

Identifying Galaxy Mergers in Observations and Simulations with Deep Learning

W. J. Pearson^{1,2}, L. Wang^{1,2}, J. W. Trayford³, C. E. Petrillo², F. F. S. van der Tak^{1,2}

¹ SRON Netherlands Institute for Space Research, Landleven 12, 9747 AD, Groningen, The Netherlands
e-mail: w.j.pearson@srn.nl

² Kapteyn Astronomical Institute, University of Groningen, Postbus 800, 9700 AV Groningen, The Netherlands

³ Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA Leiden, The Netherlands

Received DD Month YYYY / Accepted DD Month YYYY

ABSTRACT

Context. Mergers are an important aspect of galaxy formation and evolution. With large upcoming surveys, such as *Euclid* and LSST, accurate techniques that are fast and efficient are needed to identify galaxy mergers for further study.

Aims. We aim to test whether deep learning techniques can be used to reproduce visual classification of observations, physical classification of simulations and highlight any differences between these two classifications. With one of the main difficulties of merger studies being the lack of a truth sample, we can use our method to test biases in visually identified merger catalogues.

Methods. A convolutional neural network architecture was developed and trained in two ways: one with observations from SDSS and one with simulated galaxies from EAGLE, processed to mimic the SDSS observations. The SDSS images were also classified by the simulation trained network and the EAGLE images classified by the observation trained network.

Results. The observationally trained network achieves an accuracy of 91.5% while the simulation trained network achieves 74.4% on the visually classified SDSS and physically classified EAGLE images respectively. Classifying the SDSS images with the simulation trained network was less successful, only achieving an accuracy of 64.3%, while classifying the EAGLE images with the observation network was very poor, achieving an accuracy of only 49.7% with preferential assignment to the non-merger classification. This suggests that most of the simulated mergers do not have conspicuous merger features and visually identified merger catalogues from observations are incomplete and biased towards certain merger types.

Conclusions. The networks trained and tested with the same data perform the best, with observations performing better than simulations, a result of the observational sample being biased towards conspicuous mergers. Classifying SDSS observations with the simulation trained network has proven to work, providing tantalizing prospects for using simulation trained networks for galaxy identification in large surveys.

Key words. Galaxies: interactions – Techniques: image processing – Methods: data analysis – Methods: numerical

1. Introduction

Galaxy-galaxy mergers are of fundamental importance to our current understanding of how galaxies form and evolve in cold dark matter cosmology (e.g. Conselice 2014). Dark matter halos and their baryonic counterparts merge under hierarchical growth to form the universe that we see today (Somerville & Davé 2015). Mergers play an important role in many aspects of galaxy evolution such as galaxy mass assembly, morphological transformation and growth of the central black hole (e.g. Johnston et al. 1996; Naab & Burkert 2003; Hopkins et al. 2006; Bell et al. 2008; Guo & White 2008; Genel et al. 2009). In addition, galaxy mergers are believed to be the triggering mechanism of some of the brightest infrared objects known: (ultra) luminous infrared galaxies (Sanders & Mirabel 1996). With bright infrared emission often comes high star formation rates (SFRs), hence a prevailing interpretation from early merger works is that most mergers go through a starburst phase (e.g. Joseph & Wright 1985; Schweizer 2005).

Recent studies have begun to dismantle the claim that all galaxy mergers are starbursts. In a study of 1500 galaxies, within 45 Mpc of our own, Knapen et al. (2015) have found that the increase in SFR in merging galaxies is at most a factor of two, with the majority of galaxies showing no evidence of an increase in

SFR, or even showing evidence of mergers quenching the star formation. Galaxy mergers do still cause starbursts and a higher fraction of starbursts are mergers than starbursting non-mergers (Luo et al. 2014; Knapen & Cisternas 2015; Cortijo-Ferrero et al. 2017). Claims about the importance of mergers depend critically on our ability to recognise galaxy interactions. A method to reliably identify complete merger samples among a large number of galaxies is clearly needed.

Existing automated techniques for detecting mergers include selecting close galaxy pairs or selecting morphologically disturbed galaxies. The close pair method finds pairs of galaxies that are close, both on the sky and in redshift (e.g. Barton et al. 2000; Patton et al. 2002; Lambas et al. 2003; Lin et al. 2004; De Propriis et al. 2005). This method requires highly complete, spectroscopic observations and, as a result, is observationally expensive. It can also be contaminated by flybys (Sinha & Holley-Bockelmann 2012; Lang et al. 2014). Selecting the morphologically disturbed galaxies using quantitative measurements of non-parametric morphological statistics, such as the Gini coefficient, the second-order moment of the brightest 20 percent of the light (Lotz et al. 2004) and the CAS system (e.g. Bershadsky et al. 2000; Conselice et al. 2000; Wu et al. 2001; Conselice et al. 2003), aims to detect disturbances such as strong asymmetries,

double nuclei or tidal tails. This method relies on high-quality, high-resolution imaging to detect these features beyond the local universe and has a high percentage of misclassifications ($> 20\%$), especially at high redshift (Huertas-Company et al. 2015). There is also the option to classify galaxies through visual inspection. However, visual classifications are hard to reproduce and are time consuming. Large crowd sourced methods, such as Galaxy Zoo¹ (Lintott et al. 2008), are not scalable to the sizes of the data sets expected from upcoming surveys. Visual identification can also suffer from low accuracy and incompleteness (Huertas-Company et al. 2015).

Deep learning techniques have the potential to revolutionise galaxy classification. Once properly trained, the neural networks used in deep learning can classify thousands of galaxies in a fraction of the time it would take a human, or team of humans, to classify the same objects. The use of deep learning for galaxy classification was brought to wider attention after Galaxy Zoo lead a competition on the Kaggle platform², known as The Galaxy Challenge, to develop a machine learning algorithm to replicate the human classification of the Sloan Digital Sky Survey (SDSS; York et al. 2000) images. This competition was won by Dieleman et al. (2015) using a deep neural network, the architecture of which has formed the base for subsequent deep learning algorithms for galaxy classification (e.g. Huertas-Company et al. 2015; Petrillo et al. 2017). More recently, deep learning has been applied to SDSS images from Galaxy Zoo to classify objects as merging or non-merging systems using transfer learning, that is taking a pre-trained network and retraining the output layer to classify images into a different set of classifications (Ackermann et al. 2018). There has also been work using deep learning techniques to identifying mergers and tidal features in optical data from the Canada-France-Hawaii Telescope Legacy Survey (Gwyn 2012; Walmsley et al. 2019). These techniques will have an important use in classifying galaxies in large, upcoming surveys, such as the Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al. 2009) or *Euclid* (Lau-reijs et al. 2011).

In this work, we aim to develop a neural network architecture and independently train it with two different training sets. This will result in a trained neural network that can identify visually classified mergers from the SDSS data as well as one that can identify physically classified mergers from the Evolution and Assembly of GaLaxies and their Environments (EAGLE) hydrodynamical cosmological simulation (Schaye et al. 2015). Once trained, the networks will be cross applied: SDSS images through the EAGLE trained network and images of simulated galaxies from EAGLE through the SDSS trained network. Visually identified merger catalogues constructed from surveys, such as the SDSS, are biased towards mergers that produce conspicuous features but cosmological simulations include a wide variety of merging galaxies with different mass ratios, gas fractions, environments, orbital parameters etc. Therefore, through training our neural network separately with visual classifications of real observations, physical classifications in simulations and the cross-applications of the two, we can better understand any potential biases in observations and identify problems in simulations in terms of reproducing realistic merger properties.

The paper is structured as follows: Sect. 2 describes the data sets used, Sect. 3 covers the neural networks, Sect. 4 provides the results and discussion and Sect. 5 the concluding remarks. Where necessary, Wilkinson Microwave Anisotropy Probe year

7 (WMAP7) cosmology (Komatsu et al. 2011; Larson et al. 2011) is followed, with $\Omega_M = 0.272$, $\Omega_\Lambda = 0.728$ and $H_0 = 70.4 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

2. Image Data

2.1. SDSS Images

To train the neural network, a large number of images of merging and non-merging systems are required. For the training the observational network, we create our merger and non-merger samples by following Ackermann et al. (2018) and combining the Darg et al. (2010a,b) merger catalogue with non-merging systems. The Darg et al. (2010a,b) catalogue contains 3003 merging systems selected by visually rechecking the visual classifications of all objects from Galaxy Zoo with the fraction of people who classified the object as merging greater than 0.4 and spectroscopic redshifts between 0.005 and 0.1. As a result of this thorough visual classification, the Darg et al. (2010a,b) catalogue is likely to be conservative and mainly contain galaxies with obvious signs of merger, i.e. two (or more) clearly interacting galaxies or obviously morphologically disturbed systems, and may miss more subtle mergers. The SDSS spectra were only taken for objects with apparent magnitude $r < 17.77$, or absolute magnitude $r < -20.55$ at $z = 0.1$ hence resulting in an effective mass limit of $\approx 10^{10} M_\odot$ at $z = 0.1$ (Darg et al. 2010b). For the non-merging systems, we generated a catalogue of all SDSS objects with spectroscopic redshifts in the same range as the Darg et al. (2010a,b) catalogue and the fraction of people who classified the object as merging in Galaxy Zoo less than 0.2 and then randomly selected 3003 of these to form the sample. As we also require spectroscopic redshifts, the non-merger sample will have the same effective mass limit of $\approx 10^{10} M_\odot$. Cut-outs of the merging and non-merging objects were then requested from the SDSS cut-out server for data release 7³ (DR7) to create 6006 images in the gri bands, each of 256×256 pixels. These images were then cropped to the centre 64×64 pixels for use to reduce memory requirements while training. Larger image sizes were tested but showed no clear improvements over 64×64 pixel images. Examples of the central 64×64 pixels of merging and non-merging SDSS galaxies are given in Fig. 1.

The SFR and $M_\star$ for the SDSS objects were gathered from the MPA-JHU catalogue⁴; the $M_\star$ were created following the techniques of Kauffmann et al. (2003) and Salim et al. (2007), while the SFR were based on the Brinchmann et al. (2004) catalogue. The redshifts and ugriz magnitudes come from the SDSS DR7.

2.2. EAGLE Images

For the simulation network, simulated gri images from EAGLE were used. These images were generated using the py-SPHviewer code (Benítez-Llambay 2017) and do not account for dust attenuation. EAGLE galaxies from the simulation snapshots with a redshift of less than 1.0 were used. Objects with stellar mass ($M_\star$) greater than $10^{10} M_\odot$ were selected while the merging partner of the merging systems must be larger than $10^9 M_\odot$. The merging partner must also be more than 10% of the $M_\star$ of the primary galaxy. Galaxies were deemed to have merged when they are tracked as two galaxies in one simulation snapshot and then tracked as one galaxy in the following snapshot in the

¹ <http://www.galaxyzoo.org/>

² <https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge>

³ <http://cas.sdss.org/dr7/en/tools/chart/default.asp>

⁴ <https://www.mpa.mpg.de/SDSS/DR7/>

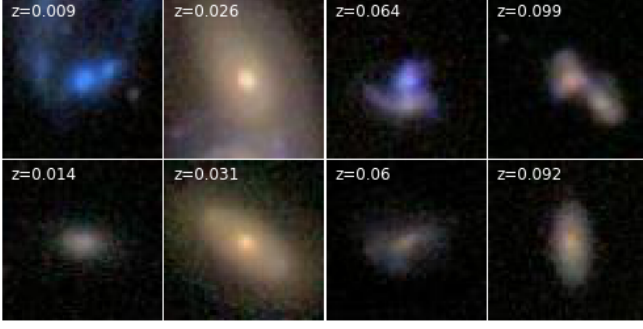


Fig. 1. Examples of the central 64×64 pixels of SDSS gri, as blue, green and red respectively, galaxy images, corresponding to an angular size of 25.3×25.3 arcsec. The top row shows merging galaxies from the Darg et al. (2010a,b) catalogue while the bottom row shows non-merging galaxies.

