MNRAS **000**, 1–13 (2019) Preprint 13 September 2019 Compiled using MNRAS LATEX style file v3.0

Deep learning classifications of streams and shells in galaxies from the Hyper-Suprime Cam DR1

Connor Bottrell[1]★, Ryan Hausen[2], Helena Domíngez Sánchez[3], Ivana Damjanov[4,5], Marc Huertas-Company[6,7], Kathryn V. Johnston[8], & Brant E. Robertson[9]

[1]*Department of Physics and Astronomy, University of Victoria, Victoria, British Columbia V8P 1A1, Canada*
[2]*Department of Computer Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA*
[3]*Department of Physics and Astronomy, University of Pennsylvania,Philadelphia, PA 19104, USA*
[4]*Department of Astronomy and Physics, Saint Mary's University, 923 Robie Street, Halifax, NS B3H 3C3, Canada*
[5]*Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA*
[6]*Instituto de Astrofísica de Canarias (IAC); Departamento de Astroísica, Universidad de La Laguna (ULL), E-38200, La Laguna, Spain*
[7]*LERMA, Observatoire de Paris, CNRS, PSL, Université Paris Diderot, France*
[8]*Department of Astronomy, Columbia University, 550 West 120th Street, New York, NY 10027, USA*
[9]*Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Comparison of mergers with non-merging galaxies while controlling for environment, gas fractions, and total stellar mass have highlighted the critical roll of mergers in triggering key processes in galaxy formation and evolution. However, different merger scenarios (e.g. orbital histories and virial accretion times) yield a corresponding range in the enhancement (e.g. star-formation, AGN activity) or suppression (e.g. central gas metallicity) of merger-triggered phenomena. Given that the initial merger scenarios are encoded in the morphologies of stellar debris found around merger remnants (shells/streams), these features can be used to further separate merger samples and make more detailed comparisons between observations and numerical predictions. In this paper, we use convolutional neural networks (CNNs) to first identify stream and shell features in galaxies from the Subaru Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP). The CNNs are trained using an unprecedentedly large stream/shell search catalogue containing 1,201 galaxies with stream and shell feature detections (with 987 streams, 214 shells, and a handful with both labels) and 20,010 non-detections. Using i-band imaging alone, we achieve test accuracies and F_1 statistics which are consistent with those reported by other works using smaller samples (TPR or recall 80%, FPR or contamination 26%). We find that one possible limitation to the performance of the CNNs (and the source of overfitting) is the large number of misclassified images in the base catalogue. Many of the images classified as detections by our CNNs which are non-detections according to the catalogue (false positives) exhibit strong stream and shell features. Applying a CNN which detects and distinguishes between stream and shell features to the full HSC-SSP imaging campaign will require careful consideration of misclassification and possible redshift and resolution biases.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

Galaxy mergers have a fundamental and critical role within the Lambda cold dark matter (ΛCDM) concordance cosmogony (e.g. White & Rees 1978; White & Frenk 1991). In this paradigm, large structures are assembled through continuous and diverse merging events between smaller structures (e.g. Lacey & Cole 1993).

The role of galaxy mergers is not limited to the *ex-situ* assembly of stellar mass in galaxies. The last five decades of theoretical, observational and numerical investigations demonstrates that gas-rich mergers are also responsible for *in-situ* growth. This in-situ growth is facilitated through central starbursts triggered by the tidal torques and shocks involved in galaxy interactions (e.g. Toomre & Toomre 1972; Hernquist 1989; Barnes & Hernquist 1991; Ellison et al. 2008; Patton et al. 2013; Blumenthal & Barnes 2018). Mergers garner interest from both observational and theoretical perspectives because

★ E-mail: cbottrel@uvic.ca

they are laboratories in which many of the poorly-constrained aspects of galaxy formation theory can be studied. The star formation triggered in gas-rich mergers simultaneously drives outflows and regulates further star formation by injecting energy and turbulence into the interstellar medium (ISM, e.g. Hayward & Hopkins 2017; Moreno et al. 2019). The circum-galactic medium (CGM) is enriched by such outflows – with observable signatures in its size and covering fractions for various gas-phase species (Hani et al. 2018). The same gas inflows that can boost central star-formation can also be accreted onto the central super-massive black hole (SMBH) and trigger an active galactic nucleus (AGN) (e.g. Keel et al. 1985; Hernquist 1989; Koss et al. 2010; Satyapal et al. 2014; Ellison et al. 2019). These *particular* merger-induced processes have galactic-scale consequences but originate from spatial scales which cannot be simultaneously and explicitly modelled with numerical hydrodynamical simulations. Consequently, sub-grid recipes must be adopted. Well-motivated observational constraints on these processes are therefore of great interest because they enable validation or rejection of those specific aspects of the models.

However, to compare between observations and numerical predictions, one must be able to distinguish between different merger scenarios or histories. Numerical simulations demonstrate that the enhancement in star-formation, AGN accretion, and subsequent cold gas depletion are sensitive to the initial orbits of mergers at fixed mass ratios and gas fractions (e.g. Cox et al. 2008). Intuitively, mergers with small impact parameters (low eccentricities) generate the largest responses. Mergers with more eccentric orbits have suppressed responses with respect to low-eccentricity orbits. If these initial orbital conditions can be estimated for observed mergers, then the merger-induced responses seen in the observations and numerical simulations can be compared without suppressing the responses in each dataset by statistically averaging over all merger scenarios.

