Advancing Machine Learning in Astrophysics

Mike Walmsley
Christ Church
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2021

Abstract

This thesis explores four projects applying supervised deep learning to help answer astrophysical questions.

I first consider faint tidal features. Tidal features are a long-lasting signature of galaxy mergers, making them useful for measuring merger rates. However, current automated methods struggle to detect faint tidal features in complex galaxies. I use convolutional neural networks to identify galaxies with tidal features in the CFHTLS-Wide Survey, improving on previous methods applied to the same dataset. I show that my networks can identify which pixels are associated with tidal features, potentially enabling researchers to not only identify but also characterise tidal features.

I then turn to Galaxy Zoo, a citizen science project measuring galaxy morphology. Galaxy Zoo is being gradually outpaced by the increasing scale of new surveys. Automated classifiers can be trained using volunteer responses; however, such classifiers are often unable to consider uncertainty in either volunteer responses or predictions, leading to wasted volunteer effort and overconfident classifications. I introduce a probabilistic approach that allows classifiers to flexibly express uncertainty. I use this probabilistic approach to build a machine learning system that 'asks' volunteers to label the galaxies it could best learn from. I relaunch Galaxy Zoo with images from the Dark Energy Camera Legacy Survey and run my system live, collecting 1.8 million volunteer responses. My final models are around 99% accurate on every question for galaxies with confident volunteer answers and are otherwise correctly uncertain.

Next, I help the Canadian Hydrogen Intensity Mapping Experiment (CHIME) detect fast radio bursts. CHIME only attempts to detect FRB above a signal-to-noise threshold of $\sigma = 8.5$, in part for lack of expert time to review candidates. I created and launched a citizen science project to classify the $7.8 \leq \sigma < 8.5$ signal-to-noise candidates detected by CHIME each week. Candidates found by this project may be the most distant fast radio bursts ever detected, which I hope will serve as useful cosmological probes of the intergalactic medium.

Finally, I show that neural network emulation can efficiently recover posteriors of galaxy parameters from photometry. Galaxy SED simulators are too slow to use MCMC inference on large samples. I train a neural network to emulate an SED simulator, providing both faster likelihood evaluations and known gradients. These gradients can then be used for efficient Hamiltonian Monte Carlo inference.

Together, these projects show how deep learning can help astronomers make effective use of limited and uncertain data.

Statement of Originality

I carried out the work presented in this thesis as a student at the University of Oxford between October 2017 and February 2021, supervised by Prof. Chris Lintott. Prof. Chris Lintott provided guidance throughout this work. It was funded by a Science Technology Facilities Council Studentship Grant Code ST/R505006/1. I hereby declare that no part of this thesis has been submitted in support of another degree, diploma or other qualification at the University of Oxford or other higher learning institute. Except where otherwise stated or where reference is made to the work of others, the work in this thesis is entirely my own.

The work in Chapter 2 is based on the peer-reviewed paper Walmsley et. al 2019, MNRAS 483(3):2968-2982 [443], for which I am the lead author and contributed all of the technical work. The classifier was developed and early results obtained as part of my MPhys degree at the University of Edinburgh, under the supervision of Prof. Annette Ferguson and Prof. Robert Mann. The results were finalised and the paper written and accepted at Oxford.

The work in Chapter 3 is based on the peer-reviewed paper Walmsley et. al. 2020, MNRAS 491(2):1554-1574 [445], for which I am the lead author and contributed all of the technical work except for the derivation of the Binomial loss function (which was done jointly with Lewis Smith) and the appendix on variational inference (not included here). Lewis Smith and Prof. Yarin Gal both made substantial contributions through helpful technical discussions.

The work in Chapter 4 is based on the peer-reviewed paper Walmsley et. al. 2021 [446], recently accepted by MNRAS and available on arxiv. I am the lead author and wrote at least 95% of the text. I was responsible for the relaunch and day-to-day operation of the Galaxy Zoo project until October 2020. I carried out all of the data analysis and developed the automated classifier. Dr. Kyle Willett and Dr. Coleman Krawcyzk each ran previous iterations of the project; I prepared and released those classifications as part of the data release. Dr. Sandor Kruk, Dr. Lee Kelvin and Prof. Steven Bamford carried out redshift debiasing work important to the data release but not included here. Tobias Géron assisted with preparing data. The Galaxy Zoo science team made substantial contributions through helpful technical discussions and ad-hoc technical assistance.

The work in Chapter 5 was carried out in collaboration with Canada Hydrogen Intensity Mapping Experiment (CHIME) scientists. Dr. Paul Scholz provided guidance and Arecibo data to support the project proposal. The text for the citizen science

project was based on an unpublished prototype project by Robert Archibald. Chitrang Patel helped to prepare the telescope data. Dr Shriharsh Tendulkar provided expert reviews of 24 candidate fast radio bursts.

The work in Chapter 6 was carried out in collaboration with Dr. Sotiria Fotopoulou, with helpful discussions from Dr. Ivana Damjanov, Dr. Nesar Ramachandra, Dr. Nic Ross, and Prof. Yuan-Sen Ting. Dr. Sotiria Fotopoulou contributed the initial dataset.

I would like to thank the Galaxy Zoo and Bursts from Space volunteers, without whom neither project would have been possible.

This research made use of the open-source Python scientific computing ecosystem, including SciPy [205], Matplotlib [195], scikit-learn [336], scikit-image [437] and Pandas [298]. This research made use of Astropy, a community-developed core Python package for Astronomy [419]. This research made use of TensorFlow [2] and PyTorch [331]. This research used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, and other institutions acknowledged at www.legacysurvey.org. The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration. The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.

Acknowledgements

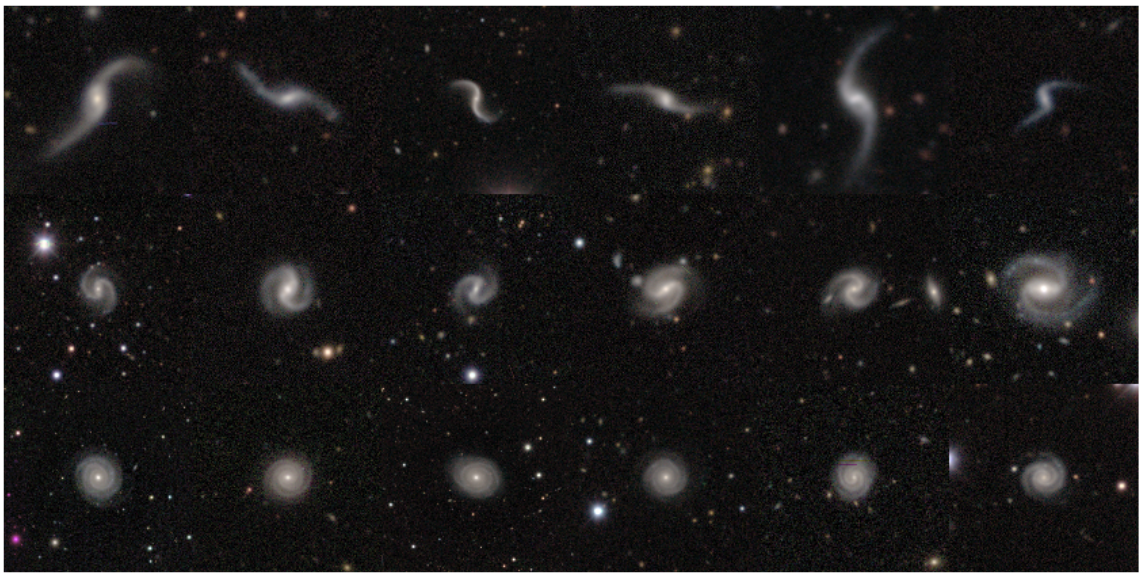
No-one writes a thesis alone. I am grateful for the guidance of my supervisor, Prof. Chris Lintott. I sent my final thesis draft to Chris around midnight; he sent comments back by 9am, which tells you everything you need to know about him as a supervisor. I am grateful for the care and support of Bianca, who has encouraged, cheered up, dragged forwards, and gently mocked me over many years. And I am grateful for my family, particularly my parents, John and Lisa, and my brother, David, for helping me make my way through the world.

This thesis is dedicated to my grandfather, Murray Mackson. Murray always wanted to see me graduate, but sadly died of COVID-19 earlier this year. I hope that we learn from this pandemic to seek truth, listen to evidence, and value life.

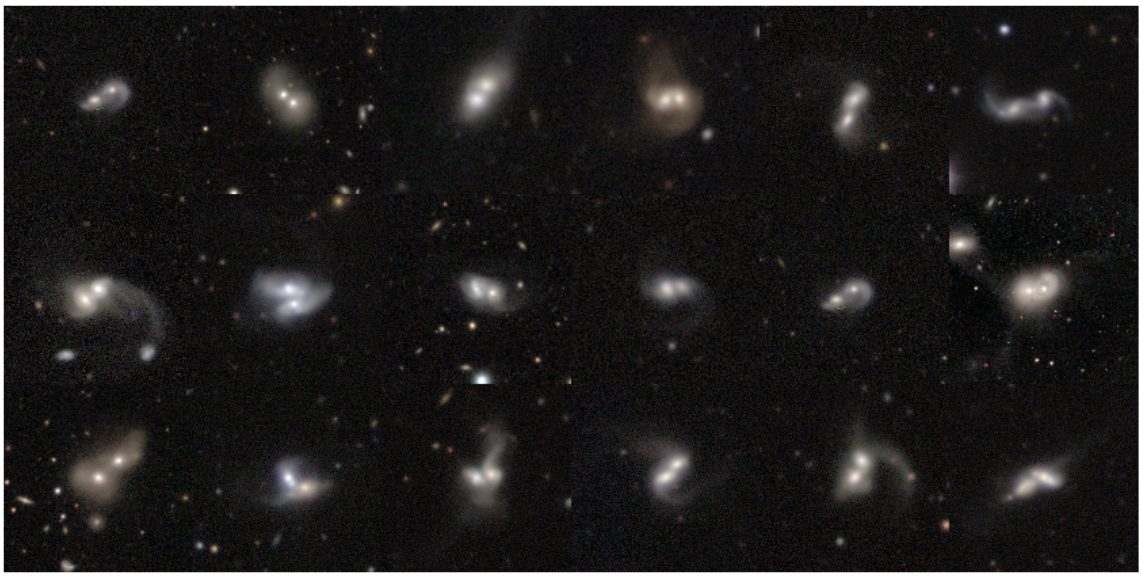
Contents

"*Some other areas have been identified in which neural networks might potentially be applied to astronomical problems. The most obvious of these are supervised classification of spectra or, more ambitiously, morphological classification of galaxies.*"

H.-M. Adorf, 1991 [8]



(a) GZ DECaLS galaxies automatically classified as most likely (highest mean posterior) to be two-armed spirals with loose (upper), medium (middle) or tight (lower) winding. See Chapter 4.



(b) GZ DECaLS galaxies automatically classified as most likely (highest mean posterior) to be mergers. See Chapter 4.

Chapter 1

Introduction

This thesis explores my application of supervised deep learning to help answer astrophysical questions. I have worked on questions in three distinct areas - detailed galaxy morphology, fast radio bursts, and photometric inference - into which this thesis is divided. I will introduce the scientific background for each area at the start of each section, explaining why an automated solution is desirable and why I thought supervised deep learning might work well. First, it is necessary to introduce supervised deep learning itself.

Below, I give a practical overview of supervised learning in general and highlight how supervised deep learning has several inherent limitations - overfitting, uncertainty, and shortcut learning - which motivate much of my work in later chapters. I then describe convolutional neural networks, a specific class of supervised deep learning model that I use throughout this thesis. Finally, I step back and consider the broad history of neural networks in astronomy and the impact that convolutional neural networks have recently had on the field.

1.1 Supervised Learning

Consider the problem of making a judgement about an image. We make many judgements effortlessly (it's a galaxy, it has two spiral arms, there's a little green man in the corner, etc.) but it is difficult to express exactly *how* we do so. We cannot easily write down a series of steps to reliably replicate our responses for other images. Instead, supervised learning aims to discover such steps automatically based on example images (or other data) for which the answers are known.

More formally, given data $\mathcal{D}$, supervised learning aims to fit (train) a mathematical model to a subset $\mathcal{D}_{\text{train}}$ where the answers (labels) are known, in the hope that the same model will also be able to make generalised predictions on the unknown

remainder. The model $f(\mathbf{x};\theta)$ makes predictions on datapoint $\mathbf{x}$ using parameters θ. The deviation between the predictions of the model and the known labels are measured using a loss function $L(f(\mathbf{x};\theta),y)$. For example, for binary classification problems, a standard choice is the binary cross-entropy

$$L = y\log p + (1-y)\log(1-p) \tag{1.1}$$

where I have suggestively set $f(\mathbf{x};\theta) = p$ to highlight that, if one treats the model output as a probability, the total cross-entropy over a training dataset of size N can be shown to be proportional to the negative log-likelihood of a series of independent events $\{y_1,...,y_N\}$ each occuring with probability $\{p_1,...,p_N\}$. Using the negative log-likelihood as a loss function is crucial to training probabilistic models; this is the starting point for my derivation of loss functions for Galaxy Zoo (Chapter 3).

To fit the model (i.e. to make a maximum likelihood estimate of the model parameters θ assuming a suitable loss function), one might minimise the total loss over the training set with respect to θ using numerical methods (typically, specialised variations of stochastic gradient descent e.g. Adam, [230]). However, calculating $\frac{\partial L}{\partial \theta}$ over the full training set has a computational cost of order N but only decreases our uncertainty in $\frac{\partial L}{\partial \theta}$ by a factor of $\sqrt{N}$. Instead, one typically calculates the expected loss for the training dataset (i.e. the empirical risk) over a training subset of size M:

$$\mathbb{E}_{(\mathbf{x},y)\sim\hat{p}_{\text{train}}}[L(f(\mathbf{x};\theta),y)] = \frac{1}{m}\sum_{i=1}^{m} L(f(\mathbf{x}^{(i)};\theta),y^{(i)}) \tag{1.2}$$

One then minimises Eqn. 1.2 with respect to θ repeatedly for many training subsets (known as batches). The additional noise introduced in $\frac{\partial L}{\partial \theta}$ has been shown empirically to help avoid overfitting (see below) and improve performance [149].

The overall process to train and evaluate a model consists of two nested loops: the inner minimisation loop described above, and an outer validation loop (Figure 1.1). With every iteration of the minimisation loop, the network is gradually fit to the training data. A batch of labelled data is given as input to the model, the model returns predictions, and the quality of these predictions is measured using the loss function. The gradient of the loss function for that batch with respect to the model parameters is computed, and the model parameters are then updated to minimise the loss function. The loop then repeats for a new batch of labelled images. Once a specified number of minimisation loops have elapsed (often equal to the number of batches in the training dataset, referred to as an 'epoch'), the validation loop is executed.

For every iteration of the validation loop, the network makes predictions for a batch of 'unseen' validation data where the labels are known but not used in the minimisation process. Metrics for the quality of these predictions (for example, the mean accuracy) are recorded. The training process (i.e. multiple minimisation loops) is then restarted. The algorithm continues until a stopping criterion is reached; typical criteria are when a fixed number of validation loops has elapsed, or when no further decrease in the validation loss is recorded ('early stopping'). Ideally, one then tests performance on a final unseen labelled 'test' dataset as a proxy for new unlabelled data.

Several drawbacks to supervised learning are already clear. For instance, we cannot calculate the loss function over the full dataset $\mathcal{D}$ as we do not know the value of y for all datapoints. We can only train the model on the labelled subset (training dataset). A model may therefore learn features of the training dataset which do not generalise to the test dataset. In the limiting case, a model with more parameters than training labels may memorise those labels, leading to excellent predictions on the training dataset but poor predictions on the test dataset [25, 318][1]. This problem is known as overfitting and is familiar from the general statistical problem of fitting an overly complex model to limited data. Overfitting is a significant challenge for learning to classify detailed galaxy morphology from limited expert data; I focus on this issue in Chapter 2.

Overfitting is closely related to uncertainty. Some unlabelled datapoints may be sufficiently unlike the training dataset that our model cannot make good predictions; these are known as 'out-of-distribution' examples. We would like our model to express uncertainty or otherwise flag these unusual datapoints (which, by virtue of being unusual, are often of particular scientific interest). But complex models which have overfit (learned rules that generalise poorly) may make confidently wrong predictions, where the actual prediction depends in part on random choices (e.g. weight initialisation) before or during training [137]. Such random choices may in principle be marginalised over by training many models, but supervised deep learning is computationally expensive and so most authors typically train a single model or a few models. Nonetheless, it is possible to use this effect to our advantage to detect unlabelled datapoints which are unlike the training dataset, label them, and thereby improve model

[1]This memorisation ability was recently exploited by Carlini et al. 2020 [70] to extract various names and phone numbers included in the training corpus of GPT-2, a large language model, by finding input prompts which caused the model to repeat the memorised personal data.

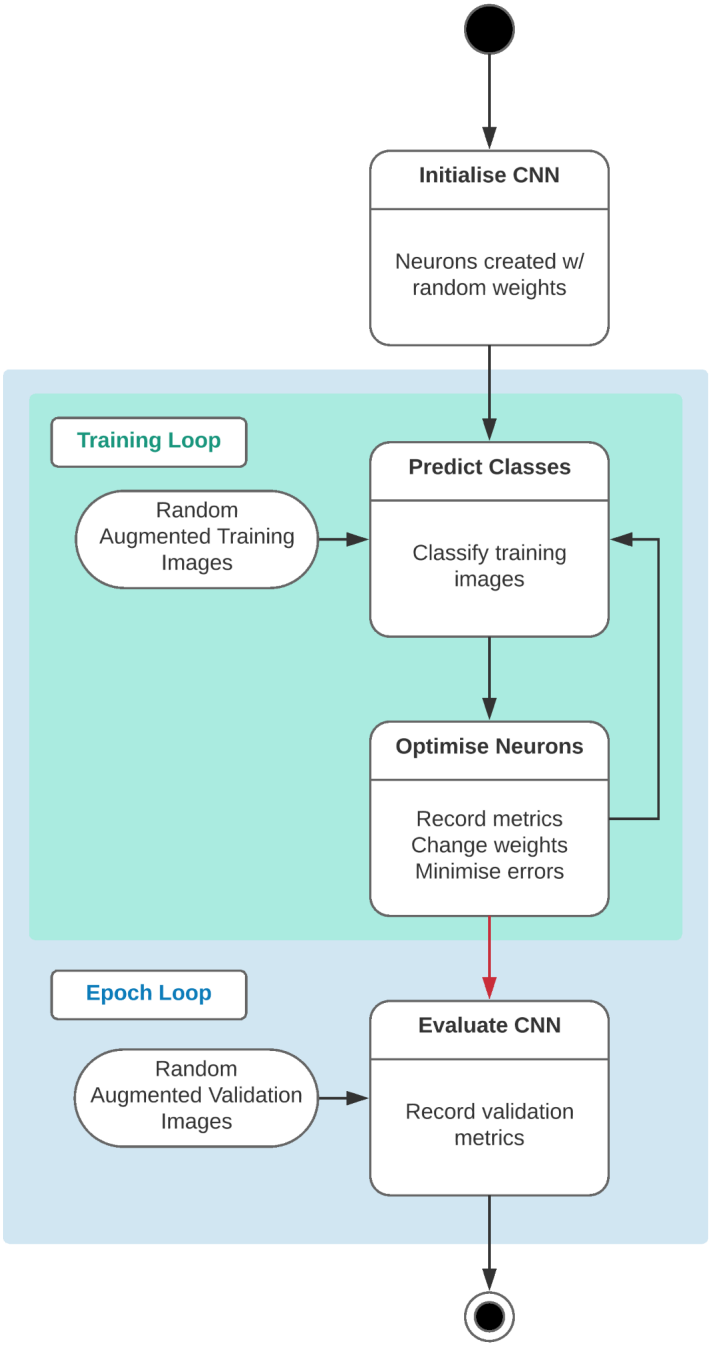


Figure 1.1: Flow chart of generic supervised learning approach. Here, a convolutional neural network is trained to make predictions on images. Red arrows denote steps which only occur after specified iterations have elapsed.

performance on similar datapoints. This is a form of active learning [182], which I apply in Chapter 3.

We may also be uncertain about the training labels; we often do not know the correct answer with absolute confidence. Much computer science research focuses on benchmark datasets explicitly constructed to only include confident examples, such as CIFAR [236] and ImageNet [365]. In my view, this has led to the importance of label uncertainty being somewhat overlooked. State-of-the-art models trained on ImageNet perform significantly worse on comparable images for which human annotators were fractionally less confident[2], even in 'the benign environment of a carefully-controlled reproducibility experiment' (Recht et al. 2019 [353]). Managing label uncertainty is crucial to the practical use of supervised deep learning. Scientific applications (as well as many important practical applications, e.g. in medical diagnosis [111, 234] or self-driving car perception [222]) often have highly heteroskedastic datasets where the desired answers for some datapoints are confidently known and some are significantly uncertain. Galaxy Zoo volunteer classifications (Sec. 3.1.1) are a prime example. Whether or not a particular question is asked depends on the volunteer's answers to previous questions. In consequence, very different numbers of volunteers may be asked a question for different galaxies, and so the typical volunteer response (from which the target label is derived, e.g. [100, 370]) will be well-constrained for some galaxies and poorly-constrained for others. I introduce a specifically-designed loss function to address this issue in Chapter 3, and further develop this loss function in Chapter 4.

Finally, having made our predictions, the lack of an explicit decision rule makes it difficult to understand why those particular predictions were made. Model interpretability is particularly important in a scientific context when predictions are used to control data collection where decisions cannot be changed later (e.g. to guide follow-up [198, 224] or reduce temporary raw data [213, 309]), or to search for correlations in large datasets where subtle biases in the data may introduce spurious correlations in the model (e.g. for weak lensing shape analysis [199, 356] or galaxy morphology [265, 445, 456]). In the extreme case, a model may learn to make predictions through undesired 'shortcuts' that work effectively but undermine the intended use. Common shortcuts include classifying based on the background (e.g. detecting sheep by identifying grassy fields [381]), based on texture [28, 55], and based on highly predictive

[2]Specifically, on images selected by an average of 71% versus 73% of human labellers as belonging to a given class. See Recht et al. 2019 [353] for further details.

yet imperceptible features [197]. In Chapter 5, I discover and then mitigate shortcut learning in a model used by the CHIME telescope [18] to detect fast radio bursts.

1.2 Convolutional Neural Networks

Having outlined the broad concept and drawbacks of supervised learning, I now describe the specific class of model $f(\mathbf{x};\theta)$ used throughout this thesis: neural networks, particularly convolutional neural networks (CNNs). Neural networks were introduced by Rosenblatt 1957 [363] and CNNs by LeCun et al. 1989 [252]. Here, I focus on a practical overview; I give a historical review in the next section (Sec. 1.3).

Neural networks are composed of repeated tensor operations called layers. The output of layer l, $\mathbf{x}^l$, is the input to layer $l+1$ and the arrangement and connectivity of layers are called the architecture. The net effect of a neural network is a non-linear mapping from input tensor to final layer output (i.e. prediction), with the aim being to learn the mapping which gives the true predictions.

Each type of layer performs a different operation. For example, the most basic is the fully-connected layer which performs the operation:

$$\mathbf{x}^{(l)} = f(\mathbf{w}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)}) \tag{1.3}$$

Consider the classification of an image using fully-connected layers. The image is encoded by the tensor of pixel values $\mathbf{x}^{(\mathbf{0})}$. This input propagates forward through the layers and is modified by the weights $\mathbf{w}^{(l)}$ and biases $\mathbf{b}^{(l)}$ of each layer through repeated operations of Eqn. 1.3. The output of the final layer is interpreted as predictions for that image. The weights and biases can then be optimised to minimise the empirical risk (Eqn. 1.2 through stochastic gradient descent, where the gradients are efficiently calculated by repeated application of the chain rule (known as backpropogation because the gradients are calculated layer-by-layer from the output [329, 364, 452]).

By designing a model as a sequence of non-linear transformations (layers of neurons), the model is (in principle) able to learn a hierarchical representation of the data, starting from the raw input and adding layers of gradually increasing abstraction [253]. For example, a hierarchical representation of an image might start with layers representing the textures and edges, then the shapes and areas, then semantic components. Convolutional neural networks are designed to learn such representations from images. More formally, CNNs are a neural network variant frequently used

to identify patterns in tensors (i.e. n-dimensional arrays) where the spatial arrangement of values is important. These tensors are most commonly the RGB pixel values of images, but may equally well be two-dimensional time series (such as a set of stock prices over time, or a lightcurve measured in several bands), abstract encodings (e.g. the positions of pieces in a board game [386]), and so forth. They routinely show state-of-the-art performance on various image classification benchmarks that require making subtle distinctions between classes and ignoring background effects [365].

Convolutional neural networks typically include two additional types of layer: convolutional and pooling. The convolutional layer operation can be described as

$$\mathbf{x}_n^{(l)} = f(\sum_i \mathbf{w}_{ij}^{(l)} * \mathbf{x}_i^{(l-1)} + \mathbf{b}_i^{(l)}) \tag{1.4}$$

where $\mathbf{w}_{ij}^{(l)}$ the filter of layer l.

Convolutional layers identify features with a fixed scale relative to the filter size. On the other hand, pooling layers reduce the size of a feature map by aggregation, for example by preserving only the local 2x2 maxima (as in this work). When alternated with convolutional layers, pooling layers allow for features of increasing spatial scale to be detected. Together, convolutional and pooling layers create increasingly abstract feature maps that encapsulate the image content. These features may then be classified using fully-connected layers. A toy CNN illustrating each operation is shown in Figure 1.2.

Convolutional neural networks are designed to learn hierarchical representations of spatially-arranged data which can then be used for tasks like classification. hierarchical representations have a major advantage over hand-designed features and 'shallow' machine learning, such as Random Forests [54], because the conversion from pixels to representation need not be a single step. This is thought to explain why, in general, deeper neural networks outperform shallower neural networks [253]. The degree to which CNNs succeed in learning hierarchical representations may vary; empirical work shows that CNNs can perform well when constrained to learning textures alone [55]. However, methods to visualise the learned representations provide strong empirical evidence that CNNs do indeed learn to classify based on recognising edges and textures, then shapes, and then semantic components [410].

Identifying hierarchical representations in images is a general purpose tool that can be applied across many domains, while handcrafted-features are problem-specific by definition. However, applying neural networks (and machine learning in general)

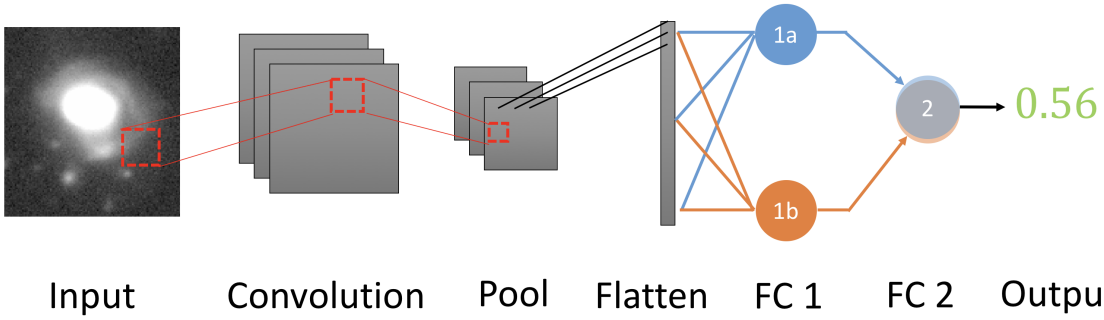


Figure 1.2: Illustrative diagram of a toy CNN. The pixel values of a galaxy image are taken as input. These are convolved with filter matrices to create feature maps. The feature maps are reduced in size by a pooling operation, then 'flattened' and concatenated into one dimension. This flattened list of abstract features is the input for two fully-connected layers with two and one neurons respectively. The final fully-connected layer outputs a prediction. In practice, the convolutional and pooling operations would repeat several times and the first fully-connected layer would include of order 100 or more neurons.

still benefits from domain expertise; I hope to show in this thesis that thoughtfully considering the scientific goal, context, and data is important for practical use.

1.3 A Brief History of Neurons in Astronomy

Astronomers have been interested in neural networks for almost as long as modern neural networks have existed. The single-layer neural network was introduced in the mid-20th century [363] to significant acclaim; the New York Times described it as 'the embryo of an electronic computer that [the US Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence' [423]. However, subsequent research showed that shallow neural networks were fundamentally limited in expressiveness [305]. Interest through the 1970s shifted towards 'knowledge-based' models able to reason with formal logic and symbols; the 'AI winter' of the 1980s followed the limited[3] success of such models [253]. Neural networks then began a resurgence in the late 1980s, partly due to the invention and populatisation of the backpropogation algorithm [329, 364, 452] which enabled deeper and hence more expressive networks to be trained.

This growth in interest was quickly matched by astronomers. The first published use of neural networks in astronomy may be Adorf and Meurs (1988) [9], who aimed

[3]It would be unfair to say symbolic models were unsuccessful; IBM's Deep Blue [68], for example, was arguably a symbolic model.

to classify objects in the IRAS Point Source Catalog [36, 319]. Just five years later, the first review of 'neural network applications in astronomy', Miller (1993) [304], cited 29 published projects, primarily in adaptive optics, telescope scheduling, object classification, and event filtering. One such project [407] is the first instance of galaxy morphology classification with neural networks. Miller 1993 also wrote that 'manual classification of large numbers of objects for training and testing networks represents by far the greatest overhead in developing such classifiers', presaging the usefulness of citizen science for creating training data. By 2009, Ball and Brunner 2009 [30] described neural networks as 'the most widely known and well-used machine learning algorithm in astronomy to-date'.

Meanwhile, new computer science research showed it was possible to train much deeper network designs (now recognised as 'deep learning') [173, 175]. One design in particular was easy to train and generalised well - convolutional neural networks (Sec. 1.2). Convolutional neural networks showed early promise with tasks including handwriting and facial recognition [250, 387], and exploded in popularity following dramatic success on the 2012 ImageNet computer vision competition[4] [253]. The winning solution, AlexNet [237], included many elements that remain standard practice today. CNNs were first used in astronomy shortly afterwards in seminal work by Dieleman et al. 2015 [100], also to win a competition - in this case, the Kaggle 'Galaxy Challenge'[5], where participants were invited to replicate classifications made by Galaxy Zoo volunteers. Dieleman et al. 2015's results encouraged other authors - including this one, as an undergraduate - to apply convolutional neural networks to classify other galaxies [191] and then to address other vision tasks.

It is hard to overstate the breadth of astronomical problems to which CNNs have subsequently been applied. Classifying images may be the most obvious use. CNNs continue to dominate the field of general galaxy morphology classification; I focus on this task in Chapter 3. But CNNs have also been used for more specific galaxy classification tasks. Star-galaxy separation based on images was another early application [229]. CNNs have been used to identify mergers at high [79, 125] and low [335] redshift, and to distinguish low-surface-brightness galaxies from artifacts [417]. Identifying low surface brightness galaxies is the focus of Chapter 2.

CNNs have also been widely applied to infer parameters from images (i.e. regression). Estimating photometric redshifts, a crucial task for the cosmology science goals

[4]Specifically, the ImageNet Large Scale Visual Recognition Challenge, ILSVRC, which is based on the ImageNet dataset but uses a reduced list of 1000 classes

[5]https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

of future surveys [248, 275], was among the first CNN applications. Hoyle (2016) [184] aimed to improve on existing work (see e.g. Salvato et al. 2019 [369] for a recent review) by exploiting additional information available in the raw images, beyond the tabular data of derived source colours. D'Isanto and Polsterer (2018) and Pasquet et al. (2019) extend this approach with additional attention towards probabilistic predictions and biases. Note that meaningful images may not be available for all sources (or might be avoided due to bias concerns); dense neural networks [367], Gaussian processes [16] and boosted trees [93] are all popular and well-suited to low-dimensional regression based only on source colours [375]. Combining both catalogue and image data (more formally, data fusion) is similarly helpful in other topics; for example, to improve colour-based mass-to-light estimates using galaxy morphologies [103].

CNNs are particularly powerful when training data can be simulated, alleviating the issue of gathering sufficiently large and diverse training sets. Strong gravitational lenses, for example, are straightforward to simulate, and so the strong lensing community have achieved significant success in training CNN to recognise strong lenses with simulated data [94, 246, 341, 373]. Hezaveh et al. 2017 [171] pioneered using CNN to estimate strong lens parameters (position, Einstein radius, ellipticity) from images with confirmed lenses, ten million times faster than with traditional maximum-likelihood methods. More recent work has, as with redshift estimation, focused on making estimates with reliable uncertainties [49, 256]. Weak lensing is also particularly amenable to simulations. Similarly to the previously-mentioned 'Galaxy Challenge', several public competitions for shear inference with simulated data have been held [57, 159, 232], the last of which was won by using dense neural networks to tweak conventional maximum-likelihood fits [233]. CNNs have been used extensively in recent years. For example, Springer et al. 2020 [403] demonstrate inference on mock HST observations of real background sources undergoing simulated lensing by the CLASH galaxy cluster [347]. Jeffrey et al. 2019 [199] learned to predict dark matter distributions from simulated DES [420] observations, with promising performance on validation simulations. Ribli et al. 2019 [356] also learned to predict dark matter distributions with DES observations, but instead of using simulations as labels, they used the partial overlap with deeper CFHTLens [118] observations where shape measurements are confidently known.

CNN can themselves act as proxies for simulations, predicting the outputs a simulation would produce given certain inputs. This approach, known as emulation, is often done to replace slow simulators with far faster models. Neural networks are effective emulators; in Chapter 6, I use a dense neural network for low-dimensional

emulation of photometric galaxy observations. CNN-based approaches can be used to emulate images, whether literal or representational; emulating cosmological dark matter N-body simulations with CNN is a particularly active field [96, 163, 214, 466]. The general topic of using machine learning to speed up simulations for inference is advancing rapidly, thanks in part to differentiable probabilistic programming tools (see e.g. Cranmer et al. 2020 [89] for a review).

CNNs are also widely used to classify transients. Naive subtraction of images taken at different times often results in artificial 'bogus' residuals from e.g. cosmic rays or satellites; CNNs are often employed to distinguish these from 'real' transients [110, 460]. Sequences of images - effectively the two-dimensional generalisation of lightcurves - may also be classified using CNN modified for sequence processing with e.g. recurrent units [71, 148]. Just as with estimating photometric redshift (above), the hope is that images of the candidate transients contain more information than the derived lightcurve. Transients need not be literal images. In the radio domain, spectrograms are often classified by CNN to identify pulsars, RFI, or, increasingly, fast radio bursts [226, 442, 447]. Identifying faint fast radio bursts is the focus of Chapter 5.

I have so far discussed CNN as if they were a standard fixed model design; in my original description, convolutional and pooling layers followed by fully-connected layers (Sec. 1.2). This traditional design, used by AlexNet, has since evolved and diversified substantially. The latest classification models (such as EfficientNet [414], which I use in Chapter 4) include new components like bottleneck layers [196] alongside convolutional and pooling layers. Beyond classification, other types of model include convolutional layers to achieve different goals.

One important type of model is variational auto-encoders (VAE) [231], which use convolutional layers to reduce (encode) input images into a lower-dimensional latent space before expanding them (decoding) back to their original size with upsampling. VAE are then trained to compress images into that latent space by requiring the recovered image to be similar to the original image. The latent space can then be used for similarity searches (e.g. to find galaxy mergers similar to other mergers, [65]) or, by decoding new points in that space, to create new images of e.g. galaxies [401]. The closely-related U-Net [361] model type also uses an encode/decode structure with convolutions and upsampling, and is often used for segmentation (assigning pixels to classes); Boucaud et al. 2020 [52], for example, uses a U-Net to deblend simulated galaxies and hence provide improved photometry measurements of the deblended sources.

Another noteworthy type of model involving convolutions is generative adversarial networks (GANs) [150], in which a generative model (such as the decoder half of a VAE) creates 'counterfeit' examples which a classifier (such as a CNN) must then distinguish from real data. This adversarial competition, if correctly balanced, challenges the generator to learn the real data distribution and create increasingly realistic examples. The trained generator can then be used to create new examples, potentially based on specific real examples. Schawinski et al. 2018 [374] and Buncher et al. 2020 [62] both use this to 'reconstruct' how a deeper survey might have observed galaxies imaged by a shallower survey, although the reliability of these reconstructions remains unclear. Reiman and Göhre 2019 [354] uses GANs to deblend galaxies, where the GAN helps 'fill in' missing pixels (unlike Boucaud et al. 2020 's previously-mentioned U-Net, which aimed only at segmentation) Perhaps more surprisingly, generative adversarial networks can also be useful anomaly detectors. While the generator does not typically have an explicit likelihood, one can still identify images (for example) that occupy unusual regions of the generator's internal latent space. Margalef-Bentabol et al. 2020 [287] and Storey-Fisher et al. 2020 [406] both use this approach to identify unusual galaxies.

The field of deep learning, and its applications in astronomy, continues to expand rapidly. Nonetheless, I hope to have covered the key methods and historical context for this thesis. In the following chapters, I introduce my own work using neural networks to address astrophysical questions.

Chapter 2

Identifying Low Surface Brightness Tidal Features

2.1 Introduction

How do galaxies grow? In hierarchical models of galaxy formation, dark matter halos grow gravitationally by merging with smaller halos, carrying baryons - stars and gas - with them [337, 455]. This suggests that present-day galaxies assembled their mass through two closely-related pathways. First, galaxies may acquire mass through mergers; the repeated aggregation of smaller systems. As smaller dark matter halos fall into more massive ones, their central galaxies may eventually merge with the dominant central galaxy [450, 454], sharing both stars and gas (in so-called 'wet' mergers, [259]). Second, galaxies may acquire mass through secular accretion; the smooth gravitational capture of gas not yet bound to a system. This gas then fuels *in situ* star formation (e.g. [1, 455]) provided it can either avoid shock heating [216] or cool sufficiently from the virial temperature [434]. The addition of stars via mergers has long been uncontroversial (e.g. [427, 474]), while smooth accretion also has strong observational evidence; for example, many massive star-forming galaxies at $z = 2 - 3$ have extended disks thought to be inconsistent with recent mergers, suggesting smooth accretion as the dominant growth mode [98]. However, the relative contribution of *in situ* star formation and directly accreted stellar mass remains an open question (e.g. [128, 254, 349, 360]).

The physics of dark matter gravitational clustering is relatively straightforward and hence is amenable to simulation; indeed, N-body simulations by Press and Schechter (1974) [348] (in which N=1000) provide some of the earliest evidence that, in their words, 'larger-mass objects form from the non-linear interaction of smaller

masses'. However, the resulting evolution of baryons is far more challenging to simulate, for familiar reasons; the underlying physics is uncertain and efficient numerical schemes require compromises. For example, in galaxy-scale zoom-in simulations of mergers, stellar feedback is 'the most important property determining the galaxy's formation history' (Hopkins et al. (2018) [181]) and the choice of sub-grid physics to model stellar feedback can cause order-of-magnitude changes in galaxy mass without careful treatment [180]. Qu et al. 2017 [349], introducing the EAGLE hydrodynamical simulation, writes that 'such models are inevitably approximate and uncertain' and 'must be calibrated by comparison to observational data'. Observational constraints on merger rates would therefore be highly useful not only in their own right (as fundamental measurements of galaxy formation) but also for improving subgrid physics prescriptions and hence the reliability of numerical cosmological simulations.

It is well established that galaxy mergers leave long-lasting observational signatures in the form of low surface brightness tidal streams, shells and perturbations (e.g. [86, 301, 350, 427]). In galaxy outskirts, where the dynamical timescales are several gigayears or longer, these features are predicted to be particularly apparent [87, 203]. Indeed, much stellar substructure of this nature has already been detected in the peripheral regions of the Milky Way [37], M31 [124], and other nearby galaxies [108, 290]). Low surface brightness tidal features are therefore a powerful means to identify systems which have undergone mergers, and hence to estimate the rates of such events. The morphology and properties of these features are also useful probes for the nature of the preceding merger. For example, tidal shells are thought to be created by satellites on near-radial orbits while tidal streams are created by satellites on less eccentric orbits, and so the relative frequency of such structures probes the orbital distribution of past satellites [167, 345].

One of the main obstacles in such studies is the difficulty in reliably identifying faint tidal features. Part of this problem stems from the fact that morphological merger signatures only persist for a finite duration after an interaction has taken place, with the exact timescale dependent on the details of the orbital interaction as well as the properties of the host galaxy [272]. In particular, minor mergers (which we define as mergers with a mass ratio $\leq 1:4$) are thought to be far more common than major mergers and have been argued to be more important for driving star formation [217, 218] but are also harder to investigate because they generate faint signatures which are typically detectable over much shorter timescales [272, 273, 320]. Indirect evidence for minor mergers from resulting morphological transformations (e.g. bulge growth) can provide sensitivity to events that have occurred over longer timescales

but it is often difficult to distinguish these transformations from secular processes (e.g. [84, 178, 235]). The difficulty in distinguishing whether a feature is merger-driven or secular may be somewhat alleviated by searching for the *absense* of a feature, which is then evidence that neither a merger nor secular process has taken place.

Simmons et al. 2013 [390] use such an approach to identify galaxies which are unlikely to have undergone recent mergers due to a lack of a bulge. This is one example of the use of morphology measurements as an indirect tracer of galactic history; later in this thesis (chapters 3 and 4), I will detail how we combine volunteer effort and deep learning to construct catalogues of such measurements.

Tidal features are generally faint and come in many classes [26], making them often hard to spot against background noise and hard to distinguish from general galaxy morphology. Identifying tidal features in deep galaxy images is therefore a crucial yet challenging visual classification problem. Most work to date has focused on visual inspection of individual galaxies or relatively small samples (e.g. [284, 290, 384]), often on images that have been specifically manipulated to enhance the appearance of low surface brightness features (e.g. [177, 207, 306, 308]). However, the role of interactions and mergers in driving galaxy evolution is likely to change with mass [257], morphology [228], environment [260], and redshift [259]; unpicking the effects of these variables will require large statistical samples (i.e. several thousand systems or more) for which expert human classification becomes impractical. Unfortunately, there has been relatively little effort to date in devising automatic methods to detect and characterise low surface brightness emission in galaxies and the methods invoked are not particularly well-suited to detecting the faint tidal features typical of minor mergers. Such automatic techniques may be broadly grouped into two categories – those which rely on model subtraction and those which appeal to non-parametric feature extraction.

Model subtraction methods work by fitting a parametric light profile to the galaxy. This parametric light profile is then subtracted from the galaxy flux to leave, ideally, residuals corresponding to tidal features. The total residual flux is then used as a proxy measure for how tidally disturbed a galaxy is [7, 412, 438]. This approach works best on galaxies with smooth radially-symmetric morphologies where the non-tidal morphology is easily described. However, for galaxies with all but the simplest morphologies, the non-tidal morphology is hard (if not impossible) to describe in parametric form. Encoding the light profile of even relatively common and 'simple' galaxies like grand design (two-armed) spirals is itself a major research topic (see e.g. discussion in [264]). Further, because the non-tidal morphology is typically much

brighter than the tidal features, any minor errors in encoding the non-tidal morphology will often have a much larger contribution to the total residual flux than actual tidal features would, leading to galaxies with more complex morphologies being falsely identified as having tidal features.

Non-parametric feature extraction methods measure one or several hand-crafted image parameters thought to correlate with post-merger disruption. Astronomers have long sought to design non-parametric statistics that describe the 'physical' properties of a galaxy (typically contrasted with 'descriptive' classification based on visual classification schemes such as the Hubble Sequence, a debate I return to in chapter 3). One particularly relevant feature is asymmetry. Kalnajs 1983 [209] introduced a 'photographic trick' of rotating a (film) negative 180 degrees halfway through exposure to create a symmetric image that highlights the 2-fold symmetry of spiral arms. This procedure was first performed on computer by Elmegreen et al. 1992 [115] at IBM Research, in an early parallel to the software collaborations with industry on which modern deep learning in astronomy relies. The total flux of the residual image was first used as a statistic by Schade et al. 1995 [372]. Conselice et al. 2000 [85] added modifications for finding an appropriate rotation centroid and accounting for background flux, and argued that asymmetry can be used to identify galaxies with interactions or mergers.

More recent work tends towards increasing complexity. Lotz et al. 2004 [270] introduced the Gini coefficient, a measure of distribution inequality borrowed from economics [147]), and M_{20}, defined as the second-order moment of the brightest 20% of a galaxy's flux, for identifying tidal features in combination with traditional statistics (asymmetry as well as concentration and clumpiness, see [84] for a full review). Freeman et al. 2013 [139] introduced three further statistics including Deviation, designed to complement asymmetry by identifying bright groups of pixels far from the galaxy centroid. Pawlik et al. 2016 [333] introduced the shape asymmetry parameter, calculated as with asymmetry but replacing flux with a binary mask selecting contiguously-connected pixels with flux at least a specified standard deviation above background (after smoothing with a 3x3 pixel kernel).

Having defined a set of non-parametric features, one can identify the most likely candidates using selection cuts. One can also create more complex decision boundaries using low-dimensional machine learning methods such as Random Forests [54], as in Freeman et al. 2013 [139]. Note that, with such an approach, the machine learning input features have been calculated following a sequence of human-specified steps from the original image, in contrast to the deep learning approaches described below.

Non-parametric methods allow for a broader range of morphologies to be classified as they do not require an explicit parametric light profile to be specified. However, they can be easily confused by complex asymmetric features such as spiral arms [211]. and are typically only sensitive to certain major merger stages [271, 273, 395]. Elmegreen et al. 1993 [116] captured the essential difficulty of non-parametric methods when, considering a method for identifying spiral arms, they wrote: 'we could not help but believe that our eyes were a better judge of structural peculiarities and regularities than a few highly reduced numbers'. It is hard to define non-parametric features because identifying tidal features in an image is a perceptual task not immediately amenable to mathematics. Ideally, then, we would like to directly automate our visual perception without first defining reduced statistics. This is a key practical distinction between image classification with low-dimensional machine learning and with deep learning, which led me to consider a deep learning approach for this problem.

2.2 Application of Convolutional Neural Network

Both model subtraction and non-parametric feature extraction methods struggle to reliably detect faint tidal features in galaxies with complex morphology. My collaborators recognised the need to develop an effective method for tidal debris detection and classification which can be applied to detect faint tidal features in large statistical samples. Motivated by the need to learn to distinguish subtle classes without explicitly specifying image features to consider, and encouraged by the exceptional performance of deep learning models such as convolutional neural networks (CNNs) capable of learning hierarchical representations of terrestrial images (see Chapter 1), I decided to explore a classification approach based on CNNs.

In this section (Sec. 2.2), I introduce the core approach: I describe the sample of galaxies under study; the motivation for the architecture; and the training and performance of a single network. I then extend this core approach using an ensemble of several networks, significantly improving performance (Sec. 2.3). Finally, I compare the performance of both our single network and ensemble of networks with the current approaches of WND-CHARM [380] and shape asymmetry [333] (Sec. 2.4).

2.2.1 Data

This analysis is based on data products from the Wide component of the Canada-France-Hawaii Telescope Legacy Survey, hereafter CFHTLS-Wide [156]. This survey covers approximately 170 deg^2 of sky in four patches and uses filters u*, g', r', i' and

z' with an exposure time of approximately one hour per filter per field. Atkinson et al. 2013 [26] (hereafter A13) used visual classifications to study the incidence of faint tidal features in a sample of $\sim$ 1800 luminous galaxies drawn from this survey, making it an ideal sample against which to benchmark the performance of CNNs.

The A13 sample contains 1781 galaxies that were selected to lie within the redshift range $0.04 < z < 0.2$ and to have magnitude $15.5 < r' < 17$. These cuts were adopted to allow for comparison with previous work on tidal feature classification, to minimise contamination from stars misidentified as galaxies and to limit the sample size to a manageable number for visual inspection. As discussed in A13, this sample is heavily biased towards bright systems, with most galaxies lying in the range $-23 < M_{r'} < -20$ mag. The typical half-light radii of the galaxies is 2-6 arcsec.

The A13 study used thumbnails in the g', r' and i' bands as these were the highest signal-to-noise images. These thumbnails were stacked together to increase contrast. A13 estimate a limiting g-band surface brightness of ≈ 27.7 mag arcsec^{-2} over small scales. Each stacked image was visually inspected and placed into one of five categories depending on the confidence of the inspector that a tidal feature was present. These ranged from very high confidence of the presence of a feature (level four) to a feature with around 75% certainty (level three) and so on, until very high confidence was reached that no tidal features were present to the depth of the data (level zero). If tidal structure was deemed to be present then it was further classified into six non-exclusive tidal feature classes – shells, streams, miscellaneous diffuse structure, arms, linear features and broad fans. Each of these classes may trace the physical properties of the original galaxies, though definitive connections are elusive; tidal tails (arms, streams, linear features) and shells likely reflect different pre-merger orbits (Sec. 2.1), and fans may indicate dry mergers [438]. Identifying large samples of such classes with the methods presented here may ultimately help uncover these relationships.

Roughly 10% of the A13 sample was classified independently by three experts to ensure that the visual classification scheme leads to consistent answers by multiple experts and is therefore reproducible. Following this, the entire sample was classified by a single inspector (Atkinson) to maximise consistency. In this work, these single expert labels are used as a ground truth against which to measure automated methods; I address the implications for reproducibility in Section 2.5. The archetypal examples provided by A13 of these feature classes are reproduced in Figure 2.1.

As the thumbnails utilised in the A13 study were not available, I had to recreate these from scratch, in an identical manner, so as to guarantee that the automated

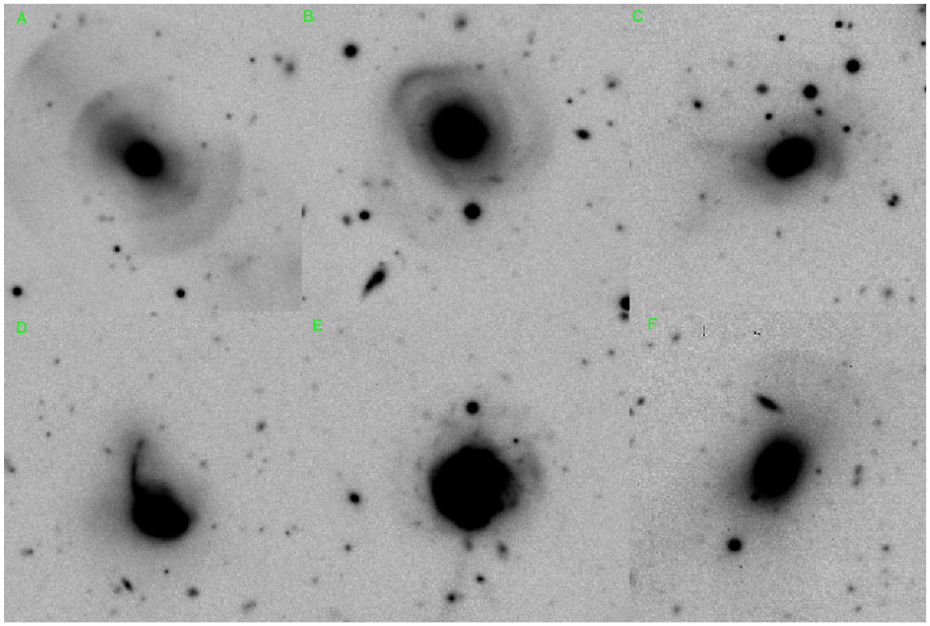


Figure 2.1: Tidal feature classes defined by Atkinson et al (2013). Clockwise from top left: shell (A), stream (B), misc. diffuse (C), arm (D), linear (E), fan (F). Reproduced by permission of the original authors and the AAS.

classifier had access to the same information as the human experts. To this end, I extracted 256×256 pixel regions in the *g'*, *r'* and *i'* bands around the galaxy centroid coordinates provided in the A13 catalogue using the CFHTLS cut-out service [118]. These images were subsequently manipulated in a variety of ways, as will be described in Section 2.2.2.

To reduce the complexity of the classification problem, tidal confidence labels were binned into binary classes. The choice to restrict the problem to a binary classification was motivated by the limited training data available (see Section 2.2.3) rather than any fundamental constraint. Non-tidal (0) was matched to confidence $\leq 25\%$ (levels zero and one in the A13 scheme) whereas tidal (1) was matched to confidence $\geq 75\%$ (levels three and four in the A13 scheme). Galaxies with a tidal confidence of 50% were deemed to provide no useful information for our purpose and were cut from the sample. Of the 1781 galaxies in the original A13 sample, 24 could not be downloaded in all three bands from the CFHTLS cut-out service, giving an initial data sample of 1757 imaged galaxies. Of those, 1316 galaxies are re-labelled False (non-tidal) and 305 are re-labelled True (tidal). 136 have a confidence of 50% and are therefore removed, leaving a final sample of 1621 galaxies with binary labels. I discuss the ability of the method to adapt to more subtle classes given sufficient training data in Section 2.5.

2.2.2 Preprocessing

Preprocessing input images can substantially improve neural network performance - in my view, typically more so than the design of the network itself (see e.g. [366]). I believed preprocessing might be particularly important for this problem after observing that when my co-authors and I attempted to visually identify LSB tidal features using SAOImage DS9 [206], we often made changes to the scaling function and scaling limits of each FITS image. These scaling changes made low surface brightness features dramatically more visible by compensating for the extreme dynamic range in the raw images (from bright galaxy core to background). Inspired by this domain expert behaviour, I implemented and experimented with various preprocessing options, listed below in order of operation.

1. **Aggregation** The *g'*, *r'* and *i'* band images provide three tensors of pixel flux values, each of shape (height, width). These are combined to create a single tensor, which includes all pixel information on each galaxy, to be used as input to the network. The bands can be pixel-averaged to create a tensor of shape (height, width). Alternatively, the bands can be concatenated (i.e. placed next to one another) along a third colour dimension to create a tensor of shape (height, width, 3) in analogy with RGB images.

2. **Background estimation** This estimate is required for the pixel intensity clipping and masking procedures described below. To estimate the sky background, I used the functions `sigma_clipped_stats` and `make_source_mask` from the Python package Photutils ([53]). `sigma_clipped_stats` estimates background from the statistics of all unmasked pixels within a given σ of the median unmasked pixel value. `sigma_clipped_stats` is called by `make_source_mask` to make an initial background estimate. `make_source_mask` then uses this estimate to detect and mask sources. The masked image is passed back to `sigma_clipped_stats` for an updated background estimate. This procedure iterates five times, giving a final background estimate.

3. **Pixel intensity clipping** The extreme intensity variation (dynamic range) between the inner galaxy core and the tidal features can interfere with rescaling algorithms (see below). Retaining only pixels with intensities lower than 6σ above the background avoids this issue.

4. **Pixel intensity rescaling** Rescaling the pixels to reduce the dynamic range of the image ensures that the tidal features contribute to the first layer values. We apply to each tensor x a rescaling mapping, for example $\sinh(x)$, x^a, or $\ln(x)$. Since the values of the first network layer are proportional to the input image pixel values, this avoids the untrained network initially seeing only the bright galaxy cores.

5. **Masking** The thumbnails have foreground and background objects, as well as occasional image artefacts, within the field-of-view. This introduces additional noise that could be mistaken for tidal features by the classifier. To mitigate this, pixels outside the contiguous galaxy light distribution can be masked. To identify which pixels to mask, I used a combination of background estimation and mean convolutions to estimate which pixels are plausibly part of the galaxy. This process is described in detail in Section 2.4.1.

6. **Local smoothing** This can enhance the appearance of faint tidal features near the signal-to-noise limit, albeit at the cost of a reduction in spatial resolution. The kernel size used controls the degree of smoothing; a 3x3 average kernel was found to give good visual results.

As the optimal combination of these various preprocessing options for tidal feature detection was not initially obvious, I approached this problem empirically by using a grid search (see Section 2.2.5).

2.2.3 Network Architecture

A significant challenge with our CNN approach to tidal feature identification is our exceptionally small training sample. Because every neuron connection is assigned a weight, CNNs typically have $> 10^5$ free parameters to learn (i.e. to fit to the data). Having many free parameters allows the learning of more complex features, but increases the entropic capacity of the classifier. Without a correspondingly large training sample to provide constraints, overfitting degrades performance. CNNs are typically applied to samples of 10^4 to 10^{10} images - for astrophysical examples, see [100, 191, 229, 341, 391] - while our CFHTLS-Wide sample contains only 1621 galaxies, of which a mere 305 have tidal features. The classifier therefore needs to operate approximately two orders-of-magnitude below the minimum sample sizes normally used by CNNs.

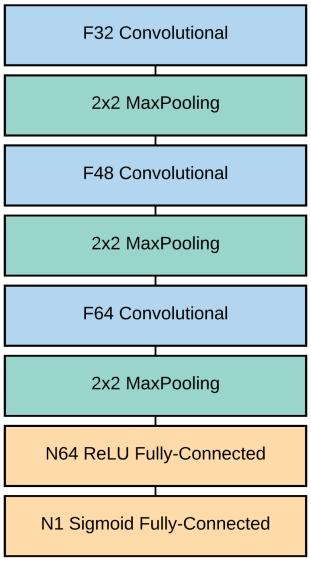


Figure 2.2: Network architecture for single CNN. Input image (tensor) at top. Output prediction at bottom. Convolutional layers have Fn (e.g. F32) 3x3 convolutional matrices (i.e. filters), each with a convolution step size (i.e. stride) of 1x1. Fully-connected layers have Nn (e.g. N64) neurons. The final layer is a single neuron whose output represents the continuous-valued class prediction.

To reduce the model's capacity for overfitting, I initially chose the architecture of Chollet 2016 [77], a relatively shallow design with (only) 3,714,593 free parameters. The architecture is presented in Figure 2.2. I then verified with a grid search that this architecture outperforms three convolutional layer networks with significantly higher or lower numbers of convolutional filters or layers. Three convolutional layers provide enough depth for high performance without becoming computationally intractable, while the relatively low number of filters helps prevent overfitting.

To further minimise overfitting, I chose to apply dropout [404], a regularisation method, to this layer. Dropout temporarily removes random selections of neurons. This encourages neurons to learn parameters that remain discriminative for many different combinations of other neurons in the network. Alternatively, dropout may be interpreted as taking the trained model and permuting it into a different one [404]. Dropout therefore approximates (with an unbounded error) training many unique networks [141]. This behaviour is crucial to the Bayesian convolutional neural networks I use in Chapters 3 and 4.

A neuron and all associated connections (weights) are referred to as a unit. For each training epoch, each unit in the fully-connected network layer has probability p to be removed for that epoch. The operation of a fully-connected layer with dropout

is

$$x_i^{(l)} = f(\mathbf{w}_i^{(l)}\mathbf{x}'^{(l-1)} + \mathbf{b}_i^{(l)}) \tag{2.1}$$

where $\mathbf{w}_i^l$ denotes unit i in layer l, $x_i^{(l)}$ is the (scalar) output of unit i in layer l and $\mathbf{x}'$ is the elementwise product $\mathbf{x}' = \mathbf{x}\ (*)\ \mathbf{B}(p)$ with $\mathbf{B}(p)$ being an $\mathbf{x}$-shaped matrix with binary elements according to the Bernoulli distribution (i.e. 1 with probability p, 0 with probability $1 - p$).

The thinned network (following dropout) is trained for a single epoch before $\mathbf{B}(p)$ is re-evaluated, causing different units to be active and a new thinned network to be created (sharing weights with the predecessor). I selected the hyperparameter p to be 0.5, based on the results of the grid search described below.

2.2.4 Augmentation

Galaxy morphological classifications should be invariant under certain transformations, such as flips, rotations, minor zooms, and minor translations. CNNs lack our intuitive understanding of transform invariance and require sufficiently diverse examples to infer which transforms are not discriminative. I therefore chose to artificially expand our training set by including many variants of the original input images [1]. By inputting many randomly-transformed images with unchanged labels, I teach the network to be insensitive to those transforms. By applying these transforms dynamically when each input image is read by the network, the effective training set becomes arbitrarily large and the network always sees a unique image. Note that augmented images are less informative than truly new images; once the network has learned the invariance, further augmented images do not improve performance.

I randomly apply all of the following transforms every time an image is loaded:

1. Horizontal flip
2. Vertical flip
3. Random resampled rotation uniformly chosen from the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$
4. Random resampled zoom uniformly chosen from the interval (90%, 110%)
5. Horizontal translation uniformly chosen from the interval (-5%, 5%)

[1] It is possible to construct CNNs with specific invariances, either by modifying the input (e.g. [100]) or the architecture [119]. However, achieving invariance through augmentations was a simple and effective alternative to these more complex approaches.

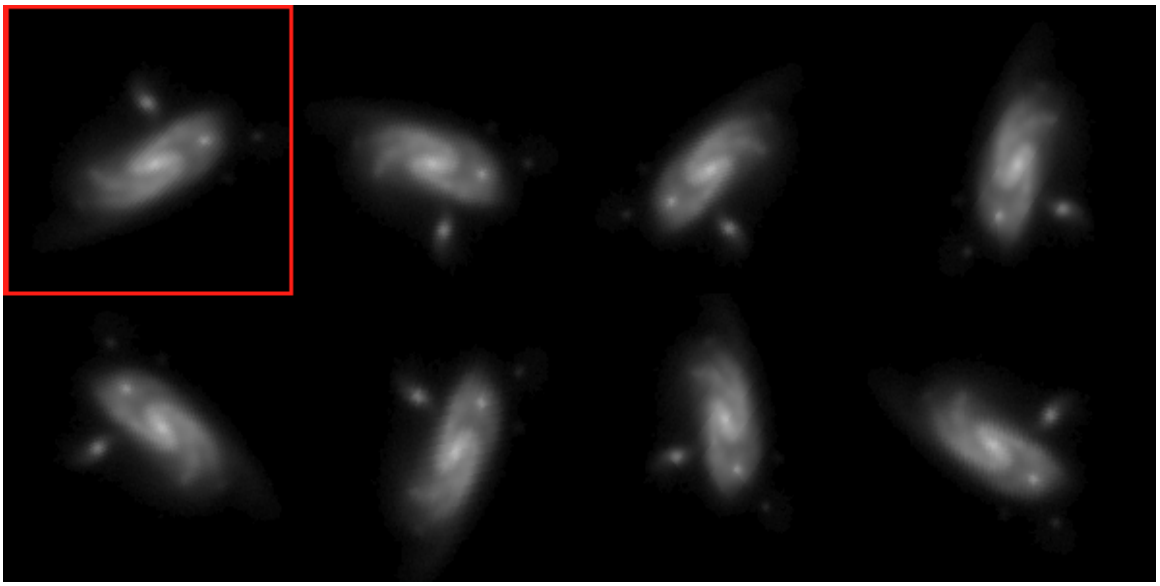


Figure 2.3: Mosaic of illustrative augmented images of a single non-tidal galaxy. Images are mean-averaged, masked (3σ) and shown in log scale. The images are cropped from 256 to 150 pixels for illustration only. The original image is shown in the top left (red highlight).

6. Vertical translation uniformly chosen from the interval (-5%, 5%)

To avoid unnecessary information loss from pixel resampling after each step, the transforms are applied through a single net transformation. I verified with a grid search (Section 2.2.5) that the augmentations improve performance, implying that resolution degradation from the net transformation has a less significant impact on prediction quality than the corresponding learned invariance.

Figure 2.3 shows a single galaxy with seven different augmentations applied. The same random augmentation process creates a unique image each time. The network is trained on approximately 84,000 uniquely-augmented images before convergence.

2.2.5 Grid Search

CNNs have tuneable design values (e.g layer count = 4, first layer width = 256) called hyperparameters [276]. The choice of hyperparameters may have a significant impact on classifier performance, but the optimal choice is not known *a priori*. Estimates can be made with heuristics (rule-of-thumb guesses) based on previous generic image classification work. However, images of galaxies with faint tidal features are unlike 'terrestrial' pictures in that they have extreme contrast, high noise levels and indistinct subject shapes, and so borrowing from such work is unlikely to be optimal.

I improved the heuristic hyperparameter estimates using grid searches. Through this procedure, many possible network configurations are trained and the performance of each is measured. I chose to separate hyperparameters into related groups and then identify the optimal choice within that group through an exhaustive grid search. For example, we assume the optimal number of layers is independent of pixel rescaling and proceed to test many possible numbers of layers with a single rescaling. This approach makes the grid search computationally feasible without needing to specify any hyperparameters with heuristics.

I used three groups of hyperparameters: preprocessing (see Section 2.2.2), architecture (see Section 2.2.3), and augmentation (see Section 2.2.4). The best performing preprocessing configuration was found to be band-stacked images with 3σ masking, logarithmic pixel intensity scaling and no mean convolutions. Performance is invariant under physically reasonable choices of pixel clipping values and, provided that mean convolutions are not used, also invariant under pixel intensity rescaling. This latter result is intuitively surprising given the impact that rescaling has on human perception. The best performing architecture and augmentation configurations have already been discussed in Sections 2.2.3 and 2.2.4, respectively.

2.2.6 Training and Evaluation

I implemented our network using the deep learning library Keras [78], with TensorFlow [2] as a backend[2].

I used a batch size of 75 images, identified as the optimal number by the grid search (see Section 2.2.5). One epoch was arbitrarily set as 14 batches or 1050 training images, roughly corresponding to the total number of labelled galaxies. Batch images were randomly selected without replacement (i.e. selected only once) in equal proportion from the tidal and non-tidal galaxy training subsets. Once all images from a subset had been selected once and removed, the subset was refilled. I chose to select the images in this manner for two reasons. Firstly, this approach provides the network with sufficient tidal examples to learn from. Secondly, it allows the network prediction to be interpreted as the probability that a given image is tidal and not merely a reflection of the base rate (i.e. the relative number of tidal versus non-tidal galaxy training examples seen by the network). Excluding the base rate during training ensures that predictions on a new sample will not be biased towards the training

[2]Keras was absorbed within the TensorFlow library subsequent to this work

base rate. Each selected image was randomly transformed to augment the dataset (see Section 2.2.4). I trained the network using a binary cross-entropy loss (Eqn. 1.1).

Any initial partitioning of data into training and validation images is arbitrary; one could have selected any set of images as validation images. We therefore needed to check if the network is fortuitously performing better on those validation images than it would on a large set of new data. Smaller datasets are particularly susceptible to such accidental overperformance as small number statistics make this scenario more likely. I used five-fold cross-validation to ensure our prediction quality metrics do not depend on an arbitrary division of data into training and validation subsets. In n-fold cross-validation, the complete data sample is split into n random subsets. $n-1$ are used to train the classifier from scratch, and the remaining subset is used as validation data. This is repeated for all n permutations. All prediction quality metrics in this chapter were averaged from each of the five-fold cross-validation runs.

2.2.7 Results

I selected completeness and contamination as metrics to evaluate the performance of our network. Conceptually, completeness is the probability for a visually-classified tidal galaxy to be correctly identified by the CNN as tidal, and contamination is the probability that a visually-classified non-tidal galaxy is incorrectly identified by the CNN as tidal. Mathematically, these are the true positive rate (TPR) and false positive rate (FPR), respectively:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}. \tag{2.2}$$

The Receiver Operating Characteristic (ROC) curve illustrates the completeness and contamination of the classifications. The ROC curve of our best-performing single CNN classifier is plotted in Figure 2.4. The completeness and contamination may be selected along any point on the curve, corresponding to varying the confidence threshold used to classify images as tidal. For example, one might choose a completeness of 70% and therefore a contamination of 22%. Random guessing would provide equal completeness and contamination.

Figure 2.5 shows the accuracy of a single classifier with and without dropout and augmentations, averaged over five runs. Shaded regions denote the 90% Bayesian credible interval [324]. Without dropout and augmentations, the training accuracy increases with the number of galaxies that the network sees while the validation accuracy remains low. This is because the network is overfitting to random patterns

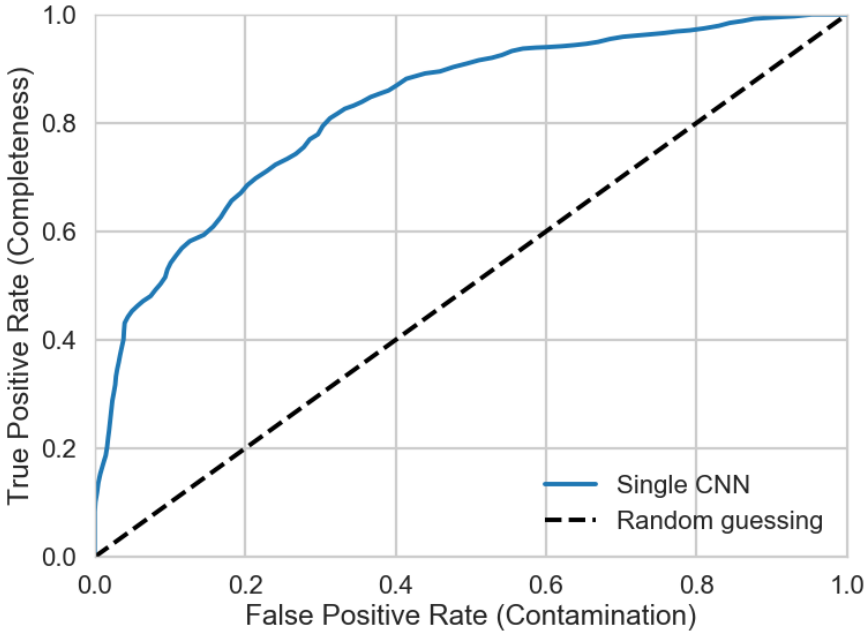


Figure 2.4: The ROC curve for a single CNN classifier on CFHTLS-Wide images. The dashed line indicates the expectation for random guesses.

in the training data. These patterns do not generalise to new data so the validation accuracy remains low. In contrast, with dropout and augmentations, the network is learning patterns present in both the training and validation data, causing both accuracies to rise.

Figure 2.6 shows performance broken down by class of tidal feature, following the schema introduced by A13. Every prediction is made by a network that has not been trained on that galaxy, following the cross-validation strategy described in Section 2.2.6. Networks perform best (i.e. have the lowest mean absolute error) on fan features, a surprising result given the relative rarity of such features. In general, performance is higher for dispersed features (fan, diffuse, shell) than small-scale structural features (arm, stream, linear). We speculate that this may be because such features are unlikely to be mimicked by contaminant objects in the field-of-view, and therefore easier to learn from our relatively small dataset.