EAGLE merger trees catalogue (Qu et al. 2017). Systems that are projected to merge, using a closing velocity extrapolation, within the next 0.3 Gyr (pre-merger) or are projected to have merged, again using a closing velocity extrapolation of the progenitors, within the last 0.25 Gyr (post-merger) were selected, along with a number of non-merging systems, and gri band images were created of these systems. Springel et al. (2005) have shown that the effects of a merger are visible for approximately 0.25 Gyr after the merger event while the pre-merger stage is much longer. However, we chose to have the pre and post merger period approximately equal as tests conducted with longer pre-merger times showed no improvement, see discussion in Sect. 4.2. We note, however, that the merger timing may suffer from imprecision as a result of the coarse time resolution of the EAGLE simulation, i.e. the time between snapshots, which becomes coarser at lower simulation redshift. Each galaxy is imaged at an assumed distance of 10 Mpc and each image contains all material within 100 kpc of the centre of the target galaxy and is 256×256 pixels, where 256 pixels corresponds to a physical size of 60 kpc. There are 537 pre-merger, 339 post-merger and 335 non-merging systems, each with six random projections to increase the size of the training set. Each of the six projections are treated as individual galaxies resulting in 3222 pre-merger, 2034 post-merger and 2010 non-merging galaxy images for training. The pre-mergers and post-mergers were combined to form the merger class, keeping the pre-merger image if the same galaxy appears in both sets.

To make the raw EAGLE images look like SDSS images (processed EAGLE images), a number of operations were performed. For each projection of each system, a redshift was randomly chosen from the redshifts of the objects in the Darg et al. (2010a,b) catalogue and the surface brightness of the galaxy was corrected to match this redshift. The image was also re-binned using interpolation with the python `reproject` package (Robitaille 2018) so that the physical resolution of the EAGLE image matches that physical resolution of an SDSS galaxy at the selected redshift. The resulting apparent r-band magnitudes are less than 17.77 for all but 58 of the 10 134 galaxy projections, meaning that the brightness of the simulated galaxies is consistent with the observed SDSS galaxies. Once the surface brightness and physical resolution correction was completed, the observed SDSS point spread functions (PSFs) for the gri bands were created using the stand alone PSF tool⁵ and the simulated

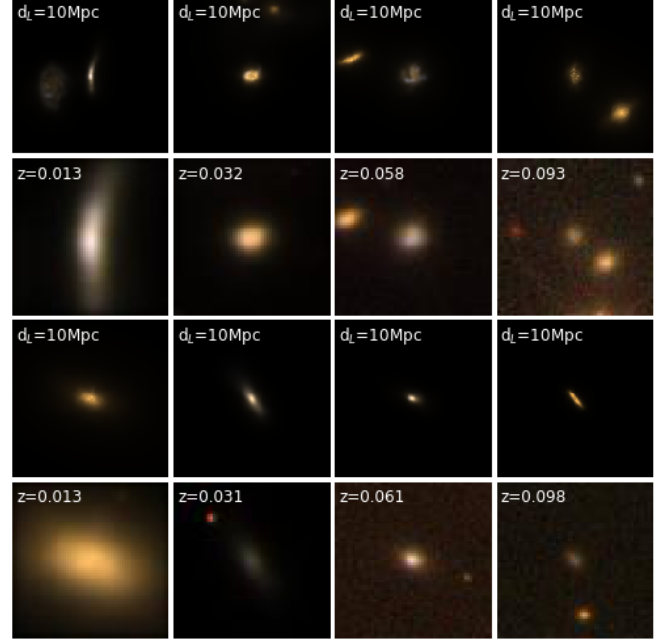


Fig. 2. Examples of the raw (first and third rows) and processed (second and fourth rows) EAGLE images for merging (first and second rows) and non-merging (third and fourth rows) systems. The raw images shown are 128×128 pixels and imaged at 10 Mpc, corresponding to a physical size of 30×30 kpc or an angular size of 621×621 arcsec, while the processed images are 64×64 pixel images corresponding to an angular size of 25.3×25.3 arcsec. The redshifts are those that the EAGLE images have been projected to.

images were convolved with these PSFs. Finally, the EAGLE galaxies were injected into real SDSS images to add realistic noise.

To get real SDSS noise, the position of all known SDSS objects from DR7 in three SDSS images were collected. The noise images were generated by offsetting from the position of the objects in these images by a random distance between 6.329 and 18.986 arcsec (i.e. between 0.25 and 0.75 times the average separation of SDSS objects) and with a random angle in the RA-dec plane. Then 256×256 pixel cut-outs were made, centred on these offset positions, and were used as noise in the EAGLE images. The code used to make the EAGLE images SDSS like and get the noise cut-outs can be downloaded from GitHub⁶ while examples of the raw and processed EAGLE images can be found in Fig. 2.

The M_* , star formation rate (SFR), ugriz absolute magnitudes, galaxy asymmetry, merger mass ratio and time to or since the merger event of the EAGLE galaxies are from the simulation. For the merging systems, M_* and SFR are calculated for the merger remnant. The galaxy asymmetry is the 3D asymmetry and is calculated as described in Trayford et al. (2019). Uniform bins of solid angle were created about the galaxy centre and the M_* within each bin is summed. The asymmetry is then the sum of the absolute mass difference between diametrically opposed bins divided by the total M_* . Thus, the higher the asymmetry value, the more asymmetric the galaxy is.

⁵ https://www.sdss.org/dr12/algorithms/read_psf/

⁶ <https://github.com/wjpearson/SDSS-EAGLE-mergers>

3. Deep Learning

3.1. Convolutional Neural Networks

Deep learning neural networks are a type of machine learning that aim to loosely mimic how a biological system processes information by using a series of layers of non-linear mathematical operations, known as neurons, each with its own weight and bias value. Here we use a type of deep learning known as convolutional neural networks (CNN). The majority of layers in a CNN are made up from kernels that are convolved with the output from the layer below with the top layer being one dimensional and fully connected, that is all the mathematical functions (neurons) that make up a layer are connected to all the neurons in the layer below. Each neuron in a network has an activation function that determines if the result should be passed to higher layers or not. The dimensionality of the network can be reduced by applying a pooling layer after an output to group the outputs from a number neurons into a single output. The kernels and weights in the layers of a CNN are trained by passing a large number of classified images through the network such that the output classifications match, or closely match, the known input classifications. A thorough description of how CNNs work is beyond the scope of this paper; further information on CNNs can be found in Lecun et al. (1998).

When discussing neural networks, some terms are used whose definitions may differ from what is expected or be unfamiliar. Also, concepts have a number of different names. To prevent confusion, terms used in this paper are defined in Table 1, taking a positive result to mean a merger and a negative result to mean a non-merger.

3.2. Architecture

The CNN used in this work was built using the Tensorflow framework (Abadi et al. 2015). As the task we are attempting to complete is similar to that of The Galaxy Challenge, we base our network on the winning Dieleman et al. (2015) architecture but apply some tweaks. The input image is 64 by 64 pixels with three colour channels. We then apply a series of four, two dimensional convolutional layers with 32, 64, 128 and 128 kernels of 6×6, 5×5, 3×3 and 3×3 pixels for the first, second, third and fourth layers respectively. The strides of the kernels, how far the kernel is moved as it scans the input, is set at 1 pixel for all layers and the zero padding is set to “same” to pad each edge of the image with zeros evenly (if required). Batch normalisation (Ioffe & Szegedy 2015) is applied after each layer, scaling the output between zero and one, and we use Rectified Linear Units (ReLU; Nair & Hinton 2010) for activation. ReLU returns $\max(x, 0)$ when passed x . Dropout (Srivastava et al. 2014) is also applied after each activation, to help reduce overfitting, with a dropout rate of 0.2, randomly setting the output of neurons to zero 20% of the time during training. The output from the first, second and fourth convolutional layer has a 2×2 pixel max-pooling applied to reduce dimensionality. After the fourth convolutional layer, we use two one-dimensional, fully connected layers of 2048 neurons, again applying ReLU activation, batch normalisation and dropout. The output layer has two neurons⁷, one for each class, and uses a softmax output which provides probabilities for each class, in the interval [0, 1], that sum to one, i.e. softmax maps the un-normalised input into it into a probability distribution over the output classes. Thus there is one output that can be consid-

⁷ It is possible to do this with a single output but this setup makes it easier to add more classes in the future.

ered the probability the input image is of a merging system and one output that can be considered to be the probability the input image is of a non-merging galaxy. In this paper, we will use the output for the merger class, although with our binary classification the non-merger class can be considered equivalent as it is 1-(merger class output). The full network can be seen in Table 2. Loss of the network is determined using softmax cross entropy and is optimised using the Adam algorithm (Kingma & Ba 2015). A learning rate, i.e. how fast the weights and biases in the network can change, of 5×10^{-5} is used as it resulted in a more accurate network.

3.3. Training, Validation and Testing

If there are an unequal number of images in the two classes, the larger class size is reduced by randomly removing images until the classes are the same size. The images were then subdivided into three groups: 80% were used for training, 10% for validation and 10% for testing. The training set was the set used to train the network while the validation was used to see how well the network was performing as training progressed. Each network was trained for 200 epochs, an epoch is showing each image to the network once, and the epoch with simultaneously the highest accuracy and lowest loss with the validation set was selected for use. Using 200 epochs is long enough as by this point the loss for the validation set has begun to increase as the network starts to over-train and learn the training set, not the features in the training set. The testing set was used once, and once only, to test the performance of the network deemed to be the best from the validation. Testing images are not used for validation to prevent accidental training on the test data set. To reduce sensitivity to galaxy orientation, the images were also augmented as they were loaded for training (and only training): the images were randomly rotated by 0°, 90°, 180° or 270°. We also crop the images to the centre 64×64 pixels and scale the images between zero and one. The code used to create, train, validate and test the networks can be downloaded from GitHub⁸.

4. Results and Discussion

We use the receiver operating characteristic (ROC) curve to determine how well the network has performed for a binary classification. The ROC curve is a plot of the recall against fall-out (see Table 1 for definitions) with each point along the curve corresponding to a different value for the output (threshold) above which an input image is considered to be of a merging system. Higher recall and lower fall-out means a better threshold while the (0,0) and (1,1) positions correspond to assigning all objects to the non-merger and merger classes respectively. The threshold with recall and fall-out closest to the (1,0) position, calculated as least squared difference, is the preferred threshold for splitting mergers from non-mergers. Also, the area under the ROC curve is unity for an infallible network, and close to unity for good networks, while a truly random network will have an area of 0.5.

The two-sample Kolmogorov-Smirnov test (KS-test; Smirnov 1939) is also used to compare the distributions of correctly and incorrectly identified objects to see if they are likely sampled from the same distribution. The null hypothesis that the two distributions are the same is rejected at level $\alpha = 0.05$ if the KS-test statistic, $D_{N,M}$, is greater than $\text{Crit}_{N,M} = c(\alpha) \sqrt{\frac{n+m}{nm}}$,

⁸ <https://github.com/wjpearson/SDSS-EAGLE-mergers>

Table 1. Terms used when describing the performance of neural networks

Term	Definition	
Positive (P)	An object classified in the catalogues or identified by a network as a merger.	
Negative (N)	An object classified in the catalogues or identified by a network as a non-merger.	
True Positive (TP)	An object classified in the catalogues as a merger that is identified by a network as a merger.	
False Positive (FP)	An object classified in the catalogues as a non-merger that is identified by a network as a merger.	
True Negative (TN)	An object classified in the catalogues as a non-merger that is identified by a network as a non-merger.	
False Negative (FN)	An object classified in the catalogues as a merger that is identified by a network as a non-merger.	
Recall	Fraction of objects correctly identified by a network as a merger with respect to the total number of objects classified in the catalogues as mergers.	$TP / (TP+FN)$
Fall-out	Fraction of objects incorrectly identified by a network as a merger with respect to the total number of objects classified in the catalogues as mergers.	$FP / (TP+FN)$
Specificity	Fraction of objects correctly identified by a network as a non-merger with respect to the total number of objects classified in the catalogues as non-mergers.	$TN / (TN+FP)$
Precision	Fraction of objects correctly identified by a network as a merger with respect to the total number of objects identified by a network as a merger.	$TP / (TP+FP)$
Negative Predictive Value (NPV)	Fraction of objects correctly identified by a network as a non-merger with respect to the total number of objects identified by a network as a non-merger.	$TN / (TN+FN)$
Accuracy	Fraction of objects, both merger and non-merger, correctly identified by a network.	$(TP+TN) / (TP+FP+TN+FN)$

where $c(\alpha) = 1.224$ for $\alpha = 0.05$ and n and m are the sizes of samples N and M .