Fortunately, the orbital scenarios are encoded in the morphological features produced in galaxy interactions. The seminal analysis demonstrating this encoding is Johnston et al. (2008, in particular, see their Figure 3) building on previous analytic and numerical works (e.g. Hernquist & Quinn 1989; Helmi & White 1999) . Johnston et al. (2008) showed that one can interpret the initial orbital conditions of a merger using the morphological features in the stellar haloes of merging galaxies. Mergers between galaxies with high-eccentricity orbits tend to produce stellar streams (as long as the initial orbital energy has had sufficient time to decay through dynamical friction). In contrast, mergers along low-eccentricity (more radial) orbits tend to produce shells as the companion passes through a much deeper potential and is consequently more heavily disrupted. These findings have enormous implications for the task of estimating the initial orbital histories of mergers. Additionally, using methods which forward-model galaxies from hydrodynamical simulations into realistic synthetic observations (e.g. SKIRT, Baes et al. 2011; Camps & Baes 2015; RealSim, Bottrell et al. 2017, Bottrell et al. 2019 (submitted)), one can (1) calibrate a method which identifies and distinguishes between streams and shells *where the merger initial conditions are known* and (2) analyze both mock-observed simulations and observed data with the same methodology. But first, the limitations (if any) of generating a model which identifies streams and shells in galaxies based strictly on observational data should be known.

A remaining challenge to stream and shell detection is their detectability within survey surface brightness and resolution limitations. Streams and shells are often low surface brightness features and are consequently the first features to be lost as image quality degrades or galaxy redshift increases (e.g. Lotz et al. 2008, 2010). Therefore, a deep survey with sufficient sky coverage that statistically large samples of stream and shell hosts is required. In an effort to satisfy both of these criteria, we use imaging from the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP). The HSC-SSP is an ongoing wide-field survey being carried out with the 8.2m Subaru Telescope (Miyazaki et al. 2012; Furusawa et al. 2018; Komiyama et al. 2018; Kawanomoto et al. 2018; Miyazaki et al. 2018). Once completed, the HSC-SSP will have a sky coverage of 1400 deg^2 in five broadband filters which span the optical and near-infrared ($grizy$). The HSC-SSP has a target 5σ point source depth of 26 in the HSC i-band and a median i-band seeing of 0.56 arcseconds (Aihara et al. 2018a,b). The HSC-SSP thus has a potential for producing statistical samples of low surface=brightness features that can be further characterized quantitatively (for example, using the methods presented in Hendel & Johnston 2015).

Presently, observational stream and shell detection is dominated by visual classification (e.g. Malin & Carter 1983; Tal et al. 2009; Nair & Abraham 2010; Atkinson et al. 2013; Hood et al. 2018; Morales et al. 2018). Given the sensitivity of tidal feature detection to photometric limitations, these studies are by no means homogeneous and exhibit variability in (1) the incidence of stream and shell features in galaxies which are shared by multiple studies and consequently (2) the fractional incidence of galaxies which exhibit tidal features reported by these studies. A large and homogeneous sample of expertly classified galaxies is therefore desired (and one that satisfies the photometric requirements from the last paragraph). However, this endeavour is limited by the rate at which expert visual classifications can be performed. Efforts that combine visual classifications with the flexibility and speed of deep learning models offer a solution to this problem (e.g. Ackermann et al. 2018; Walmsley et al. 2019). In particular, convolutional neural networks (CNNs) are a class of deep learning models which have proven useful for morphological identification/classification tasks (e.g. Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018; Huertas-Company et al. 2019). For a homogeneous survey, a sufficiently representative subsample of galaxy images can be used to train a CNN. The trained CNN can then rapidly classify galaxies in the rest of the survey at a rate that is unachievable with human classifiers. The speed with which galaxies can be classified with CNNs is also necessary given the size of modern observing programs like the HSC-SSP.

In this paper, we aim to improve detection and characterization of stream and shell features by training CNNs on deep HSC-SSP imaging. We use the Kado-Fong et al. (2018) visual classification sample for our training and validation images. Kado-Fong et al. (2018) performed their visual classifications with 512×512 pixel cutouts (87×87 arcsec^2) using all $grizy$ bands, centred on galaxies in the overlap between Data Release (DR) 1 of the HSC-SSP Wide layer (Aihara et al. 2018b) and the Sloan Digital Sky Survey DR12 spectroscopic galaxy sample (Alam et al. 2015). These selection criteria yield a total of 21,208 galaxies making the Kado-Fong et al. (2018) catalogue the largest available tidal structure search catalogue and an excellent candidate for training CNNs. To aid their visual classifications, Kado-Fong et al. (2018) perform a wavelet decomposition on their images – which is designed to identify high spatial frequency features. For each galaxy, Kado-Fong et al. (2018) then use the five original $grizy$ images (where available) and five images which enhance high spatial frequency components to make a final visual classification: (1) *stream*, (2) *shell*, or (3) *non-detection*. A handful of galaxies carry both the *stream* and *shell* labels. We use these visually classified galaxy images from the HSC-SSP DR1 to train CNNs with the following goals: (a) *identify* stream and

shell features in the HSC-SSP images; (b) *characterize* the detected features as streams or shells; and (c) apply our trained CNNs to the whole of the HSC-SSP Wide layer DR2 (Aihara et al. 2019).

2 DATA & METHODS

2.1 The Hyper Suprime-Cam Subaru Strategic Program

The HSC-SSP is a wide-field optical imaging survey on the 8.2m Subaru Telescope. The salient information about the instrumentation and survey planning can be found in a series of dedicated papers: Miyazaki et al. (2012); Furusawa et al. (2018); Komiyama et al. (2018); Kawanomoto et al. (2018); Miyazaki et al. (2018). We use the co-add images from the Wide layer DR2 (Aihara et al. 2019) which combines 170 nights of observations since January, 2018 and covers 300 deg^2 in all five HSC broadband filters ($grizy$) using the hscPipe software (Bosch et al. 2018). hscPipe performs source detection and sky-subtraction.

