the individual HTs, the accuracy of LBS remains unchanged, while that of types E and S0–Sa decrease. The individual TPRs of types Sab–Scd and Sd–Irr reflect marginal increases.

Removing $g - i$ and $u - r$ colours resulted in an increase in the overall TPR value, to 76.8 per cent. This increase was contributed by the increases in the TPRs of galaxy types LBS, Sab–Scd. and Sd–Irr. The accuracies of types E and S0–Sa was found to drop marginally.

The slight increases and decreases in the TPR values when the parameters are removed one by one are largely within the error margins defined for the TPRs from the original run and therefore are not deemed significant. Similar is the case when redundancies in parameters are removed.[9] Therefore we conclude that, while the individual HTs might be sensitive to certain parameters more than the others, all parameters contribute to some extent in the overall classification process of the CTRF algorithm.

4.3 CTRF classifier for binary classification

With the same training, test, and parameter sets that we have employed in multiclass classification, we constructed a binary CTRF classifier with two classes, spheroid-dominated and disc-dominated.[10] The galaxies which were visually classified as ellipticals (type 1, E), LBS (type 2), and early-type spirals (type 1112, S0–Sa) were considered as spheroid-dominated, while the intermediate-type spirals (type 1314, Sab–Scd) and late-type spirals and irregulars (type 15, Sd-Irr) were considered as disc-dominated.

This binary CTRF classifier returned a total success ratio of 89.8 per cent$^{+0.7}_{-0.8}$ with individual TPRs of 84.9 per cent$^{+1.4}_{-1.7}$ and 92.5 per cent$^{+0.8}_{-0.9}$ for the spheroid- and disc-dominated classes, respectively. This significant increase from the original CTRF classifier's TPRs proves that as the number of classes into which classification is made increases, the classification accuracy decreases. This might also be directly related to the size of the data set, and how well each class is represented in the training set.

Similar to the analysis in Section 4.2, we also explored the impact the different parameters might have on the classification performance of the classifier constructed by the CTRF algorithm. The results of this are given in Table 10.

Removing half-light radius from the parameter set used for training and testing the CTRF algorithm seems to be the have the most impact on the performance of the binary CTRF classifier. While the overall success rate drops by 2.46 per cent, the values for spheroid- and disc-dominated systems fall by ∼6 per cent and ∼0.4 per cent, respectively. This points at the greater significance of half-light radius in the classification of spheroid-dominated galaxies rather than the disc-dominated ones. This is in agreement with the results represented in Table 9 in which the classification accuracies fall consistently for these three classes (E, LBS, and S0–Sa) in the case of multiclass classification.

Ellipticity and $g - i$ colour, and absolute magnitude and $u - r$ colour seem to have similar overall effect on the classification process, drops by ∼0.8 per cent and ∼0.6 per cent for the respective pairs. The fluctuations in the TPR values are most significant in the case of ellipticity for disc-dominated systems. The entire contribution to the change in TPR while ellipticity is removed as a classifying criterion comes from disc-dominated systems. This is a very interesting development because, in the case of multiclass CTRF classification discussed in Section 4.2, ellipticity is one of the parameters that cause the TPR to decrease for all three galaxy types collectively called as spheroid-dominated. This might indicate cross-contamination between these three galaxy types in the visually classified sample which confuses the classifier.

Table 10. Panel 1 shows the results of parameter sensitivity test performed with the binary CTRF classifier. The results with all chosen parameters (Table 2) are shown in panel 2.

Parameter removed	Spheroid -dominated	Disc -dominated	All
Half-light radius in kpc	78.7	92.1	87.4
Sérsic index	82.5	90.4	87.6
Kron radius in kpc (semimajor axis)	82.6	91.4	88.3
Kron radius in kpc (semiminor axis)	83.8	91.2	88.6
Mass-to-light ratio	83.6	91.4	88.7
Stellar mass	83.0	92.0	88.8
Ellipticity	84.9	91.3	89.0
$g - i$ colour	84.3	91.6	89.0
Absolute magnitude	84.8	91.6	89.2
$u - r$ colour	83.8	92.1	89.2
All chosen parameters	$84.9^{+1.4}_{-1.7}$	$92.5^{+0.8}_{-0.9}$	$89.8^{+0.7}_{-0.8}$

[9] It is interesting to see that the TPR values for Sab–Scd, the class that performs the worst during classification by all our algorithms, experience significant increases when the redundant parameters are removed. However, since this does not make a noteworthy change in the overall rate of accuracy, we have decided to overlook this improvement and keep the parameter set as is.

[10] We use this terminology based on the visual classification of the data set. Since lenticular galaxies are gathered under the same umbrella as Sa-type galaxies, an early- to late-type galaxy split would involve reclassifying the entire visual sample, which is beyond the scope of this work.

The accuracy rates (both overall and individual) fall beyond the error margins when parameters such as Sérsic index, Kron radii (major and minor axes), mass-to-light ratios, and stellar mass are removed. According to this study, the parameters that influence our CTRF algorithm the most are half-light radius, Sérsic index, Kron radii, mass-to-light ratio, and stellar mass.