4.1. Observation Trained Network

The 97th epoch of the network trained with SDSS images (observation network) is used. This epoch has an accuracy (see Table 1 for definition) at validation of 0.932 cutting at a threshold of 0.5 to separate mergers from non-merger classification. Using the validation set, we plot the ROC curve for this network in blue in Fig. 3. This has an area of 0.966 and provides an ideal cut threshold of 0.57. At this threshold, the accuracy of the validation set increases to 0.935. To determine the true accuracy of the network, we perform the same analysis for the test data set. The area under the ROC curve, shown in Fig. 3 in yellow, remains constant at 0.966. With the threshold set at 0.57, the final accuracy of the network is 0.915, with recall, precision, specificity and NPV of 0.920, 0.911, 0.910 and 0.919 respectively (see Table 1 for definitions). It is possible to increase the accuracy, and other cut dependent statistics, by changing the cut threshold for the training set. However, this risks accidentally using the test set for training and thus not giving a true representation of the network.

Our results can be compared to those of Ackermann et al. (2018), who performed a similar study using the same Darg et al. (2010a,b) merger catalogue. Ackermann et al. (2018) have a recall of 0.96, a precision of 0.97 and the area under the ROC curve is 0.9922. All these values are slightly larger than those we find, demonstrating that their network performs somewhat bet-

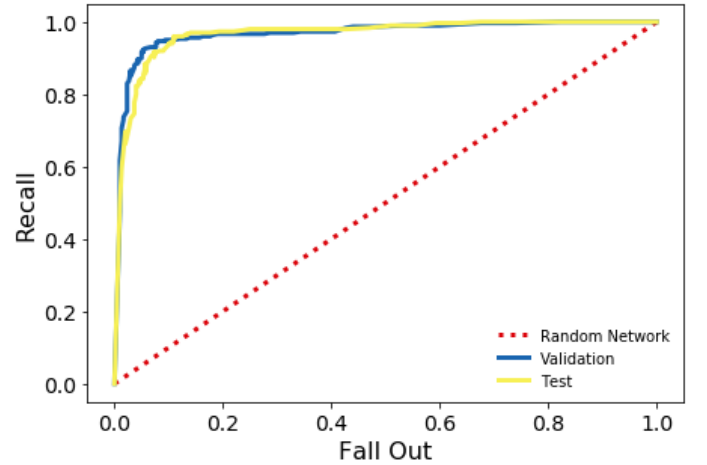


Fig. 3. ROC curve for the observation network used on visually classified SDSS images at validation (blue) and testing (yellow). The area under each curve is 0.966. The dashed red line shows the position of a truly random network.

ter. However, there are some differences between the two studies. The architecture of the CNN used here is different from that used by Ackermann et al. (2018), who use the Xception architecture (Chollet 2017), and they perform transfer learning: using a network pre-trained on the non-astronomical ImageNet images (Deng et al. 2009) and then continuing to train using the merger and non-merger images. The non-merger set is also dif-

Table 2. Architecture of the CNN. The first column in the type of layer while the second column contains the associated properties. The input is a 64×64 pixel, 3 channel image and the output is two probabilities, one for the probability the input is a merger and one for the probability the input is a non-merger. Further details on what the properties of the layers mean can be found in Sect. 3.2.

Layer	Properties				
Input	64×64 pixels	3 channels			
Convolutional	32, 6×6 pixel kernels	1 pixel stride	“same” padding	Batch normalisation	ReLU activation
Dropout	Dropout rate of 0.2				
MaxPooling	2×2 pixel	2 pixel stride			
Convolutional	64, 5×5 pixel kernels	1 pixel stride	“same” padding	Batch normalisation	ReLU activation
Dropout	Dropout rate of 0.2				
MaxPooling	2×2 pixel	2 pixel stride			
Convolutional	128, 3×3 pixel kernels	1 pixel stride	“same” padding	Batch normalisation	ReLU activation
Dropout	Dropout rate of 0.2				
Convolutional	128, 3×3 pixel kernels	1 pixel stride	“same” padding	Batch normalisation	ReLU activation
Dropout	Dropout rate of 0.2				
MaxPooling	2×2 pixel	2 pixel stride			
Flatten					
Fully Connected	2048 neurons	Batch normalisation	ReLU activation		
Dropout	Dropout rate of 0.2				
Fully Connected	2048 neurons	Batch normalisation	ReLU activation		
Dropout	Dropout rate of 0.2				
Output	2 neurons	Softmax activation			

ferent, with Ackermann et al. (2018) using 10 000 non-merging galaxies as opposed to our 3003. We use an equal number of mergers and non-mergers to prevent accidental bias against the class with fewer images. Finally, the Ackermann et al. (2018) study does not change the cut threshold from 0.5 to improve the recall or precision, suggesting that these values may be able to be improved.

Another study, by Walmsley et al. (2019), trains a CNN on data from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS; Gwyn 2012). Here, they aim to identify galaxies with tidal features, which are likely due to galaxy interactions. The performance of our SDSS trained network is much better than that of the CFHTLS network: we achieve recall of 0.920 while Walmsley et al. (2019) achieve 0.760. However, the differences in data set and network architecture will have an effect on the results.

To determine if certain physical properties of the galaxies are the cause of the misclassification, the specific SFR (SFR/M_* , $s\text{SFR}$), M_* , redshift and u griz band magnitudes of the misclassified objects have been compared to their correctly classified counterparts. This will allow us to determine if, for example, all of the high mass, non-mergers have been classified as mergers. There are no trends in any of these properties: the distribution of the misclassified objects is the same as the distribution of the correctly classified objects. The confusion matrix, showing the number of TP, FP, TN and FN, for the SDSS images classified by the observation network can be found in Table 3 while the KS-test statistics comparing the distributions of correctly and incorrectly identified galaxies with the physical properties can be found in Table 4. See Table 1 for definitions of TP, FP, TN and FN.

The images of the misclassified objects have also been visually inspected. Over half of the FP objects (16 of 27) have a close chance projection or a second galaxy projected into the disk of the primary galaxy, possibly fooling the network into believing that the two galaxies are merging. Four further galaxies fill the entire 64×64 pixel image, two of which also have a chance projection of a second galaxy into the disk of the primary galaxy.

Table 3. Confusion matrix for SDSS images classified by the observation network.

		Network Classification		Total
		Merger	Non-merger	
Catalogue Classification	Merger	276 TP	24 FN	300
	Non-merger	27 FP	273 TN	300
Total		303	297	

Table 4. KS-test statistic, $D_{N,M}$, and the critical value, $\text{Crit}_{N,M} = c(\alpha) \sqrt{\frac{n+m}{nm}}$, for the SDSS images classified by the observation network. If $D_{N,M} > \text{Crit}_{N,M}$, the null hypothesis that the two distributions are the same is rejected at level $\alpha = 0.05$. Here, $c(\alpha) = 1.224$ for $\alpha = 0.05$ and n and m are the sizes of samples N and M .

Physical Parameter	$D_{TP,FN}$	$\text{Crit}_{TP,FN}$	$D_{TN,FP}$	$\text{Crit}_{TN,FP}$
M_*	0.144	0.261	0.091	0.247
$s\text{SFR}$	0.203	0.261	0.141	0.247
u -magnitude	0.190	0.260	0.195	0.247
g -magnitude	0.324	0.260	0.168	0.247
r -magnitude	0.236	0.260	0.178	0.247
i -magnitude	0.196	0.260	0.193	0.247
z -magnitude	0.199	0.260	0.197	0.247

For six of the FP, there is no clear reason why they are misclassified: they appear to be isolated galaxies without signs of morphological disturbance. The final FP is a large grand design spiral that has been identified off centre in the original 256×256

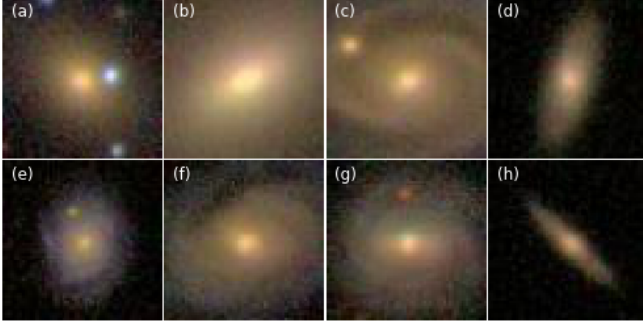


Fig. 4. Examples of FP galaxies from the observation network for (a) a chance projection, (b) a galaxy filling the image, (c) a galaxy filling the image with a chance projection and (d) an isolated, non-interacting galaxy. Panels (e) to (h) show TN galaxies that are visually similar to those shown in (a) to (d).

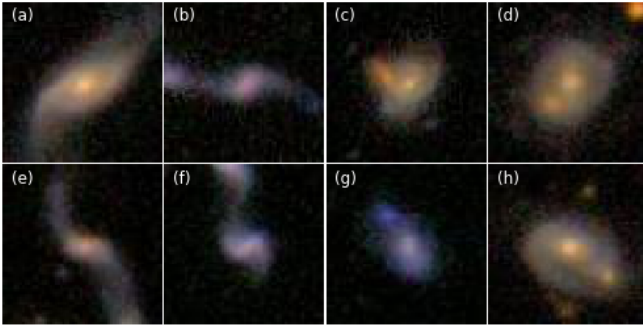


Fig. 5. Examples of FN galaxies from the observation network for (a) a galaxy with its merging companion outside the image, (b) a galaxy with its merging companion on the edge of the image, (c) a merging system and (d) a merging system where the minor galaxy is almost point-like. Panels (e) to (h) show TP galaxies that are visually similar to those shown in (a) to (d).

pixel image. When cropped, the image contains only the arms of the spiral that appear like a disturbed system. Examples of the FP are shown in Fig. 4a-d. However, and unsurprisingly with so few misclassified objects, galaxies that are visually similar to the FP have also been correctly identified, as seen in Fig. 4e-h.

For the FN objects, six of the 24 have a merging companion that is either outside the 64×64 pixel image or on the very edge, indicating that a larger image may reduce the FP rate. The remaining images show a clear morphological disturbance or a clear merger companion. It is possible that these companions are being identified by the network as chance projections, especially the companions that are almost point-like in the image. Examples of the FN are shown in Fig. 5a-d. As with the FP objects, there are also example TP that are visually similar to the FN galaxies, presented in Fig. 5e-h.