2.2 Galaxy image sample

We use a catalogue of 21,208 galaxies that were visually classified for containing streams or shells or as non-detections by Kado-Fong et al. (2018). The sample of galaxies were selected from cross-matching the HSC-SSP Wide internal data release S16A catalogue (which covered $\sim$ 200 deg^2 of the 300 deg^2 presented in the second public data release) to the SDSS DR12 spectroscopic galaxy sample. The redshift limits of the sample are $0.05 < z < 0.45$. Figures 3 and 7 of Kado-Fong et al. (2018) nicely summarize the sample selection. In particular, Figure 3 of Kado-Fong et al. (2018) shows that there are no particularly strong differences between the empirical density functions of redshift and stellar mass for stream/shell hosts and the density distributions for galaxies without stream/shell detections.

Each galaxy in the Kado-Fong et al. (2018) catalogue has a descriptive label for whether it was identified as a stream/shell host or whether it was a non-detection. These labels are used as the target classes for training and validation of our CNNs. There are 1,201 detection labels, of which 987 are streams and 214 are shells (with a handful of galaxies with both labels). The remaining 20,010 targets are non-detections. Random examples of the non-detections and stream/shell hosts in cutouts of the HSC-SSP i-band images are shown in Figure 1 and 2, respectively. Mislabeling will always occur in visual classifications and this dataset is not an exception – for both the non-detection and detection samples. Many of the apparent misclassifications may be driven by changes in (1) the sky-subtraction method and (2) completeness of the mosaics between the original S16A classification images and the PDR2 images we use (Aihara et al. 2019). Nonetheless, the data clearly satisfies the survey objectives. The images are deep with high resolution (5σ point source depth of 26 in the HSC i-band and a median i-band seeing of 0.56 arcseconds) – revealing low surface-brightness structures with great clarity.

2.3 Training images

The size of the raw i-band cutout images (87 arcsec, the same size as was used for the original classifications) is large for the CNNs which have been used for other morphological classification tasks in astronomy (e.g. Huertas-Company et al. 2015; Ackermann et al. 2018; Walmsley et al. 2019; Domínguez Sánchez et al. 2018; Huertas-Company et al. 2019). Therefore, we performed several tests aimed at reducing the size of the images without overly degrading the spatial resolution or cropping away the stream and shell structures. However, given the enormous range of radial separations of stream and shell features from their hosts, we ultimately performed a simple rebinning onto a 128×128 pixel grid. Through visual comparison of the raw and rebinned images, we determined that the loss of resolution did not significantly hamper the visibility of streams, shells, or other low surface-brightness features. In our current tests, we use only the i-band, which has the best depth and resolution of the HSC-SSP observations.

Training with only a single band also affords other advantages in terms of limiting the information that a neural network can use to make classifications. A CNN that is trained on a single band cannot exploit colour information. This colour-insensitivity is particularly relevant for detecting features related to galaxy interactions. Colour correlates strongly with star-formation and mergers between gas-rich galaxies exhibit enhanced star-formation (e.g. Hernquist 1989; Barnes & Hernquist 1991; Ellison et al. 2008; Patton et al. 2013; Blumenthal & Barnes 2018). Therefore, classifications *could* undesirably be made on the basis of colour rather than the presence of stream or shell features. In contrast, a CNN that is single band cannot make such a connection and makes classifications that are, by construction, unbiased by colours of the host.

The image intensities are placed on a logarithmic scale in such a way that the contrast is optimized for the central target in each image following the approach of Bottrell et al. (2019, submitted). First, we take the logarithm of the calibrated, sky-subtracted HSC-SSP image in AB nanomaggies (Oke & Gunn 1983). All pixels which become undefined under this operation are set to -3. We then compute the 99$^{\text{th}}$ percentile inside a (20,20) pixel window centred on the central target galaxy. Everything in the full image that exceeds the 99$^{\text{th}}$ percentile is clipped to this value. We take 0.001 nanomaggies (-3 in our log images) as a lower bound. Everything below this lower bound is clipped to this value. We then shift and subsequently scale the full image to a number between zero and one using these upper and lower bounds. Figures 1 and 2 show that this normalization is optimized to highlight low surface-brightness structures while preserving the bulk galaxy light. We will explore other normalizations in future work.

Given the large imbalance between the detection and non-detection images, we apply augmentations to the images. Eight augmentations of every image are generated which span all possible flip and 90° rotation transformations, shifts the image by up to 10% of its rebinned size, and crops and rebins the image by down to 70% of its original size. These augmentations are simple but do not generate edge artifacts from cutting off other sources that would arise with more rotational freedom. Since we are looking for regions with high spatial frequency, such edge artifacts would be problematic. We oversample the detection images by storing all of these augmentations while only holding a single augmentation per non-detection image. Augmentation brings the class imbalance between positives and negatives to $9,608 : 20,010$. We then downsample the non-detections by selecting the first 10,000 in the Kado-Fong et al. (2018) catalogue. For a given model, we adopt a train, validation, and test split of (70, 15, 15)%. The test images are all originals and augmentations of the test images are discarded. We also clarify (given our unconventional procedure for generating augmentations) that we apply the training/validation/test split such that augmentations of original images within a given subset only appear in that subset. As such, no augmentations of the images in the training data can appear in the validation or test data and vice versa.