5 DISCUSSION

In this section, we discuss in greater detail our previously recovered results. To begin, we note that type 15 (Sd–Irr galaxies), account for almost 50 per cent of our test set, and the associated TPR success values are above 80 per cent for all considered automated classification methods. This could indicate one of three scenarios; (1) as the percentage of objects in a certain class increases, the accuracy of classification increases as well, (2) the algorithms that we tested are more effective in classifying a particular HT (type 15 in our case) using the parameters that we have prescribed or (3) the human classifications may be biased towards being able to more accurately classify Sd–Irr type galaxies.

The first scenario is not generally supported by our own results. The TPR values for type 1314 are consistently low across all four considered methods and yet it is the second most populous type in both our training and test sets. This warrants additional analysis in future works; by testing the codes on larger data samples and by fine-tuning the classification algorithms by introducing techniques such as cross-validation.

As to the second scenario, the successful utilization of our adopted functions are directly linked to our choice of parameters. It may be that one or more of the parameters that we have chosen are more effective in classifying certain HTs while falling short in others. For example, the complexity in the structure of the galaxy might not be well defined by the parameters that we have chosen. As can be seen in Table 3, the TPR values are considerably higher for single component systems such as ellipticals (type 1, E) and late-type spirals/irregulars (type 15, Sd–Irr) compared to multicomponent systems such as early- and intermediate-type spirals (types 1112, S0-Sa and 1314, Sab–Scd, respectively).

All four algorithms are in agreement with each other in 1040 out of the 1506 galaxies in our test set. And out of these 1040 objects, 143 (i.e. ∼10 per cent of the total test set) disagree with the classification by visual inspection. Of these, ∼9 per cent are ellipticals, ∼9 per cent are LBS, ∼14 per cent are early-type spirals, ∼21 per cent are intermediate-type spirals, and ∼4 per cent are late-type spirals and irregulars. These are illustrated in Fig. 13. There seems to be an element of symmetry in this occurrence. For instance, as can be seen from the figure, no objects that have been visually classified as S0–Sa are machine classified unanimously as Sd–Irr, this pattern holds true in converse as well. But this is not always the case. No visual LBS galaxies have been unanimously machine classified as S0–Sa objects, but one visual S0–Sa galaxy has been machine classified unanimously as LBS. This, along with the possibility of unanimous disagreement being a potential indicator of human error in classification by visual inspection are interesting paths to follow in future works that extend this study.

When we train a machine, for e.g. to classify galaxies (our case) based on visual classifications, what we essentially do is train it to reproduce our classification strategy, replete with our human biases. For instance, if, beyond a certain redshift, the human eye is ineffective in distinguishing between certain classes of galaxies, the data set that we apply to the algorithms will reflect the same bias. Therefore, we propose that the disagreement between the machine and the visual classifications could be due to one of two reasons, (1) the visual classification is inaccurate, and based on the values of the parameters used to train and test the algorithms, the galaxy belongs to one of the other classes, or (2) the parameters do not sufficiently characterize what we see while classifying by eye.

In Figs 16, 17, and 20, it can be seen that the CTRF method replicates the visual classification to a greater extent than in Figs 18 and 19. This leads us to speculate that our algorithms in their present configuration might be more effective in classifying single-component systems such as ellipticals and late-type spirals rather than multicomponent systems like early- and intermediate-type spirals.

One of the methods that we have used in our work is SVM with a tree structure. With this approach, the accuracy obtained on our entire test set is 75.8 per cent. The accuracies for the different HTs are represented in Table 3. This value seems encouraging when we compare our results to Huertas-Company et al. (2007), who also used an SVM approach in their work to obtain morphological classification to a sample of 1500 galaxies from the SDSS (500 to train and 1000 to test). Their method was a generalization of C-A system using non-linear SVM boundaries with 12 dimensions. The mean accuracy of the method was ∼80 per cent. We note that the Huertas-Company et al. (2007) method only classifies galaxies into early and late types, while our algorithm classifies galaxies into five distinct morphological types, which may explain why their success ratio is ∼4 per cent higher than ours.

In our NN method, we reproduce the classifications learned on the training set to an accuracy of 76.0 per cent on the test set. Banerji et al. (2010) applied artificial NN to a sample of almost one million objects from the SDSS previously classified by human eye by volunteers as part of the Galaxy Zoo project. Their training set consisted of 75 000 objects, classifying the test set into three morphological classes (early-types, spirals, and point sources/artefacts) with 12 parameters. The accuracy of their approach was close to 90 per cent. Considering that our training set and test sets are much smaller compared to Banerji et al. (2010) and that we use a larger range of classification types, our value of 76.0 per cent is highly promising.

Our CT algorithm uses classification (decision) trees to attain morphological classification with an accuracy of 69.0 per cent on our entire test set. The size of the data set and the number of classification types for the method of Owens et al. (1996) is comparable to our own. They use a sample of 5217 galaxies from the ESO-LV[11] catalogue (Lauberts & Valentijn 1989) using 13 parameters to discern between five morphological types (ellipticals, lenticulars, early-type spirals, late-type spirals, and irregulars). With a fivefold cross-validation on their approach, they achieved an average accuracy of 63 per cent on a test set which amounted to 1/5th of the whole set. They have compared their results with Storrie-Lombardi et al. (1992) which applied an artificial NN approach to the same data with an accuracy of 64.1 per cent and Lauberts & Valentijn (1989) whose automated classifier reproduced classifications to an accuracy of 56.3 per cent. We note however, that Storrie-Lombardi et al. (1992) have used ∼30 per cent of their total data sample as the training set and 70 per cent as the test set in contrast to our method of adopting a larger training set and smaller test set as detailed in Section 2.5. The improvement of 69.0 per cent accuracy that we observe is undoubtedly due to this reason. Furthermore, we have 2000 more objects in the data sample which will influence the classification accuracy.