4.2. Simulation Trained Network

The 14th epoch of the network trained with EAGLE images (simulation network) is used. This epoch has an accuracy of 0.692 at validation, cutting at a threshold of 0.5 to separate mergers from non-merger classification. Using the validation set, we plot the ROC curve for this network in dot-dashed yellow in Fig. 6. This has an area of 0.747 and provides an ideal cut threshold of 0.46. At this threshold, the accuracy of the validation set in-

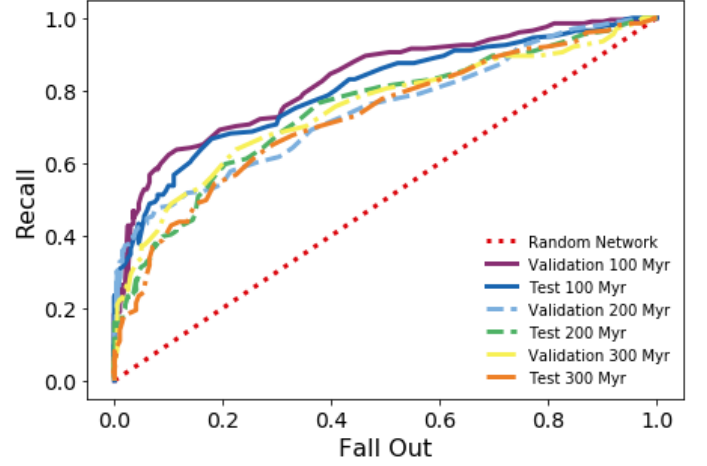


Fig. 6. ROC curve for the simulation networks at validation (purple, light blue, yellow) and testing (dark blue, green, orange) for 100 Myr (solid), 200 Myr (dashed) and 300 Myr (dot-dashed) from the merger event. The areas under the curves can be found in Table 5. The dashed red line shows the position of a truly random network.

creases to 0.706. To determine the true accuracy of the network, we perform the same analysis for the test data set. The area under the ROC curve, the dot-dashed orange curve in Fig. 6, decreases to 0.727. With the threshold set at 0.46, the final accuracy of the network is 0.679, with recall, precision, specificity and NPV of 0.652, 0.689, 0.706 and 0.670 respectively.

The lower accuracy of the simulation trained network relative to the observation trained network (discussed in Sect. 4.1) is a result of the difference in the training sample. The SDSS merger sample has been thoroughly checked to verify there are visible indications of a merger, as can be seen in the examples in Fig. 1. The EAGLE merger sample, however, contains physically classified mergers in the simulation without visual inspection to check whether there are any obvious signs of merging. As such, the EAGLE merger sample includes a wide variety of merger types (in terms of their mass ratios, orbital parameters, gas fractions, etc.) and hence some of the mergers are bound to have inconspicuous merging signs and will therefore be harder to discern, resulting in a lower accuracy. We have checked the merging galaxies misclassified by the network visually and confirm that most EAGLE mergers are indeed not as conspicuous as the ones in the SDSS catalogue, where the mergers from Darg et al. (2010a,b) have been selected to be conspicuous.

If the time before and after the merger event is decreased, the accuracy of the network increases. Performing the same analysis as with the full EAGLE data set, we find that using galaxies that are within 200 Myr of the merger event results in a network that has a test accuracy of 0.684 at a cut threshold of 0.39 in the 28th epoch. Using galaxies that are within 100 Myr of the merger event the network as a test accuracy of 0.744 at a cut threshold of 0.37 in the 60th epoch. The full statistics for these two networks can be found in Table 5 and the ROC curves can be found in Fig. 6. As the 100 Myr network has the greatest accuracy and largest area under the ROC curve, the majority of the remainder of the paper will now focus on the 100 Myr network when discussing the simulation network. The confusion matrix for the 100 Myr network can be found in Table 6.

A similar study has been performed by Snyder et al. (2018) using simulated galaxy images from the Illustris simulation (Vogelsberger et al. 2014), although their technique is somewhat dif-

Table 5. Statistics for the SDSS and the 100 Myr, 200 Myr and 300 Myr EAGLE trained networks at testing.

	SDSS	100 Myr	200 Myr	300 Myr
Epoch used	97	60	28	14
Cut threshold	0.57	0.37	0.39	0.46
ROC area	0.966	0.800	0.745	0.727
Recall	0.920	0.667	0.612	0.652
Precision	0.911	0.788	0.715	0.689
Specificity	0.910	0.821	0.756	0.706
NPV	0.919	0.711	0.661	0.670
Accuracy	0.915	0.744	0.684	0.679

Table 6. Confusion matrix for EAGLE images classified by the simulation network.

		Network Classification		Total
		Merger	Non-merger	
Catalogue Classification	Merger	134 TP	67 FN	201
	Non-merger	36 FP	165 TN	201
Total		170	232	

ferent. In their study, Snyder et al. (2018) train Random Forests using non-parametric morphology statistics, such as concentration, asymmetry, Gini and M_{20} , as inputs, with these statistics derived from Illustris galaxies processed to look like Hubble Space Telescope images. They select galaxies that will, or have, merge within 250 Gyr. The recall of Snyder et al. (2018) is slightly higher than this work, they achieve ≈ 0.70 compared to our 0.667, but their precision is much lower, at ≈ 0.30 compared to 0.788. Comparing the Snyder et al. (2018) results to our 300 Myr trained network, a more fair comparison, shows similar results: Snyder et al. (2018) has higher recall, ≈ 0.70 compared to 0.652, but lower precision, ≈ 0.30 compared to 0.689.

As the galaxies are generated from a simulation, we know the physical properties of these systems. As with the SDSS objects, we can compare the physical properties of the galaxies that are correctly and incorrectly identified. KS-test statistics comparing the distributions of correctly and incorrectly identified galaxies with the physical properties can be found in Table 7.

Many FN objects appear to have low simulation snapshot redshifts when compared to the TP, see Fig. 7a, potentially a result of coarser time resolution of the simulation at low redshift. Note that the simulation snapshot redshift is different from the redshift used when making the EAGLE galaxies look like SDSS images. For the TN and FP populations, higher snapshot redshifts have a higher fraction of FP sources relative to the TN, see Fig. 7b. This suggests that simulated non-mergers in the local universe look different from simulated non-mergers in the higher-z universe.

The asymmetry of the non-merger population also has an effect: non-merging objects with higher asymmetry are preferentially being identified as merging systems. It is worth noting that

Table 7. KS-test statistic, $D_{N,M}$, and the critical value, $\text{Crit}_{N,M} = c(\alpha) \sqrt{\frac{n+m}{nm}}$, for the EAGLE images classified by the simulation network. If $D_{N,M} > \text{Crit}_{N,M}$, the null hypothesis that the two distributions are the same is rejected at level $\alpha = 0.05$. Here, $c(\alpha) = 1.224$ for $\alpha = 0.05$ and n and m are the sizes of samples N and M .

Physical Parameter	$D_{TP, FN}$	$\text{Crit}_{TP, FN}$	$D_{TN, FP}$	$\text{Crit}_{TN, FP}$
Projection	0.156	0.189	0.223	0.192
Redshift				
Simulation	0.262	0.189	0.299	0.192
Redshift				
Asymmetry	0.169	0.189	0.251	0.192
Time since	0.092	0.188	-	-
Merger				
u-magnitude	0.462	0.189	0.448	0.192
g-magnitude	0.423	0.189	0.393	0.192
r-magnitude	0.385	0.189	0.326	0.192
i-magnitude	0.347	0.189	0.306	0.192
z-magnitude	0.323	0.189	0.278	0.192
Mass ratio	0.174	0.189	-	-
$M_{\star}$	0.209	0.189	0.209	0.192
sSFR	0.324	0.189	0.391	0.196

the time to/from the merger event does not appear directly correlated with the asymmetry of the galaxy.

In $M_{\star}$, there is a slight trend for the low mass, merging systems to be identified as non-mergers, although the non-merging galaxies are typically slightly lower mass than the merging systems so this is not overly unexpected. For sSFR, there is a splitting, with low sSFR merging systems being preferentially assigned the non-merger classification and the high sSFR non-merging galaxies preferentially identified as mergers, as shown in Fig. 8.

The apparent magnitude of the simulated galaxy after redshift projection has the largest effect, compared to the other parameters investigated, on the correct identification. For the merging systems, faint objects are preferentially classified as non-merging systems while the bright non-mergers are more likely to be misclassified as mergers. An example of this in the g-band is presented in Fig. 9. Misclassification for the merging systems is likely a result of the merging systems being brighter, on average, than the non-merger systems while the majority of the merging systems are fainter, hence the high misclassification rate for non-merging systems at these magnitudes.

As with the SDSS images, the misclassified EAGLE images have also been visually inspected. The majority of the FP galaxies, 29 of 36, contain a chance projection generated when the real SDSS noise is added. Three FP galaxies have a projected galaxy that is much brighter than the EAGLE galaxy, resulting in the EAGLE galaxy becoming extremely faint in the image and (almost) impossible to see by eye. There are also correctly identified galaxies that also suffer from the same image suppression, suggesting that this issue is not the sole cause of the misclassification. Of the remaining seven FP, one is at a low projection redshift, resulting in the features and inhomogeneities of the galaxy appearing as morphological disturbances, although, again, there are examples of these low projection redshift galaxies that have been correctly identified as non-merging. The other six objects show no signs of asymmetry or morphological disturbances. Examples of these galaxies can be found in Fig. 10a-d while example TN galaxies that are visually similar can be found in Fig. 10e-h.

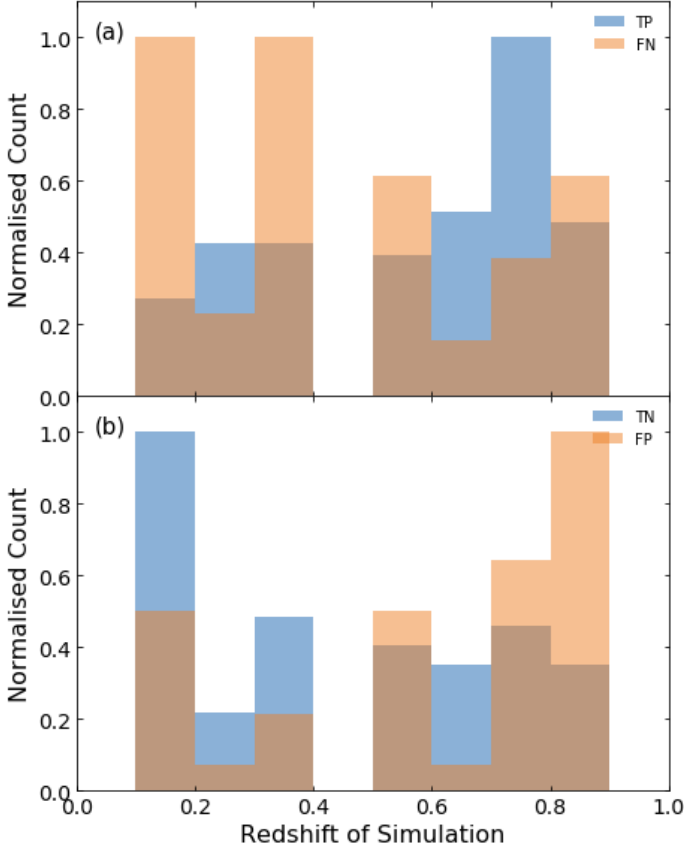


Fig. 7. Distributions for the correctly (blue) and incorrectly (orange) identified EAGLE objects in the simulation network for (a) mergers and (b) non-mergers as a function of simulation snapshot redshift. Merging objects with low snapshot redshifts are disproportionately assigned a non-merger classification while non-merging objects with high simulation redshifts are often seen as mergers.