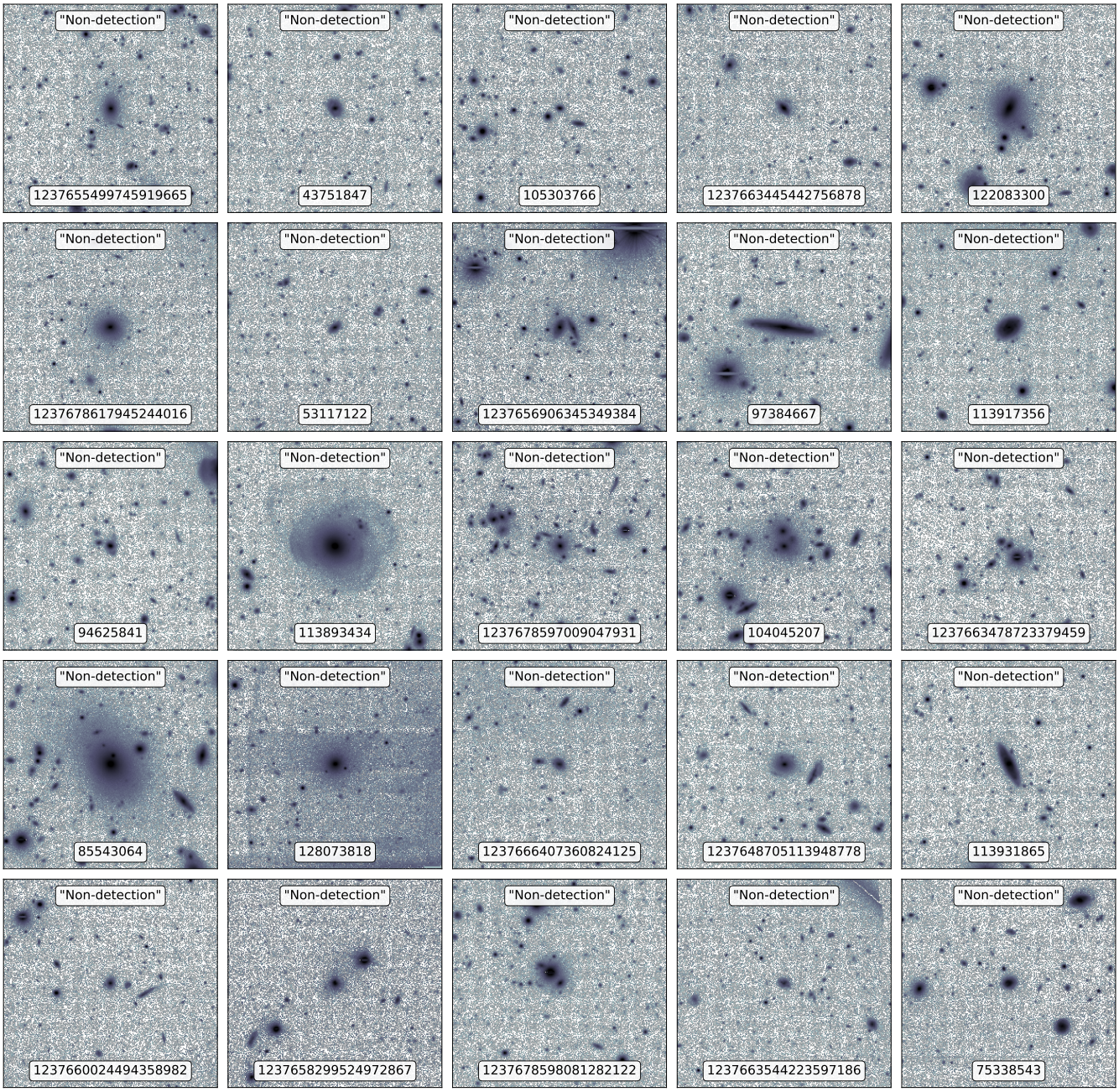

Figure 1. Example "non-detections" from the Kado-Fong et al. (2018) catalogue of stream/shell hosts in the HSC-SSP *i*-band (25 selected randomly). There are 20,010 galaxies classed as non-detections in the Kado-Fong et al. (2018) catalogue in total. The cutouts are 512×512 pixels which correspond to an angular field of view of 87 arcsec. The upper and lower labels are target classes for each object and the cross-matched object ID to the survey providing the spectroscopic confirmation and redshifts, respectively. Given the subjective nature of visual classification (even boosted by an algorithm such as the one used in Kado-Fong et al. 2018), misclassifications are possible. In particular, the galaxy at (column 2, row 3) is a host of *both* stellar stream and shell structures. Indeed, these misclassifications are fairly common in the non-detections sample – likely owing to the large sensitivity of classifications to the choice of contrast and whether the high spatial frequency features were picked up by the Kado-Fong et al. (2018) wavelet decomposition. Nonetheless, these images demonstrate the remarkable image quality of HSC-SSP.

2.4 Convolutional neural networks

Though our goal is to characterize the *type* of tidal feature, we first test whether we can train a model which distinguishes the hosts of streams and shells from galaxies that do not host these tidal features. If successful, we can then move to either: (1) a nested CNN – in which one CNN first classifies targets as detections or non-detections and then a subsequent CNN classifies detections as streams or shells; or (2) a multi-class or multilabel CNN which performs the classifications with a single model.

For the binary task of classifying systems as detections or non-detections, we use a convolutional neural network architecture similar to ones used in other morphological classification analyses