All classes except fan (which is both rare and has a low mean error) have at least one prediction with an error close to one. This reflects the probabilistic nature of the method; our probabilistic metrics of success do not imply that every prediction is approximately correct. Figure 2.7 shows the galaxies with the highest and lowest absolute error (matching the highest and lowest horizontal bars across all columns

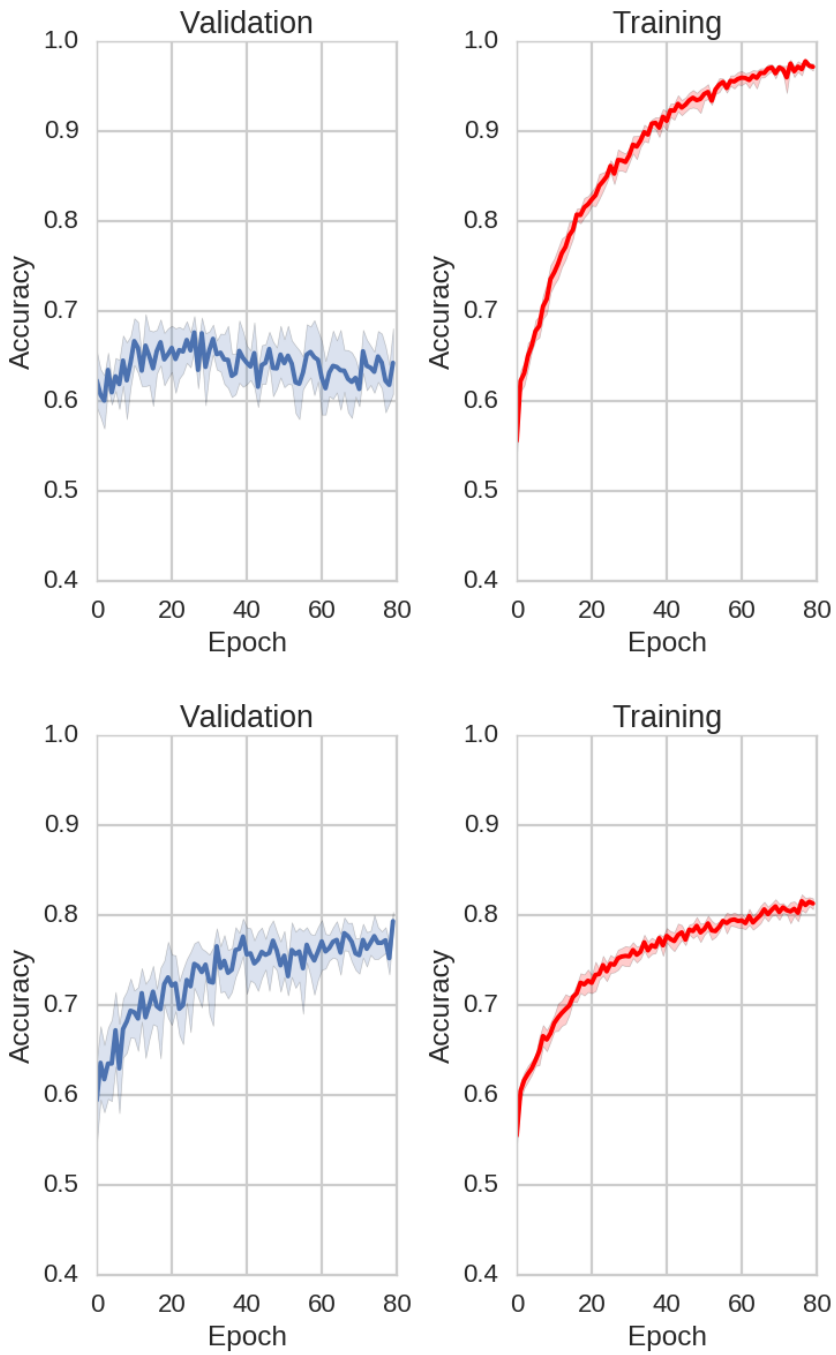


Figure 2.5: Mean training metrics for the same network architecture with dropout and augmentations off (top) vs. on (bottom), over five runs (trained and validated on each five-fold cross-validation permutation). Shaded regions denote the 90% Bayesian credible interval. 80 epochs for convergence (defined as a non-decreasing validation loss by eye) correspond to interactions with 84,000 uniquely-augmented training images.

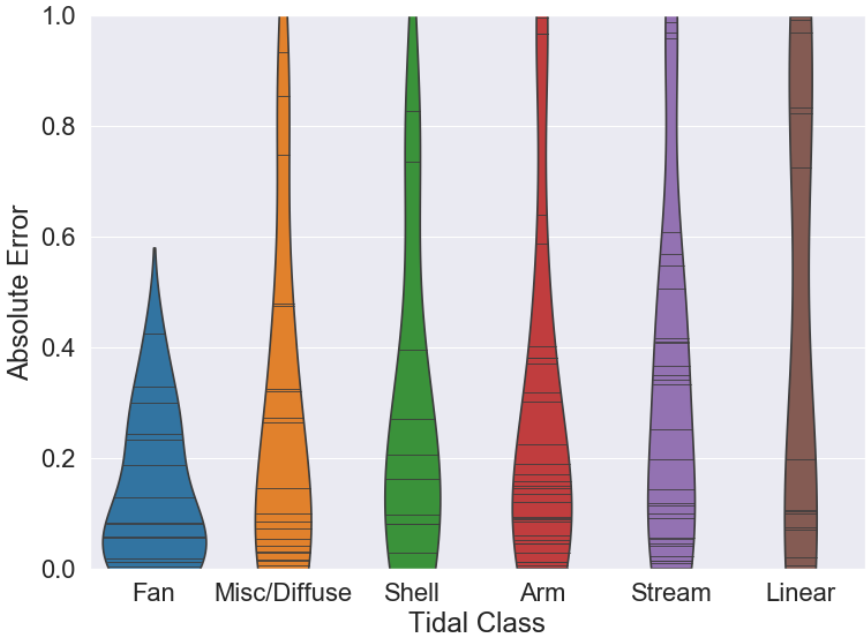


Figure 2.6: Single network validation performance by class of tidal feature. Each column is a class of tidal feature, ordered left to right by increasing mean absolute error on galaxies with that feature. Horizontal black bars denote individual galaxies: for example, a galaxy with a fan feature on which the network prediction had an absolute error of 0.2. More galaxies with lower absolute error indicates better performance. The area of each column illustrates the probability density, inferred (by kernel density estimate) from all galaxies of that class. Feature classes follow the schema introduced by A13.

in Figure 2.6). Failures show no obvious pattern, underscoring how the operation of convolutional neural networks is not always immediately interpretable by humans. I investigate the behaviour of the network in Section 2.5.1.

2.3 Ensemble of Convolutional Neural Networks

2.3.1 Configurations

The predictions of an ensemble of independent classifiers are well-known to typically outperform those of a single classifier, assuming all classifiers have similar individual performance. Framed as statistics, taking an average will partially cancel the random errors in each prediction. Framed as information theory, each independent prediction provides new information on the input image. Indeed, ensemble methods are routinely used to improve image classification performance e.g. [100].

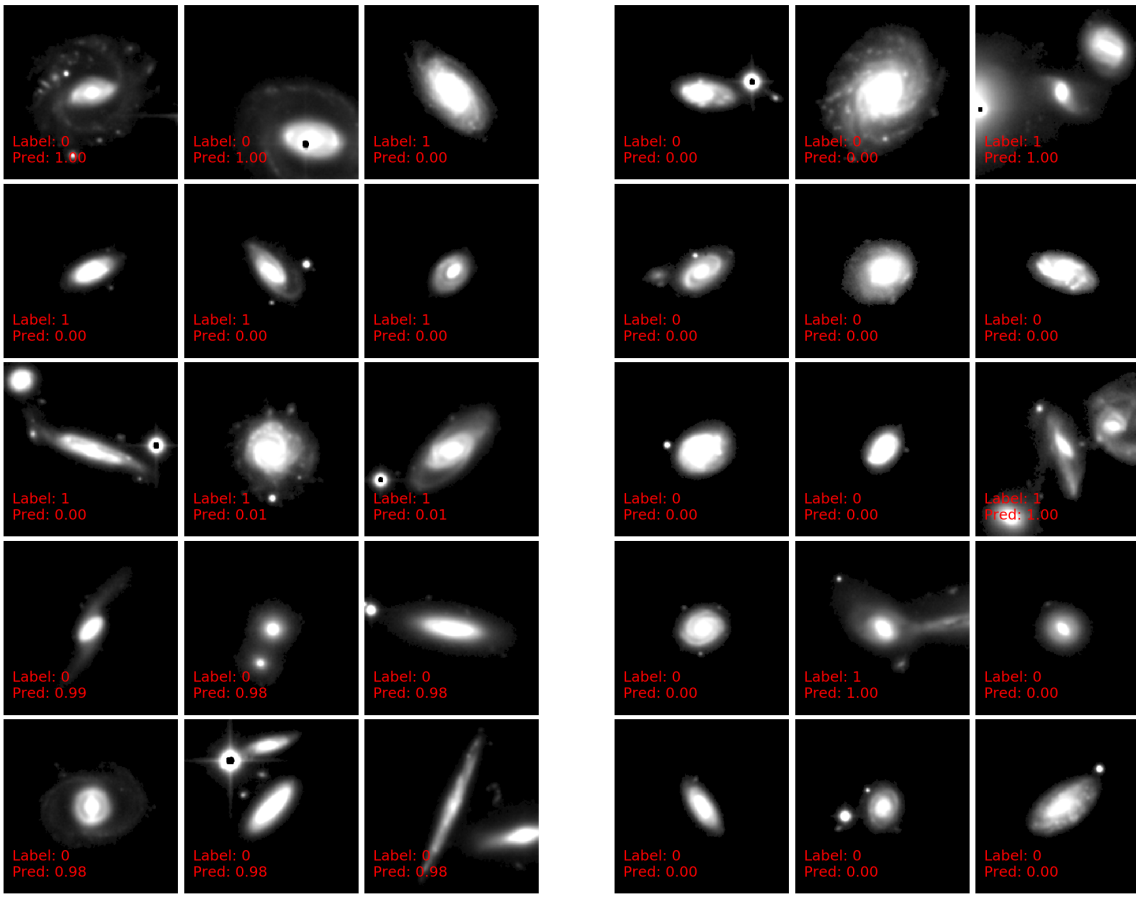


Figure 2.7: Galaxies with the highest (left) and lowest (right) absolute error in validation predictions, as presented to the network following the preprocessing strategy identified as optimal (including pixel rescaling and background masking, see Section 2.2.2). Brightness and contrast have been further adjusted for human viewing of tidal features.

I investigated two different ensemble configurations – CNN using optimal preprocessing (configuration A), and CNN using varied preprocessing (configuration B) – as a means to generate more accurate faint tidal feature classifications for our sample.

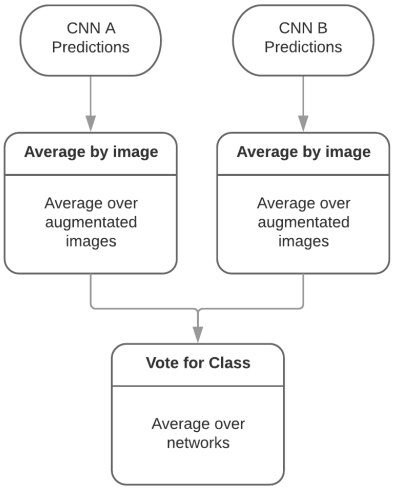
In configuration A, each CNN is in the optimal hyperparameter configuration identified in Section 2.2.5. The random order of input training images and the random initialisation of weights and bias prior to training may cause the CNN to converge to different local minima during training, particularly when applied to smaller training sets [253]. This leads to identically-configured CNNs making slightly different predictions, which is described as stochastic independence.

In configuration B, each CNN uses varied preprocessing hyperparameters, as detailed below. This introduces further independence between CNNs. Different preprocessing hyperparameters might lead a CNN to advantageously detect different tidal features. For example, more restrictive masking thresholds will reduce the number of contaminant objects in the field-of-view but may also reduce the spatial extent of particularly faint tidal features. However, hyperparameters that are different to the optimal hyperparameters will degrade the performance of a single model. By comparing each ensemble configuration, we test if (for our problem) it is more effective to ensemble individually stronger classifiers with lower independence (configuration A) or individually weaker classifiers with higher independence (configuration B).

I selected the following set of preprocessing hyperparameters for the five CNNs comprising the configuration B:

1. Logarithmic rescaling, 3σ mask threshold (i.e. optimal)
2. Logarithmic rescaling, 5σ mask threshold
3. No rescaling, 3σ mask threshold
4. No rescaling, 5σ mask threshold
5. No rescaling, band-stacked (un-masked) image

These were chosen for being high-performing combinations identified with the grid search described in Section 2.2.5, and for spanning visually distinct preprocessing steps.



Galaxy	CNN A Prediction	CNN B Prediction	True Label
a	0.86	0.71	1
a	0.83	0.82	1
a	0.92	0.81	1
b	0.43	0.35	0
b	0.54	0.39	0
b	0.32	0.42	0

Galaxy	CNN A Prediction	CNN B Prediction	True Label
a	0.87	0.77	1
b	0.46	0.39	0

Galaxy	Prediction	True Label
a	0.86	1
b	0.43	0

Figure 2.8: Flow chart of each stage for ensemble classifier. Red text illustrates the values being combined at each stage.

2.3.2 Training and Evaluation

To decide which configuration has the best performance, I trained and evaluated all five CNN comprising each configuration. Specifically, I trained each CNN on images that are randomly drawn in equal measure from 80% of the tidal and non-tidal classes, as described in Section 2.2.6. I then made predictions with each CNN on the remaining 'unseen' 20% of galaxies. Finally, I calculated an overall prediction for the configuration by combining the predictions of each CNN. Figure 2.8 illustrates how the predictions of each CNN are combined.

First, for each CNN, I averaged over all predictions made by that CNN on augmented images of the same galaxy. The true label is invariant under our augmentations but the CNN may not have completely learned to ignore them. Averaging over predictions of the same galaxy ensures that the final configuration prediction will not depend on any particular augmentation.

After recording the augmentation-averaged prediction on each galaxy by all five independently-trained CNN, I then averaged those single-CNN predictions to exploit any independence in those predictions to improve performance, as explained in Section 2.3.1.

2.3.3 Results

Figure 2.9 shows the average ROC for each ensemble configuration, and overplots the ROC of the individual optimal CNN shown in Figure 2.4. The completeness and contamination of the two CNN ensemble configurations are notably improved over the single CNN for galaxies with more ambigious scores, leveraging residual independence between classifiers to increase performance. For example, a single CNN achieves a completeness of 70% with a contamination of 22%. With ensembling, this improves to a completeness of $76\% \pm 2\%$ with the same level of contamination. For galaxies with more extreme (confident) scores, the network ensembles show relatively little improvement. This may be a consequence of there being relatively little disagreement between ensemble classifiers for the most obvious examples. The prediction quality of the two ensemble configurations is approximately equal within the expected statistical variation. That is, for our problem, both configurations are equally effective.

The ROC curve measures performance when classifying all galaxies, which is appropriate for understanding the overall performance of a method. In practice, one might instead choose to classify only a subset of galaxies for which the model is reasonably confident, and refer the remainder to experts or citizen scientists. We can measure model confidence using the continuous prediction score output by the model. By optimising our model using the binary cross-entropy loss (Equation 1.1), which heavily penalises mistaken scores near 0 or 1, we can interpret scores near 0 or 1 as confident predictions and scores close to 0.5 as uncertain predictions [418]. Therefore, we can select galaxies with confident predictions by requiring a score at least some minimum difference from 0.5.

However, because the model was trained on an equal number of tidal and non-tidal galaxies (Section 2.2.6), our scores on the full imbalanced sample are uncalibrated; the model does not know that non-tidal galaxies are common. To account for this, when calculating confidence, I calibrated our scores with Platt's Scaling [132]. Specifically, I used logistic regression to fit a correction to the fraction of true positives on 25% of galaxies, such that the scores match the empirical probability that a galaxy is tidal, and then applied that correction to the scores of the remaining galaxies.

Having calibrated our scores, it is now possible measure how performance varies on increasingly confident subsamples. The results show that performance can be dramatically improved, at the cost of leaving some galaxies unlabelled. Figure 2.10 shows how the accuracy increases as one considers only galaxies where the model is increasingly confident. On the full sample, the calibration causes the model to predict 'non-tidal' on three out of four galaxies, leading to an accuracy similar to a baseline

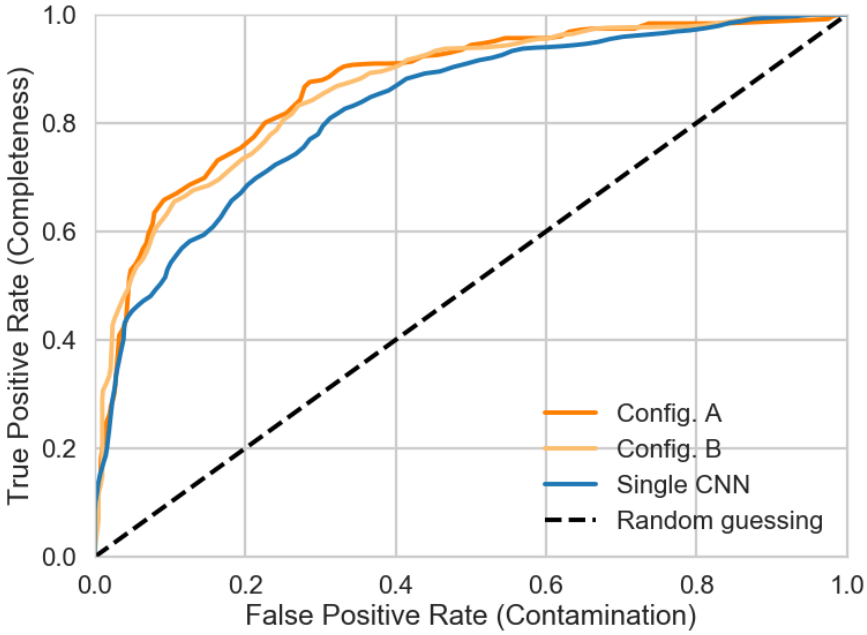


Figure 2.9: Comparison of the ROC curves of a single CNN, and our two ensemble CNNs. Our ensembles (each of five CNN) substantially outperform the single CNN.

classifier that always predicts non-tidal. However, by using the score to identify galaxies where the model is more confident, one can make useful predictions on the bulk of the sample. For example, the 72% of galaxies with a minimum score difference of at least ± 0.33 can be classified with 97% accuracy, compared to 84% accuracy on all galaxies. This suggests that our prototype model can be used to identify the bulk of a survey with near-perfect accuracy, reducing the human labelling effort required to create extensive science-ready catalogues of galaxies with or without tidal features.

I next investigated the independence of the single classifiers within each ensemble by measuring the correlation between each possible pair of classifiers. I calculated the Pearson r correlation coefficients between the continuous-valued predictions of each classifier. The resulting matrices are shown in Figure 2.11. The matrices are symmetric due to the symmetry of the correlation coefficien; unitary diagonal elements result from pairwise comparisons between a CNN and itself, and may therefore be neglected.

Recall that configuration A combines five classifiers all using the same optimal preprocessing configuration (logarithmic rescaling and a 3-sigma pixel mask, see Section 2.2.5), while configuration B combines five classifiers with varied preprocessing configurations. The average (non-unitary) correlation coefficient is lower for the ensemble

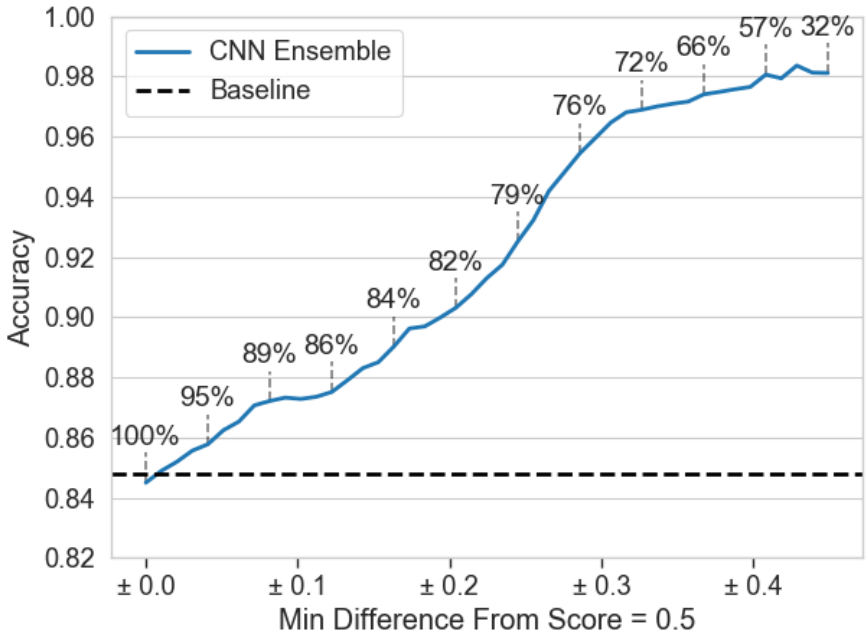


Figure 2.10: Accuracy of CNN ensemble (Config. A) on subsamples with increasingly confident predictions. Accuracy is measured for galaxies where the classifier score is a given minimum difference from 0.5. The greater the minimum difference, the more confident the classifier is. The percentage of galaxies with at least that confidence is annotated. Also shown is a baseline classifier that always predicts non-tidal (the majority class). We find that confident galaxies are more far likely to have correct classifications, For example, 72% of galaxies have a minimum score difference of at least ± 0.33 (i.e. a score above 0.83 or below 0.17) and can be classified with 97% accuracy. This suggests referring the least confident galaxies to experts or citizen scientists could be an effective strategy.

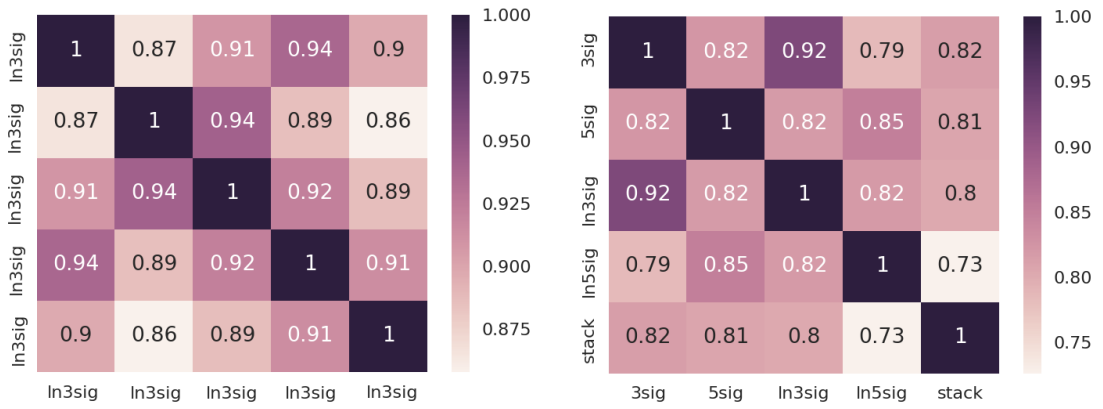


Figure 2.11: Pearson correlation coefficients between the predictions of single CNN (rows, columns) classifiers acting within ensembles using optimal preprocessing (A, left) and varied preprocessing (B, right). Labels denote the preprocessing used for that CNN, with 'ln' denoting logarithmic pixel rescaling and 'Nsig' denoting the masking threshold used. Configuration A combines five classifiers all using the same optimal preprocessing configuration (logarithmic rescaling and a 3-sigma pixel mask) while configuration B combines five classifiers with varied preprocessing configurations.

with varied preprocessing (B, $\bar{r}$=0.82) than with optimal preprocessing (A, $\bar{r}$=0.90), indicating that *additional independence can be introduced by altering the preprocessing process*. In particular, altering the masking threshold has a greater effect on classifier predictions than changing from logarithmic to linear rescaling. This is consistent with our earlier finding that prediction accuracy is invariant within statistical uncertainty under pixel rescaling.

2.4 Comparison with Current Methods

As discussed in Section 2.1, most current methods of automated feature detection are not well-suited to recovering the typical low surface brightness tidal features that arise from minor mergers and accretions. To accurately measure our performance compared to existing work, my co-authors and I selected two alternative methods from the recent literature and applied them to the A13 sample to benchmark their performance against that of the CNNs. These are:

1. Shape asymmetry [333], an example of a method based on non-parametric feature extraction;

2. WND-CHARM [378, 380], an alternative unsupervised machine learning approach previously shown to be successful in identifying peculiar and interacting galaxies.

Detecting tidal features by any method is dependent on:

1. The nature of the sample under study. The varying depths, bandpasses and spatial resolutions of different datasets can lead to incomparable detection rates;

2. The author's definition of what is tidal. The context of the paper often sets the definition for a tidal feature, and different authors may reasonably have different definitions.

For example, [56] and A13 both use data from the CFHTLS to identify tidal features through visual inspection. However, [56] uses data from the Deep component of the survey, which covers less sky area but is sensitive to more distant galaxies than the Wide component used by A13. Furthermore, they select different features to define which galaxies are tidal (tidal tails and bridges vs. the more subtle debris features outlined by A13). Directly comparing the detection rates (and underlying methodology) of these two papers is therefore not meaningful as they measure different things.

By applying all three methods to the same galaxy sample, with the same binary labels, we sidestep many of the complications that arise when comparing results that have appeared in the literature. This also ensures that the ability of each classifier to detect *the same* tidal features is tested fairly. Below, I describe each method and motivate why we have selected that particular method for comparison.

2.4.1 Application of the Shape Asymmetry method

Shape asymmetry was introduced by [333] as a method to automatically detect faint asymmetric tidal features in galaxies that experienced a recent merger. It is an appropriate choice for tidal feature detection in galaxies with complex morphologies since, unlike residual-based methods, it does not require a parametric fit of the underlying galaxy light profile. The measure is only sensitive to morphological asymmetry and does not contain information about the asymmetry of the light distribution. When applied to a sample of 70 starburst and post-starburst galaxies imaged by the Sloan Digital Sky Survey [3], [333] report an accuracy of 95% in detecting post-merger tidal features.

The method works as follows. First, following [85], the minimal asymmetry centroid is identified and asymmetry parameter A is recorded.

$$A = \frac{\sum | I_0 - I_{180} |}{2 \sum | I_{180} |} - A_{bgr} \tag{2.3}$$

where I_0 is the value of a galaxy pixel, I_{180} is the value of the pixel at the same position after the image is rotated 180 deg, A_{bgr} is the estimated contribution to asymmetry from background noise, and all sums act over all pixels. Note that low surface brightness pixels will have small I_0 and hence will make only minimal contributions to A. As a result, A is relatively insensitive to faint tidal features.

Next, a 3x3 mean convolution is applied to the galaxy image to enhance low surface brightness features. A binary mask is then created with values of 1 where the corresponding pixel count is both some chosen Nσ above the original measured sky background and contiguously eight-connected to the central pixel. The intuitive effect is to create a silhouette of the galaxy outline that includes faint structure - see Fig 2.12. For my re-implementation, background estimation is done with the procedure described in Section 2.2.2. I found a pixel masking threshold of $N = 3$ gives optimal results.

Finally, the shape asymmetry parameter A_s is calculated in analogy to A but with I_0 and I_{180} replaced by the pixel values of the binary mask, rather than the original galaxy image:

$$A_s = \frac{\sum | M_0 - M_{180} |}{2 \sum | M_{180} |} \tag{2.4}$$

where M (M_{180}) is the value of a mask pixel at some (rotated 180 deg) position on the binary mask.

To ensure tidal features at the image extremities are included, the selection radius used to calculate both A and A_s is taken as the minimum radius that encloses the full binary mask. By plotting A against A_s, an empirical selection cut can be made to identify galaxies with tidal features.

Figure 2.13 shows the resulting asymmetry space for our CFHTLS-Wide sample where 250 random examples are plotted per binary class. On the basis of visual inspection, [333] chose an empirical cut of $A_s > 0.2$ to select tidal galaxies. However, we wanted to measure how the shape asymmetry method balances completeness and contamination. To do this, I generated ROC curves using two methods (each generating a slightly different curve). In the first method, I generalized the $A_s > 0.2$ sample cut by making many sample cuts separated by δA_s. The ROC curve is then calculated in

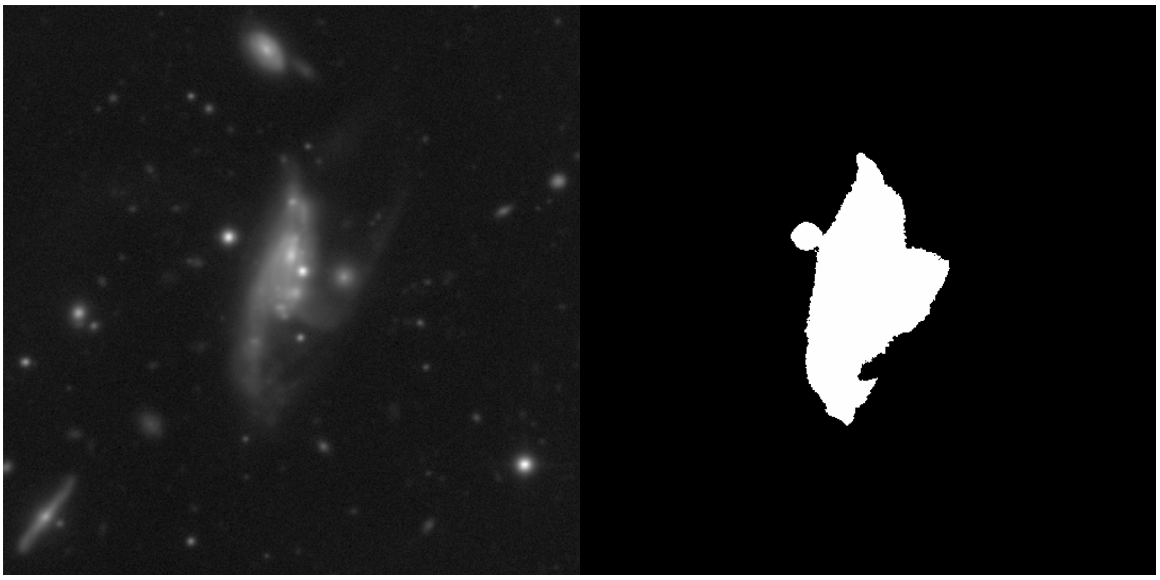


Figure 2.12: Illustration of the non-parametric shape asymmetry method of Pawlik et al 2016, applied to a galaxy in the CFHTLS-Wide sample. Left: stacked *gri* galaxy image (logarithmically rescaled for illustration only). Right: binary mask of pixels above 3σ used to calculate shape asymmetry A_s.

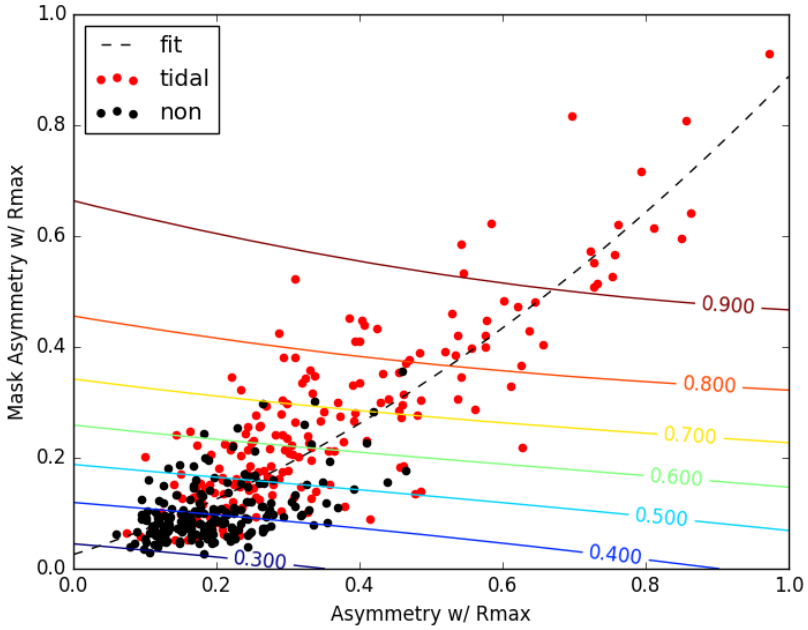


Figure 2.13: Probability space generated by Pawlik method on 500 CFHTLS-Wide galaxies from the A13 sample, illustrated by contours. Mask asymmetry is the shape asymmetry A_s. Galaxies are observed to follow a clear linear trend on the mask asymmetry/asymmetry space, which we fit for interest only.

the continuous limit $\delta A \to 0$. In the second method, I divided the galaxies into five subsets, train a logistic regression classifier [336] implemented in `scikit-learn` on four subsets, and make predictions on the remaining test partition. This is repeated for each combination of partitions (i.e. five-fold cross-validation). The ROC curve is calculated as the mean ROC curve over the test predictions for combination. To verify that the logistic regression classifiers are functioning correctly, Figure 2.13 shows a contour plot of the mean estimated tidal probabilities.

2.4.2 Application of the WND-CHARM algorithm

WND-CHARM [326] is a general-purpose image classification algorithm. Like CNNs, WND-CHARM was originally developed for other uses [326] and was only later applied to astronomy. It has been successfully used to classify peculiar galaxies [380] and general galaxy morphology [378] and so could be reasonably expected to perform well on the problem of faint tidal debris.

WND-CHARM [326] calculates a pre-specified feature vector (list) of 1025 image statistics, chosen to measure the contrast, texture and variation of the image. The statistics used range from the coefficients of Chebyshev [153] polynomials approximating the image to Tamura textures measuring contrast, coarseness, and directionality [413]. This may be seen as a generalisation of algorithms that use image features specified by domain experts (for example, random forests - see [127]). WND-CHARM then uses the labels to identify which features are most discriminative between classes in the training sample using the Fisher discriminant score [44], with the least discriminative (lowest Fisher score) features discarded. Those features are then used to classify test images using a novel variant of nearest neighbour classification [109] where distances along each dimension (feature) are weighted according to how discriminative each feature is for the training set. WND-CHARM is publicly available as both a command-line tool and Python API from `https://github.com/wnd-charm/wnd-charm` (for which we thank the authors).

The augmentation procedure I created for our convolutional network (see Section 2.2.4) is designed to improve classifier performance. To provide a fair comparison, I trained and tested WND-CHARM on two subsets of 25,000 images preprocessed and augmented through the same procedure. I used a train-test split of 80% and 20% respectively when selecting the original images used to generate these augmented subsets.

2.4.3 Overall Comparison

Figure 2.14 shows the completeness and contamination achieved by the three approaches (CNNs, shape asymmetry, and WND-CHARM) over many confidence thresholds (not shown). Figure 2.15 summarises overall performance with the area under each curve, known as the AUC score, for each method. The AUC score can be shown to equal the probability that a randomly-chosen positive example (tidal galaxy) is ranked higher than a randomly-chosen negative example (non-tidal galaxy), making it a useful and common scalar summary metric of classifier quality [188]. For our problem, the AUC score measures the probability that a random galaxy with tidal features will correctly be recognised as being more likely to have such features than another random galaxy without tidal features.

It is readily apparent that our CNNs have higher completeness and lower contamination than either of the alternative methods investigated in this paper. The ensemble configurations show the best overall performance, followed by the single-classifier configuration. Of the alternative methods, shape asymmetry outperforms WND-CHARM. All the methods tested definitively outperform random guessing.

The shape asymmetry method is found to be moderately effective in identifying galaxies with faint tidal structure. However, the shape asymmetry method performs less well for galaxies with minor A_s and A, causing the gradient to subsequently flatten as less confident predictions are included. Intuitively, this suggests that minor asymmetries are not an effective distinguishing feature between tidal and non-tidal galaxies. Extending shape asymmetry to use logistic regression rather than cuts provides a small improvement.

WND-CHARM is found to be the least effective method investigated for identifying galaxies with faint tidal features in CFHTLS-Wide sample. I speculate that its poor performance may be a consequence of the macroscopic feature extraction step employed in the algorithm. The ratio of image information content corresponding to faint tidal features may be sufficiently low that WND-CHARM struggles to identify a genuinely predictive feature set amongst the 'noise' of the general morphology. With an ability to investigate 1025 image features for correlations with labels, WND-CHARM could be overfitting to image features that do not relate to tidal features in the test data. However, WND-CHARM does show better performance than shape asymmetry when using generous score thresholds to select a highly complete sample; WND-CHARM achieves higher completeness at fixed contamination for contamination less than around 0.5, but still underperforms all the CNN approaches.

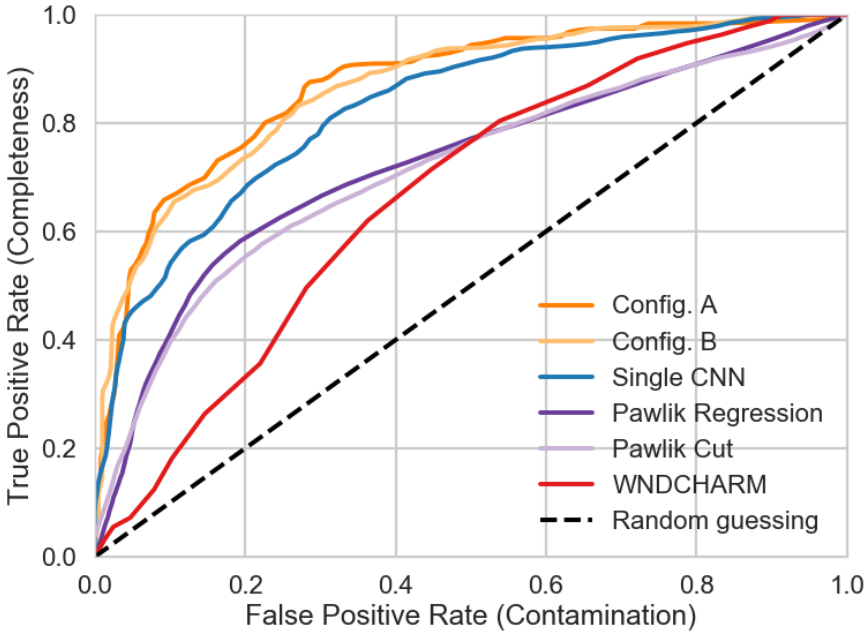


Figure 2.14: The ROC curves for all classifiers tested on the A13 sample. All CNN-based approaches substantially outperform other current methods. CNN ensembles (Config. A, Config. B) outperform a single CNN.

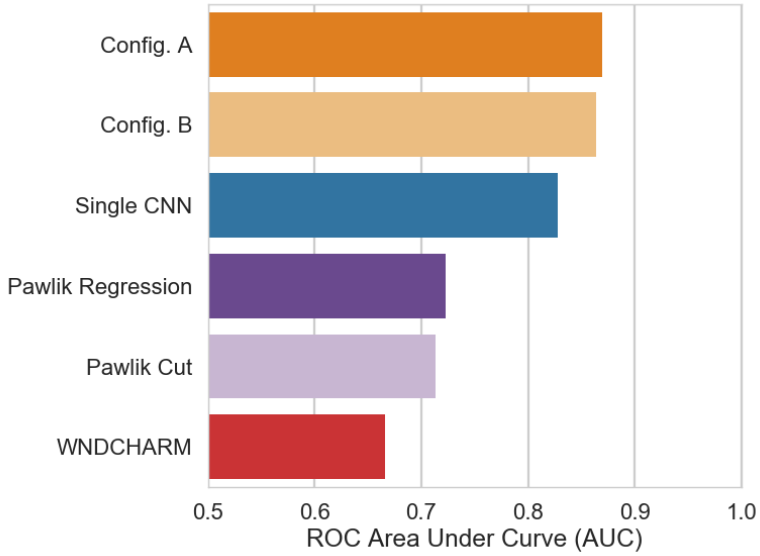


Figure 2.15: The ROC area-under-curve (AUC) values for all classifiers tested on the A13 sample, summarising classifier performance. All CNN-based approaches substantially outperform other current methods. CNN ensembles (Config. A, Config. B) outperform a single CNN.

2.5 Discussion

2.5.1 Heatmaps

A common criticism of CNNs, and deep learning in general, is that they are 'black box' algorithms which are difficult to interpret. While the resultant classification is readily apparent, how it was arrived at is usually less so. There is no clear link from the properties of the galaxy features to the prediction made.

In order to establish if our method is truly identifying faint tidal features in the way we intend, I decided to use prediction heatmaps [470]. Having established that each ensemble offers comparable performance, I arbitrarily investigate Configuration A (similar individual classifiers).

For a single image, a synthetic low surface brightness tidal structure is added into a small area. First, a 5x5 grid of pixel values is sampled from a Gaussian distribution with background variance and a mean 3σ above the background which represents the synthetic structure. Second, a random 5x5 pixel area in the original image is replaced with our new structure.

Each time the structure is added, the new image (original plus synthetic structure) is reclassified with an ensemble classifier and the change in tidal prediction from the original image is recorded. By plotting the tidal predictions as a heatmap where each pixel is the tidal prediction given a 5x5 synthetic structure at the location, one can identify in which image regions the ensemble sensitive to small changes. The basic assumption is that adding a tiny synthetic structure to a region that the network prediction is highly sensitive (one might say, 'suspects' as being tidal) causes a much greater increase in the tidal prediction for the whole image than adding such structure to an otherwise non-tidal region.

Figure 2.16 shows one example. The input image is shown at the top left. After a brief (5 epoch) training period, the heatmap is approximately a pixel-count-weighted distribution. After training is complete (epoch 125), the heatmap shows the network to have identified a linear feature at the bottom left corner of the image. Redisplaying the original image on a logarithmic scale, we verify that there is indeed a low surface brightness linear feature present at that location. This feature is detected and localised by the network despite being sufficiently faint to be invisible to the eye on the unscaled input image.

Our prediction heatmap demonstrates that the CNNs are identifying which image pixels are associated with low surface brightness tidal features. This shows that the CNNs are learning to perform the prediction task based on the image features

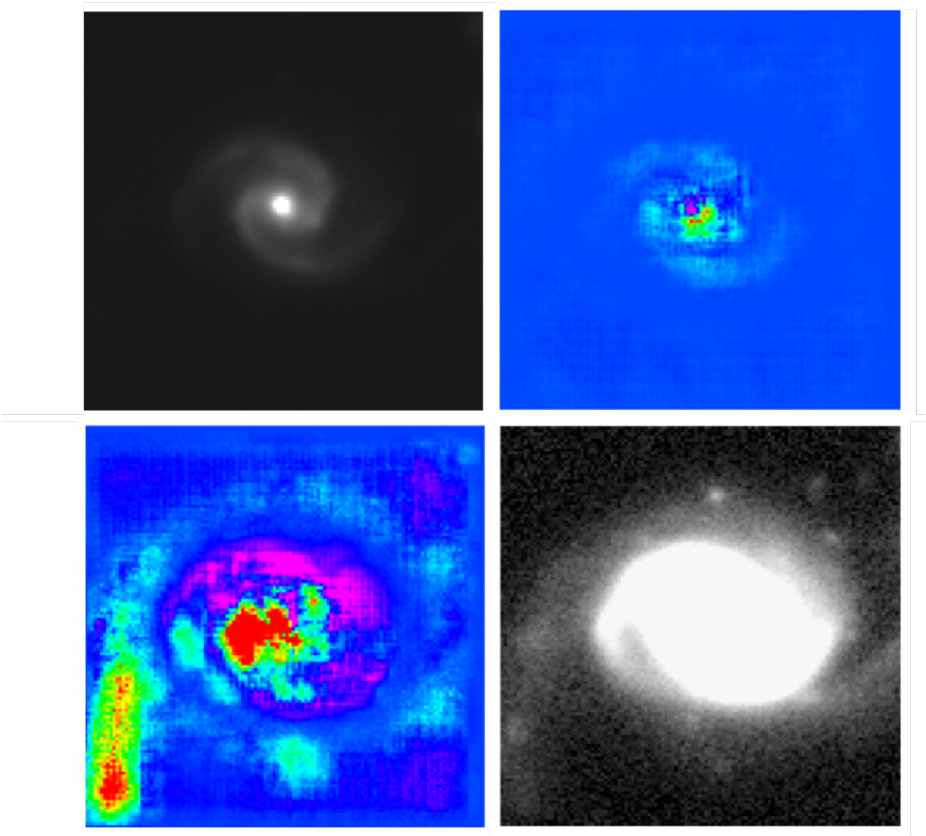


Figure 2.16: Top left: a cleaned galaxy image without rescaling. Top right: the heatmap from epoch 5. Bottom left: the heatmap from epoch 125. Bottom right: cleaned image with logarithmic rescaling. Magenta denotes regions the network considers non-tidal. Blue denotes neutral. Green through yellow through red denotes increasing tidal confidence. Note that the synthetic tidal structure is only added temporarily to alter the network predictions, and is not shown in any of the images above.

intended, rather than exploiting other features ('shortcuts', 1.1) which may cause biases or other undesired behaviour.

Further, if the pixel associations are sufficiently reliable, this offers the potential for automatic measurement of the sizes and shapes of tidal features. Once our CNNs identify which pixels belong to tidal features, it would be straightforward to write simple rule-based algorithms to calculate properties of interest to researchers. For example, the length of tidal tails probes the shape of dark matter halo potential [33, 303], and would be easily calculated given the pixels comprising those tails. One could verify the reliability of such measurements with citizen science projects following the style of e.g. Galaxy Zoo 3D (in prep - see Masters 2019 [291] for an outline). Hendel and Johnston 2015 , citing simulation work by [204], writes 'the population of streams with different extents and surface brightnesses could conceptually be used to ascertain the rates of minor mergers with different mass ratios'. Our CNN approach could ultimately provide the measurements needed to do this in practice.

2.5.2 Training Data

The sophistication of the CNNs used in this paper is limited by the size of the training data. The expert labels from A13 contain 305 tidal galaxies spread over six non-exclusive morphological classes of tidal feature. This scarcity of labelled examples places a fundamental limit on how much a convolutional network can generalise and learn to recognise such features. Pre-processing, shallow network design, augmentation and dropout are all necessary to achieve our classification performance.

Larger training sets would provide constraining information to support CNNs with more free parameters. This in turn would allow for more complex predictions about the input images. In principle, a CNN could directly localise tidal features with bounding boxes ([189]), provide predictions for many different classes of tidal features ([391]), and estimate tidal parameters like the length of a tidal tail ([428]). Our heatmap experiment (Section 2.5.1) provides compelling evidence for the plausibility of these applications if a sufficiently large training sample can be realised. I discuss three possibilities for expanding the size of the training set, and ultimately enabling these useful scientific applications, below.

Visually identifying large samples of galaxies with faint tidal structure is a daunting task given the relative rarity of such features at the typical surface brightness levels of current wide-field datasets. Most studies agree that to a surface brightness of $\mu \sim 26.5 - 28$ mag arcsec^{-2}, roughly 10-20% of galaxies show evidence for faint tidal features (e.g. A13, [177, 207, 308]. In order to create a training sample of even

$\sim 10,000$ tidal systems, more than 100,000 galaxies would need to be visually inspected. Crowd-sourcing efforts like Galaxy Zoo [265, 456] could be an effective way to accomplish this. I note that tidal features from minor mergers and accretions are often rather subtle in appearance and visual identifications typically require some degree of interactive manipulation of pixel scaling and contrast. The success of projects such as Planet Hunters TESS [112, 113] show that citizen scientists are more than able to effectively manipulate interactive elements to solve complex classification tasks; I think that future citizen science projects aiming to identify faint tidal features would benefit greatly from including these elements. The accuracy of resulting tidal labels would need to be carefully verified, perhaps by checking against a smaller expert catalogue. Citizen science would also help mitigate the risk of a single expert producing classifications that systematically deviate from other experts.

Alternatively, or in conjunction, one could use synthetic training data from simulations. Individual tidal features can be simulated in exquisite detail (e.g. [167, 204] and large-scale hydrodynamical simulations of galaxy formation now have the resolution to resolve these features in populations of several thousand galaxies (e.g. [345]). With simulated data, mock observations could be made at many viewing angles and surface brightness thresholds to provide an arbitrarily large training sample. However, while simulations provide perfect information on tidal labels, the simulated images may not be completely equivalent to real labelled images. The obvious challenge is fidelity; simulations are unlikely to fully capture the development and evolution of real tidal features, impairing the ability of the classifier to detect such features. More subtly, the process of creating the synthetic images from the simulation may also cause deviations between real and simulated galaxy images. Subsequent to the publication of this work, [51] experimented with training CNN classifiers to detect mergers using synthetic images from hydrodynamical simulations, and found that a realistic treatment when inserting the simulated galaxies into sky images was important to achieve good performance - even more so than the radiative transfer model in the simulations themselves. That said, while these caveats suggest simulated images may not be fully realistic, they may still be highly useful. Convolutional networks are now routinely trained on simulated images for identifying and measuring strong lenses (see e.g. [171, 246, 256, 334, 341]). As the quality of simulations continue to improve, I expect the use of simulated morphology images as supplementary training data to become more common.

Finally, transfer learning provides an indirect method to include training data. First, a convolutional network is trained to solve a related problem on an indepen-

dent training set. The convolutional layers of the network become able to extract features relevant to that related problem. Second, those feature-extracting layers are used to construct a new convolutional network aimed at solving the target problem. The filters learned by those feature-extracting layers may be useful to re-apply. For example, learned filters that detect shapes and orientation on the related problem may be helpful for the target problem (see [468]). The features learned by CNNs trained on general galaxy morphology problems with far larger samples ([100, 191, 456]) could be particularly relevant for detecting faint tidal features. Ackermann et al. 2018 [6], concurrently with the work shared in this chapter, used transfer learning in conjunction with CNNs to automatically identify images of galaxy mergers.

2.5.3 Scaling

The ultimate purpose of our convolutional neural networks is to detect tidal features in a large galaxy sample not previously classified. It is therefore important to ensure that that this method scales.

Each ensemble classifier makes tidal predictions on the order of 100 galaxies per second on a standard 2.4Ghz CPU, or approximately eight million galaxies per CPU-day. This means that classifying forthcoming samples from LSST and Euclid, which will be several orders-of-magnitude larger than the A13 sample, is computationally feasible. Using GPUs would likely further increase speed, as GPUs are far more efficient at calculating the matrix multiplications underlying deep learning; I use GPUs extensively in the following chapters.

A13 manually removed images contaminated by stars, which would not be feasible for a large sample. However, automatic identification of contaminating stars is straightforward [64, 223, 379, 397]. Current methods reach an AUC score exceeding 0.99 on comparable CFHT imaging [229]. For LSST-scale samples, one could use such methods to automatically remove contaminating stars prior to application of our convolutional neural network.

We decided to remove as uninformative 8% of images (136 of 1757) with expert labels of exactly 50% confidence in tidal features. The performance metrics reported apply only to this slightly cleaner sample. Assuming classifiers guess randomly for such uncertain galaxies, and the true labels are equally random, the AUC scores of all the methods discussed would be slightly lower. This does not affect our demonstration of the relative strength of convolutional neural networks at detecting tidal features.

2.5.4 Potential Bias

Scalability is only meaningful if we understand the biases involved in the classifications. Below, I discuss two important sources of bias that may be introduced by the classifier.

In the first case, the classifier may perform particularly poorly at recognising some classes of tidal features (e.g. streams or shells). It is crucial to understand these biases so that they may be distinguished from genuine trends in the galaxy population. One way to approach this would be to construct a 'calibration' catalogue where the true tidal feature labels are known. This could be achieved through using multi-expert visual classifications, or even synthetic data. Given a calibration catalogue, one can measure how classifier performance varies for each tidal feature class. I presented the performance of our classifier by tidal debris class in Section 2.2.7. Should some classes be poorly recognised, one could either apply an appropriate correction or search for additional examples of that tidal feature class to improve performance.

On the other hand, within any given dataset, bias may be triggered by the image context. Experts understand that they should not consider bright foreground or background objects, diffraction spikes or any other 'artefacts' when making a classification. CNNs have no such expertise unless inferred from the training data. Further domain-specific augmentations could help the classifier avoid confusion from these context biases. Adding synthetic observational effects would provide training examples to teach the classifier to ignore such effects and better handle, for example, classifications of galaxies in crowded images.

2.6 Conclusion

In this chapter, I described using CNNs with dropout and augmentation to identify galaxies in the CFHTLS-Wide Survey that have faint tidal features in their outer regions. Learning the ideal features to extract from the pixel data and gradually increasing the pixel scale of feature maps make CNNs effective at classifying features in complex images. I have shown that appropriate preprocessing and augmentation combined with a relatively shallow network architecture is key to avoiding overfitting of the data. Randomised five-fold cross-validation verifies that these results are independent of which images are selected for training and which for testing. Training and testing five uniquely-instantiated CNNs in two different ensemble configurations confirms that these results are statistically reliable and do not result from a fortuitous instantiation of initial weights. Through adding mock tidal features, I have

shown that this method highlights image features that are found to be discriminatory without applying a parametric model.

Comparing the performance of our classifiers against previously-published expert visual classifications, our method achieves high (76%) completeness and low (20%) contamination. It also performs considerably better than other automated methods recently applied in the literature, namely the shape asymmetry method, a non-parametric approach developed for identifying post-merger galaxies by [333]), and WND-CHRM, a generic machine learning approach previously applied to image classification in astronomy ([380]).

This demonstration of the effectiveness of CNNs represents a significant step forward in developing a fully-automated method for faint tidal feature detection in galaxies. Indeed, most work in detecting and classifying low surface brightness tidal features in galaxies was, at the time our work was first published, wholly or partially dependent on expert visual identification (e.g. Kado-Fong et al. 2018 [207], Hood et al. 2018 [177], Morales et al. 2018 [308]) - and this remains true today (e.g. Bílek et al. 2020 [42]). Expert visual classification alone is completely inadequate for the next-generation of deep wide field surveys, such as LSST and Euclid, which will cover $\sim 15,000-20,000$ square degrees at unprecedented photometric depth [248, 359]. Encouragingly, the use of deep learning for detecting mergers has become common (see e.g. Bottrell et al. 2019 [51], Pearson et al. 2019 [335], Ferreira et al. 2020 [125], Ciprijanovic et al. 2020 [79]) and I expect this trend will ultimately extend to low surface brightness features. While a limiting factor is the lack of currently-available training data, the use of either citizen science labels, simulation data or transfer learning are potential ways to address this. The development of a robust and efficient method to not only identify, but also characterise, faint tidal features around galaxies will enable the record of minor mergers and interactions to be mined in very large statistical samples. This will provide unique and previously inaccessible insight into the history of the galaxy population over cosmic time and facilitate the much-anticipated revolution that next generation facilities promise in terms of quantitative low surface brightness science.

Chapter 3

Probabilistic Galaxy Morphology through Bayesian CNNs and Active Learning

3.1 Probabilistic Classification

3.1.1 Introduction

Morphology is a key driver and tracer of galaxy evolution. For example, bars are thought to move gas inwards [368] either driving or shutting down star formation [200, 385], and bulges are linked to global quenching [45, 122, 293] and inside-out quenching [261, 400]. Morphology also records other key star formation drivers such as the merger history of a galaxy. Mergers contribute to galaxy assembly [288, 448], though their relative contribution is an open question [74], and may create tidal features, bulges, and disks, allowing past mergers to be identified [58, 133, 179, 219]. I introduced a new method to detect such tidal features in Chapter 2

Unpicking the complex interplay between morphology and galaxy evolution requires measurements of detailed morphology in large samples. While modern surveys reveal exquisite morphological detail, they image far more galaxies than scientists can visually classify. Galaxy Zoo solves this problem by asking members of the public to volunteer as 'citizen scientists' and provide classifications through a web interface. Galaxy Zoo has provided morphology measurements for surveys including SDSS [265, 456] and large HST programs [389, 459].

Knowing the morphology of homogeneous samples of hundreds of thousands of galaxies supports science only possible at scale. The catalogues produced by the collective effort of Galaxy Zoo volunteers have been used as the foundation of a large number of studies of galaxy morphology [295]. Galaxy Zoo measures subtle effects

in large populations [158, 292, 458]; identifies unusual populations that challenge standard astrophysics [240, 390, 426]; and finds unexpected and interesting objects that provide unique data on broader galaxy evolution questions [69, 221, 266].

Galaxy Zoo was created because SDSS-scale surveys could not be visually classified by professional astronomers [265]. In turn, Galaxy Zoo is being gradually outpaced by the increasing scale of modern surveys like DES [130], PanSTARRS [208], the Kilo-Degree Survey [95], and Hyper Suprime-Cam [13]. Each of these surveys can each image galaxies as fast or faster than those galaxies are being classified by volunteers. For example, DECaLS [99], the focus of chapter 4, contains (as of Data Release 5) approximately 350,000 galaxies suitable for detailed morphological classification (applying $r < 17$ and `petroR90_r`[1] > 3 arcsec, the cuts used for Galaxy Zoo 2 [456]). Collecting 40 independent volunteer classifications for each galaxy, as for Galaxy Zoo 2, would take approximately eight years without further promotion efforts - by which time we expect new surveys to start. The Galaxy Zoo science team must therefore both judiciously select which surveys to classify and, for the selected surveys, reduce the number of independent classifications per galaxy. The speed at which galaxies can be classified severely limits the scale, detail, and quality of morphology catalogues, diminishing the scientific value of such surveys.

The next generation of surveys will make this speed limitation even more stark. Euclid[2], the Vera Rubin Observatory[3] and the Roman Space Telescope [4] are expected to resolve the morphology of unprecedented numbers of galaxies. This could be revolutionary for our understanding of galaxy evolution, but only if such galaxies can be classified. The future of morphology research therefore inevitably relies on automated classification methods. Given a working prediction algorithm, the time required to classify new galaxies becomes proportional to computation, rather than human time, and computation is cheap. The challenge is typically in creating that algorithm; deep learning methods in particular are often data-hungry due to dimensionality (Sec. 1.2). In the previous chapter, I mitigated this need for large training sets by using data augmentations and a relatively small model (Sec. 2.2.3-2.2.4). In this and the following chapter, I develop more sophisticated approaches. Automated classification also brings new opportunities which I feel are likely to dramatically reshape the field of galaxy morphology; see Sec. 4.8 for a discussion.

[1] `petroR90_r` is the Petrosian radius which contains 90% of the r-band flux
[2] 15,000 deg^2 at $0.30''$ half-light radius PSF from 2022, [248]
[3] 18,000 deg^2 to $0.39''$ half-light radius PSF from 2023, [275]
[4] 2,000 deg^2 at $0.12''$ half-light radius PSF from approx. 2025, [399]

Automated classification of galaxies has a long history. I previously described (Sec. 2.1) non-parametric methods designed to distinguish galaxies of different morphologies; notably Concentration C and Asymmetry A by Abraham et al. 1994 [5] [5], Smoothness by Conselice 2003 [84], Gini and M_{20} by Lotz et al. 2004 [270], and Multi-Mode, Intensity, and Deviation by Freeman et al. 2013 [139]. As the number of possible non-parametric measurements grew, researchers began combining them by finding decision boundaries with low-dimensional machine learning methods including principal component analysis [340, 371] and (at the time, recently invented) support-vector machines [190, 190, 193].

In parallel, starting around 1990, astronomers became interested in neural networks (Sec. 1.3). Classifying galaxy morphology was one of the first applications. Storrie-Lombardi et al. 1992 [407] were the first to use neural networks to classify galaxies, writing that 'such automated procedures are the only practical way of classifying the enormous amounts of data produced by machine scans of Schmidt plates' - illustrating how 'big data' is not a new problem. Early neural networks [29, 31, 242, 312] primarily used parametric features such as colours, axial ratios and de Vaucouleurs exponents [97] previously derived for galaxy catalogues (e.g. [247, 467]).

Both parametric and non-parametric features suffer from intrinsic limitations in their ability to detect specific morphological features, as I previously discussed in Sec. 2.1. The basic issue is that images of galaxies are difficult to summarise with a few numbers. Adding machine learning to interpret or combine those numbers does not resolve this. Odewahn et al. 2002 writes 'it must be conceded that none of these systems are truly morphological in nature'[6].

Convolutional neural networks (CNNs), in contrast, learn to detect features directly from pixel data. I previously gave a general introduction to the advantages of CNNs (Sec. 1.2) and their use in astronomy, including for inference on galaxy images (Sec. 1.3); here, I focus on the application of CNNs for classifying galaxy morphology. The first use of CNNs in astronomy was for morphology; Dieleman et al. 2015 [100] reproduced Galaxy Zoo 2 votes to win the Kaggle 'Galaxy Challenge'[7]. Huertas-Company et al. 2015 [191] followed shortly after, using almost the same architecture to predict expert classifications of Kartaltepe et al. 2015 for higher redshift ($z = 0.5 - 3$)

[5] Abraham et al. 1994 was partly motivated by practicality, writing: 'automated morphological classifications based on the central concentration of light are much simpler to implement than automated classifications based on the Hubble system'

[6] Odewhan et. al. writes in relation to machine learning with non-parametric features, but the same argument applies to parametric features

[7] https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

CANDELS galaxies. CNNs have since become the standard approach. More recent work applies CNNs to classify morphology in new surveys [227, 339, 370] and simulations [194], or to identify interesting populations like mergers [6] and high-redshift starforming galaxies [192]. CNNs have also been used at other wavelengths, being used to divide galaxies by radio morphology features like Fanaroff-Riley jets [279, 416] and even for IR/radio source cross-matching and classification [461]. Recent comparison experiments have conclusively shown that CNNs make more accurate predictions of Galaxy Zoo volunteer responses than low-dimensional models [32, 76].

However, despite major progress in raw performance, the increasing complexity of automated classification methods poses a problem for scientific inquiry (Sec. 1.1). The CNNs used in previous work do not account for uncertainty in training labels, limiting their ability to learn from all available labelled galaxies (one common approach is to train only on 'clean' subsets). Further, these CNNs were not typically designed to make probabilistic predictions (though they have sometimes been interpreted as such), limiting the reliability of conclusions drawn using such methods.

In this chapter, first published in Walmsley et al. 2020 [445], my co-authors and I combine a novel generative model of volunteer responses with Monte Carlo dropout [141] to create Bayesian CNNs that predict *posteriors* for the morphology of each galaxy. Posteriors are crucial for drawing statistical conclusions that account for uncertainty, and so including posteriors significantly increases the scientific value of morphology catalogues. Posteriors also allow our Bayesian CNN to account for varying (i.e. heteroskedastic) uncertainty in volunteer responses, making maximal use of the available data.

3.1.2 Motivating Example - Overconfident Bar Predictions

CNN predictions are not (in general) well-calibrated probabilities [155, 243]. Given the extensive success of CNN at accurately predicting galaxy morphologies, one might wonder - does that matter? The reason that accuracy is insufficient without calibration is that poorly-calibrated probabilities may cause systematic errors in later analysis, even if the model is quite accurate overall. To illustrate this problem, and emphasise the advantages of using probabilistic methods, I describe here an experiment showing that the CNN probabilities published in DS+18 [370] significantly overestimate the prevalence of expert-classified barred galaxies. I chose DS+18 as the most recent deep learning morphology catalogue made publicly available, and thank the authors for their openness. I do not believe this issue is unique to DS+18.

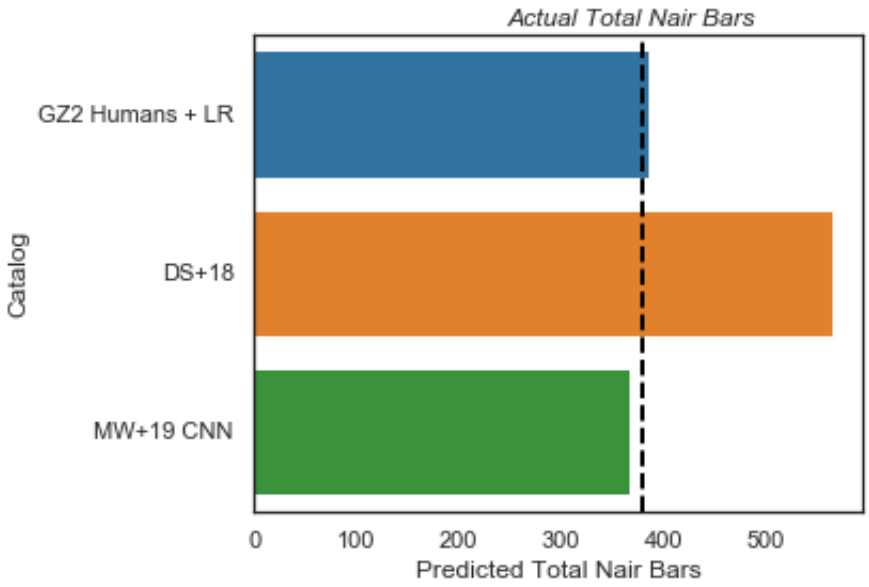


Figure 3.1: Predictions for the total number of galaxies labelled as 'Bar' by human expert NA10 in test galaxy subset (correct answer: 379). The predictions of DS+18 overestimate the number of Nair Bars (559). GZ2 vote fractions from volunteers can be used to make an improved estimate (396) with a rescaling correction calculated via logistic regression (GZ2 Humans + LR). Applying the same correction to the vote fractions predicted by the Bayesian CNN in this work (MW+19) also produces an improved estimate (372). By accurately predicting the vote fractions, and then applying a correction to map from vote fractions to expert responses, one can predict what NA10 would have said for the full SDSS sample.

DS+18 trained a CNN to predict the probability that a galaxy is barred (DS+18 Section 5.3). Barred galaxies were defined as those galaxies labelled as having any kind of bar (weak/intermediate/strong) in the expert catalogue by Nair and Abraham (2010) (NA10). I refer to such galaxies as Nair Bars. I chose to investigate this particular DS+18 model because it explicitly aims to reproduce the (expert) NA10 classifications, allowing for a direct comparison of the predicted probabilities against the true labels. Such a comparison is not straightforward for other DS+18 models because they aim to *improve* upon (crowdsourced) labels.

For my experiment, I selected a random subset of 1211 galaxies classified by NA10 (this subset is motivated below). How many barred galaxies are in this subset? The DS+18 Nair Bar 'probabilities' p_i (for each galaxy i) predict $\sum_i p_i = 559$ Nair Bars. However, only 379 are actually Nair Bars (Figure 3.1). This error is caused by the DS+18 Nair Bar 'probabilities' being, on average, skewed towards predicting 'Bar', as shown by the calibration curve of the DS+18 Nair Bar probabilities (Figure 3.2).

How can we better predict the total number of Nair Bars? GZ2 collected volunteer responses for many galaxies classified by NA10 (6,051 of 14,034 match within 5″, after filtering for total 'Bar?' votes $N_{\rm bar} > 10$ - see chapter 3.1.8). The fraction of volunteers who responded 'Bar' to the question 'Bar?' is predictive of Nair Bars, but is not a probability [265]. For example, volunteers are less able to recognise weak bars than experts [294], and hence the 'Bar' vote fraction only slightly increases for galaxies with weak Nair Bars versus galaxies without. We need to rescale the GZ2 vote fractions. To do this, I divided the NA10 catalogue into 80% train and 20% test subsets and used the train subset to fit (via logistic regression) a rescaling function (Figure 3.3) mapping GZ2 vote fractions to p(Nair Bar|GZ2 Fraction). I then evaluated the calibration of these probabilities on the test subset, which is the subset of 1211 galaxies used above. The rescaled volunteer votes predict 396 Nair Bars, which compares well with the correct answer of 379 versus the DS+18 answer of 559 (Figure 3.1). This directly demonstrates that our rescaled GZ2 predictions are correctly calibrated over the full test subset. The calibration curve shows no systematic skew, unlike DS+18 (Figure 3.2).

The rescaled volunteer votes imply the correct number of expert bars, but we cannot expect to always have such votes; ultimately, we need an automated classifier (Sec. 3.1.1). Since the GZ2 vote fractions can be rescaled to Nair Bar probabilities, automated vote fraction predictions could be converted to Nair Bar probabilities using the same rescaling function. In the remainder of this chapter, I will introduce a Bayesian CNN which makes vote fraction predictions. Those vote fraction predictions from my final Bayesian CNN ultimately correctly estimate the count of Nair Bars (372 bars predicted versus 379 observed bars, Figure 3.1).

3.1.3 Probabilistic Framework for Galaxy Zoo

I have shown above that probabilistic predictions are important for drawing reliable scientific conclusions. Next, I describe how I approached making probabilistic predictions for Galaxy Zoo volunteers.

Within the context of machine learning, the uncertainty in a prediction is often divided into two parts; epistemic uncertainty and aleatoric uncertainty [222]. Epistemic uncertainty is uncertainty from the model's limited knowledge about the world - perhaps the model has not 'seen' a particular kind of galaxy before, for example. Expanding the training set will better inform the model and better constrain its parameters, reducing epistemic uncertainty. Accounting for epistemic uncertainty is challenging for supervised deep learning; I return to this issue in Sec. 3.1.5. Aleatoric

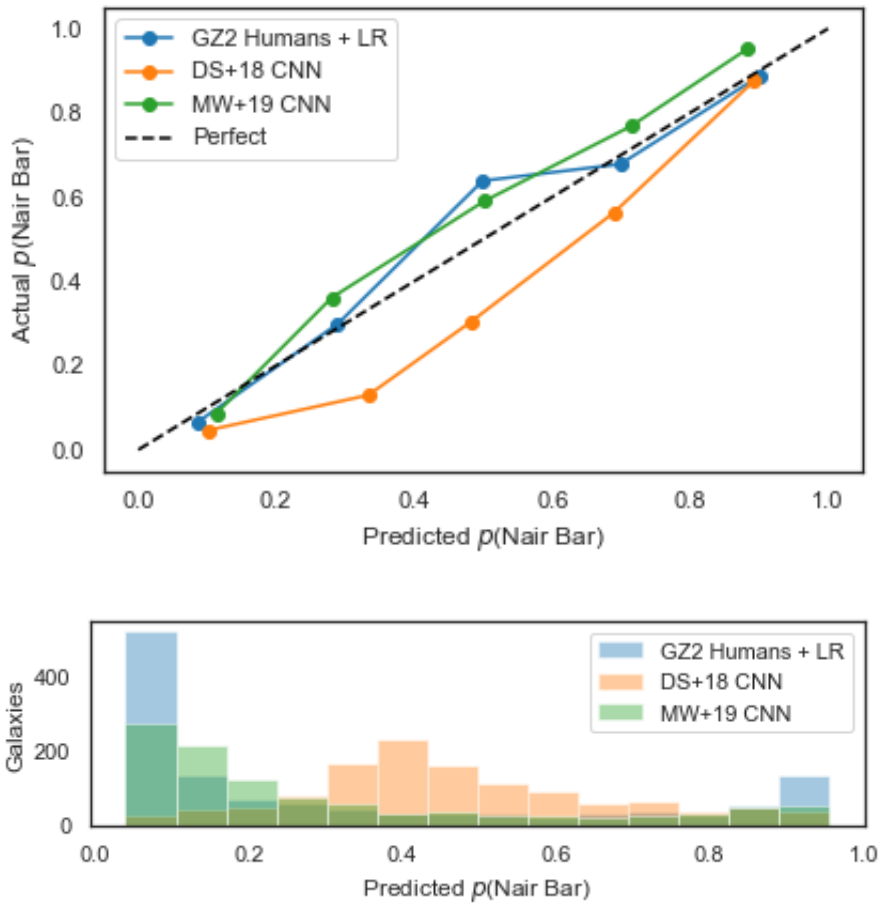


Figure 3.2: Above: Comparison of calibration curves for each predictive model. The calibration curve is calculated by binning the predicted probabilities and counting the fraction of Nair Bars in each bin. The fraction of Nair Bars in a given bin approximates the true (frequentist) probability of each binned galaxy being a Nair Bar. Points compare the predicted fraction of Nair Bars (x axis) with the actual fraction (y axis) for 5 equally-spaced bins. For well-calibrated probabilities, the predicted and actual fractions are equal (black dashed line). Below: the distribution of Nair Bar predictions from each model. DS+18 typically predicts $p \sim 0.4$ (below) and has a relatively poor calibration near $p \sim 0.4$ (above), leading to a significant overestimate of the total number of Nair Bars.