[11] http://archive.eso.org/wdb/help/eso/esolv.html

Table 11. Summary of the results from this study (top) alongside results from several other studies from the literature using a variety of statistical learning methods (bottom).

Statistical learning method	Total sample	Training set	Test set	Number of classes	Dimensions	Accuracy	Reference
SVM	7528	6022 (80 per cent)	1506 (20 per cent)	5	10	75.8 per cent	Results from our work
NN						76.0 per cent	
CT						69.0 per cent	
CTRF						76.2 per cent	
SVM	∼1500	500 (33 per cent)	1000 (67 per cent)	2 (early-type, late-type)	12	80 per cent	Huertas-Company et al. (2007)
NN	∼1 000 000	∼75 000 (7.5 per cent)	∼925 000 (92.5 per cent)	3 (early-type, spirals, point sources/artefacts)	12	90 per cent	Banerji et al. (2010)
Oblique CT	5217	∼4174 (80 per cent)	∼1043 (20 per cent)	5 (E, S0, Sa+Sb, Sc+Sd, Irr)	13	63 per cent	Owens et al. (1996)
Three CT algorithms including CTRF	75 000	67 500 (90 per cent)	7500 (10 per cent)	3 (ellipticals, spirals, unknown)	13	96.2 per cent	Gauci et al. (2010)
ConvNet	58 000	47 700 (∼82 per cent)	5000 (∼9 per cent)	5 (probablities[a])	Run on images	∼99 per cent	Huertas-Company et al. (2015)
			5300 (∼9 per cent) used for real-time evaluation during training				Dieleman et al. (2015)

Note. [a]Probabilities for each galaxy having a disc or a spheroid, being a point source, having an irregularity or being unclassifiable are the outputs.

Among our methods, the CTRF algorithm which employs an RF of 100 trees was found to have an accuracy of 76.2 per cent. This method has a marginal, but encouraging higher accuracy among all four methods that we have tested. Gauci et al. (2010) performed a comparison of different CT algorithms to a data set of 75 000 objects from the SDSS previously classified by the Galaxy Zoo project. The algorithms of CART, C4.5, and RF are tested with a tenfold cross-validation technique where, in each run, nine subsets of the data are used for training and one for testing. The success rate was 97.33 per cent for an RF algorithm with 50 trees and 96.2 per cent over all the methods. However, Gauci et al. (2010) have only three classification types (elliptical, spiral, and unknown morphology) compared with five in this study.

We trained a binary CTRF classifier that classifies our data sample to spheroid- and disc-dominated systems. For this, we consider galaxy types E, LBS, and S0–Sa as spheroid-dominated and galaxy types Sab–Scd and Sd–Irr as disc-dominated. The overall accuracy rate for this classifier is ∼90 per cent with individual TPRs for spheroid- and disc-dominated systems to be ∼85 per cent and ∼93 per cent, respectively.

The results from our binary CTRF classifier has clarified certain aspects about the effectiveness of our overall study. The results indicate that the number of data types into which the classification is done is a very important criterion for accuracy. There is an increase of almost 14 per cent overall accuracy when the number of types changed from 5 to 2. It is conceivable that the size of the data set and how comprehensively the different galaxy types are represented in the training and test sets play a role in the performance accuracy as well. This can be seen in the higher accuracy of the disc-dominated galaxies which make up ∼67 per cent of the total data set. So a way to address the decrease in accuracy as the galaxy types increase might be to increase the size of the data set accordingly.

To facilitate future studies and to aid in comparison with other works (see Table 11), the machine learning algorithms employed in this study have not been significantly modified beyond their default setups as detailed in Section 3. There are several avenues that could be pursued in order to make them more precise. Applying the SVM method for multiclass classification using error-correcting output codes is one such approach (Dietterich & Bakiri 1995). There are indications in literature that this technique could be more accurate than the tree structure that we have considered in this work. Assigning probabilities to our classifications rather than binary values may be a useful tool to see the effectiveness of the classification process. Owens et al. (1996) posit that differentiating between neighbouring classes of galaxies (for e.g. types 1112 and 1314 in our sample) is more complicated than differentiating between non-neighbouring classes of galaxies. By analysing the probabilities assigned to each class by the classification algorithms and manual examination, it might be possible to define criteria or introduce parameters that provide a more robust delineation between neighbouring galaxy types. Introducing PCA as a means to choose a robust set of parameters and extensive error analysis of the parameters that we have chosen are other interesting prospects, allowing for the introduction (e.g. some measure of asymmetry) or elimination of parameters (e.g. ellipticity) which do not seem to be vital in predicting morphology. The methods that we have chosen construct classifiers with different mathematical structures. Therefore, each constructed classifier may capture different aspects of the ideal classifier effectively. Using a combination of classifiers constructed using different statistical learning methods may give rise to a new classifier with better

accuracy (and closer to an ideal classifier) than each classifier taken individually (Chen, Pereverzyev & Xu 2015; Kriukova et al. 2016). The design of appropriate combination strategies is another avenue that we may explore in the future.