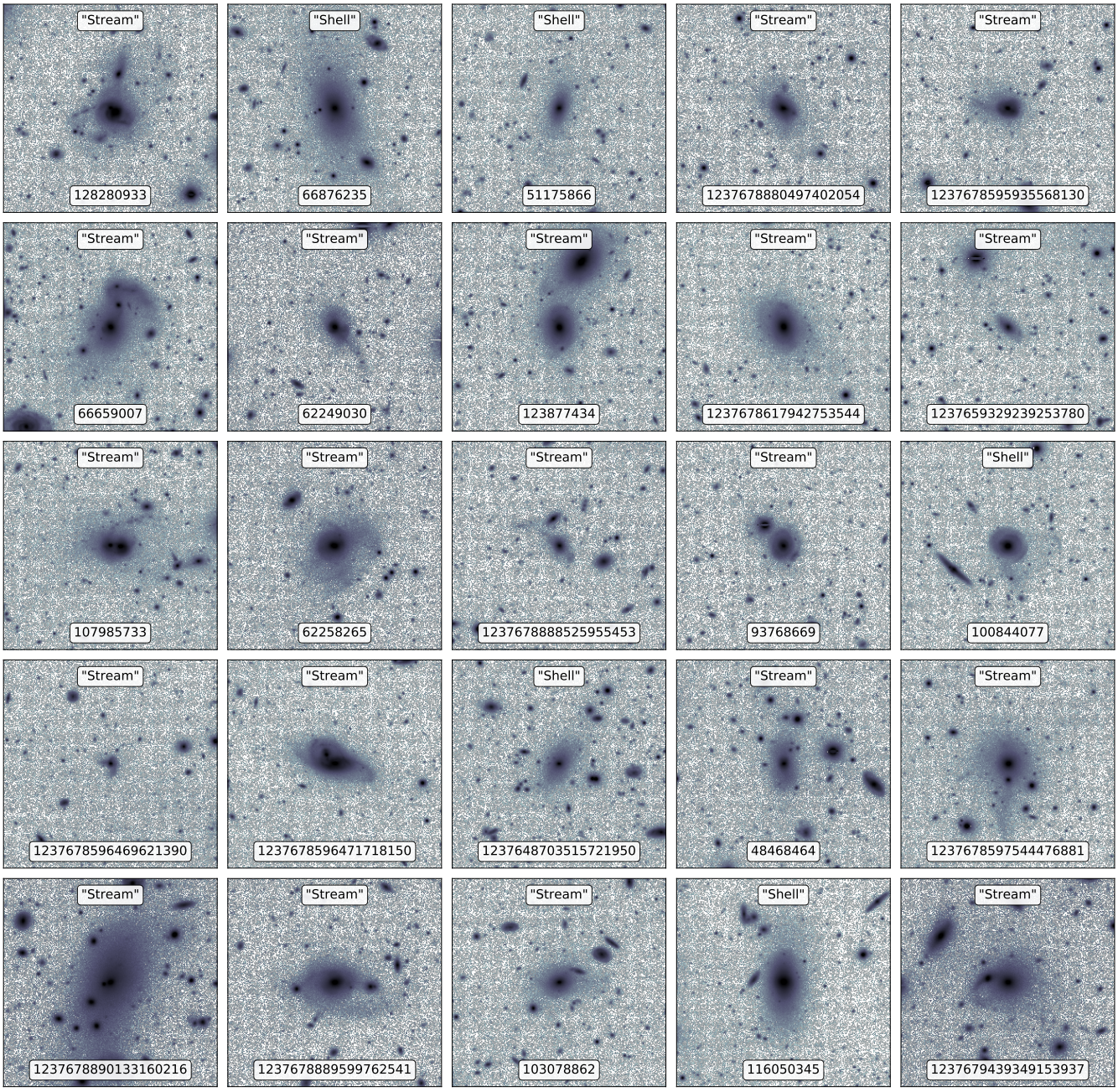

Figure 2. Example "detections" from the Kado-Fong et al. (2018) catalogue of stream/shell hosts in the HSC-SSP i-band (25 selected randomly). There are 1,201 galaxies with a detection label of stream or shell. As with the negatives, there are some targets with less certain classifications. For example, (column 4, row 4) shows no signs of hosting stream or shell features, yet is classed as a stream host. However, given that the stream/shell classes require some criterion to be classified as such, it is expected that the detection sample is more pure than the non-detection sample – though perhaps with low completeness. Combined with Figure 1, these misclassifications in the catalogue represent challenges to training neural networks which will automate the classifications.

(e.g. Huertas-Company et al. 2015; Ackermann et al. 2018; Walmsley et al. 2019; Domínguez Sánchez et al. 2018; Huertas-Company et al. 2019). The model consists of four convolution layers – the first three of which include (2,2) max pooling. These four convolution layers have kernel sizes of (5,5), (3,3), (2,2), (2,2) and each use REctified Linear Unit (RELU) non-linear activation functions. We do not use dropouts in the convolution layers. The output from the fourth convolution layer is flattened to a (32,768) feature array and fed to two fully connected (FC) dense layers and a binary classification layer (sizes: 64; 16; 1). The FC layers use 50% dropout rates and RELU activation functions. The classification layer uses a sigmoid activation function. We use a training batch size of 32 and the ADADELTA optimizer (Zeiler 2012) with a binary cross-entropy loss function. We use the `Keras` API (Chollet et al. 2015) with `tensorflow` as the backend (Abadi et al. 2015) for model creation and training.

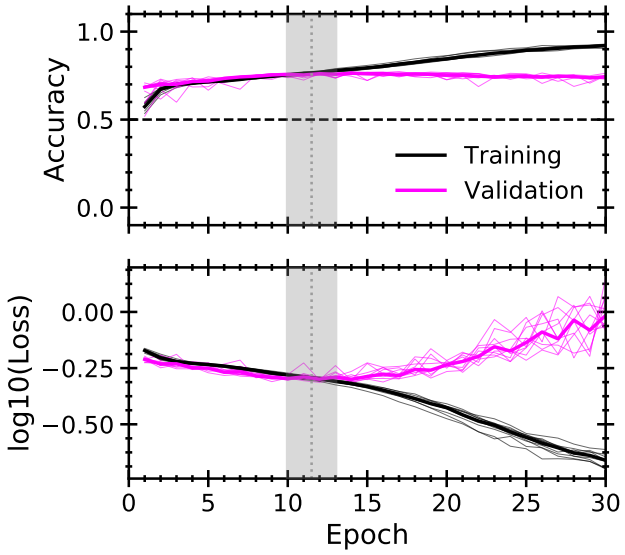

Figure 3. Accuracy and log-loss for CNNs trained distinguishing systems which host streams and/or shells from systems which do not, plotted against training epoch. Thin coloured lines correspond to the training (black) and validation (magenta) accuracy and log-loss for an individual CNN model during training. Each individual model was trained using different random split of training, validation, and test data to guarantee the robustness of the model to the choice of training set – given the sparsity of the detection sample. The thick lines show the median values computed from each individual model. The grey dotted line shows the mean epoch at which the optimal (best-fit) model was saved before overfitting begins to occur. The best-fit model is taken to be the model epoch that gives the maximum accuracy for the validation set. The shaded regions shows the standard deviation in the best-fit epoch. The median maximum validation accuracy is 77%.