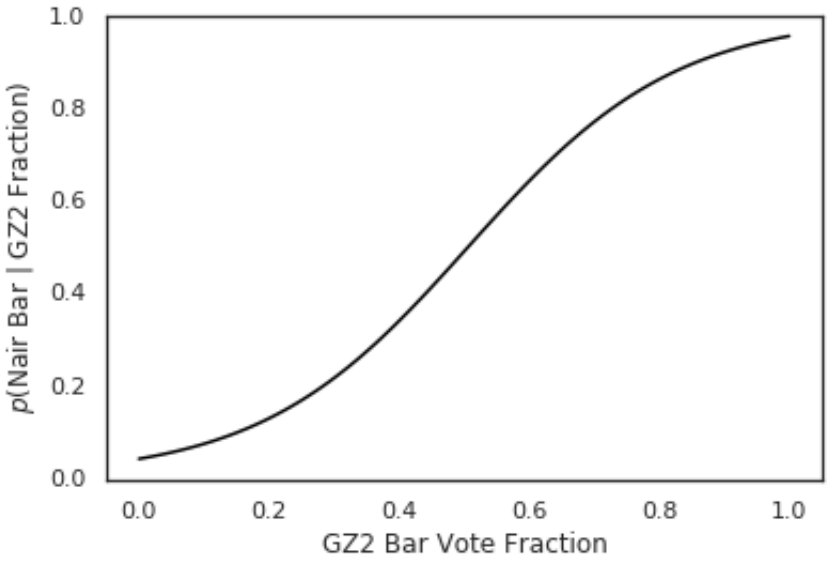


Figure 3.3: The rescaling function used to map GZ2 vote fractions to p(Nair Bar|GZ2 Fraction), estimated via logistic regression. This rescaling function is also used (without modification) to map Bayesian CNN GZ2 vote fraction predictions to p(Nair Bar|BCNN-predicted GZ2 Fraction)

Not that the rank of the vote fractions is unchanged.

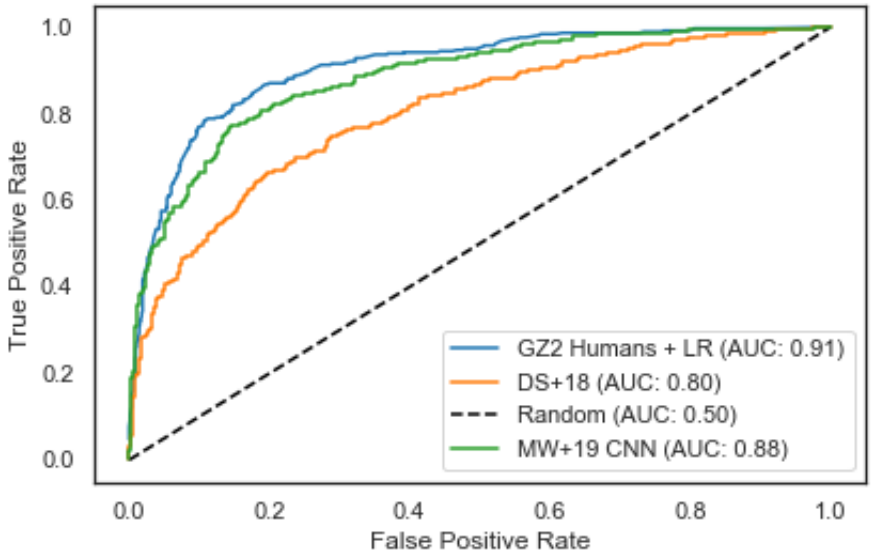


Figure 3.4: Comparison of ROC curves for predicting Nair Bars using each model.

uncertainty is uncertainty intrinsic to the data ('aleator' is Latin for 'dice-thrower'). Aleatoric uncertainty is fixed - one cannot predict a dice throw no matter how much training data is collected. [8] Aleatoric uncertainty is an important aspect of predicting galaxy morphology with Galaxy Zoo because volunteer vote fractions are heteroskedastic; they have different aleatoric uncertainties. For a given question, many galaxies receive answers from only a few volunteers (see Figure 3.5) and so have a far higher variance in their vote fractions. However, previous work has overlooked this heteroskedasticity. For example, the Kaggle Galaxy Zoo Challenge [100] asked participants to minimise the root-mean-square error (summed over all answers). Minimising the closely-related mean-squared-error loss function is trivially equivalent to maximising the likelihood of the observed labels under Gaussian noise of constant variance. The assumption of constant variance is problematic here.

Nix and Weigend 1994 [321] introduced the idea of explicitly interpreting model outputs as Gaussian distributions, predicting both a mean and variance for each datapoint. This allowed the variance to be learned as function of the data; in other words, for the model to express degrees of confidence in each prediction. This idea has since used widely, including for Gaussian Processes [251] and CNNs [222]. Predicting Gaussians for Galaxy Zoo would not have been ideal as the vote fractions are defined on the [0, 1] unit interval, while Gaussians are defined on $\mathbb{R}$. I could have truncated them and constrained the means with a sigmoid, but this felt inelegant. Instead, I chose to use a more appropriate distribution; a Binomial.

The statistical model I introduced is derived as follows. Formally, each Galaxy Zoo decision tree question asks N_i volunteers to view galaxy image x_i and select the most appropriate answer A_j from the available answers $\{A\}$. This reduces to a binary choice; where there are more than two available answers ($|\{A\}| > 2$), one can consider each volunteer response as either A_j (positive response) or not A_j (negative response). A binary model is therefore useful for questions with any number of answers.

Let k_{ij} be the number of volunteers (out of N_i) observed to answer A_j for image x_i. Aassume that there is a true fraction ρ_{ij} of the population (i.e. all possible volunteers) who would give the answer A_j for image x_i. Also assume that volunteers are drawn uniformly from this population, so that if one asks N_i volunteers about image x_i, one expects that the distribution over the number of positive answers k_{ij} to be binomial:

$$k_{ij} \sim \text{Bin}(\rho_{ij}, N_i) \tag{3.1}$$

[8] The physicist reader may note that dice throws are entirely predictable given sufficient measurements. Aleatoric uncertainty might be better described to physicists as measurement uncertainty.

$$p(k_{ij}|x_{ij}, N_i) = \binom{N_i}{k_i} \rho_{ij}^{k_{ij}} (1 - \rho_{ij})^{N_i - k_{ij}} \tag{3.2}$$

This is the model I used for how each volunteer response k_{ij} was generated. Note that ρ_{ij} is a latent variable: one only observe the responses k_{ij}, never ρ_{ij} itself.

3.1.4 Probabilistic Prediction with CNNs

I used this novel generative model to infer the likelihood of observing a particular k for each galaxy x, described below. For brevity, I omit subscripts.

Consider the scalar output from a neural network $f^w(x)$ as a (deterministic) prediction for ρ, and hence a probabilistic prediction for k:

$$p(k|x, w) = \text{Bin}(k|f^w(x), N) \tag{3.3}$$

For each labelled galaxy, k positive responses are recorded from Galaxy Zoo volunteers. Ideally, one would like to find the network weights w such that $p(k|x, N)$ is maximised (i.e. to make a maximum likelihood estimate given the observations):

$$\max_w [p(k|x, w)] = \max_w [\text{Bin}(k|f^w(x), N)] \tag{3.4}$$

$$= \max_w [\log \binom{N}{k} + k \log f^w(x) + (N - k) \log(1 - f^w(x))] \tag{3.5}$$

The combinatorial term is fixed and hence the objective function to minimise is

$$\mathcal{L} = k \log f^w(x) + (N - k) \log(1 - f^w(x)) \tag{3.6}$$

One can create a probabilistic model for k by optimising the network to make maximum likelihood estimates $\hat{\rho} = f^w(x)$ for the latent parameter ρ from which k is drawn.

In short, each network w predicts the response probability ρ that a random volunteer will select a given answer for a given image.

3.1.5 From Probabilistic to Bayesian CNN

My approach above leads to a model that accounts for the aleatoric counting uncertainty in the volunteer vote fractions. I now return to considering epistemic uncertainty; uncertainty from lack of knowledge about the world, or more formally, lack of constraints on the model weights. Handling epistemic uncertainty is the focus of *Bayesian deep learning*, which aims to place distributions over the model weights $p(w|\mathcal{D})$ (intuitively, how likely each possible model is). This is important because to

predict a Bayesian posterior of k given $\mathcal{D}$, one should marginalise over these weight distributions:

$$p(k|x, \mathcal{D}) = \int p(k|x, w)p(w|\mathcal{D})dw \tag{3.7}$$

Unfortunately, in practice, you do not know the weight distributions. You only observe the single model that was actually trained. Bayesian deep learning practitioners have attempted to work around this limitation through various means, including placing approximating distributions q_θ over the weights (variational inference, [46, 154, 174]) and specialised forms of Hamiltonian Monte Carlo sampling [316, 451]. These achieved some success but generally suffer from a high computational cost for all but the simplest networks or approximating distributions. Instead, Gal 2016 [141] suggested using dropout [404] as an alternative. Dropout is a regularization method that temporarily removes random neurons according to a Bernoulli distribution, where the probability of removal ('dropout rate') is a hyperparameter to be chosen. Dropout may be interpreted as taking the trained model and permuting it into a different one [404]. Gal 2016 [141] introduced the approach of approximating the distributions of models one might have trained, but didn't, with the distribution of models from applying dropout:

$$p(w|\mathcal{D}) \approx q^* \tag{3.8}$$

removing neurons according to dropout distribution q^*. This is the Monte Carlo Dropout approximation (hereafter MC Dropout). It can be shown that MC Dropout is theoretically equivalent to performing variational inference over the network weights, albeit with an unbounded approximation error [141]. Of more interest to the pragmatic astronomer, MC Dropout uncertainty estimates are both straightforward to calculate and moderately accurate [143].

Choosing the dropout rate affects the approximation; greater dropout rates lead the model to estimate higher uncertainties (on average) [142]. Following convention, I arbitrarily chose a dropout rate of 0.5. I discuss the implications of using an arbitrary dropout rate, and opportunities for improvement, in Section 3.3.

Applying MC Dropout to marginalise over models (Eqn. 3.7):

$$p(k|x, \mathcal{D}) = \int p(k|x, w)q^*dw \tag{3.9}$$

In practice, following Gal 2016 , predictions are made by sampling from q^* with T forward passes using dropout *at test time* (i.e. Monte Carlo integration):

$$\int p(k|x,w)q^*dw \approx \frac{1}{T}\sum_t p(k|x,w_t) \tag{3.10}$$

MC Dropout improves predicted posteriors by (approximately) marginalising over the possible models that might have been trained.

To demonstrate the probabilistic model and the use of MC Dropout, I trained models to predict volunteer responses to the 'Smooth or Featured' and 'Bar' questions on Galaxy Zoo 2 (Section 3.1.7).

3.1.6 Data - Galaxy Zoo 2

Galaxy Zoo 2 (GZ2) classified all 304,122 galaxies from the Sloan Digital Sky Survey (SDSS) DR7 Main Galaxy Sample [3, 408] with $r < 17$ and `petroR90_r`[9] > 3 arcsec. Classifying 304,122 galaxies required ~ 60 million volunteer responses collected over 14 months.

GZ2 is the largest homogenous galaxy sample with reliable measurements of detailed morphology, and hence was an ideal data source for this work. GZ2 has been extensively used as a benchmark to compare machine learning methods for classifying galaxy morphology. The original GZ2 data release [456] included comparisons with (pre-CNN) machine learning methods by Baillard et al. 2011 [27] and Huertas-Company et al. 2011 [193]. GZ2 subsequently provided the data for seminal work on CNN morphology classification [100] and continues to be used for validating new approaches [227, 370].

I used the 'original' subset of the 'GZ2 normal-depth sample' catalogue (hereafter 'GZ2 catalogue'), available from data.galaxyzoo.org [157]. This excludes galaxies fainter than $r < 17$, explicitly excludes galaxies in stripe82, and requires spectroscopic redshift measurements. I downloaded the images shown to volunteers using an internal catalogue of the image URLs; the images have subsequently been made available at [457].

The GZ2 catalogue provides aggregate volunteer responses at each of the three post-processing stages: raw vote counts (and derived vote fractions), consensus vote fractions, and redshift-debiased vote fractions. The raw vote counts are simply the number of users who selected each answer. The consensus vote fractions are calculated by iteratively re-weighting each user based on their overall agreement with other users. The debiased fractions estimate how the galaxy would have been classified if viewed at $z = 0.03$ [157]. Unlike recent work [227, 370], I chose to use the raw vote counts.

[9] `petroR90_r` is the Petrosian radius which contains 90% of the r-band flux

The redshift-debiased fractions estimate the *true* morphology of a galaxy, not what the image actually *shows*. To predict what volunteers would say about an image, in my view, one should only consider what the volunteers see. I believe that debiasing is better applied after predicting responses, not before. The performance metrics in this chapter are therefore not directly comparable to those of [370] and [227], who use the debiased fractions as ground truth.

3.1.7 Application

3.1.7.1 Tasks

To test the probabilistic CNNs, I predicted volunteer responses for the 'Smooth or Featured' and 'Bar' questions.

The 'Smooth or Featured' question asks volunteers 'Is the galaxy simply smooth and rounded, with no sign of a disk?' with (common[10]) answers 'Smooth' and 'Featured or Disk'. As 'Smooth or Featured' is the first decision tree question, this question is always asked, and therefore every galaxy has ~ 40 'Smooth or Featured' responses[11]. With N fixed to ~ 40 responses, the loss function (Eqn. 3.6) depends only on k (for a given model w).

The 'Bar' question asks volunteers 'Is there a sign of a bar feature through the centre of the galaxy?' with answers 'Bar (Yes)' and 'No Bar'. Because 'Bar' is only asked if volunteers respond 'Featured' and 'Not Edge-On' to previous questions, each galaxy can have anywhere from 0 to 40 total responses – typically around 10 (Figure 3.5). This scenario is common; only 2 questions are always asked, and most questions have $N << 40$ total responses (Figure 3.5). Building probabilistic CNNs that learn better by appreciating the varying count uncertainty in volunteer responses is a key advantage of the design developed here.

3.1.7.2 Architecture

The CNN architecture I used is shown in Figure 3.6. This architecture was inspired by VGG16 [391], but scaled down to be shallower and narrower to fit our computational budget. I used a softmax final layer to ensure the predicted typical vote fraction ρ lies between 0 and 1, as required by our binomial loss function (Equation 3.6).

[10]'Smooth or Featured' includes a third 'Artifact' answer. However, artifacts are sufficiently rare (0.08% of objects have 'Artifact' as the majority response) that predicting 'Smooth' or 'Not Smooth' is sufficient to separate smooth and featured galaxies in practice.

[11]Technical limitations and rare artifacts during GZ2 caused 26,530 galaxies to have $N < 36$ (excluding 'Artifact' responses). We exclude these galaxies for simplicity.

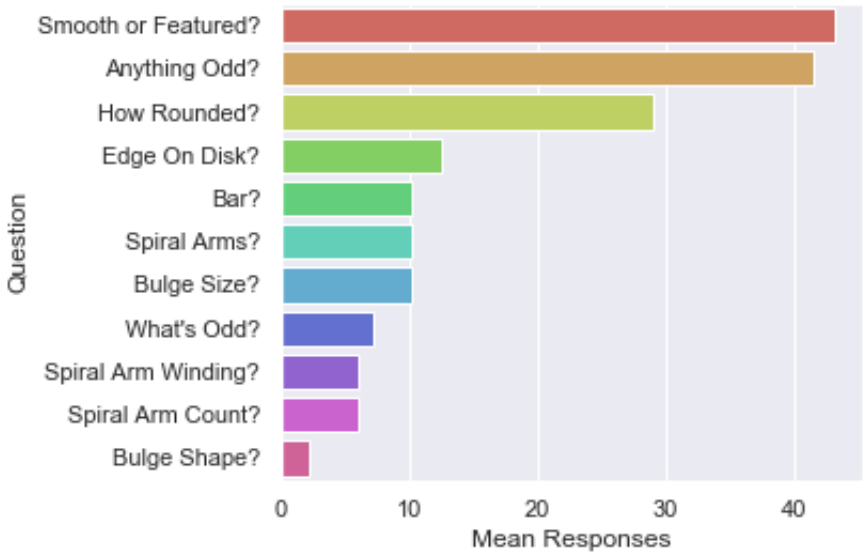


Figure 3.5: Mean responses (N) by GZ2 question. Being the first question, 'Smooth or Featured' has an unusually high (~ 40) number of responses. Most questions (6 of 11), including 'Bar', are only asked for 'Featured' galaxies, and hence have only ~ 10 responses. Training CNNs while accounting for the label uncertainty caused by low N responses is a key goal of this work.

During this work, I was primarily concerned with accounting for label uncertainty and predicting posteriors, rather than maximising performance metrics. That said, the final model is competitive with, or outperforms, previous work (Section 3.1.9.1). Overall performance can likely be significantly improved with more recent architectures [162, 187, 411] and a correspondingly larger computational budget. I go on to improve the architecture in chapter 4.

3.1.7.3 Augmentations

To generate the training and test images, I resized the original 424x424x3 pixel GZ2 `png` images shown to volunteers into 256x256x3 `uint8` [12] matrices and saved these matrices in TFRecords (to facilitate rapid loading). When serving training images to the model, the following transformations are applied to each image:

1. Average over channels to create a greyscale image

2. Perform random horizontal and/or vertical flips

[12]Unsigned 8-bit integer i.e. 0-255 inclusive. After rescaling, this is sufficient to express the dynamic range of the images (as judged by visual inspection) while significantly reducing memory requirements versus the original 32-bit float flux measurements.

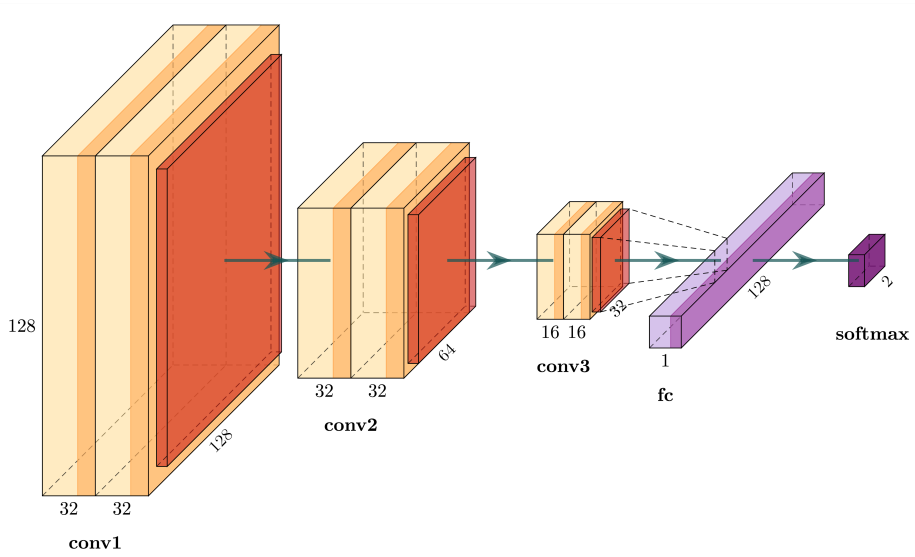


Figure 3.6: The CNN architecture used throughout. The input image, after applying augmentations (Section 3.1.7.3), is of dimension 128x128x1. The first pair of convolutional layers are each of dimension 128x128x32 with 3x3 kernels. These are reduced by max-pooling for a second pair of convolutional layers of dimension 64x64x32 with 3x3 kernels, then again to a final pair of dimension 32x32x16 with 3x3 kernels. The network ends with a 128-neuron linear dense layer and a 2-neuron softmax dense layer.

3. Rotate through an angle randomly selected from $0 to 90$ (using nearest-neighbour interpolation to fill pixels)

4. Adjust the image contrast to a contrast uniformly selected from 98% to 102% of the original contrast

5. Crop either randomly ('Smooth or Featured') or centrally ('Bar') according to a zoom level uniformly selected from 1.1x to 1.3x ('Smooth or Featured') or 1.7x to 1.9x ('Bar')

6. Resize to a target size of 128x128(x1)

I trained the network on greyscale images because colour is often predictive of galaxy type (E and S0 are predominantly redder, while S are bluer, [358]) and I wanted to ensure that our classifier does not learn to make biased predictions from this correlation. For example, a galaxy should be classified as smooth because it appears smooth, and not because it is red and therefore more likely to be smooth. Otherwise, any later research investigating correlations between morphology and colour would be biased.

Random flips, rotations, contrast adjustment, and zooms (via crops) help the CNN learn that predictions should be invariant to these transformations - predictions should not change because the image is flipped, for example. I chose a higher zoom level for 'Bar' because the original image radius for GZ2 was designed to show the full galaxy and any immediate neighbours [456] yet bars are generally found in the centre of galaxies [240]. The 'Bar' classification should be invariant to all but the central region of the image, and so I chose to sacrifice the outer regions in favour of increased resolution in the centre. Cropping and resizing are performed last to minimise resolution loss due to aliasing. Images are resized to match the computational budget available.

These augmentations are also applied at test time. This marginalises over any unlearned invariance using MC Dropout, as part of marginalising over networks (Section 3.1.5). Each permuted network makes predictions on a uniquely-augmented image. The aggregated posterior (over many forward passes T) is therefore independent of orientation (for example), enforcing domain knowledge.

3.1.8 Experimental Setup

For each question, 2500 galaxies are randomly selected as a test subset and the model is trained on the remaining galaxies (following the selection criteria described in Section 3.1.6). Unlike [370] and [227], I did not select a 'clean' sample of galaxies with extreme vote fractions on which to train. Instead, I took full advantage of the responses collected for every galaxy by carefully accounting for the vote uncertainty in galaxies with fewer responses (Eqn 3.6).

For 'Smooth or Featured', I used a final training sample of 176,328 galaxies. For 'Bar', I trained and tested only on galaxies with $N_{\text{bar}} \geq 10$ (56,048 galaxies). Without applying this cut, models fail to learn; performance fails to improve from random initialisation. This may be because galaxies with $N_{\text{bar}} < 10$ must have $k_{\text{featured}} < 10$ and so are almost all smooth and unbarred, leading to increasingly unbalanced typical vote fractions ρ. I solve this issue with a new statistical description of volunteers in chapter 4.

I trained the models on an Amazon Web Services (AWS) `p2.xlarge` EC2 instance with an NVIDIA K80 GPU. Training a model from random initialisation takes approximately eight hours. I then used the trained models to make predictions $\hat{\rho}$ for the typical vote fraction ρ of each galaxy in the test subsets. Finally, I evaluated performance by comparing $p(k|\hat{\rho}, N)$, the posterior for k positive responses from N volunteers, with the observed k from the N Galaxy Zoo volunteers asked.

3.1.9 Results

The probabilistic CNNs I developed produce posteriors that are reliable and informative.

Figures 3.7 and 3.8 show our posteriors for 'Smooth or Featured' and 'Bar', respectively. For each question, I show a random selection of posteriors from either 1 or 30 MC Dropout forward passes (i.e. 1 or 30 MC-dropout-approximated 'networks').

Without MC Dropout, the posteriors are binomial. The spread of each posterior reflects two effects. First, the spread reflects the extremity of $\hat{\rho}$ that previous authors have expressed as 'volunteer agreement' or 'confidence' [100, 370]. $\text{Bin}(k|\hat{\rho}, N)$ is narrower where $\hat{\rho}$ is close to 0 or 1. Second, the spread reflects N, the number of volunteers asked. For 'Smooth or Featured', where N is approximately fixed, this second effect is minor. For 'Bar', where N varies significantly between 10 and ~ 40, the posteriors are more spread (less precise) where fewer volunteers have been asked.

With MC Dropout, the posteriors are a superposition of Binomials from each forward pass, each centered on a different $\hat{\rho}_t$. In consequence, the MC Dropout posteriors are more uncertain. This is expected; I believe that marginalising over weights and augmentations causes the predictions to broaden.

Given that each single network is relatively confident and the MC-dropout-marginalised model is relatively uncertain, which should be used? I argued in section 3.1.2 that one should prefer predictions that are well-calibrated i.e. which reflect the true uncertainty in the predictions. To quantify calibration, I introduced a novel method; comparing the predicted and observed vote fractions $\frac{k}{N}$ within increasing ranges of acceptable error. I outline this procedure below.

Choose some maximum acceptable error ϵ in predicting each vote fraction $v = \frac{k}{N}$. Over all galaxies, sum the total probability (from our predicted posteriors) that $v_i = \hat{v}_i \pm \epsilon$ for each galaxy i. This is the expected count: how many galaxies the posterior suggests should have v within ϵ of the model prediction $\hat{v}$. For example, the 'Bar' model expects 2320 of 2500 galaxies in the 'Bar' test set to have an observed v within ± 0.20 of $\hat{v}$.

$$C_{\text{expected}} = \sum_{i}^{N_{\text{galaxies}}} \sum_{j>\hat{k}-N\epsilon}^{j<\hat{k}+N\epsilon} p(j|\hat{\rho}_i, N_i) \tag{3.11}$$

Next, over all galaxies, count how often v_i is within that maximum error $v_i = \hat{v}_i \pm \epsilon$. This is the 'actual' count: how many galaxies are actually observed to have v_i within

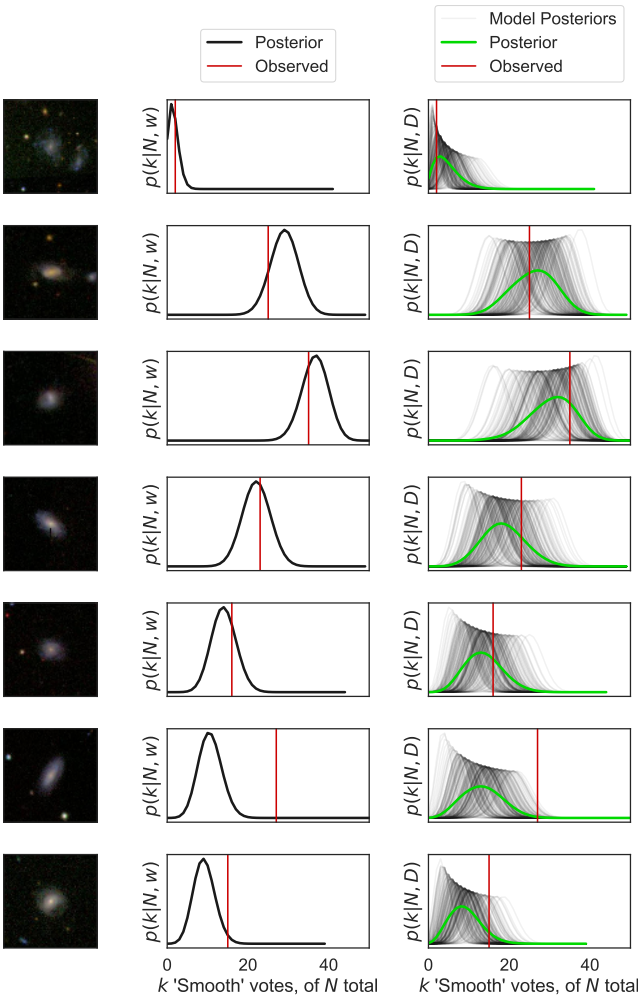


Figure 3.7: Posteriors for k of N volunteers answering 'Smooth' to the question 'Smooth or Featured?'). Each row is a randomly selected galaxy. Overplotted in red is the actual k measured from $N \sim 40$ volunteers. The left column shows the galaxy in question, shown in colour here but greyscale to the network. The central column shows the posterior predicted by a single network (black), while the right column shows the posterior marginalised (averaged) over 30 MC-dropout-approximated 'networks' (green) as well as from each 'network' (grey). While the posterior from a single network is fixed to a binomial form, the marginalised posteriors from many 'networks' can take any form. The posterior from a single network is generally more confident (narrower); I later show that a single network is overconfident, and many 'networks' are better calibrated.

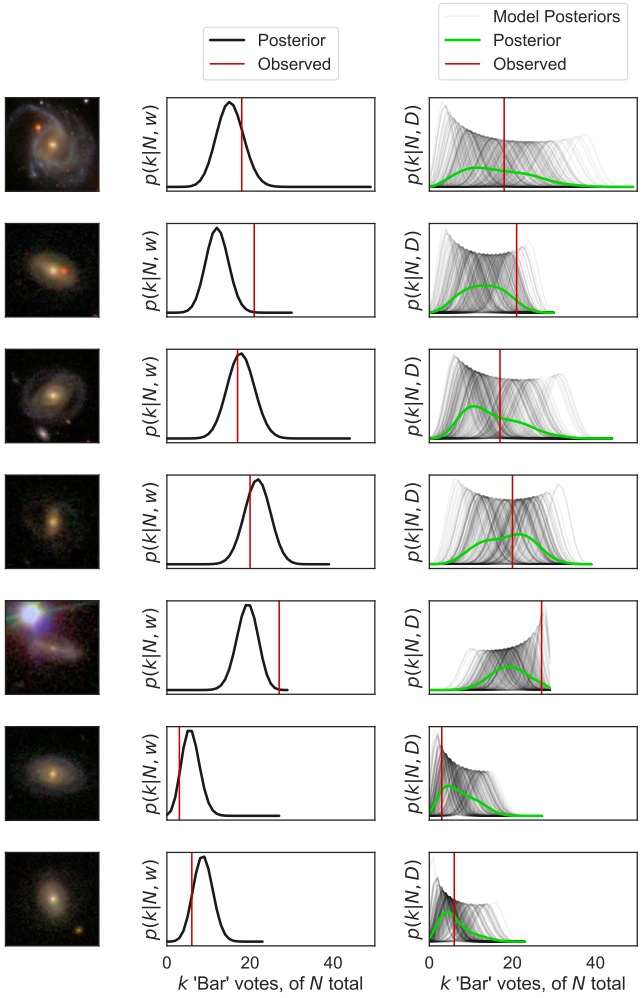


Figure 3.8: As for Figure 3.7, but showing posteriors for k of N volunteers answering 'Bar (Yes)' to the question 'Bar?'. Unlike 'Smooth or Featured', N varies significantly between galaxies, and hence so does the spread (uncertainty in k) and absolute width (highest possible k) of the posterior.

Table 3.1: Calibration results for predicting the probability that $v \pm \epsilon$ fraction of volunteers respond 'Smooth', with and without applying MC Dropout.

Max Error ϵ	Coverage Error without MC	Coverage Error with MC
0.02	49.6%	16.5%
0.05	38.5%	13.4%
0.10	26.1%	9.4%
0.20	7.9%	5.4%

Table 3.2: Calibration results for predicting the probability that $v \pm \epsilon$ fraction of volunteers respond 'Bar', with and without applying MC Dropout.

Max Error ϵ	Coverage Error without MC	Coverage Error with MC
0.02	92.2%	45.5%
0.05	85.5%	42.4%
0.10	57.8%	29.2%
0.20	22.6%	11.8%

ϵ of the model prediction $\hat{v}_i$. For example, 2075 of 2500 galaxies in the 'Bar' test set are observed to have v_i within ± 0.20 of $\hat{v}$.

$$C_{\text{actual}} = \sum_{i}^{N_{\text{galaxies}}} \sum_{j > \hat{k}_i - N\epsilon}^{j < \hat{k}_i + N\epsilon} \delta(k_i - j) \tag{3.12}$$

For a perfectly calibrated posterior, the actual and expected counts would be identical: the model would be correct (within some given maximum error) as often as it expects to be correct. For an overconfident posterior, the expected count will be higher, and for an underconfident posterior, the actual count will be higher.

The predicted posteriors of volunteer votes are shown to be fairly well-calibrated; the model is correct approximately as often as it *expects* to be correct. Figure 3.10 compares the expected and actual counts for the model, choosing ϵ between 0 and 0.5. Tables 3.1 and 3.2 show calibration results for the 'Smooth' and 'Bar' models, with and without MC Dropout, evaluated on their respective test sets. Coverage error is calculated as:

$$\text{Coverage error} = \frac{C_{\text{expected}} - C_{\text{actual}}}{C_{\text{actual}}}. \tag{3.13}$$

For both questions, the single network (without using MC Dropout) is substantially overconfident. The MC-dropout-marginalised network shows a significant improvement in calibration over the single network. I interpret this as evidence for

the importance of marginalising over both networks and augmentations in accurately estimating uncertainty (Section 3.1.5).

When making precise predictions, the MC-dropout-marginalised network remains somewhat overconfident. However, as the acceptable error ϵ is allowed to increase, the network is increasingly well-calibrated. For example, the predicted probability that $v \pm 0.02$ (i.e. $\epsilon = 0.02$) k of N volunteers respond 'Bar' is over-estimated by $\sim$ 45%. In contrast, the predicted probability that $k \pm 0.2$ (i.e. $\epsilon = 0.2$) of N volunteers respond 'Bar' is $\sim 10\%$ of the true probability. I discuss possible approaches to further improve calibration in Section 3.3, and go on to apply them in Chapter 4.

A key method for galaxy evolution research is to compare the distribution of some morphology parameter across different samples (e.g. are spirals more common in dense environments, [449], do bars fuel AGN, [144], do mergers inhibit LERGs, [152], etc.) Therefore, the distribution of predicted $\hat{\rho}$ and $\hat{k}$ over all galaxies should ideally approximate the observed distribution of ρ[13] and k. In short, the predictions should be *globally unbiased*. Figure 3.11 compares the predicted and actual distributions of ρ and k. The predicted distributions for ρ and k are shown to match well with the observed distributions for most values of ρ and k. The model appears somewhat reticent to predict extreme ρ (and therefore extreme k) for both questions. This may be a consequence of the difficulty in predicting the behaviour of single volunteers. Again, I discuss this further in Section 3.3 and mitigate this in chapter 4.

Reliable research conclusions also require that model performance should not depend strongly on non-morphological galaxy parameters (mass, colour, etc). For example, if a researcher would like to investigate correlations between galaxy mass and bars, it is important that the model is equally able to recognise bars in high-mass and low-mass galaxies. To check if the model is sensitive to non-morphological parameters, I used an Explainable Boosting Machine (EBM) model [73, 274]. EBM aim to predict a target variable based on tabular features by separating the impact of those features into single (or, optionally, pairwise) effects on the target variable. They are a specific[14] implementation of Generalised Additive Models (GAM, [160]). GAM are of the form:

$$g(y) = f_1(x) + ... + f_n(x_n) \tag{3.14}$$

[13]The 'observed' ρ is approximated as $\rho_{\text{proxy}} = \frac{k}{N}$, which has a similar distribution to the true (latent, unobserved) ρ over a large sample.

[14]https://github.com/microsoft/interpret

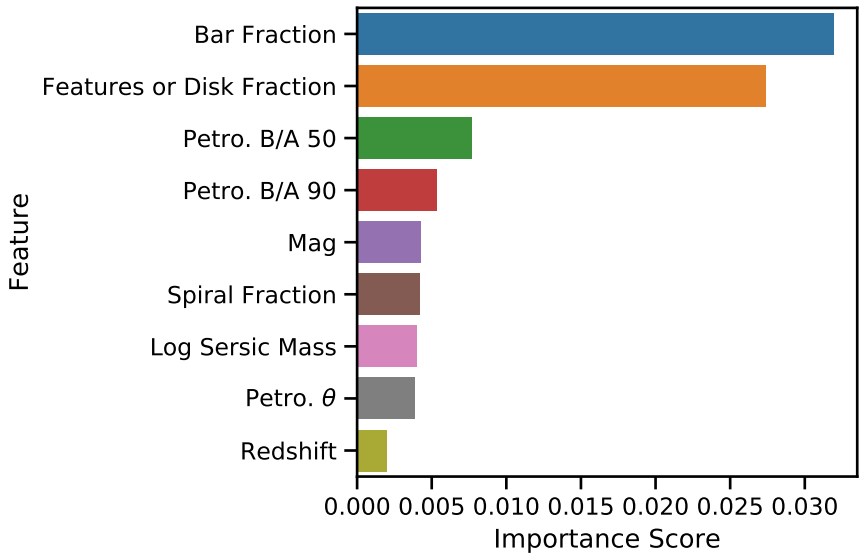


Figure 3.9: Relative importance of morphological (Features or Disk, Bar, Spiral) and non-morphological (Petro B/A, Mag, etc.) features for BCNN performance. Morphology fractions are the (volunteer-reported) $\frac{k}{N}$ values from Galaxy Zoo 2. Petro B/A 50 and Petro B/A 90 measure the axial ratios at 50% and 90% of the half-light radius. Mag is the estimated B magnitude. Sersic mass is the approximate stellar mass, estimated from the single-component Sersic fit flux. Petro θ is the (r-band) Petrosian radius. Redshift is measured spectroscopically. The effect of each component is additive and independent; for example, the measured effect of spiral features does not include the effect of being featured in general. BCNN performance varies much less from the effect of non-morphological features than from morphological features.

where g is identity for regression problems and f_i is any learnable function. For EBM, each f_i is learned using gradient boosting with bagging of shallow regression trees. They aim to answer the question 'What is the effect on the target variable of *this particular feature alone*?' I trained an EBM to predict the surprise[15] of our 'Bar' model when making test set predictions (Sec. 3.1.8), using the volunteer-reported morphologies and key non-morphological parameters reported in the NASA Sloan Atlas (v1.01, [15]).

The interested reader can find my full investigation at https://doi.org/10.5281/zenodo.4545335, recorded as a Jupyter Notebook. Figure 3.9 shows the key result; the relative importance of each feature on BCNN model surprise. Performance variation with respect to non-morphological parameters is shown to be much smaller than variation with re-

[15]Recall that we quantify surprise as the likelihood of our prediction given the observed votes $\frac{k}{N}$ (Eqn 3.3).

spect to morphology. The network performs better on smooth galaxies and unbarred galaxies (plausibly because there are more training examples of such galaxies to learn from). Inclination is the non-morphological parameter with the strongest effect on performance, and this effect is approximately 3.5-4 times weaker than the effect of either smoothness or barredness above. This suggests that the model introduces no new major biases with respect to key non-morphological parameters.

3.1.9.1 Comparison to Previous Work

The key goals of this chapter are to introduce probabilistic predictions for votes and (in the following section) to apply this to perform active learning. However, by reducing the model's probabilistic predictions to point estimates, I also provide conventional predictions and performance metrics.

Previous work has focused on deterministic predictions of either the vote fractions [100] or (more commonly) the majority response [32, 76, 227, 370]. While differences in sample selection and training data prevent a precise comparison, the model performs well at both tasks.

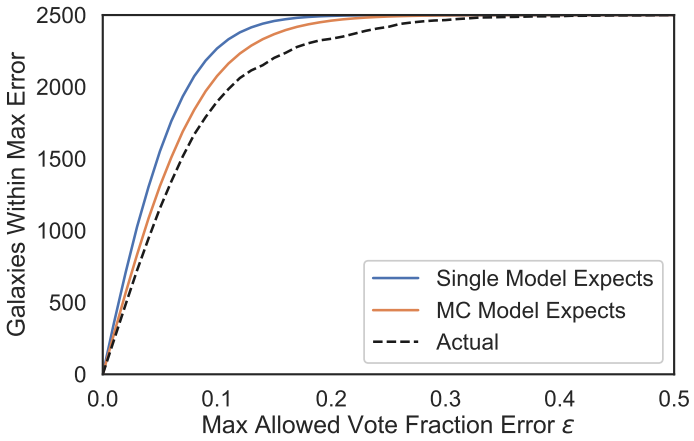
I reduced the posteriors to the most likely vote count $\hat{k}$, achieving a root-mean-square error of 0.10 (approx. ± 3 votes) for 'Smooth or Featured' and 0.15 for 'Bar'. I also reduced the same posteriors to the most likely majority responses. Here, I present the results in the style of the ROC curves in [370] (hereafter DS+18, Figure 3.12) and the confusion matrices in [227] (hereafter K+18, Figure 3.13) using the reduced posteriors. The results show that the model presented here likely outperforms [370] and is likely comparable with [227].

Overall, these conventional metrics demonstrate that the models presented here are sufficiently accurate for practical use in galaxy evolution research even when reduced to point estimates.

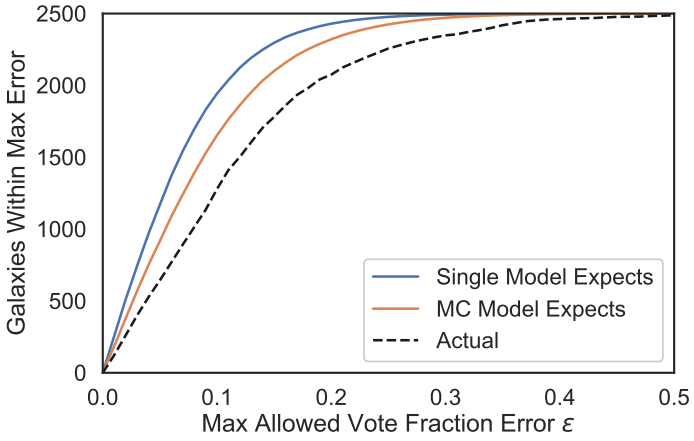
3.2 Active Learning

In the previous sections of this chapter, I presented Bayesian CNNs that predict posteriors for the morphology of each galaxy. Limited volunteer classification speed remains a hurdle; one still needs to collect enough responses to train these Bayesian networks. How do we train Bayesian networks to perform well while minimising the number of new responses required?

Previous approaches in morphology classification have largely used fixed datasets of labelled galaxies acquired prior to model training. This is true both for authors

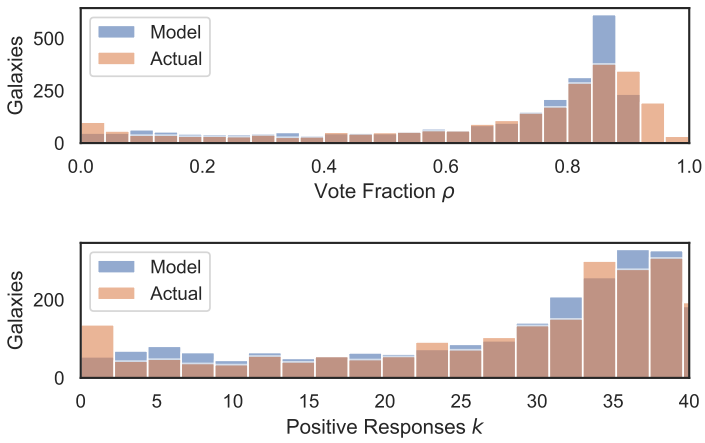


(a) Calibration for 'Smooth or Featured'

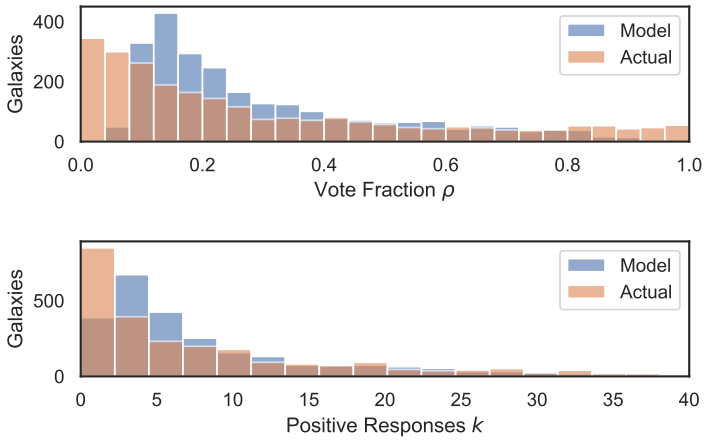


(b) Calibration for 'Bar'

Figure 3.10: Calibration of CNN-predicted posteriors, showing the expected versus actual count of galaxies within each acceptable maximum vote fraction error range (ϵ). The probabilistic model is fairly well-calibrated (similar expected and actual counts), with a significant improvement from applying MC Dropout.

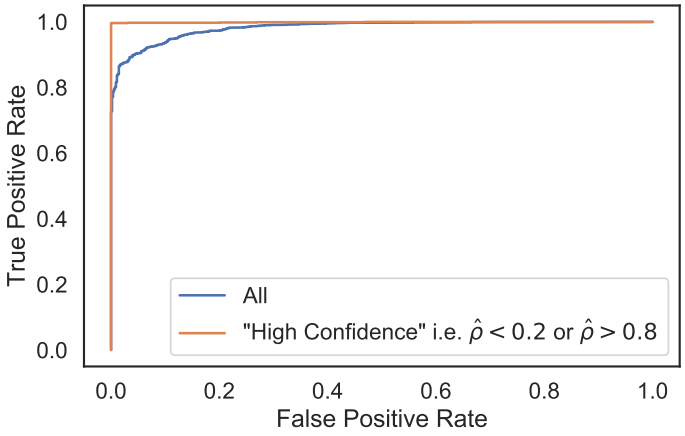


(a) Distribution of k and ρ for 'Smooth or Featured'

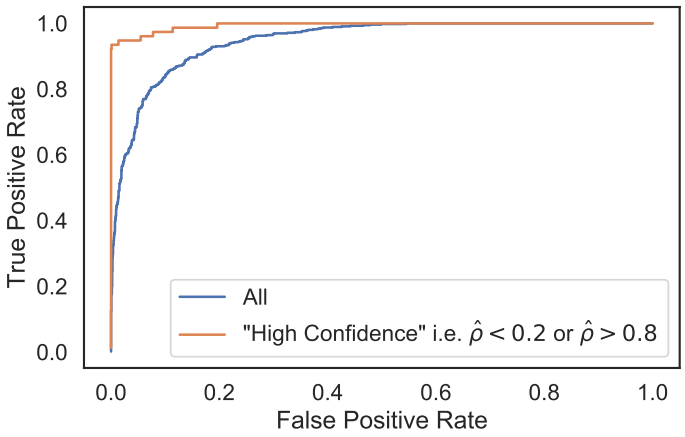


(b) Distribution of k and ρ for 'Bar'

Figure 3.11: Comparison between the distribution of predicted or observed ρ and k over all galaxies, for each question. Upper: comparison for 'Smooth or Featured'. Lower: comparison for 'Bar'. The observed ρ is approximated as $\rho_{\text{proxy}} = \frac{k}{N}$ The distributions of predicted ρ and k closely match the observed distributions, indicating the models are globally unbiased. The only significant deviation is near extreme ρ and k, which the models are 'reluctant' to predict.



(a) ROC curve for the 'Smooth or Featured' question.



(b) ROC curve for the 'Bar' question.

Figure 3.12: ROC curves for the 'Smooth or Featured' (above) and 'Bar' (below) questions, as predicted by the probabilistic model. To generate scalar class predictions on which to threshold, I reduced the posteriors to mean vote fractions. For comparison to DS+18, I also include ROC curves of the subsample they describe as 'high confidence' – galaxies where the class probability (for us, $\hat{\rho}$) is extreme (1420 galaxies for 'Smooth', 1174 for 'Bar')

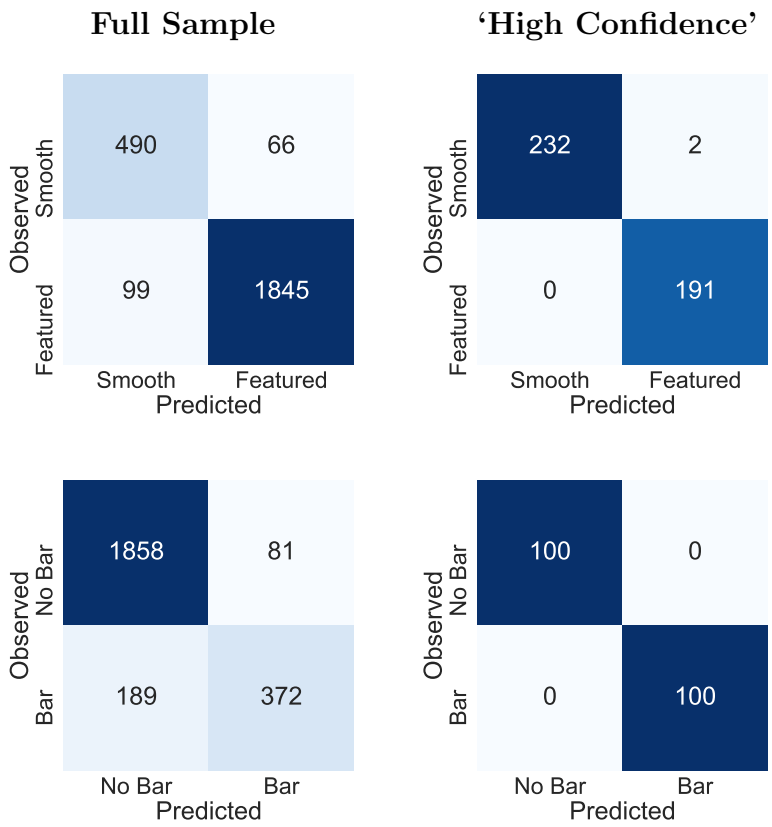


Figure 3.13: Confusion matrices for 'Smooth or Featured' (upper row) and 'Bar' (lower row) questions. For comparison to K+18, I also include confusion matrices for the most confident predictions (right column) Following K+18, I include the most confident $\sim$ 7.7% of spirals and $\sim$ 9.3% of ellipticals (upper right). Of the two galaxies where humans select 'Smooth' ($\frac{k}{N} > 0.5$) and the model selects 'Featured' ($\hat{\rho} < 0.5$), one is an ongoing smooth/featured major merger and one is smooth with an imaging artifact. Generalising (K+18 do not consider bars), I also show the most confident $\sim$ 8% of barred and unbarred galaxies. The probabilistic model achieve perfect classification for 'Bar'.

applying direct training [127, 191, 192, 370, 443] and those applying transfer learning [6, 104, 339]. Instead, I thought we should ask: to train the best model, which galaxies should volunteers label? Each galaxy, if labelled, provides information to the model; they are *informative*. My hypothesis was that all galaxies are informative, but some galaxies are more informative than others. To exploit this, I used the Bayesian CNN galaxy morphology posteriors to apply an active learning strategy [182]: *intelligently selecting the most informative galaxies for labelling by volunteers*. By prioritizing the galaxies that this strategy suggests would, if labelled, be most informative to the model, one can create or fine-tune models with even less newly-labelled data.

In the rest of this chapter, I describe simulating using the posteriors to select the most informative galaxies for labelling by volunteers.

Selecting the most informative data to label is known as active learning. Active learning is useful when acquiring labels is difficult (expensive, time-consuming, requiring experts, private, etc). This scenario is common for many, if not most, real-world problems. Terrestrial examples include detecting cardiac arrhythmia [352], sentiment analysis of online reviews [473], and Earth observation [267, 432]. Astrophysical examples include stellar spectral analysis [396], variable star classification [357], telescope design and time allocation [462], redshift estimation [185] and spectroscopic follow-up of supernovae [198].

3.2.1 Active Learning Approach for Galaxy Zoo

Given that only a small subset of galaxies can be labelled by humans, we should intelligently select which galaxies to label. The aim is to make CNNs that are just as accurate without having to label as many galaxies. To do so, I took the following approach. First, I trained a Bayesian CNN (introduced above) on a small randomly chosen initial training set. Then, I repeated the following active learning loop:

1. Measure the CNN prediction uncertainty on all currently-unlabelled galaxies (excluding a fixed test set)

2. Apply an acquisition function (Section 3.2.2) to select the most uncertain galaxies for labelling

3. Upload these galaxies to Galaxy Zoo and collect volunteer classifications (in this work, simulated with historical classifications)

4. Re-train the CNN and repeat

Other astrophysics research has combined crowdsourcing with machine learning models. Wright et al. 2017 [460] classified supernovae in PanSTARRS [208] by aggregating crowdsourced classifications with the predictions of expert-trained CNN and show that the combined human/machine ensemble outperforms either alone. However, this approach is not directly feasible for Galaxy Zoo, where scale prevents us from recording substantial numbers of crowdsourced classifications for every image. Beck et al. 2018 [34] developed a 'decision engine' to allocate galaxies for classification by either human or machine (via a random forest), unlike the system presented here which only requests responses for informative galaxies. Beck et al. 2018 also train their model exclusively on galaxies which can be confidently assigned to a class, while the model presented here learns from every classified galaxy.

The work in this chapter is the first time active learning has been used for morphological classification, and the first time in astrophysics that active learning has been combined with CNNs or crowdsourcing.

3.2.2 BALD and Mutual Information

Active learning requires an acquisition strategy - a method to estimate which unlabelled examples will be most helpful to label (acquire). I base mine on the general information-theoretic acquisition strategy Bayesian Active Learning by Disagreement (BALD) [182, 280]. BALD selects subjects to label by maximising the mutual information between the model parameters θ and the probabilistic label prediction y. BALD is derived from the mutual information as follows.

Assume you have observed data $\mathcal{D} = (x_i, y_i)_{i=1}^n$. Here, x_i is the ith subject and y_i is the label of interest. Further assume there are (unknown) parameters θ that model the relationship between input subjects x and output labels y, $p(y|x, \theta)$. One would like to infer the posterior of θ, $p(\theta|\mathcal{D})$. Once $p(y|x, \theta)$ is known, one can make predictions on new galaxy images.

The mutual information measures how much information some random variable A carries about another random variable B, defined as:

$$\mathbb{I}[A, B] = H[p(A)] - E_{p(B)}H[p(A|B)] \tag{3.15}$$

where H is the entropy operator and $E_{p(B)}H[p(A|B)]$ is the expected entropy of $p(A|B)$, marginalised over $p(B)$ [310].

One would like to know how much information each label y provides about the model parameters θ. One could then pick subjects x to maximise the mutual information $\mathbb{I}[y, \theta]$, helping to learn θ efficiently. Substituting A and B for x and y:

$$\mathbb{I}[y, \theta] = H[p(y|x, \mathcal{D})] - \mathbb{E}_{p(\theta|\mathcal{D})}[H[p(y|x, \theta)]] \tag{3.16}$$

The first term is the entropy of our prediction for x given the training data, implicitly marginalising over the possible model parameters θ. I refer to this as the predictive entropy. The predictive entropy reflects our overall uncertainty in y given the training data available.

The second term is the expected entropy of our prediction made with a given θ, sampling over each θ that might have been inferred from $\mathcal{D}$. The expected entropy reflects the typical uncertainty of each particular model on x. Expected entropy has a lower bound set by the inherent difficulty in predicting y from x, regardless of the available labelled data.

Confident disagreement between possible models leads to high mutual information. For high mutual information, we should be highly uncertain about y after marginalising over all the models we might infer (high $H[p(y|x, \mathcal{D})]$), but have each particular model be confident (low expected $H[p(y|x, \theta)]$). If we are uncertain overall, but each particular model is certain, then the models must confidently disagree.

Throughout this chapter, when I refer to galaxies as informative, I mean specifically that they have high mutual information; they are *informative for the model*. These are not necessarily the galaxies that are the most *informative for science*; any overlap will depend upon the research question at hand. The scientific benefit of the approach presented here is that the Bayesian CNN learns to make accurate morphological predictions for all galaxies using minimal newly-labelled examples.

3.2.3 Estimating Mutual Information

Rewriting the mutual information explicitly, replacing y with our labels k and θ with the network weights w:

$$\mathbb{I}[k, w] = \mathbb{H}[\int p(k|x, w)p(w|\mathcal{D})dw] - \int p(w|\mathcal{D})\mathbb{H}[p(k|x, w)]dw \tag{3.17}$$

As I noted in Sec. 3.1.5, Gal et al. 2017 [143] showed that we can use Eqn. 3.8 to replace $p(w|\mathcal{D})$ in the mutual information (Eqn. 3.17):

$$\mathbb{I}[k, w] = \mathbb{H}[\int p(k|x, w)q^*dw] - \int q^*\mathbb{H}[p(k|x, w)]dw \tag{3.18}$$

and again sample from q^* with T forward passes using dropout at test time (i.e. Monte Carlo integration), allowing the mutual information to be calculated:

$$\mathbb{I}[k, w] = \mathbb{H}[\frac{1}{T}\sum_t p(k|x, w)] - \frac{1}{T}\sum_t \mathbb{H}[p(k|x, w)] \tag{3.19}$$

I derive an acquisition function for Galaxy Zoo by combining the approach of Gal et al. 2017 above with the probabilistic predictions $p(k|x, w)$ I introduced in Sec. 3.1.4. Recall that I trained Bayesian networks to make probabilistic predictions for k by estimating the latent parameter ρ from which k is Binomially drawn (Eqn. 3.3). Substituting the probabilistic predictions of Eqn. 3.3 into the mutual information:

$$\mathbb{I}[k, w] = \mathbb{H}[\frac{1}{T}\sum_t \text{Bin}(k|f^w(x), N)] - \frac{1}{T}\sum_t \mathbb{H}[\text{Bin}(k|f^w(x), N)] \tag{3.20}$$

Or concisely:

$$\mathbb{I}[k, w] = \mathbb{H}[\langle \text{Bin}(k|f^w(x), N)\rangle] - \langle \mathbb{H}[\text{Bin}(k|f^w(x), N)]\rangle \tag{3.21}$$

A novel complication is that we do not know N, the total number of responses, prior to labelling. In GZ2, each subject is shown to a fixed number of volunteers, but (due to the decision tree) N for each question will depend on responses to the previous question. Further, technical limitations mean that even for the first question ('Smooth or Featured'), N can vary (Figure 3.5). I (implicitly, for clarity) approximate N with the expected $\langle N\rangle$ for that question. In effect, I am calculating the acquisition function with N set to the value that, *were we to ask volunteers to label this galaxy, we would expect N responses.*

To summarise, Eqn. 3.21 asks: how much additional information would be gained about network parameters that we use to predict ρ and k, were we to ask $\langle N\rangle$ people about subject x?

3.2.4 Entropy Evaluation

Having approximated $p(w|\mathcal{D})$ with dropout and calculated $p(k|x, w)$ with the probabilistic model, all that remains is to calculate the entropies $\mathbb{H}$ of each term.

k is discrete and hence we can directly calculate the entropy over each possible state:

$$\mathbb{H}[\text{Bin}(k|f^w(x), N)] = -\sum_{k=0}^{N} \text{Bin}(k|f^w(x), N) \log[\text{Bin}(k|f^w(x), N)] \tag{3.22}$$

For $\mathbb{H}[\langle \text{Bin}(k|f^w(x), N)\rangle]$, we can also enumerate over each possible k, where the probability of each k is the mean of the posterior predictions (sampled with dropout) for that k:

$$\begin{aligned}\mathbb{H}[\langle \text{Bin}(k|f^w(x), N)\rangle] = \\ -\sum_{k=0}^{N} \langle \text{Bin}(k|f^w(x), N)\rangle \log[\langle \text{Bin}(k|f^w(x), N)\rangle]\end{aligned} \tag{3.23}$$

and hence the final expression for the mutual information is:

$$\begin{aligned}\mathbb{I}[k, w] = & \\ & - \sum_{k=0}^{N} \langle \text{Bin}(k|f^w(x), N) \rangle \log[\langle \text{Bin}(k|f^w(x), N) \rangle] \\ & + \sum_{k=0}^{N} \text{Bin}(k|f^w(x), N) \log[\text{Bin}(k|f^w(x), N)]\end{aligned} \tag{3.24}$$

3.2.5 Application

To evaluate the active learning approach described above, I simulated applying active learning during GZ2. I compared the performance of the models when trained on galaxies selected using the mutual information versus galaxies selected randomly. For simplicity, each simulation trained a model to predict either 'Smooth or Featured' responses or 'Bar' responses.

For the 'Smooth or Featured' simulation, I began with a small initial training set of 256 random galaxies. I trained a model to predict $p(k|\rho, N)$ (where N is the expected number of volunteers to answer the question, calculated as the mean total number of responses for that question over all previous galaxies - see Figure 3.5). I then used the BALD-based acquisition function I derived above (Eqn. 3.21) to identify the 128 most informative galaxies to label. To simulate uploading the informative galaxies to GZ and receiving classifications, I retrieved previously collected GZ2 classifications for the most informative galaxies. Finally, I added the newly-labelled informative galaxies to the training set. I refer to each execution of this process (training the model, selecting new galaxies to label, and adding them to the training set) as an *iteration*. I repeated this procedure for 20 iterations, recording the performance of the model throughout.

I selected 256 initial galaxies and 128 further galaxies per iteration, to match the training data size over which the 'Smooth or Featured' model performance varies. The relatively shallow model reaches peak performance on around 3000 random galaxies; more galaxies do not significantly improve performance.

For the 'Bar' simulation, I found that performance saturates after more galaxies (approx. 6000) and so I doubled the scale; I started with 512 galaxies and acquired 256 further galaxies per iteration. This matches previous results (and intuition) that 'Smooth or Featured' is an easier question to answer than 'Bar'. Identifying bars, particularly weak bars, is challenging for both humans [241, 294] and machines (including CNNs, Sánchez et al. (2018)).

To measure the effect of the active learning strategy, I also trained a baseline classifier by providing batches of randomly selected galaxies. This allowed me to compare the acquisition strategy I developed (selecting galaxies with maximal mutual information via BALD and MC Dropout) against selecting randomly (baseline). I evaluated performance on a fixed test set of 2500 random galaxies. I repeated each simulation four times to reduce the risk of spurious results from random variations in performance.

3.2.6 Results

For both 'Smooth' and 'Bar' simulations, the probabilistic models achieved equal performance on fewer galaxies using active learning versus random galaxy selection. Below, I show model performance by iteration for the 'Smooth' (Figure 3.14) and 'Bar' (Figure 3.15) simulations. These figures display three metrics: training loss (model surprise on previously-seen images, measured by Eqn. 3.6), evaluation loss (model surprise on unseen images), and root-mean-square error (RMSE). I measured the RMSE between the maximum-likelihood-estimates $\hat{\rho}$ and $\rho_{\text{proxy}} = \frac{k}{N}$ as ρ itself is never observed and hence cannot be used for evaluation. Due to the high variance in metrics between batches, I smoothed our metrics via LOWESS [80] and averaged across 4 simulation runs.

For 'Smooth', the models achieved equal RMSE scores with, at best, $\sim 60\%$ fewer newly-labelled galaxies (RMSE of 0.117 with 256 versus 640 new galaxies, Figure 3.14). Similarly for 'Bar', the models achieved equal RMSE scores with, at best, $\sim$ 35% fewer newly-labelled galaxies (RMSE of 0.17 with 1280 versus 2048 new galaxies, Figure 3.15). Active learning outperformed random selection in every run.

Given sufficient ($\sim$ 3000 for 'Smooth', $\sim$ 6000 for 'Bar') galaxies, the models eventually converged to similar performance levels – regardless of galaxy selection. I speculate that this is because our relatively shallow model architecture places an upper limit on performance. In general, model complexity should be large enough to exploit the information in the training set yet small enough to avoid fitting to spurious patterns (Section 2.2.3). Model complexity increases with the number of free parameters and decreases with regularization [140]. The model is both shallow and well-regularized (recall that dropout was originally used as a regularization technique, Section 3.1.5). A more complex (deeper) model may be able to perform better by learning from additional galaxies.

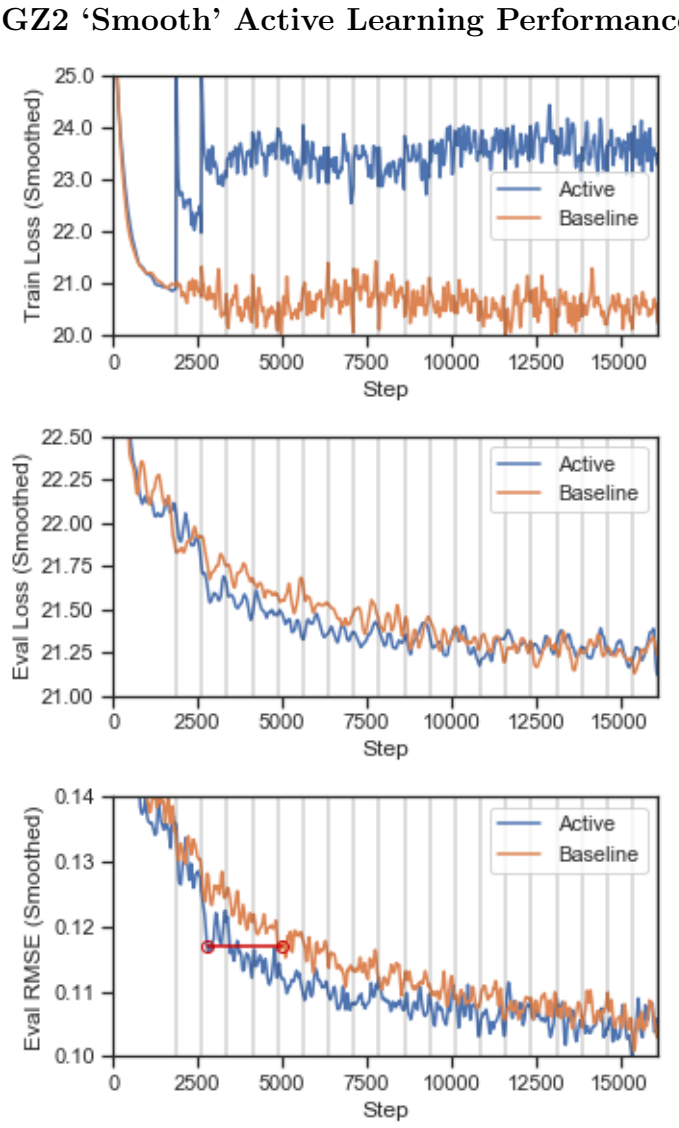


Figure 3.14: Training loss (upper), evaluation loss (middle), and RMSE (lower) of model performance on 'Smooth or Featured' during active learning simulations, by iteration (set of new galaxies). Vertical bars denote new iterations, where new galaxies are acquired and added to the training set. Prior to 2000 training iterations, both the random selection (baseline) models and active learning models trained on only the initial random training set of 256 galaxies, and hence showed similar performance. Around 2000 to 3500 iterations, after acquiring 128-256 additional galaxies, the active learning model showed a clear improvement in evaluation performance over the baseline model. I annotate in red where each model achieves the maximal relative RMSE improvement, highlighting the reduction in newly-labelled galaxies required (vertical bars = 128 new galaxies). Note that active learning leads to a dramatically higher training loss, indicating that more challenging galaxies are being identified as informative and added to the training set.

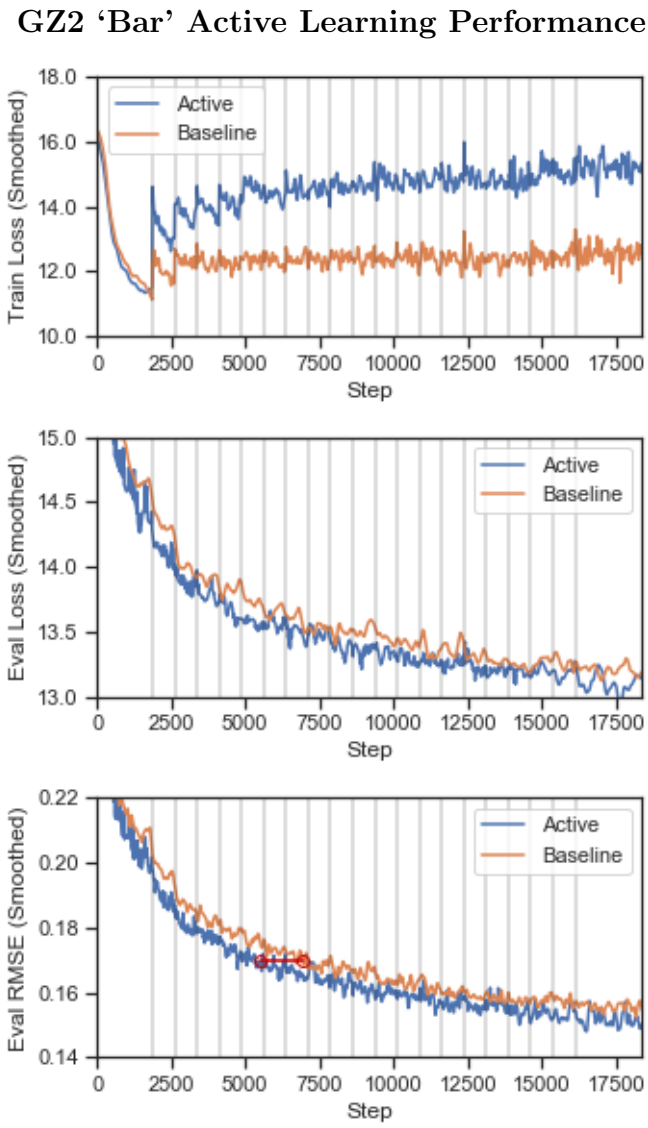


Figure 3.15: As with 3.14, but for the 'Bar' active learning simulations. Again, active learning led to a clear improvement in evaluation performance and a dramatically higher training loss (indicating challenging galaxies are being selected). I annotate in red where each model achieves the maximal relative RMSE improvement, highlighting the reduction in newly-labelled galaxies required (vertical bars = 256 new galaxies).

3.2.6.1 Selected Galaxies

Which galaxies did the models identify as informative? To investigate, I randomly selected one 'Smooth or Featured' and one 'Bar' simulation.

For the 'Smooth or Featured' simulation, Figure 3.16 shows the observed 'Smooth' vote fraction distribution, per iteration (set of new galaxies) and in total (summed over all new galaxies). Highly smooth galaxies are common in the general GZ2 catalogue. Random selection therefore led to a training sample skewed towards highly smooth galaxies. In contrast, the BALD-derived acquisition function was far more likely to select galaxies that are featured, leading to a more balanced sample. This was especially true for the first few iterations; I speculate that this counteracts the skew towards smooth in the randomly selected initial training sample. By the final training sample, featured galaxies became moderately more common than smooth (mean $\frac{k_{\text{smooth}}}{N} = 0.38$). This suggests that featured galaxies are (on average) more informative for the model – over and above correcting for the skewed initial training sample. I believe that featured galaxies may be more visually diverse, which led to a greater challenge in fitting volunteer responses, more disagreement between dropout-approximated-models, and ultimately higher mutual information.

For the 'Bar' simulation, Figure 3.17 shows the 'Bar' vote fraction distribution, per iteration and in total, as well as the total redshift distribution. Again, the BALD-derived acquisition function selected a more balanced sample by prioritising (rarer) barred galaxies. This selection remained approximately constant (within statistical noise) as more galaxies are acquired. With respect to redshift, the acquisition function preferentially selected galaxies at lower redshifts. Based on inspection of the selected images (Figure 3.19), I suspect that these galaxies are more informative to the model because such galaxies are better resolved (i.e. less ambiguous) and more likely to be barred.

I show the most and least informative galaxies from the (fixed and never labelled) test subset for 'Smooth' (Figure 3.18 and Bar (Figure 3.19), as identified by the novel acquisition function and the final models from each simulation.