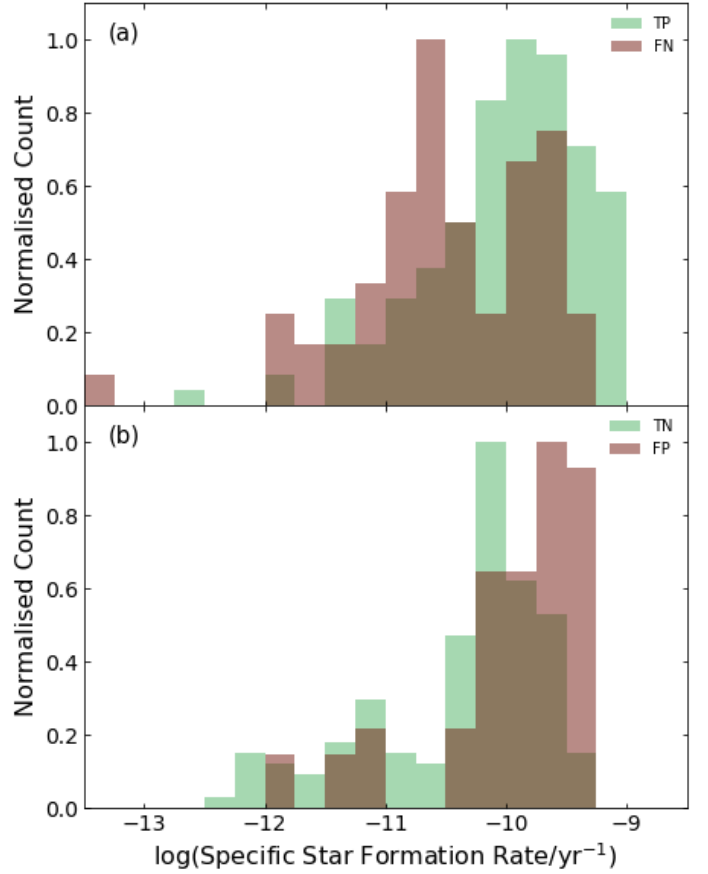


Fig. 8. Distribution of EAGLE galaxies from the simulation network of the correctly (green) and incorrectly (brown) mergers (a) and non-mergers (b) as a function of EAGLE sSFR. Merging galaxies with low sSFR are often misclassified as non-merging while high sSFR non-mergers are often identified as mergers.

For the FN objects, twelve of the 67 have a bright chance projection from the added SDSS noise that results in the EAGLE galaxy becoming (almost) impossible to see in the image. As these types of images are present in the FN, FP, TP and TN, it is unlikely that the bright counterpart is causing the misclassifications. 24 FNs do not appear morphologically disturbed or asymmetric. This is likely a result of the PSF convolution and redshift re-projection smoothing out the visual merger indicators resulting in what appears to be a single, smooth galaxy. The remaining 11 objects do have clearly identifiable merger counterparts or asymmetry. Examples of these galaxies can be found in Fig. 11a-d. As with the EAGLE FP, Fig. 11e-h show there are also examples of visually similar galaxies that have been correctly identified by the simulation network.

The CNN architecture was also trained on simulation images that had only been partially processed to look like SDSS images. This will allow us to determine if there is a specific part of the process that results in the lower accuracy for the simulation network with respect to the observation network. For this, we use EAGLE galaxies that are within 100 Myr of the merger event and perform one of the following processes: convolve the EAGLE image with the SDSS PSF (C), inject the EAGLE image into the real SDSS noise (N), match the EAGLE resolution to that of the SDSS images (R), adjust the EAGLE magnitude to be the correct apparent magnitude for a chosen redshift (Z) or a combination of three (CNR, CNZ, CRZ, NRZ). We also train

the network on the EAGLE images that have not been processed. As with training with SDSS or fully processed EAGLE images, the epoch with simultaneously the lowest validation loss and accuracy is chosen and the cut threshold with fall-out and recall closet to (0,1) is used. The statistics are then calculated for the test set and are presented in Table 8. Individually, C, N, R and Z do not notably change the accuracy of the trained network: all the accuracies are within a few percentage points of the accuracy of the un-processed EAGLE images, 87% although R does produce the single largest difference. Similarly, CRN and ZRC are within a few percentage points of 87%. However, CNZ and NRZ have much lower accuracies, around 75%, and are consistent with the fully processed EAGLE images. This suggests that the combination of N and Z is resulting in the lower accuracy, possibly because when changing from absolute to apparent magnitude, the fainter objects are becoming harder to discern when injected into the real SDSS noise. We note, however, that only 58 of the original 10 134 processed EAGLE images have an apparent r-band magnitude greater than the limit applied to the SDSS images.

4.3. Cross application of the networks

Here we pass the images through the other network, that is we pass all 6006 SDSS images through the simulation network and all 4020 EAGLE images through the observation network. For

Table 8. Statistics for the network trained with partially processed EAGLE images at testing. C is convolving the EAGLE image with the SDSS PSF, N is injecting the EAGLE image into the real SDSS noise, R is matching the EAGLE resolution to that of SDSS and Z is changing the EAGLE magnitude to apparent from absolute.

Processing	None	C	N	R	Z	CNR	CNZ	CRZ	NRZ
Epoch used	162	183	166	120	171	107	50	176	36
Cut threshold	0.45	0.43	0.47	0.44	0.55	0.39	0.35	0.39	0.36
ROC area	0.941	0.939	0.936	0.904	0.934	0.917	0.821	0.927	0.813
Recall	0.831	0.816	0.841	0.771	0.781	0.841	0.726	0.816	0.721
Precision	0.903	0.911	0.885	0.866	0.940	0.858	0.777	0.886	0.759
Specificity	0.910	0.920	0.891	0.881	0.950	0.861	0.791	0.896	0.771
NPV	0.843	0.833	0.848	0.794	0.813	0.844	0.743	0.829	0.735
Accuracy	0.871	0.868	0.866	0.856	0.866	0.851	0.759	0.856	0.746

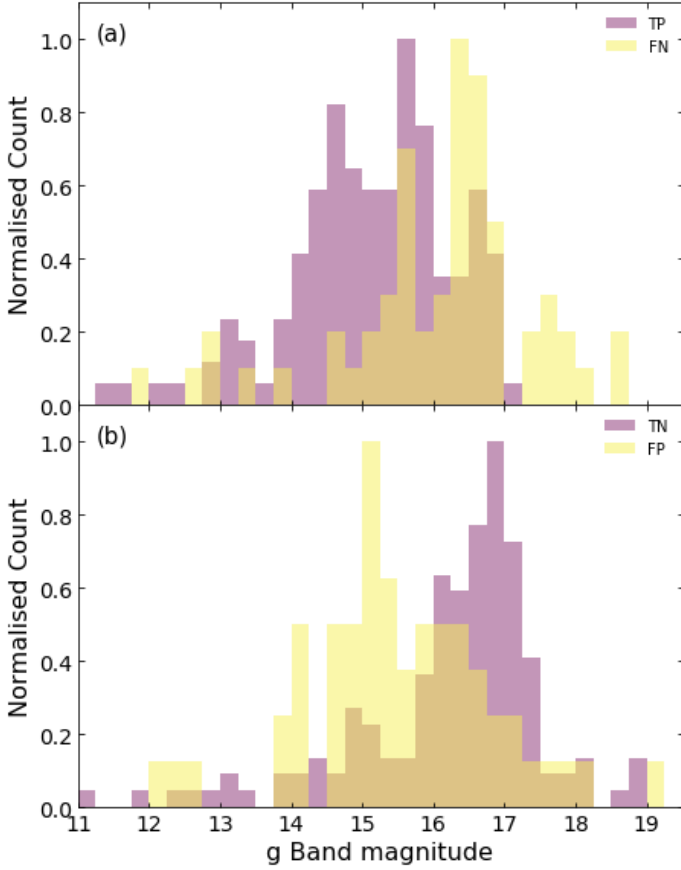


Fig. 9. Distributions for the correctly (purple) and incorrectly (yellow) identified objects for (a) mergers and (b) non-mergers as a function of g-band magnitude for EAGLE galaxies classified by the simulation network. Faint mergers are preferentially classified as non-mergers while the distribution of misclassified non-mergers is at intermediate magnitudes.

this, we use the same cut threshold as for passing through the “correct” images. This is done so that we can understand any biases and incompleteness in the two data sets. For example, the visually classified mergers from the SDSS data consist only of certain types of mergers with conspicuous merging signs, e.g., two massive galaxies obviously interacting with strong tidal features. However, the EAGLE simulation contains a much more complete merger sample. So, one would expect the neural network trained with the visually classified SDSS merger sample to perform poorly on simulated images of EAGLE mergers. We also perform the cross application so that any SDSS objects clas-

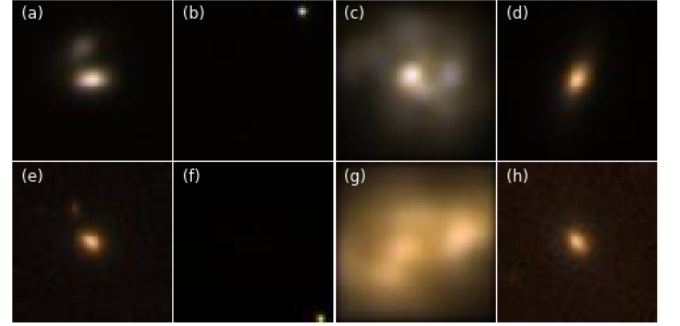


Fig. 10. Examples of FP EAGLE galaxies from the simulation network for (a) a chance projection, (b) a galaxy where the chance projection from the SDSS noise has resulted in the EAGLE galaxy appearing faint in the image, (c) a galaxy at low projection redshift and (d) an isolated, non-interacting galaxy. Panels (e) to (h) show TN galaxies that are visually similar to those shown in (a) to (d).

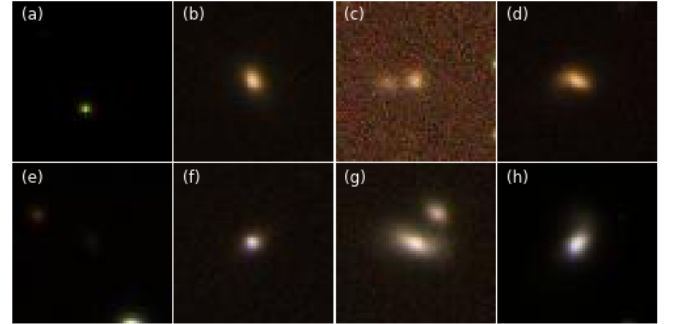


Fig. 11. Examples of FN EAGLE galaxies from the simulation network for (a) a galaxy where the chance projection from the SDSS noise has resulted in the EAGLE galaxy appearing faint in the image, (b) a merging system that appears as a single, smooth galaxy, (c) a galaxy with a clearly identifiable counterpart and (d) as asymmetric galaxy. Panels (e) to (h) show TP galaxies that are visually similar to those shown in (a) to (d).

sified as merging systems in the visual classification but not identified by the simulation trained network can be identified and help improve our understanding of the limitations of simulations so that they become more representative of the real universe in future developments.

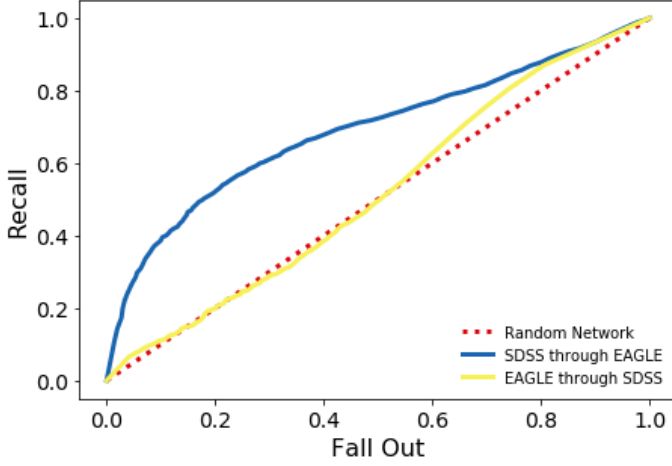


Fig. 12. ROC curve for the SDSS images classified by the simulation network (blue) and the EAGLE images classified by the observation network (yellow). The area under the EAGLE through observation network is 0.515 while the area under the SDSS through simulation network is 0.689. The dashed red line shows the position of a truly random network.

Table 9. Confusion matrix for EAGLE images classified by the observation network.

		Network Classification		Total
		Merger	Non-merger	
Catalogue Classification	Merger	470 TP	1540 FN	2010
	Non-merger	481 FP	1529 TN	2010
Total		951	3069	

4.3.1. EAGLE images through the observation network

Passing all the EAGLE images through the observation network resulted in an accuracy of 0.497, consistent at first glance with random assignment of objects. Similarly, precision and NPV are also close to random at 0.494 and 0.498. However, the recall is low, at 0.234, and the specificity is high, at 0.761, demonstrating that the network preferentially assigns objects to the non-merger class but with each class containing approximately half correct identifications and half incorrect identifications. As to be expected, the area under the ROC curve is close to 0.5 at 0.515, depicted in yellow in Fig. 12. The confusion matrix can be found in Table 9.