6 CONCLUSION

In this study, we have used the statistical machine learning algorithms of SVM, CT, CTRF, and NN to carry out morphological classifications for 7528 galaxies from the GAMA survey. These galaxies were previously visually classified independently by three classifier teams and the majority vote has been included in the GAMA catalogue. The algorithms are trained on a set of 6022 objects (80 per cent of the data set) using 10 distance independent parameters. These algorithms are subsequently tested on the remaining 20 per cent of the data set (1506 objects) to classify them into five galaxy types: elliptical (type 1, E), LBS (type 2), early-type spirals (type 1112, S0–SBa), intermediate-type spirals (type 1314, Sab–SBcd), and late-type spirals and irregulars (type 15, Sd–Irr). We draw the following conclusions from our study.

(i) The success rates on the entire test set are 69.0 per cent, 76.2 per cent, 75.8 per cent and 76.0 per cent for the CT, CTRF, SVM, and NN algorithms respectively. While the performance of the SVM, CTRF, and NN algorithms are very similar, the CTRF algorithm has a marginally better success rate and a simpler mathematical structure. We therefore recommend this algorithm to provide robust, automated HT classifications when applied to future extragalactic surveys.

(ii) Our algorithms have a greater success rate in case of single-component systems such as ellipticals, and late-type spirals and irregular galaxies. This is especially clear when we look at the success rate of type 15 galaxies (Sd–Irr). They form 47 per cent of our entire sample. The success rates of all four algorithms are above 80 per cent for this galaxy type and close to 90 per cent for CTRF, SVM, and NN algorithms.

(iii) We find that the success rates decrease with increasing stellar mass. This trend seems drastic in case of HTs S0–Sa, Sab–Scd, and Sd–Irr. This apparent phenomenon warrants further investigation.

(iv) We do not find a universal trend in the success rates with respect to redshift, however, we find that there is some redshift dependence within each galaxy type. This is especially apparent in the case of type Sab–Scd, for which, the success rates are lower at lower redshifts and increase towards higher redshifts.

(v) In the cases where all four machine learning algorithms agree with each other, they disagree with the visual classification $\sim$10 per cent of the time, with $\sim$9 per cent being ellipticals, $\sim$9 per cent LBS, $\sim$14 per cent S0–Sa, $\sim$21 per cent Sab–Scd, and $\sim$4 per cent Sd–Irr. These unanimous disagreement fractions could be a potential indicator for human error in visual classifications. Further exploration of this is an interesting path to investigate for future work.

(vi) When we decrease the number of galaxy types into which classification is done, the accuracy of classification increases considerably. Our binary CTRF classifier achieved an overall accuracy of 89.8 per cent with the spheroid- and disc-dominated classes achieving accuracies of 84.9 per cent and 92.5 per cent, respectively. This hints that a way to cope with the decrease in classification accuracy as the galaxy types increase might be to use larger data sets.

(vii) There are many possible avenues to pursue following from this study. These include introducing analysis methods such as PCA or cross-validation to create a robust data set of input features, foregoing the SVM tree structure in favour of error-correcting codes, and using an ensemble of classifiers constructed using different statistical learning methods.

ACKNOWLEDGEMENTS

GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the SDSS and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including *GALEX* MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP providing ultraviolet to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is http://www.gama-survey.org/.

SPJgratefully acknowledges the support of the Austrian Science Fund (FWF): project P 29514-N32.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Abraham R. G., van den Bergh S., Glazebrook K., Ellis R. S., Santiago B. X., Surma P., Griffiths R. E., 1996, ApJs, 107, 1
Adelman-McCarthy J. K. et al., 2008, ApJS, 175, 297
Baldry I. K., 2008, Astron. Geophys., 49, 5.25
Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469
Baldry I. K. et al., 2012, MNRAS, 421, 621
Baldry I. K. et al., 2014, MNRAS, 441, 2440
Balogh M. L., Navarro J. F., Morris S. L., 2000, ApJ, 540, 113
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Banerji M. et al., 2010, MNRAS, 406, 342
Boyd S., Vandenberghe L., 2004, Convex Optimization. Cambridge Univ. Press, Cambridge
Breiman L., 2001, Mach. Learn., 45, 5
Breiman L., Friedman J., Stone C. J., Olshen R. A., 1984, Classification and Regression Trees. CRC Press, Boca Raton, FL
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Cameron E., 2011, PASA, 28, 128
Cameron E., Driver S. P., Graham A. W., Liske J., 2009, ApJ, 699, 105
Campbell C., 2001, Stud. Fuzziness Soft Comput., 66, 155
Chabrier G., 2005, in Corbelli E., Palla F., Zinnecker H., eds, Astrophysics and Space Science Library, Vol. 327, The Initial Mass Function 50 Years Later. Springer-Verlag, Dordrecht, p. 41
Chen J., Pereverzyev S., Jr, Xu Y., 2015, Inverse Probl., 31, 075005
Colless M. et al., 2001, MNRAS, 328, 1039
Conselice C. J., 2003, ApJS, 147, 1
Cristianini N., Shawe-Taylor J., 2000, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge Univ. Press, Cambridge
de Vaucouleurs G., 1959, Handbuch Phys., 53, 275
Diaferio A., Kauffmann G., Balogh M. L., White S. D. M., Schade D., Ellingson E., 2001, MNRAS, 323, 999
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Dietterich T. G., Bakiri G., 1995, J. Artif. Intell. Res., 2, 263
Doi M., Fukugita M., Okamura S., 1993, MNRAS, 264, 832
Dressler A., 1980, ApJ, 236, 351
Driver S. P., Liske J., Cross N. J. G., De Propris R., Allen P. D., 2005, MNRAS, 360, 81
Driver S. P. et al., 2006, MNRAS, 368, 414
Driver S. P. et al., 2009, Astron. Geophys., 50, 12
Driver S. P. et al., 2011, MNRAS, 413, 971