3.3 Discussion

Learning from fewer examples is an expected benefit of both probabilistic predictions and active learning. The models approach peak performance on remarkably few examples: 2816 galaxies for 'Smooth' and 5632 for 'Bar'. With this system, volunteers

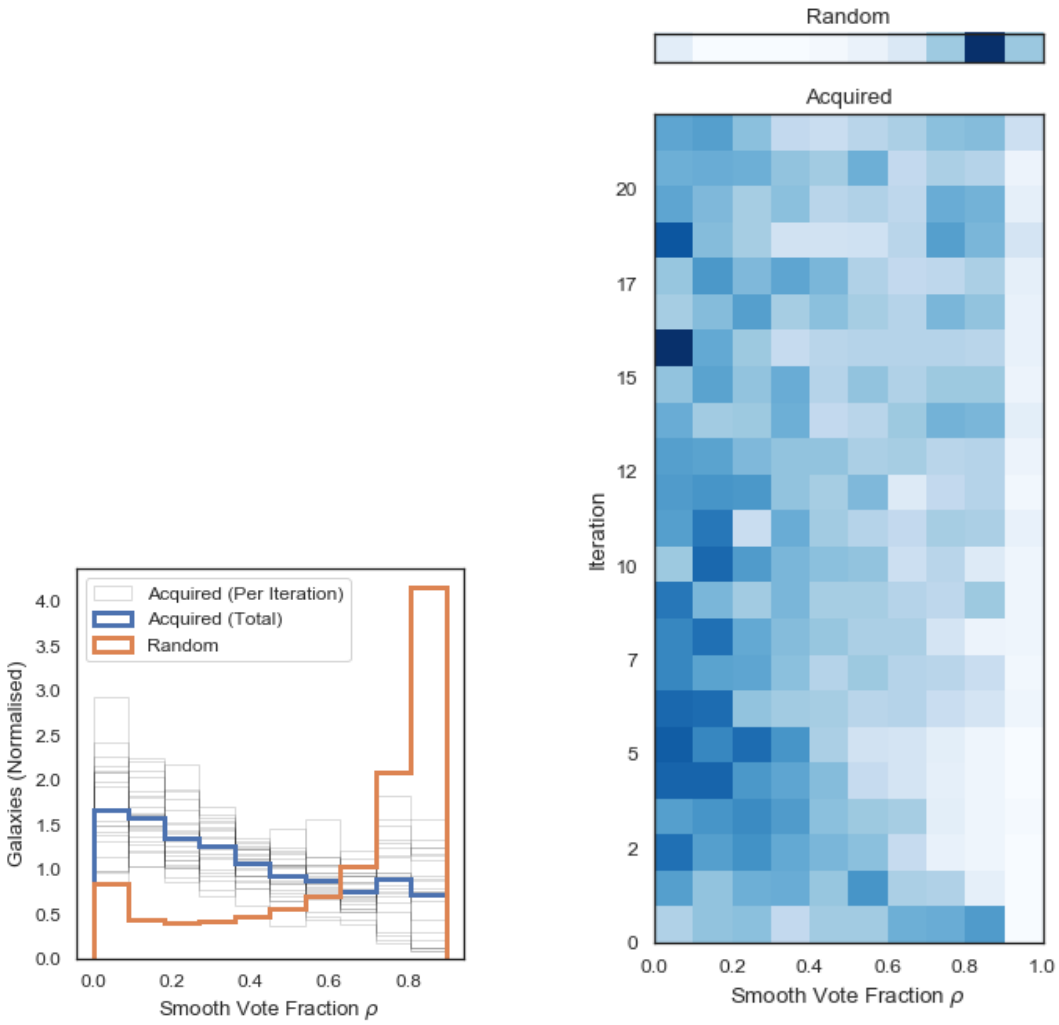


Figure 3.16: Distribution of observed 'Smooth' vote fraction p in galaxies acquired during Galaxy Zoo 'Smooth or Featured' active learning simulation. Left: Distribution of acquired p over all iterations, compared against random selection. While randomly selected galaxies are highly smooth, our acquisition function selects galaxies from across the p range, with a moderate preference towards featured. Right: Distribution of p by iteration, compared against random selection (upper inset). The acquisition function strongly prefers featured galaxies in early ($n <\sim 7$) iterations, and then selects a more balanced sample. This likely compensates for the initial training sample being highly smooth.

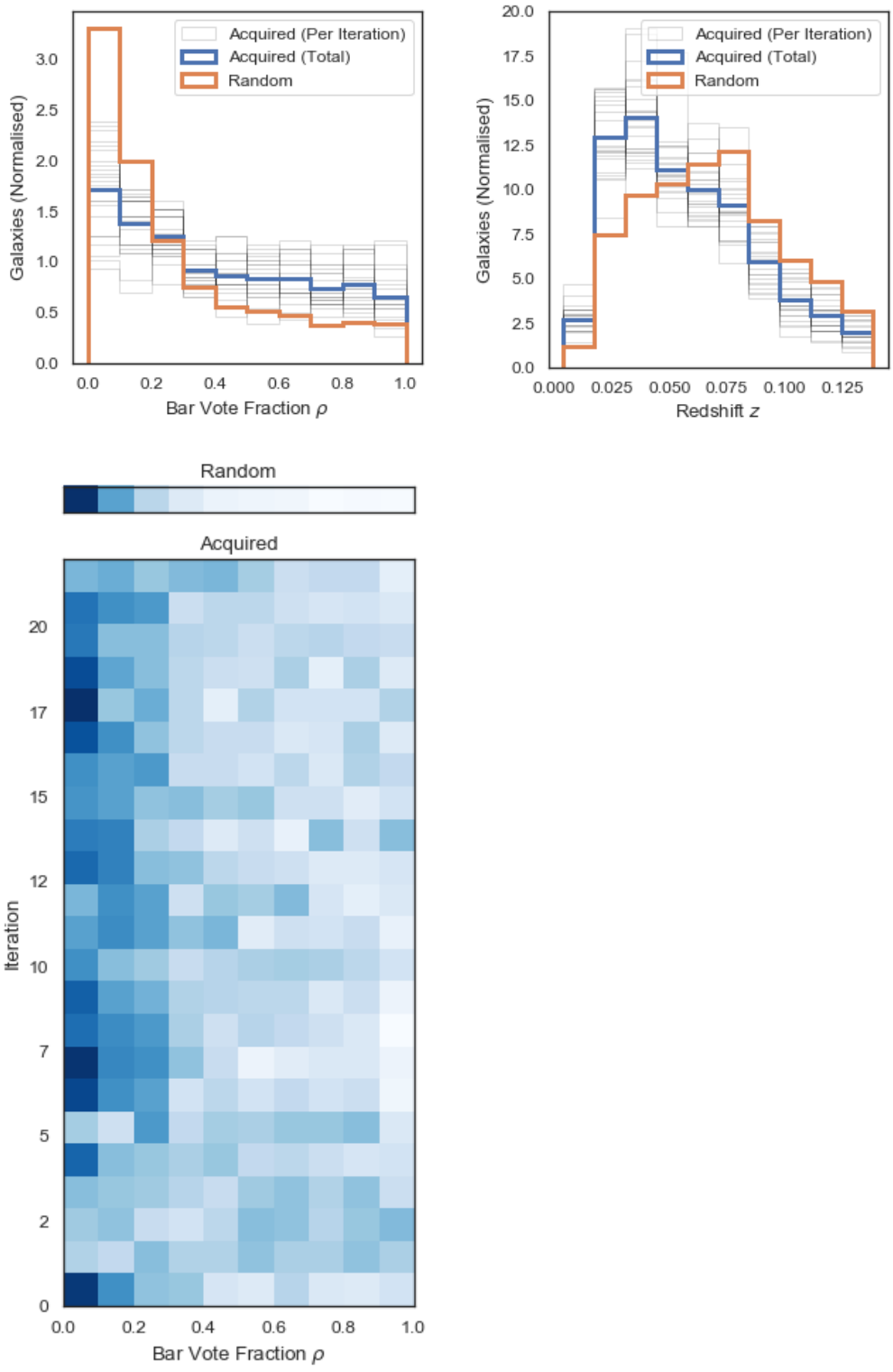


Figure 3.17: Upper left: Distribution of observed 'Bar' vote fraction p in galaxies acquired during Galaxy Zoo 'Bar' active learning simulation. Upper right: Redshift distribution of acquired galaxies over all iterations, compared against random selection. The 'Bar' model selects lower redshift galaxies, which are both more featured and better resolved (i.e. less visually ambiguous). While randomly selected galaxies are highly non-barred, the 'Bar' model selects a more balanced sample. Lower: Distribution of 'Bar' p by iteration, compared against random selection (upper inset). The acquisition function selects a similar *rho* distribution at each iteration.

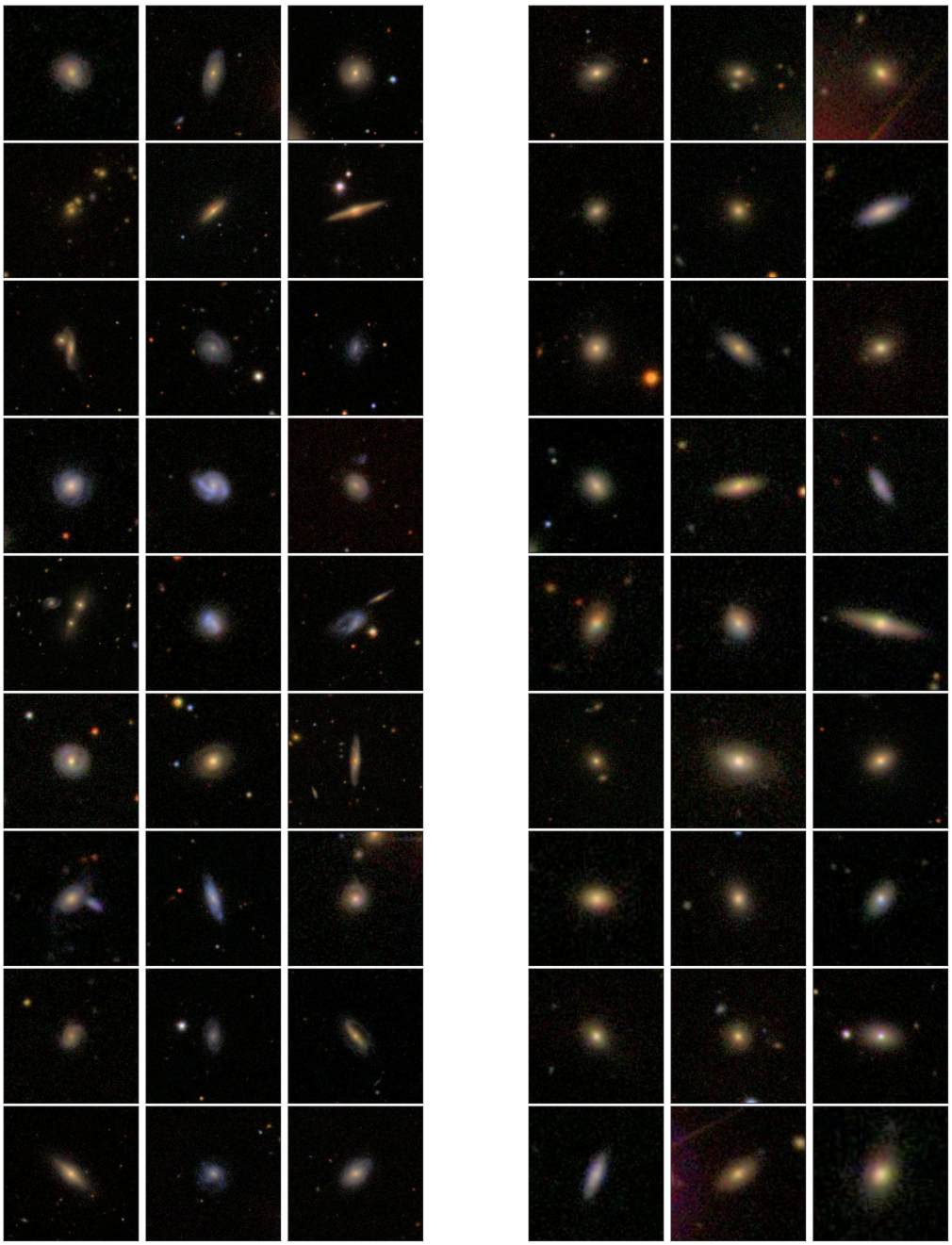


Figure 3.18: Informative and uninformative galaxies from the (hidden) test subset, as identified by the novel acquisition function and the final model from a 'Smooth or Featured' simulation. If active learning were applied to Galaxy Zoo, volunteers would be more frequently presented with the most informative images (left) than the least (right).

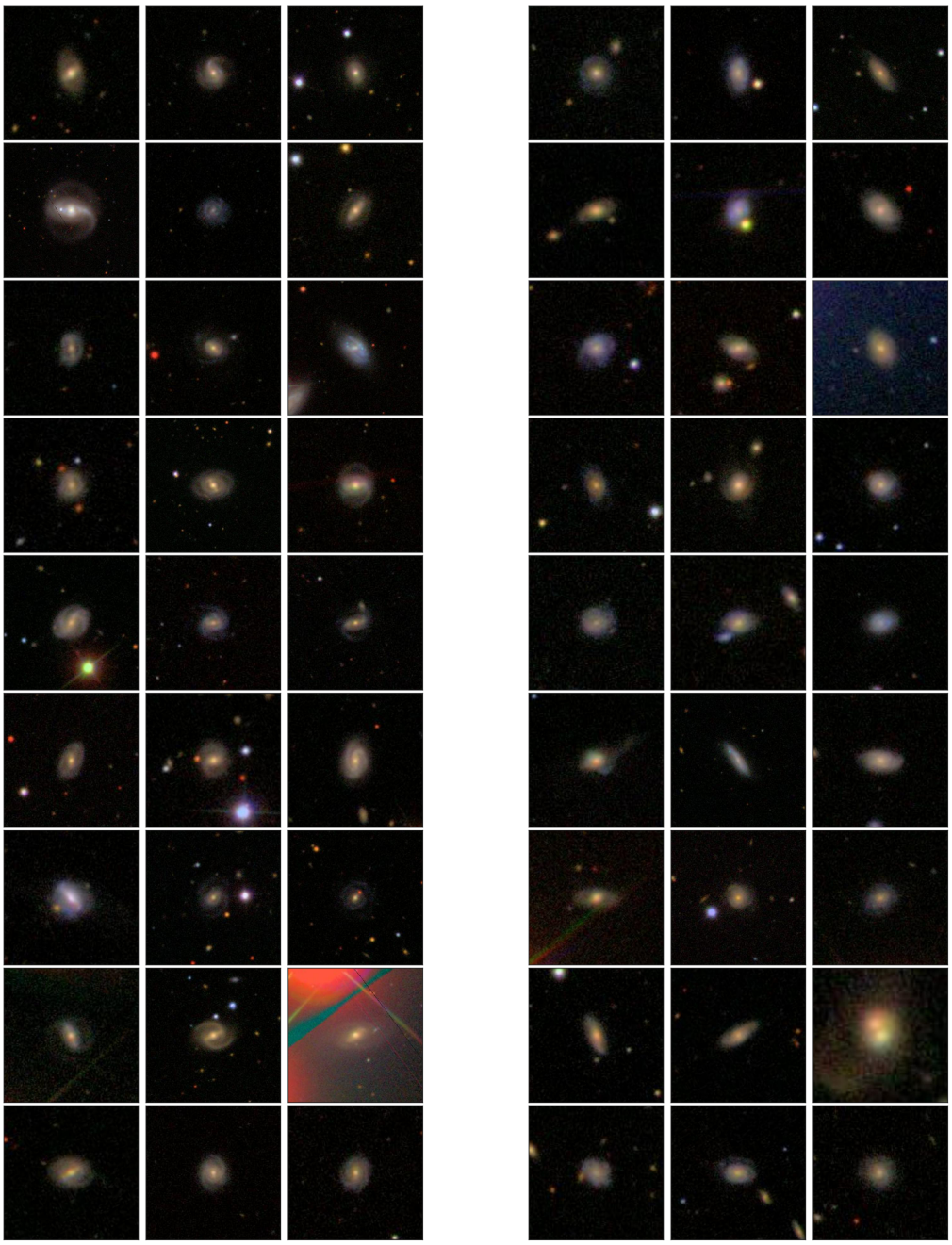


Figure 3.19: As with Figure 3.18 above, but showing galaxies identified by the final model from a 'Bar' simulation. Galaxies found to be informative for learning to classify bars (left) have well-resolved centres, unlike those found to be uninformative (right).

could complete Galaxy Zoo 2 in weeks[16] rather than years if the peak performance of the models would be sufficient for their research. Further, reaching peak performance on relatively few examples indicates that an expanded model with additional free parameters is likely to perform better [310]. I go on to use such an expanded model in the following chapter.

My results motivate various improvements to the probabilistic morphology models I introduced. In Section 3.1.9, I showed that the models were approximately well-calibrated, particularly after applying MC Dropout. However, the calibration was imperfect; even after applying MC Dropout, the models remained slightly overconfident (Figure 3.10). I suggest two reasons for this remaining overconfidence. First, within the MC Dropout approximation, the dropout rate is known to affect the calibration of the final model [143]. I chose the dropout rate arbitrarily (0.5); however, this rate may not sufficiently vary the model to approximate training many models. One solution might be to 'tune' the dropout rate until the calibration is correct [143]. However, the MC Dropout approximation is itself imperfect; removing random neurons with dropout is not identical to training many networks [137]. As an alternative, one could simply train several models and ensemble the predictions [243]. One could even have the best of both worlds; train several models, and use MC dropout for each model. This is the approach I chose to take in the following chapter.

I also showed the distribution of model predictions over all galaxies generally agrees well with the distribution of predictions from volunteers (i.e. the models are globally unbiased, Section 3.1.9). However, I noted that the models are 'reluctant' to predict extreme ρ (the typical response probability, Section 3.1.3). I suggest that this is a limitation of the statistical description of volunteer responses. The binomial likelihood becomes narrow when p (here, ρ) is extreme, and hence models are heavily penalised for incorrect extreme p estimates. If volunteer responses were precisely binomially distributed (i.e. N independent identically-distributed trials per galaxy, each with a fixed p of a positive response), this heavy penalty would correctly reflect the significance of the error. However, the binomial description of volunteers is only approximate; one volunteer may give consistently different responses to another. In consequence, the true likelihood of non-extreme k responses given ρ is wider than the binomial likelihood from the 'typical' response probability ρ suggests, and the network is penalised 'unfairly'. The network therefore learns to avoid making risky extreme predictions. If this suggestion is correct, the risk-averse prediction shift will

[16]For example, classifying $\sim$ 10,000 galaxies (sufficient to train the models to peak performance) at the mean GZ2 classification rate of $\sim$ 800 galaxies/day would take $\sim$ 13 days.

be monotonic (i.e. extreme galaxies will have slightly different ρ but still be ranked in the same order) and hence researchers selecting galaxies near extreme ρ may simply choose a slightly higher or lower $\hat{\rho}$ threshold. To mitigate this issue, one could apply a monotonic rescaling to the network predictions (as I did in Section 3.1.2) or calibrate the loss to reflect the scientific utility of extreme predictions [81]. However, I felt it would be best to address the fundamental issue directly by introducing a more flexible model of volunteer behaviour, and do so in the following chapter. This has the added benefit that the model is better able to express its own uncertainty, as I explain in section 4.5.

During Galaxy Zoo 2 (GZ2), N, the number of responses to a galaxy for a particular question, depends on the answers to previous questions and so cannot (in general) be known beforehand. For this work, I relied on GZ2 data when simulating a (historical) classification request. Therefore, when deriving the acquisition function, I approximated N as the expected number of responses $\langle N \rangle$. However, during live application of this system, one could control the Galaxy Zoo classification logic to collect exactly N responses per image, for any desired N. This would allow the model to request (for example) one more classification for *this* galaxy, and three more for *that* galaxy, before retraining. Precise classification requests from the model would enable us to ask volunteers exactly the right questions, helping them make an even greater contribution to scientific research.

I also hope that this human-machine collaboration provides a better experience for volunteers. Inspection of informative galaxies (Figures 3.16, 3.17) suggests that more informative galaxies are more diverse than less informative galaxies. I hope that volunteers find these (now more frequent) informative galaxies interesting and engaging. When applied in the following chapter, approximately 95% of volunteers choose to see the 'Enhanced' active-learning-prioritised galaxies over a random selection.

Finally, I would like to highlight that this approach is highly general. I hope that Bayesian CNNs and active learning can contribute to the wide range of astrophysical problems where CNNs are applicable (e.g. images, time series), uncertainty is important, and the data is expensive to label, noisy, imbalanced, or includes rare objects of interest. In particular, imbalanced datasets (where some labels are far more common than others) are common throughout astrophysics. Topics include transient classification [460], exoplanet detection [327], and fast radio burst searches [472] - the topic of chapter 5. Active learning is known to be effective at correcting such imbalances [198]. The results presented here suggest that this remains true when active learning is combined with CNNs. Recall that smooth galaxies are far more common

in GZ2 but featured galaxies are strongly preferentially selected by active learning – automatically, without our instruction – apparently to compensate for the imbalanced data (Figure 3.16). If this behaviour proves to be general, Bayesian CNNs and active learning may be an effective method for intelligent data collection to overcome research challenges throughout astrophysics.

In the following chapter, I go on to apply active learning alongside an improved Bayesian CNN approach to make predictions on newly-observed galaxies.

Chapter 4

Galaxy Zoo DECaLS

4.1 Introduction

I warned in 3.1.1 that surveys like the Dark Energy Camera Legacy Survey (DECaLS, [99]) would image more galaxies than could be comprehensively classified by volunteers alone. I used existing Galaxy Zoo 2 data to develop and demonstrate a solution: using Bayesian CNNs to predict galaxy morphology posteriors, and then using those posteriors to select the most informative galaxies for volunteers to label (active learning). In this chapter, I describe using an improved version of that approach to classify DECaLS galaxies. These classifications compose the Galaxy Zoo DECaLS data release.

Galaxy Zoo DECaLS is the first systematic engagement of volunteers with low-redshift images as deep as those provided by DECaLS, enabling detailed classification of fainter features. Classifications include the presence and strength of bars and bulges, the count and winding of spiral arms, and the indications of recent or ongoing mergers. Volunteer classifications were sourced over three separate Galaxy Zoo DECaLS (GZD) classification campaigns, GZD-1, GZD-2, and GZD-5, which classified galaxies first released in DECaLS Data Releases 1, 2, and 5 respectively[1]. Data collection for GZD-1 and GZD-2 ended before my DPhil; I was responsible for relaunching Galaxy Zoo to carry out the GZD-5 campaign. For GZD-5, an improved decision tree was added aimed at exploiting the deeper DECaLS images for better identification of mergers and weak bars. Galaxy Zoo volunteers contributed 1.8 million responses to this improved tree, with 139,919 galaxies receiving 30 or more classifications (prioritised partly by active learning) and the remaining 173,870 receiving approximately 5. I then trained an updated Bayesian CNN (sec. 4.5) to predict the improved tree

[1]Legacy Survey data releases are cumulative; data releases 3 and 4 are therefore also included.

answers for all galaxies, including both the GZD-1/2 galaxies classified with the older tree and the GZD-5 galaxies with fewer classifications.

4.2 Imaging

4.2.1 Observations

The galaxy images were created from data collected by the DECaLS survey [99]. DECaLS used the Dark Energy Camera (DECam, [129]) at the 4m Blanco telescope at Cerro Tololo Inter-American Observatory, near La Serena, Chile. DECam has a roughly hexagonal 3.2 square degree field of view with a pixel scale of 0.262 arcsec[2] per pixel. The median point spread function FWHM is $1\farcs29$, $1\farcs18$ and $1\farcs11$ for g, r, and z, respectively.

The DECaLS survey contributed targeting images for the upcoming Dark Energy Spectroscopic Instrument (DESI). DECaLS was responsible for the DESI footprint in the Southern Galactic Cap (SGC) and the $\delta \leq 34$ region of the Northern Galactic Cap (NGC), totalling 10,480 square degrees[2]. 1130 square degrees of the SGC DESI footprint were already being imaged by DECam through the Dark Energy Survey (DES, [420]) so DECaLS did not repeat this part of the DESI footprint. DECaLS implemented a 3-pass strategy to tile the sky. Each pass was slightly offset (approx 0.1-0.6 deg). The choice of pass and exposure time for each observation was optimised in real time based on the observing conditions recorded for the previous targets, as well as the interstellar dust reddening, sky position, and estimated observing conditions of possible next targets. This led to a near-uniform depth across the survey. In DECaLS DR1, DR2, and DR5, from which our images are drawn, the median 5σ point source depths for areas with 3 observations was approximately (AB) g=24.65, r=23.61, and z=22.84[3]. The DECaLS survey completed observations in March 2019.

4.2.2 Selection

Galaxies were identified in the DECaLS imaging using the NASA-Sloan Atlas v1.0.0 (NSA). As the NSA was derived from SDSS DR8 imaging [12], Galaxy Zoo DECaLS only includes galaxies that are within both the DECaLS and SDSS DR8 footprint. In effect, Galaxy ZOO DECaLS uses deeper DECaLS imaging of the galaxies previously imaged in SDSS DR8. This ensures the morphological measurements have a wealth

[2]The remaining DESI footprint is being imaged by DECaLS' companion surveys, MzLS and BASS [99]

[3]See https://www.legacysurvey.org/dr5/description/ and related pages

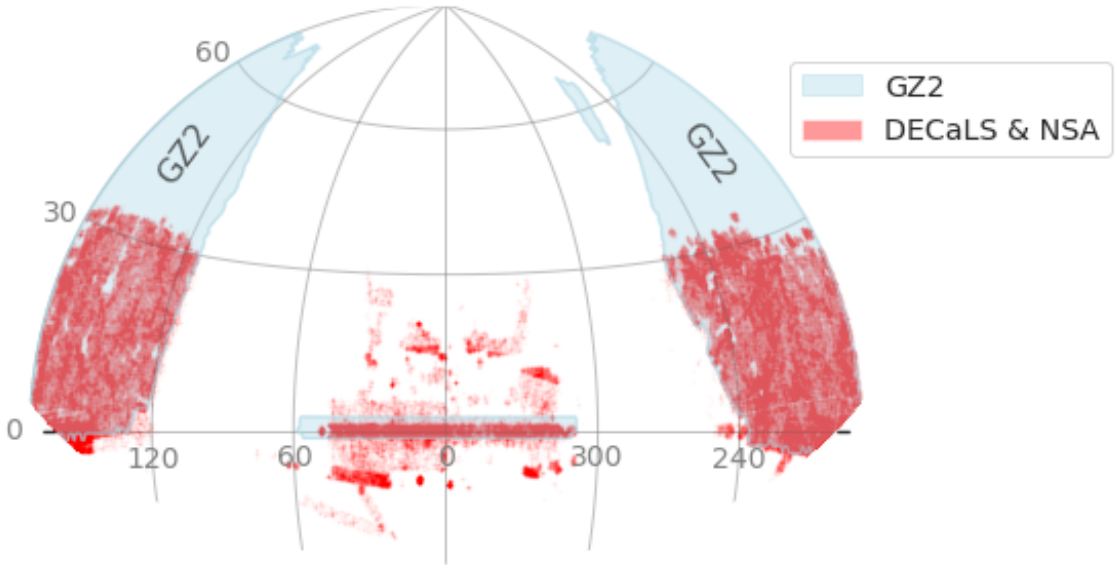


Figure 4.1: Sky coverage of GZ DECaLS (equatorial coordinates), resulting from the imaging overlap of DECaLS DR5 and SDSS DR8, shown in red. Darker areas indicate more galaxies. Sky coverage of Galaxy Zoo 2, which used images sourced from SDSS DR7, shown in light blue. The NSA includes galaxies imaged by SDSS DR8, including galaxies newly imaged at the Southern Galactic Cap (approx. $2500\,\mathrm{deg}^2$)

of ancillary information derived from SDSS and related surveys, and allows for the measurement of any shift in classifications versus Galaxy Zoo 2 using the subset of SDSS DR8 galaxies classified both in Galaxy Zoo DECaLS and in Galaxy Zoo 2 (Sec. 4.4). Figure 4.1 shows the resulting GZ DECaLS sky coverage. NSA v1.0.0 was not published but the columns used here are identical to those in NSA v1.0.1, released in SDSS DR13 [15].

Selecting galaxies with the NSA introduces two implicit cuts. First, the NSA primarily includes galaxies brighter than $m_r = 17.77$, the SDSS spectroscopic target selection limit. Galaxies fainter than $m_r = 17.77$ are included only if they are in deeper survey areas (e.g. Stripe82) or were measured using 'spare' fibres after all brighter galaxies in a given field were covered. Second, the NSA only covers redshifts of $z = 0.15$ or below. An explicit cut requiring a Petrosian radius (`PETROTHETA`) of at least 3 arcseconds was added to these implicit cuts to ensure the galaxy is sufficiently extended for meaningful classification.

For each galaxy, if the coordinates had been imaged in the g, r and z bands, and the galaxy passed the selection cuts above, a combined FITS cutout of the grz bands was acquired from the DECaLS cutout service (www.legacysurvey.org).

Volunteers were presented with 424×424 pixel square galaxy images. GZD-1

and GZD-2 acquired 424 × 424 pixel square FITS cutouts directly from the cutout service. To ensure that galaxies typically fit well within a 424 pixel image, cutouts were downloaded with an interpolated pixel scale s of

$$s = \max(\min(p_{50} * 0.04, p_{90} * 0.02), 0.1) \tag{4.1}$$

where p_{50} is the Petrosian 50%-light radius and p_{90} is the Petrosian 90%-light radius.

For GZD-5, to avoid banding artifacts caused by the interpolation method of the DECaLS cutout service, I downloaded each FITS image at the fixed native telescope resolution of 0.262 arcsec^2 per pixel [4] , with enough pixels to cover the same area as 424 pixels at the interpolated pixel scale s. I then resized these individually-sized FITS up to the interpolated pixel scale s by Lanczos interpolation [244]. Image processing was otherwise identical between campaigns. Galaxies with incomplete imaging, defined as more than 20% missing pixels in any band, were discarded. I retrospectively applied this requirement to GZD-1 and GZD-2, which had required no more than 20% missing pixels *over all bands*; this failed to remove galaxies only partially imaged in one band. For GZD-1/2, 92,960 of 101,252 galaxies had complete imaging (91.8%). For GZD-5, 216,106 of 247,746 galaxies not in DECaLS DR1/2 had complete imaging (87.2%)[5].

4.2.3 RGB Image Construction

The measured *grz* fluxes were converted into RGB images. To use the *grz* bands as RGB colours, the flux values in each band were multiplied by 125.0, 71.43, and 52.63, respectively. These numbers are chosen by eye (by Dustin Lang for GZD-1) such that typical subjects show an appropriate range of colour once mapped to RGB channels.

For background pixels with very low flux, and therefore high variance in the proportion of flux per band, naively colouring by the measured flux creates a speckled effect [459]. As an extreme example, a pixel with 1 photon in the g band and no photons in r or z would be rendered entirely red. To remove these colourful speckles, pixels with very low flux were desaturated. The total per-pixel photon count N was estimated assuming an exposure time of 90 seconds per band and a mean photon frequency of 600nm. Poisson statistics imply the standard deviation on the total mean flux in that pixel is proportional to $\sqrt{N}$. Pixels with a standard deviation below 100 had their per-band deviation from the mean per-pixel flux multiplied (and hence

[4] Up to a maximum of 512 pixels per side. Highly extended galaxies were downloaded at reduced resolution such that the FITS had exactly 512 pixels per side.

[5] Note that these numbers do not sum to the total number of galaxies classified across both campaigns because some galaxies are shared between campaigns.

reduced) by a factor of 1% of the standard deviation. The effect is to reduce the saturation of low-flux pixels in proportion to the standard deviation of the total flux. Mathematically,

$$X'_{ijc} = \overline{X_{ij}} + \alpha X_{ijc} \quad \text{where} \quad \alpha = \min(0.01\sqrt{\overline{X_{ij}}T/\lambda}, 1) \tag{4.2}$$

where X_{ijc} and X'_{ijc} are the flux at pixel ij in channel c before and after desaturation, $\overline{X_{ij}}$ is the mean flux across bands at pixel ij, T is the mean exposure time (here, 90 seconds) and λ is the mean photon wavelength (here, 600 nm).

Pixel values were scaled by $\sinh^{-1}(x)$ to compensate for the high dynamic range typically found in galaxy flux, creating images that can show both bright cores and faint outer features. To remove the very brightest and darkest pixels, the pixel values were linearly rescaled to lie on the $(-0.5,\ 300)$ interval and then clipped to 0 and 255 respectively. These final values were then used to create an RGB image using `pillow` [439]. The images are available on Zenodo at https://doi.org/10.5281/zenodo.4196266.

I experimented with other methods for constructing RGB images that might better reveal detailed morphology. A mosaic of my experiments is shown in Figure 4.2. I ultimately decided that being consistent with GZD-1 and GZD-2 was more important than improving the images, and so I did not change the method described above for GZD-5. However, I hope that these experiments encourage similar experimentation before the launch of the next Galaxy Zoo project. My code is available on Github here.

4.3 Volunteer Classifications

Volunteer classifications for GZ DECaLS were collected during three campaigns. GZD-1 and GZD-2 classified all 99,109 galaxies passing the criteria above from DECALS DR1 and DR2, respectively. GZD-1 ran from September 2015 to February 2016, primarily managed by Kyle Willett, and GZD-2 from April 2016 to February 2017, primarily managed by Coleman Krawczyk. GZD-5 classified 262,000 DECALS DR5-only galaxies passing the criteria above. GZD-5 ran from March 2017 to October 2020, primarily managed by myself.

4.3.1 Decision Trees

The questions and answers which Galaxy Zoo asks define the morphology measurements published. It is therefore critical that the Galaxy Zoo decision tree matches the science goals of the research community.

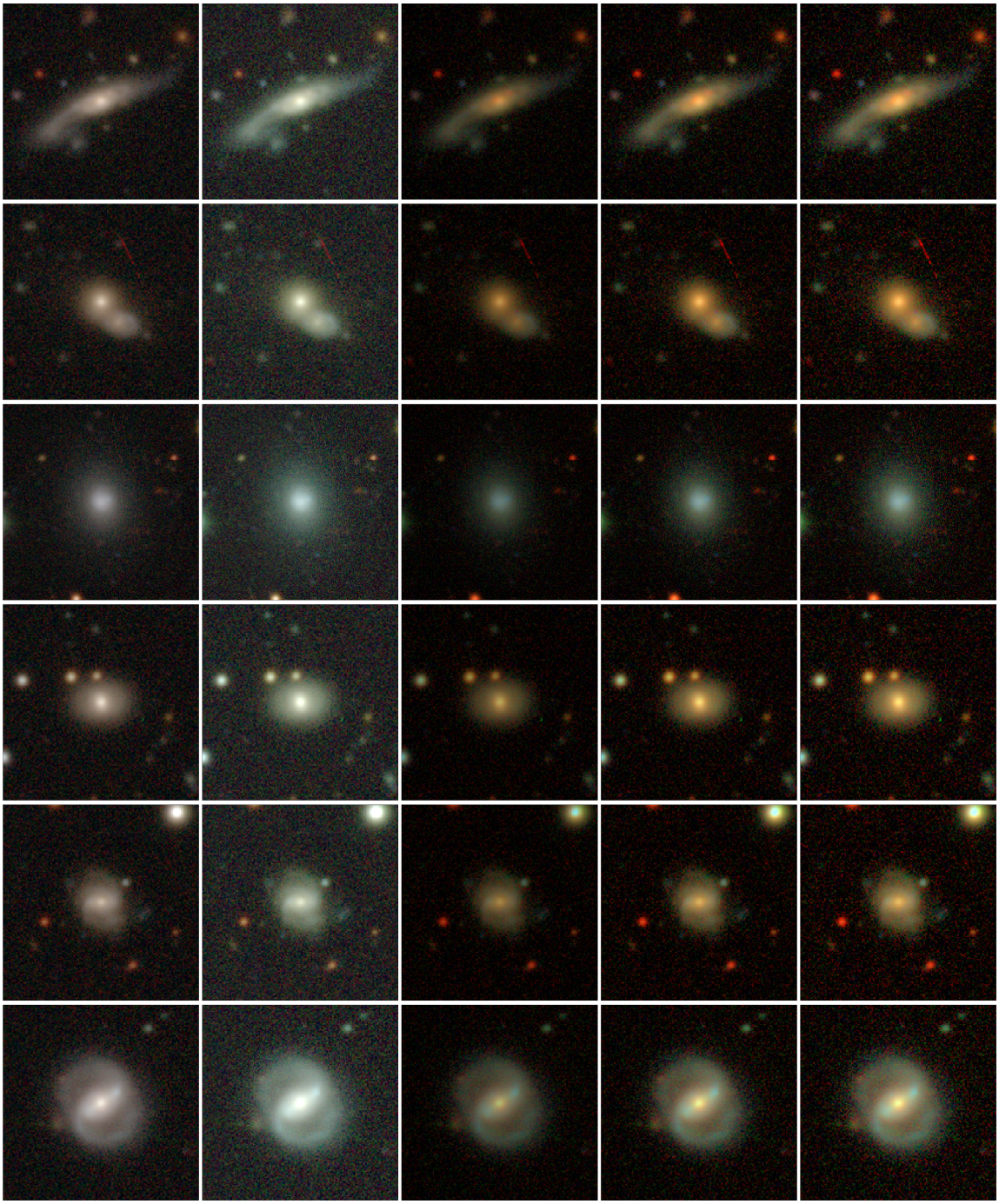


Figure 4.2: Comparison of galaxy images converted to RGB with different methods. Columns from left: images with the method described here and used in all GZ DECalS campaigns; images with a method developed by the DECaLS collaboration (https://github.com/legacysurvey/legacypipe/tree/master/py/legacypipe); images with an alternative method I developed based on Lupton scaling [277] for three sets of parameter choices (see the code), not used in GZ DECaLS. The choice of colour and scaling when creating RGB images can affect the visibility of detailed morphological features such as bars.

The decision tree used for GZD-1 and GZD-2 included three modifications from the Galaxy Zoo 2 decision tree [456]. The 'Can't Tell' answer to 'How many spiral arms are there?' was removed, the number of answers to 'How prominent is the central bulge?' was reduced from four to three, and 'Is the galaxy currently merging, or is there any sign of tidal debris?' was added as a standalone question.

For GZD-5, the Galaxy Zoo science team and I decided to make three further changes. Several Galaxy Zoo studies (e.g. [241, 294, 393, 456]) found that galaxies selected with $0.2<p_{\rm bar}<0.5$ in GZ2 correspond to 'weak bars' when compared with expert classification such as those in [313]. Therefore, to increase the detection of bars, we changed the possible answers to the 'Does this galaxy have a bar?' question from 'Yes' or 'No' to 'Strong', 'Weak' or 'No'. We defined a strong bar as one that is clearly visible and extending across a large fraction of the size of the galaxy. A weak bar is smaller and fainter relative to the galaxy, and can appear more oval than the strong bar, while still being longer in one direction than the other. Our definition of strong vs. weak bar is similar to that of Nair and Abraham (2010) [313], with the exception that they also have an 'intermediate' classification. We added examples of galaxies with 'weak bars' to the Field Guide and provided a new icon for this classification option, as shown in Figure 4.3.

Second, to allow for more fine-grained measurements of bulge size, we increased the number of "How prominent is the central bulge?" answers from three ('No', 'Obvious', 'Dominant') to five ('No Bulge', 'Small', 'Moderate', 'Large', 'Dominant'). We also re-included the 'Can't Tell' answer.

Third, we modified the 'Merging' question from 'Merging', 'Tidal', 'Both', or 'None', to the more phenomenological 'Merging', 'Major Disturbance', 'Minor Disturbance', or 'No'. I argued for this change in the hope of better distinguishing major and minor mergers, motivated by the science case I outlined in Chapter 2. We made this final "merger" change two months after launching GZD-5; 6722 GZD-5 galaxies were fully classified before that date and so do not have responses from volunteers to this question.

Several improvements were made to the illustrative icons shown for each answer. These icons are the most visible guide for volunteers as to what each answer means (complementing the tutorial, help text, field guide, and 'Talk' forum). Figure 4.3 shows the GZD-5 decision tree with new icons as shown to volunteers. The new icons were made by Dr. Brooke Simmons and Becky Rother.

Changes to the decision tree complicate comparisons with other Galaxy Zoo projects. As I show in the following sections, the available answers affect the sensitiv-

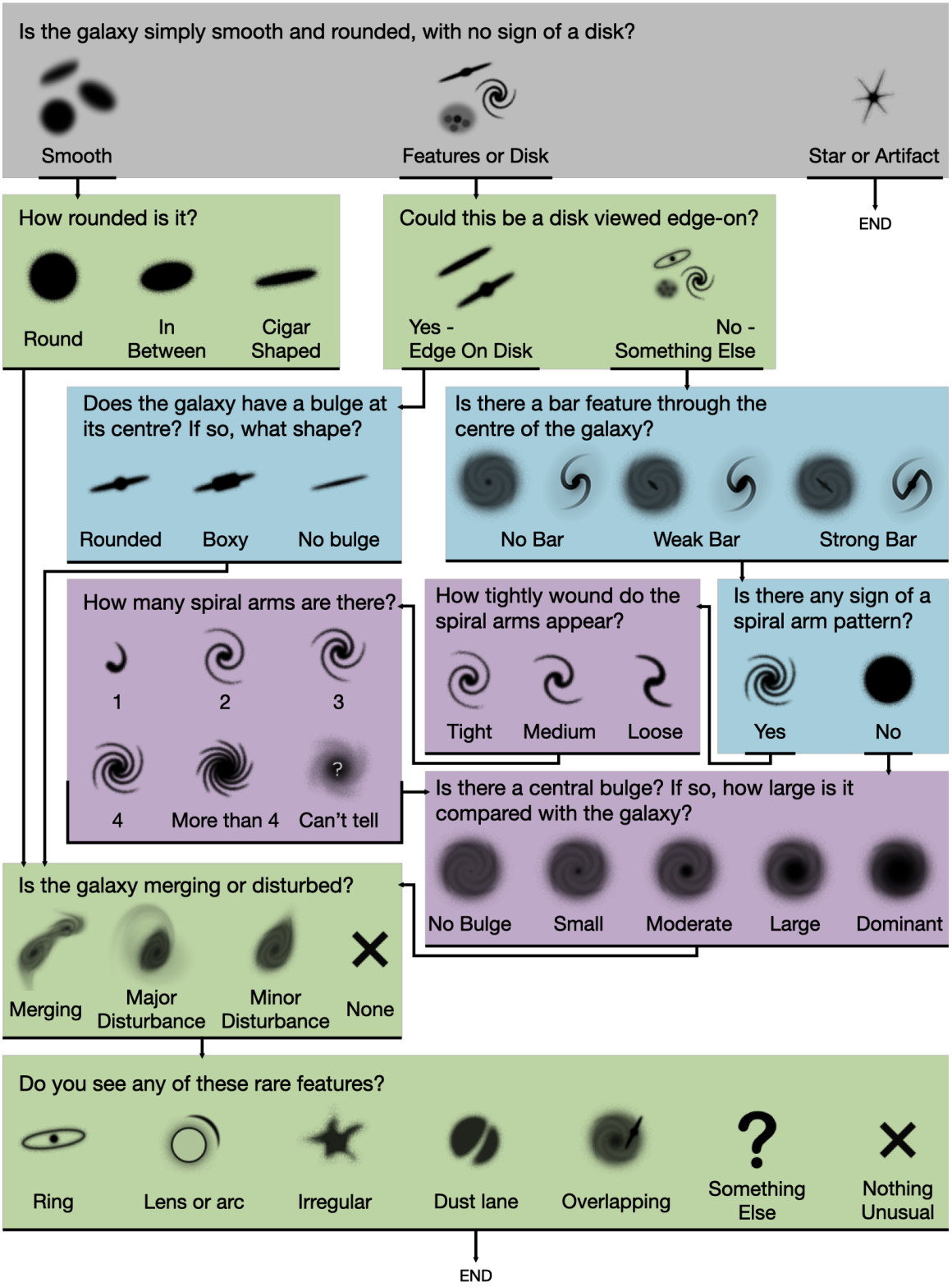


Figure 4.3: Classification decision tree for GZD-5, with new icons as shown to volunteers. Questions shaded with the same colours are at the same level of branching in the tree; grey questions have zero dependent questions, green one, blue two, and purple three.

ity of volunteers to certain morphological features, and so morphology measurements made with different decision trees may not be directly comparable. This difficulty in comparison has historically required the Galaxy Zoo science team to be conservative in making changes to the decision tree. However, the advent of effective automated classifications allows for retrospective classifications using any preferred decision tree. Specifically, in the work I present in this chapter, I trained our automated classifier to predict what volunteers would have said using the GZD-5 decision tree, for galaxies that were originally classified by volunteers using the GZD-1/2 decision tree (Section 4.5.1).

4.4 Volunteer Analysis

4.4.1 Improved Feature Detection from DECaLS imagery

The images used in GZ DECaLS are deeper and higher resolution than were available for GZ2. The GZ2 primary sample [456] used images from SDSS DR7 [3], which are 95% complete to $r = 22.2$ with a median seeing of $1\farcs4$ and a plate scale of $0\farcs396$ per pixel [467]. In contrast, GZ DECaLS used images from DECaLS DR2 to DR5, which have a median 5σ point source depth of $r = 23.6$, a seeing better than $1\farcs3$ for at least one observation, and a plate scale of $0\farcs262$ per pixel [99][6].

I hoped the improved imaging would reveal morphology not previously visible, particularly for features which are faint (e.g. tidal features, low surface brightness spiral arms) or intricate (e.g. weak bars, flocculent spiral arms). The GZD-5 changes to the decision tree (Sec. 4.3.1) were partly made to better exploit this improved imaging.

To investigate the consequences of improved imaging, I compared galaxies classified in both GZ2 and GZ DECalS. Galaxies will typically be classified by both projects if they are inside both the SDSS DR7 Legacy catalogue (i.e. the source GZ2 catalogue) and DECaLS DR5 footprints (broadly, North Galactic Cap galaxies with $-35 < \delta < 0$) and match the selection criteria of each project (see [456] and Sec. 4.2.2). GZ2's $r < 17.0$ cut, with no corresponding GZ DECaLS magnitude cut, means that the odds of any given GZ2 galaxy being in GZ DECaLS is close to random (for an isotropic sky) but only the brighter half of suitably-located GZ DECaLS galaxies are in GZ2. To exclude the effect of modifying the decision tree in GZD-5 (addressed separately in Sec 4.4.2), I only compared GZ DECaLS classifications from GZD-1 and GZD-2. 33,124 galaxies were classified in both GZ2 and GZD-1 or GZD-2.

[6]See also http://www.legacysurvey.org/dr5/description/

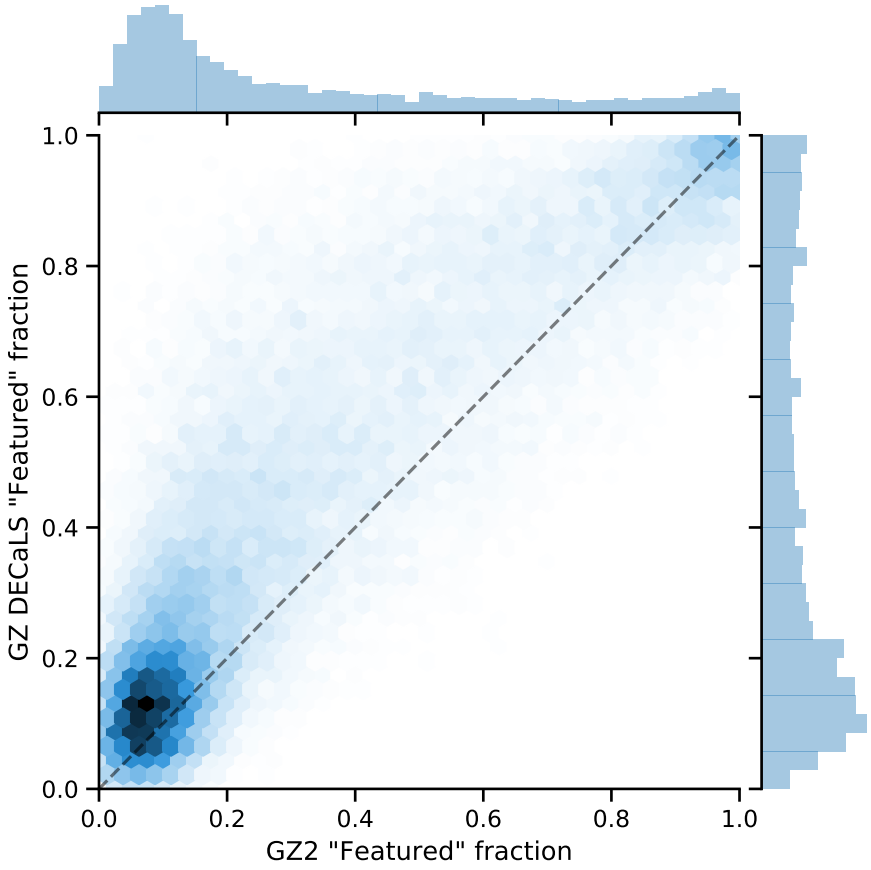


Figure 4.4: Comparison of 'Featured' fraction for galaxies classified in both GZ2 and GZ DECaLS. Ambiguous galaxies are consistently reported as more featured in GZ DECaLS, which I attribute to the significantly improved imaging depth of DECaLS.

I found that volunteers successfully recognise newly-visible morphology features. Figure 4.4 compares the distribution of vote fractions to 'Is this galaxy smooth or featured?' for GZ2 and GZ DECaLS. Ambiguous galaxies, with 'featured' fractions between approx. 0.25 and 0.75 are consistently reported as more featured (median absolute increase of 0.13, median percentage increase of 22%) with the deeper GZ DECaLS images.

The shift towards featured galaxies is an accurate response to the new images, rather than systematics from (for example) a changing population of volunteers. Figure 4.5 compares the GZ2 and GZ DECaLS images of a random sample of galaxies drawn from the 1000 cross-classified galaxies with the largest increase in 'featured' fraction. In all of these galaxies (and for a clear majority of galaxies in similar samples), volunteers are correctly recognising newly visible detailed features.

One can observe a similar pattern in the vote fractions of spiral arms and bars

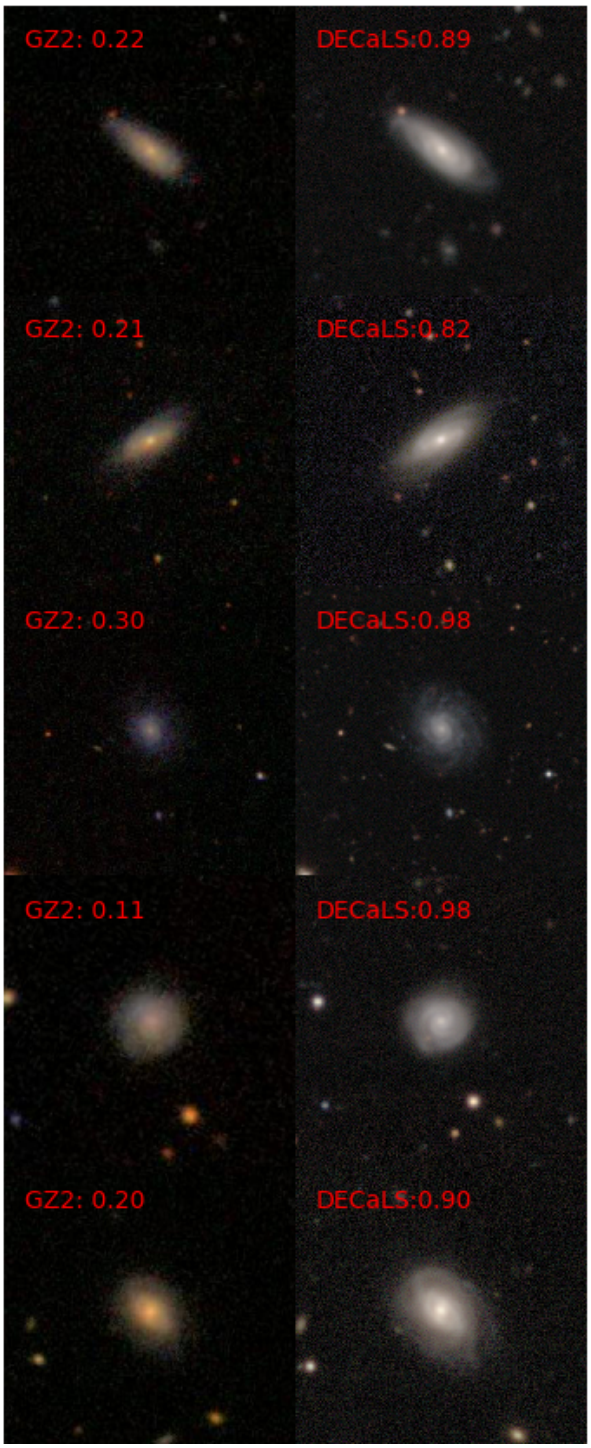


Figure 4.5: GZ2 and GZ DECaLS images for 5 galaxies drawn randomly from the 1000 galaxies classified in both projects with the largest increase in 'featured' vote fraction (reported fractions shown in red). The increased fraction accurately reflects the increased visibility of detailed morphology from improved imaging.

for featured galaxies. For galaxies consistently considered featured (i.e. where both projects reported a 'featured' vote fraction of at least 0.5), the median vote fraction for spiral arms increased from 0.84 to 0.9, and for bars from 0.21 to 0.24. This suggests that even for galaxies where some details were already visible (and hence were considered featured), improved imaging makes our volunteers more likely to identify specific features.

I believe the improved depth of DECaLS ($r = 23.6$ vs $r = 22.2$ for SDSS) is revealing low surface brightness features that were previously ambiguous. There may also be contributions from the modified image processing approach and from the shift between using gri bands (SDSS) to grz bands (DECaLS), which might make older stars more prominent.

Comparing classifications made using the same possible answers on the same galaxies shows how improved DECaLS imaging leads to ambiguous galaxies being correctly reported as more featured, and to spiral arms and bars being reported with more confidence. However, volunteers are also sensitive to which questions are asked and how those questions are asked. I consider the impact of our changes to the decision tree 'Bar' question for GZD-5 in the next section.

4.4.2 Improved Weak Bar Detection from GZD-5 Decision Tree

To measure the effect of the new decision tree on bar sensitivity, I compared the classifications made using each tree against expert classifications. Nair and Abraham 2010 [313] (hereafter NA10) classified all 14,034 SDSS DR4 galaxies at $0.01 < z < 0.05$ with $g < 16$. Of those, 1497 were imaged by DECaLS DR1/2 and classified by volunteers during GZD-1/2. I chose to re-classify these galaxies during GZD-5 to measure the effect of the new bar answers, as compared to the expert classifications of NA10. Note that because NA10 used shallower SDSS images, NA10's classifications are best used as positive evidence; while NA10 finding a bar in SDSS images implies a visible bar in DECaLS images, NA10 not finding a bar may not always exclude a visible bar in DECaLS. To exclude smooth galaxies, which are unbarred by definition in all Galaxy Zoo trees, I made a cut of $f_{\text{featured}} > 0.25$ (as measured by GZD-5), selecting a featured sample of 807 galaxies classified by NA10, GZD-1/2, and GZD-5.

Figure 4.6 compares volunteer classifications for expert-labelled calibration galaxies made using each tree. Barred and unbarred galaxies are shown to be significantly better separated with the Strong/Weak/None answers than with Yes/No answers. Of

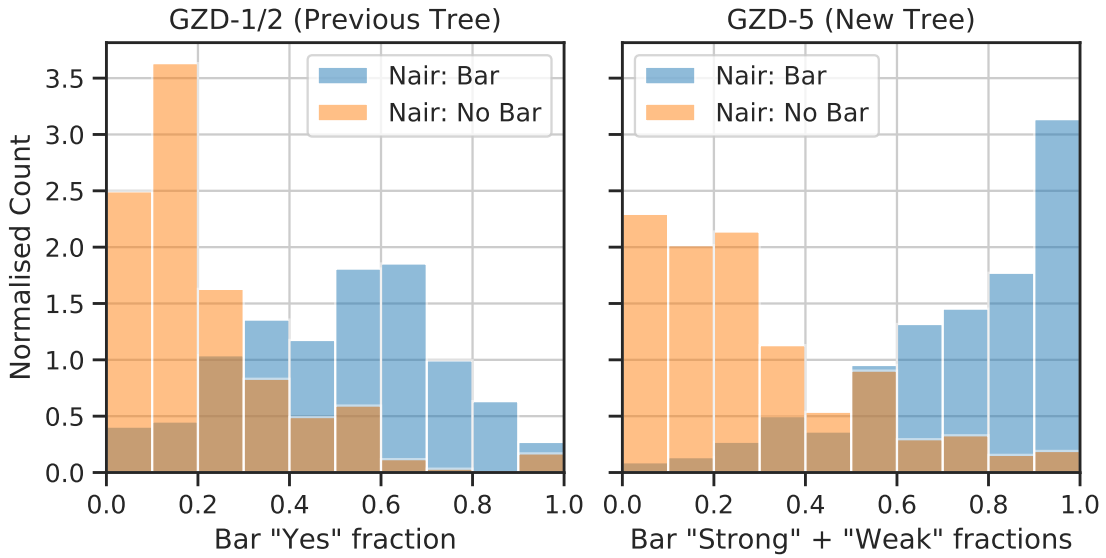


Figure 4.6: Left: Distribution of fraction of GZD-1/2 volunteers answering 'Yes' (not 'No' to 'Does this galaxy have a bar?', split by expert classification from NA10 of barred (blue) or unbarred (orange). Right: as left, but for GZD-5 volunteers answering 'Strong' or 'Weak' (not 'No'). Volunteers are substantially better at identifying barred galaxies using the GZD-5 three-answer question.

220 Nair-identified bars (of any type), 184 (84%) received a majority vote for being barred by volunteers using the new tree, up from 120 (55%) with the previous tree.

NA10 classified barred galaxies into five subtypes: Strong, Intermediate, Weak, Nuclear, Ansae, and Peanut (plus None, implicitly). I used the first three subtypes as a measurement of expert-classified bar strength to evaluate how volunteers respond to bars of different strengths. Following the approach for defining summary metrics of Masters et al. 2019 [292], I summarised the bar vote fractions into a single volunteer estimate of bar strength, $B_{\rm vol} = f_{\rm strong} + 0.5 f_{\rm weak}$. Figure 4.7 compares the distribution of B for each expert-classified bar strength. The volunteer bar strength estimates increase smoothly with expert-classified bar strength, though individual galaxies vary substantially. This suggests that typical bar strength in galaxy samples can be successfully inferred from volunteer votes.

The addition of the 'weak bar' answer in GZD-5 significantly improves sensitivity to bars compared with previous versions of the decision tree. Additionally, volunteer votes across the three answers may be used to infer bar strength. I hope that the detailed bar classifications in our catalogue will help researchers better understand the properties of strong and weak bars and their influence on host galaxies. The classifications are already being used within the Galaxy Zoo science team by Tobias Géron

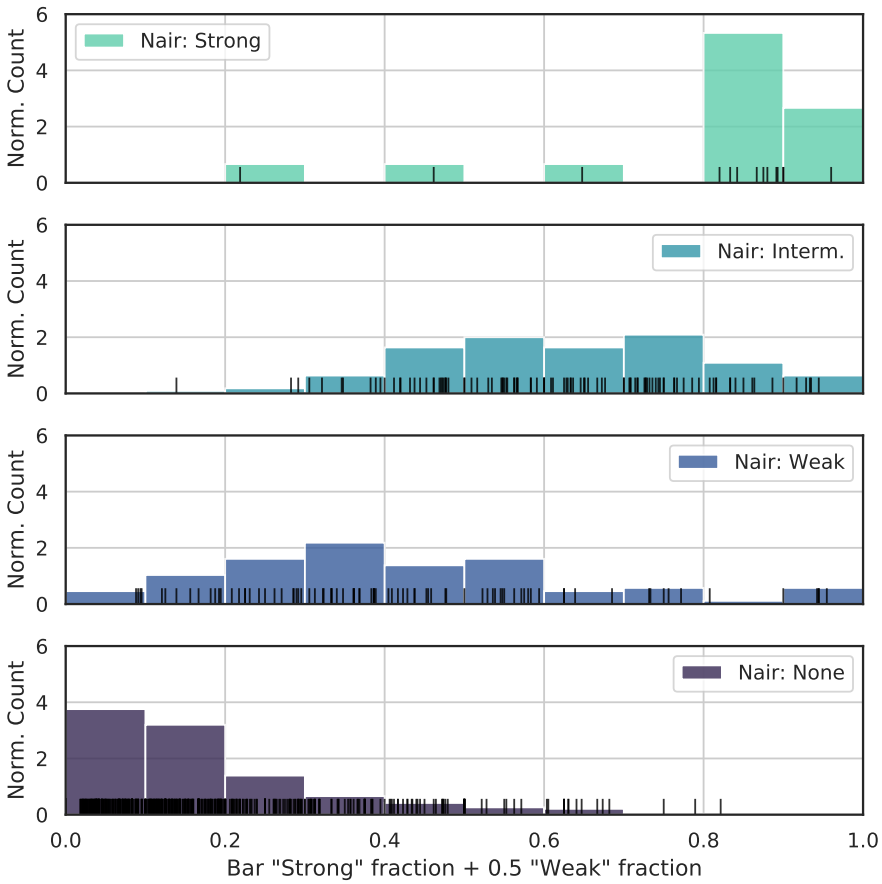


Figure 4.7: Distributions of volunteer bar strength estimates, $B_{\rm vol} = f_{\rm strong} + 0.5 f_{\rm weak}$, split by expert-classified (NA10) bar strength. Individual galaxies are shown with rug plots (15 Strong, 110 Intermediate, 87 Weak, and 377 None). Volunteer bar strength estimates increase smoothly with expert-classified bar strength, though individual galaxies vary substantially.

to evaluate whether strong or weak bars are distinct classes or part of a continuum, and how each affects quenching (Géron, private communication).

4.4.3 Classification Modifications

Galaxy Zoo data releases have previously included two post-hoc modifications to the volunteer classifications; volunteer weighting, to reduce the influence of strongly atypical volunteers, and redshift debiasing, to estimate the vote fractions a galaxy might have received had it been observed at a specific redshift. Redshift debiasing was carried out by Sandor Kruk, Steven Bamford and Lee Kelvin using the method described in Hart et al. 2016 [157], and is fully described in the paper[7] on which this work is based; I describe volunteer weighting below.

Volunteer weighting, as introduced in Galaxy Zoo 2 [456], assigns each volunteer an aggregation weight of (initially) one, and iteratively reduces that weight for volunteers who typically disagree with the consensus. This method affects relatively few volunteers and therefore causes only a small shift in vote fractions - in Galaxy Zoo 2, for example, approximately 95% of volunteers had a weighting of one (i.e. unaffected), 94.8% of galaxies had a change in vote fraction of no more than 0.1 for any question, and the mean change in vote fraction across all questions and galaxies was 0.0032.

The most significant change in final vote fractions is typically caused by down-weighting rare (approx. 1%) volunteers who repeatedly disagree with consensus by answering 'artifact' at implausibly high rates (including 100%) for many galaxies. Answering artifact ends the classification and shows the next galaxy, and so it may be that these rare volunteers are primarily interested in seeing many galaxies rather than contributing meaningful classifications. There are very few such volunteers, but because answering artifact allows classifications to be submitted very quickly, they have an outsize effect on the aggregate vote fractions.

Figure 4.8 shows the distribution of reported artifact rates for volunteers with at least 150 total classifications. The true fraction of artifacts is expected to be less than 0.1, and the vast majority of volunteers report artifact rates consistent with this. However, the distribution is bimodal, with a small second peak around 1.0 (i.e. volunteers reporting every galaxy as an artifact). To remove the implausible mode, I chose to discard the classifications of volunteers with at least 150 total classifications and reported artifact rates greater than 0.5. In GZD-1/2, 1.1% (643) of volunteers are excluded, discarding 11% (483,081) of classifications. In GZD-5, 0.03% (543)

[7]Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies, Walmsley et. al, under review and available on arxiv.

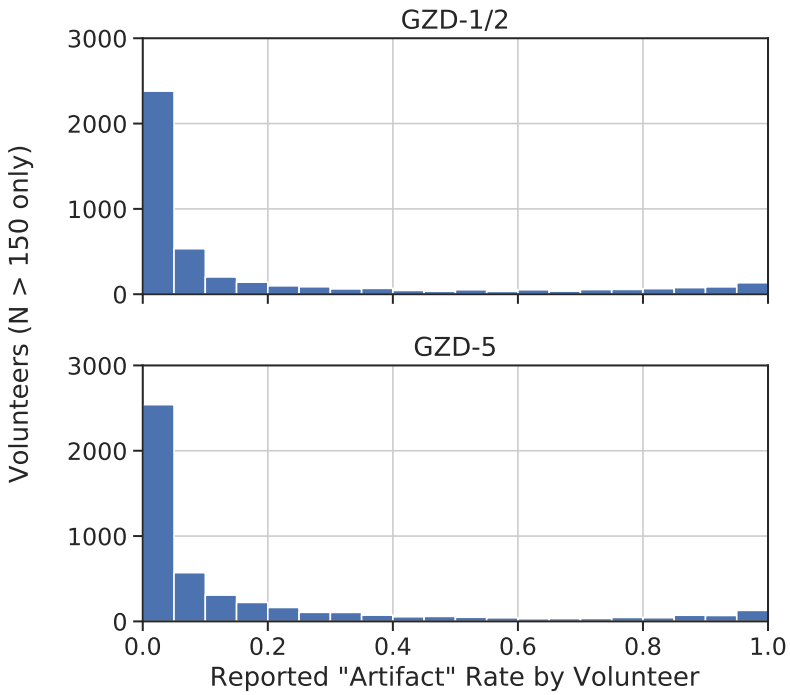


Figure 4.8: Distribution of reported 'artifact' rates by volunteer (i.e. how often each volunteer answered 'artifact' over all the galaxies they classified). The vast majority report artifact rates consistent with those of the authors (below 0.1), but a very small subset report implausibly high artifact rates (> 0.5) and consequently have their classifications discarded. Only volunteers with at least 150 classifications are shown; the distribution for volunteers with fewer classifications is not bimodal.

volunteers are excluded, discarding 5.3% (249,592) of classifications. Interestingly, the bimodal distribution only appears when considering only volunteers with at least approximately 100 total classifications, suggesting that artifact-clicking volunteers tend to either stop after a handful (and so are indistinguishable from the volunteers who happen to genuinely only see several artifacts) or click artifact on very large numbers of galaxies.

I investigated the possibility of other groups of atypical volunteers giving similar answers across questions by analysing the per-user vote fractions with either a two-dimensional visualisation using UMAP [297] or with clustering using HDBSCAN [296]. I found no strong evidence that such clusters exist. The frequency of each answer does vary substantially between volunteers, suggesting that more sophisticated weighting may be important; I discuss this further in Sec. 4.8.

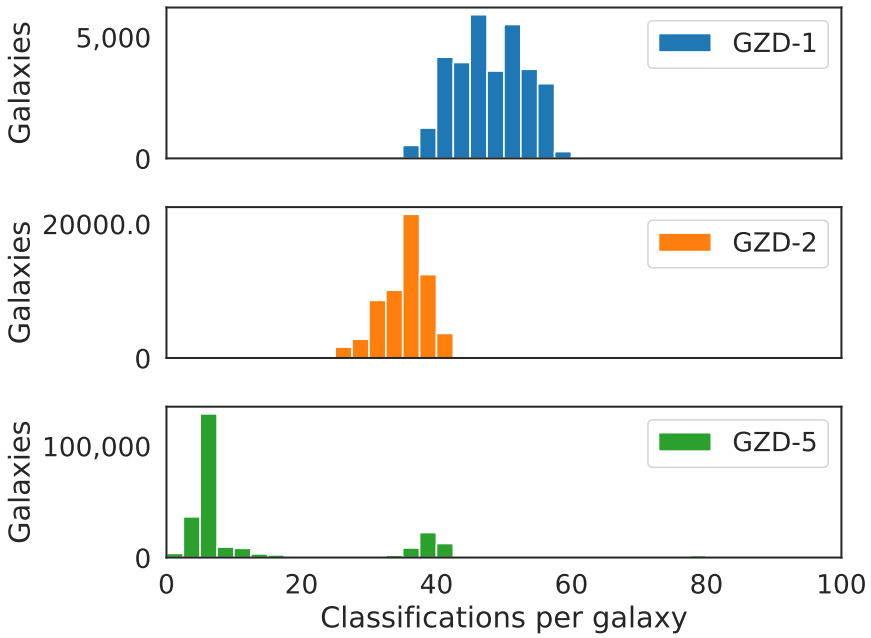


Figure 4.9: GZD-1, GZD-2 and GZD-5 classification counts, excluding implausible classifications (Sec. 4.4.3). GZD-1 has approximately 40-60 classifications, GZD-2 has approximately 40, and GZD-5 has either approximately 5 or approximately 30-40. 5.9% of GZD-5 galaxies received more than 40 classifications due to mistaken duplicate uploads.

4.4.4 Retirement

For GZD-1 and GZD-2, all galaxies received at least 40 classifications (as with previous data releases). GZD-1 galaxies have between 40 and 60 classifications, selected at random, while GZD-2 galaxies all have approximately 40. For GZD-5, galaxies classified before June 2019 also received approximately 40 classifications while I developed the active learning system described in Sec. 3.2. I activated the system in June 2019. Using active learning, galaxies expected to be the most informative for training our deep learning model received 40 classifications, and all remaining galaxies received at least 5 classifications.

The active learning system used a Bayesian CNN trained to predict the 'Smooth or Featured' question following the same architecture and training procedure already described in Sec. 3.2. The initial training set was all GZD-5 galaxies fully classified ($N > 36$) by the time of activation. Each week, the model was retrained with all galaxies fully classified by that date. Next, unlabelled galaxies were ranked with the BALD-derived acquisition function (see Sec. 3.2), and the most informative 1000 of a random 32768[8] galaxies were uploaded. I chose to select from a subset of galaxies not yet classified for two reasons. The first was for computational efficiency: calculating the acquisition function requires making 5 predictions per galaxy. The second was that my ad hoc experiments showed that galaxies with the very highest acquisition function values were often artifacts or otherwise highly unusual, and I was concerned these might be *too* unusual to learn from effectively. I also added a retirement rule to retire galaxies receiving 5 classifications of 'artifact', to help avoid volunteers being presented with these prioritised artifacts. While active learning was running, I continued to improve my Bayesian CNN approach based on the results of my experiments with Galaxy Zoo 2 (Sec. 3.2). In the next section, I describe these improvements.

4.5 Automated Classifications

In Chapter 3 I showed how, through careful consideration of uncertainty, models can both learn from uncertain volunteer responses and predict posteriors (rather than point estimates) for new galaxies. However, my work also revealed some practical limitations (Sec.3.3). Here, I describe improving my approach to overcome these limitations. I then share results measuring the performance at predicting volunteer votes for DECALS galaxies fully ($N > 36$) labelled during GZD-5. The improved

[8]To allow for shuffling the larger-than-memory galaxy dataset, I stored encoded galaxy images in 'shards' of 4096 galaxies each. 32,768 corresponds to 8 such shards

classification approach is then used to predict volunteer votes for all GZ DECaLS galaxies.

4.5.1 Improved Bayesian Deep Learning Classifier

My overall goals remain the same as in the previous chapter. I aimed to develop a classifier that can:

1. Learn efficiently from volunteer responses of varying (i.e. heteroskedastic) uncertainty.

2. Predict posteriors for those responses on new galaxies, for every question.

In Chapter 3, my collaborators and I modelled volunteer responses as being binomially distributed and trained our model to make maximum likelihood estimates using the loss function of Eqn. 3.6:

$$\mathcal{L} = k \log f^w(x) + (N - k) \log(1 - f^w(x))$$

where, for some target question, k is the number of responses (successes) of some target answer, N is the total number of responses (trials) to all answers, and $f^w(x) = \hat{\rho}$ is the predicted probability of a volunteer giving that answer.

This Binomial assumption, while broadly successful, broke down for galaxies with vote fractions $\frac{k}{N}$ close to 0 or 1, where the Binomial likelihood is extremely sensitive to $f^w(x)$, and for galaxies where the question asked was not appropriate (e.g. predict if a featureless galaxy has a bar). Instead, in this chapter, the model predicts a distribution $p(\rho|f^w(x))$ and ρ is then drawn from that distribution.