As before, the physical properties of the EAGLE images can be examined to determine if they are affecting the classification by the network, a brief summary of which can be seen in the KS-test results in Table 10. One property that has an obvious splitting between correct and incorrect assignment is the redshift of projection. As is evident in Fig. 13, objects with high projection redshifts are preferentially being classified as non-merger systems, see Fig. 13a, while objects with low projected redshifts are classified as merging systems, see Fig. 13b. The distribution of redshifts used to re-project the EAGLE galaxies

Table 10. KS-test statistic, $D_{N,M}$, and the critical value, $\text{Crit}_{N,M} = c(\alpha) \sqrt{\frac{n+m}{nm}}$, for the EAGLE images classified by the observation network. If $D_{N,M} > \text{Crit}_{N,M}$, the null hypothesis that the two distributions are the same is rejected at level $\alpha = 0.05$. Here, $c(\alpha) = 1.224$ for $\alpha = 0.05$ and n and m are the sizes of samples N and M .

Physical Parameter	$D_{TP, FN}$	$\text{Crit}_{TP, FN}$	$D_{TN, FP}$	$\text{Crit}_{TN, FP}$
Projection	0.261	0.060	0.479	0.059
Redshift	0.060	0.060	0.032	0.059
Simulation	0.154	0.060	0.055	0.059
Redshift	0.330	0.60	-	-
Asymmetry	0.273	0.060	0.471	0.059
Time since	0.287	0.060	0.486	0.059
Merger	0.297	0.060	0.480	0.059
u-magnitude	0.298	0.060	0.476	0.059
g-magnitude	0.293	0.060	0.474	0.059
r-magnitude	0.211	0.060	-	-
i-magnitude	0.197	0.060	0.097	0.059
z-magnitude	0.169	0.060	0.075	0.060
Mass ratio				
$M_{\star}$				
sSFR				

is nearly identical to the SDSS distribution: the redshifts used to re-project the galaxies were drawn randomly from the redshifts of the SDSS observations. Thus this effect is not a result of a mismatch in the redshift distributions between observations and simulations. The issue of misclassified mergers at high redshift may arise while matching the physical resolution (i.e. kpc per pixel) of the EAGLE images to the SDSS images. At high redshift, this could result in a loss of finer detail that would be expected in merging systems, resulting in these systems being classified as non-mergers. The main misclassification of non-merging systems happens at low projection redshift. The physical resolution of EAGLE images matches the physical resolution of the SDSS images at $z \approx 0.03$. Objects assigned a redshift lower than this value are increased in physical resolution using a bicubic interpolation. This interpolation may result in the creation of artifacts that appear, to the CNN, like features of merging systems. Alternatively, it is possible that at low redshifts the individual particles of the simulation are detectably disturbing the light profile of the galaxies and resulting in misclassification.

There is also a trend with the mass ratio of the merging systems. Although the TP and FN do not split into two distinct distributions, the low mass ratio merger systems, i.e. major mergers, are more often misclassified as non-merging galaxies. This is the opposite to what would be expected: minor mergers would be expected to be misclassified more often as the disturbances from the smaller galaxy would be expected to be less obvious. Similarly, low mass mergers have a slight preference to be assigned the non-merger class and vice versa, although again this is unsurprising as the merger sample is typically higher mass than the non-merger sample for the EAGLE galaxies. The mass ranges for the EAGLE and SDSS data are not quite comparable: while the SDSS data has an effective mass limit of $10^{10} M_{\odot}$ at $z = 0.1$, there are lower mass galaxies in the sample, while the mass limit for EAGLE is $10^{10} M_{\odot}$ at all projected redshifts. For the sSFR, the high sSFR merging galaxies are often misclassified as non-mergers while there is no obvious misclassification for the non-merging galaxies.

As with the EAGLE images through the simulation network, the apparent magnitude of the object has the largest effect on

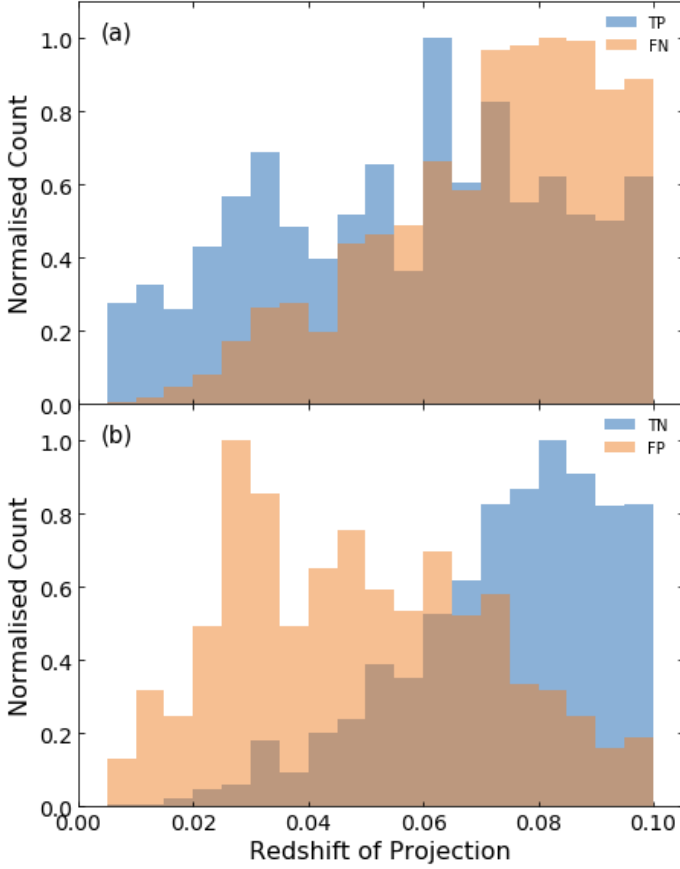


Fig. 13. Distributions for the correctly (blue) and incorrectly (orange) identified EAGLE objects for (a) mergers and (b) non-mergers as a function of redshift used for projection after being classified by the observation network. High redshift, merging systems are preferentially classified as non-merging while low redshift non-merging systems are preferentially classified as merging.

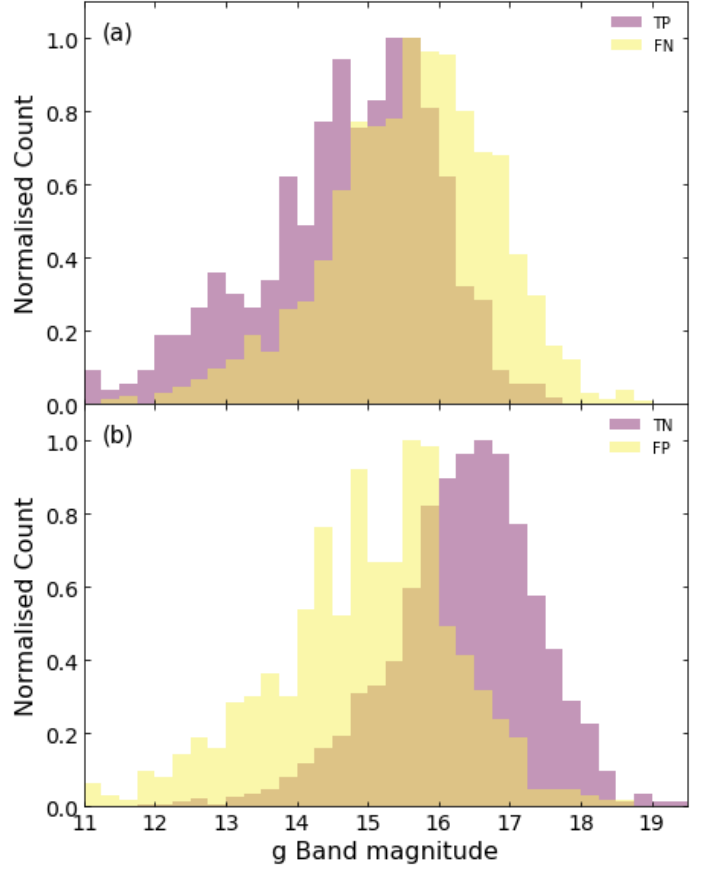


Fig. 14. Distributions for the correctly (purple) and incorrectly (yellow) identified EAGLE objects for (a) mergers and (b) non-mergers as a function of g-band magnitude after being classified by the observation network. Faint, merging systems are preferentially classified as non-merging while bright non-merging systems are preferentially classified as merging.

the correct classification. Like the simulation network, the observation network also identifies faint, merging systems as non-merging systems and identifies bright, non-merging systems as merging systems. This is true for all five of the ugriz bands. An example in the g-band is presented in Fig. 14. This is consistent with the results seen with the projected redshift above and is likely a result of the re-projection to the projection redshift.

It is also more likely that the more complete classification for the EAGLE galaxies is causing the low accuracy. The SDSS classifications are for objects that are clearly visually merging systems while the EAGLE classifications will include systems that are not obviously, visually merging. Thus, the observation trained network has not been trained to identify merging systems that are not obviously, visually merging and hence assign these objects the non-merger classification, increasing the number of FN.

Visual inspection of the 481 FP shows that the majority of these objects (293) appear to be isolated, non-interacting systems. A further 66 objects have a close chance projection that may be being mistaken for a merging partner by the CNN. 36 objects are galaxies that have been projected into a larger angular size than the original, raw image from EAGLE. This often results in the internal structure of the galaxy being expanded and could appear to the network to be morphological disturbances or multiple galaxies. The remaining 32 objects have a bright chance projection in the SDSS noise and, as a result, are (almost) im-

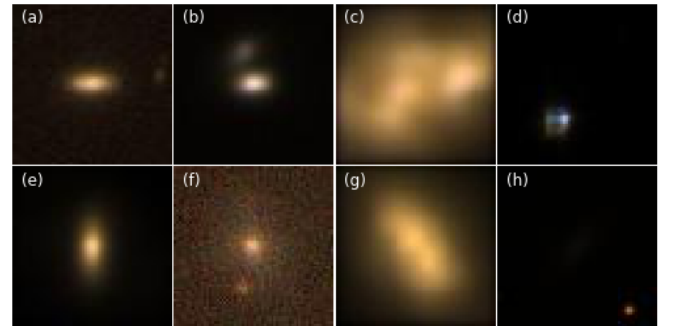


Fig. 15. Examples of EAGLE FP galaxies (a to d) from the observation network for (a) an isolated, non-interacting galaxy, (b) a chance projection, (c) a galaxy at low projection redshift and (d) a galaxy where the chance projection from the SDSS noise has resulted in the EAGLE galaxy appearing faint in the image. Panels (e) to (h) show TN galaxies that are visually similar to those shown in (a) to (d).

possible to see in the image. Examples of these galaxies can be found in Fig. 15a-d. With accuracy, specificity and NPV all being almost equivalent to 0.5, it is unsurprising to find examples of visually similar galaxies that have been correctly identified and are shown in Fig. 15e-h.

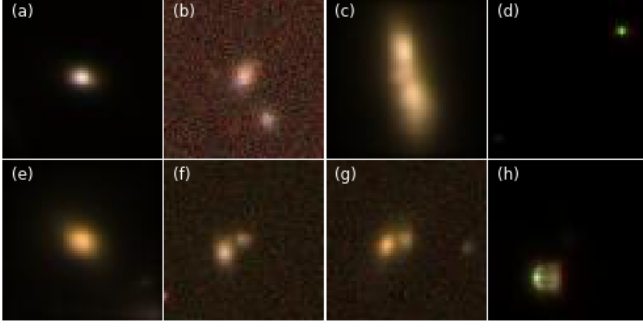


Fig. 16. Examples of EAGLE FN galaxies (a to d) from the observation network for (a) an apparent single object, (b) a galaxy with a counterpart, either a merger counterpart from EAGLE or a chance projection from the SDSS noise, (c) an unambiguous merger and (d) a galaxy where the chance projection from the SDSS noise has resulted in the EAGLE galaxy appearing faint in the image. Panels (e) to (h) show TP galaxies that are visually similar to those shown in (a) to (d).