3 RESULTS

3.1 Long-term stability of the CNN models

We trained CNNs using the architecture described in the last section to perform the binary classification task of distinguishing systems which host streams and/or shells from systems which do not. The accuracy (the number of correctly classified images as a percentage of the total number of images) and log-loss as a function of training epoch on the training and validation data are shown in Figure 3. Here, we force the model to continue training for 30 epochs even where the optimal model has already been reached and place a checkpoint at the optimal model. We define the optimal (best-fit) model as the model epoch at which the validation accuracy is maximized. To test the stability and variation in the CNNs' performance to the particular training set, we train 10 models using unique random training/validation/test splits of the data. The thin magenta (validation) and black (training) lines show results of each of the 10 models on the corresponding data. The thick lines show the medians of the 10 models. The grey dotted line and shaded region show the mean and standard deviation for the epoch at which the best-fit checkpoint was made and the model was saved.

Figure 3 shows that our model reaches maximum accuracy quite quickly and with little variation based on the training set that was used. The maximum validation accuracy that is reached is quite poor (77%) but is comparable with the validation accuracy for the models from Walmsley et al. (2019) for their much cleaner and smaller sample (∼ 77%, see their Figure 7)[1]. Figure 3 also shows that a degree of overfitting is occurring. Until around epoch 11 or 12, the validation accuracy and loss track the training accuracy and loss. However, instead of plateauing at this point, the validation accuracy begins to decrease slightly while the training accuracy continues to increase to 90% by epoch 30. This result demonstrates that after epoch 11 or 12 the classifiers begin to lose generalizability to data on which the CNNs were not trained. To avoid overfitting, the models are saved at the point where the validation and training scores begin to diverge. Each saved model has a unique test set which corresponds to the unique random seed used to generate the training/validation/test split. We discuss overfitting and possible sources in Sections 4.1 and 4.4.

3.2 Class distribution functions

The overall accuracy does not reveal whether particular class is preferred by the classifier. Figure 4 shows the distributions in $P(X)$, the "probability score" that an image contains streams or shells, for the 10 CNN models (thin lines) and the median over these 10 individual models (thick lines) for images with detection target classes (blue) and non-detections targets (orange). The left, centre, and right panels correspond to the training, validation, and test data, respectively. The trained CNNs behave predictably. The $P(X)$ distribution for images with detection targets is skewed to high $P(X)$ and the images with non-detection targets are skewed to low $P(X)$. Also, while individual CNN models vary (note the differences in the low/high-$P(x)$ tails of the distribution, specially for the positive samples at $P(x) > 0.9$), the median training, validation, and test distributions track each other – showing that there is no overfitting to the training data due to our criterion for selecting the best-fit models. We discuss the issue of overfitting in training epochs beyond the best-fitting model in Sections 4.1 and 4.4.

3.3 Receiver operating characteristics

The receiver operating characteristic (ROC curve) for a binary classification relates the true positive rate or "recall," to the false positive rate or "fall-out," for various thresholds of $P(X)$. The true positive rate is computed as $TPR = TP/(TP + FN)$ where TP is the number of true positives (images which are correctly classified by the CNN as detections) and FN is the number of false negatives (images which are misclassified as non-detections whose target classes are detections). The false positive rate is $FPR = FP/(TN + FN)$ where FP is the number of false positives (images which are misclassified as detections whose target classes are non-detections) and TN is the number of true negatives (the number of non-detections that are correctly classified). The relationship between these two quantities shows the purity and completeness that is achieved by defining a value of $P(X)$ which would separate non-detections and detections.

Figure 5 shows the ROC curves for our CNNs for the training, validation, and test data. The first three panels show the individual ROC curves for each CNN (thin lines) and the corresponding median

[1] Walmsley et al. (2019) used the visual classification sample from Atkinson et al. (2013) for a total of 1,781 luminous galaxies with magnitudes $15.5 < r < 17$ and redshifts $0.04 < z < 0.2$ drawn from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS-Wide, Gwyn 2012). As discussed by Atkinson et al. (2013), their selection is heavily biased towards bright galaxies predominantly in the range $-23 < M_{r'} < -20$ and with half-light radii in a range of 2-6 arcsec.