Driver S. P. et al., 2012, MNRAS, 427, 3244
Gauci A., Zarb Adami K., Abela J., 2010, preprint (arXiv:1005.0390)
Gnedin O. Y., 2003a, ApJ, 589, 752
Gnedin O. Y., 2003b, ApJ, 582, 141
Graham A. W., Trujillo I., Caon N., 2001, AJ, 122, 1707
Graham A. W., Ciambur B. C., Savorgnan G. A. D., 2016, ApJ, 831, 132
Gunn J. E., Gott J. R., III, 1972, ApJ, 176, 1
Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning. Springer Series in Statistics. Springer-Verlag, New York
Hiemer A., Barden M., Kelvin L. S., Häußler B., Schindler S., 2014, MNRAS, 444, 3089
Hill D. T. et al., 2011, MNRAS, 412, 765
Holmberg E., 1958, Meddelanden fran Lunds Astron. Obs. Ser. II, 136, 1
Hopkins A. M. et al., 2013, MNRAS, 430, 2047
Hubble E. P., 1936, Realm of the Nebulae. Yale University Press, New Haven
Hubble E., Humason M. L., 1931, ApJ, 74, 43
Huertas-Company M., Rouan D., Soucail G., Le Fèvre O., Tasca L., Contini T., 2007, A&A, 468, 937
Huertas-Company M. et al., 2015, ApJS, 221, 8
Jeans J. H., 1919, Problems of Cosmogony and Stellar Dynamics. University Press, Cambridge
Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201
Kelvin L. S. et al., 2012, MNRAS, 421, 1007
Kelvin L. S. et al., 2014a, MNRAS, 439, 1245
Kelvin L. S. et al., 2014b, MNRAS, 444, 1647
Kriukova G., Panasiuk O., Pereverzyev S. V., Tkachenko P., 2016, Neural Netw., 73, 26
Lange R. et al., 2015, MNRAS, 447, 2603
Larson R. B., Tinsley B. M., Caldwell C. N., 1980, ApJ, 237, 692
Lauberts A., Valentijn E. A., 1989, The Surface Photometry Catalogue of the ESO-Uppsala Galaxies. European Southern Observatory, Garching
Liller M. H., 1966, ApJ, 146, 28
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Liske J. et al., 2015, MNRAS, 452, 2087
Moffett A. J. et al., 2016, MNRAS, 457, 1308
Moller M. F., 1993, Neural Netw., 6, 525
Moore B., Katz N., Lake G., Dressler A., Oemler A., 1996, Nature, 379, 613
Moore J. A., Pimbblet K. A., Drinkwater M. J., 2006, PASA, 23, 135
Morgan W. W., 1958, PASP, 70, 364
Moss C., Whittle M., 2000, MNRAS, 317, 667
Oemler A., Jr, 1974, ApJ, 194, 1
Okamura S., Kodaira K., Watanabe M., 1984, ApJ, 280, 7
Owens E. A., Griffiths R. E., Ratnatunga K. U., 1996, MNRAS, 281, 153
Park C., Gott J. R., III, Choi Y.-Y., 2008, ApJ, 674, 784
Pearson K. F., 1901, Phil. Mag., 2, 559
Reynolds J. H., 1920, MNRAS, 80, 746
Sandage A., 1961, The Hubble Atlas of Galaxies. Carnegie Institution, Washington, DC
Sandage A., Binggeli B., 1984, AJ, 89, 919
Sandage A., Sandage M., Kristian J., 1975, Galaxies and the Universe. University of Chicago Press, Chicago
Savorgnan G. A. D., Graham A. W., 2016, MNRAS, 457, 320
Shapley H., Paraskevopoulos J. S., 1940, Proc. Natl. Acad. Sci., 26, 31
Shapley H., Sawyer H. B., 1927, Harvard College Obs. Bull., 846, 1
Smail I., Dressler A., Couch W. J., Ellis R. S., Oemler A., Jr, Butcher H., Sharples R. M., 1997, ApJS, 110, 213
Smee S. A. et al., 2013, AJ, 146, 32
Smith G. P., Treu T., Ellis R. S., Moran S. M., Dressler A., 2005, ApJ, 620, 78
Sokolova M., Lapalme G., 2009, Inf. Process. Manage., 45, 427
Spitzer L., Jr, Baade W., 1951, ApJ, 113, 413
Storrie-Lombardi M. C., Lahav O., Sodre L., Jr, Storrie-Lombardi L. J., 1992, MNRAS, 259, 8P
Taylor E. N. et al., 2011, MNRAS, 418, 1587
Toomre A., 1977, in Tinsley B. M., Larson D., Campbell R. B. G., eds, Evolution of Galaxies and Stellar Populations. Yale University Observatory, New Haven, p. 401
van der Wel A., 2008, ApJ, 675, L13
Welikala N., Connolly A. J., Hopkins A. M., Scranton R., Conti A., 2008, ApJ, 677, 970
Welikala N., Connolly A. J., Hopkins A. M., Scranton R., 2009, ApJ, 701, 994
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
Wolf M., 1908, Publikationen des Astrophysikalischen Instituts Koenigstuhl-Heidelberg, 3, 109