Images of the FN are more useful in understanding why the EAGLE images are poorly classified by the observation network. Of the 1540 FN, the vast majority (931) appear to be a single object when visually inspected. This suggests that these objects have had the visible signatures of merger suppressed while being processed to look like SDSS images, likely by the re-projection and PSF matching, or that these mergers are not obvious, even without the processing steps to make them look like SDSS images. It could also be that the merging companion is hidden behind the galaxy it is merging with, as the angle the galaxy is viewed at is picked randomly, so it cannot be seen within the image. Of the remaining objects, 383 had at least one counterpart, either from the simulation or random projections from the SDSS noise, that could potentially be merging with the central galaxy and 52 were unambiguously merging systems. As with the FP, there are a small number of images (173) whose simulated galaxies have been suppressed by bright chance projections from the SDSS noise. Only one FN object has been projected into a larger angular size than can fit within the 64×64 pixel image. Example FN galaxies can be found in Fig. 16a-d and their visually similar but correctly identified counterparts can be found in Fig. 16e-h.

4.3.2. SDSS images through the simulation network

Passing all the SDSS images through the simulation network was more successful than passing all the EAGLE images through the observation network. While still not as good as SDSS images through the observation network, the SDSS images classified by the simulation network had an accuracy of 0.639. Like passing the EAGLE images through the observation network, the precision and specificity are reasonably similar to the accuracy at 0.630 and 0.603. However, unlike passing the EAGLE images through the observation network, the recall and NPV are also reasonably similar to each other, at 0.676 and 0.651, showing that the network is not preferentially assigning the objects to a single class. The area under the ROC curve is 0.689, see the blue line in Fig. 12. The statistics for the cross application of the networks can be found in Table 11. The confusion matrix, showing the number of correctly and incorrectly identified objects, can be found in Table 12.

As with the SDSS images identified by the observation network, we can examine the estimated physical parameters of

Table 11. Statistics for the EAGLE images classified by the observation network and the SDSS images classified by the simulation network.

Images	EAGLE	SDSS
Network	Observation	Simulation
Cut threshold	0.57	0.37
ROC area	0.515	0.689
Recall	0.234	0.676
Precision	0.494	0.630
Specificity	0.761	0.603
NPV	0.498	0.651
Accuracy	0.497	0.639

Table 12. Confusion matrix for SDSS images classified by the simulation network.

		Network Classification		Total
		Merger	Non-merger	
Catalogue Classification	Merger	2031 TP	972 FN	3003
	Non-merger	1193 FP	1810 TN	3003
Total		3224	2782	

the SDSS images that were classified by the simulation network. The KS-test statistics comparing the distributions of correctly and incorrectly identified galaxies with the physical properties can be found in Table 13. There is an obvious splitting in the distributions of M_* for correctly and incorrectly identified objects: the high mass merging objects are preferentially assigned the non-merger classification, while the intermediate mass non-merging objects are preferentially assigned the merger classification. Although no low mass mergers being assigned the non-merging class is reassuring as there are no low mass non-merging objects. This splitting may arise from the training sample having non-merging systems as preferentially high mass and merging objects as preferentially intermediate and low mass. A similar, but opposite, split is seen with sSFR: low sSFR mergers are identified as non-mergers, as seen in Fig. 17a, while high sSFR non-mergers have a higher misclassification rate than low sSFR non-mergers, as seen in Fig. 17b. This suggests that the EAGLE images for merging systems may preferentially show boosted sSFR.

The trend of the ugriz band magnitudes of the SDSS images is also interesting. As the band becomes more red, from g through to z, the distributions of correctly and incorrectly identified objects become more and more split, as can be seen by the increasing KS-test statistic in Table 13. Thus, as the band becomes redder, more and more bright mergers are classified as non-mergers while the faint objects are correctly classified more often. Similarly, as the bands become redder, the distribution of incorrectly identified non-mergers moves to the fainter end. An example of the z-band magnitude distribution is shown in Fig. 18. The trend that is seen for misclassification in the merging systems is the opposite of the effect seen in the EAGLE test set when classified by the simulation network.

Table 13. KS-test statistic, $D_{N,M}$, and the critical value, $\text{Crit}_{N,M} = c(\alpha) \sqrt{\frac{n+m}{nm}}$, for the SDSS images classified by the simulation network. If $D_{N,M} > \text{Crit}_{N,M}$, the null hypothesis that the two distributions are the same is rejected at level $\alpha = 0.05$. Here, $c(\alpha) = 1.224$ for $\alpha = 0.05$ and n and m are the sizes of samples N and M .

Physical Parameter	$D_{TP, FN}$	$\text{Crit}_{TP, FN}$	$D_{TN, FP}$	$\text{Crit}_{TN, FP}$
$M_\star$	0.542	0.045	0.492	0.055
sSFR	0.508	0.045	0.537	0.055
u-magnitude	0.153	0.045	0.164	0.055
g-magnitude	0.163	0.045	0.116	0.055
r-magnitude	0.359	0.045	0.281	0.055
i-magnitude	0.407	0.045	0.325	0.055
z-magnitude	0.447	0.045	0.365	0.055

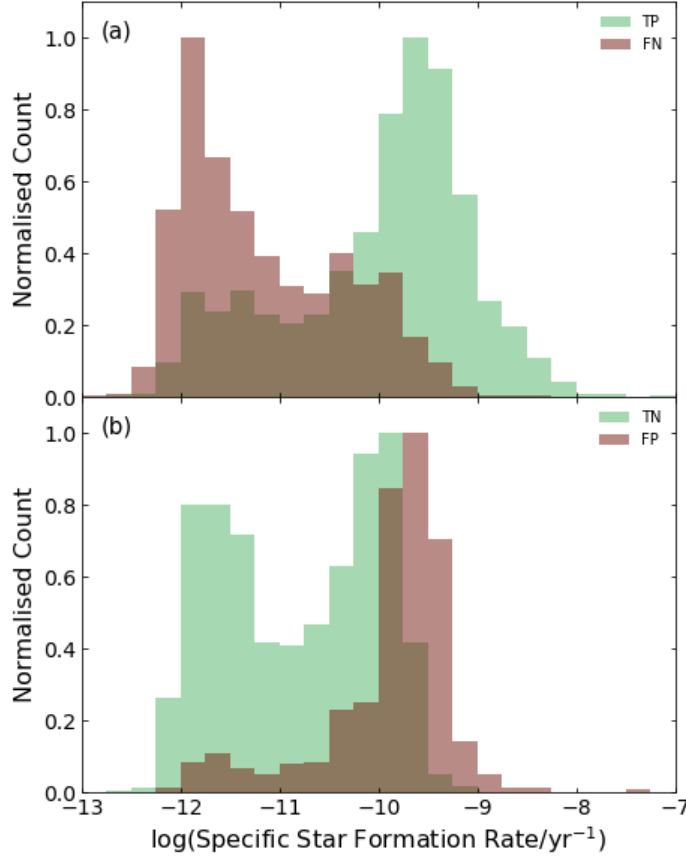


Fig. 17. Distributions for the correctly (green) and incorrectly (brown) identified SDSS objects for (a) mergers and (b) non-mergers as a function of sSFR after being classified by the simulation network. Low sSFR merging systems are preferentially classified as non-merging while high sSFR non-merging systems have a higher misclassification rate than low sSFR non-merger.

The 1193 FP have been visually inspected. 467 of the FP have at least one other galaxy that lie close to the primary galaxy but are not visually interacting with the primary. These secondary galaxies are likely being identified as a merging companion to the primary or they are possibly merging systems that appear in simulations but are not identified as such in Galaxy Zoo. The majority of the FP (684) are unambiguous, non-interacting, isolated galaxies. This is possibly a result of many merging systems in the EAGLE training set visually looking like single, undisturbed galaxies. However, that does not exclude these

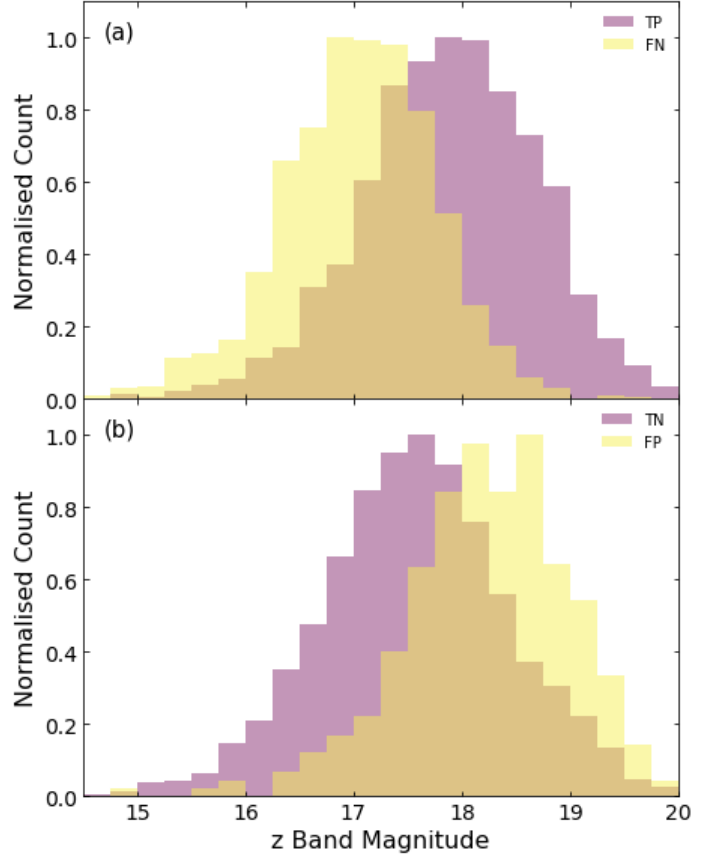


Fig. 18. Distributions for the correctly (purple) and incorrectly (yellow) identified SDSS objects after being classified by the simulation network for (a) mergers and (b) non-mergers as a function of z-band magnitude. Bright mergers are preferentially classified as non-mergers while the distribution of misclassified non-mergers is skewed towards the faint end of the distribution. This trend becomes less pronounced as the bands become more blue, from z to u-band.

galaxies from being true mergers as the EAGLE training set should be more complete than the SDSS images. A further 28 objects show signs of asymmetry or morphological disturbances. As with the misidentified chance projections, this may be a result of the strict selection for merging SDSS systems ignoring these galaxies but the more complete selection from EAGLE identifying these as mergers. The remaining 14 galaxies contain a non-physical artifact, typically a single pixel width black line through the galaxy, although there are also a number of TN that also have similar artifacts, so this is unlikely to be causing the misclassification. Example FP galaxies can be found in Fig. 19a-d and the visually similar TN in Fig. 19e-h.

Alongside the 1193 FP, there are 972 FN. The majority of these systems (731) clearly show two interacting galaxies, which may be a result of the network identifying these as chance projections. A further 36 objects show clear evidence of morphological disturbances or asymmetry. 205 of the FN have their counterpart of the edge or outside the image cut-out. This suggests, like the observation network, that a larger cut-out may help identify these objects. Examples of these objects can be found in Fig. 20a-d and examples of TP galaxies that look similar in Fig. 20e-h.