One could parametrise $p(\rho|f^w(x))$ with the Beta distribution (chosen for being flexible and defined on the unit interval), and predict the Beta distribution parameters $f^w(x) = (\hat{\alpha}, \hat{\beta})$ by minimising

$$\mathcal{L} = \int \mathrm{Bin}(k|\rho, N)\mathrm{Beta}(\rho|\alpha, \beta) d\alpha d\beta \tag{4.3}$$

where the Binomial and Beta distributions are conjugate and hence this integral can be evaluated analytically. This would work for binary questions, and hence (by training a model to predict either 'this answer' or 'not this answer' for each answer, see Sec. 3.1.3) for all questions. However, one would need to manage a menagerie of models. It would be far more convenient to directly predict the responses to questions with more than two answers. I therefore decided to replace each distribution with

its multivariate counterpart; Beta($\rho|\alpha, \beta$) with Dirichlet($\rho|\alpha$), and Binomial($k|\rho, N$) with Multinomial($\mathbf{k}|\rho, N$).

$$\mathcal{L}_q = \int \text{Multi}(\mathbf{k}|\rho, N)\text{Dirichlet}(\rho|\alpha)d\alpha \tag{4.4}$$

where $\mathbf{k}, \rho$ and α are now all vectors with one element per answer.

The Dirichlet-Multinomial distribution loss function is much more flexible than the Binomial, allowing the model to express uncertainty through wider posteriors and confidence through narrower posteriors. This is a major qualitative improvement over the Binomial distribution loss function, where the model could only specify a typical response probability ρ and the uncertainty was then completely determined by the Binomial distribution shape given that ρ and the number of volunteers asked. I believe this is a novel approach.

For the base architecture, I chose to use the EfficientNet B0 model [414]. The EfficientNet family of models includes several architectural advances over the standard architectures used commonly within astrophysics (e.g. [76, 100, 125, 191, 227] and my own previous work [444, 445]), including depthwise convolutions [183], bottleneck layers [196], and squeeze-and-excitation optimisation [186]. The EfficientNet B0 model was identified using multi-objective neural architecture search [164, 415], optimising for both accuracy and FLOPS (i.e. computational cost of prediction). This balancing of accuracy and FLOPS is particularly useful for astrophysics researchers with limited access to GPU resources. I briefly experimented with a larger EfficientNet (B2) and found no significant improvement in performance.

I modified the final EfficientNet B0 layer output units to give predictions smoothly between 1 and 100 (using softmax activation), which is appropriate for Dirichlet parameters α. α elements below 1 can lead to bimodal 'horseshoe' posteriors, and α elements above approximately 100 can lead to extremely confident predictions in extreme ρ, both of which are implausible for galaxy morphology posteriors. These constraints may cause galaxies where one answer is clearly the most appropriate to have predicted vote fractions which are slightly less extreme than volunteers would record, but I do not anticipate this to affect practical use; whether a galaxy is extremely likely to have a bar or merely highly likely is rarely of scientific consequence. I experimented with less restrictive output ranges and found they did not improve performance.

Lastly, motivated by recent empirical work showing that multi-task learning (where a model learns to optimise several objectives at once) can counter-intuitively often lead to better performance than 'single-task' learning [90], I wanted to predict the

answers to every question with a single model. This is similar to the seminal work of Dieleman et al. 2015 [100], and unlike more recent work e.g. [227, 370, 445]). Multi-task learning is thought to improve performance because the model learns a shared representation between tasks [72] - intuitively, knowing how to recognise spiral arms can also help you count them. Learning from every galaxy to predict every answer uses our valuable volunteer effort as efficiently as possible. This is particularly effective because I aimed to predict detailed morphology, and so needed detailed representations of each galaxy.

To predict the answer to every question, my EfficientNet architecture has one output unit per answer (i.e. for 13 questions with a total of 20 answers, I use 20 output units). The (negative log) likelihood is calculated per question as above (Eqn. 4.4), and then, treating the errors in the model's answers to each question as independent events, the total loss is calculated as

$$\log \mathcal{L} = \sum_q \mathcal{L}_q(\mathbf{k_q}, N_q, \mathbf{f_q^w}) \tag{4.5}$$

where, for question q, N_q is the total answers, $\mathbf{k_q}$ is the observed votes for each answer, and $\mathbf{f_q^w}$ is the values of the output units corresponding to those answers (which we interpret as the Dirichlet α parameters in Eqn. 4.4).

I trained the model using the GZD-5 volunteer classifications. Because the training set includes both active-learning-selected galaxies receiving at least 40 classifications and the remaining GZD-5 galaxies with around 5 classifications, it was crucial that the model could learn efficiently from labels of varying uncertainty. Unlike the previous chapter, where I trained one model per question and needed to filter galaxies where that question asked may not be appropriate, this new model predicts answers to all questions and learns from all labelled galaxies.

The galaxy images shown to volunteers (Section 4.2.3) were modified as follows. I took an average over channels to remove colour information and avoid biasing our morphology predictions (see Sec. 3.1.7.3), then resized and saved the images as 300x300x1 matrices in binary TFRecord format for rapid loading into memory. When loading each image I apply random augmentations, creating a unique randomly-modified image to be used as input to the network. I first apply random horizontal and vertical flips, followed by an aliased rotation by a random angle in the range $(0, \pi)$, with missing pixels being filled by reflection on the boundaries. Finally, I crop the image about a random centroid to 224x224 pixels, effectively zooming in slightly towards a

random off-centre point. I also apply these augmentations at test time to marginalise the posteriors over any unlearned variance.

I trained and evaluated models using the 249,581 (98.5%) GZD-5 galaxies with at least three volunteer classifications. Learning from galaxies with even fewer (one or two) classifications should be possible in principle, but I did not attempt it as I do not expect galaxies with so few classifications to be significantly informative.

The Dirichlet concentrations (distribution parameters) used to calculate the metrics were predicted by three identically-trained models, each making 5 forward passes with random dropout configurations and augmentations. The motivation for using multiple models, over and above the usual benefits of ensembling (see Sec. 2.3), is that recent empirical work [137] suggests that model initialisation has a substantial effect on predictions even when using MC Dropout. Using several models (with different random initialisations) mitigates this. I ensembled all 15 forward passes by simply taking the mean posterior given the total votes recorded, which may be interpreted as the posterior of an equally-weighted mixture of Dirichlet-Multinomial distributions. This mean posterior can then be used to calculate credible intervals (error bars) and in standard statistical analyses. I developed my approach using a conventional 80/20 train-test split, and made a new split before calculating the final metrics reported here.

For the published automated classifications, where I simply aimed to make the best predictions possible rather than to test performance, I trained models on all 249,581 galaxies with at least 3 votes. I trained five rather than three models to maximise performance. Training each model on an NVIDIA V100 GPU took around 24 hours. I then made predictions (using the updated GZD-5 schema) on all 313,789 galaxies in all campaigns.

When I previously argued (3.1.1) for the importance of automated morphological classification, I focused on the need to scale: being able to quickly classify large surveys. My ensemble of networks achieves this goal: each prediction (forward pass) takes approx. 6ms, equating to approx. 160ms for each published posterior and so around half a day to classify all GZ DECaLS galaxies. The speed of classification enables a further practical benefit: I can retroactively update the Galaxy Zoo decision tree. Because our classifier learns to make predictions from GZD-5 classifications, using the improved tree with better detection of mergers and weak bars, I can predict what our volunteers would have said for the GZD-1 and GZD-2 galaxies *had they been using the improved tree at that time.*

4.5.2 Results

The classifier successfully predicts posteriors for volunteer votes to each question. I show example posteriors for a question with two answers, 'has spiral arms', in Fig. 4.10, and a question with three answers, 'bulge size', in Fig. 4.11.

To aid intuition for the typical performance, I reduced both the vote fraction labels and the posteriors down to discrete classifications, and calculated classification metrics (Table 4.1) and confusion matrices (Figures 4.12-4.13). Here and throughout this section, to remove galaxies for which the question is not relevant, I only count galaxies where at least half the volunteers were asked that question. I report two sets of classification metrics; metrics for all (relevant) galaxies, and only for galaxies where the volunteers are confident (defined as having a vote fraction for one answer above 0.8, following [104]).

The performance on confident galaxies is useful to measure because such galaxies have a well-known correct label. For such galaxies, performance is near-perfect; we achieve better than 99% accuracy for most questions, with the lowest accuracy (for spiral arm count) being 98.6%. The confusion matrices reflect this, showing little notable confusion for any question.

Reported performance on all galaxies will be lower than on confident galaxies as the correct labels are uncertain. The measured vote fractions are approximations of the theoretical 'true' vote fractions (as one cannot ask infinitely many volunteers), and many galaxies are genuinely ambiguous and do not have a meaningful 'correct' answer. No classifier should achieve perfect accuracy on galaxies where the volunteers themselves are not confident. Nonetheless, performance is more than sufficient for scientific use; accuracy ranges from 77.4% (spiral arm count) to 98.7% (disk edge on). We observe some moderate confusion between similar answers, particularly between No or Weak bar, Moderate or Large bulges, and Two or Three spiral arms, which matches our intuition for the answers that volunteers might confuse and so likely reflects ambiguity in the training data. More surprisingly, there is also confusion between Two spiral arms and Can't Tell. Figure 4.16 shows random examples of spirals where the most common volunteer answer was Two, but the classifier predicted Can't Tell, and vice versa. In both cases, the galaxies generally have diffuse or otherwise subtle spiral arms embedded in a bright disk, confusing both human and machine. This highlights the difficulty in using discrete classification metrics to assess performance on ambiguous galaxies.

To mitigate the ambiguity in classifications of galaxies, one can instead measure regression metrics on the vote fractions without rounding to discrete classifications.

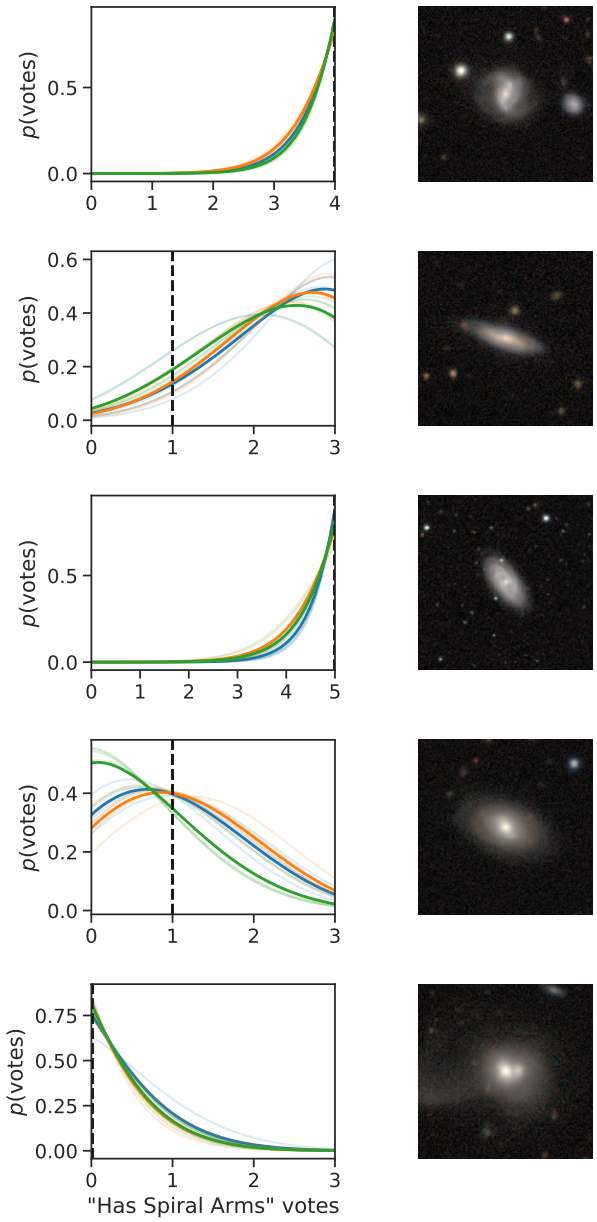


Figure 4.10: Posteriors for 'Does this galaxy have spiral arms?', split by ensemble model (bold colours) and, within each model, dropout forward passes (faded colours). The number of volunteers answering 'Yes' (not known to classifier) is shown with a black dashed line. Galaxies selected at random from the test set, provided the spiral question is relevant (defined as a vote fraction of 0.5 or more to the preceding answer, 'Featured'. The image presented to volunteers is shown to the right. The model input is a cropped, downsized, greyscale version (Sec 4.5.1).

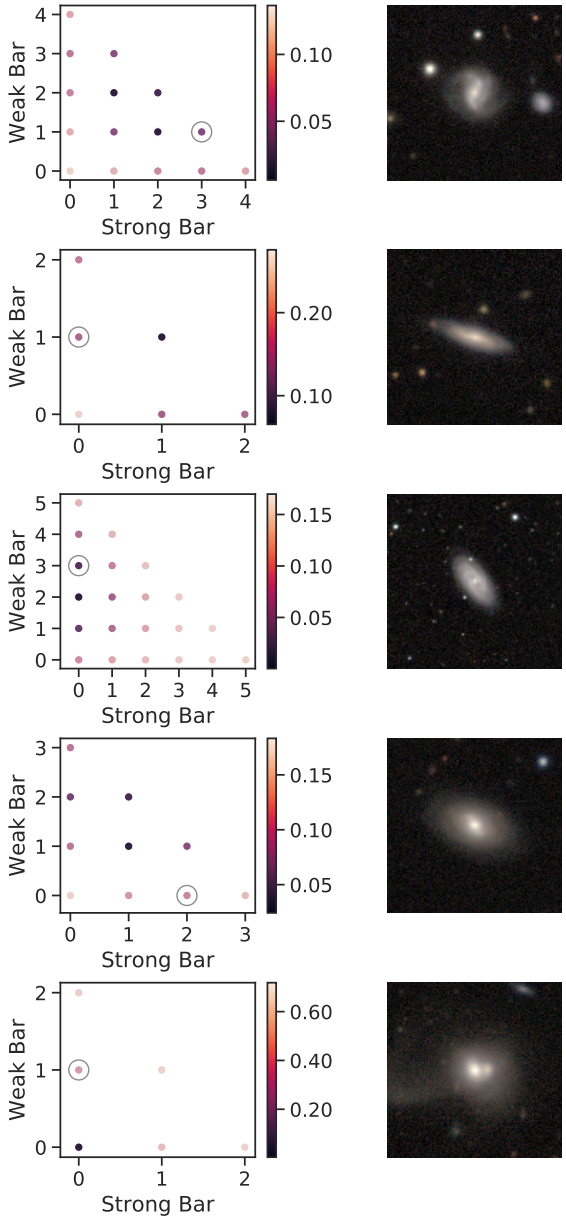


Figure 4.11: Posteriors for 'Does this galaxy have a bar?', for the same random galaxies selected in Fig. 4.10. Each point is coloured by the predicted probability of volunteers giving that many 'Strong', 'Weak', and (implicitly, as the total answers is fixed) 'None' votes. The volunteer answer (not known to classifier) is circled. For clarity, only the mean posterior across all models and dropout forward passes is shown.

Question	Count	Accuracy	Precision	Recall	F1
Smooth Or Featured	11346	0.9352	0.9363	0.9352	0.9356
Disk Edge On	3803	0.9871	0.9871	0.9871	0.9871
Has Spiral Arms	2859	0.9349	0.9364	0.9349	0.9356
Bar	2859	0.8185	0.8095	0.8185	0.8110
Bulge Size	2859	0.8419	0.8405	0.8419	0.8409
How Rounded	6805	0.9314	0.9313	0.9314	0.9313
Edge On Bulge	506	0.9111	0.9134	0.9111	0.8996
Spiral Winding	1997	0.7832	0.8041	0.7832	0.7874
Spiral Arm Count	1997	0.7742	0.7555	0.7742	0.7560
Merging	11346	0.8798	0.8672	0.8798	0.8511

(a) Classification metrics for all galaxies

Question	Count	Accuracy	Precision[a]	Recall[b]	F1[c]
Smooth Or Featured	3495	0.9997	0.9997	0.9997	0.9997
Disk Edge On	3480	0.9980	0.9980	0.9980	0.9980
Has Spiral Arms	2024	0.9921	0.9933	0.9921	0.9924
Bar	543	0.9945	0.9964	0.9945	0.9951
Bulge Size	237	1.0000	1.0000	1.0000	1.0000
How Rounded	3774	0.9968	0.9968	0.9968	0.9968
Edge On Bulge	258	0.9961	0.9961	0.9961	0.9961
Spiral Winding	213	0.9906	1.0000	0.9906	0.9953
Spiral Arm Count	659	0.9863	0.9891	0.9863	0.9871
Merging	3108	0.9987	0.9987	0.9987	0.9987

(b) Classification metrics for galaxies where volunteers are confident

[a]The proportion of predicted positives that are actually positive $\frac{TP}{TP+FP}$

[b]The proportion of positives predicted as such $\frac{TP}{TP+FN}$

[c]The harmonic mean of precision and recall $\frac{TP}{TP+0.5(FP+FN)}$, acknowledging the tradeoff between completeness and contamination

Table 4.1: Classification metrics on all galaxies (above) or on galaxies where volunteers are confident for that question (i.e. where one answer has a vote fraction above 0.8). Multi-class precision, recall and F1 scores are weighted by the number of true galaxies for each answer. Classifications on confident galaxies are near-perfect.

Figure 4.14 shows the mean deviations between the model predictions (mean posteriors) and the observed vote fractions, by question, for retired test set galaxies. Performance is again excellent, with the predictions typically well within 10% of the observed vote fractions. Predicting spiral arm count is relatively challenging, as noted above. Predicting answers to the 'Merger' question of 'None' (i.e. not a merger) is also challenging, perhaps because of the rarity of counter-examples.

Even the volunteer vote fractions are themselves somewhat uncertain for many galaxies. The ultimate aim is to predict the true vote fraction, i.e. the vote fraction from $\lim_{N\to\infty}$ volunteers, but one can only even measure the vote fraction from N volunteers. However, 387 pre-active-learning galaxies were erroneously uploaded twice or more, and so received more than 75 classifications each. I therefore compared the predictions against only these confidently-known galaxies. Specifically, I calculated the deviations from asking fewer ($N << 75$) volunteers by artificially truncating the number of votes collected. I can then ask 'how many volunteer responses to that question would we need to have errors similar to that of our model?' Figure 4.15 shows the model and volunteer deviations for a representative selection of questions; the model predictions are as accurate as asking that question to around 10 volunteers [9]. The actual number of volunteers needed to be shown that galaxy to achieve equivalent accuracy will be higher for questions only asked given certain previous answers (i.e. all but 'Smooth or Featured?' and 'Merger?'), as some will give different answers to preceding questions and so not be asked that question.

I also measured if the posteriors correctly estimate this uncertainty. As a qualitative test, Figure 4.17 shows a random selection of galaxies binned by 'Smooth or Featured' vote fraction prediction entropy, measuring the model's uncertainty. Prediction entropy was calculated as the (discrete) Shannon entropy over all possible combinations of votes, assuming 10 total votes for this question (the results are robust to other choices of total votes). Unusual, inclined or poorly-scaled galaxies have highly uncertain (high entropy) votes, while smooth and especially clearly featured galaxies have confident (low entropy) votes. The most uncertain galaxies (not shown) are so poorly scaled (due to incorrect estimation of the Petrosian radius in the NASA-Sloan Atlas) that they are barely visible. These are the galaxies I would intuitively expect to be uncertain.

More quantitatively, Figure 4.18 shows the calibration of our posteriors for the two binary questions in GZD-5 - 'Edge-on Disk' and 'Has Spiral Arms'. A well-calibrated posterior dominated by data (i.e. where the prior has a minimal effect) will include

[9] The model is, in this strict sense, slightly superhuman.

the measured value within any bounds as often as the total probability within those bounds. I calculated calibration by, for each galaxy, iterating through each symmetric highest posterior density credible interval (i.e. starting from the posterior peak and moving the bounds outwards) and recording both the total probability inside the bounds and whether the recorded volunteer vote is inside the bounds. I then grouped (binned) by total probability and recorded the empirical frequency with which the votes lie within bounds of that total probability. Calibration is found to be excellent. Our classifier is correctly uncertain.

Such confident galaxies are expected to have a clearly correct label, making correct and incorrect predictions straightforward to measure but also making the classification task easier.

The ultimate measure of success is whether the predictions are useful for science. Masters et al. 2019 [295] (hereafter M19) used GZ2 classifications to investigate the relationship between bulge size and winding angle and found - contrary to a conventional view of the Hubble sequence - no strong correlation. I repeated this analysis using our (deeper) DECaLS data, using either volunteer or automated classification, to check if the automated classifications lead to the same science results as the volunteers.

Specifically, I selected a clean sample of face-on spiral galaxies using M19's vote fraction cuts of $f_{\text{feat}} > 0.43$, $f_{\text{not-edge-on}} > 0.715$, and $f_{\text{spiral-yes}} > 0.619$. I also made a cut of $f_{\text{merging=none}} > 0.5$, analogous to M19's f_{odd} cut, to remove galaxies with ongoing mergers or with otherwise disturbed features. For the volunteer vote fractions, I could have used either GZD-1/2 or GZD-5 classifications, since the former decision tree had three bulge size answers and the latter had five; I chose GZD-5 to benefit from the added precision of additional answers. I only used galaxies classified prior to active learning being activated. I used the same automated classifications with which the other results in this section are calculated; predictions from an ensemble of three models, each making five dropout forward passes. This expands the sample size from 5,378 galaxies using GZD-5 volunteers only to 43,672 galaxies using automated classification[10].

I calculated bulge size and spiral winding following Eqn. 1 and 3 in M19, trivially generalising the bulge size calculation to allow for five bulge size answers:

[10]Selected with the requirements above from all 313,798 GZ DECaLS galaxies

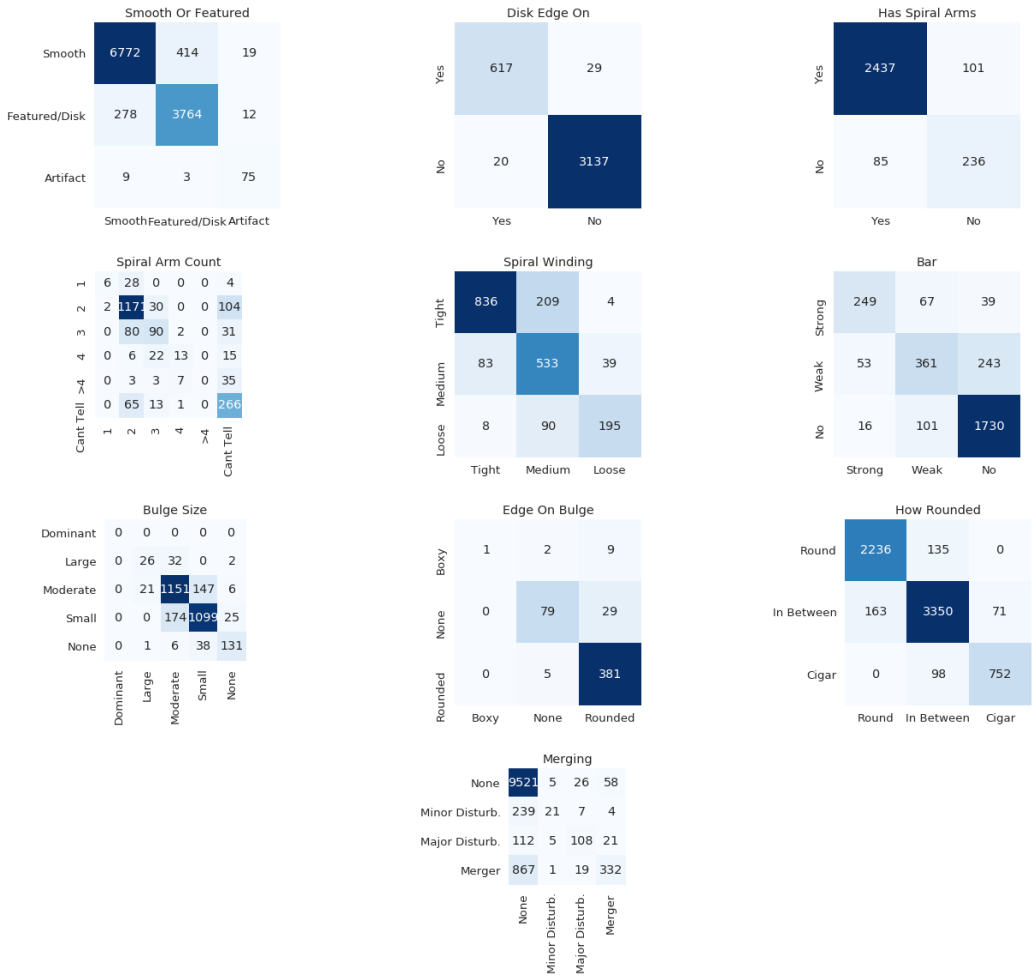


Figure 4.12: Confusion matrices for each question, made on the test set of 11,346 galaxies in the (random) test set with at least 34 votes. Discrete classifications are made by rounding the vote fraction (label) and mean posterior (prediction) to the nearest integer. The matrices then show the counts of rounded predictions (x axis) against rounded labels (y axis). To avoid the loss of information from rounding, we encourage researchers not to treat Galaxy Zoo classifications as discrete, and instead to use the full vote fractions or posteriors where possible.

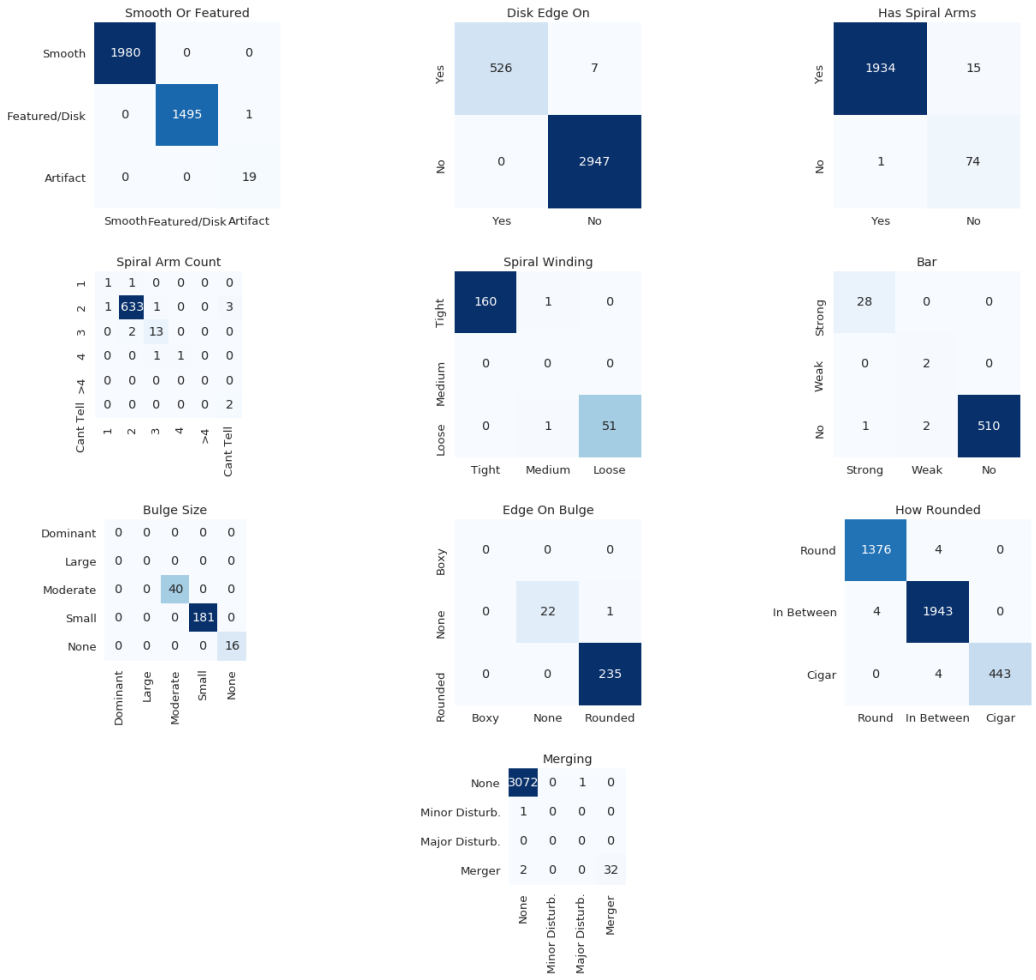


Figure 4.13: Confusion matrices for test set galaxies where the volunteers are confident in that question, defined as having the vote fraction for one answer above 0.8. Such confident galaxies are expected to have a clearly correct label, making correct and incorrect predictions straightforward to measure but also making the classification task easier. To avoid the loss of information from rounding, we encourage researchers not to treat Galaxy Zoo classifications as discrete, and instead to use the full vote fractions or posteriors where possible.

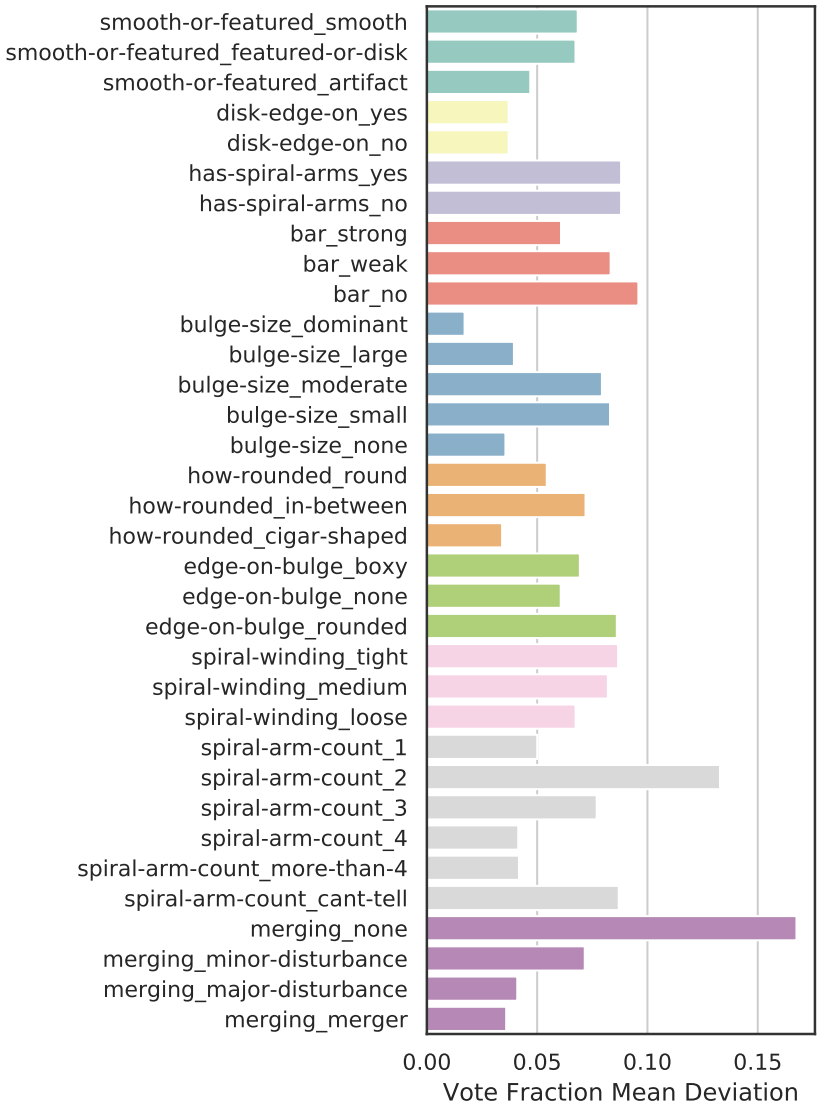


Figure 4.14: Mean absolute deviations between the model predictions and the observed vote fractions, by question, for the retired test set galaxies. The model is typically well within 10% of the observed vote fractions.

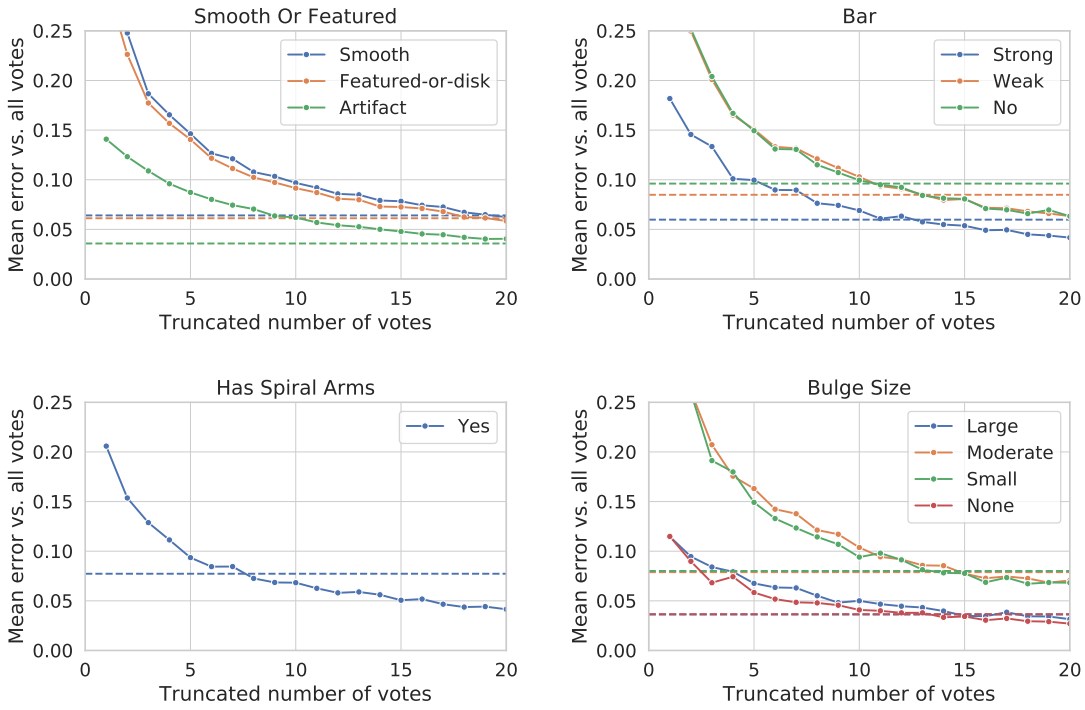


Figure 4.15: Mean errors vs. the true ($N > 75$) vote fractions for either a truncated ($N = 0$ to $N = 20$) number of volunteers (solid) or the automated classifier (dashed). Asking only a few volunteers gives a noisy estimate of the true vote fraction. Asking more volunteers reduces this noise. For some number of volunteers, the noise in the vote fraction is similar to the error of the automated classifier, meaning they make classifications of similar accuracy; this number is where the solid and dashed lines intersect. We find the automated classifier has a similar accuracy to approx. 5 to 15 volunteers, depending on the question.

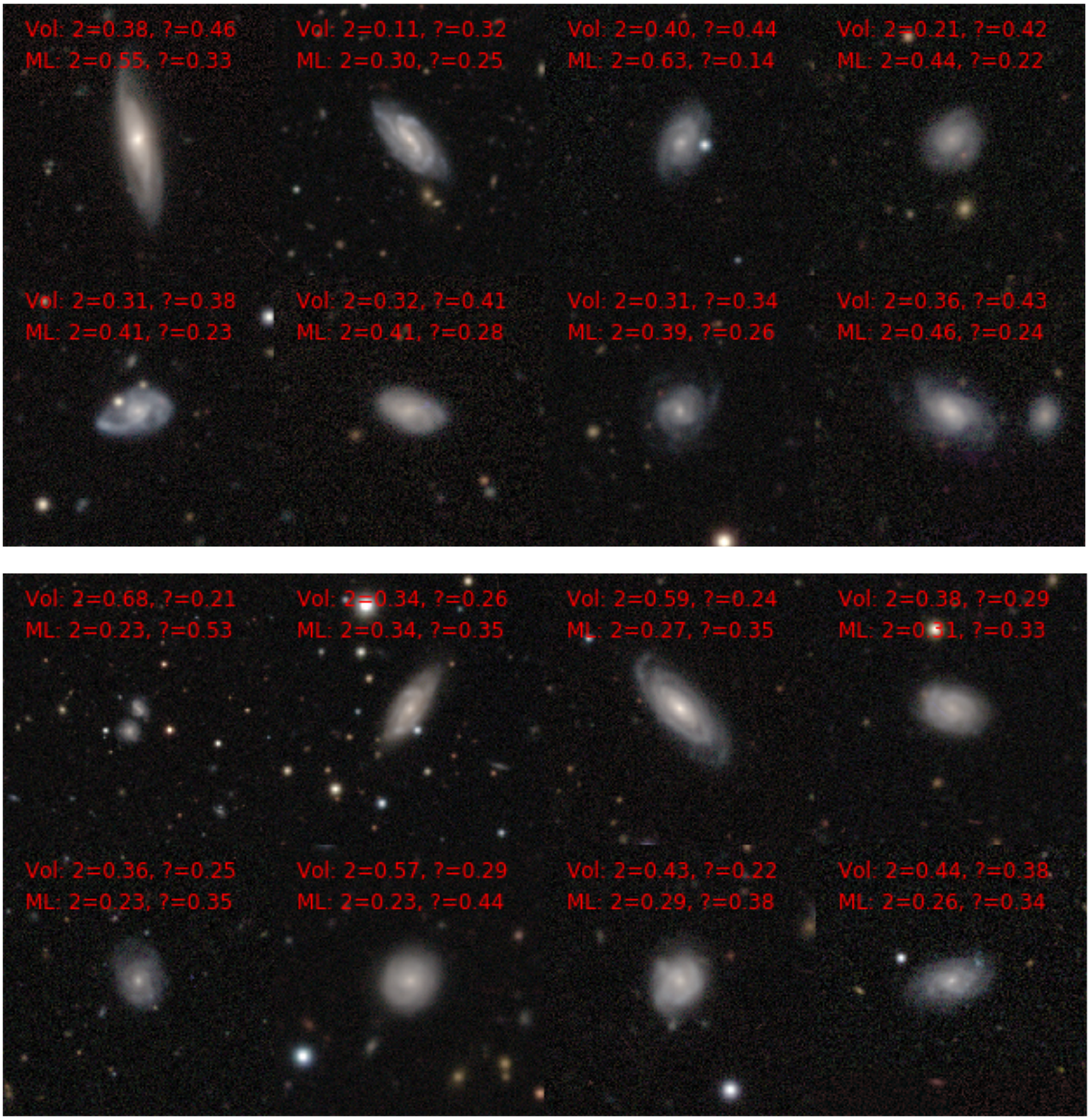


Figure 4.16: Random spiral galaxies where the classifier confuses the most likely volunteer vote for spiral arm count between '2' and 'Can't Tell'. Above: galaxies where the classifier predicted '2' but more volunteers answered 'Can't Tell'. Below: vice versa, galaxies where the classifier predicted 'Can't Tell' but more volunteers answered '2'. Red text shows the volunteer (vol.) and machine-learning-predicted (ML) vote fractions for each answer. Counting the spiral arms is challenging, even for the authors. This highlights the difficulty in assessing performance by reducing the posteriors to classifications and then comparing against uncertain true labels.

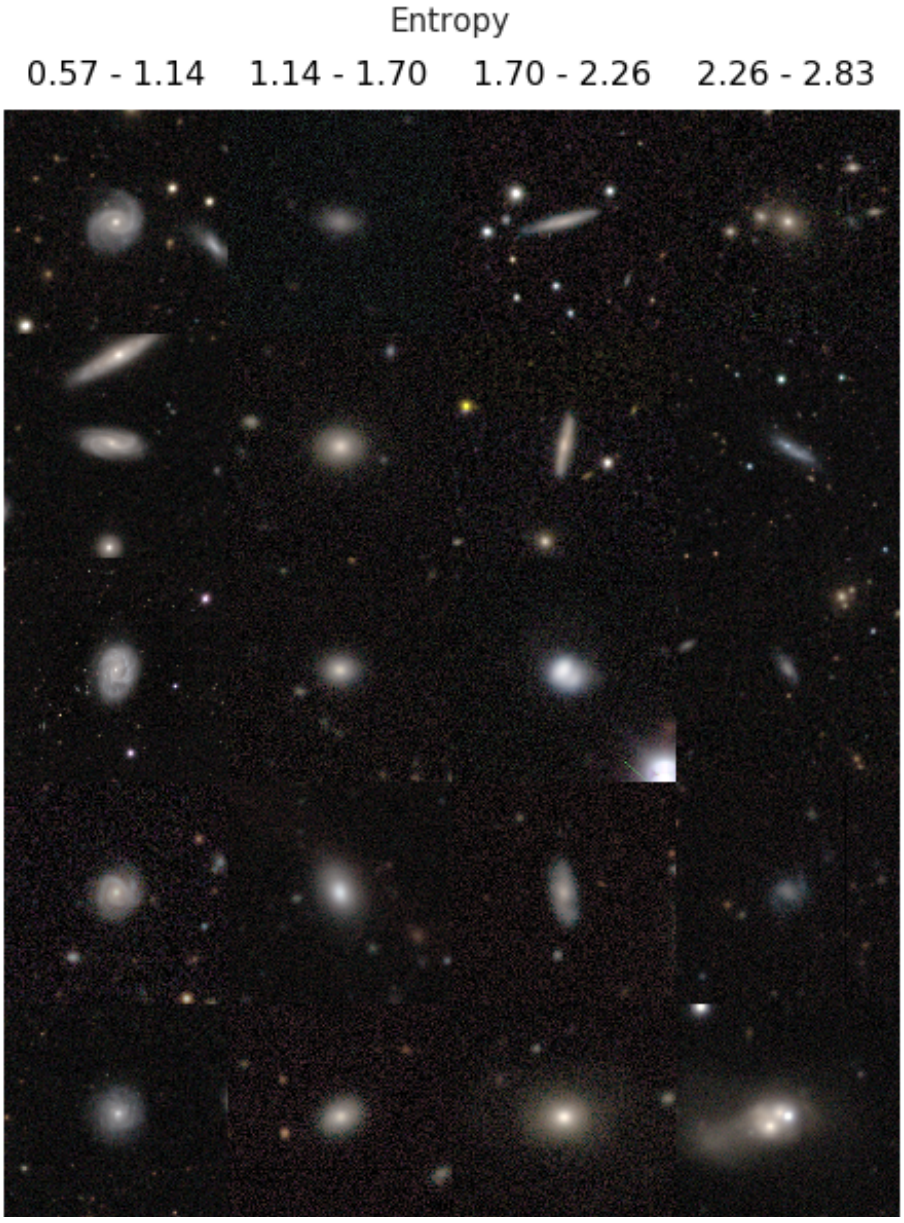


Figure 4.17: Galaxies binned by 'Smooth or Featured' vote prediction entropy, measuring the model's uncertainty in the votes. Bins (columns) are equally spaced (boundaries noted above). Five random galaxies are shown per bin. Unusual, inclined or poorly-scaled galaxies have highly uncertain (high entropy) votes, while smooth and especially clearly featured galaxies have confident (low entropy) votes, matching our intuition and demonstrating that our posteriors provide meaningful uncertainties.

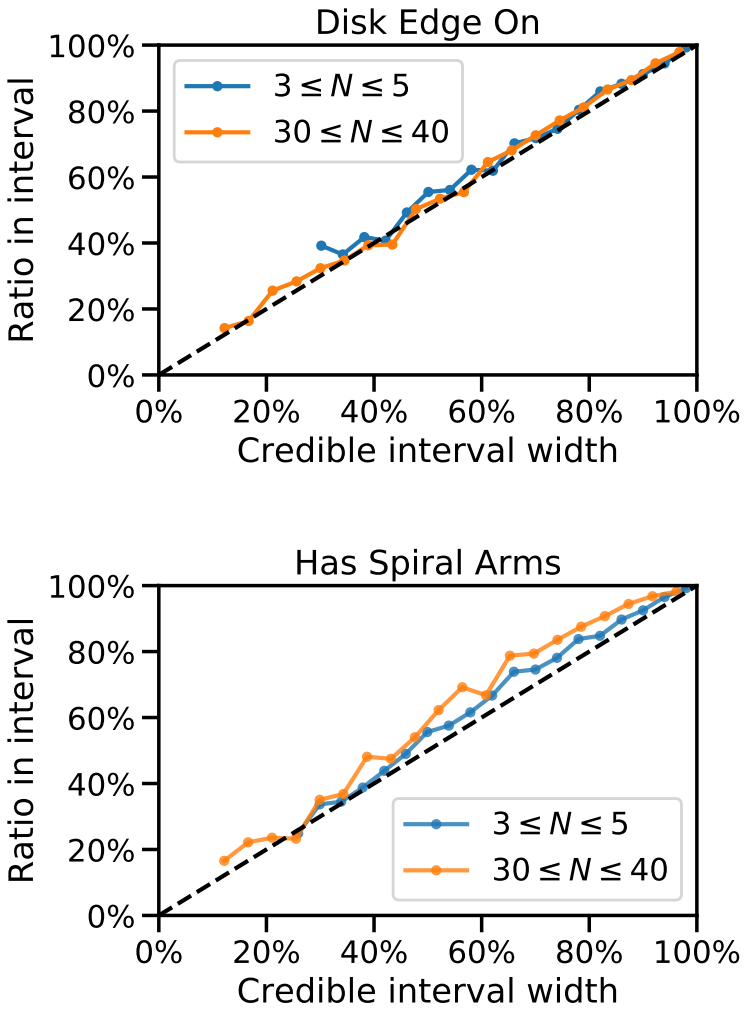


Figure 4.18: Calibration curves for the two binary GZ DECaLS questions. The x-axis shows the credible interval width - for data-dominated posteriors, roughly (e.g.) 30% of galaxies should have vote fractions within their 30% credible interval. The y-axis shows what percentage actually do fall within each interval width. We split calibration by galaxies with few votes (and hence typically wider posteriors) and more votes (narrower posteriors). Only credible interval bins with at least 100 galaxies are shown. Calibration for both questions is excellent.

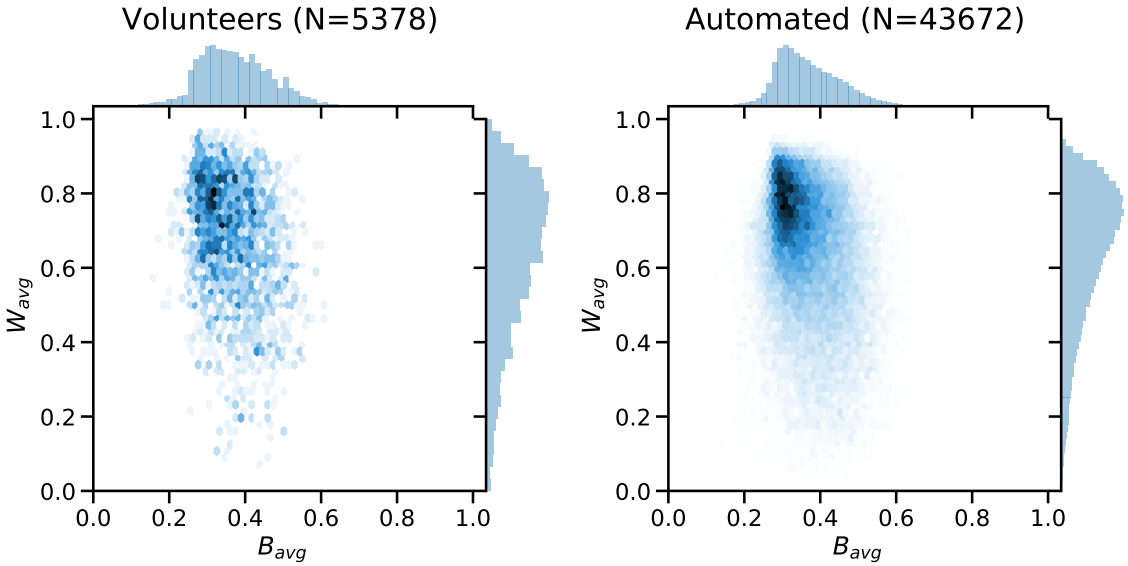


Figure 4.19: Distribution of bulge size vs. spiral winding, using responses from volunteers (left) or our automated predictions (right). We observe no clear correlation between bulge size and spiral winding, consistent with M19. The distributions are consistent between volunteers and our automated method. We hope this demonstrates the accuracy and scientific value of our automated classifier.

$$W_{\text{avg}} = 0.5 f_{\text{medium}} + 1.0 f_{\text{tight}} \tag{4.6}$$

$$B_{\text{avg}} = 0.25 f_{\text{small}} + 0.5 f_{\text{moderate}} + 0.75 f_{\text{large}} + 1.0 f_{\text{dominant}} \tag{4.7}$$

Both human and automated classification methods found no correlation between bulge size and spiral winding, consistent with M19. Figure 4.19 shows the distribution of bulge size against spiral winding using either volunteer predictions (fractions) or the deep learning predictions (expected fractions) for the sample of featured face-on galaxies selected above. The distributions are indistinguishable, with the automated method offering a substantially larger (approx 8x) sample size. I hope this demonstrates the accuracy and scientific value of my automated classification approach.

4.6 Usage

4.6.1 Catalogues

The Galaxy Zoo DECaLS data release includes two volunteer catalogues and two automated catalogues, available on Zenodo at https://doi.org/10.5281/zenodo.4196266

(along with the galaxy images).

`gz_decals_volunteers_ab` includes the volunteer classifications for 92,960 galaxies from GZD-1 and GZD-2. Classifications were made using the GZD-1/2 decision tree. All galaxies received at least 40 classifications, and consequently have approximately 30-40 after volunteer weighting (Sec. 4.4.3). This catalogue is ideal for researchers needing standard morphology measurements on a reasonably large sample, with minimal complexity. 33,124 galaxies in this catalogue were also previously classified in GZ2; the GZD-1/2 classifications are better able to detect faint features due to deeper DECaLS imaging, and so should be preferred.

`gz_decals_volunteers_c` includes the volunteer classifications from GZD-5. Classifications are made using the improved GZD-5 decision tree which adds more detail for bars and mergers (Sec. 4.4.2). This catalogue includes 253,286 galaxies, but each galaxy does not have the same number of classifications. 59,337 galaxies have at least 30 classifications (after denoising), and the remainder have far fewer (approximately 5). This catalogue may be useful to researchers who prefer a larger sample than `gz_decals_volunteers_ab` at the cost of more uncertainty and the introduction of selection effects, or who need detailed bar or merger measurements for a small number of galaxies.

The automated classifications are made using the Bayesian deep learning classifier, trained on `gz_decals_volunteers_c` to predict the answers to the GZD-5 decision tree for all GZ DECaLS galaxies (including those in GZD-1 and GZD-2). `gz_decals_auto_posteriors` contains the predicted posteriors for each answer - specifically, the Dirichlet concentration parameters that encode the posteriors. I hope this catalogue will be helpful to researchers analysing galaxies in Bayesian frameworks similar to those used to great effect in cosmology.

`gz_decals_auto_fractions` reduces those posteriors to the automated equivalent of previous Galaxy Zoo data releases, containing the expected vote fractions (mean posteriors). Note that not all vote fractions are relevant for every galaxy; I suggest assessing relevance using the estimated fraction of volunteers that would have been asked each question, which I also include. I hope this catalogue will be useful to researchers seeking detailed morphology classifications on the largest possible sample, who might benefit from error bars but do not need full posteriors.

The automated classifications may be interactively explored at https://share.streamlit.io/mwalm
poster/gz_decals_mike_walmsley.py.

4.7 Discussion

What does a classification mean? The comparison of GZ2 and GZ DECaLS images (Fig. 4.5) highlights that Galaxy Zoo classifications aim to characterise the clear features of an image, and not what an expert might infer from that image. For example, volunteers might see an image of a galaxy that is broadly smooth, and so answer smooth, even though our astronomical understanding might suggest that the faint features around the galaxy core are likely indicative of spiral arms that would be revealed given deeper images. This situation occurs in several galaxies in Fig. 4.5. These 'raw' classifications will be most appropriate for researchers working on computer vision or on particularly low-redshift, well-resolved galaxies. I hope that publishing the original images alongside the data release (https://doi.org/10.5281/zenodo.4196266) will support such computer vision work. The redshift-debiased classifications, which are effectively an estimate of galaxy features *not clearly seen* in the image, will be most appropriate for researchers especially interested in fainter features or studying links between our estimated intrinsic visual morphologies and other galaxy properties.

I showed in Sec. 4.4.2 that changing the answers available to volunteers significantly improves our ability to identify weak bars. This highlights how classifications are only defined in the context of the answers presented. One cannot straightforwardly compare classifications made using different decision trees. Our scientific interests and our understanding of volunteers both evolve, and so the Galaxy Zoo decision trees must also evolve to match them. However, only the last few years of volunteer classifications will use the latest decision tree (based on previous data releases), placing an upper limit on the number of galaxies with compatible classifications at any one time. Our automated classifier resolves this here by allowing us to retrospectively apply the GZD-5 decision tree (with better weak bar detection, among other changes) to galaxies only classified by volunteers in GZD-1 and GZD-2. This flexibility ensures that Galaxy Zoo will remain able to answer the most pertinent research questions at scale.

I have shown (4.5.2) that our automated classifier is generally highly accurate, well-calibrated, and leads to at least one equivalent science result. However, we cannot exclude the possibility of unexpected systematic biases or of adversarial behaviour from particular images. Avoiding subtle biases and detecting overconfidence on out-of-distribution data remain open computer science research questions, often driven by important terrestrial applications [120, 145, 168, 287, 355, 394, 409, 465]. Volunteers also have biases (e.g a slight preference for recognising left-handed spirals, [245])

and struggle with images of an adversarial nature (e.g. confusing edge-on disks with cigar-shaped ellipticals), though these can often be discovered and resolved through discussion with the community and by adapting the website. I discuss this further in the main Conclusion (Chapter 7).

4.8 Future of Morphology Classification

This chapter presented an automated classifier that predicts the majority volunteer response with near-perfect accuracy [11] when considering only galaxies where the volunteers are confident. Any further raw progress in automated classification (from better architectures, etc.) must therefore show improved performance on galaxies where volunteers are *not* confident. But when the volunteers are not confident, we do not know which prediction is correct - so how can we measure performance? I have mitigated this issue using the vote fractions instead of majority response and using galaxies with unusually many ($N > 75$) volunteers (Sec. 4.5.2). However, these do not fully address the fundamental problem. I believe that the limiting factor in automatic classifications is now our statistical knowledge of the volunteers. Extensive work in citizen science [172, 441] and even specifically with Galaxy Zoo [392] shows that statistical models of individual volunteers can dramatically improve the reliability of the final aggregated measurements, or at least better quantify the uncertainty of those measurements. Such statistical models could be combined with representations of each galaxy extracted from deep learning models to understand how volunteers react to different galaxy features - to recognise that a volunteer may often confuse rounded with boxy bulges but be excellent at spotting weak bars, for example. These models could also be straightforwardly extended to include redshift debiasing in a principled manner, rather than as an ad-hoc matching of distributions.

Galaxy Zoo aims to provide detailed morphology classifications suitable for common science cases and has been extensively used as such. Galaxy Zoo is also often the basis for developing automated classification methods, perhaps because of the large size of the labelled dataset, the breadth of available questions, and the continuing influence of the 2015 Kaggle competition [100]. However, there are many other projects attempting to solve important morphology classification problems not addressed by Galaxy Zoo. I would therefore like to review the progress in automated morphology classification more broadly, and then consider how my work with Galaxy Zoo and the methods I presented in the previous chapters might best fit into a larger system.

[11] The classifier achieves 99% accuracy for every question - see Sec. 4.5.2

Many deep learning tools have been used to interpret galaxy images. Self-supervised methods, where a generative model learns to predict existing data without the need for external labels, appears a natural fit for galaxy morphology because examples of galaxies are plentiful. Being able to exploit far more data helps models learn more sophisticated representations, a crucial benefit I return to shortly. Generative models with either an explicit likelihood (e.g. pixelCNN, [435, 436]) or methods to estimate a likelihood (e.g. generative adversarial networks, [150]) are well-suited to detecting anomalies[12]; images which are statistically unlike those seen before. Both methods have been successfully applied to galaxy images [287, 354, 374, 406, 469]. However, because self-supervised models learn representations based only on the image content, those representations will include features that affect the image but are not astrophysically meaningful. For example, Spindler et al. 2020 [401] successfully trains a variational autoencoder to generate SDSS-like galaxy images, but find that a classification schema derived from the autoencoder's representations places undue importance on foreground/background companion objects. This lack of semantic understanding poses a problem for purely self-supervised classification approaches. Instead, much classification work - including these past chapters - has focused on supervised approaches. I have already summarised the successes of supervised galaxy classification (3.1.1) and introduced several advances to use human labels more efficiently (Sections 3.1.3, 3.2, 4.5) However, it may possible to combine both supervised and self-supervised approaches.

Recent empirical research suggests that the performance of deep natural language models with Transformer architectures [440] follows fundamental scaling relations. Broadly speaking, performance increases approximately as a power law with respect to either the number of model parameters, the size of the training dataset, or the computational budget, provided the two fixed variables are sufficient [210]. For example, increasing the number of model parameters will likely increase performance provided one has access to effectively unlimited data and compute. Most researchers have neither, and so the best-performing models are increasingly created by a few well-resourced groups such as OpenAI (in partnership with Microsoft) [59] and Google Brain [123]. These natural language models are trained to predict missing words in sentences (along with related tasks) and so effectively all digitised writing is potentially useful training data. Having learned an effective representation of language, the models can then be fine-tuned on so-called domain tasks: tasks of practical interest

[12]More formally, out-of-distribution data

such as summarising news articles, coding websites, or writing poetry. Crucially, because the fundamental language representation is already learned, fine-tuning requires far more modest data and compute.

Could such an approach work for galaxy morphology? CNN probably follow similar scaling laws [383] as do vision-based Transformers [169]. Recent work by Hayat et al. 2020 [161] uses self-supervised learning to learn galaxy representations explicitly for generic downstream tasks. Fine-tuning in a supervised context has been extensively used to make galaxy-related predictions on new surveys using models pretrained on ImageNet [6, 289, 461] or other surveys [105, 279, 339, 416] If we combine all these ideas, we can have the best of every method. To summarise the strategy:

1. Use contrastive learning to train a self-supervised model on large galaxy survey. Invest the time and computation to use the plentiful data to train the largest feasible model.

2. Gradually retrain that model to predict Galaxy Zoo votes for all questions using the methods I have introduced, allowing the representation (weights) to evolve. By using Galaxy Zoo votes, we encourage the representation to be physically meaningful for a broad range of tasks.

3. Share the model with the community. Other researchers can then create very effective models fine-tuned for their specific galaxy morphology tasks using minimal data (perhaps on the order of hundreds of examples).

The first two steps could be done combined using the 'noisy student' approach presented in Xie et al. 2020 [463].

I believe the future of morphology classification is in such thoughtful combination of volunteers and probabilistic deep learning. These combinations will be more than just faster; they will be replicable, uniform, error-bounded, and quick to adapt to new tasks. They will let us ask new questions - draw the spiral arms, select the bar length, separate the merging galaxies pixelwise - which would be infeasible with volunteers alone for all but the smallest samples (e.g. [264]). And they will find the interesting, unusual and unexpected galaxies which challenge our understanding and inspire new research directions.

Chapter 5

Finding Faint Fast Radio Bursts with CHIME

5.1 Introducton

In the preceding two chapters (3-4), I showed how machine learning could be combined with citizen science to efficiently classify large galaxy samples. Active learning was key to this approach; since not all galaxies can be labelled, one should label the galaxies that would be most informative for the machine learning model to learn from. But what if we can't even record all the data to label in the first place? Such a situation is common in radio transient astronomy, where the raw data rates often vastly exceed any realistic hope of storage. In this chapter, I attempt to address such a problem: identifying faint fast radio bursts in data that would otherwise be discarded.

Fast radio bursts (FRBs) are as-yet-unexplained extragalactic signals with pulsar-like profiles. The first identified example was the Lorimer burst, found in an archival search of Parkes data by Lorimer et al. (2007) . The Lorimer burst was remarkable for being both extremely bright (it saturated the primary Parkes receiver) and distant (with a dispersion measure[1] of 375 pc cm^{-}3). While initially controversial, the Lorimer burst was established as belonging to a broader population by subsequent detections of similar extragalactic bursts such as those reported by Thornton et al. (2013) [425]. At the time of writing, approximately 285 such bursts have been identified and published[2]. 19 have been observed to repeat [19, 20, 131, 402]; whether these represent a distinct population is an intriguing open question. The geographic

[1]Free electrons between source and observer cause radio signals to arrive with a time delay proportional to ν^{-2}. Dispersion measure is the integrated free electron density between source and observer inferred from the measured delays.

[2]An up-to-date list can be found at https://www.wis-tns.org/, based on work by Petroff et al. 2016 [342]

diversity of detectors (Parkes, Green Bank, Arecibo, ASKAP, UTMOST, etc.) rules out the possibility of Peryton-like false positives [269]. Localisations of various bursts to host galaxies [19, 20, 40, 131, 165, 220] confirms the bursts are of extragalactic origin. However, beyond their distant existence, much remains unknown.

Many theories for their origin have been proposed [215] including compact mergers (e.g. neutron stars [429]), neutron star collapse [121], and 'artificial beams for driving light sails' [263]. Magnetar flares are currently the leading candidate (see e.g. [278, 338]). The short timescale and significantly (pulsar-like) polarized pulses suggest the sources are compact objects with strong magnetic fields, and the repeating nature of some bursts require a non-cataclysmic origin [215]. Whether magnetars can generate the extreme energies implied by the cosmological distances of FRBs was a major barrier for the theory until recently. In April 2020, CHIME and STARE2 detected an FRB-like signal from galactic magnetar SGR 1935+2154 [21, 47]. The evidence that SGR 1935+2154 was the source is compelling; the source was localised to within 1 deg of SGR 1935+2154, the CHIME-estimated dispersion measure (approximately 330 pc cm^{-}3) is within the Milky Way and consistent with SGR 1935+2154, and the radio detection was coincident (1ms) with independent gamma and X-ray detections matching in time and dispersion [300, 311]. The detected radio fluence implied a burst energy of $10^{34}-10^{35}$ ergs, three orders of magnitude brighter than previously observed for magnetars and approaching that of typical FRB (the closest FRB with well-determined distances, FRB 181030.J1054+736 and FRB 141113, had approximately $10^{36} - 10^{38}$ ergs). The discovery of fainter FRB might close the remaining gap; however, as Andersen et al. 2020 [21] writes, 'the detection of these faint FRB is limited by the sensitivity of our instruments' . Identifying such faint FRB is a major goal of my work.

Identifying new FRB can also help constrain theories. Comparing the rates and host galaxies of FRB with the rates and hosts predicted by each theory is one approach. For example, the host galaxies of the first four bursts localised by ASKAP [40] disfavour SLSNe magnetars[3], but core-collapse magnetars 'remains plausible'. New large samples of FRB may also reveal if repeating FRB share the same origin as FRB not observed to repeat. The repeating pulses detected to date show statistically significant $(4-5\sigma)$ increases in burst widths, lending credence to the idea that repeaters may have a different origin. The distribution of dispersion measures, in contrast, is

[3]90% of low-z SLSNe magnetars lie on the starforming main sequence, but the ASKAP FRB do not [40].

entirely consistent. Much larger samples should establish which burst features, if any, are definitively distinct [343].

Fast radio bursts are also useful for cosmology as probes of ionized baryons. The dispersion of fast radio bursts as a function of redshift, DM(z), directly depends on the electron column density (and hence ionized baryon density) of the IGM. McQuinn 2014 [299] argues that the scatter in DM(z) between different sightlines (sources) is driven by the number of collapsed halos each sightline encounters, as diffuse ionised baryons will make similar contributions for any sightline. The scatter of the fast radio burst DM(z) distribution should therefore measure the proportion of ionised baryons in collapsed halos versus the diffuse IGM.

The distribution of dispersion measures p(DM) is itself useful for measuring helium reionization, with no redshifts required. Reionization occurs at approximately $z \approx 3$, and so bursts above that redshift are able to probe reionization [41, 262]; these correspond to dispersions of around 3000-5000 pc cm^{-}3, around twice the highest published FRB dispersion [131]. Finding high-dispersion FRBs is directly connected to finding faint FRBs, and so finding high-dispersion FRBs is similarly a major goal of this chapter.

Unfortunately, the scarcity of published detections restricts our ability to constrain source theories or to probe cosmic baryon distributions. This will change. Considering the narrow sky coverage of the instruments currently used to make detections, detecting 285 events to date implies that FRB must be common: over the full sky, one expects approximately $10^3 - 10^4$ FRB per day with a fluence of > 1 Js ms [346]. Wide-area surveys are therefore likely to dramatically increase the number of detections.

CHIME (Canada Hydrogen Intensity Mapping Experiment) is one such wide-area survey, operating at 0.4 - 0.8GHz [18]. Originally designed to map hydrogen density, CHIME has been extended[4] to include a synchronous scan for FRBs. Since starting operations in 2018, CHIME has discovered 18 of 19 published repeating FRBs [19, 20, 131] (all but the first) as well as at least 700 unpublished single pulse FRBs [131].

The CHIME instrument paper [18] states that CHIME archives $\sigma \geq 10$ events due to data storage constraints and the time required to manually review candidates. Such signals are identified by calculating the signal-to-noise at each leaf of a tree search through possible dispersion measures [18]. Assuming FRB are approximately uniformly-distributed standard candles, back-of-the-envelope calculations I introduce

[4]in the literal sense - with a shielded shipping container of computing hardware

below (Sec. 5.3) suggest that slightly fainter FRB are likely to be common. However, CHIME does not currently have algorithms to search for FRBs in such low σ events, ruling out detection of faint FRBs by definition. I wrote a proposal to CHIME to attempt to address this.

5.2 Proposal

I wrote a proposal to CHIME suggesting that a combination of citizen science and machine learning could identify likely FRBs within the low σ candidates that would otherwise be discarded. I suggested the following approach.

1. Lower the CHIME trigger threshold to collect a portion of $\sigma < 10$ candidates
2. Rank the most promising candidates using machine learning
3. Present these promising candidates to volunteers through the Zooniverse citizen science platform to be labelled as FRB-like or not
4. Refer the best volunteer-labelled candidates for expert internal review by CHIME
5. Use the new volunteer labels to continually improve (train) the ranking algorithms, which could ultimately be used as a baseband trigger for CHIME

In the proposed Zooniverse project, volunteers would be asked to judge if spectrograms of candidate signals detected by CHIME contain possible FRBs or just RFI. I was encouraged to suggest this approach by the past successes of Zooniverse volunteers at identifying transients in complex and abstract data. In the LIGO collaboration project Gravity Spy [471], volunteers divide interferometer spectrograms into different classes of noise, helping LIGO identify and address the sources. And in the Planet Hunters projects [126] volunteers have discovered exoplanets from lightcurves - most recently, 90 promising candidates found in data from NASA's Transiting Exoplanet Survey Satellite TESS [112, 113].

I suggested using machine learning to prioritise candidates because this allows limited volunteer effort to be focused on the candidates most likely to be FRB. The trained algorithms could ultimately be run commensally at CHIME, allowing CHIME to save full baseband data for candidates the volunteers would have identified as promising. I discuss the extensive benefits of such data in 5.6.

To avoid the need for simplistic pulse/noise models, leading to failures on more complex observations, one would like to learn representations of FRB and RFI directly

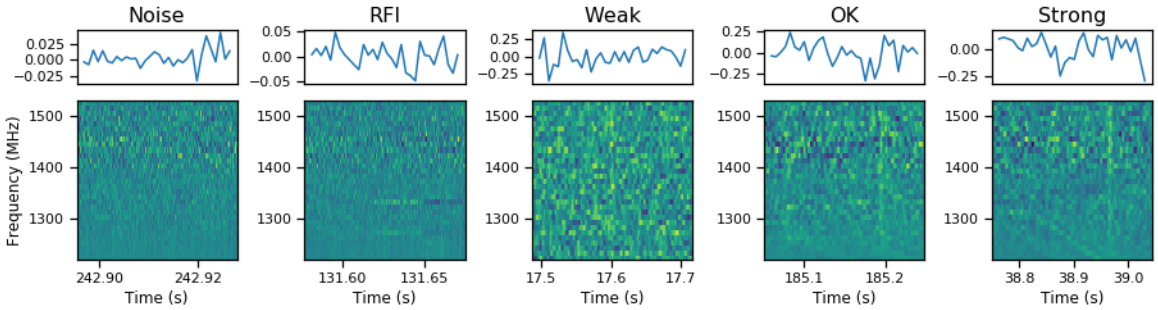


Figure 5.1: PALFA candidate events were labelled by experts as Noise, RFI, Weak, OK, or Strong. Here, I show representative examples of each class. The prototype CNN aimed to rank promising (Strong/OK/Weak) candidates highly, for hypothetical referral to experts or volunteers.

from the data. Learning complex representations from high-dimensional data is the key advantage of deep learning and (for image-like data) of convolutional neural networks (CNN) in particular, as I have already discussed (Sec. 1.2). CNNs have been recently shown to be highly effective at detecting simulated FRB [82]), re-detecting previously-known FRB [10], and at finding new pulses in focused observations of a known repeater [472]. It therefore seemed plausible that CNN might be effective on CHIME FRB observations.

The CNN would need to remain at least somewhat effective at $\sigma < 10$. To verify this and to provide a ballpark performance estimate for my proposal, I trained a CNN to rank FRB candidates observed by PALFA [332] with $\sigma \geq 7$ (and DM 300-3000 pc cm^{-3}). PALFA was a pulsar survey at Arecibo (single 305m dish) exploring the galactic plane ($|b| > 5$) at 1.4GHz [88]. PALFA collected millions of candidate signals since 2004, the overwhelming majority of which are RFI. Previous analysis searching for single pulses identified a single non-repeating probable FRB, in line with statistical expectations from sky coverage and detection limits [332]. In the course of that search, PALFA researchers visually inspected $\approx$ 4000 candidate signals and labelled them by astrophysical plausibility: Strong, OK, Weak, RFI, or Noise.

I trained a standard 'off-the-shelf' CNN (using the same architecture as in Chapter 3) to classify candidates as either promising (Strong/OK/Weak) or not (RFI/Noise). This CNN could then be used to rank candidates for hypothetical review by citizen scientists or experts. Figure 5.2 shows the CNN candidate ranking of the candidates. The prototype CNN is shown to be highly effective at prioritising Promising candidates, ranking 91% of Promising candidates in the top 10% of all candidates and 98% in the top 20%. This suggests that even standard CNN are capable of effectively prioritising lower σ candidates.

Figure 5.2: Ranking of candidates by our prototype 'off-the-shelf' CNN. Each candidate is represented by a vertical line. Promising (Strong/OK/Weak) candidates are shown in red, and other (RFI/Noise) candidates in grey. Percentages indicate relative ranking. For example, a red line at 88% indicates a Promising candidate ranked higher than 88% of other candidates. Our prototype CNN is highly effective at prioritising Promising candidates, ranking 91% of Promising candidates in the top 10% of all candidates and 98% in the top 20%.

5.3 Performance Estimate

In my proposed system, candidate signals with $\sigma < 10$ identified by the CHIME pipeline would be ranked by machine learning, and the most promising candidates reviewed by Zooniverse volunteers. To persuade CHIME the system was worthwhile, I estimated how low in σ could this system reach and how many additional FRB would be detected.

First, one needs to calculate how many candidates CHIME might observe at or above some fixed lower σ. This can be rephrased as how many signals must be reviewed to reach that σ, and hence how many aggregate volunteer classifications would be needed.

Patel et al. [332] Figure 7 shows the distribution of expert-classified RFI signals received by the PALFA survey, $N_{\mathrm{RFI}}(\sigma)$, down to $\sigma = 7$ (DM between 300 and 3000 pc cm^{-}3). Astrophysical signals are a small fraction of the total candidate signals and hence $N_{\mathrm{candidates}}(\sigma) \approx N_{\mathrm{RFI}}(\sigma)$. I used this distribution as a rough proxy for candidate signals received by CHIME, with the substantial caveat that PALFA covers a different frequency range and potentially has a different RFI environment. Inspection of Patel et al. [332] Figure 7 gives

$$N(\sigma) = A \times 10^{-0.42\sigma} \tag{5.1}$$

where the constant A sets the overall event rate and varies by survey according to field-of-view etc. I set A to match the approximate rate of candidate events at CHIME: 1000 events per day (365,000 per year) between $\sigma = 7$ and $\sigma = 10$ (CHIME Collaboration, private communication).