APPENDIX A: METHODS IN DETAIL

A1 Support Vector Machines

In SVM, the constructed classifier f is of the following form:

$$f(\boldsymbol{x}) = \mathrm{sign}\,(g(\boldsymbol{x})) = \mathrm{sign}\left(b + \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)\right) \tag{A1}$$

where K is a positive definite kernel,[12] and b, α_i, $i = 1, 2, \ldots, N$ are certain coefficients from $\mathbb{R}$. We set $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots, \alpha_N)^\top$.

The coefficients b, $\boldsymbol{\alpha}$ are chosen as the solution of the following minimization problem:

$$\sum_{i=1}^{N} (1 - z_i g(\boldsymbol{x}_i))_+ + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} \longrightarrow \min_{b,\boldsymbol{\alpha}}, \tag{A2}$$

where $(a)_+ := \max(0, a)$, $\boldsymbol{K}$ is the $N \times N$ kernel matrix with entries $\boldsymbol{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $\lambda > 0$ is a penalty parameter. Note that the minimization problem in equation (A2) is convex, and therefore, various methods of convex optimization (e.g. Boyd & Vandenberghe 2004) can be used to solve it. We employ the *sequential minimal optimization* method, which is suggested in MATLAB as the standard method to solve this. The first term in equation (A2) measures the closeness of $f(\boldsymbol{x}_i) = \mathrm{sign}\,(g(\boldsymbol{x}_i))$ to z_i, i.e. it tells us how well the classification is predicted on the training set, while the second term penalizes coefficients in $\boldsymbol{\alpha}$, and λ gives a trade-off between the two terms. We take the default value for λ, which is $\lambda = 1$.

Due to the nature of the function $(\cdot)_+$ in equation (A2), many values of α_i are equal to 0. Therefore, in the representation in equation (A1), the linear combination involves functions from a subset of $\{K(\cdot, \boldsymbol{x}_i),\ i = 1, 2, \ldots, N\}$, and the corresponding $\boldsymbol{x}_i$ are called *support vectors*.

The kernel that we have chosen is the *Gaussian Radial Basis Function* given as:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right), \tag{A3}$$

where σ is the scaling factor, whose default value we have retained as $\sigma = 1$.

The SVM method is illustrated by an example in Fig. A1. In this example, the feature vector is $(x_1, x_2) \in [-1, 1] \times [-1, 1]$ and $D := \{(x_1, x_2) \mid x_1^2 + x_2^2 \leq (1/2)^2\}$ is a disc with its centre as (0, 0) and radius 1/2. The ideal classifier f^* assigns the feature vector to class 1 if it belongs to D, and to class -1 otherwise. We generate a training set $\mathcal{Z}$ that consists of 100 feature vectors $\boldsymbol{x}_i = (x_{1,i}, x_{2,i})$. Features $x_{1,i}$ and $x_{2,i}$ are randomly sampled using the uniform distribution over $[-1, 1]$. The classes for the feature

[12] Recall that a symmetric function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ (here symmetric means that $K(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2) = K(\bar{\boldsymbol{x}}_2, \bar{\boldsymbol{x}}_1)$ for any $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2 \in \mathbb{R}^p$) is called *positive definite kernel* if for any $m \in \mathbb{N}$ and any distinct $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_m \in \mathbb{R}^p$, the $m \times m$ matrix $\bar{\boldsymbol{K}}$ with entries $\bar{\boldsymbol{K}}_{ij} = K(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}_j)$ is positive definite.