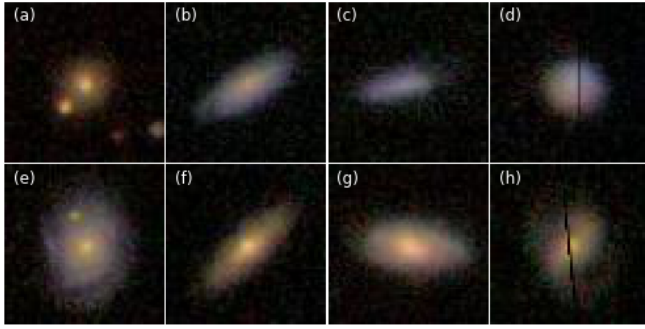


Fig. 19. Examples of SDSS FP galaxies from the simulation network for (a) a galaxy with a close (in projection) companion, (b) a non-interacting, isolated galaxy, (c) a galaxy showing asymmetry or morphological disturbance and (d) a galaxy with a non-physical artifact within the image. Panels (e) to (h) show TN galaxies that are visually similar to those shown in (a) to (d).

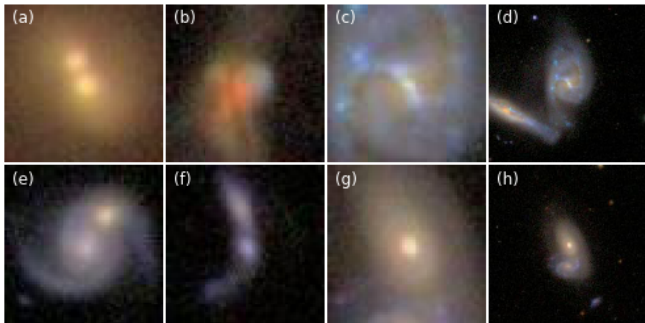


Fig. 20. Examples of SDSS FN galaxies from the simulation network for (a) a galaxy with a clear merging counterpart, (b) a clearly disturbed system, (c) a galaxy whose merger companion is outside of the 64×64 pixel image and (d) the larger 256×256 pixel image showing the merger companion outside of panel (c). Panels (e) to (h) show TP galaxies that are visually similar to those shown in (a) to (d).

5. Conclusions

Training and applying a CNN on SDSS images has been successful, achieving an accuracy of 91.5%. This clearly demonstrates that CNNs can be used to reproduce visual classification. There is no clear indication of a specific type of object that is incorrectly identified from the physical or observable parameters. Training and applying a CNN on the EAGLE images was also reasonably successful, with an accuracy of 74.4% when trained using mergers that will or have occurred within 100 Myr of the image snapshot. Using a longer time between the image snapshot and the merger reduces the accuracy of the network. This relatively lower accuracy suggests that some EAGLE mergers do not have visible merging features that can be picked up by the CNN. The incorrectly identified mergers are primarily at low simulation snapshot redshifts as well as faint apparent magnitude. The combination of real noise added to the EAGLE images and converting the absolute magnitude to apparent magnitude also reduces the effectiveness of the CNN, which demonstrates the importance of image quality (in terms of, for example, signal-to-noise and resolution) in merger identification. Within the image, chance projections result in a large number of non-merging galaxies being identified as mergers.

The lower accuracy of the EAGLE trained network is most likely a result of the difference in the training sample. The SDSS merger sample has been selected to contain conspicuous mergers and so the features of a merger are more easily identified but will miss subtler mergers. Meanwhile, the EAGLE sample has fewer conspicuous mergers but should be more complete (including mergers with a wide range of mass ratios, gas fractions, viewing angles, environments, orbital parameters, etc.), resulting in less obvious merger features, in pixel space, that are harder for a CNN to recognise.

Passing the SDSS images through the EAGLE trained network has proven to work, although with only 64.3% accuracy. This relatively low accuracy appears to be a result of high mass or low sSFR objects being identified as non-mergers and low mass or high sSFR objects being identified as mergers. This could suggest that simulations show evidence of high sSFR in the merging systems when this may not necessarily be true. However, the EAGLE trained network may also be identifying merging systems that the visual classification missed. The EAGLE classification will be more complete, as we know which systems are merging, and so the EAGLE trained network may be identifying these objects in the SDSS images that have been missed by the less complete, but more visually obvious, SDSS classification. The result may be a lower specificity, that is a smaller fraction of non-mergers are being correctly identified, when using the SDSS classifications as the truth when in fact the EAGLE trained network is correctly identifying merging systems missed by the human visual classification. However, the relatively low recall, the fraction of mergers correctly identified, suggests that EAGLE has relatively few conspicuous mergers.

This has a tantalizing prospect for large upcoming surveys, such as LSST and *Euclid*. It is possible to train a CNN with images from simulations and apply it to observations of galaxies from the real universe. Presently, the simulation trained network could be used to generate a set of galaxy merger candidates, which would need to be checked by a human expert, for use in training an observation network. However, with further refinement to the training images from simulations it is not beyond the realm of possibility to reduce the need for an observation training set and apply a simulation trained CNN directly to images from an entire survey, massively speeding up identification.

Passing the EAGLE images through the SDSS trained network was unsuccessful, with the network preferentially assigning objects to the non-merger class. This suggests that some EAGLE mergers are not representative of the SDSS selected mergers, although this appears to be primarily due to how re-projecting the galaxies to their assigned redshift has been done, so it may not be that the EAGLE mergers themselves do not look like observable mergers. The mergers in EAGLE are also less conspicuous than those in the SDSS training set so the observational network has not been trained to identify these less obvious merger events, resulting in a large number of EAGLE mergers being identified as non-mergers.

Improvements for the simulation galaxies in future work would be to increase the mass resolution, which can affect the appearance of galaxies and galaxy mergers (Sparre et al. 2015; Torrey et al. 2015; Trayford et al. 2015; Sparre & Springel 2016), and exactly match the stellar mass distributions with those of observations. Increasing the time resolution, for example by using the snapshots⁹ instead of snapshots from EAGLE, should also provide improvement along with improving the estimates of time to or since the merger event by tracing when the central black

⁹ High time resolution output from EAGLE.

holes merge. It would also be informative to include the effects of dust attenuation.

As has been shown in this work, chance projections are a major problem in merger classification. In future work, we could train the network to recognise chance projections better by constructing training samples of galaxies which appear close together in the sky but are actually far away from each other. Another area to improve in the future is to come up with a more refined merger classification system, rather than just a binary classification of merger versus non-mergers. For example, have separate classes of early mergers, late mergers, minor mergers, major mergers etc.

Acknowledgements. We would like to thank S. Ellison and L. Bignone for helpful discussions that have improved this paper.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is <http://www.sdss.org/>. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

References

- Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from [tensorflow.org](https://www.tensorflow.org)
- Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K., & Turp, M. D. 2018, MNRAS, 479, 415
- Barton, E. J., Geller, M. J., & Kenyon, S. J. 2000, ApJ, 530, 660
- Bell, E. F., Zucker, D. B., Belokurov, V., et al. 2008, ApJ, 680, 295
- Benítez-Llambay, A. 2017, Py-SPHViewer: Cosmological simulations using Smoothed Particle Hydrodynamics, Astrophysics Source Code Library
- Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645
- Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
- Chollet, F. 2017, in 2017 IEEE Conference on Computer Vision and Pattern Recognition
- Conselice, C. J. 2014, ARA&A, 52, 291
- Conselice, C. J., Bershady, M. A., Dickinson, M., & Papovich, C. 2003, AJ, 126, 1183
- Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886
- Cortijo-Ferrero, C., González Delgado, R. M., Pérez, E., et al. 2017, A&A, 607, A70
- Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010a, MNRAS, 401, 1552
- Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010b, MNRAS, 401, 1043
- De Propris, R., Liske, J., Driver, S. P., Allen, P. D., & Cross, N. J. G. 2005, AJ, 130, 1516
- Deng, J., Dong, W., Socher, R., et al. 2009, in 2009 IEEE Conference on Computer Vision and Pattern Recognition
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
- Genel, S., Genzel, R., Bouché, N., Naab, T., & Sternberg, A. 2009, ApJ, 701, 2002
- Guo, Q. & White, S. D. M. 2008, MNRAS, 384, 2
- Gwyn, S. D. J. 2012, AJ, 143, 38
- Hopkins, P. F., Hernquist, L., Cox, T. J., et al. 2006, ApJS, 163, 1
- Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
- Ioffe, S. & Szegedy, C. 2015, International Conference on Machine Learning
- Johnston, K. V., Hernquist, L., & Bolte, M. 1996, ApJ, 465, 278
- Joseph, R. D. & Wright, G. S. 1985, MNRAS, 214, 87
- Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 33
- Kingma, D. P. & Ba, J. 2015, 3rd International Conference for Learning Representations [arXiv:1412.6980]
- Knapen, J. H. & Cisternas, M. 2015, ApJ, 807, L16
- Knapen, J. H., Cisternas, M., & Querejeta, M. 2015, MNRAS, 454, 1742
- Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18
- Lambas, D. G., Tissera, P. B., Alonso, M. S., & Coldwell, G. 2003, MNRAS, 346, 1189
- Lang, M., Holley-Bockelmann, K., & Sinha, M. 2014, ApJ, 790, L33
- Larson, D., Dunkley, J., Hinshaw, G., et al. 2011, ApJS, 192, 16
- Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, in Proceedings of the IEEE, Vol. 86, Issue 11, 2278–2324
- Lin, L., Koo, D. C., Willmer, C. N. A., et al. 2004, ApJ, 617, L9
- Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179
- Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163
- LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, ArXiv e-prints [arXiv:0912.0201]
- Luo, W., Yang, X., & Zhang, Y. 2014, ApJ, 789, L16
- Naab, T. & Burkert, A. 2003, ApJ, 597, 893
- Nair, V. & Hinton, G. E. 2010, in Proceedings of the 27th international conference on machine learning (ICML-10), 807–814
- Patton, D. R., Pritchett, C. J., Carlberg, R. G., et al. 2002, ApJ, 565, 208
- Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129
- Qu, Y., Helly, J. C., Bower, R. G., et al. 2017, MNRAS, 464, 1659
- Robitaille, T. 2018, reproject: Image reprojection (resampling), <https://reproject.readthedocs.io/en/stable/>, accessed: 2019-02-12
- Salim, S., Rich, R. M., Charlot, S., et al. 2007, ApJS, 173, 267
- Sanders, D. B. & Mirabel, I. F. 1996, ARA&A, 34, 749
- Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521
- Schweizer, F. 2005, in Astrophysics and Space Science Library, Vol. 329, Starbursts: From 30 Doradus to Lyman Break Galaxies, ed. R. de Grijs & R. M. González Delgado, 143
- Sinha, M. & Holley-Bockelmann, K. 2012, ApJ, 751, 17
- Smirnov, N. V. 1939, Bulletin Moscow University, 2, 3
- Snyder, G. F., Rodriguez-Gomez, V., Lotz, J. M., et al. 2018, arXiv e-prints [arXiv:1809.02136]
- Somerville, R. S. & Davé, R. 2015, ARA&A, 53, 51
- Sparre, M., Hayward, C. C., Springel, V., et al. 2015, MNRAS, 447, 3548
- Sparre, M. & Springel, V. 2016, MNRAS, 462, 2418
- Springel, V., Di Matteo, T., & Hernquist, L. 2005, MNRAS, 361, 776
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, Journal of Machine Learning Research, 15, 1929
- Torrey, P., Snyder, G. F., Vogelsberger, M., et al. 2015, MNRAS, 447, 2753
- Trayford, J. W., Frenk, C. S., Theuns, T., Schaye, J., & Correa, C. 2019, MNRAS, 483, 744
- Trayford, J. W., Theuns, T., Bower, R. G., et al. 2015, MNRAS, 452, 2879
- Vogelsberger, M., Genel, S., Springel, V., et al. 2014, MNRAS, 444, 1518
- Walmsley, M., Ferguson, A. M. N., Mann, R. G., & Lintott, C. J. 2019, MNRAS, 483, 2968
- Wu, K. L., Faber, S. M., & Lauer, T. R. 2001, in Deep Fields, ed. S. Cristiani, A. Renzini, & R. E. Williams, 170
- York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579