The total number of additional events recorded between the current threshold $\sigma_{current} \approx 10$ and a lower threshold σ_{low} is therefore

$$N_{\text{candidates}} = A \int_{\sigma_{low}}^{\sigma_{now}} 10^{-0.42\sigma} d\sigma \tag{5.2}$$

Zooniverse volunteers could provide N_{human} aggregated classifications of some chosen fraction f of the most promising candidates (discarding the rest). Trivially, the system could review $\frac{N_{\text{human}}}{f} = N_{\text{candidates}}$. Therefore,

$$N_{\text{human}} = fA \int_{\sigma_{low}}^{\sigma_{now}} 10^{-0.42\sigma} d\sigma \tag{5.3}$$

which can be solved to find $N_{\text{human}}(\sigma_{low})$, and then rephrased as $\sigma_{low}(N_{\text{human}})$, the lowest σ reached as a function of the number of aggregated human reviews (i.e. available volunteer effort). Figure 5.3 (left) shows $\sigma_{low}(N_{\text{human}})$ for several possible choices of f, the fraction of ML-ranked candidates for which to request volunteer classifications.

How many additional FRB would CHIME detect at or above that new lowest σ? Lawrence et al. 2017 [249] quotes the total expected count of Poisson-distributed events of fixed source brightness, uniformly distributed in local (Euclidean) space, with a fluence above S, as:

$$\Lambda(s) = Cs^{-\frac{3}{2}} \tag{5.4}$$

Holding the instrument and pulse properties fixed, fluence S is proportional to σ [332, 343]. The relative rate at of events at σ_{low} or higher vs. σ_{now} or higher is therefore expected to be

$$\frac{\Lambda(s_{\text{low}})}{\Lambda(s_{\text{now}})} = (\frac{s_{\text{now}}}{s_{\text{low}}})^{\frac{3}{2}} = (\frac{\sigma_{\text{now}}}{\sigma_{\text{low}}})^{\frac{3}{2}} \tag{5.5}$$

Figure 5.3 (right) shows the percentage increase in FRB rate (vs. $\sigma = 10$) which might therefore result as the σ_{low} cutoff is reduced.

Combining $\sigma_{low}(N_{\text{candidates}})$ (Eqn. 5.3) and the expected percentage increase in FRB rate as a function of σ (Eqn. 5.5), one can estimate the percentage increase in total FRB rate directly as a function of N_{human} aggregated reviews. However, not all of these additional FRB would be identified as such due to candidate classification errors by both automated and human classifiers. Figure 5.4 shows the percentage increase in *identified* FRB (vs. $\sigma = 10$) as estimated for perfect classifiers, poor classifiers (see caption), and the prototype CNN (Sec. 5.2), given several possible choices of f.

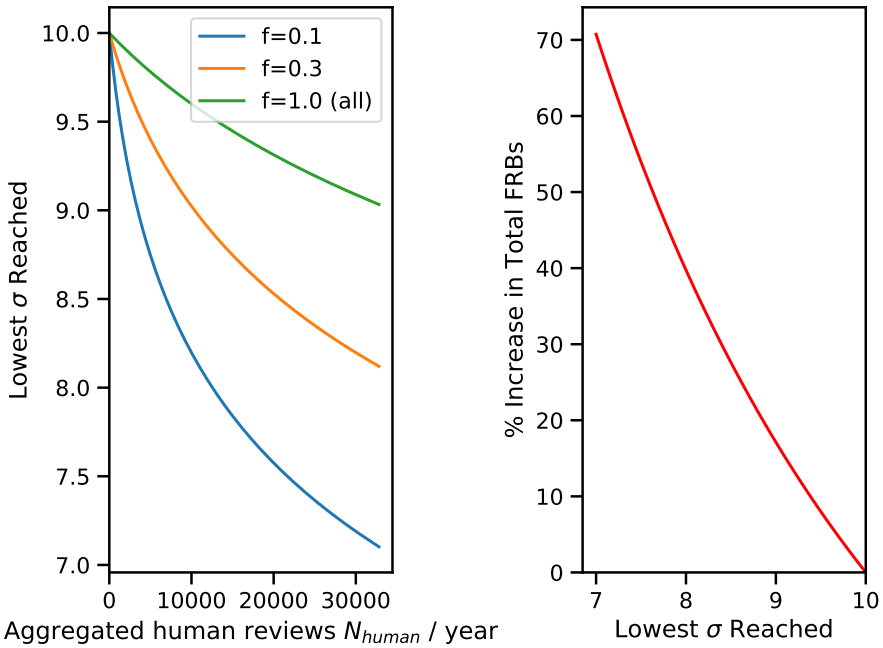


Figure 5.3: Left: lowest σ reached by prioritising candidates with automated classifier(s) and then, for the highest-ranked fraction f of candidates, collecting N_{human} aggregated reviews. Right: Percentage increase in total FRBs at or above σ, vs. $\sigma = 10$, assuming FRB are standard candles evenly distributed in local (Euclidean) space.

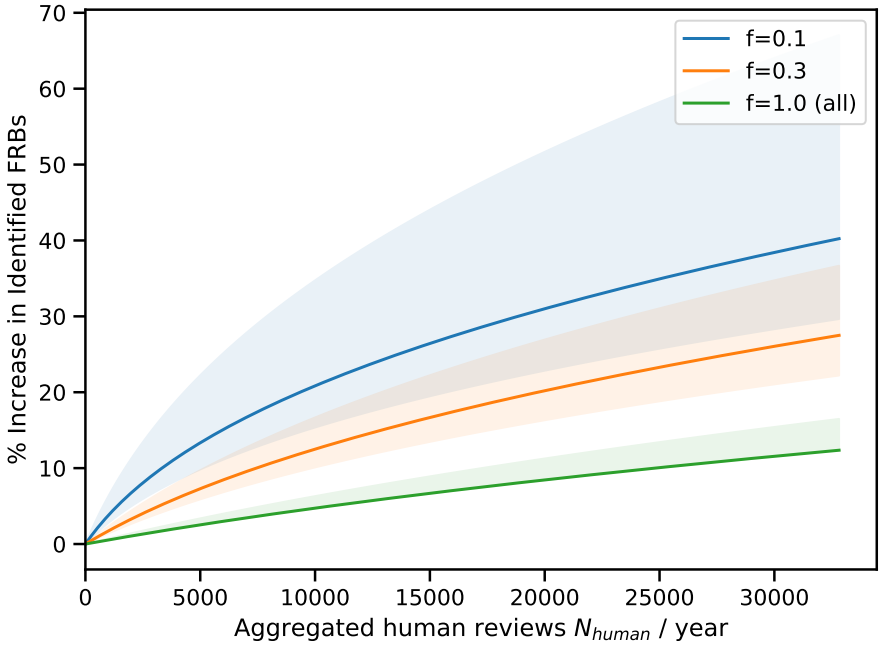


Figure 5.4: Percentage increase in identified FRBs (vs. $\sigma = 10$) a function of the aggregated human reviews available, for various f. Shaded areas represent uncertainty around the performance of automated and human classifiers. The upper bound assumes perfect classifiers (both automated and human). The lower bound assumes 75% human accuracy (in line with existing Zooniverse projects) and an automated classifier constructed (by ranking following a truncated half-normal distribution with std=0.43) to also be 75% accurate. The solid line assumes 75% human accuracy and uses the actual ranking performance of our prototype CNN.

The optimal choice of f is important; a lower f allows the system to review more events, reach lower σ, and potentially find far more FRB, but also increases the chance for real FRB to be incorrectly discarded due to imperfect automated candidate ranking. Better automated classifier(s) allow one to choose a lower f, leading to substantially more identified FRB given fixed available volunteer effort.

My preparatory work showed that there are likely substantial numbers of FRB below a $\sigma = 10$ cutoff (this Section) and that CNN might be able to detect them (Sec. 5.2), and so I submitted my proposal to attempt this project. In subsequent discussions, CHIME researchers told us they were already experimenting with using CNN to identify fast radio bursts in an as-yet-unpublished pilot project, with the aim of reducing false positive detections from RFI. I revised the proposal to build directly on that pilot project; the proposal was subsequently accepted.

In the remainder of the chapter, I describe the work I carried out using CHIME

data. I first investigate the behaviour of CHIME's prototype CNN. I then develop and launch a citizen science project to find faint FRB, and use the resulting classifications to train a new CNN.

5.4 Investigating CHIME's Prototype CNN

Ultimately published as Yadav 2020 [464], CHIME's pilot CNN project proceeded as follows. A CNN was trained to classify candidate spectrograms as either RFI or astrophysical using expert classifications of candidates with $\sigma \geq 8.5$, a cut lowered for this purpose from the previous $\sigma \geq 9 - 10$ cut. The CNN architecture followed a standard design, with one major exception. The first layers were modified to use fixed Sobell and Prewitt kernels of various scales. These act as edge detectors, increasing any coherent deviations from random noise in the hope of better detecting RFI and astrophysical candidates, respectively.

Data augmentation was used during training to improve performance. The authors created spectrograms dedispersed to 10 DMs (dispersion measures, see Sec. 5.1) uniformly sampled from the CHIME pipeline's DM range estimate, in order to account for expected small errors in the candidate DM estimate. The authors also applied noise augmentation following Zhang et al. 2018 [472], adding weighted spectrograms of random blank sky spectrograms (selected by shifting the trigger time of each genuine candidate) to candidates labelled by experts as FRB. The authors did not apply noise augmentation to candidates labelled as RFI out of concern the resulting spectrograms would not be 'realistic'.

The classifier achieved 99.2% accuracy on the test set - effectively perfect performance. Unfortunately, I argue here that the classifier likely achieved this performance through a 'shortcut' that does not generalise to lower σ candidates, and consequently makes it unable to detect faint FRB.

Empirical research shows that convolutional neural networks, along with other deep learning methods, take so-called shortcuts: decision rules that classify both training and test data well but do not generalise to practical application [146]. Shortcuts are not unique to deep learning, or even to algorithms. Most famously, a German horse known as Clever Hans garnered international attention (including a 1904 article in the New York Times, [421]) for his ability to count and solve multiplication problems. An expert committee was 'baffled' and suggested further study; controlled experiments eventually showed the horse was reacting to subconscious cues by the

questioner [422] [5]. Deep learning is plausibly more susceptible than horses to such shortcuts because they seek useful patterns in data without a corresponding semantic understanding of what that data means. [6]. Convolutional neural networks have been observed to classify based on image background rather than content [35], prioritise texture over shape [145], and to exploit patterns that are brittle and broadly incomprehensible to humans [197]. In this case, I believe the network learned to classify based on the background noise in the spectrograms.

To illustrate this, I first show that one can create a 92% accurate 'classifier' based solely on the total standard deviation of the spectrograms. Figure 5.5 shows the distributions of standard deviations for astrophysical and RFI classes. The astrophysical spectrograms have total standard deviations very close to 1, while the RFI spectrograms have consistently higher standard deviations. One can therefore make appropriate cuts on the total standard deviation to create a 92% accurate 'classifier'. Why should this be the case? I speculate that this is because the spectrograms were normalised by channel such that each channel was scaled to have a mean of 0 and a standard deviation of 1. For candidates where any correlations (signals) were faint and brief, this normalisation scaling had a similar effect on all channels. But for candidates where correlations were substantial or of extended duration, the normalisation causes bands intersecting the signal to be scaled very differently to bands not intersecting the signal. This then leads to a high standard deviation in the full normalised spectrogram.

The effectiveness of 'classifying' purely based on standard deviation suggests that the noise properties of the spectrogram are highly informative. I suspected the CNN might be using these noise patterns rather than the morphology of the signals. To investigate this, I took candidate spectrograms and added a random blank sky spectrogram with increasing weight, mimicking the effect of the candidate signal becoming fainter and the background more uniform. Figure 5.6 shows the effect on two random candidates. I then increased the weight of the blank sky and measured how the response of the network varied. Figure 5.7 shows the results. Before adding blank sky (i.e. when blank sky is added with weight 0), the network predicts the classes of all signals extremely confidently and extremely well. As the weight of the blank sky increases, the predictions on RFI signals rapidly shift towards astrophysical, while

[5]Perhaps the truly impressive feat of Clever Hans was that he learned to distinguish such cues from so many human questioners.

[6]Much research has focused on combining computer vision with language tasks such as 'caption this image' [351] in an effort to encourage semantic understanding

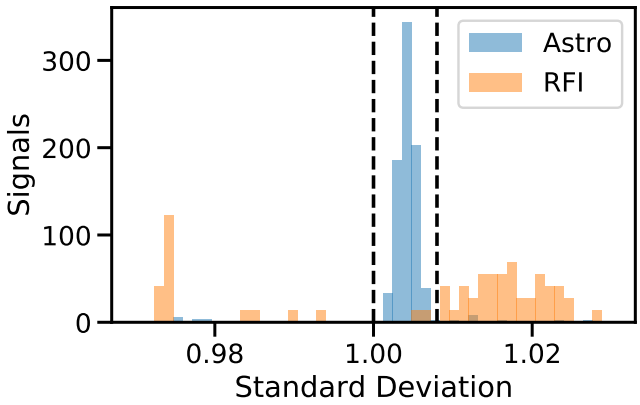


Figure 5.5: Distribution of total standard deviations for spectrograms labelled as either astrophysical (blue) or RFI (yellow). Cuts for 92% accuracy shown in black.

predictions on the astrophysical signals remain similar. Finally, as the blank sky begins to dominate, both predictions tend in the direction of the blank sky prediction. These results exactly match what one would expect to see if the network is distinguishing signals based on noise characteristics rather than on morphology. Essentially, the network likely classifies faint signals as astrophysical and strong signals as RFI.

To automatically identify faint fast radio bursts, we need a classifier that does not use the brightness of the signal as the critical feature. To avoid this issue, we need to retrain on labelled candidates which all have low signal-to-noise and so force the classifier to find other distinguishing features. [7] Labelling thousands of candidates is time-intensive and so beyond the capacity of the CHIME team, who are already occupied labelling higher signal-to-noise candidates. Instead, continuing with my proposed system, I turned to citizen science.

5.5 Citizen Science Project

To develop the citizen science project, I first copied (with permission) the 'Bursts from Space' project drafted (but never launched) by Robert Archibald to search for fast radio bursts in Arecibo survey PALFA [88]. Robert Archibald had used the Zooniverse project builder, an interactive tool to quickly create citizen science projects, to make a project with a single question: is this signal from space, or RFI? As well as rewriting and expanding the associated text, I made several technical changes. I introduced

[7] I would like to stress that this was also planned as future work by Yadav 2020 [464], who looked ahead to 'training on events with σ lower than the pipeline's default threshold' to 'help lower the...threshold and increase the overall FRB detection rate'.

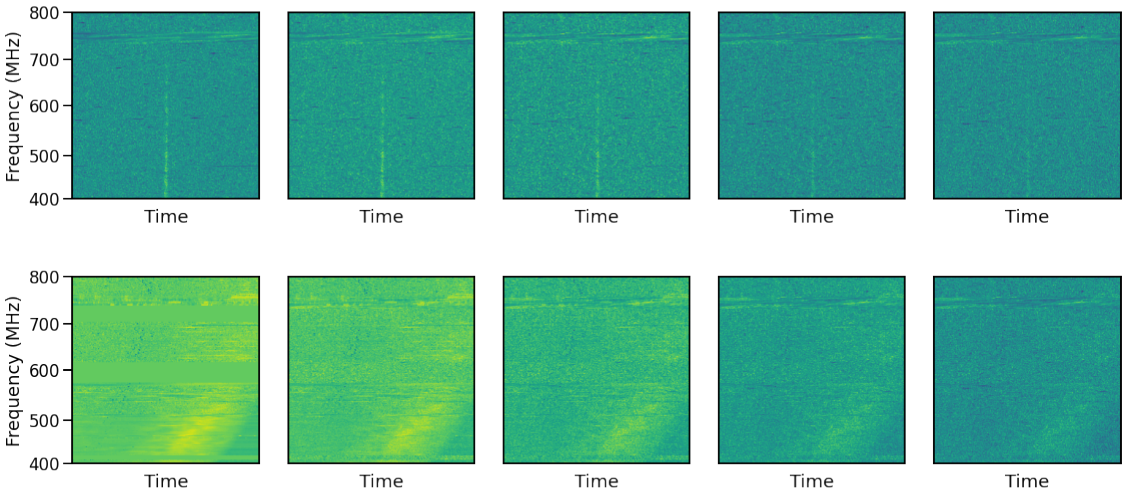


Figure 5.6: Example spectrograms for astrophysical (upper) and RFI (lower) candidates, with increasingly-weighted sky noise added (left to right).

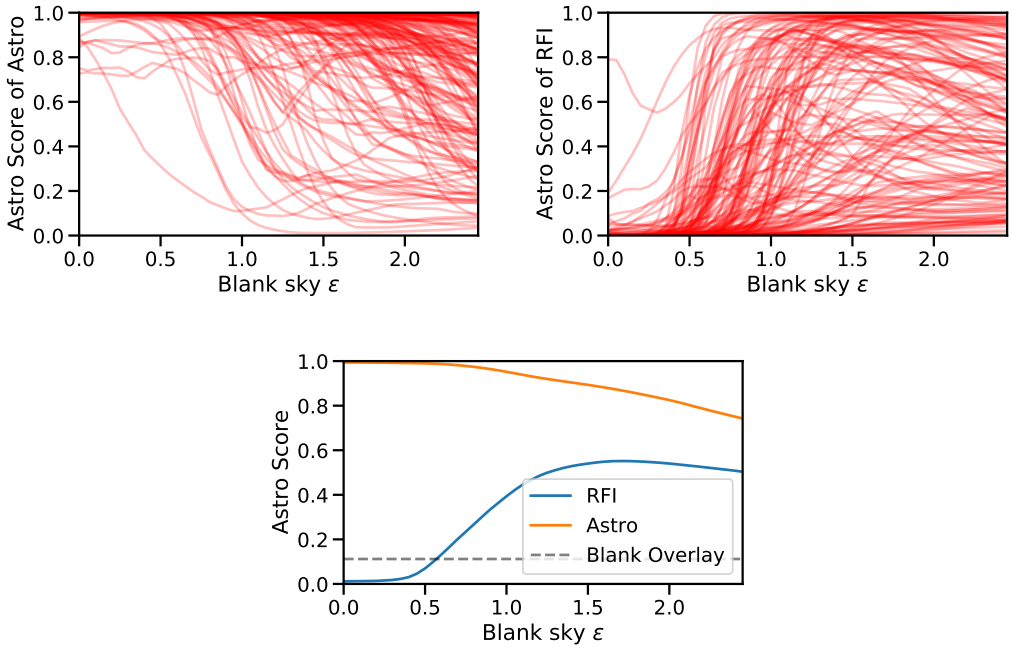


Figure 5.7: Upper: change in network predictions for either expert-labelled astronomical (left) or RFI (right) as blank sky weight is increased. Each candidate waterfall is shown as a red trace. Lower: Mean change in network predictions over all candidates.

'feedback' subjects; subjects with known classifications that displayed messages to volunteers depending on if the volunteers agreed with the known classification. These subjects were first generated from the expert-labelled high σ spectrograms, and then (after launch) extended to include low σ spectrograms classified by other volunteers with high confidence and selected by myself. Feedback subjects were shown to new logged-in volunteers at initially high rates, to help teach the task, before decreasing to a constant low rate (to measure agreement)[8]. The images shown to volunteers were generated from normalised CHIME spectrograms by clipping the intensity to lie in the 1%-99% range and then generating an image with the 'YlGnBu' `Matplotlib` colormap (after substantial experimentation). Volunteers were later given the option of also viewing spectrograms with 3x3 mean smoothing to better reveal faint bursts at the cost of resolution.

Inspired by Supernova Hunters [460], I decided to publish data from CHIME in weekly batches. Supernova Hunters found that regular small batches of new data helped keep volunteers motivated. Weekly batches were also useful for providing prompt volunteer classifications back to CHIME in order to minimise the additional storage overhead of recording faint candidates.

Volunteers on the Zooniverse beta-testing mailing list were invited to try the project in September 2020, using the high σ data already prepared for Yadav 2020 and labelled by experts. These volunteers were extremely capable at distinguishing astronomical signals from RFI; the aggregated votes from 15 or more volunteers agreed with CHIME experts in 99.3% of cases (99.1% for exactly 10 volunteers, 98.6% for exactly 5 volunteers). Of the top 10 candidates labelled by experts as astronomical but most confidently labelled by volunteers as RFI, follow-up inspection revealed three to be mislabelled by the experts (the remainder were considered relatively astronomical by volunteers, with a mean 'Space' vote fraction of 0.4 - compare to Fig. 5.8).

Prior to public launch, CHIME began saving intensity data for events with σ of $7.8 - 8.5$ (8.5 being the previous limit) for volunteers to classify. However, my inspection of the first low σ candidates showed that these low σ candidates were qualitatively different from the ≥ 8.5 signals used by Yadav 2020 and the Bursts from Space beta volunteers. While the ≥ 8.5 signals were roughly balanced between RFI and astronomical signals, the $7.8 - 8.5$ signals were dominated by visually blank spectrograms. My own classifications suggested around 10% looked plausibly astrophysical and around 5% were likely RFI, with the remainder having no visible signals.

[8]Specifically, the chance of encountering a feedback subject was 1 for the first 4 subjects, 30% for the following 4, 15% for the next 4, and 3% thereafter.

To account for this, I added a third answer to the first question - 'Nothing Here' (i.e. blank). I also added a fourth answer, 'It's Complicated', to handle signals that didn't fit well into the standard answers. This fourth answer was then further broken down by a follow-up question into 'Human and Space' (in response to questions from beta volunteers on handling spectrograms where an astrophysical-like signal is coincident with distinct RFI), 'Repeating Space' (similarly), and 'Something Weird' (in the hope of identifying bursts with unusual morphology, as well as data processing errors).

Bursts from Space launched in October 2020. Since launch, as of February 15th 2021, 1,482 registered volunteers have contributed 258,338 classifications. The rate of responses is sufficient for all candidate signals reported by CHIME at $7.8-8.5\sigma$ (17,264 to date) to receive 10 volunteer classifications without the need for prioritisation with machine learning. Figures 5.8 shows the σ distribution of the bursts. The median σ is 8.1. Figure 5.8 also shows the distribution of the fraction of volunteer votes for 'Space'. The vast majority receive very few 'Space' votes, but a significant minority receive several. My own visual inspection suggests that candidates with vote fractions of $0.3 - 0.5$ or above typically appear promisingly astronomical, based purely on the spectrogram. 814 (4.7%) of candidates have a Space vote fraction of 0.5 or above.

These promising candidates are typically at higher dispersion measures than have previously been reported. Figure 5.9 compares the distribution of promising ('Space' fraction ≥ 0.5) bursts against all published bursts recorded on the Transient Name Server[9]. The highest dispersion burst previously published by CHIME is 1300 pc cm^{-3} [131], and the highest dispersion burst from any survey is 2596 pc cm^{-3} (detected at the more sensitive Parkes Radio Telescope, [39]) Bursts from Space finds 87 promising candidates with dispersions above 4000 pc cm^{-3}. Verification of the first 24 candidates was carried out by Shriharsh Tendulkar (private communication), who found that — applying the same process as for other CHIME bursts — 4 are not explained by any known terrestrial or astronomical sources and are consistent with being fast radio bursts. Figure 5.10 shows these four FRB-consistent high-dispersion candidates. The highest dispersion candidate has a dispersion of 5571 pc cm^{-3}.

Further expert verification is urgently needed and underway at the time of writing. However, it is possible to make rough estimates for likely discoveries based on the verification thus far. If the fraction of promising candidates that pass expert verification is similar for the subsequently-discovered high-dispersion candidates, one expects discoveries of approximately 14 high-dispersion bursts to date. Assuming the fraction

[9]https://wis-tns.weizmann.ac.il/

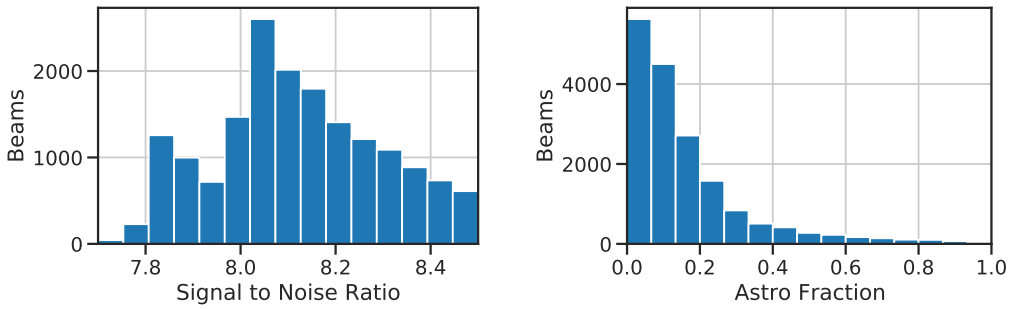


Figure 5.8: Distribution of CHIME beams sent to Bursts from Space by signal to-noise (left) and volunteer 'Space' response fractions (right).

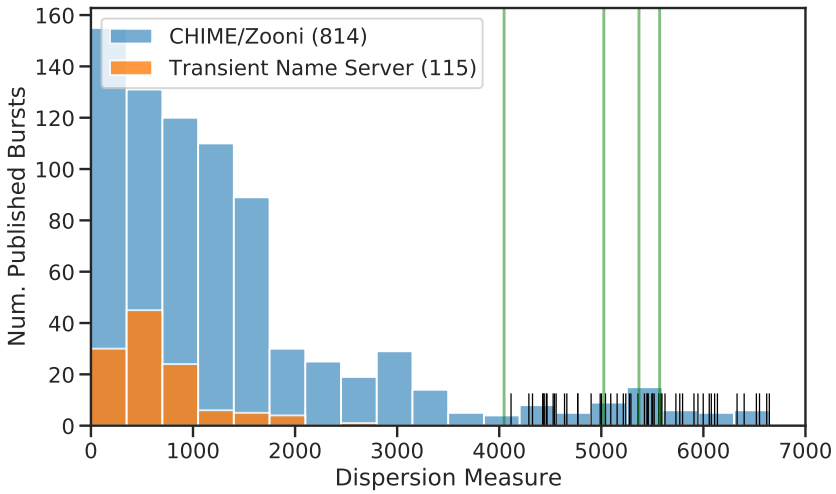


Figure 5.9: Dispersion measures of promising ('Space' > 0.5) CHIME candidates identified by Bursts from Space volunteers, compared against all previously published bursts on the Transient Name Server (excluding repeats). Black rug lines show candidates that underwent expert verification (24). Green vertical lines show candidates appear consistent with fast radio bursts (4)

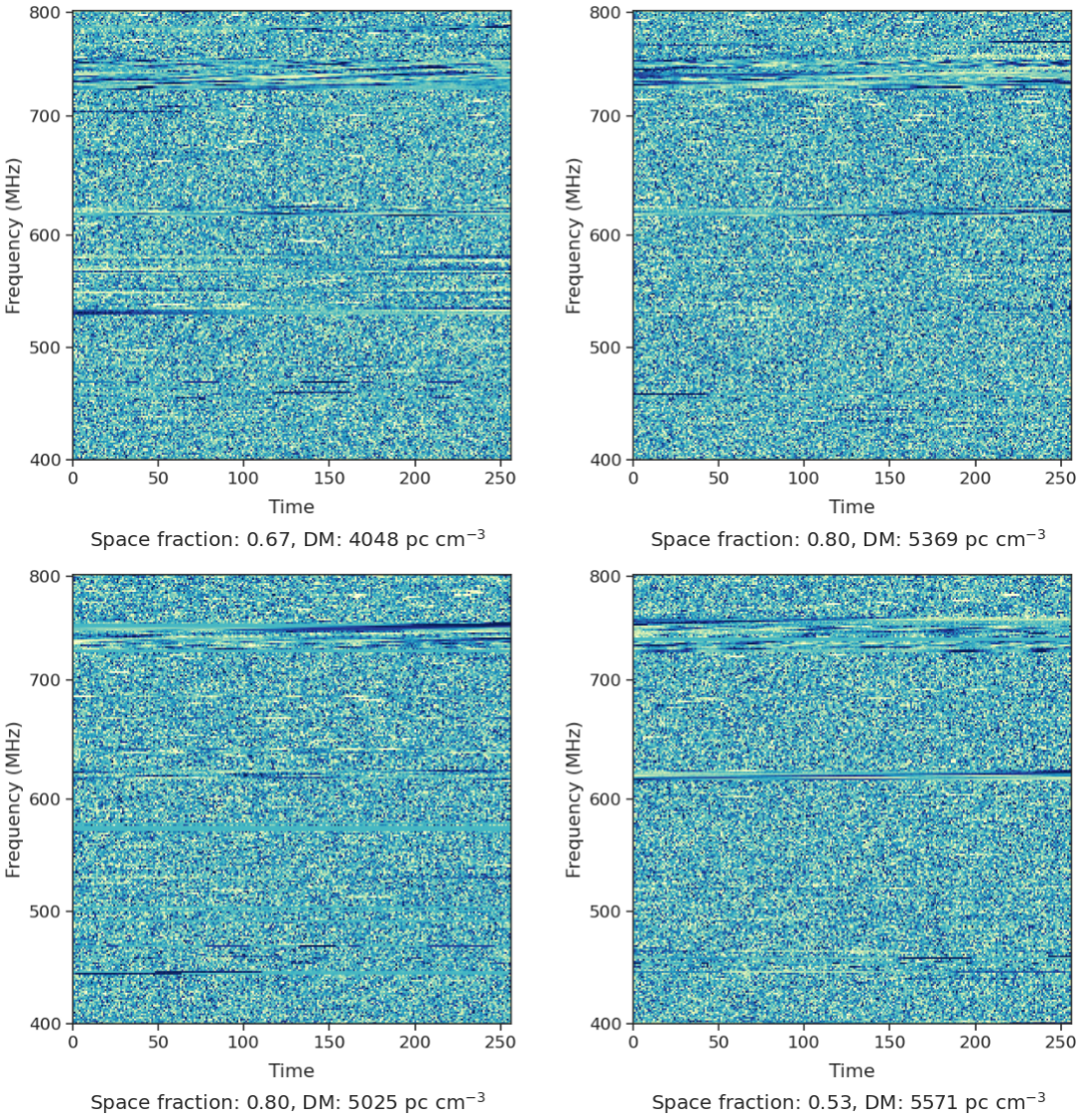


Figure 5.10: High-dispersion ($\geq$ 4000 pc cm^{-}3) promising ('Space > 0.5') bursts found following expert verification to be consistent with fast radio bursts.

is comparable for all promising candidates, not only high-dispersion ones, 661 promising candidates suggests approximately 130 FRB-consistent detections. Fonseca and Lopes 2017 [132] suggests an event detection rate of 700 fast radio bursts per year for CHIME during commissioning; 130 additional bursts over four months of operation suggests Bursts from Space is increasing the CHIME detection rate by approximately 40%.

5.6 Automated Classifier Retraining

My original proposal suggested using machine learning to prioritise which candidates should be reviewed by volunteers. The enthusiastic participation of Zooniverse volunteers made this unnecessary; each week, all candidates are classified by 10 volunteers. These classifications can be used to train a model able to predict what volunteers would have said rapidly enough for CHIME to save baseband data on the event.

Baseband data records the raw voltages at each antenna, in contrast to intensity-only data. This is useful for several reasons. First, baseband data allows for precise host localisation [10]. As I summarised in Sec. 5.1, identifying host galaxies both constrains FRB origin theories and provides independent redshifts for cosmological baryon measurements. Second, baseband data allows for far higher time resolution. Intensity-only data must be incoherently dedispersed by appropriately time-shifting channels, causing a smearing effect within each channel. Coherent dedispersion with baseband data corrects for this effect. This high-resolution data reveals the intriguing fine structure of each burst [170]. Finally, baseband data includes polarization information that helps characterise the magnetic environment of the as-yet-unknown sources [343]. Michilli et al. 2021 [302] developed an automated baseband analysis pipeline able to localise signals with sub-arcminute precision and approx 0.1ms time resolution. However, the computational cost is significant; each triggered event requires an average of 100 GB of storage. A citizen-trained machine learning algorithm could be used as a trigger to record baseband data for faint candidates which would otherwise be far too numerous to save.

To test this, I retrained the classifier introduced by Yadav 2020 on volunteer responses. I reframed the prediction task from classification (astronomical or RFI) to regression (volunteer 'Space' vote fraction) and switched to the loss function from binary cross-entropy to mean-squared-error. This was to gain information from the

[10] Particularly from the experimental CHIME 'outriggers' - separate radio telescopes designed to help localise bursts

level of volunteer confidence implied by the vote fraction, rather than discarding this information by binning. I only bin to calculate performance metrics after training, choosing a cut of 'Space' fraction ≥ 0.3 to match the approximate point where candidates are plausibly FRB-like (Sec. 5.5). I could not apply DM-augmentation as alternative dedispersion spectrograms were not available from CHIME. However, I introduced a new augmentation method, channel augmentation, where a random minority (10%, here) of channels are set to 0 each time a spectrogram is loaded into memory. This ensures predictions cannot strongly depend on any one channel, which is undesirable given that bursts cover many channels. I randomly divided the 13,440 candidates classified to date into 70% training, 10% validation, and 20% test subsets.

Figure 5.11 compares the $(7.8 - 8.5\ \sigma)$ test set performance of the retrained model with Yadav 2020's original model (trained only on expert $\geq 8.5\ \sigma$ labels). The retrained model is dramatically better at prioritising faint candidates the volunteers consider promising. In a sense, this is unsurprising; the original model was never trained on faint candidates and I previously showed (Sec. 5.4) that it is likely strongly biased towards labelling any faint signal as astrophysical. The purpose of the experiment is to highlight that one cannot assume models trained only on high σ data will generalise to lower σ, even when they are essentially perfect within their own regime. Retraining using labels from volunteers within the target σ range solves this issue. The most confident predictions of the retrained model are very likely to be promising faint candidates (85 of the top 100 ranked candidates are promising, compared to a base rate of 14%) and so the model could already be used to trigger baseband saves for such candidates.

5.7 Discussion

While I had hoped to find somewhat more distant bursts, discovering FRB-consistent signals at dispersions several times greater than have previously been published is surprising and exciting. As the project continues, the existence of this population will become clearer. The CHIME team will carry out more careful checks including verifying that adjacent non-overlapping beams were not activated (which would indicate RFI) and that the pulses show appropriate pulse broadening. Should the bursts be established as real, they may become unique tools for investigating helium reionization. The rate of reionization could be directly measured using the dispersion measure distribution of $z = 3 - 6$ FRBs [41].

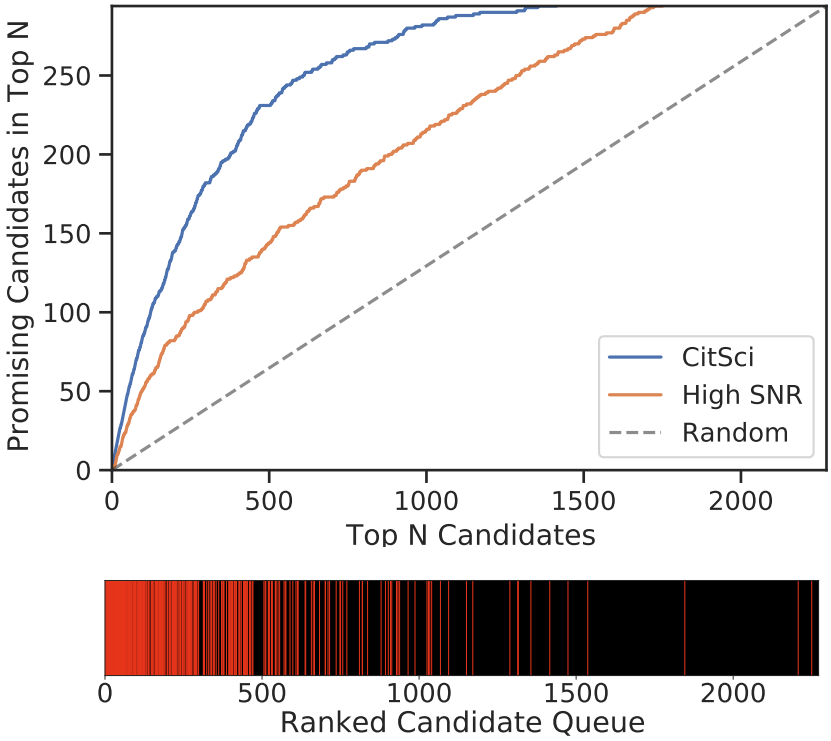


Figure 5.11: Above: number of promising ('Space' fraction > 0.3) candidates in the top N ranked candidates, with (CitSci) and without (High σ) retraining the model on citizen scientists. Below: corresponding visualisation of the retrained model's ranked queue of promising candidates (red vertical lines) versus other candidates (black lines). Candidates would be selected from the left (queue position 0) first.

Testing the sensitivity and specificity of the volunteers with injected pulses remains important outstanding work. Such measurements are necessary before one can draw statistical conclusions. Measuring volunteer performance with simulated data is well-supported by the Zooniverse infrastructure, but creating the simulations is not trivial due to the details of the CHIME pipeline. This should be a priority for future work, alongside the verification of specific high-dispersion bursts

My original proposal included two other classification approaches designed to complement a convolutional neural network. One was a PyTorch-based [331] optimisation method that aimed to rapidly find the most likely fit of a simple pulse profile to dedispersed data. For PALFA (with which the proposal was designed) and for CHIME at high σ, the main challenge is distinguishing different signals (bursts from RFI). But the data collected at lower σ in the course of this work suggests that in that regime, the task is more commonly to distinguish if a dedispersed signal is actually present. Optimisation methods are well-suited to such a task and bring the added benefits of being interpretable and statistically robust, with no concerns over shortcut learning (Sec. 5.4). I suggest further experimentation in this direction.

The CHIME pipeline itself includes a support vector machine (a basic machine learning classifier) to make live decisions on which events are astrophysical or RFI [18]. This support vector machine (SVM) was retrained by CHIME during Bursts from Space which resulted in a dramatic increase in the number of events being classified as astrophysical and sent to the Zooniverse. The features used for the SVM are simply the σ when dedispersing with slightly shifted dispersions, and 'the distribution of above threshold locations in the DM-time plane' [18]. The save/reject decisions made by this SVM are crucial and irreversible. I think that the results of this project could improve these save/reject decisions, either by providing additional training data or by ensembling classifiers together rather than stacking one classifier atop another. Redesigning a complete system that includes citizen-science-supported machine learning should help CHIME best find bursts that would otherwise be lost.

Chapter 6

Fast Photometry Inference Through Neural Emulation

6.1 Introduction

I have so far described training supervised machine learning models on labels derived from human responses - experts for tidal features (Chapter 2), volunteers for Galaxy Zoo (Chapters 3-4) and volunteers for Bursts from Space (Chapter 5). However, supervised learning need not require humans. Learning to emulate a simulation is one example. Models can learn to predict simulation outputs using a training set of many simulation runs using different initial parameters. The goal is typically to replace a slow simulator with a faster emulator. The emulator can then be used as a drop-in replacement for the simulator to enable applications for which the original simulator would be infeasibly slow.

One such application is MCMC sampling. The computational cost of drawing each sample is dominated by evaluating the likelihood function, which in turn is usually dominated by calculating the forward model (simulation). Replacing the forward model with an emulator therefore often speeds up MCMC runs by a similar factor. Supervised machine learning has been used to emulate simulations and hence apply MCMC to infer cosmological parameters from the matter [11, 166] and CMB power spectra [377], infer constraints on the epoch of reionisation from 21cm measurements [225, 376], and infer intrinsic stellar parameters from stellar spectra [91].

Emulation with neural networks in particular offers a further and less commonly exploited benefit. Neural networks are differentiable by design. This allows one to replace a non-differentiable simulator with a differentiable neural network emulator. One can then take advantage of gradients to explore the parameter space more efficiently, as is done in Hamiltonian Monte Carlo (HMC) [314]. In this chapter, I

describe using a neural network emulator for HMC inference of galaxy photometry.

Inferring galaxy parameters is an important and common task for extra-galactic surveys. Physical parameter estimation has led to significant discoveries in the domain of galaxy evolution such as the cosmic star-formation history [114, 282, 283], the mass-metallicity relation [117] and the starforming galaxy main sequence [398, 453]. Two main strategies have been developed in the literature to estimate physical parameters from galaxy SEDs or photometry. The first approach is to use a fixed set of galaxy models (or 'templates') chosen to represent the galaxy population analysed. These models are created using known physical parameters through stellar population synthesis models [60, 286]. Such methods typically include codes that perform χ^2 minimization over a large grid of models (e.g. Hyperz [48], LePHARE [24]) to find the best matching template or templates. Template matching struggles to account for degeneracy, where two or more templates with different intrinsic parameters may look similar (and so have similar χ^2). The lack of an explicit probability for each template makes it hard to meaningfully resolve such cases. MCMC sampling, the second approach, uses continuous forward models (rather than discrete templates) and so avoids this issue. Where different parameters yield observations of similar likelihood, the resulting probability distributions over each parameter should [1] account for this uncertainty. Unfortunately, MCMC is prohibitively slow for large galaxy samples; inference by various MCMC-based codes (MagPhys [92], Cigale [50, 63, 322], Prospector [255], AGNFitter [66], Fortesfit [362]) takes minutes to days per galaxy (depending on the model complexity) compared to seconds for template methods. Neural network emulation and gradient-based sampling offer a potential solution.

6.2 Forward Model

To investigate the use of emulation and gradient-based sampling for photometric inference, I first needed to create a forward model to emulate. Forward models specify the photometric measurements x observed given a galaxy with parameters θ. For my forward model, I used and extended the Python package `Prospector` [202, 255]. `Prospector` is a galaxy SED fitting tool designed to simulate realistic galaxy SEDs with a user-specified model, to make mock observations of those SEDs, and to infer galaxy parameters based on the SED model and some (real) observation. I use the SED and mock observation features of `Prospector` to create my forward model.

[1] Assuming the chains are fully mixed

Designing a model with the appropriate number of free parameters is difficult. Any useful model must have enough flexibility to fit real observations. Adding free parameters increases flexibility; however, too many free parameters will be difficult to constrain from photometry alone and the posteriors will become prior-dominated. From a pragmatic perspective, too many free parameters would also eventually defeat standard gradient-free MCMC approaches, preventing the direct comparisons between gradient and gradient-free sampling methods that I ultimately make in Sec. 6.4. I therefore aimed to use free parameters for the variables with the greatest effect on photometry, and to make physically reasonable fixed assumptions for the other variables based on the experience of my specialist collaborators.

I was particularly interested in modelling the effect of AGN. Photometric estimation of AGN flux is important for understanding how AGN affect galaxy evolution. Investigating the physics behind how AGN interact with their hosts often relies on measuring correlations between AGN and other galaxy properties (mass, star formation, quenching, merger history, morphology, etc). Estimating AGN flux is also crucial to accurately measure cosmological parameters with weak lensing. AGN introduce systematics in such measurements because they alter galaxy colours and therefore bias the photometric redshift estimates often required for weak lensing [369]. At the billion-source scale of Euclid, for example, such systematics are the limiting factor for precision cosmology [248]. It is crucial to identify galaxies likely to be 'contaminated' with AGN flux, allowing them to be cleaned from the weak lensing sample. Explicitly modelling AGN allows for this.

The task of estimating AGN flux is often replaced with the simpler task of determining whether or not the source flux is stellar-dominated, AGN dominated, or composite [328]. I aimed instead to directly measure the continuous flux contributions of each component. Accurate physical models of AGN SEDs remain an ongoing research challenge - in part due to the extensive variation of observed SEDs [61]. By focusing on inference from photometry, this difficulty is somewhat mitigated; only the most crucial of these variations will substantially affect the observed colours. I created an AGN SED model composed of accretion disk and dusty torus components, where the normalisation of each can vary independently. This allows the forward model to simulate AGN where the disk dominates the SED, and AGN with a heavily obscured disk but bright torus.

In short, the overall galaxy model I created includes flux from three components: the stars, the AGN disk, and the AGN torus. The stellar component is simulated with FSPS [83] and has three free parameters: the stellar mass, the star formation history

exponent τ, and the dust extinction. The AGN disk is based on a template [382] and has two free parameters: the luminosity and the (disk-specific) dust extinction. The AGN torus is based on torus simulations [317] and also has two free parameters: the luminosity and the inclination.

Below, I describe each component of the model in detail. I then emulate the model with a neural network (Sec. 6.3) and compare three sampling on a realistic set of simulated galaxies (Sec. 6.4).

6.2.1 Stellar Flux

The source frame galaxy SEDs simulated by Prospector are created using `FSPS` [83] (called via `pyFSPS` [134]). FSPS simulates the spectra of (composite) stellar populations, based on stellar evolution models, and (optionally) applies corrections for dust, IGM absorption, and other details. I assumed the galaxy model below (each entry corresponds to an FSPS argument).

I parametrise the star formation with a delay-τ model[2], with a log-uniform τ prior of [0.1, 30]. I scale the stellar populations using a stellar mass with a log-uniform prior of $[10^9, 10^{12}]$ $M_\odot$. For dust attenuation, parametrised with the dust optical depth at 5500Å, I use a uniform prior of [0, 2].

I then make several assumptions for the remaining necessary parameters, creating a simplified version of the model used in Leja et al. 2017 [255]. I use a Kroupa [239] initial mass function, a Calzetti [67] dust (extinction) law, dust emission following Draine and Li (2007) [106][3], and assume solar metallicity.

6.2.2 Accretion Disk

To model the accretion disk, I used the median composite radio-quiet quasar reported in Shang et al. 2011 [382] as a template and added a parameter to control the normalisation factor. For quasars, the accretion disk is expected to dominate in wavelengths short of 1 micron. As the torus is modelled independently (below), I applied an arbitrary power-law damping to the template at wavelengths above 1 micron to exclude any contribution at longer wavelengths. Specifically, I multiplied the disk flux at wavelengths above 1 micron by a damping factor given by $d = 10^5\lambda(A) + 10^4$.

[2]Delay-tau models have been noted to introduce systematics by coupling early and late star formation [388]; however, a full non-parametric model is outside the scope of this work

[3]Specifically, I assume a minimum incoming intensity $U_{min} = 1.$ (i.e. MW-like), exposed dust fraction $\gamma_e = 0.04$, and PAH dust fraction $q_{pah} = 0.1\%$

I applied an independent Calzetti extinction law to the disk component to allow for different typical dust optical depths for the (galaxy) stellar and (AGN) disk environments. Following Calzetti et al. 2000 [67], this extinction is given by:

$$f_{\text{reddened}} = f_0 \cdot 10^{-0.4 \cdot k \cdot \text{EB-V}} \tag{6.1}$$

where k is the wavelength.

6.2.3 AGN Dusty Torus

`Prospector` offers AGN torus templates based on the `CLUMPY` simulations by Nenkova et al. 2008 [317]. However, in the Prospector implementation, the normalisation of these templates is defined as a fraction of the stellar flux (i.e. $L_{AGN} = CL_*$ where C is a fixed or free variable). To allow for galaxies that are dominated by the quasar, the torus model must be able to contribute independently from the galaxy emission. I decided to introduce the torus model as a separate component, using the same `CLUMPY` simulations.

Theory [238] and observations of nearby AGN [430] both suggest that AGN are composed of dusty clumps. `CLUMPY` simulates the rest-frame SED that would be observed from the (re-)emission from such a clumpy torus under a pre-defined geometrical configuration. The main parameters are: the inner radius R_d (set by the dust sublimation temperature T_d), the outer radius R_0, the total number of clouds N_0, the opening angle, σ, and the inclination with respect to the observer i. The clumps are of equal optical depth τ_V and distributed with radial density profile r^{-q} out to $Y = \frac{R_0}{R_d}$ and various possible angular distributions. The average number of clouds along an equatorial ray is parameterised by N_0.

Nenkova et al. 2008 [317] provides a grid of SEDs calculated at each possible combination of these parameters. Photometric observations do not provide sufficient information to constrain all 6 torus parameters (in addition to the galaxy and accretion disk components). To restrict the free parameter space, my forward model assumes the physically reasonable values of opening angle $\sigma = 30\,\text{deg}$, cloud radial distribution: $q = 3$, disk size: $Y = 30$, and number of clouds $N_0 = 5$.

Inclination has the most significant effect on the resulting photometry and so the inclination is allowed to vary. Given the fixed parameters above, I interpolated (in log wavelength, log flux space) between the varied-inclination SEDs to create an SED component model as a function of arbitrary inclination $f_{torus}(i)$.

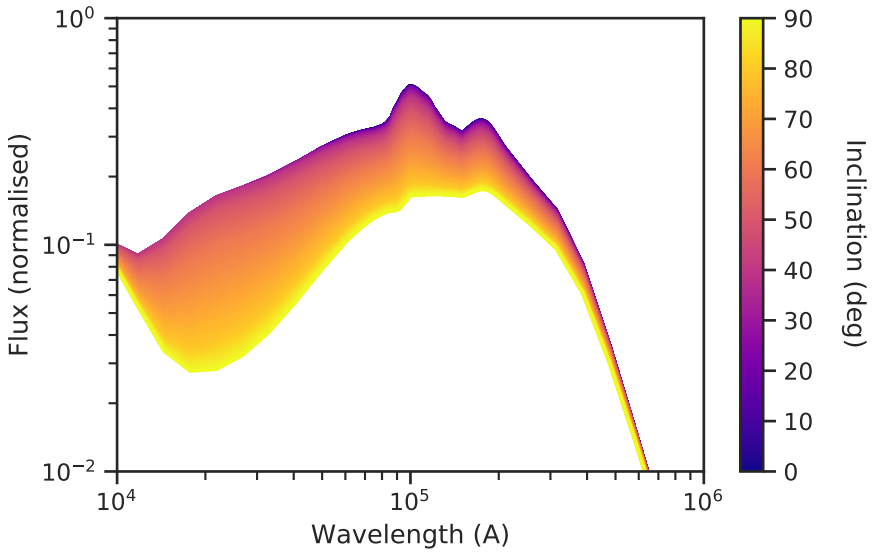


Figure 6.1: Dusty torus SED components by inclination. Interpolated between simulations by Nenkova et al (2008).

6.2.4 Mock observations

Having individually modelled the SED components (stellar light, by FSPS, and the AGN disk and torus light, as above), one can make mock photometric observations of the net SED. First, using `Prospector`, the SED is redshifted and absorption from the intergalactic medium is applied (I assume Madau absorption [283] at the default FSPS level). The flux through each desired filter is then calculated then via `sedpy` [201] bandpasses.

This forward model, particularly the stellar component, is prohibitively slow for practical inference on large samples of galaxies. Calculating e.g. 100,000 MCMC samples on 100,000 galaxies requires (at least) 10^{10} forward model evaluations. At 8.3ms per evaluation, this requires approx. 1,400,000 CPU-hours (160 CPU-years), neglecting the overhead of the sampling algorithm itself. In the next section, I show how neural network emulation can radically speed up evaluating the forward model.

6.3 Neural Network Emulation

Evaluating the forward model $f(\theta)$ dominates the computational cost of MCMC sampling and so it is helpful to replace $f(\theta)$ with some approximating function $\tilde{f}(x)$ that is cheaper to evaluate. To do this, I trained a neural network $f_w(\theta)$ to approximate $f(\theta)$. I generated many $\{\theta, f(\theta)\}$ pairs to learn from using the (non-emulated) forward

model. Parameters were selected with Latin hypercube sampling following Schneider et al. 2011 [377]. Latin hypercube sampling divides the space defined by the possible ranges of each parameter into an evenly-spaced grid and then randomly selects exactly one point within each cell. This pseudo-random sampling ensures approximately uniform coverage of possible parameters. I calculated 2 million $\{\theta, f(\theta)\}$ pairs for galaxies at redshifts from $0 \geq z \geq 4$ and a further 2 million pairs at redshifts from $0 \geq z \geq 1$. The $0 \geq z \geq 4$ redshift range covers the full range of the catalogue that I use later (see Sec. 6.4) and includes high redshift QSO, while the additional $0 \geq z \geq 1$ pairs improve performance in the low redshift regime where detected sources are far more numerous. The computational cost was roughly equivalent to (standard) MCMC sampling of a few hundred galaxies, which was easily feasible on a desktop workstation. Naively, four million pairs sounds excessive. However, with 9 free parameters, this is only a $5 \times 5 \times ...5$ grid of galaxies (the 'curse of dimensionality'). Emulator performance was found to consistently improve as additional pairs were added and so further increasing the number of training pairs would likely help. A random 10% of the pairs were set aside as a test set.

I identified the optimal neural network architecture using Hyperband [258], a Bayesian optimisation algorithm designed for bandit problems, as implemented for neural network architecture search by `keras-tuner` [325]. In a bandit problem, limited resources must be divided between exploration and exploitation; in neural network search, limited compute is divided between exploring new architectures and further training known promising architectures. The best performing networks had alternating wide and narrow layers, suggesting the physical relationships between galaxy parameters and photometry may be being learned through a series of encodings. Within this requirement, performance did not depend strongly on architecture design.

The emulator needed to be extremely accurate. For meaningful posteriors, the acceptable error in photometry predictions is not just a small fraction of the observed photometry, but a small fraction of the *uncertainty* in the photometry. Figure 6.2 illustrates the issue with predictions on a simulated (test set) galaxy. The emulator may be near-perfect (by eye) at predicting the photometry, but still make errors of comparable scale to the photometric error bars (of the order 10^{-9} maggies[4]). These errors then cause later inference to have incorrectly-centered posteriors, as the ob-

[4]A source with flux f in nanomaggies has a magnitude of m =[22.5 mag] - $2.5 \log_{10} f$. See http://www.sdss3.org/dr8/algorithms/magnitudes.php

Method	Parallelism	Time per Galaxy
Original Simulation	1 CPU core	8.3 ms
Original Simulation	16 CPU cores	≥ 0.5 ms
Neural Emulator	1 GPU, 1 galaxy	≈ 0.14 ms
Neural Emulator	1 GPU, 32 galaxies	6.0 μs
Neural Emulator	1 GPU, 1024 galaxies	1.26 μs

Table 6.1: Comparison of time to calculate observed photometry with the original simulation and with the neural emulator.

served and expected photometry must match on the scale of the error bars in order for a set of parameters to be plausible.

I found that naive prediction of the curve was not sufficient to reach the required accuracy of 10^{-9} maggiesThe emulator would commonly predict approximately the correct shape of the photometry - giving visually persuasive results - but with a fractionally offset normalisation leading to substantially incorrect posteriors. To rectify this, I modified the emulator to predict only the shape of the photometry (by normalising the outputs to sum to 1) and left scaling the emulated photometry as a (trivial) MCMC parameter. This allowed the emulator to reach significantly higher accuracies. Figure 6.3 shows the resulting emulator performance over the test set; per-band accuracy is typically better than 0.0008 mags. However, this performance comes at a cost. By reparametrising to separate out the scale parameter, the remaining parameters are no longer directly physically interpretable. The stellar mass and AGN disk and torus 'mass' (flux) parameters now express the relative contributions to the photometry rather than the absolute contributions.

While making the emulator sufficiently accurate is challenging, making it sufficiently fast is straightforward. TensorFlow takes advantage of GPU hardware design to make parallel predictions on many inputs, known as batches. Predicting photometry for (here) 1024 galaxies is therefore much quicker[5] than predicting photometry for one galaxy 1024 times. The speed of the emulator at predicting photometry for a single galaxy is roughly comparable to that of the original simulation, and so the emulator is thousands of times faster when making parallel predictions. Specifically, calculating photometry takes 8.3 ms h^{-1} with full model, 6.0 μs h^{-1} with the emulator in batches of 32, and 1.26$\mu\mathrm{s}^{-1}$ h^{-1} in batches of 1024 [6].

[5]More formally, the computation required is $\mathcal{O}(1)$

[6]h is a hardware unit, either a 3.2GHz CPU core (full model) or a GeForce GTX 1070 GPU (neural emulator).

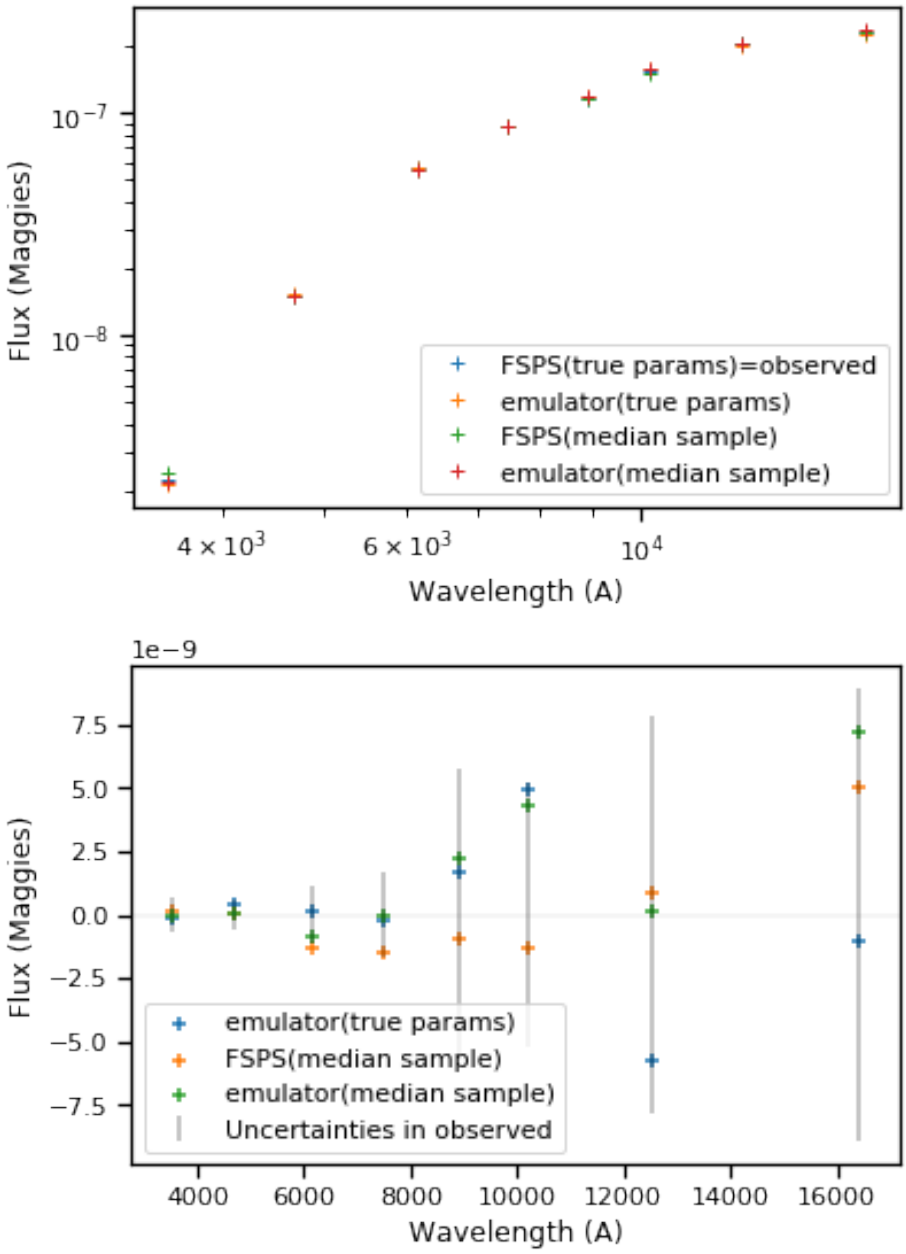


Figure 6.2: Photometry calculated by the forward model $f(\theta)$ and by the neural emulator $f_w(\theta)$ for a random simulated galaxy. Photometry is calculated four ways: from the true parameters with each model (blue and orange), and from the median of the HMC-sampled posterior (Sec. 6.4) (green and red). The photometry appears to match well (above). However, the residuals compared to the forward model and true parameters (below) are substantial for all three other methods.

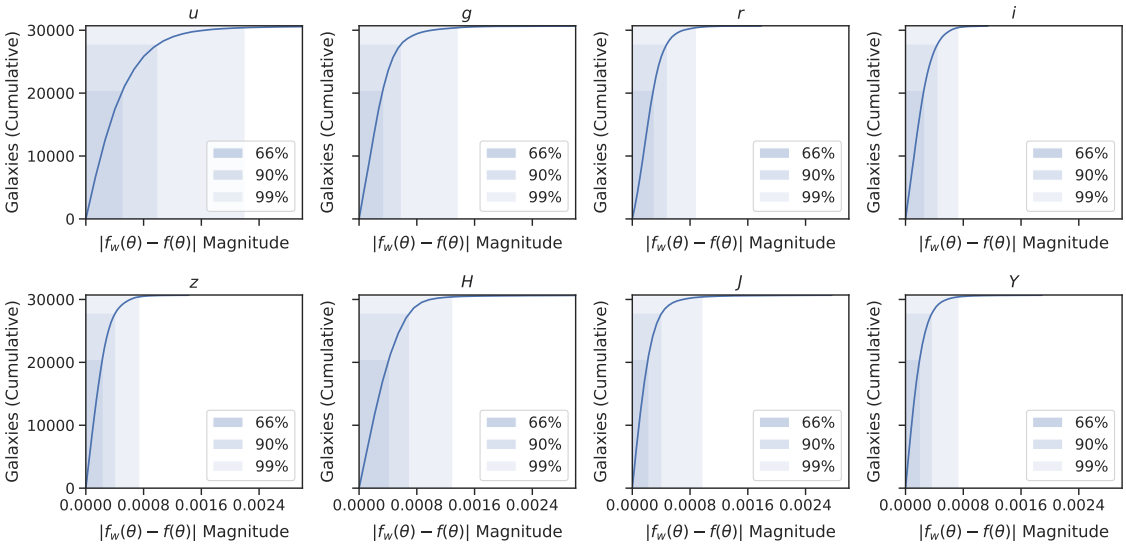


Figure 6.3: Absolute errors in predicting magnitudes on test set galaxies. The emulator-predicted flux is scaled to equal the observed flux so that the emulator need only predict the correct 'shape' of the photometry. The scaled predictions are shown to be extremely accurate.

The neural emulator $f_w(x)$ is ≈ 6000 times faster than the full forward model $f(x)$ while remaining accurate to within approximately 0.0008 mags.

6.4 Sampling

Above, I described defining a forward model $\hat{x} = f(\theta)$ (Sec. 6.2) and creating a fast, accurate neural emulator $\hat{x} = f_w(x)$ for that forward model (Sec. 6.3) . I used these to measure how emulation and gradient-based sampling affect the speed and accuracy of inference on a realistic problem.

To separate the effect of emulation from that of gradient-based sampling, I implemented and tested three inference approaches.

- EM - affine-invariant sampling (via `emcee`) with the full forward model $f(\theta)$
- EM-NN - affine-invariant sampling (via `emcee`) with the neural emulator $f_w(\theta)$
- HMC-NN - Hamiltonian Monte Carlo sampling (via `tensorflow-probability`) with the neural emulator $f_w(\theta)$

Comparing EM and EM-NN measures the effect (in speed and posterior quality) of neural emulation. Comparing EM-NN and HMC-NN measures any change in

performance between affine-invariant sampling and gradient-based sampling when using neural emulation.

Below, I first discuss setting up the realistic inference problem, and then describe each sampling method in detail.

6.4.1 Simulated Galaxies

To evaluate success, I measured how well each method recovered the known true parameters of a set of simulated galaxies with mock photometric observations. These simulated galaxies were selected from the $\{\theta, f(\theta)\}$ galaxy hypercube pairs (Sec. 6.3) to create a mock catalogue as similar as possible to the `CPz` catalogue of Fotopoulou and Paltani 2018 [138], which includes both AGN and QSO galaxy populations.

The CPz catalogue was constructed to act as a practical test of photometric redshift estimation in the presence of AGNs and QSOs. It was created by cross-matching sources detected in various original surveys, leading to a matched catalogue of galaxies with photometric and spectroscopic redshifts as well as wide photometric wavelength coverage. Measurement uncertainties are also quoted from the original surveys. In this work, I use the SDSS *ugriz* and VISTA *HJY* photometry, redshifts and corresponding uncertainties reported in CPz.

Each real `CPz` galaxy was matched to the simulated galaxy with the most similar photometry, defined as the simulated galaxy not yet matched that minimises the Euclidean distance in magnitude space between real and simulated galaxies. Galaxies were only allowed to match once at most. Only simulated galaxies in the test set (Sec. 6.3) were allowed to match, to avoid the network making predictions on galaxies it had already been trained to simulate well. To remove outliers (and ensure realistic uncertainties, below) I excluded real galaxies with all observed fluxes (by band) in the highest or lowest 2%. The marginal magnitude distributions of real and simulated catalogues are shown in Figure 6.4.

Each simulated galaxy also required realistic measurement uncertainties. I chose these uncertainties by fitting the empirical magnitude-σ[7] relation in the CPz catalogue using a non-parametric LOWESS smoother [80]. The resulting magnitude-σ fits (by band) are shown in Figure 6.5.

[7]The CPZ catalogue [138] quotes the uncertainties σ on the flux measurements for each band from each original survey composing the catalogue; refer to Alam et al. 2015 [14] and Arnaboldi et al. 2007 [23] for details of the SDSS DR12 and VISTA flux measurement error estimates, respectively.

The simulated catalogue is observationally indistinguishable from the real CPz catalogue, having the same joint probability distribution of magnitudes and uncertainties. Figure 6.4 compares the marginal distributions of photometry, by band.

Because I selected simulated galaxies based only on having similar magnitudes to the real CPz catalogue, the distribution of parameters in our simulated catalogue ($p(\theta|\mathcal{D}_{\text{test}})$ will be different to our priors $p(\theta)$ (which are uniform or log-uniform, see Sec. 6.2). This will be true for any real application and so sampling methods should be robust to this.

6.4.2 Sampling Implementation Details

Treating the photometric measurement errors as independent Gaussian variables, the log-likelihood of photometric observations x given parameters θ is:

$$\log p(\theta|x) = -\frac{1}{2}(\log(2\pi\sigma^2) + \frac{(x-\hat{x})^2}{\sigma^2}) \tag{6.2}$$

where $\hat{x} = f(\theta)$ using the full forward model, or $\hat{x} = f_w(\theta)$ using the neural emulator.

The task for each inference method was to find the posterior over the galaxy parameters given the observations, $p(\theta|x)$, given by:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{6.3}$$

This is analytically intractable (because calculating $p(x) = \int(p(x|\theta)d\theta$ requires one to consider all possible models) and so we need to estimate $p(\theta|x)$ using MCMC sampling. The essence of MCMC sampling is to randomly 'step' between θ values according to the changing value of $p(x|\theta)p(\theta)$. The θ position after each step is recorded (a 'sample'). The distribution of samples is guaranteed (under certain conditions) to converge to $p(\theta|x)$ after sufficient steps [281]. However, the number of steps required can be arbitrarily large. Worse, in high θ dimensions, posterior probability mass is concentrated in a relatively small volume, called the typical set. Any proposed step in a random direction becomes exponentially likely to 'land' at a θ outside the typical set and hence be rejected. Shrinking the step size will help make proposals within the typical set, but the chain will then mix increasingly slowly. Ensemble sampling (as implemented in emcee) and Hamiltonian Monte Carlo sampling (as implemented in TFP and my own code) address this issue in different ways.

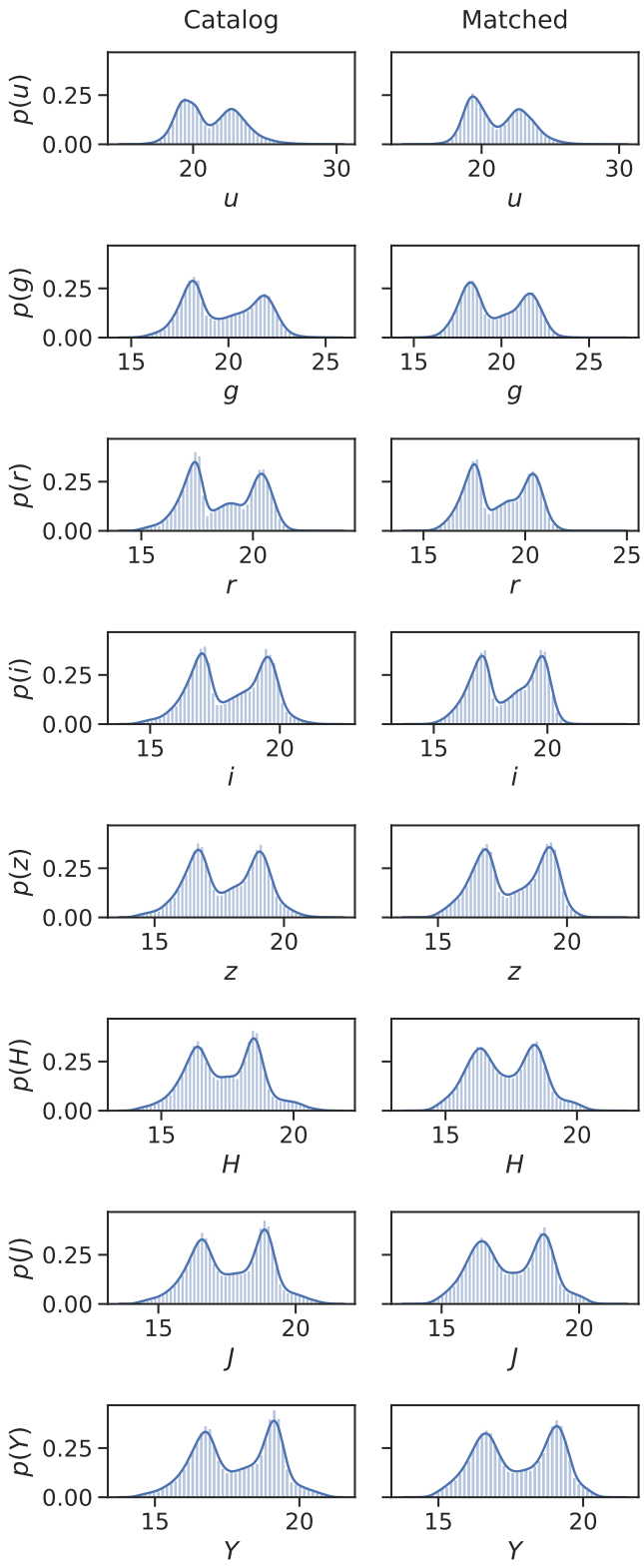


Figure 6.4: Marginal magnitude distributions of (real) CPz catalogue (left) and matched simulated catalogue (right).