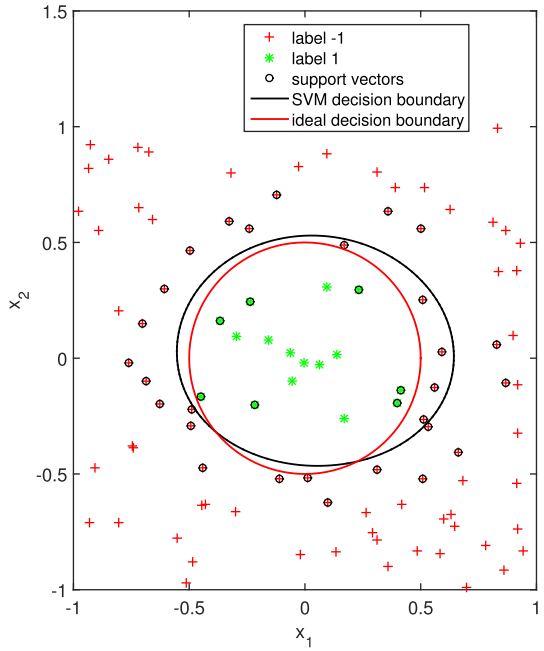

Figure A1. Illustrative example for the SVM method in the case of two features x_1 and x_2. The feature vector belongs to the class 1 if it is inside the red curve, which is the ideal decision boundary. Otherwise, the feature vector is in the class -1. The black curve is the decision boundary constructed by the SVM method using the training data in the picture. The corresponding constructed SVM classifier assigns the feature vectors inside this black curve to class 1, and outside the black curve to class -1. The training feature vectors that are additionally marked by a small surrounding circle are the support vectors.

vectors $\boldsymbol{x}_i$ in the training set are determined using the ideal classifier f^*. Then, we construct the function g using the described SVM method with kernel equation (A3) and $\sigma = 1$.

The red curve is the boundary of D, which can be called as the *ideal decision boundary*. The black curve consists of feature vectors $\boldsymbol{x}$ for which $g(\boldsymbol{x}) = 0$. This curve is called the *SVM decision boundary*. For the feature vectors inside this curve, we have $g(\boldsymbol{x}) > 0$, and therefore, these feature vectors will be classified by the constructed SVM classifier f as class 1. The other feature vectors satisfy the condition $g(\boldsymbol{x}) < 0$, and therefore, they are assigned by f to class -1. The training feature vectors inside the small circles are the support vectors.

In general, the SVM decision boundary may have an arbitrary shape, and it may also consist of several closed curves. The support vectors are located near the SVM decision boundary, in a way supporting and defining its shape.

A2 Classification Trees with hyper-rectangular partitions

In the CT method, the feature space is split into the rectangular partitions by a recursive binary method. First, it is split into two regions, $\{\boldsymbol{x} \in \mathbb{R}^p \mid x_i < s\}$ and $\{\boldsymbol{x} \in \mathbb{R}^p \mid x_i \geq s\}$ using a selected feature x_i and a split point s. Then, one or both of these regions are split similarly into two more regions. this process continues until a certain stopping condition is fulfilled.

An example of the above-described partition for two features (x_1, x_2) with values in the unit square is presented in Fig. 8. In this example, the first split is made at $x_1 = s_1$. Then, the region $\{\boldsymbol{x} \in \mathbb{R}^p \mid x_1 < s_1\}$ is split at $x_2 = s_2$, and the region $\{\boldsymbol{x} \in \mathbb{R}^p \mid x_1 \geq s_1\}$ at $x_1 = s_3$. In the end, the region $\{\boldsymbol{x} \in \mathbb{R}^p \mid x_1 \geq s_3\}$ is split at $x_2 = s_4$. Thus the partition of the feature space into five rectangular regions $R_1, R_2, \ldots, R_5$ shown in Fig. 8 is obtained.

The nodes of the CT are split based on the *impurity measure* of the node. We represent the region that corresponds to the node t as R_t. Let $N_t := \#\{\boldsymbol{x}_i \in R_t\}$ denote the number of training feature vectors in R_t. The mathematical notation $\#R$ is used for the number of elements of a set R. We further define $p_k(t) := \#\{\boldsymbol{x}_i \in R_t \mid y_i = k\}/N_t$ as the proportion of the training feature vectors in the node t (or, which is the same, in the region R_t) that belong to class k. Impurity measure $I(t)$ of the node t is a function of the proportions $p_k(t)$. It tells us how even the distribution of the feature vectors in the node t are over the classes. It has a maximum value when the feature vectors are distributed evenly over the classes in the node t, i.e. when $p_k(t) = 1/T$, $k = 1, 2, \ldots, K$. In contrast, when the node t contains feature vectors only from one class, say class ℓ, i.e. when $p_\ell(t) = 1$, and $p_k(t) = 0$, $k \neq \ell$, then the impurity measure $I(t)$ has a minimal value, and the node is called *pure*. As the impurity measure, we consider the Gini index : $I(t) = 1 - \sum_{k=1}^{T} p_k^2(t)$.

The goal of the node splitting is to obtain new nodes with smaller impurity measures. This is achieved by defining a characteristic called impurity gain, and the splitting is then done such that this gain is maximized. Let $P(R_t) = N_t/N$ denote the proportion of the training feature vectors in the node t. Consider a particular splitting candidate of the node t, i.e. a particular splitting feature and a split point, and denote the corresponding new left node as t_1 and the new right node as t_2. Then, the impurity gain is defined as :

$$\Delta I = P(R_t)\, I(t) - P\left(R_{t_1}\right) I(t_1) - P\left(R_{t_2}\right) I(t_2), \tag{A4}$$

and then, the splitting candidate for which this impurity gain is maximum is chosen.