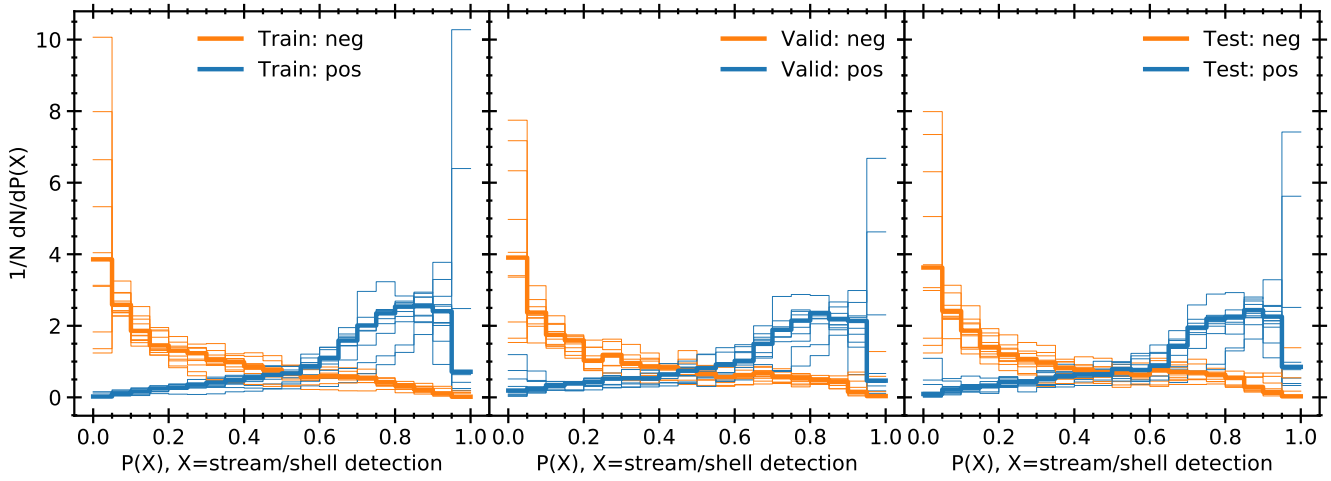

Figure 4. Class distribution functions for the training, validation, and test sets for each of the 10 CNNs trained on different random train/valid/test splits of the data. Thin lines show the distributions of stream and shell detection "probabilities" estimated by the CNN for images with non-detection target classes (orange) and detection target classes (blue). The thick lines are derived from the median in each bin, after which the distribution functions are renormalized to unity. The training, validation, and test distributions track each other very well and $P(X) = 0.5$ seems to be an appropriate threshold for separate detections from non-detections. We do not explore threshold optimization. Threshold optimization may make a difference on the level of a few percent in terms of classification accuracy and F1 statistics.

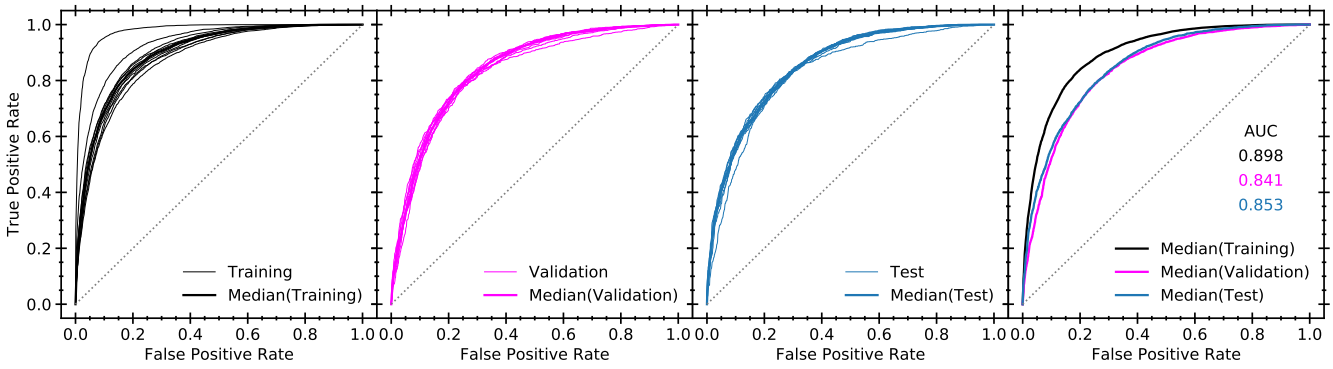

Figure 5. Receiver operating characteristic (ROC) curves for training (first panel), validation (second panel), and test (third panel). Thin lines show the results for each of the 10 individual CNNs and solid lines show the median relation. The fourth panel shows the median ROC curves for the training, validation, and test data all together along with the area under the curve (AUC) characteristics for each dataset. The thin dotted line shows $TPR = FPR$, for which the classifications would be completely random and the AUC=0.5 (a fair two-sided coin flip between detection and non-detection). The validation and test ROC curves track each other very well given the diversity of stream/shell and host morphologies that are possible.

ROC curves. The rightmost panel shows the median ROC curves for each dataset along with the colour-coded area under the curve (AUC) characteristics. The validation and test AUC (for which 1 is a perfect classifier and 0.5 is completely random) track each other very well owing to our early-stopping of the CNNs before they begin overfitting to the training data as shown in Figure 3. Still, Figure 5 further illustrates that there is overfitting to the training data that is not generalized to the validation or test data – resulting in higher AUC and better TPR for any given FPR.

Walmsley et al. (2019) report a TPR of 76% (which they term "completeness") for a corresponding false positive rate of 22% (which they term "contamination") with their ensemble CNN – which combines the predictions of multiple CNNs by averaging the output detection probability, $P(X)$, from each CNN for a given image. For the same FPR we have a slightly poorer test TPR of 75% with our single model. However, as was seen in Walmsley et al. (2019), it might be anticipated that averaging the results from an ensemble of CNNs with different architectures may improve our results. We will explore this in future work.