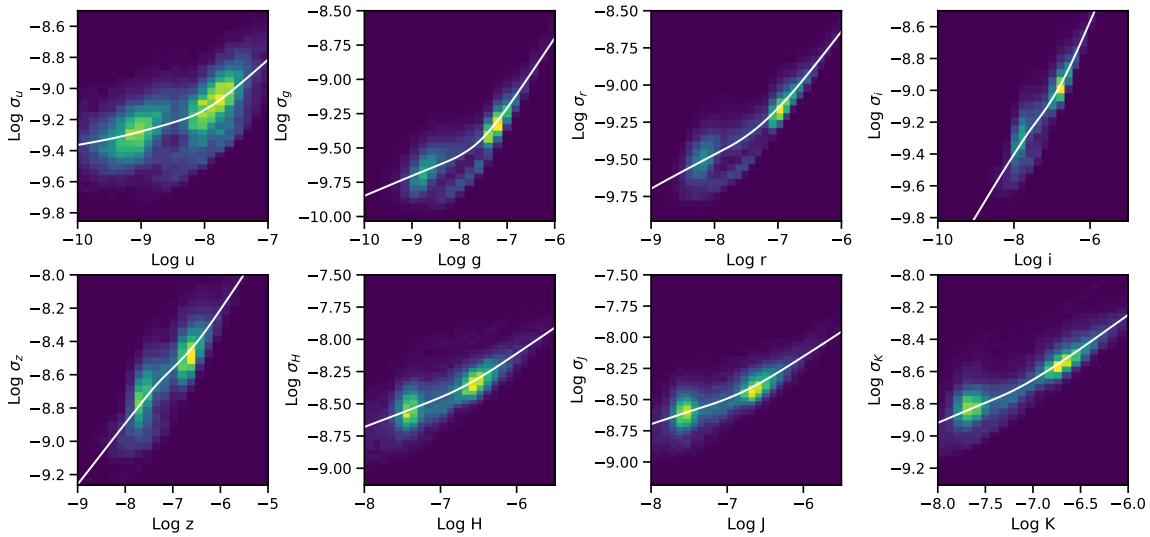


Figure 6.5: Magnitude-σ fits, by band, calculated for the CPz catalogue and used to estimate realistic uncertainties for the simulated (test) catalogue. Non-parametric fitting was performed with a LOWESS smoother [80]

6.4.3 Affine-Invariant Ensemble Sampling

Affine-invariant sampling with `emcee` is standard practice for many astronomers [136]. Affine-invariant sampling was introduced in Goodman and Weare 2010 [151] and is summarised by Foreman-Mackey et al. 2013 [135]. Ensemble sampling mitigates the issue of rejected proposals by using many parallel chains, called walkers. Walkers make move proposals along the vector towards another random walker, with the distance (and sign) randomised. The typical distance depends on the scale factor a; I used the default scale factor $a = 2$, which is thought to be appropriate for most problems [151]. In emcee's parallelised implementation, the walkers are divided into two subsets and the walkers in one subset each make a move proposal towards a random walker in the other subset. By proposing moves based on the relative position of other walkers, the sampling becomes affine-invariant i.e. performance becomes independent of any affine transformation to the problem. This is extremely helpful because many practical problems have correlated parameters which, in higher dimensions, make the typical set small; affine-invariant sampling naturally accounts for such correlations. Ensemble sampling is therefore suitable for a wide range of problems with minimal tuning (hence, `emcee` 'hammer')

To reduce the burn-in time required, I initialised the chains/walkers from the highest identified likelihood maxima. For the original forward model, such maxima were identified through Levenberg-Marquardt optimisation (following Prospector, [255]).

For the neural emulator, I took advantage of the known gradients by using Adam [230], an adaptive gradient descent optimiser. Finding global maxima in high dimensions is difficult due to the presence of many local maxima which nonetheless have unacceptably low likelihoods (corresponding to poor fits). To resolve this, many optimisation attempts are made starting from random θ (uniformly distributed along the range of the priors) and only the best are used as starting points. By visually comparing the quality of the resulting fits, I set a minimum likelihood of $\log L \geq -2.5$ as a cut for acceptable fits. 50 attempts with Levenberg-Marquardt or 100 attempts with Adam[8] are typically sufficient to identify at least one acceptable θ in over 99% of galaxies. Rare galaxies where no initial point of acceptable likelihood was found were flagged and not sampled, to avoid poorly-behaved chains slowing down parallel sampling.

The marginal posterior distributions (MPD) are typically the key measurement of interest. These change rapidly during the early stages of sampling, then stabilise after convergence, with additional samples gradually refining the estimated posterior. I selected the number of samples and chains (for `emcee`, walkers) empirically based on the rate of change of the MPD with respect to each. Samples and chains were added until the MPDs stabilised sufficiently. Figure 6.6 shows an illustrative example for HMC. The MPD-motivated choices are also supported by visual inspection of the corner plots, θ traces, and log-likelihood traces. 256 walkers collecting a total of 1 million samples per galaxy are found to be appropriate for `emcee`. The high number of samples is an expected consequence of the relatively high dimensionality ($D = 9$) of the inference problem.

6.4.4 Hamiltonian Monte-Carlo Sampling

The fundamental issue with Metropolis-Hastings Monte Carlo is that making a 'guess and check' move proposal in a random direction will likely lead to rejection. Instead, HMC (introduced as 'hybrid' Monte Carlo by Duane et al. 1987 [107]) makes move proposals based on the geometry, and specifically the gradient, of the target probability density. The gradient of the density points towards the mode, but - through a physical analogy where an auxilliary momentum variable is added and the resulting path is integrated - these gradients can be used to calculate proposals aligned with the typical set. An intuitive (astrophysical) analogy from Betancourt 2017 [38] is that of a satellite orbiting a planet; given the gradient of the potential energy (probability

[8]Fewer might be sufficient, but Adam is fast and parallel so I do not attempt to improve this

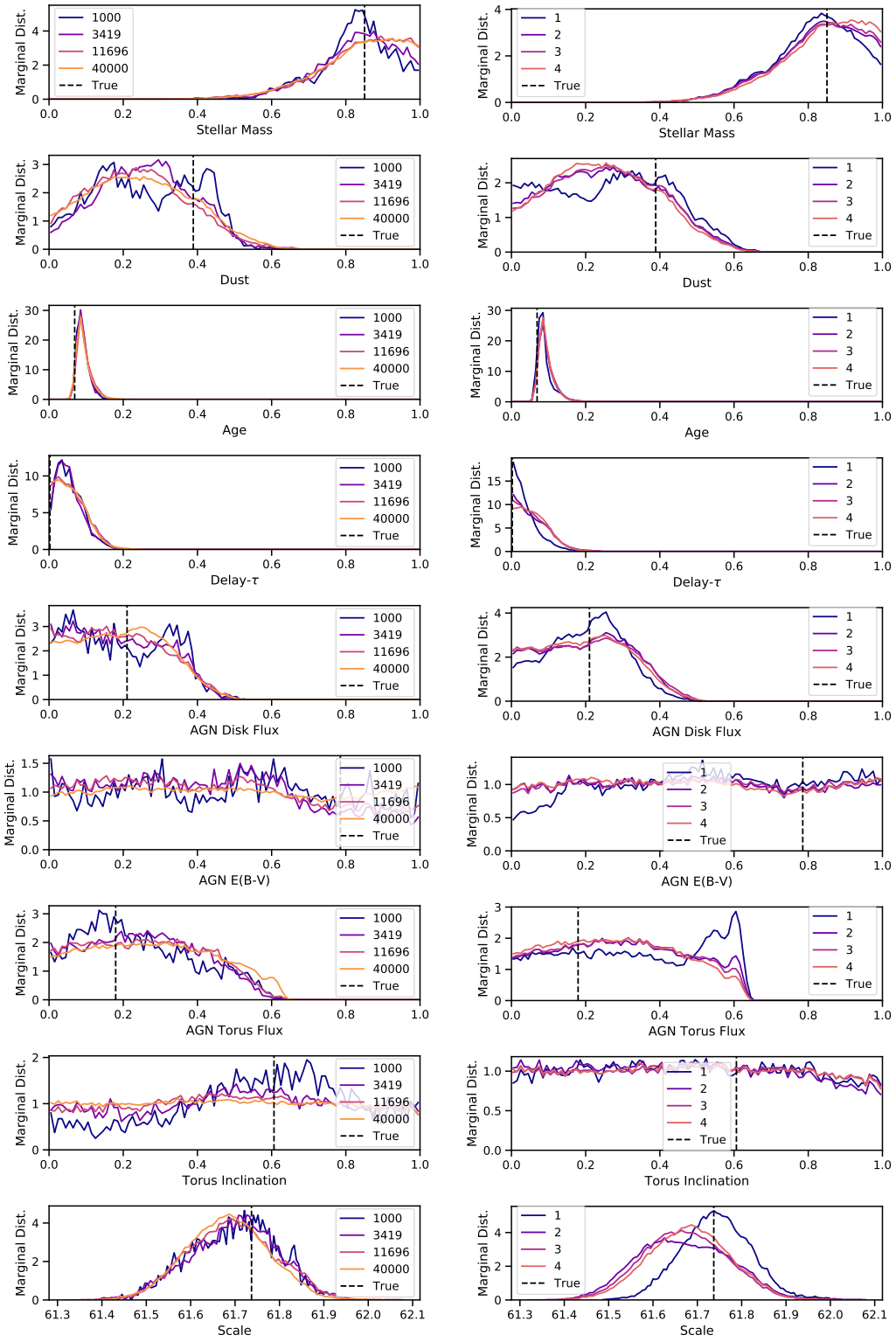


Figure 6.6: Marginal posteriors (unnormalised) estimated by HMC-NN for a random simulated galaxy, with increasing numbers of samples (left) or chains (right). I used the stability of these marginal posteriors to estimate the appropriate numbers of samples and chains typically required for convergence. Sample cuts are equally spaced on a log scale.

density), one can calculate the momenta and paths to remain in orbit (inside the typical set).

HMC is not affine-invariant and so correlated parameters will be difficult to sample. It is therefore important to rescale the parameters, or, equivalently, carefully choose the momentum distribution. I do so following the procedure outlined by Neal 2011 [315] and Betancourt 2017 [38]. Continuing the physical analogy, one can define a Euclidean metric M, $\Delta(q, q') = (q - q')^T \cdot M \cdot (q - q')$, where M is chosen to decorrelate the parameters. This metric implies distances in momentum space, $\Delta(q, q') = (p - p')^T \cdot M^{-1} \cdot (p - p')$, that in turn allows for momentum proposal distributions such as (commonly) $p \sim \mathcal{N}(0, M)$. Choosing the inverse metric to equal the covariances of the target distribution, $M^{-1} = \mathbb{E}[(q - \mu)(q - \mu)^T]$, will best decorrelate the parameters under such a distribution [9] .

Having selected a momentum distribution, one also needs to choose both the step size used during path integration, ϵ, and for how many steps to integrate the resulting path, L.

With respect to the step size ϵ, short step sizes will be computationally expensive, while long step sizes will lead to incorrect paths and rejected proposals. I set the step size using dual-averaging step size adaptation, as introduced by Hoffman and Gelman 2014 [176] (along with NUTS, below), with a target mean acceptance ratio of 0.75. Detailed balance is not satisfied during step size adaption and so a multi-phase approach is needed, which I detail after discussing L.

With respect to the number of steps L, short paths will fail to benefit from following the gradients along the typical set, while long paths may 'orbit' and repeatedly travel through parameter space already explored. The No U-Turn Sampler, NUTS, introduced by Hoffman and Gelman 2014 [176], is often used to set L dynamically by terminating a path if it doubles back. However, for technical reasons (see my investigation at this GitHub issue), the TFP implementation of NUTS is significantly slower than HMC with a fixed L. Instead, having fixed M and ϵ as above, I manually tuned L; I found $L = 10$ to be appropriate for this problem.

Bringing together the various approaches above, my final HMC sampling procedure was as follows. The first phase was an initial burn-in period of 2500 samples, initialised from the optimised starting point found as with EM-NN (Sec. 6.4.3), followed by 1000 samples where the step size was allowed to vary (initially 0.05). The

[9] If one allows more sophisticated metrics, e.g. Riemannian metrics $\Sigma(q)$ that vary locally, one can further decorrelate parameters at the cost of increased complexity

Method	Sampling Time ($s^{-1}h^{-1}$)	Speedup vs. EM ($h^{-}1$)
EM	2.0×10^6	-
EM-NN	1.6×10^3	1,000x
HMC	1.5×10^2	10,000x

Table 6.2: Comparison of total sampling time for posterior recovery, by sampling method.

second phase was 5000 samples at a fixed step size set to the mean of the previously-identified step sizes. These samples are used to estimate the parameter covariance and hence construct a non-unity M. Finally, the third phase is a further burn-in period of 2500 samples (where the step size is again allowed to vary) followed by the 10,000 samples ultimately used to measure the posterior. The various hyperparameters above were generally found through extensive trial-and-error rather than from theoretical considerations. Based on the evolution of the marginal posterior distributions (as for `emcee`), 32 chains and 100,000 samples per chain were found to be appropriate for HMC.

6.5 Results

I showed (Table 6.1) that one can dramatically speed up the calculation of photometry from θ by emulating our forward model with a neural network. Comparing the speed of `emcee` with the full forward model (EM) and with the neural emulator (EM-NN) shows the practical effect of this speed-up. EM-NN is approximately 1000 times faster than EM, a similar ratio to the relative speed-up of the emulator vs. the full forward model. This is expected given that sampling the forward model is the main computational cost; speeding up the forward model should therefore proportionally speed up the overall inference.

By using the known emulator gradients to perform HMC sampling, and using the efficient batchwise HMC implementation in `tensorflow-probability`, I achieved a further factor of 10 in speed.

Figure 6.7 shows a visualisation of the posteriors as a function of true parameter value. One can qualitatively verify the accuracy of each method by comparing the posteriors estimated via each sampling method with each other and with the true parameter values used when simulating the galaxies. All three approaches provided posteriors that are very similar to one another and consistent with the known true parameter values. The posteriors are often poorly-constrained, as expected from the limited (*ugrizHJY* photometry only) data. Crucially, and unlike template-matching,

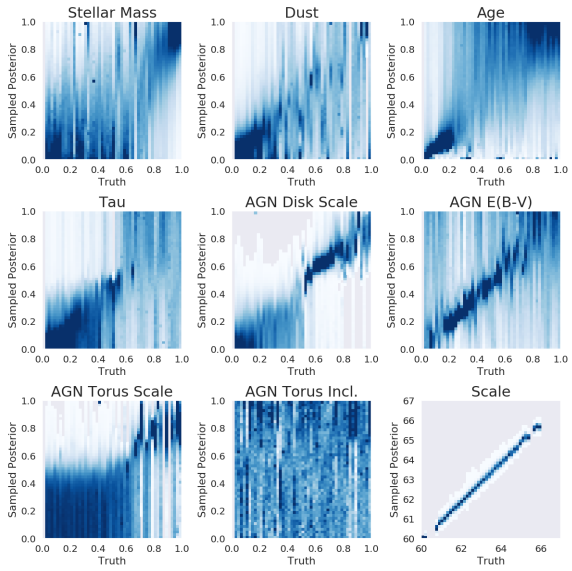
the stacked (marginal) posteriors reveal the true mean parameter values in regions of parameter space where the photometry is sufficient to provide weak constraining information. For example, the relative flux contributions of each component (stellar mass, AGN disk scale, and AGN torus scale) are poorly constrained when low (and dominated by other components), but well-constrained when high (and significant).

These results demonstrate both that Bayesian inference can recover useful measurements from photometry alone, and that neural emulation can radically speed up such inference without reducing accuracy. This project remains in progress; I hope to perform more quantitative tests on the quality of the posteriors and to investigate adding redshift as a free parameter.

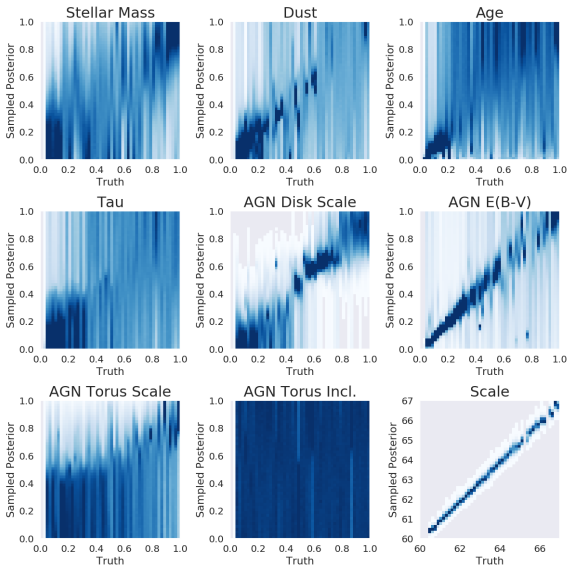
6.6 Discussion

Much of the work in this chapter related to the technical details of effectively using TensorFlow Probability's HMC sampler. Being more familiar with the robust 'hammer' design of `emcee`, I was surprised that HMC required a carefully designed multi-stage process and significant tuning to perform efficiently. This is partly a consequence of TFP being a relatively new package. As TFP and related packages mature, running HMC with neural emulators will likely become far easier. `Stan` [405] is an example of a mature HMC package for graphical models. I would favour the creation of a 'Stan with neural networks' package which combines automatic emulation with full-featured HMC. Meanwhile, my implementation is available at https://github.com/mwalmsley/agnfinder for other researchers interested in the specific context of photometric inference.

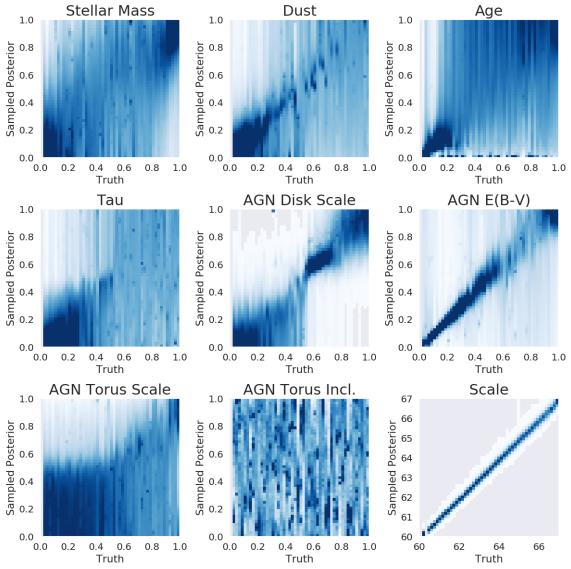
The further speedup from using HMC is likely to be increasingly pronounced in high dimensions. However, the curse of dimensionality makes training accurate emulators difficult as the volume per example increases rapidly. Fitting an explicit functional form, as opposed to using the network outputs directly, may relieve this. Taken further, one could likely improve performance by emulating each component of the forward model separately. Many practical astrophysical simulations, including those here, are composed of complex sources (e.g. stellar spectra) and simple rules to combine or observe them (e.g. redshifting, IGM absorption). One could emulate the complex source functions, as above, and straightforwardly rewrite the remaining simple rules in probabilistic languages. This would both better encode our astrophysical knowledge and drastically reduce the dimensionality of the emulation problems.



(a) `emcee` with original simulation (EM). 500 galaxies.



(b) `emcee` with neural emulator (EM-NN). 500 galaxies.



(c) Hamiltonian Monte-Carlo with neural emulator (HMC-NN). 1400 galaxies.

Figure 6.7: Mean posterior vs. true parameter value, for EM (above), EM-NN (centre) and HMC (below), by parameters. Parameter values (x axis) and posterior values (y axis) are each split into 50 evenly-spaced bins. White to dark blue represents zero to maximal posterior mass per bin (all sampled posteriors are normalised to integrate to 1). True values lie on the diagonals, but the individual posteriors should typically be only weakly constrained. All three methods produce very similar posteriors, supporting the reliability of the (far faster) HMC-NN approach. Tick labels are rescaled from the physical parameter ranges to the 0-1 interval.

Alternatively, it may be possible to compress the forward model outputs into lower dimensions, as was done by [17] to predict spectra.

This chapter has investigated how to speed up sampling from a posterior. But what if sampling was not needed? Likelihood-free inference (LFI) methods aim to directly predict the posterior from observations, by learning from $\{\theta, f(\theta)\}$ pairs like those simulated in Sec. 6.3. This would be several orders of magnitude faster than sampling as one avoids making thousands of forward model predictions per posterior. However, LFI is presently a non-interpretable 'black box' method that lacks (for example) the established MCMC convergence diagnostics, and so one would need some guarantee that the posteriors are reliable. This is especially important given the difficulty in accurately emulating photometry that this chapter shows, as emulating photometry might be considered a necessary step implicit in LFI. A solution might be a 'cosmic ladder' for inference: using emcee for small samples, HMC for large samples, and LFI for very large samples, while checking that each produces consistent posteriors for galaxies in specifically-chosen overlapping subsets.

Deep learning is often criticised for being incompatible with making precision astrophysical measurements. This is a fair critique; even the techniques I use for inferring Galaxy Zoo posteriors in Chapter 4 likely fall short of the level of precision in posteriors that a weak lensing study, for example, might require. However, I hope that this chapter demonstrates how deep learning may, through combination with traditional techniques, play an integral part in drawing robust statistical conclusions.

Chapter 7

Conclusion

This thesis has covered four different scientific topics, each addressed through different methods. My choice of approach has been guided by considering how to make the best use of limited data. For detecting low surface brightness tidal features (Chapter 2), I used custom augmentations and a small model to learn from scarce expert labels. For classifying galaxy morphology (Chapters 3-4), I introduced novel probabilistic loss functions to learn from uncertain volunteer labels. For CHIME (Chapter 5), the promising FRB candidates would have been discarded for being too faint were it not for my project. And for fast inference with galaxy photometry (Chapter 6), I used emulated Hamiltonian Monte-Carlo sampling to handle degeneracies from having only colour measurements rather than full spectra.

Efficiently using limited data requires careful consideration of uncertainty. Judging what can and cannot be inferred is a familiar statistical question, and so one can greatly benefit from applying established statistical tools in the new context of deep learning. For example, my final galaxy morphology classifier - the part of this thesis that I consider to have the most interesting methods - combines Dirichlet-Multinomial distributions and entropy with MC Dropout and ensembled networks. One can also apply deep learning to support established statistical tools, as I did with my emulator in Chapter 6. Combinations in either direction are only possible thanks to probabilistic programming libraries like `Tensorflow Probability` [101] and `Pyro` [43, 344]. I therefore encourage scientists to use, give feedback, and where possible contribute to such libraries.

Exploiting uncertain information helps you build more accurate models. To do so so, you build models which themselves make predictions with uncertainty. This win-win scenario is especially important in the broader context of ML in astronomy.

I believe that deep learning in astronomy is moving from 'will it work?' to 'can we trust it?' We have seen many successful applications of deep learning in astronomy

since being introduced around 2015 (Sec. 1.3). While this undoubtedly reflects some selection bias (the most recent unsuccessful application I came across while writing this thesis was Thonnat and Bijaoui 1989 [424]), the broad effectiveness of deep learning is indisputable. However, it is notable that papers showing a deep learning approach is *possible* (i.e. predictions are generally accurate) remain more common than papers that actually *use* deep learning to answer a science question or publish new measurements. I think this is partly attributable to well-founded concerns around reliability; generally accurate predictions are often not sufficient for science[1]. In the context of large surveys (where machine learning is often described as vital for scaling up analysis), systematic errors increasingly dominant statistical errors from limited sample sizes. I opened this thesis with an argument that deep learning is both prone to a variety of systematic errors and difficult to inspect (Sec. 1.1). Several well-known limitations of deep learning - overfitting, uncertainty, and shortcut learning - must be considered seriously for scientific applications. I hope this thesis shows that it is possible to make progress in overcoming them.

I think that future progress will require an attitude of 'trust but verify'. In experimental sciences, controlled trials are crucial. The performance of a deep learning method is ultimately an experimental question that deserves deliberate comparisons in a controlled environment. I applaud recent work testing machine learning approaches in this manner, such as PLAsTiCC [285] and the LSST photometric redshift experiment [375]. To facilitate such comparisons, astronomical deep learning models must be transparent, open-source, reproducible, and portable (i.e. straightforwardly applied to related datasets). To this end, I have included the images shown to volunteers in the Galaxy Zoo DECaLS data release[2] in the hope that it can provide a useful benchmark on which to compare new methods, much like the Galaxy Challenge did in first demonstrating the potential of CNN. I have also open-sourced my code should it be useful as a baseline.

Computer scientists continue to make progress on estimating uncertainties [4, 431, 433] and on shining light into black box models [22, 75, 307]. Astronomers must search the developing literature for new tools that might help address our practical concerns. Further, we should engage with computer scientists to argue for the importance of our requirements and to recruit them into astronomical work. Astronomy is an ideal

[1]There are specific cases such as adaptive optics control or observational scheduling where machine learning acts more like an engineering tool than a scientific instrument.

[2]Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies. Currently under review and available on arxiv.

domain for researchers interested in ML problems, partly for our large and diverse datasets but also - and perhaps more importantly - for our interesting questions.

Bibliography

[1] Mario G. Abadi, Julio F. Navarro, Matthias Steinmetz, and Vincent R. Eke. Simulations of Galaxy Formation in a Lambda CDM Universe I: Dynamical and Photometric Properties of a Simulated Disk Galaxy. *The Astrophysical Journal*, 591(2):499–514, jul 2002. ISSN 0004-637X. doi: 10.1086/375512. URL http://stacks.iop.org/0004-637X/591/i=2/a=499.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2016. URL http://arxiv.org/abs/1603.04467.

[3] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, Marcel A. Agüeros, et al. the Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 182(2):543–558, 2009. ISSN 0067-0049. doi: 10.1088/0067-0049/182/2/543. URL http://stacks.iop.org/0067-0049/182/i=2/a=543?key=crossref.bc07496fb06b943bcf82755687fa84b4.

[4] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges, nov 2020. ISSN 23318422. URL http://arxiv.org/abs/2011.06225.

[5] Roberto G. Abraham, Francisco Valdes, H. K. C. Yee, and Sidney van den Bergh. The morphologies of distant galaxies. 1: an automated classification system. *The Astrophysical Journal*, 432:75, 1994. ISSN 0004-637X. doi: 10.1086/174550.

[6] Sandro Ackermann, Kevin Schawinski, Ce Zhang, et al. Using transfer learning to detect galaxy mergers. *Monthly Notices of the Royal Astronomical Society*, 479(1):415–425, sep 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1398. URL http://arxiv.org/abs/1805.10289.

[7] Scott M. Adams, Dennis Zaritsky, David J. Sand, et al. The environmental dependence of the incidence of galactic tidal features. *Astronomical Journal*, 144(5):128, 2012. ISSN 00046256. doi: 10.1088/0004-6256/144/5/128. URL http://iopscience.iop.org/1538-3881/144/5/128.

[8] H Adorf. Connectionism and Neural Networks. In *Knowledge-Based Systems in Astronomy*, chapter 13, pages 213–245. Springer, 1991. URL https://link.springer.com/chapter/10.1007/3-540-51044-3_25.

[9] H. M. Adorf and E. J. A. Meurs. Supervised and unsupervised classification - The case of IRAS point sources. In *Large-Scale Structures in the Universe Observational and Analytical Methods*, pages 315–322. Springer, Berlin, Heidelberg, 1988. doi: 10.1007/3-540-50135-5_86. URL https://link.springer.com/chapter/10.1007/3-540-50135-5_86.

[10] Devansh Agarwal, Kshitij Aggarwal, Sarah Burke-Spolaor, et al. FETCH: A deep-learning based classifier for fast transient classification. *Monthly Notices of the Royal Astronomical Society*, 497(2):1661–1674, feb 2020. ISSN 13652966. doi: 10.1093/mnras/staa1856. URL http://arxiv.org/abs/1902.06343.

[11] Shankar Agarwal, Filipe B. Abdalla, Hume A. Feldman, et al. PkANN - I. Non-linear matter power spectrum interpolation through artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 424(2): 1409–1418, aug 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2012.21326.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2012.21326.x.

[12] Hiroaki Aihara, Carlos Allende Prieto, Deokkeun An, et al. THE EIGHTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY: FIRST DATA FROM SDSS-III. *The Astrophysical Journal Supplement Series*, 193(2):29, 2011. ISSN 0067-0049. doi: 10.1088/0067-0049/193/2/29.

[13] Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, et al. The Hyper Suprime-Cam SSP survey: Overview and survey design. *Publications of the Astronomical Society of Japan*, 70(Special Issue 1), jan 2018. ISSN 2053051X. doi: 10.1093/pasj/psx066. URL https://academic.oup.com/pasj/article/doi/10.1093/pasj/psx066/4103292.

[14] Shadab Alam, Franco D. Albareti, Carlos Allende Prieto, et al. THE ELEVENTH and TWELFTH DATA RELEASES of the SLOAN DIGITAL SKY SURVEY: FINAL DATA from SDSS-III. *Astrophysical Journal, Supplement Series*, 219(1):12, jul 2015. ISSN 00670049. doi: 10.1088/0067-0049/219/1/12.

[15] Franco D. Albareti, Carlos Allende Prieto, Andres Almeida, et al. The 13th Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-IV Survey Mapping Nearby Galaxies at Apache Point Observatory. *The Astrophysical Journal Supplement Series*, 233(2):25, dec 2017. ISSN 1538-4365. doi: 10.3847/1538-4365/aa8992. URL http://stacks.iop.org/0067-0049/233/i=2/a=25?key=crossref.0e5e2c912566adf4ffaa0bcb41b9a5c4.

[16] Ibrahim A. Almosallam, Sam N. Lindsay, Matt J. Jarvis, and Stephen J. Roberts. A sparse Gaussian process framework for photometric redshift estimation. *Monthly Notices of the Royal Astronomical Society*, 455(3):2387–2401, jan 2016. ISSN 13652966. doi: 10.1093/mnras/stv2425. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stv2425.

[17] Justin Alsing, Hiranya Peiris, Joel Leja, et al. SPECULATOR: Emulating stellar population synthesis for fast and accurate galaxy spectra and photometry. *The Astrophysical Journal Supplement Series*, 249(5), nov 2020. ISSN 23318422. doi: 10.3847/1538-4365/ab917f. URL http://dx.doi.org/10.3847/1538-4365/ab917f.

[18] M. Amiri, K. Bandura, P. Berger, et al. The CHIME Fast Radio Burst Project: System Overview. *The Astrophysical Journal*, 863(1):48, mar 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aad188. URL http://dx.doi.org/10.3847/1538-4357/aad188.

[19] M. Amiri, K. Bandura, M. Bhardwaj, et al. A second source of repeating fast radio bursts. *Nature*, 566(7743):235–238, feb 2019. ISSN 14764687. doi: 10.1038/s41586-018-0864-x. URL http://www.nature.com/articles/s41586-018-0864-x.

[20] B. C. Andersen, K. Bandura, M. Bhardwaj, et al. CHIME/FRB discovery of eight new repeating fast radio burst sources. *The Astrophysical Journal Letters*, 885(1):L24, aug 2019. ISSN 23318422. doi: 10.3847/2041-8213/ab4a80.

[21] B. C. Andersen, K. M. Bandura, M. Bhardwaj, et al. A bright millisecond-duration radio burst from a Galactic magnetar. *Nature*, 587(7832):54–58, may 2020. ISSN 14764687. doi: 10.1038/s41586-020-2863-y. URL `http://arxiv.org/abs/2005.10324`.

[22] Javier Antorán, Umang Bhatt, Tameem Adel, et al. Getting a CLUE: A Method for Explaining Uncertainty Estimates, 2020. ISSN 23318422. URL `http://arxiv.org/abs/2006.06848`.

[23] M Arnaboldi, M.∼J. Neeser, L.∼C. Parker, et al. ESO Public Surveys with the VST and VISTA. *The Messenger*, 127:28, mar 2007.

[24] S. Arnouts and Olivier Ilbert. Le PHARE Photometric Analysis for Redshift Estimate. *Astrophysics Source Code Library*, 2011. URL `http://cdsads.u-strasbg.fr/abs/2011ascl.soft08009A`.

[25] Devansh Arplt, Stanislaw Jastrzebskl, Nicolas Bailas, et al. A closer look at memorization in deep networks. *34th International Conference on Machine Learning, ICML 2017*, 1:350–359, 2017. ISSN 2640-3498.

[26] Adam M. Atkinson, Roberto G. Abraham, and Annette M N Ferguson. Faint tidal features in galaxies within the Canada-France-Hawaii telescope legacy survey wide fields. *Astrophysical Journal*, 765(1), 2013. ISSN 15384357. doi: 10.1088/0004-637X/765/1/28.

[27] A Baillard, E Bertin, V De Lapparent, et al. The EFIGI catalogue of 4458 nearby galaxies. *Astronomy and Astrophysics*, 532(74), 2011.

[28] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14(12):e1006613, dec 2018. ISSN 15537358. doi: 10.1371/journal.pcbi.1006613. URL `https://dx.plos.org/10.1371/journal.pcbi.1006613`.

[29] N. M. Ball, J. Loveday, M. Fukugita, et al. Galaxy types in the Sloan Digital Sky survey using supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 348(3):1038–1046, mar 2004. ISSN 00358711. doi: 10.1111/j.1365-2966.2004.07429.x. URL `https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2004.07429.x`.

[30] Nicholas M. Ball and Robert J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, jul 2009. ISSN 0218-2718. doi: 10.1142/S0218271810017160. URL http://www.worldscientific.com/doi/abs/10.1142/S0218271810017160.

[31] Manda Banerji, Ofer Lahav, Chris J. Lintott, et al. Galaxy Zoo: Reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353, 2010. ISSN 00358711. doi: 10.1111/j.1365-2966.2010.16713.x.

[32] P.H. H. Barchi, R.R. R. de Carvalho, R.R. R. Rosa, et al. Machine and Deep Learning applied to galaxy morphology - A comparative study. *Astronomy and Computing*, 30:100334, jan 2020. ISSN 2213-1337. doi: 10.1016/J.ASCOM.2019.100334. URL https://www.sciencedirect.com/science/article/pii/S2213133719300757.

[33] Joshua E. Barnes and Lars Hernquist. Dynamics of interacting galaxies. *Annual Review of Astronomy and Astrophysics*, 30(1):705–742, 1992. ISSN 00664146. doi: 10.1146/annurev.aa.30.090192.003421.

[34] Melanie R Beck, Claudia Scarlata, Lucy F Fortson, et al. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society*, 476(4):5516–5534, jun 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty503. URL https://academic.oup.com/mnras/article/476/4/5516/4923080.

[35] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11220 LNCS: 472–489, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01270-0_28.

[36] C. A. Beichman, G. Neugebauer, H. J. Habing, et al. IRAS Explanatory Supplement, 1988. URL https://lambda.gsfc.nasa.gov/product/iras/docs/exp.sup/index.html.

[37] V. Belokurov, D. B. Zucker, N. W. Evans, et al. The Field of Streams: Sagittarius and Its Siblings. *The Astrophysical Journal*, 642(2):L137–L140, 2006. ISSN 0004-637X. doi: 10.1086/504797. URL http://stacks.iop.org/1538-4357/642/i=2/a=L137.

[38] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo, jan 2017. ISSN 23318422. URL http://arxiv.org/abs/1701.02434.

[39] S. Bhandari, E. F. Keane, E. D. Barr, et al. The survey for pulsars and extragalactic radio bursts - II. new FRB discoveries and their follow-up. *Monthly Notices of the Royal Astronomical Society*, 475(2):1427–1446, nov 2018. ISSN 13652966. doi: 10.1093/mnras/stx3074. URL http://dx.doi.org/10.1093/mnras/stx3074.

[40] Shivani Bhandari, Elaine M. Sadler, J. Xavier Prochaska, et al. The host galaxies and progenitors of Fast Radio Bursts localized with the Australian Square Kilometre Array Pathfinder, may 2020. ISSN 23318422. URL http://dx.doi.org/10.3847/2041-8213/ab672e.

[41] Mukul Bhattacharya, Pawan Kumar, and Eric V. Linder. Fast radio burst dispersion measure distribution as a probe of helium reionization. *Physical Review D*, 103(10), oct 2021. ISSN 24700029. doi: 10.1103/PhysRevD.103.103526. URL http://arxiv.org/abs/2010.14530.

[42] Michal Bílek, Pierre Alain Duc, Jean Charles Cuillandre, et al. Census and classification of low-surface-brightness structures in nearby early-type galaxies from the MATLAS survey. *Monthly Notices of the Royal Astronomical Society*, 498(2):2138–2166, 2020. ISSN 13652966. doi: 10.1093/mnras/staa2248.

[43] Eli Bingham, Jonathan P Chen, Martin Jankowiak, et al. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

[44] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[45] Asa F.L. Bluck, J. Trevor Mendel, Sara L. Ellison, et al. Bulge mass is king: The dominant role of the bulge in determining the fraction of passive galaxies in the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 441(1):599–629, jun 2014. ISSN 13652966. doi: 10.1093/mnras/stu594.

[46] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *32nd International Conference on Machine Learning, ICML 2015*, 2:1613–1622, 2015.

[47] C. D. Bochenek, V. Ravi, K. V. Belov, et al. A fast radio burst associated with a Galactic magnetar. *Nature*, 587(7832):59–62, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2872-x. URL http://dx.doi.org/10.1038/s41586-020-2872-x.

[48] M. Bolzonella, J. M. Miralles, and R. Pelló. Photometric redshifts based on standard SED fitting procedures. *Astronomy and Astrophysics*, 363(2):476–492, 2000. ISSN 00046361.

[49] C. R. Bom, J. Poh, B. Nord, et al. Deep learning in wide-field surveys: Fast analysis of strong lenses in ground-based cosmic experiments, nov 2019. ISSN 23318422. URL http://arxiv.org/abs/1911.06341.

[50] M. Boquien, D. Burgarella, Y. Roehlly, et al. CIGALE: A python Code Investigating GALaxy Emission. *Astronomy and Astrophysics*, 622:A103, feb 2019. ISSN 14320746. doi: 10.1051/0004-6361/201834156.

[51] Connor Bottrell, Maan H. Hani, Hossen Teimoorinia, et al. Deep learning predictions of galaxy merger stage and the importance of observational realism. *Monthly Notices of the Royal Astronomical Society*, 490(4):5390–5413, oct 2019. ISSN 13652966. doi: 10.1093/mnras/stz2934. URL http://arxiv.org/abs/1910.07031.

[52] Alexandre Boucaud, Marc Huertas-Company, Caroline Heneka, et al. Photometry of high-redshift blended galaxies using deep learning. *Monthly Notices of the Royal Astronomical Society*, 491(2):2481–2495, may 2020. ISSN 13652966. doi: 10.1093/mnras/stz3056. URL http://dx.doi.org/10.1093/mnras/stz3056.

[53] Larry Bradley, Brigitta Sipocz, Thomas Robitaille, et al. astropy/photutils: v0.5, aug 2018. URL https://doi.org/10.5281/zenodo.1340699.

[54] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001. ISSN 08856125. doi: 10.1023/A:1010933404324.

[55] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *Seventh International Conference on Learning Representations (ICLR 2019)*. arXiv, mar 2019. URL http://arxiv.org/abs/1904.00760.

[56] C. R. Bridge, R. G. Carlberg, and M. Sullivan. The CFHTLS-DEEP catalog of interacting galaxies. I. Merger rate evolution to z = 1.2. *Astrophysical Journal*, 709(2):1067–1082, 2010. ISSN 15384357. doi: 10.1088/0004-637X/709/2/1067. URL http://stacks.iop.org/0004-637X/709/i=2/a=1067?key=crossref.c8033cb2ea950006c5ea35a265c45213.

[57] Sarah Bridle, John Shawe-Taylor, Adam Amara, et al. Handbook for the GREAT08 challenge: An image analysis competition for cosmological lensing. *Annals of Applied Statistics*, 3(1):6–37, mar 2009. ISSN 19326157. doi: 10.1214/08-AOAS222.

[58] Alyson Brooks and Charlotte Christensen. Bulge formation via mergers in cosmological simulations. In *Galactic Bulges*, volume 418, pages 317–353. Springer, 2015. ISBN 9783319193786. doi: 10.1007/978-3-319-19378-6_12.

[59] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners, may 2020. ISSN 23318422. URL http://arxiv.org/abs/2005.14165.

[60] G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344(4):1000–1028, oct 2003. ISSN 00358711. doi: 10.1046/j.1365-8711.2003.06897.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1046/j.1365-8711.2003.06897.x.

[61] Catherine L. Buchanan, Jack F. Gallimore, Christopher P. O'Dea, et al. Spitzer IRS Spectra of a Large Sample of Seyfert Galaxies: A Variety of Infrared Spectral Energy Distributions in the Local Active Galactic Nucleus Population . *The Astronomical Journal*, 132(1):401–419, jul 2006. ISSN 0004-6256. doi: 10.1086/505022. URL http://stacks.iop.org/1538-3881/132/i=1/a=401.

[62] Brandon Buncher, Awshesh Nath Sharma, and Matias Carrasco Kind. Survey2Survey: A deep learning generative model approach for cross-survey image mapping, nov 2020. ISSN 23318422.

[63] Denis Burgarella, Veronique Buat, and J. Iglesias-Páramo. Star formation and dust attenuation properties in galaxies from a statistical ultraviolet-to-far-infrared analysis. *Monthly Notices of the Royal Astronomical Society*, 360(4):1413–1425, apr 2005. ISSN 00358711. doi: 10.1111/j.1365-2966.2005.09131.x. URL http://dx.doi.org/10.1111/j.1365-2966.2005.09131.x.

[64] L. Cabayol, I. Sevilla-Noarbe, E. Fernández, et al. The PAU survey: Star-galaxy classification with multi narrow-band data. *Monthly Notices of the Royal Astronomical Society*, 483(1):529–539, feb 2019. ISSN 13652966. doi: 10.1093/mnras/sty3129. URL https://academic.oup.com/mnras/article/483/1/529/5188687.

[65] Maxwell X. Cai, Jeroen Bédorf, Vikram A. Saletore, et al. DeepGalaxy: Deducing the Properties of Galaxy Mergers from Images Using Deep Neural Networks. *Arxiv preprint*, oct 2020. URL http://arxiv.org/abs/2010.11630.

[66] Gabriela Calistro Rivera, Elisabeta Lusso, Joseph F. Hennawi, and David W. Hogg. AGNfitter: A Bayesian MCMC Approach to Fitting Spectral Energy Distributions of AGNs. *The Astrophysical Journal*, 833(1):98, jun 2016. ISSN 1538-4357. doi: 10.3847/1538-4357/833/1/98. URL http://dx.doi.org/10.3847/1538-4357/833/1/98.

[67] Daniela Calzetti, Lee Armus, Ralph C. Bohlin, et al. The Dust Content and Opacity of Actively Starforming Galaxies. *The Astrophysical Journal*, 533(2):682–695, apr 2000. ISSN 0004-637X. doi: 10.1086/308692. URL http://stacks.iop.org/0004-637X/533/i=2/a=682.

[68] Murray Campbell, A. Joseph Hoane, and Feng Hsiung Hsu. Deep Blue. *Artificial Intelligence*, 134(1-2):57–83, jan 2002. ISSN 00043702. doi: 10.1016/S0004-3702(01)00129-1.

[69] Carolin Cardamone, Kevin Schawinski, Marc Sarzi, et al. Galaxy Zoo Green Peas: Discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, 399(3):1191–1205, nov 2009. ISSN 00358711. doi: 10.1111/j.1365-2966.2009.15383.x.

[70] Nicholas Carlini, Florian Tramer, Eric Wallace, et al. Extracting Training Data from Large Language Models. *arXiv*, dec 2020. URL http://arxiv.org/abs/2012.07805.

[71] Rodrigo Carrasco-Davis, Guillermo Cabrera-Vives, Francisco Förster, et al. Deep learning for image sequence classification of astronomical events. *Publications of the Astronomical Society of the Pacific*, 131(1004), jul 2019. ISSN 00046280. doi: 10.1088/1538-3873/aaef12. URL http://arxiv.org/abs/1807.03869.

[72] Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997. ISSN 08856125. doi: 10.1023/A:1007379606734.

[73] Rich Caruana, Yin Lou, Johannes Gehrke, et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2015-Augus, pages 1721–1730, Sydney, Australia, 2015. ACM. ISBN 9781450336642. doi: 10.1145/2783258.2788613.

[74] Kevin R V Casteels, Christopher J. Conselice, Steven P. Bamford, et al. Galaxy and Mass Assembly (GAMA): Refining the local galaxy merger rate using morphological information. *Monthly Notices of the Royal Astronomical Society*, 445 (2):1157–1169, 2014. ISSN 13652966. doi: 10.1093/mnras/stu1799.

[75] Chaofan Chen, Oscar Li, Chaofan Tao, et al. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. arXiv, jun 2019. URL `http://arxiv.org/abs/1806.10574`.

[76] Ting Yun Cheng, Christopher J. Conselice, Alfonso Aragon-Salamanca, et al. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging. *Monthly Notices of the Royal Astronomical Society*, 493(3):4209–4228, aug 2020. ISSN 13652966. doi: 10.1093/mnras/staa501. URL `http://arxiv.org/abs/1908.03610`.

[77] François Chollet. Building powerful image classification models using very little data, 2016. URL `https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html`.

[78] François Chollet and Others. Keras, 2015. URL `https://github.com/fchollet/keras`.

[79] A. Ciprijanovic, G. F. Snyder, B. Nord, and J. E.G. Peek. DeepMerge: Classifying high-redshift merging galaxies with deep neural networks, apr 2020. URL `http://arxiv.org/abs/2004.11981`.

[80] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, dec

1979. ISSN 1537274X. doi: 10.1080/01621459.1979.10481038. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481038.

[81] Adam D. Cobb, Stephen J. Roberts, and Yarin Gal. Loss-Calibrated Approximate Inference in Bayesian Neural Networks. *arXiv*, may 2018. URL http://arxiv.org/abs/1805.03901.

[82] Liam Connor and Joeri van Leeuwen. Applying Deep Learning to Fast Radio Burst Classification. *The Astronomical Journal*, 156 (6):256, nov 2018. ISSN 1538-3881. doi: 10.3847/1538-3881/aae649. URL http://stacks.iop.org/1538-3881/156/i=6/a=256?key=crossref.57b4c8ce730ab35cd2847321c97e3b8d.

[83] Charlie Conroy, James E. Gunn, and Martin White. The propagation of uncertainties in stellar population synthesis modeling. I. the relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies. *Astrophysical Journal*, 699(1):486–506, jul 2009. ISSN 15384357. doi: 10.1088/0004-637X/699/1/486. URL http://stacks.iop.org/0004-637X/699/i=1/a=486?key=crossref.3ce94867f8570c0a13cac7032dd20cea.

[84] Christopher J. Conselice. The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. *The Astrophysical Journal Supplement Series*, 147(July):1–28, 2003. ISSN 0067-0049. doi: 10.1086/375001. URL https://iopscience.iop.org/article/10.1086/375001.

[85] Christopher J. Conselice, Matthew A. Bershady, and Anna Jangren. The Asymmetry of Galaxies: Physical Morphology for Nearby and High-Redshift Galaxies. *The Astrophysical Journal*, 529(2):886–910, 2000. ISSN 0004-637X. doi: 10.1086/308300. URL https://iopscience.iop.org/article/10.1086/308300.

[86] A. P. Cooper, S. Cole, C. S. Frenk, et al. Galactic stellar haloes in the CDM model. *Monthly Notices of the Royal Astronomical Society*, 406(2): 744–766, aug 2010. ISSN 00358711. doi: 10.1111/j.1365-2966.2010.16740.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2010.16740.x.

[87] Andrew P. Cooper, Richard D'Souza, Guinevere Kauffmann, et al. Galactic accretion and the outer structure of galaxies in the CDM model. *Monthly Notices of the Royal Astronomical Society*, 434(4):3348–3367, 2013. ISSN 00358711. doi: 10.1093/mnras/stt1245.

[88] J. M. Cordes, P. C. C. Freire, D. R. Lorimer, et al. Arecibo Pulsar Survey Using ALFA. I. Survey Strategy and First Discoveries. *The Astrophysical Journal*, 637(1):446–455, jan 2006. ISSN 0004-637X. doi: 10.1086/498335. URL `http://stacks.iop.org/0004-637X/637/i=1/a=446`.

[89] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48):30055–30062, 2020. ISSN 10916490. doi: 10.1073/pnas.1912789117. URL `https://arxiv.org/abs/1911.01429`.

[90] Michael Crawshaw. Multi-task learning with deep neural networks: A survey, sep 2020. ISSN 23318422. URL `http://arxiv.org/abs/2009.09796`.

[91] Ian Czekala, Sean M. Andrews, Kaisey S. Mandel, et al. Constructing a flexible likelihood function for spectroscopic inference. *Astrophysical Journal*, 812(2):128, oct 2015. ISSN 15384357. doi: 10.1088/0004-637X/812/2/128.

[92] Elisabete Da Cunha, Stéphane Charlot, and David Elbaz. A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. *Monthly Notices of the Royal Astronomical Society*, 388(4):1595–1617, aug 2008. ISSN 00358711. doi: 10.1111/j.1365-2966.2008.13535.x.

[93] N. Dalmasso, T. Pospisil, A. B. Lee, et al. Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362, jan 2020. ISSN 22131337. doi: 10.1016/j.ascom.2019.100362.

[94] Andrew Davies, Stephen Serjeant, and Jane M. Bromley. Using convolutional neural networks to identify gravitational lenses in astronomical images. *Monthly Notices of the Royal Astronomical Society*, 487(4):5263–5271, may 2019. ISSN 13652966. doi: 10.1093/mnras/stz1288. URL `http://arxiv.org/abs/1905.04303`.

[95] Jelte T. A. de Jong, Gijs A. Verdoes Kleijn, Danny R. Boxhoorn, et al. The first and second data releases of the Kilo-Degree Survey. *Astronomy and Astrophysics*, 582:A62, 2015. ISSN 0004-6361. doi: 10.1051/0004-6361/201526601. URL http://www.aanda.org/10.1051/0004-6361/201526601.

[96] Renan Alves de Oliveira, Yin Li, Francisco Villaescusa-Navarro, et al. Fast and Accurate Non-Linear Predictions of Universes with Deep Learning. *arXiv*, nov 2020. URL http://arxiv.org/abs/2012.00240.

[97] Gerard De Vaucouleurs. Revised Classification of 1500 Bright Galaxies. *The Astrophysical Journal Supplement Series*, 8:31, 1963.

[98] a Dekel, Y Birnboim, G Engel, et al. Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature*, 457(7228):451–4, 2009. ISSN 1476-4687. doi: 10.1038/nature07648. URL http://www.ncbi.nlm.nih.gov/pubmed/19158792.

[99] Arjun Dey, David J. Schlegel, Dustin Lang, et al. Overview of the DESI legacy imaging surveys. *The Astronomical Journal*, 157(5):168, apr 2019. ISSN 23318422. doi: 10.3847/1538-3881/ab089d. URL http://arxiv.org/abs/1804.08657.

[100] S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv632. URL http://arxiv.org/abs/1503.07077.

[101] Joshua V. Dillon, Ian Langmore, Dustin Tran, et al. Tensorflow distributions, nov 2017. ISSN 23318422. URL http://arxiv.org/abs/1711.10604.

[102] A. D'Isanto and K. L. Polsterer. Photometric redshift estimation via deep learning Generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy and Astrophysics*, 609:A111, jan 2018. ISSN 14320746. doi: 10.1051/0004-6361/201731326.

[103] Wouter Dobbels, Serge Krier, Stephan Pirson, et al. Morphology-assisted galaxy mass-to-light predictions using deep learning. *Astronomy and Astrophysics*, 624:A102, apr 2019. ISSN 14320746. doi: 10.1051/0004-6361/201834575. URL https://www.aanda.org/10.1051/0004-6361/201834575.

[104] H Domínguez Sánchez, M Huertas-Company, M Bernardi, et al. Transfer learning for galaxy morphology from one survey to another. *Monthly Notices of the Royal Astronomical Society*, 484(1):93–100, mar 2019. ISSN 0035-8711. doi: 10.1093/mnras/sty3497. URL https://academic.oup.com/mnras/article/484/1/93/5266389.

[105] H. Doáÿ£inguez Sanchez, M. Huertas-Company, M. Bernardi, et al. Transfer learning for galaxy morphology from one survey to another. *Monthly Notices of the Royal Astronomical Society*, 484(1):93–100, jul 2019. ISSN 13652966. doi: 10.1093/mnras/sty3497. URL http://arxiv.org/abs/1807.00807.

[106] B. T. Draine and Aigen Li. Infrared Emission from Interstellar Dust. IV. The Silicate-Graphite-PAH Model in the Post-Spitzer Era. *The Astrophysical Journal*, 657(2):810–837, mar 2007. ISSN 0004-637X. doi: 10.1086/511055. URL http://stacks.iop.org/0004-637X/657/i=2/a=810.

[107] Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, sep 1987. ISSN 03702693. doi: 10.1016/0370-2693(87)91197-X.

[108] Pierre Alain Duc, Jean Charles Cuillandre, Emin Karabal, et al. The ATLAS3D project - XXIX: The new look of early-type galaxies and surrounding fields disclosed by extremely deep optical images. *Monthly Notices of the Royal Astronomical Society*, 446(1):120–143, jan 2015. ISSN 13652966. doi: 10.1093/mnras/stu2019. URL http://academic.oup.com/mnras/article/446/1/120/1310555/The-ATLAS3D-project-XXIX-The-new-look-of-earlytype.

[109] Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, USA, 2000. ISBN 0471056693.

[110] Dmitry A. Duev, Ashish Mahabal, Quanzhi Ye, et al. DeepStreaks: Identifying fast-moving objects in the Zwicky Transient Facility data with deep learning. *Monthly Notices of the Royal Astronomical Society*, 486(3):4158–4165, apr 2019. ISSN 13652966. doi: 10.1093/mnras/stz1096. URL http://dx.doi.org/10.1093/mnras/stz1096.

[111] Michael W. Dusenberry, Dustin Tran, Edward Choi, et al. Analyzing the role of model uncertainty for electronic health records. In *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, pages

204–213, 2020. ISBN 9781450370462. doi: 10.1145/3368555.3384457. URL http://arxiv.org/abs/1906.03842.

[112] N. L. Eisner, O. Barragán, S. Aigrain, et al. Planet Hunters Tess I: TOI 813, a subgiant hosting a transiting Saturn-sized planet on an 84-day orbit. *Monthly Notices of the Royal Astronomical Society*, 494(1):750–763, sep 2020. ISSN 13652966. doi: 10.1093/mnras/staa138. URL http://arxiv.org/abs/1909.09094.

[113] Nora L. Eisner, Oscar Barragán, Chris Lintott, et al. Planet hunters TESS II: Findings from the first two years of TESS. *Monthly Notices of the Royal Astronomical Society*, 501(4):4669–4690, nov 2021. ISSN 23318422. doi: 10.1093/mnras/staa3739. URL http://arxiv.org/abs/2011.13944.

[114] D. Elbaz, E. Daddi, D. Le Borgne, et al. The reversal of the star formation-density relation in the distant universe. *Astronomy and Astrophysics*, 468(1):33–48, jun 2007. ISSN 00046361. doi: 10.1051/0004-6361:20077525. URL http://www.aanda.org/10.1051/0004-6361:20077525.

[115] Bruce G. Elmegreen, Debra M. Elmegreen, and Luis Montenegro. Optical tracers of spiral wave resonances in galaxies. II - Hidden three-arm spirals in a sample of 18 galaxies. *The Astrophysical Journal Supplement Series*, 79:37, mar 1992. ISSN 0067-0049. doi: 10.1086/191643.

[116] Bruce G. Elmegreen, Debra M. Elmegreen, and Luis Montenegro. Computer analysis of galactic symmetry. *Publications of the Astronomical Society of the Pacific*, 105(688):644, jun 1993. ISSN 0004-6280. doi: 10.1086/133210.

[117] Dawn K. Erb, Alice E. Shapley, Max Pettini, et al. The MassâĂŘMetallicity Relation at z >2. *The Astrophysical Journal*, 644(2):813–828, jun 2006. ISSN 0004-637X. doi: 10.1086/503623.

[118] T. Erben, H. Hildebrandt, L. Miller, et al. CFHTLenS: The Canada-France-Hawaii telescope lensing survey - Imaging data and catalogue products. *Monthly Notices of the Royal Astronomical Society*, 433(3):2545–2563, 2013. ISSN 00358711. doi: 10.1093/mnras/8tt928.

[119] Carlos Esteves. Theoretical aspects of group equivariant neural networks, 2020. URL https://arxiv.org/abs/2004.05154.

[120] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, et al. Robust Physical-World Attacks on Deep Learning Models. In *Conference on Computer Vision and Pattern Recognition*, jul 2018. URL http://arxiv.org/abs/1707.08945.

[121] Heino Falcke and Luciano Rezzolla. Fast radio bursts: The last sign of supramassive neutron stars. *Astronomy and Astrophysics*, 562:A137, feb 2014. ISSN 00046361. doi: 10.1051/0004-6361/201321996.

[122] Jerome J. Fang, S. M. Faber, David C. Koo, and Avishai Dekel. A link between star formation quenching and inner stellar mass density in Sloan digital sky survey central galaxies. *Astrophysical Journal*, 776(1):63, oct 2013. ISSN 15384357. doi: 10.1088/0004-637X/776/1/63.

[123] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *arXiv*, pages 1–31, 2021.

[124] Annette M. N. Ferguson and A. D. Mackey. Substructure and Tidal Streams in the Andromeda Galaxy and its Satellites. In *Tidal Streams in the Local Group and Beyond*, chapter 8, pages 191–217. Springer, Cham, 2016. doi: 10.1007/978-3-319-19336-6_8. URL http://link.springer.com/10.1007/978-3-319-19336-6_8.

[125] Leonardo Ferreira, Christopher J. Conselice, Kenneth Duncan, et al. Galaxy merger rates up to z $\sim$ 3 using a Bayesian deep learning model âĂŤ A major-merger classifier using IllustrisTNG simulation data. *The Astrophysical Journal*, 895(2):115, may 2020. ISSN 23318422. doi: 10.3847/1538-4357/ab8f9b. URL http://arxiv.org/abs/2005.00476.

[126] Debra A. Fischer, Megan E. Schwamb, Kevin Schawinski, et al. Planet Hunters: The first two planet candidates identified by the public using the Kepler public archive data. *Monthly Notices of the Royal Astronomical Society*, 419(4): 2900–2911, feb 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2011.19932.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2011.19932.x.

[127] J. L. Fischer, H. Domínguez Sánchez, and M. Bernardi. SDSS-IV MaNGA PyMorph Photometric and Deep Learning Morphological catalogues and implications for bulge properties and stellar angular momentum. *Monthly Notices of the Royal Astronomical Society*, 483(2):2057–2077, nov 2019. ISSN

13652966. doi: 10.1093/mnras/sty3135. URL https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/sty3135/5188692.

[128] Alex Fitts, Michael Boylan-Kolchin, James S Bullock, et al. No assembly required: Mergers are mostly irrelevant for the growth of low-mass dwarf galaxies. *Monthly Notices of the Royal Astronomical Society*, 479(1):319–331, sep 2018. ISSN 13652966. doi: 10.1093/mnras/sty1488. URL https://academic.oup.com/mnras/article/479/1/319/5033708.

[129] B. Flaugher, H. T. Diehl, K. Honscheid, et al. The Dark Energy Camera. *Astronomical Journal*, 150(5):150, nov 2015. ISSN 00046256. doi: 10.1088/0004-6256/150/5/150.

[130] Brenna Flaugher. The Dark Energy Survey. *International Journal of Modern Physics A*, 20(14):3121–3123, jun 2005. ISSN 0217-751X. doi: 10.1142/S0217751X05025917. URL http://www.worldscientific.com/doi/abs/10.1142/S0217751X05025917.

[131] E. Fonseca, B. C. Andersen, M. Bhardwaj, et al. Nine new repeating fast radio burst sources from CHIME/FRB. *The Astrophysical Journal Letters*, 891(1):L6, jan 2020. ISSN 23318422. doi: 10.3847/2041-8213/ab7208.

[132] Pedro G. Fonseca and Hugo D. Lopes. Calibration of Machine Learning Classifiers for Probability of Default Modelling. *Arxiv preprint*, oct 2017. URL http://arxiv.org/abs/1710.08901.

[133] Fabio Fontanot, Gabriella de Lucia, David Wilman, and Pierluigi Monaco. The other side of bulge formation in a Λ cold dark matter cosmology: Bulgeless galaxies in the local Universe. *Monthly Notices of the Royal Astronomical Society*, 416(1):409–415, sep 2011. ISSN 00358711. doi: 10.1111/j.1365-2966.2011.19047.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2011.19047.x.

[134] Dan Foreman-Mackey, Jonathan Sick, and Ben Johnson. Python-FSPS: Python Bindings To Fsps (v0.1.1), oct 2014. URL https://zenodo.org/record/12157#.Wd0TrUyZMo8.

[135] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee : The MCMC Hammer. *Publications of the Astronomical Society of*

the Pacific, 125(925):306–312, feb 2013. ISSN 00046280. doi: 10.1086/670067. URL http://dx.doi.org/10.1086/670067.

[136] Daniel Foreman-Mackey, Will M. Farr, Manodeep Sinha, et al. emcee v3: A Python ensemble sampling toolkit for affine-invariant MCMC, nov 2019. ISSN 2475-9066. URL http://dx.doi.org/10.21105/joss.01864.

[137] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2019. ISSN 23318422. URL http://arxiv.org/abs/1912.02757.

[138] S. Fotopoulou and S. Paltani. CPz: Classification-aided photometric-redshift estimation. *Astronomy and Astrophysics*, 619:A14, nov 2018. ISSN 14320746. doi: 10.1051/0004-6361/201730763. URL https://www.aanda.org/10.1051/0004-6361/201730763.

[139] P. E. Freeman, R. Izbicki, A. B. Lee, et al. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282–295, 2013. ISSN 00358711. doi: 10.1093/mnras/stt1016.

[140] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2001.

[141] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[142] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete Dropout. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 3581–3590, 2017. URL http://papers.nips.cc/paper/6949-concrete-dropout.

[143] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. URL http://arxiv.org/abs/1703.02910.

[144] Melanie A. Galloway, Kyle W. Willett, Lucy F. Fortson, et al. Galaxy Zoo: The effect of bar-driven fuelling on the presence of an active galactic nucleus in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 448(4):3442–3454, 2015. ISSN 13652966. doi: 10.1093/mnras/stv235.

[145] Robert Geirhos, Claudio Michaelis, Felix A. Wichmann, et al. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, nov 2019. URL http://arxiv.org/abs/1811.12231.

[146] Robert Geirhos, Jörn-Henrik Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, apr 2020. ISSN 25225839. doi: 10.1038/s42256-020-00257-z. URL http://arxiv.org/abs/2004.07780.

[147] C Gini. *Mutuabilità: Contributo Allo Studio Delle Distribuzioni E Delle Relazioni Statistiche*. Bologna, 1912.

[148] Catalina Gómez, Mauricio Neira, Marcela Hernández Hoyos, et al. Classifying image sequences of astronomical transients with deep neural networks. *Monthly Notices of the Royal Astronomical Society*, 499(3):3130–3138, apr 2020. ISSN 13652966. doi: 10.1093/mnras/staa2973. URL http://arxiv.org/abs/2004.13877.

[149] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[150] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 3, pages 2672–2680. Neural information processing systems foundation, 2014. doi: 10.3156/jsoft.29.5_177_2.

[151] Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, 2010. ISSN 21575452. doi: 10.2140/camcos.2010.5.65.

[152] Yjan A. Gordon, Kevin A. Pimbblet, Sugata Kaviraj, et al. The effect of minor and major mergers on the evolution of low excitation radio galaxies. *arXiv*, 878 (2):88–101, apr 2019. ISSN 0004-637X. doi: 10.3847/1538-4357/ab203f. URL http://arxiv.org/abs/1905.00018.

[153] I.S Gradshteyn and I.M Rzyhik. *Table of Integrals, Series, and Products*. Elsevier, 1980. doi: 10.1016/c2013-0-10754-4.

[154] Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems*, 2011.

[155] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *International Conference on Machine Learning*, jun 2017. ISSN 01956574. doi: 10.1007/s001320050185. URL `https://arxiv.org/abs/1706.04599http://arxiv.org/abs/1706.04599`.

[156] Stephen D J Gwyn. The Canada-France-Hawaii telescope legacy survey: Stacked images and catalogs. *Astronomical Journal*, 143(2), 2012. ISSN 00046256. doi: 10.1088/0004-6256/143/2/38.

[157] Ross E. Hart, Steven P. Bamford, Kyle W. Willett, et al. Galaxy Zoo: Comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society*, 461(4):3663–3682, 2016. ISSN 13652966. doi: 10.1093/mnras/stw1588.

[158] Ross E Hart, Steven P Bamford, Kevin R V Casteels, et al. Galaxy Zoo: star formation versus spiral arm number. *Monthly Notices of the Royal Astronomical Society*, 468(March):1850–1863, 2017. doi: 10.1093/mnras/stx581. URL `https://academic.oup.com/mnras/article/468/2/1850/3063901`.

[159] D. Harvey, T. D. Kitching, J. Noah-Vanhoucke, et al. Observing Dark Worlds: A crowdsourcing experiment for dark matter mapping. *Astronomy and Computing*, 5:35–44, jul 2014. ISSN 22131337. doi: 10.1016/j.ascom.2014.04.003.

[160] Trevor J. Hastie and R.J Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1 edition, 1990. ISBN 9781351414234. doi: 10.1201/9780203738535.

[161] Md Abul Hayat, George Stein, Peter Harrington, et al. Self-Supervised Representation Learning for Astronomical Images. *arXiv*, dec 2020. URL `http://arxiv.org/abs/2012.13083`.

[162] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, number 3, pages 770–778. IEEE Computer Society, dec 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90. URL `http://arxiv.org/abs/1512.03385`.

[163] Siyu He, Yin Li, Yu Feng, et al. Learning to predict the cosmological structure formation. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28):13825–13832, jul 2019. ISSN 10916490. doi: 10.1073/pnas.1821458116.

[164] Xin He, Kaiyong Zhao, and Xiaowen Chu. AutoML: A Survey of the State-of-the-Art. *Arxiv preprint*, aug 2019. URL http://arxiv.org/abs/1908.00709.

[165] Kasper Elm Heintz, J. Xavier Prochaska, Sunil Simha, et al. Host galaxy properties and offset distributions of fast radio bursts: Implications for their progenitors, sep 2020. ISSN 23318422. URL http://dx.doi.org/10.3847/1538-4357/abb6fb.

[166] Katrin Heitmann, David Higdon, Charles Nakhleh, and Salman Habib. Cosmic Calibration. *The Astrophysical Journal*, 646(1):L1–L4, jul 2006. ISSN 0004-637X. doi: 10.1086/506448.

[167] David Hendel and Kathryn V. Johnston. Tidal debris morphology and the orbits of satellite galaxies. *Monthly Notices of the Royal Astronomical Society*, 454(3):2472–2485, dec 2015. ISSN 13652966. doi: 10.1093/mnras/stv2035. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stv2035.

[168] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, oct 2017. URL http://arxiv.org/abs/1610.02136.

[169] Tom Henighan, Jared Kaplan, Mor Katz, et al. Scaling Laws for Autoregressive Generative Modeling. *arXiv*, 2020. URL http://arxiv.org/abs/2010.14701.

[170] J. W.T. Hessels, L. G. Spitler, A. D. Seymour, et al. FRB 121102 bursts show complex time-frequency structure. *The Astrophysical Journal Letters*, 876 (2):L23, 2018. ISSN 23318422. doi: 10.3847/2041-8213/ab13ae. URL http://dx.doi.org/10.3847/2041-8213/ab13ae.

[171] Yashar D. Hezaveh, Laurence Perreault Levasseur, and Philip J. Marshall. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature*, 548(7669):555–557, aug 2017. ISSN 14764687. doi: 10.1038/nature23463. URL http://www.nature.com/articles/nature23463.

[172] Greg Hines, Alexandra Swanson, and Margaret Kosmala. Aggregating User Input in Ecology Citizen Science Projects. *Proceedings of the Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence*, pages 3975–3980, 2015.

[173] Geoffrey E. Hinton. What kind of a graphical model is the brain? *IJCAI International Joint Conference on Artificial Intelligence*, pages 1765–1775, 2005. ISSN 10450823.

[174] Geoffrey E. Hinton and Drew van Camp. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the sixth annual conference on Computational learning theory*, 1993.

[175] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527. URL `https://doi.org/10.1162/neco.2006.18.7.1527`.

[176] Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, nov 2014. ISSN 15337928. URL `http://arxiv.org/abs/1111.4246`.

[177] Callie E. Hood, Sheila J. Kannappan, David V. Stark, et al. The Origin of Faint Tidal Features Around Galaxies in the RESOLVE Survey. *The Astrophysical Journal*, 857(2):144, apr 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aab719. URL `http://arxiv.org/abs/1803.05447`.

[178] Philip F. Hopkins, Rachel S. Somerville, Thomas J. Cox, et al. The effects of gas on morphological transformation in mergers: Implications for bulge and disc demographics. *Monthly Notices of the Royal Astronomical Society*, 397(2):802–814, 2009. ISSN 00358711. doi: 10.1111/j.1365-2966.2009.14983.x.

[179] Philip F. Hopkins, Kevin Bundy, Darren Croton, et al. Mergers and bulge formation in λcDM: Which mergers matter? *Astrophysical Journal*, 715(1):202–229, may 2010. ISSN 15384357. doi: 10.1088/0004-637X/715/1/202. URL `http://stacks.iop.org/0004-637X/715/i=1/a=202?key=crossref.d956c23377616fd48555c7b449c86600`.

[180] Philip F. Hopkins, Andrew Wetzel, Dušan Kereš, et al. How to model supernovae in simulations of star and galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 477(2):1578–1603, jun 2018. ISSN 13652966. doi: 10.1093/mnras/sty674.

[181] Philip F Hopkins, Andrew Wetzel, Dusan Keres, et al. FIRE-2 Simulations: Physics versus Numerics in Galaxy Formation. *Monthly Notices of the Royal Astronomical Society*, 480(1):800–863, oct 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1690. URL https://academic.oup.com/mnras/article/480/1/800/5046474.

[182] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. *Bayesian Active Learning for Classification and Preference Learning*. PhD thesis, University of Cambridge, dec 2011. URL http://arxiv.org/abs/1112.5745.

[183] Andrew G. Howard, Menglong Zhu, Bo Chen, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *Arxiv preprint*, apr 2017. URL http://arxiv.org/abs/1704.04861.

[184] B. Hoyle. Measuring photometric redshifts using galaxy images and Deep Neural Networks. *Astronomy and Computing*, 16:34–40, jul 2016. ISSN 22131337. doi: 10.1016/j.ascom.2016.03.006.

[185] Ben Hoyle, Kerstin Paech, Markus Michael Rau, et al. Tuning target selection algorithms to improve galaxy redshift estimates. *Monthly Notices of the Royal Astronomical Society*, 458(4):4498–4511, jun 2016. ISSN 13652966. doi: 10.1093/mnras/stw563. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw563.

[186] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7132–7141. IEEE Computer Society, sep 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00745. URL http://arxiv.org/abs/1709.01507.

[187] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 2261–2269, aug 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.243. URL http://arxiv.org/abs/1608.06993.

[188] Jin Huang and Charles X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3): 299–310, 2005. ISSN 10414347. doi: 10.1109/TKDE.2005.50.

[189] Jonathan Huang, Vivek Rathod, Chen Sun, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:3296–3305, 2017. ISSN 16113349. doi: 10.1109/CVPR.2017.351. URL http://arxiv.org/abs/1611.10012.

[190] M. Huertas-Company, D. Rouan, L. Tasca, et al. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images: I. Method description. *Astronomy and Astrophysics*, 478 (3):971–980, feb 2008. ISSN 00046361. doi: 10.1051/0004-6361:20078625. URL http://www.aanda.org/10.1051/0004-6361:20078625.

[191] M. Huertas-Company, R. Gravet, G. Cabrera-Vives, et al. A catalog of visual-like morphologies in the 5 candels fields using deep learning. *Astrophysical Journal, Supplement Series*, 221(1):8, 2015. ISSN 00670049. doi: 10.1088/0067-0049/221/1/8. URL http://dx.doi.org/10.1088/0067-0049/221/1/8.