There is a finite number of splitting candidates. For each feature x_q, $q = 1, \ldots, p$, possible splitting points are obtained from the training data by sorting $x_{i,q}$ in the ascending order. Note that $x_{i,q}$ is the qth component of the feature vector $\boldsymbol{x}_i$, and those feature vectors $\boldsymbol{x}_i$ are considered that belong to the splitting node. Then, the maximization of the impurity gain (equation A4) is done by checking through all possible splitting candidates.

A3 Classification Trees with Random Forest

Using random sampling *with replacement*, B random samples $\mathcal{Z}_b$, $b = 1, \ldots, B$ of the training set $\mathcal{Z}$ are created. Consider the set $\tilde{\mathcal{Z}} = \{1, 2, \ldots, 10\}$ with 10 elements. Then, four random samples of $\tilde{\mathcal{Z}}$ that are made using the random sampling with replacement of 10 elements from $\tilde{\mathcal{Z}}$ can be, for example, the following:

$$\tilde{\mathcal{Z}}_1 = (3, 4, 9, 6, 3, 8, 4, 9, 3, 10),$$
$$\tilde{\mathcal{Z}}_2 = (2, 8, 4, 7, 5, 5, 2, 4, 7, 8),$$
$$\tilde{\mathcal{Z}}_3 = (1, 1, 2, 4, 10, 5, 6, 2, 8, 10),$$
$$\tilde{\mathcal{Z}}_4 = (2, 3, 8, 4, 7, 6, 6, 7, 2, 8).$$

On each training sample $\mathcal{Z}_b$, a CT classifier f_b is trained using a modified CT learning algorithm. In this modified algorithm, at each node split, possible splitting features are taken from a random sample of all used features. Typically, these random samples contain $\sqrt{p}$ (rounded down) features (p is the number of features, Section 3.1). The resulting CTRF classifier assigns a classification to a feature vector $\boldsymbol{x}$ using the majority vote of the constructed CT classifiers $\{f_b,\ b = 1, \ldots, B\}$.

A4 Single hidden layer feed-forward neural networks

The arrows in the network diagram, Fig. 10 indicate the dependence between network units, and this dependence is modelled as:

$$w_{\rm m} = g_1\left(\alpha_{0\rm m} + \boldsymbol{\alpha}_{\rm m}^{\top}\boldsymbol{x}\right),\ m = 1, 2, \ldots, M,$$
$$\bar{v}_{\rm k} = \beta_{0\rm k} + \boldsymbol{\beta}_{\rm k}^{\top}\boldsymbol{w},\ {\rm k} = 1, 2, \ldots, T,$$
$$v_{\rm k} = g_{2,\rm k}\left(\bar{\boldsymbol{v}}\right) =: f_{\rm k}\left(\boldsymbol{x}\right),\ k = 1, 2, \ldots, T,$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_{\rm M})^{\top}$, and $\bar{\boldsymbol{v}} = (\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{\rm T})^{\top}$. The numbers $\alpha_{0\rm m}, \beta_{0\rm k}$ and vectors $\boldsymbol{\alpha}_{\rm m} \in \mathbb{R}^p, \boldsymbol{\beta}_k \in \mathbb{R}^M$ are model parameters called weights. The complete set of these weights is denoted by $\boldsymbol{\theta}$. The functions g_1 and $g_{2,\rm k}$ are called transfer functions. For g_1, we take the tan-sigmoid transfer function:

$$g_1(s) = \frac{(\exp(s) - \exp(-s))}{(\exp(s) + \exp(-s))},$$

and for $g_{2,k}$, we take the softmax transfer function,

$$g_{2,\rm k}\left(\bar{\boldsymbol{v}}\right) = \frac{\exp\left(\bar{v}_{\rm k}\right)}{\sum_{\ell=1}^{T}\exp\left(\bar{v}_{\ell}\right)}.$$

The softmax transfer function ensures that the unit values $v_{\rm k}$ belong to the interval (0, 1) and satisfy $\sum_{k=1}^{T} v_{\rm k} = 1$, which allows $v_{\rm k}$ to be interpreted as the probability to belong to class k. The mentioned conditions on $v_{\rm k}$ require the second transfer function $g_{2,\rm k}$, in contrast to the first transfer function g_1, to vary with k.

Once the weights $\boldsymbol{\theta}$ of the NN are chosen, the NN classifier is defined as:

$$f\left(\boldsymbol{x}\right) = \underset{k}{\rm argmax}\ f_{\rm k}\left(\boldsymbol{x}\right)$$

i.e. for any feature vector, the class with the highest probability is taken.

During the training process, the weights $\boldsymbol{\theta}$ of the NN are tuned such that the error function $E(\boldsymbol{\theta})$ is minimized. The error function describes how well the NN model fits the training data. As the error function, we consider the cross-entropy function:

$$E(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{k=1}^{T} v_{\rm ik}\log f_{\rm k}\left(\boldsymbol{x}_{\rm i}\right),$$

where $v_{\rm ik} = 1$ if $y_{\rm i} = k$, and $v_{\rm ik} = 0$ otherwise. The minimization of the error function can be done by gradient-based methods. We use the scaled conjugate gradient backpropagation algorithm (Moller 1993) which is suggested in MATLAB for tuning NN used for classification problems.

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