[192] M. Huertas-Company, J. R. Primack, A. Dekel, et al. Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range. *The Astrophysical Journal*, 858(2):114, may 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aabfed. URL http://stacks.iop.org/0004-637X/858/i=2/a=114?key=crossref.4e09299072a147355484d972ba56e818.

[193] Marc Huertas-Company, J. A. L Aguerri, M. Bernardi, et al. Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available bayesian automated classification. *Astronomy and Astrophysics*, 525(157):1–13, 2011. ISSN 0004-6361. doi: 10.1051/0004-6361/201015735. URL http://arxiv.org/abs/1010.3018%0Ahttp://dx.doi.org/10.1051/0004-6361/201015735.

[194] Marc Huertas-Company, Vicente Rodriguez-Gomez, Dylan Nelson, et al. The Hubble Sequence at z $\sim$ 0 in the IllustrisTNG simulation with deep learning. *Monthly Notices of the Royal Astronomical Society*, 489(2):1859–1879, mar 2019. ISSN 13652966. doi: 10.1093/mnras/stz2191. URL http://arxiv.org/abs/1903.07625.

[195] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):99–104, may 2007. ISSN 15219615. doi: 10.1109/MCSE.2007.55. URL http://ieeexplore.ieee.org/document/4160265/.

[196] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *International Conference on Learning Representations*, feb 2017. URL http://arxiv.org/abs/1602.07360.

[197] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, et al. Adversarial examples are not bugs, they are features, may 2019. ISSN 23318422. URL http://arxiv.org/abs/1905.02175.

[198] E. E. O. Ishida, R. Beck, S González-Gaitán, et al. Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning. *MNRAS*, 000:1–18, 2018. URL http://arxiv.org/abs/1804.03765.

[199] Niall Jeffrey, François Lanusse, Ofer Lahav, and Jean Luc Starck. Deep learning dark matter map reconstructions from DES SV weak lensing data, aug 2019. ISSN 23318422. URL http://dx.doi.org/10.1093/mnras/staa127.

[200] Shardha Jogee, Nick Z. Scoville, and Jeffrey D. P. Kenney. The Central Region of Barred Galaxies: Molecular Environment, Starbursts, and Secular Evolution. *The Astrophysical Journal, Volume 630, Issue 2, pp. 837-863.*, 630(2):837–863, sep 2005. ISSN 0004-637X. doi: 10.1086/432106. URL http://stacks.iop.org/0004-637X/630/i=2/a=837.

[201] Benjamin D. Johnson, Johnson, and Benjamin D. SEDPY: Modules for storing and operating on astronomical source spectral energy distribution. *Astrophysics Source Code Library*, page ascl:1905.026, 2019. URL https://ui.adsabs.harvard.edu/abs/2019ascl.soft05026J/abstract.

[202] Benjamin D Johnson, Joel L Leja, Charlie Conroy, and Joshua S Speagle. Prospector: Stellar population inference from spectra and SEDs, may 2019. URL https://github.com/bd-j/prospector.

[203] Kathryn V. Johnston, Lars Hernquist, and Michael Bolte. Fossil Signatures of Ancient Accretion Events in the Halo. *The Astrophysical Journal*, 465:278, feb 1996. ISSN 0004-637X. doi: 10.1086/177418. URL http://dx.doi.org/10.1086/177418.

[204] Kathryn V. Johnston, James S. Bullock, Sanjib Sharma, et al. Tracing Galaxy Formation with Stellar Halos. II. Relating Substructure in Phase and Abundance Space to Accretion Histories. *The Astrophysical Journal*, 689(2): 936–957, dec 2008. ISSN 0004-637X. doi: 10.1086/592228. URL `http://stacks.iop.org/0004-637X/689/i=2/a=936`.

[205] Eric Jones, Travis Oliphant, Peterson Pearu, and Others. SciPy: Open Source Scientific Tools for Python, 2001. URL `http://www.scipy.org/`.

[206] W A Joye and E Mandel. New Features of SAOImage DS9. In H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, editors, *Astronomical Data Analysis Software and Systems XII*, volume 295 of *Astronomical Society of the Pacific Conference Series*, page 489, jan 2003.

[207] Erin Kado-Fong, Jenny E. Greene, David Hendel, et al. Tidal Features at 0.05 $<$ z $<$ 0.45 in the Hyper Suprime-Cam Subaru Strategic Program: Properties and Formation Channels. *eprint arXiv:1805.05970*, may 2018. URL `http://arxiv.org/abs/1805.05970`.

[208] Nick Kaiser, William Burgett, Ken Chambers, et al. The Pan-STARRS wide-field optical/NIR imaging survey. In Larry M. Stepp, Roberto Gilmozzi, and Helen J. Hall, editors, *Ground-based and Airborne Telescopes III*, volume 7733, page 77330E. International Society for Optics and Photonics, jul 2010. ISBN 9780819482235. doi: 10.1117/12.859188. URL `http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.859188`.

[209] Agris J Kalnajs. Theory Of Spiral Structure. *Symposium - International Astronomical Union*, 100:109–116, 1983. doi: 10.1017/S007418090003254X.

[210] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models, jan 2020. ISSN 23318422. URL `http://arxiv.org/abs/2001.08361`.

[211] Jeyhan S Kartaltepe, D B Sanders, E Le Floc'h, et al. A MULTIWAVELENGTH STUDY OF A SAMPLE OF 70 μm SELECTED GALAXIES IN THE COSMOS FIELD. II. THE ROLE OF MERGERS IN GALAXY EVOLUTION. *The Astrophysical Journal*, 721(1):98–123, 2010. ISSN 0004-637X. doi: 10.1088/0004-637X/721/1/98. URL `https://iopscience.iop.org/article/10.1088/0004-637X/721/1/98`.

[212] Jeyhan S Kartaltepe, Mark Mozena, Dale Kocevski, et al. Candels Visual Classifications: Scheme, Data Release, and First Results. *The Astrophysical Journal Supplement Series*, 221(17pp):11, 2015. ISSN 1538-4365. doi: 10.1088/0067-0049/221/1/11. URL http://dx.doi.org/10.1088/0067-0049/221/1/11.

[213] G. Kasieczka, T. Plehn, A. Butter, et al. The machine learning landscape of top taggers. *SciPost Phys*, 7(1):14, feb 2019. ISSN 23318422. doi: 10.21468/scipostphys.7.1.014.

[214] Noah Kasmanoff, Jeremy Tinker, Francisco Villaescusa-Navarro, and Shirley Ho. dm2gal: Mapping dark matter to galaxies with neural networks. In *NeurIPS 2020 Machine Learning and the Physical Sciences Workshop*, 2020.

[215] J. I. Katz. Fast radio bursts - A brief review: Some questions, fewer answers. *Modern Physics Letters A*, 31(14):1630013, 2016. ISSN 02177323. doi: 10.1142/S0217732316300135. URL http://arxiv.org/abs/1604.01799.

[216] Neal Katz, Dusan Keres, Romeel Davé, and David H. Weinberg. How do Galaxies Get Their Gas? In *The IGM/Galaxy Connection*, pages 185–192. Springer, Dordrecht, 2003. doi: 10.1007/978-94-010-0115-1_34.

[217] S. Kaviraj, S. Cohen, R. A. Windhorst, et al. The insignificance of major mergers in driving star formation at z = 2. *Monthly Notices of the Royal Astronomical Society: Letters*, 429(1), 2013. ISSN 17453925. doi: 10.1093/mnrasl/sls019.

[218] Sugata Kaviraj. The significant contribution of minor mergers to the cosmic star formation budget. *Monthly Notices of the Royal Astronomical Society: Letters*, 437(1), 2013. ISSN 17453925. doi: 10.1093/mnrasl/slt136.

[219] Sugata Kaviraj. The importance of minor-merger-driven star formation and black hole growth in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 440(4):2944–2952, 2014. ISSN 13652966. doi: 10.1093/mnras/stu338.

[220] E. F. Keane, S. Johnston, S. Bhandari, et al. The host galaxy of a fast radio burst. *Nature*, 530(7591):453–456, feb 2016. ISSN 14764687. doi: 10.1038/nature17140. URL http://www.nature.com/articles/nature17140.

[221] William C. Keel, W. Peter Maksym, Vardha N. Bennert, et al. HST Imaging of fading AGN candidates. I. Host-galaxy properties and origin of the extended gas. *Astronomical Journal*, 149(5):155, may 2015. ISSN 00046256. doi: 10.1088/0004-6256/149/5/155.

[222] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS 2017*, pages 5574–5584, Long Beach, CA, USA, mar 2017. ISBN 1070-9878. doi: 10.1109/TDEI.2009.5211872. URL http://arxiv.org/abs/1703.04977.

[223] Noble Kennamer, David Kirkby, Alex Ihler, and Javier Sánchez. ContextNet: Deep learning for Star Galaxy Classification. In Andreas Dy, Jennifer and Krause, editor, *Proceedings of the 35th International Conference on Machine Learning*, pages 2582–2590, Stockholm, 2018. PMLR. URL http://proceedings.mlr.press/v80/kennamer18a.html.

[224] Noble Kennamer, Emille E.O. Ishida, Santiago Gonzalez-Gaitan, et al. Active learning with RESSPECT: Resource allocation for extragalactic astronomical transients. *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*, pages 3115–3124, 2020. doi: 10.1109/SSCI47803.2020.9308300. URL http://arxiv.org/abs/2010.05941.

[225] Nicholas S. Kern, Adrian Liu, Aaron R. Parsons, et al. Emulating Simulations of Cosmic Dawn for 21 cm Power Spectrum Constraints on Cosmology, Reionization, and X-Ray Heating, may 2017. ISSN 0004-637X. URL http://dx.doi.org/10.3847/1538-4357/aa8bb4.

[226] Joshua Kerrigan, Paul la Plante, Saul Kohn, et al. Optimizing sparse RFI prediction using deep learning. *Monthly Notices of the Royal Astronomical Society*, 488(2):2605–2615, feb 2019. ISSN 13652966. doi: 10.1093/mnras/stz1865. URL http://arxiv.org/abs/1902.08244.

[227] Asad Khan, E. A. Huerta, Sibo Wang, et al. Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics*, 795:248–258, 2019. ISSN 03702693. doi: 10.1016/j.physletb.2019.06.009. URL https://www.sciencedirect.com/science/article/pii/S0370269319303879?via%3Dihub.

[228] S. Khochfar and J. Silk. Dry mergers: A crucial test for galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 397(1):506–510, 2009. ISSN 00358711. doi: 10.1111/j.1365-2966.2009.14958.x.

[229] Edward J. Kim and Robert J. Brunner. Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464(4):4463–4475, 2017. ISSN 13652966. doi: 10.1093/mnras/stw2672.

[230] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[231] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2014.

[232] T. D. Kitching, S. T. Balan, S. Bridle, et al. Image analysis for cosmology: Results from the GREAT10 Galaxy Challenge. *Monthly Notices of the Royal Astronomical Society*, 423(4):3163–3208, jul 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2012.21095.x. URL `https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2012.21095.x`.

[233] T. D. Kitching, J. Rhodes, C. Heymans, et al. Image analysis for cosmology: Shape measurement challenge review and results from the Mapping Dark Matter challenge. *Astronomy and Computing*, 10:9–21, apr 2015. ISSN 22131337. doi: 10.1016/j.ascom.2014.12.004.

[234] Simon A.A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, et al. A probabilistic U-net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems*, 2018-Decem:6965–6975, jun 2018. ISSN 10495258. URL `http://arxiv.org/abs/1806.05034`.

[235] John Kormendy and Robert C. Kennicutt. Secular Evolution and the Formation of Pseudobulges in Disk Galaxies. *Annual Review of Astronomy and Astrophysics*, 42(1):603–683, sep 2004. ISSN 0066-4146. doi: 10.1146/annurev.astro.42.053102.134024. URL `http://www.annualreviews.org/doi/10.1146/annurev.astro.42.053102.134024`.

[236] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/$\sim$kriz/cifar.html.

[237] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, jun 2017. ISSN 15577317. doi: 10.1145/3065386. URL https://dl.acm.org/doi/10.1145/3065386.

[238] Julian H. Krolik and Mitchell C. Begelman. Molecular Tori in Seyfert Galaxies: Feeding the Monster and Hiding It. *The Astrophysical Journal*, 329:702–711, 1988.

[239] Pavel Kroupa. On the variation of the initial mass function. *Monthly Notices of the Royal Astronomical Society*, 322(2):231–246, apr 2001. ISSN 00358711. doi: 10.1046/j.1365-8711.2001.04022.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1046/j.1365-8711.2001.04022.x.

[240] Sandor J. Kruk, Chris J. Lintott, Brooke D. Simmons, et al. Galaxy Zoo: Finding offset discs and bars in SDSS galaxies. *Monthly Notices of the Royal Astronomical Society*, 469(3):3363–3373, aug 2017. ISSN 13652966. doi: 10.1093/mnras/stx1026. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stx1026.

[241] Sandor J. Kruk, Chris J. Lintott, Steven P. Bamford, et al. Galaxy Zoo: Secular evolution of barred galaxies from structural decomposition of multiband images. *Monthly Notices of the Royal Astronomical Society*, 473(4):4731–4753, feb 2018. ISSN 13652966. doi: 10.1093/mnras/stx2605. URL http://academic.oup.com/mnras/article/473/4/4731/4411828.

[242] O. Lahav, A. Naim, R. J. Buta, et al. Galaxies, human eyes, and artificial neural networks. *Science*, 267(5199):859–862, feb 1995. ISSN 00368075. doi: 10.1126/science.267.5199.859.

[243] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Thirty-first Conference on Neural Information Processing Systems (NeurIPS 2017*, 2017. ISBN 0045-6535. doi: doi:10.1007/s11098-011-9715-4. URL http://arxiv.org/abs/1612.01474.

[244] C Lanczos. Trigonometric Interpolation of Empirical and Analytical Functions. *Journal of Mathematics and Physics*, 17(1-4):123–199, apr 1938. ISSN 00971421. doi: 10.1002/sapm1938171123. URL http://doi.wiley.com/10.1002/sapm1938171123.

[245] Kate Land, Anže Slosar, Chris Lintott, et al. Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 388(4):1686–1692, 2008. ISSN 00358711. doi: 10.1111/j.1365-2966.2008.13490.x.

[246] François Lanusse, Quanbin Ma, Nan Li, et al. CMU DeepLens: Deep learning for automatic image-based galaxy-galaxy strong lens finding. *Monthly Notices of the Royal Astronomical Society*, 473(3):3895–3906, 2018. ISSN 13652966. doi: 10.1093/mnras/stx1665. URL http://arxiv.org/abs/1703.02642.

[247] Andris Lauberts and E.∼A. Valentijn. The surface photometry catalogue of the ESO-Uppsala galaxies, 1989.

[248] R. Laureijs, J. Amiaux, S. Arduini, et al. Euclid Definition Study Report. *Arxiv preprint*, oct 2011. ISSN 15507998. doi: 10.1088/0264-9381/18/14/306. URL http://arxiv.org/abs/1110.3193.

[249] Earl Lawrence, Scott Vander Wiel, Casey Law, et al. The Nonhomogeneous Poisson Process for Fast Radio Burst Rates. *The Astronomical Journal*, 154 (3):117, nov 2017. ISSN 1538-3881. doi: 10.3847/1538-3881/aa844e. URL http://arxiv.org/abs/1611.00458.

[250] Steve Lawrence, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997. ISSN 10459227. doi: 10.1109/72.554195.

[251] Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic Gaussian process regression. In *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pages 489–496, New York, New York, USA, 2005. ACM Press. ISBN 1595931805. doi: 10.1145/1102351.1102413. URL http://portal.acm.org/citation.cfm?doid=1102351.1102413.

[252] Yann LeCun, Bernhard Boser, John S Denker, et al. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[253] Yann A. LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. ISSN 0028-0836. doi: 10.1038/nature14539. URL http://dx.doi.org/10.1038/nature1453910.1038/nature14539.

[254] Jaehyun Lee and Sukyoung K. Yi. Formation and assembly history of stellar components in galaxies as a function of stellar and halo mass. *The Astrophysical Journal*, 836(2):1–11, 2017. ISSN 0004-637X. doi: 10.3847/1538-4357/aa5b87. URL http://dx.doi.org/10.3847/1538-4357/aa5b87.

[255] Joel Leja, Benjamin D. Johnson, Charlie Conroy, et al. Deriving Physical Properties from Broadband Photometry with Prospector: Description of the Model and a Demonstration of its Accuracy Using 129 Galaxies in the Local Universe. *The Astrophysical Journal*, 837(2):170, sep 2017. doi: 10.3847/1538-4357/aa5ffe. URL http://dx.doi.org/10.3847/1538-4357/aa5ffe.

[256] Laurence Perreault Levasseur, Yashar D. Hezaveh, and Risa H. Wechsler. Uncertainties in Parameters Estimated with Neural Networks: Application to Strong Gravitational Lensing. *The Astrophysical Journal*, 850(1):L7, nov 2017. ISSN 20418213. doi: 10.3847/2041-8213/aa9704. URL http://stacks.iop.org/2041-8205/850/i=1/a=L7?key=crossref.dd8f01b687a77b74ce33239cdb39c453.

[257] B. L'Huillier, F. Combes, and B. Semelin. Mass assembly of galaxies: Smooth accretion versus mergers. *Astronomy and Astrophysics*, 544:A68, aug 2012. ISSN 00046361. doi: 10.1051/0004-6361/201117924.

[258] Lisha Li, Kevin Jamieson, Giulia DeSalvo, et al. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:1–52, mar 2018. ISSN 15337928. URL http://arxiv.org/abs/1603.06560.

[259] Lihwai Lin, David R. Patton, David C. Koo, et al. The Redshift Evolution of Wet, Dry, and Mixed Galaxy Mergers from Close Galaxy Pairs in the DEEP2 Galaxy Redshift Survey. *The Astrophysical Journal*, 681(1):232–243, 2008. ISSN 0004-637X. doi: 10.1086/587928.

[260] Lihwai Lin, Michael C. Cooper, Hung Yu Jian, et al. Where do wet, dry, and mixed galaxy mergers occur? a study of the environments of close galaxy pairs in the DEEP2 Galaxy Redshift Survey. *Astrophysical Journal*, 718(2):1158–1170, aug 2010. ISSN 15384357. doi: 10.1088/0004-637X/718/2/1158.

[261] Lihwai Lin, Bau-Ching Hsieh, Hsi-An Pan, et al. SDSS-IV MaNGA: Inside-out versus Outside-in Quenching of Galaxies in Different Local Environments. *The Astrophysical Journal*, 872(1):50, jan 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/aafa84. URL http://arxiv.org/abs/1901.05126.

[262] Eric V. Linder. Detecting helium reionization with fast radio bursts. *Physical Review D*, 101(10), jan 2020. ISSN 24700029. doi: 10.1103/PhysRevD.101.103019. URL http://dx.doi.org/10.1103/PhysRevD.101.103019.

[263] Manasvi Lingam and Abraham Loeb. Fast radio bursts from extragalactic light sails. *The Astrophysical Journal Letters*, 837(2):L23, jan 2017. ISSN 23318422. doi: 10.3847/2041-8213/aa633e.

[264] Timothy K. Lingard, Karen L. Masters, Coleman Krawczyk, et al. Galaxy Zoo Builder: Four Component Photometric decomposition of Spiral Galaxies Guided by Citizen Science. *arXiv*, jun 2020. ISSN 0004-637X. doi: 10.3847/1538-4357/ab9d83. URL http://arxiv.org/abs/2006.10450.

[265] Chris J. Lintott, Kevin Schawinski, Anže Slosar, et al. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, sep 2008. ISSN 00358711. doi: 10.1111/j.1365-2966.2008.13689.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2008.13689.x.

[266] Chris J. Lintott, Kevin Schawinski, William Keel, et al. Galaxy Zoo: Hanny's Voorwerp, a quasar light echo? *Monthly Notices of the Royal Astronomical Society*, 399(1):129–140, oct 2009. ISSN 00358711. doi: 10.1111/j.1365-2966.2009.15299.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2009.15299.x.

[267] Peng Liu, Hui Zhang, and Kie B. Eom. Active Deep Learning for Classification of Hyperspectral Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2):712–724, feb 2017. ISSN 1939-1404. doi: 10.1109/JSTARS.2016.2598859. URL http://ieeexplore.ieee.org/document/7568999/.

[268] D R Lorimer, M Bailes, M A McLaughlin, et al. A bright millisecond radio burst of extragalactic origin. *Science*, 318(5851):777–780, nov 2007. ISSN 00368075.

doi: 10.1126/science.1147532. URL http://www.ncbi.nlm.nih.gov/pubmed/17901298.

[269] Duncan R. Lorimer. A decade of fast radio bursts, oct 2018. ISSN 23973366. URL http://arxiv.org/abs/1811.00195.

[270] Jennifer M. Lotz, Joel Primack, and Piero Madau. A New Nonparametric Approach to Galaxy Morphological Classification. *The Astronomical Journal*, 128(1):163–182, 2004. ISSN 0004-6256. doi: 10.1086/421849. URL http://stacks.iop.org/1538-3881/128/i=1/a=163.

[271] Jennifer M. Lotz, M. Davis, S. M. Faber, et al. The Evolution of Galaxy Mergers and Morphology at z < 1.2 in the Extended Groth Strip. *The Astrophysical Journal*, 672(1):177–197, 2008. ISSN 0004-637X. doi: 10.1086/523659. URL http://stacks.iop.org/0004-637X/672/i=1/a=177.

[272] Jennifer M. Lotz, Patrik Jonsson, T. J. Cox, and Joel R. Primack. Galaxy merger morphologies and time-scales from simulations of equal-mass gas-rich disc mergers. *Monthly Notices of the Royal Astronomical Society*, 391(3):1137–1162, 2008. ISSN 00358711. doi: 10.1111/j.1365-2966.2008.14004.x.

[273] Jennifer M. Lotz, Patrik Jonsson, T. J. Cox, et al. The major and minor galaxy merger rates at z < 1.5. *Astrophysical Journal*, 742(2):103, 2011. ISSN 15384357. doi: 10.1088/0004-637X/742/2/103. URL http://stacks.iop.org/0004-637X/742/i=2/a=103?key=crossref.edf80af2565aabae46c3a5f4fde1177b.

[274] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158, 2012. doi: 10.1145/2339530.2339556.

[275] LSST Science Collaboration, Paul A. Abell, Julius Allison, et al. LSST Science Book, Version 2.0. Technical report, LSST Science Collaborations, dec 2009. URL http://arxiv.org/abs/0912.0201.

[276] Jie Lu, Vahid Behbood, Peng Hao, et al. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80(2002):14–23, 2015. ISSN 09507051. doi: 10.1016/j.knosys.2015.01.010. URL http://arxiv.org/abs/1709.07417.

[277] Robert Lupton, MichaelÂăR. Blanton, George Fekete, et al. Preparing RedâĂŘ-GreenâĂŘBlue Images from CCD Data. *Publications of the Astronomical Society of the Pacific*, 116(816):133–137, 2004. ISSN 0004-6280. doi: 10.1086/382245.

[278] Yuri Lyubarsky. A model for fast extragalactic radio bursts. *Monthly Notices of the Royal Astronomical Society: Letters*, 442(1): L9–L13, jul 2014. ISSN 17453933. doi: 10.1093/mnrasl/slu046. URL http://academic.oup.com/mnrasl/article/442/1/L9/2889071/A-model-for-fast-extragalactic-radio-bursts.

[279] Zhixian Ma, Haiguang Xu, Jie Zhu, et al. A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the BestâĂŞHeckman Sample. *The Astrophysical Journal Supplement Series*, 240(2):34, 2019. ISSN 0067-0049. doi: 10.3847/1538-4365/aaf9a2. URL http://dx.doi.org/10.3847/1538-4365/aaf9a2.

[280] David J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, jul 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590. URL http://www.mitpressjournals.org/doi/10.1162/neco.1992.4.4.590.

[281] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[282] Piero Madau and Mark Dickinson. Cosmic Star-Formation History. *Annual Review of Astronomy and Astrophysics*, 52(1):415–486, aug 2014. ISSN 0066-4146. doi: 10.1146/annurev-astro-081811-125615. URL http://www.annualreviews.org/doi/10.1146/annurev-astro-081811-125615.

[283] Piero Madau, Lucia Pozzetti, and Mark Dickinson. The Star Formation History of Field Galaxies. *The Astrophysical Journal*, 498(1):106–116, may 1998. ISSN 0004-637X. doi: 10.1086/305523.

[284] D. F. Malin and D. Carter. A catalog of elliptical galaxies with shells. *The Astrophysical Journal*, 274:534, nov 1983. ISSN 0004-637X. doi: 10.1086/161467. URL http://adsabs.harvard.edu/doi/10.1086/161467.

[285] A. I. Malz, R. Hložek, T. Allam, et al. The photometric lsst astronomical time-series classification challenge (plasticc): Selection of a performance metric for classification probabilities balancing diverse science goals, sep 2018. ISSN 23318422. URL http://dx.doi.org/10.3847/1538-3881/ab3a2f.

[286] Claudia Maraston. Evolutionary population synthesis: Models, analysis of the ingredients and application to high-z galaxies. *Monthly Notices of the Royal Astronomical Society*, 362(3):799–825, sep 2005. ISSN 00358711. doi: 10.1111/j.1365-2966.2005.09270.x.

[287] Berta Margalef-Bentabol, Marc Huertas-Company, Tom Charnock, et al. Detecting outliers in astronomical images with deep generative networks. *Monthly Notices of the Royal Astronomical Society*, 496(2), 2020. ISSN 23318422. doi: 10.1093/mnras/staa1647. URL https://arxiv.org/abs/2003.08263.

[288] G. Martin, S. Kaviraj, J. E.G. Devriendt, et al. The role of mergers in driving morphological transformation over cosmic time. *Monthly Notices of the Royal Astronomical Society*, 480(2):2266–2283, oct 2018. ISSN 13652966. doi: 10.1093/MNRAS/STY1936.

[289] Ana Martinazzo, Mateus Espadoto, and Nina S.T. Hirata. Deep learning for astronomical object classification: A case study. *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 5(Visigrapp):87–95, 2020. doi: 10.5220/0008939800870095.

[290] David Martínez-Delgado, R. Jay Gabany, Ken Crawford, et al. Stellar Tidal Streams in Spiral Galaxies of the Local Volume: a Pilot Survey With Modest Aperture Telescopes. *The Astronomical Journal*, 140(4):962–967, 2010. ISSN 0004-6256. doi: 10.1088/0004-6256/140/4/962. URL http://stacks.iop.org/1538-3881/140/i=4/a=962?key=crossref.9ac7f8e5ec824998edae23efb016c114.

[291] Karen L. Masters. Twelve years of Galaxy Zoo. *Proceedings of the International Astronomical Union*, 14(S353):205–212, oct 2019. ISSN 17439221. doi: 10.1017/S1743921319008615. URL http://arxiv.org/abs/1910.08177.

[292] Karen L. Masters, Robert Nichol, Steven Bamford, et al. Galaxy Zoo: Dust in spiral galaxies. *Monthly Notices of the Royal Astronomical Society*, 404

(2):792–810, may 2010. ISSN 00358711. doi: 10.1111/j.1365-2966.2010.16335.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2010.16335.x.

[293] Karen L. Masters, Robert C. Nichol, Ben Hoyle, et al. Galaxy Zoo: Bars in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 411(3): 2026–2034, mar 2011. ISSN 00358711. doi: 10.1111/j.1365-2966.2010.17834.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2010.17834.x.

[294] Karen L. Masters, Robert C. Nichol, Martha P. Haynes, et al. Galaxy Zoo and ALFALFA: Atomic gas and the regulation of star formation in barred disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 424(3): 2180–2192, aug 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2012.21377.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2012.21377.x.

[295] Karen L. Masters, Chris J. Lintott, Ross E. Hart, et al. Galaxy Zoo: Unwinding the winding problem - Observations of spiral bulge prominence and arm pitch angles suggest local spiral galaxies are winding. *Monthly Notices of the Royal Astronomical Society*, 487(2):1808–1820, apr 2019. ISSN 13652966. doi: 10.1093/mnras/stz1153. URL http://arxiv.org/abs/1904.11436.

[296] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017. ISSN 2475-9066. doi: 10.21105/joss.00205.

[297] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Arxiv preprint*, feb 2020. URL http://arxiv.org/abs/1802.03426.

[298] Wes McKinney. Data Structures for Statistical Computing in Python, 2010. URL http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

[299] Matthew McQuinn. Locating the "missing" baryons with extragalactic dispersion measure estimates. *Astrophysical Journal Letters*, 780(2):L33, jan 2014. ISSN 20418205. doi: 10.1088/2041-8205/780/2/L33.

[300] S. Mereghetti, V. Savchenko, C. Ferrigno, et al. INTEGRAL discovery of a burst with associated radio emission from the magnetar SGR 1935+2154. *The Astrophysical Journal Letters*, 898(2):L29, may 2020. ISSN 23318422. doi: 10.3847/2041-8213/aba2cf.

[301] Allison Merritt, Pieter van Dokkum, Roberto Abraham, and Jielai Zhang. The DRAGONLY Nearby Galaxies Survey. I. Substantial Variation in the Diffuse Stellar Halos Around Spiral Galaxies. *The Astrophysical Journal*, 830(2):62, oct 2016. ISSN 1538-4357. doi: 10.3847/0004-637x/830/2/62.

[302] D. Michilli, K. W. Masui, R. Mckinven, et al. An Analysis Pipeline for CHIME/FRB Full-array Baseband Data. *The Astrophysical Journal*, 910(2): 147, oct 2021. ISSN 0004-637X. doi: 10.3847/1538-4357/abe626. URL http://arxiv.org/abs/2010.06748.

[303] J Christopher Mihos, John Dubinski, and Lars Hernquist. Tidal Tales Two: The Effect of Dark Matter Halos on Tidal Tail Morphology and Kinematics. *Astrophysical Journal v.494*, 494:183, 1998. ISSN 15384357. doi: 10.1086/305179. URL https://iopscience.iop.org/article/10.1086/305179/meta.

[304] A. S. Miller. A review of neural network applications in Astronomy. *Vistas in Astronomy*, 36(PART 2):141–161, jan 1993. ISSN 00836656. doi: 10.1016/0083-6656(93)90118-4.

[305] Marvin Minsky and Seymour A Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[306] A. Miskolczi, D. J. Bomans, and R.-J. Dettmar. Tidal streams around galaxies in the SDSS DR7 archive. *Astronomy and Astrophysics*, 536:A66, 2011. ISSN 0004-6361. doi: 10.1051/0004-6361/201116716. URL http://www.aanda.org/10.1051/0004-6361/201116716.

[307] Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Methods for interpreting and understanding deep neural networks, feb 2018. ISSN 10512004. URL https://www.sciencedirect.com/science/article/pii/S1051200417302385.

[308] Gustavo Morales, David Martínez-Delgado, Eva K. Grebel, et al. Systematic search for tidal features around nearby galaxies: I. Enhanced SDSS imaging of the Local Volume. *Astronomy and Astrophysics*, 614:A143, jun 2018. ISSN

14320746. doi: 10.1051/0004-6361/201732271. URL http://dx.doi.org/10.1051/0004-6361/201732271.

[309] Eric A. Moreno, Olmo Cerri, Javier M. Duarte, et al. JEDI-net: a jet identification algorithm based on interaction networks. *European Physical Journal C*, 80(1), aug 2020. ISSN 14346052. doi: 10.1140/epjc/s10052-020-7608-4. URL http://dx.doi.org/10.1140/epjc/s10052-020-7608-4.

[310] Kevin P Murphy. *Machine Learning: A Probabilisitic Perspective*. MIT Press, Boston, MA, 2012.

[311] Zhang S. N., Xiong S. L., K. Li, C., et al. Insight-HXMT X-ray and hard X-ray detection of the double peaks of the Fast Radio Burst from SGR 1935+2154, 2020. URL https://ui.adsabs.harvard.edu/abs/2020GCN.27675....1Z/abstract.

[312] A. Naim, O. Lahav, R. J. Buta, et al. A comparative study of morphological classifications of APM galaxies. *Monthly Notices of the Royal Astronomical Society*, 274(4):1107–1125, jun 1995. ISSN 0035-8711. doi: 10.1093/mnras/274.4.1107. URL https://academic.oup.com/mnras/article/274/4/1107/1254562/A-comparative-study-of-morphological.

[313] Preethi B. Nair and Roberto G. Abraham. a Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 186(2):427–456, 2010. ISSN 0067-0049. doi: 10.1088/0067-0049/186/2/427. URL http://arxiv.org/abs/1001.2401.

[314] Radford M Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

[315] Radford M. Neal. MCMC using hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011. ISBN 9781420079425. doi: 10.1201/b10905-6.

[316] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2011.

[317] Maia Nenkova, Matthew M. Sirocky, Zeljko Ivezic, and Moshe Elitzur. AGN Dusty Tori. I. Handling of Clumpy Media. *The Astrophysical Journal*, 685

(1):147–159, jun 2008. ISSN 0004-637X. doi: 10.1086/590482. URL http://dx.doi.org/10.1086/590482.

[318] Elias Chaibub Neto. Detecting Learning vs Memorization in Deep Neural Networks using Shared Structure Validation Sets, 2018. ISSN 23318422. URL https://arxiv.org/abs/1802.07714.

[319] G. Neugebauer, H. J. Habing, R. van Duinen, et al. The Infrared Astronomical Satellite (IRAS) mission. *The Astrophysical Journal*, 278:L1, mar 1984. ISSN 0004-637X. doi: 10.1086/184209.

[320] R. Nevin, L. Blecha, J. Comerford, and J. Greene. Accurate Identification of Galaxy Mergers with Imaging. *The Astrophysical Journal*, 872(1):76, 2019. ISSN 23318422. doi: 10.3847/1538-4357/aafd34. URL http://dx.doi.org/10.3847/1538-4357/aafd34.

[321] David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *IEEE International Conference on Neural Networks - Conference Proceedings*, volume 1, pages 55–60. IEEE, 1994. doi: 10.1109/icnn.1994.374138.

[322] S. Noll, D. Burgarella, E. Giovannoli, et al. Analysis of galaxy spectral energy distributions from far-UV to far-IR with CIGALE: Studying a SINGS test sample. *Astronomy and Astrophysics*, 507(3):1793–1813, sep 2009. ISSN 00046361. doi: 10.1051/0004-6361/200912497. URL http://dx.doi.org/10.1051/0004-6361/200912497.

[323] S. C. Odewahn, S. H. Cohen, R. A. Windhorst, and Ninan Sajeeth Philip. Automated Galaxy Morphology: A Fourier Approach. *The Astrophysical Journal*, 568(2):539–557, oct 2002. ISSN 0004-637X. doi: 10.1086/339036. URL http://dx.doi.org/10.1086/339036.

[324] Travis E Oliphant. A Bayesian perspective on estimating mean , variance, and standard-deviation from data. *Bringham Young University ScholarsArchive*, dec 2006. URL https://scholarsarchive.byu.edu/facpub/278.

[325] Tom O'Malley, Elie Bursztein, James Long, et al. Keras Tuner, 2019. URL https://github.com/keras-team/keras-tuner.

[326] Nikita Orlov, Lior Shamir, Tomasz Macura, et al. WND-CHARM: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693, 2008. ISSN 01678655. doi: 10.1016/j.patrec.2008.04.013.

[327] Hugh P. Osborn, Megan Ansdell, Yani Ioannou, et al. Rapid classification of TESS planet candidates with convolutional neural networks. *Astronomy and Astrophysics*, 633, feb 2020. ISSN 14320746. doi: 10.1051/0004-6361/201935345. URL http://arxiv.org/abs/1902.08544.

[328] P. Padovani, D. M. Alexander, R. J. Assef, et al. Active galactic nuclei: what's in a name? *Astronomy and Astrophysics Review*, 25(1), 2017. ISSN 09354956. doi: 10.1007/s00159-017-0102-9. URL http://dx.doi.org/10.1007/s00159-017-0102-9.

[329] David B Parker. Learning-logic: Casting the cortex of the human brain in silicon. Technical report, Cambridge, Mass. Center for Computational Research in Economics and Management Science, 1985. URL https://www.worldcat.org/title/learning-logic-casting-the-cortex-of-the-human-brain-in-silicon/oclc/930980029?referer=di&ht=edition.

[330] Johanna Pasquet, E. Bertin, M. Treyer, et al. Photometric redshifts from SDSS images using a convolutional neural network. *Astronomy and Astrophysics*, 621:A26, jan 2019. ISSN 14320746. doi: 10.1051/0004-6361/201833617. URL https://www.aanda.org/10.1051/0004-6361/201833617.

[331] Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H Wallach, H Larochelle, A Beygelzimer, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[332] C. Patel, D. Agarwal, M. Bhardwaj, et al. PALFA Single-pulse Pipeline: New Pulsars, Rotating Radio Transients, and a Candidate Fast Radio Burst. *The Astrophysical Journal*, 869(2):181, aug 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aaee65. URL http://dx.doi.org/10.3847/1538-4357/aaee65.

[333] M. M. Pawlik, V. Wild, C. J. Walcher, et al. Shape asymmetry: A morphological indicator for automatic detection of galaxies in the post-coalescence merger stages. *Monthly Notices of the Royal Astronomical Society*, 456(3):3032–3052, 2016. ISSN 13652966. doi: 10.1093/mnras/stv2878.

[334] James Pearson, Nan Li, and Simon Dye. The use of convolutional neural networks for modelling large optically-selected strong galaxy-lens samples. *Monthly Notices of the Royal Astronomical Society*, 488(1):991–1004, apr 2019. ISSN 13652966. doi: 10.1093/mnras/stz1750. URL `http://arxiv.org/abs/1904.06199`.

[335] W. J. Pearson, L. Wang, J. W. Trayford, et al. Identifying Galaxy Mergers in Observations and Simulations with Deep Learning. *Astronomy and Astrophysics*, 626(49), feb 2019. ISSN 0004-6361. doi: 10.1051/0004-6361/201935355.

[336] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012. ISSN 15324435. doi: 10.1007/s13398-014-0173-7.2. URL `http://dl.acm.org/citation.cfm?id=2078195`.

[337] P. J. E. Peebles and J. T. Yu. Primeval Adiabatic Perturbation in an Expanding Universe. *The Astrophysical Journal*, 162:815, 1970. ISSN 0004-637X. doi: 10.1086/150713.

[338] Ue Li Pen and Liam Connor. Local circumnuclear magnetar solution to extragalactic fast radio bursts. *Astrophysical Journal*, 807(2), jan 2015. ISSN 15384357. doi: 10.1088/0004-637X/807/2/179. URL `https://arxiv.org/abs/1501.01341`.

[339] M. Pérez-Carrasco, G. Cabrera-Vives, M. Martinez-Marin, et al. Multiband galaxy morphologies for CLASH: A convolutional neural network transferred from CANDELS. *Publications of the Astronomical Society of the Pacific*, 131 (1004), oct 2019. ISSN 00046280. doi: 10.1088/1538-3873/aaeeb4. URL `http://arxiv.org/abs/1810.07857`.

[340] Michael A. Peth, Jennifer M. Lotz, Peter E. Freeman, et al. Beyond spheroids and discs: Classifications of CANDELS galaxy structure at $1.4 < z < 2$ via principal component analysis. *Monthly Notices of the Royal Astronomical Society*, 458(1):963–987, apr 2016. ISSN 13652966. doi: 10.1093/mnras/stw252. URL `http://dx.doi.org/10.1093/mnras/stw252`.

[341] C. E. Petrillo, C. Tortora, S. Chatterjee, et al. Finding strong gravitational lenses in the Kilo Degree Survey with Convolutional Neural Networks. *Monthly Notices of the Royal Astronomical Society*, 472(1):1129–1150, 2017. ISSN 13652966. doi: 10.1093/mnras/stx2052.

[342] E. Petroff, E. D. Barr, A. Jameson, et al. FRBCAT: The fast radio burst catalogue. *Publications of the Astronomical Society of Australia*, 33, jan 2016. ISSN 14486083. doi: 10.1017/pasa.2016.35. URL http://dx.doi.org/10.1017/pasa.2016.35.

[343] E. Petroff, J. W.T. Hessels, and D. R. Lorimer. Fast Radio Bursts. *Astronomy and Astrophysics Review*, 27(1), apr 2019. ISSN 09354956. doi: 10.1007/s00159-019-0116-6. URL http://arxiv.org/abs/1904.07947.

[344] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

[345] Ana-Roxana Pop, Annalisa Pillepich, Nicola C Amorisco, and Lars Hernquist. Formation and Incidence of Shell Galaxies in the Illustris Simulation. *Monthly Notices of the Royal Astronomical Society*, 480(2):1715–1739, oct 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1932. URL https://academic.oup.com/mnras/article/480/2/1715/5059587.

[346] Sergei Popov, Konstantin Postnov, and Maxim Pshirkov. Fast radio bursts: Superpulsars, magnetars, or something else? *International Journal of Modern Physics D*, 27(10), jan 2018. ISSN 02182718. doi: 10.1142/S0218271818440169. URL http://arxiv.org/abs/1801.00640.

[347] Marc Postman, Dan Coe, Narciso Benítez, et al. The cluster lensing and supernova survey with hubble: An overview. *Astrophysical Journal, Supplement Series*, 199(2):25, apr 2012. ISSN 00670049. doi: 10.1088/0067-0049/199/2/25. URL http://stacks.iop.org/0067-0049/199/i=2/a=25?key=crossref.e16849c95860e421e821a52f7f59e960.

[348] William H. Press and Paul Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *The Astrophysical Journal*, 187:425, 1974. ISSN 0004-637X. doi: 10.1086/152650. URL https://ui.adsabs.harvard.edu/abs/1974ApJ...187..425P/abstract.

[349] Yan Qu, John C. Helly, Richard G. Bower, et al. A chronicle of galaxy mass assembly in the EAGLE simulation. *Monthly Notices of the Royal Astronomical Society*, 464(2):1659–1675, jan 2017. ISSN 13652966. doi: 10.1093/mnras/stw2437. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw2437.

[350] P. J. Quinn. On the formation and dynamics of shells around elliptical galaxies. *The Astrophysical Journal*, 279:596–609, apr 1984. ISSN 0004-637X. doi: 10.1086/161924. URL http://adsabs.harvard.edu/doi/10.1086/161924.

[351] Alec Radford, Jong Wook, Kim Chris, et al. Learning Transferable Visual Models From Natural Language Supervision. *OpenAI*, page 47, 2019. URL https://github.com/openai/CLIP.

[352] M.M. Al Rahhal, Yakoub Bazi, Haikel AlHichri, et al. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, jun 2016. ISSN 0020-0255. doi: 10.1016/J.INS.2016.01.082. URL https://www.sciencedirect.com/science/article/pii/S0020025516300184.

[353] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 9413–9424. International Machine Learning Society (IMLS), feb 2019. ISBN 9781510886988. URL http://arxiv.org/abs/1902.10811.

[354] David M. Reiman and Brett E. Göhre. Deblending galaxy superpositions with branched generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 485(2):2617–2627, oct 2019. ISSN 13652966. doi: 10.1093/mnras/stz575. URL http://dx.doi.org/10.1093/mnras/stz575.

[355] Jie Ren, Peter J Liu, Emily Fertig, et al. Likelihood ratios for out-of-distribution detection. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 14707–14718, 2019.

[356] Dezso Ribli, László Dobos, and István Csabai. Galaxy shape measurement with convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 489(4):4847–4859, feb 2019. ISSN 13652966. doi: 10.1093/mnras/stz2374. URL https://academic.oup.com/mnras/article/489/4/4847/5556540?login=true.

[357] Joseph W. Richards, Dan L. Starr, Henrik Brink, et al. Active learning to overcome sample selection bias: Application to photometric variable star classification. *Astrophysical Journal*, 744(2):192, jan 2012. ISSN 15384357. doi: 10.1088/0004-637X/744/2/192. URL http://stacks.iop.org/0004-637X/744/i=2/a=192?key=crossref.c39ba28e373f1cc1d66b153e7beb0596.

[358] Morton. S Roberts and Martha P. Haynes. Physical Parameters along the Hubble Sequence. *Annual Review of Astronomy and Astrophysics*, 32(1):115–152, 1994. ISSN 00664146. doi: 10.1146/annurev.astro.32.1.115.

[359] Brant E. Robertson, Manda Banerji, Michael C. Cooper, et al. Large Synoptic Survey Telescope Galaxies Science Roadmap. *eprint arXiv:1708.01617*, pages 1–3, aug 2017. URL http://arxiv.org/abs/1708.01617.

[360] Vicente Rodriguez-Gomez, Annalisa Pillepich, Laura V. Sales, et al. The stellar mass assembly of galaxies in the Illustris simulation: Growth by mergers and the spatial distribution of accreted stars. *Monthly Notices of the Royal Astronomical Society*, 458(3):2371–2390, may 2016. ISSN 13652966. doi: 10.1093/mnras/stw456. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw456.

[361] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. Springer, Cham, 2015. ISBN 9783319245737. doi: 10.1007/978-3-319-24574-4_28. URL http://link.springer.com/10.1007/978-3-319-24574-4_28.

[362] D. J. Rosario, Rosario, and D. J. FortesFit: Flexible spectral energy distribution modelling with a Bayesian backbone, 2019. URL https://ascl.net/1904.011.

[363] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[364] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0.

[365] Olga Russakovsky, Jia Deng, Hao Su, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y. URL http://dx.doi.org/10.1007/s11263-015-0816-y.

[366] Marc Rußwurm and Marco Körner. Self-Attention for Raw Optical Satellite Time Series Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169(June):421–435, 2020. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2020.06.006. URL https://doi.org/10.1016/j.isprsjprs.2020.06.006.

[367] I. Sadeh, F. B. Abdalla, and O. Lahav. ANNZ2: Photometric redshift and probability distribution function estimation using machine learning. *Publications of the Astronomical Society of the Pacific*, 128(968):104502, oct 2016. ISSN 00046280. doi: 10.1088/1538-3873/128/968/104502.

[368] K. Sakamoto, S. K. Okumura, S. Ishizuki, and N. Z. Scoville. BarâĂŘdriven Transport of Molecular Gas to Galactic Centers and Its Consequences. *The Astrophysical Journal*, 525(2):691–701, nov 1999. ISSN 0004-637X. doi: 10.1086/307910. URL http://iopscience.iop.org/article/10.1086/307910/pdf.

[369] Mara Salvato, Olivier Ilbert, and Ben Hoyle. The many flavours of photometric redshifts. *Nature Astronomy*, 3(3):212–222, mar 2019. ISSN 23973366. doi: 10.1038/s41550-018-0478-0. URL http://www.nature.com/articles/s41550-018-0478-0.

[370] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, et al. Improving galaxy morphologies for SDSS with deep learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, may 2018. ISSN 13652966. doi: 10.1093/MNRAS/STY338. URL https://academic.oup.com/mnras/article/476/3/3661/4848300.

[371] C. Scarlata, C. M. Carollo, S. J. Lilly, et al. COSMOS morphological classification with ZEST (the Zurich Estimator of Structural Types) and the evolution since z=1 of the Luminosity Function of early-, disk-, and irregular galaxies. *The Astrophysical Journal Supplement Series*, 172(1):406–433, sep 2007. ISSN 0067-0049. doi: 10.1086/516582. URL http://stacks.iop.org/0067-0049/172/i=1/a=406.

[372] David Schade, S. J. Lilly, David Crampton, et al. Canada-France Redshift Survey: [ITAL]Hubble Space Telescope[/ITAL] Imaging of High-Redshift Field Galaxies. *The Astrophysical Journal*, 451(1):L1, sep 1995. ISSN 0004637X. doi: 10.1086/309677.

[373] C. Schaefer, M. Geiger, T. Kuntzer, and J. P. Kneib. Deep convolutional neural networks as strong gravitational lens detectors. *Astronomy and Astrophysics*, 611, 2018. ISSN 14320746. doi: 10.1051/0004-6361/201731201.

[374] Kevin Schawinski, M. Dennis Turp, and Ce Zhang. Exploring galaxy evolution with generative models. *Astronomy and Astrophysics*, 616, dec 2018. ISSN 14320746. doi: 10.1051/0004-6361/201833800. URL `http://dx.doi.org/10.1051/0004-6361/201833800`.

[375] S J Schmidt, A I Malz, J. Y.H. Soo, et al. Evaluation of probabilistic photometric redshift estimation approaches for the Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606, sep 2020. ISSN 13652966. doi: 10.1093/mnras/staa2799. URL `https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/staa2799/5905416`.

[376] Claude J Schmit and Jonathan R Pritchard. Emulation of reionization simulations for Bayesian inference of astrophysics parameters using neural networks. *Monthly Notices of the Royal Astronomical Society*, 475(1):1213–1223, jul 2018. ISSN 13652966. doi: 10.1093/mnras/stx3292. URL `http://dx.doi.org/10.1093/mnras/stx3292`.

[377] Michael D. Schneider, Oskar Holm, and Lloyd Knox. Intelligent design: On the emulation of cosmological simulations. *Astrophysical Journal*, 728(2):137, 2011. ISSN 15384357. doi: 10.1088/0004-637X/728/2/137.

[378] A. Schutter and L. Shamir. Galaxy morphology - An unsupervised machine learning approach. *Astronomy and Computing*, 12:60–66, 2015. ISSN 22131337. doi: 10.1016/j.ascom.2015.05.002.

[379] I Sevilla-Noarbe, B Hoyle, M J Marchã, et al. Star-galaxy classification in the Dark Energy Survey Y1 dataset. *Monthly Notices of the Royal Astronomical Society*, 481(4):5451–5469, sep 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2579. URL `https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/sty2579/5104406`.

[380] Lior Shamir. Automatic detection of peculiar galaxies in large datasets of galaxy images. *Journal of Computational Science*, 3(3):181–189, 2012. ISSN 18777503. doi: 10.1016/j.jocs.2012.03.004.

[381] Janelle Shane. AI Weirdness: Do neural nets dream of electric sheep?, 2019. URL https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep.

[382] Zhaohui Shang, Michael S. Brotherton, Beverley J. Wills, et al. The next generation atlas of quasar spectral energy distributions from radio to X-RAYS. *Astrophysical Journal, Supplement Series*, 196(1):2, sep 2011. ISSN 00670049. doi: 10.1088/0067-0049/196/1/2. URL http://stacks.iop.org/0067-0049/196/i=1/a=2?key=crossref.4f79dc498b421f54de2515faa2c95095.

[383] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold, 2020. ISSN 23318422. URL https://arxiv.org/abs/2004.10802.

[384] Yun Kyeong Sheen, Sukyoung K. Yi, Chang H. Ree, and Jaehyun Lee. Post-merger signatures of red-sequence galaxies in rich abell clusters at z âĽš 0.1. *The Astrophysical Journal Supplement Series*, 202(1):8, sep 2012. ISSN 00670049. doi: 10.1088/0067-0049/202/1/8. URL http://stacks.iop.org/0067-0049/202/i=1/a=8?key=crossref.a4bc03fae5edb4d10f9f5484410a018f.

[385] Kartik Sheth, Andrew W. Blain, Jean-Paul Kneib, et al. Detection of CO from SMM J16359+6612, the Multiply Imaged Submillimeter Galaxy behind A2218. *The Astrophysical Journal*, 614(1):L5–L8, oct 2004. ISSN 0004-637X. doi: 10.1086/425308. URL http://stacks.iop.org/1538-4357/614/i=1/a=L5.

[386] David Silver, Aja Huang, Chris J. Maddison, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 0028-0836. doi: 10.1038/nature16961. URL http://dx.doi.org/10.1038/nature16961.

[387] Patrice Y Simard, Dave Steinkraus, and John C Platt. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Washington, DC, 2003. IEEE Computer Society. ISBN 0769519601. URL https://dl.acm.org/doi/10.5555/938980.939477.

[388] Vimal Simha, David H. Weinberg, Charlie Conroy, et al. Parametrising Star Formation Histories. *Arxiv preprint*, apr 2014. URL http://arxiv.org/abs/1404.0402.

[389] B. D. Simmons, Chris Lintott, Kyle W. Willett, et al. Galaxy Zoo: Quantitative visual morphological classifications for 48 000 galaxies from CANDELS. *Monthly Notices of the Royal Astronomical Society*, 464(4):4420–4447, 2017. ISSN 13652966. doi: 10.1093/mnras/stw2587.

[390] Brooke Simmons, Chris Lintott, Kevin Schawinski, et al. Galaxy zoo: Bulgeless galaxies with growing black holes. *Monthly Notices of the Royal Astronomical Society*, 429(3):2199–2211, mar 2013. ISSN 00358711. doi: 10.1093/mnras/sts491. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/sts491.

[391] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. ISBN 0950-5849. doi: 10.1016/j.infsof.2008.09.005. URL http://arxiv.org/abs/1409.1556.

[392] Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. Dynamic bayesian combination of multiple imperfect classifiers. *Studies in Computational Intelligence*, 474:1–35, 2013. ISSN 1860949X. doi: 10.1007/978-3-642-36406-8_1.

[393] Ramin A. Skibba, Karen L. Masters, Robert C. Nichol, et al. Galaxy Zoo: The environmental dependence of bars and bulges in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 423(2):1485–1502, jun 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2012.20972.x.

[394] Lewis Smith and Yarin Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. *Arxiv preprint*, mar 2018. URL http://arxiv.org/abs/1803.08533.

[395] Gregory F. Snyder, Jennifer Lotz, Christopher Moody, et al. Diverse structural evolution at z $>$ 1 in cosmologically simulated galaxies. *Monthly Notices of the Royal Astronomical Society*, 451(4):4290–4310, 2015. ISSN 13652966. doi: 10.1093/mnras/stv1231.

[396] T. Solorio, O. Fuentes, R. Terlevich, and E. Terlevich. An active instance-based machine learning method for stellar population studies. *Monthly Notices of the Royal Astronomical Society*, 363(2):543–554, oct 2005. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2005.09456.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2005.09456.x.

[397] MT Soumagnac. Tipping Scales in Galaxy Surveys: Star/Galaxy Separation and Scale-Dependent Bias. *Doctoral thesis, UCL (University College London).*, jan 2015. URL http://discovery.ucl.ac.uk/1460581/.

[398] J. S. Speagle, C. L. Steinhardt, P. L. Capak, and J. D. Silverman. A highly consistent framework for the evolution of the star-forming "main sequence" from z $\sim$ 0-6. *Astrophysical Journal, Supplement Series*, 214(2), 2014. ISSN 00670049. doi: 10.1088/0067-0049/214/2/15.

[399] D. Spergel, N. Gehrels, J. Breckinridge, et al. WFIRST-2.4: What Every Astronomer Should Know. *arXiv*, may 2013. URL http://arxiv.org/abs/1305.5425.

[400] Ashley Spindler, David Wake, Francesco Belfiore, et al. SDSS-IV MaNGA: The Spatial Distribution of Star Formation and its Dependence on Mass, Structure and Environment. *Monthly Notices of the Royal Astronomical Society*, 23 (October):1–23, 2017. URL http://arxiv.org/abs/1710.05049.

[401] Ashley Spindler, James E Geach, and Michael J Smith. AstroVaDEr: Astronomical variational deep embedder for unsupervised morphological classification of galaxies and synthetic image generation, nov 2020. ISSN 23318422. URL https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/staa3670/6006284.

[402] L. G. Spitler, P. Scholz, J. W.T. Hessels, et al. A repeating fast radio burst. *Nature*, 531(7593):202–205, mar 2016. ISSN 14764687. doi: 10.1038/nature17168. URL http://www.nature.com/articles/nature17168.

[403] Ofer M Springer, Eran O Ofek, Yair Weiss, and Julian Merten. Weak lensing shear estimation beyond the shape-noise limit: A machine learning approach. *Monthly Notices of the Royal Astronomical Society*, 491(4):5301–5316, oct 2020. ISSN 13652966. doi: 10.1093/mnras/stz2991. URL https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/stz2991/5607795.

[404] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. ISSN 15337928. doi: 10.1214/12-AOS1000.

[405] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, 2019. URL `https://mc-stan.org/`.

[406] Kate Storey-Fisher, Marc Huertas-Company, Nesar Ramachandra, et al. Anomaly Detection in Astronomical Images with Generative Adversarial Networks. *arXiv*, dec 2020. URL `http://arxiv.org/abs/2012.08082`.

[407] M. C. Storrie-Lombardi, O. Lahav, L. Sodre, and L. J. Storrie-Lombardi. Morphological Classification of galaxies by Artificial Neural Networks. *Monthly Notices of the Royal Astronomical Society*, 259(1):8P–12P, nov 1992. ISSN 0035-8711. doi: 10.1093/mnras/259.1.8p.

[408] Michael A. Strauss, David H. Weinberg, Robert H. Lupton, et al. Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. *The Astronomical Journal*, 124(3):1810–1824, sep 2002. ISSN 00046256. doi: 10.1086/342343. URL `http://stacks.iop.org/1538-3881/124/i=3/a=1810`.

[409] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2014.

[410] Christian Szegedy, Wei Liu, Yangqing Jia, et al. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, sep 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298594. URL `http://arxiv.org/abs/1409.4842`.

[411] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, et al. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 2818–2826, dec 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.308. URL `http://arxiv.org/abs/1512.00567`.

[412] Tomer Tal, Pieter G. van Dokkum, Jenica Nelan, and Rachel Bezanson. the Frequency of Tidal Features Associated With Nearby Luminous Elliptical Galaxies From a Statistically Complete Sample. *The Astronomical Journal*, 138

(5):1417–1427, 2009. ISSN 0004-6256. doi: 10.1088/0004-6256/138/5/1417. URL http://stacks.iop.org/1538-3881/138/i=5/a=1417?key=crossref.4c7d94a682c78e9735e36c3d79aef91c.

[413] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, 1978. ISSN 21682909. doi: 10.1109/TSMC.1978.4309999.

[414] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 10691–10700, may 2019. ISBN 9781510886988. URL http://arxiv.org/abs/1905.11946.

[415] Mingxing Tan, Bo Chen, Ruoming Pang, et al. Mnasnet: Platform-aware neural architecture search for mobile. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 2815–2823, jul 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00293. URL http://arxiv.org/abs/1807.11626.

[416] Hongming Tang, A. M.M. Scaife, and J. P. Leahy. Transfer learning for radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 488 (3):3358–3375, mar 2019. ISSN 13652966. doi: 10.1093/mnras/stz1883. URL http://dx.doi.org/10.1093/mnras/stz1883.

[417] Dimitrios Tanoglidis, Aleksandra Ćiprijanović, and Alex Drlica-Wagner. Deepshadows: separating low surface brightness galaxies from artifacts using deep learning. *arXiv*, nov 2020. ISSN 23318422. URL http://arxiv.org/abs/2011.12437.

[418] Ambuj Tewari and Peter L. Bartlett. On the Consistency of Multiclass Classification Methods. *Journal of Machine Learning Research*, 8(May):143–157, 2005. ISSN 15324435. doi: 10.1007/11503415_10. URL http://link.springer.com/10.1007/11503415_10.

[419] The Astropy Collaboration, A. M. Price-Whelan, B. M. SipÅŚcz, et al. The Astropy Project: Building an inclusive, open-science project and status of the v2.0 core package. *The Astronomical Journal*, 156(3):123, jan 2018. ISSN 1538-3881. doi: arXiv:1801.02634v2. URL http://arxiv.org/abs/1801.02634.

[420] The Dark Energy Survey Collaboration. The Dark Energy Survey. oct 2005. URL http://arxiv.org/abs/astro-ph/0510346.

[421] The New York Times. "CLEVER HANS" AGAIN; Expert Commission Decides That the Horse Actually Reasons, oct 1904. URL https://timesmachine.nytimes.com/timesmachine/1904/10/02/120289067.html.

[422] The New York Times. A HORSE — AND THE WISE MEN, jul 1911. URL https://timesmachine.nytimes.com/timesmachine/1911/07/23/104872007.html?pageNumber=16.

[423] The New York Times. NEW NAVY DEVICE LEARNS BY DOING; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser, jul 1958. URL https://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html.

[424] Monique Thonnat and Albert Bijaoui. Knowledge based classification of galaxies. In *Knowledge-Based Systems in Astronomy*, pages 121–159. Springer Berlin Heidelberg, apr 1989. doi: 10.1007/3-540-51044-3_21.

[425] D Thornton, B Stappers, M Bailes, et al. A population of fast radio bursts at cosmological distances. *Science*, 341(6141):53–56, jul 2013. ISSN 10959203. doi: 10.1126/science.1236789. URL http://www.ncbi.nlm.nih.gov/pubmed/23828936.

[426] Rita Tojeiro, Karen L. Masters, Joshua Richards, et al. The different star formation histories of blue and red spiral and elliptical galaxies. *Monthly Notices of the Royal Astronomical Society*, 432(1):359–373, jun 2013. ISSN 00358711. doi: 10.1093/mnras/stt484.

[427] Alar Toomre and Juri Toomre. Galactic Bridges and Tails. *The Astrophysical Journal*, 178:623, dec 1972. ISSN 0004-637X. doi: 10.1086/151823. URL http://adsabs.harvard.edu/doi/10.1086/151823.

[428] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, jun 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.214. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909610.

[429] Tomonori Totani. Cosmological Fast Radio Bursts from Binary Neutron Star Mergers. *Publications of the Astronomical Society of Japan*, 65(5):L12, oct 2013. ISSN 0004-6264. doi: 10.1093/pasj/65.5.l12.

[430] K. R.W. Tristram, K. Meisenheimer, W. Jaffe, et al. Resolving the complex structure of the dust torus in the active nucleus of the Circinus galaxy. *Astronomy and Astrophysics*, 474(3):837–850, nov 2007. ISSN 00046361. doi: 10.1051/0004-6361:20078369. URL `http://www.aanda.org/10.1051/0004-6361:20078369`.

[431] Evgenii Tsymbalov, Kirill Fedyanin, and Maxim Panov. Dropout Strikes Back: Improved Uncertainty Estimation via Diversity Sampled Implicit Ensembles. *arXiv*, mar 2020. URL `http://arxiv.org/abs/2003.03274`.

[432] Devis Tuia, Michele Volpi, Loris Copa, et al. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, jun 2011. ISSN 1932-4553. doi: 10.1109/JSTSP.2011.2139193. URL `http://ieeexplore.ieee.org/document/5742970/`.

[433] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network, mar 2020. ISSN 23318422. URL `http://arxiv.org/abs/2003.02037`.

[434] Freeke van de Voort, Joop Schaye, C. M. Booth, et al. The rates and modes of gas accretion on to galaxies and their gaseous haloes. *Monthly Notices of the Royal Astronomical Society*, 414(3):2458–2478, 2011. ISSN 00358711. doi: 10.1111/j.1365-2966.2011.18565.x.

[435] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *33rd International Conference on Machine Learning, ICML 2016*, volume 4, pages 2611–2620. International Machine Learning Society (IMLS), jan 2016. ISBN 9781510829008.

[436] Aäron Van Den Oord, Nal Kalchbrenner, Oriol Vinyals, et al. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems*, pages 4797–4805. Neural information processing systems foundation, jun 2016. URL `http://arxiv.org/abs/1606.05328`.

[437] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, et al. Scikit-image: image processing in Python. *PeerJ*, 2:e453, jun 2014. ISSN 2167-8359. doi: 10.7717/peerj.453. URL https://peerj.com/articles/453.

[438] Pieter G van Dokkum. The Recent and Continuing Assembly of Field Elliptical Galaxies by Red Mergers. *The Astronomical Journal*, 130:2647, 2005. ISSN 0004-6256. doi: 10.1086/497593. URL https://iopscience.iop.org/article/10.1086/497593.

[439] Hugo van Kemenade, Wiredfool, Andrew Murray, et al. python-pillow/Pillow 7.1.2, apr 2020.

[440] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem:5999–6009, 2017. ISSN 10495258. URL http://arxiv.org/abs/1706.03762.

[441] Matteo Venanzi, John Guiver, Gabriella Kazai, et al. Community-based Bayesian aggregation models for crowdsourcing. *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web*, pages 155–164, 2014. doi: 10.1145/2566486.2567989.

[442] Kim Venn, Sébastien Fabbro, Adrian Liu, et al. Machine learning advantages in canadian astrophysics, oct 2019. ISSN 23318422. URL http://arxiv.org/abs/1910.00774.

[443] Mike Walmsley. Galaxy Zoo Bayesian CNN: Initial public release, may 2019. URL https://zenodo.org/record/2677874#.XNaQU6ZCfBI.

[444] Mike Walmsley, Annette M.N. N Ferguson, Robert G. Mann, and Chris J. Lintott. Identification of low surface brightness tidal features in galaxies using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 483(3):2968–2982, mar 2019. ISSN 13652966. doi: 10.1093/mnras/sty3232. URL https://academic.oup.com/mnras/article/483/3/2968/5218505.

[445] Mike Walmsley, Lewis Smith, Chris Lintott, et al. Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*, 491(2):1554–1574, jan 2020. ISSN 0035-8711. doi: 10.1093/mnras/stz2816. URL https://academic.oup.com/mnras/article/491/2/1554/5583078.

[446] Mike Walmsley, Chris Lintott, Geron Tobias, et al. Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies. *Monthly Notices of the Royal Astronomical Society*, dec 2021. URL `https://arxiv.org/abs/2102.08414`.

[447] Hongfeng Wang, Weiwei Zhu, Ping Guo, et al. Pulsar candidate selection using ensemble networks for FAST drift-scan survey. *arXiv*, mar 2019. ISSN 23318422. URL `http://arxiv.org/abs/1903.06383`.

[448] J. Wang, J. F. Navarro, C. S. Frenk, et al. Assembly history and structure of galactic cold dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 413(2):1373–1382, 2011. ISSN 00358711. doi: 10.1111/j.1365-2966.2011.18220.x.

[449] L. Wang, P. Norberg, S. Brough, et al. Galaxy and Mass Assembly (GAMA): The environmental dependence of the galaxy main sequence. *Astronomy and Astrophysics*, 618(2016), 2018. ISSN 14320746. doi: 10.1051/0004-6361/201832697. URL `http://arxiv.org/abs/1802.08456`.

[450] Wenting Wang and Simon D.M. White. Satellite abundances around bright isolated galaxies. *Monthly Notices of the Royal Astronomical Society*, 424(4):2574–2598, 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2012.21256.x.

[451] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[452] Paul Werbos. *Beyond regression: new tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, 1974.

[453] Katherine E. Whitaker, Pieter G. Van Dokkum, Gabriel Brammer, and Marijn Franx. The star formation mass sequence out to z = 2.5. *Astrophysical Journal Letters*, 754(2):L29, 2012. ISSN 20418205. doi: 10.1088/2041-8205/754/2/L29.

[454] S. D. M. White and M. J. Rees. Core condensation in heavy halos - A two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183:341–358, 1978. ISSN 03088146. doi: 10.1016/j.foodchem.2010.01.035. URL `http://adsabs.harvard.edu/abs/1978MNRAS.183..341W`.

[455] Simon D. M. White and Carlos S. Frenk. Galaxy formation through hierarchical clustering. *The Astrophysical Journal*, 379:52, sep 1991. ISSN 0004-637X. doi: 10.1086/170483. URL http://adsabs.harvard.edu/doi/10.1086/170483.

[456] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, et al. Galaxy Zoo 2: Detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013. ISSN 00358711. doi: 10.1093/mnras/stt1458.

[457] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, et al. Galaxy Zoo 2: Images from Original Sample, nov 2013. URL https://zenodo.org/record/3565489#.YC0ZBNYRcay.

[458] Kyle W. Willett, Kevin Schawinski, Brooke D. Simmons, et al. Galaxy Zoo: The dependence of the star formation-stellar mass relation on spiral disc morphology. *Monthly Notices of the Royal Astronomical Society*, 449(1):820–827, 2015. ISSN 13652966. doi: 10.1093/mnras/stv307.

[459] Kyle W. Willett, Melanie A. Galloway, Steven P. Bamford, et al. Galaxy Zoo: Morphological classifications for 120 000 galaxies in HST legacy imaging. *Monthly Notices of the Royal Astronomical Society*, 464(4):4176–4203, feb 2017. ISSN 13652966. doi: 10.1093/mnras/stw2568. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw2568.

[460] Darryl E. Wright, Chris J. Lintott, Stephen J. Smartt, et al. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society*, 472(2):1315–1323, dec 2017. ISSN 13652966. doi: 10.1093/MNRAS/STX1812. URL https://academic.oup.com/mnras/article/472/2/1315/3979473.

[461] Chen Wu, Oiwei Ivy Wong, Lawrence Rudnick, et al. Radio Galaxy Zoo: ClaRAN - A Deep Learning Classifier for Radio Morphologies. *Monthly Notices of the Royal Astronomical Society*, 1230(1):1211–1230, jan 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2646. URL https://academic.oup.com/mnras/article/482/1/1211/5142869.

[462] Xide Xia, Pavlos Protopapas, and Finale Doshi-Velez. Cost-Sensitive Batch Mode Active Learning: Designing Astronomical Observation by Optimizing

Telescope Time and Telescope Choice. *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 477–485, 2016. ISSN 1069-6563. doi: 10.1137/1.9781611974348.54. URL https://epubs.siam.org/doi/10.1137/1.9781611974348.54.

[463] Qizhe Xie, Minh Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10684–10695. IEEE Computer Society, nov 2020. doi: 10.1109/CVPR42600.2020.01070. URL http://arxiv.org/abs/1911.04252.

[464] Prateek Yadav. *Applying convolutional neural networks to classify fast radio bursts detected by the CHIME telescope*. PhD thesis, The University of British Columbia, 2020. URL https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0389889.

[465] Kaiyu Yang, Klint Qinami, Li Fei-Fei, et al. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558. Association for Computing Machinery, Inc, dec 2020. ISBN 9781450369367. doi: 10.1145/3351095.3375709. URL http://dx.doi.org/10.1145/3351095.3375709.

[466] Jacky H.T. Yip, Xinyue Zhang, Yanfang Wang, et al. From dark matter to galaxies with convolutional neural networks. In *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019)*. arXiv, oct 2019. URL http://arxiv.org/abs/1910.07813.

[467] Donald G. York, J. Adelman, John E. Anderson, Jr., et al. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1587, sep 2000. ISSN 00046256. doi: 10.1086/301513.

[468] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z Ghahramani, M Welling, C Cortes, et al., editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014. URL http://arxiv.org/abs/1411.1792.

[469] Lorenzo Zanisi, Marc Huertas-Company, François Lanusse, et al. The relationship between fine galaxy stellar morphology and star formation activity in cosmological simulations: a deep learning view, jun 2020. ISSN 23318422. URL http://arxiv.org/abs/2007.00039.

[470] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1):818–833, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10590-1_53. URL https://link.springer.com/chapter/10.1007/978-3-319-10590-1_53.

[471] Michael Zevin, Scott Coughlin, Sara Bahaadini, et al. Gravity Spy: Integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, 34(6), nov 2017. ISSN 13616382. doi: 10.1088/1361-6382/aa5cea. URL http://dx.doi.org/10.1088/1361-6382/aa5cea.

[472] Yunfan Gerry Zhang, Vishal Gajjar, Griffin Foster, et al. Fast Radio Burst 121102 Pulse Detection and Periodicity: A Machine Learning Approach. *The Astrophysical Journal*, 866(2):149, sep 2018. ISSN 0004-637X. doi: 10.3847/1538-4357/aadf31. URL http://arxiv.org/abs/1809.03043.

[473] Shusen Zhou, Qingcai Chen, and Xiaolong Wang. Active deep learning method for semi-supervised sentiment classification. *Neurocomputing*, 120:536–546, nov 2013. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2013.04.017. URL https://www.sciencedirect.com/science/article/pii/S0925231213004888.

[474] F. Zwicky. Luminous Intergalactic Matter. *Publications of the Astronomical Society of the Pacific*, 64:242, 1952. ISSN 0004-6280. doi: 10.1086/126484.