3.4 Confusion matrix

Figure 6 shows the confusion matrix for the test data for one of the 10 CNNs that we trained to detect stream and shell features in images. A confusion matrix shows the distribution of predicted labels (y-axis) for each true label (x-axis). Columns of this confusion matrix therefore sum to unity. Many of the F_1 diagnostic statistics for a binary classification are represented in the elements of the confusion matrix including the recall (lower right corner), specificity (upper left corner), miss rate (upper right corner), and fall-out (lower left

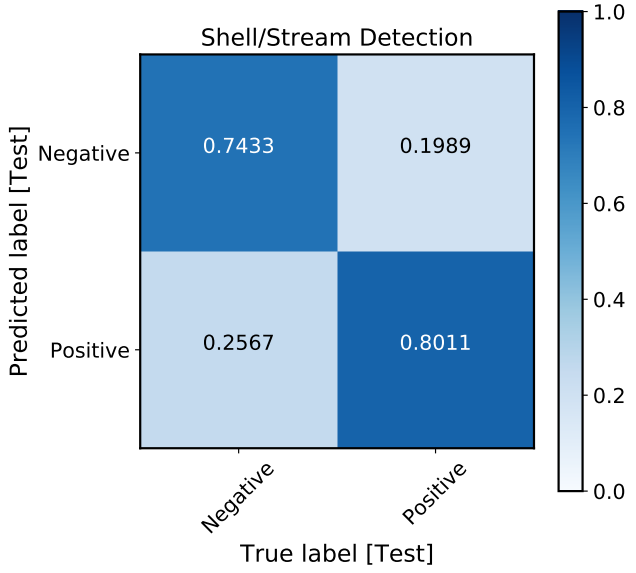

Figure 6. Test data confusion matrix for one of our CNNs, selected randomly from the 10 we generated. True labels are along the x-axis and predicted labels are along the y-axis where "positive" corresponds to detection of a stream or shell feature and "negative" corresponds to a non-detection. The upper left corner is the true negative rate or "specificity", $TNR = TN/(TN+FP)$. The lower left corner is the false positive rate or "fallout". The upper right corner is the false negative rate or "miss rate", $FNR = FN/(FN+TP)$. The lower right corner is the true positive rate. The model achieves a reasonable TPR of $\sim 80\%$ though this success is accompanied by a large amount of contamination as denoted by the FPR of $\sim 26\%$.

corner). The precision, $TP/(FP+FN)$ is 75%. The corresponding F_1-score, the harmonic mean of precision and recall, is 0.77.

The diagnostics from this section and our comparison with other works illustrates that our models are performing reasonably but are hindered by some factor that is preventing the validation and test accuracies and performances from tracking the training accuracies as seen in Figure 3. In other words, there is information in the training data that is allowing the network to continue to learn but is not generalizing to the validation and test data.

4 DISCUSSION

4.1 Overfitting

Figure 3 shows that the training accuracy and loss quickly decouple from the validation accuracy after only 11 to 12 training epochs, on average. This decoupling is a clear indicator of overfitting. Creating a model checkpoint at the epoch where this decoupling occurs prevents the model from overfitting to the training data. However, the fact that the CNN continues to improve in training accuracy indicates that there is information that is being exploited in the training set but is not translating to the validation set. Our model does not have so many trainable parameters ($\sim$ 2 million) that it should approach a dimensionality problem for our dataset (where the number of trainable parameters is of similar order as the number of data points). If the overfitting is indeed not a dimensionality problem, then there is some information that is being learned which is effective in the training set but not the validation and test sets.

It may be that the images are too large and contain too much additional information (neighbouring sources, etc.). However, the challenge of making the images any smaller is two-fold: (1) further cropping would crop away highly extended stream and shell features and remove them from the data; and (2) further degradation of the spatial resolution to 64×64 pixel images (still 87 arcsec) may degrade the tidal features sufficiently that they are no longer distinguishable from the host. After all, the primary motivation for employing the HSC-SSP images for this task was its exceptional depth and resolution for a ground-based large-program optical survey. In any case, the effects of such changes to image field of view or scale will be explored.

4.2 Hyper-parameter search

We performed a crude hyper parameter search over a few alternative CNN specifications to assess the sensitivity of network success to the details of the model and to assess whether our choice of model was well-motivated. In our presently completed search, we focused on the effect of the dropout rate in the FC layers and a different batch size, exploring values in the range (dropouts: 0.0, 0.25, 0.5, and 0.75) and ($N_{\rm batch} = 10$ compared to the 32 we used in the fiducial runs). Comparing the recall values of these CNNs to the recall from Figure 6, all do more poorly by at least Δ(Recall) = 6%. While a larger and more rigorous hyper-parameter search remains to be explored, these results demonstrate that our CNNs at least do not exhibit a poor choice of architecture. Indeed, slight variants of this architecture are proven to be effective in dealing with various galaxy morphological classification tasks (e.g. Domínguez Sánchez et al. 2018; Huertas-Company et al. 2019).

4.3 Does the model favour streams or shells?

In this section, we investigate whether the CNNs preferentially classify or misclassify images with stream or shell target classes. Streams and shells are unique signatures of the orbital scenario and time since a merger event (e.g. Johnston et al. 2008). Consequently, knowing whether a CNN is a better/poorer classifier of either streams *or* shells is important if we want sufficient purity and completeness of both signatures. Since our CNNs are binary classifiers, we cannot estimate the FPR of either streams or shells because we cannot estimate the number of false positives for each type individually, FP(streams) and FP(shells). However, we can compute the TPR (recall) and FNR (miss rate) for the CNNs.

We select one of our models at random (the same for which the confusion matrix is shown in Figure 6). For this model, we split the test data for which the target classes are detections into streams and shells according to the catalogues. We compute $TPR(\text{streams}) = TP(\text{streams})/(TP(\text{streams})+FN(\text{streams})) = 80\%$, where TP(streams) is the number of true positives and FN(streams) is the number of false negatives with "stream" target classifications. Similarly, TPR(shells) = 81%. These results indicate that neither tidal feature is particularly "favoured" by the CNNs. The true positive rate for each morphological signature is the same. The corresponding miss rates for streams and shells are 20% and 19%, respectively.

4.4 Limitations of the target classifications

Figure 1 revealed the possibility of misclassifications in the Kado-Fong et al. (2018) catalogue. Such a misclassification as the one in

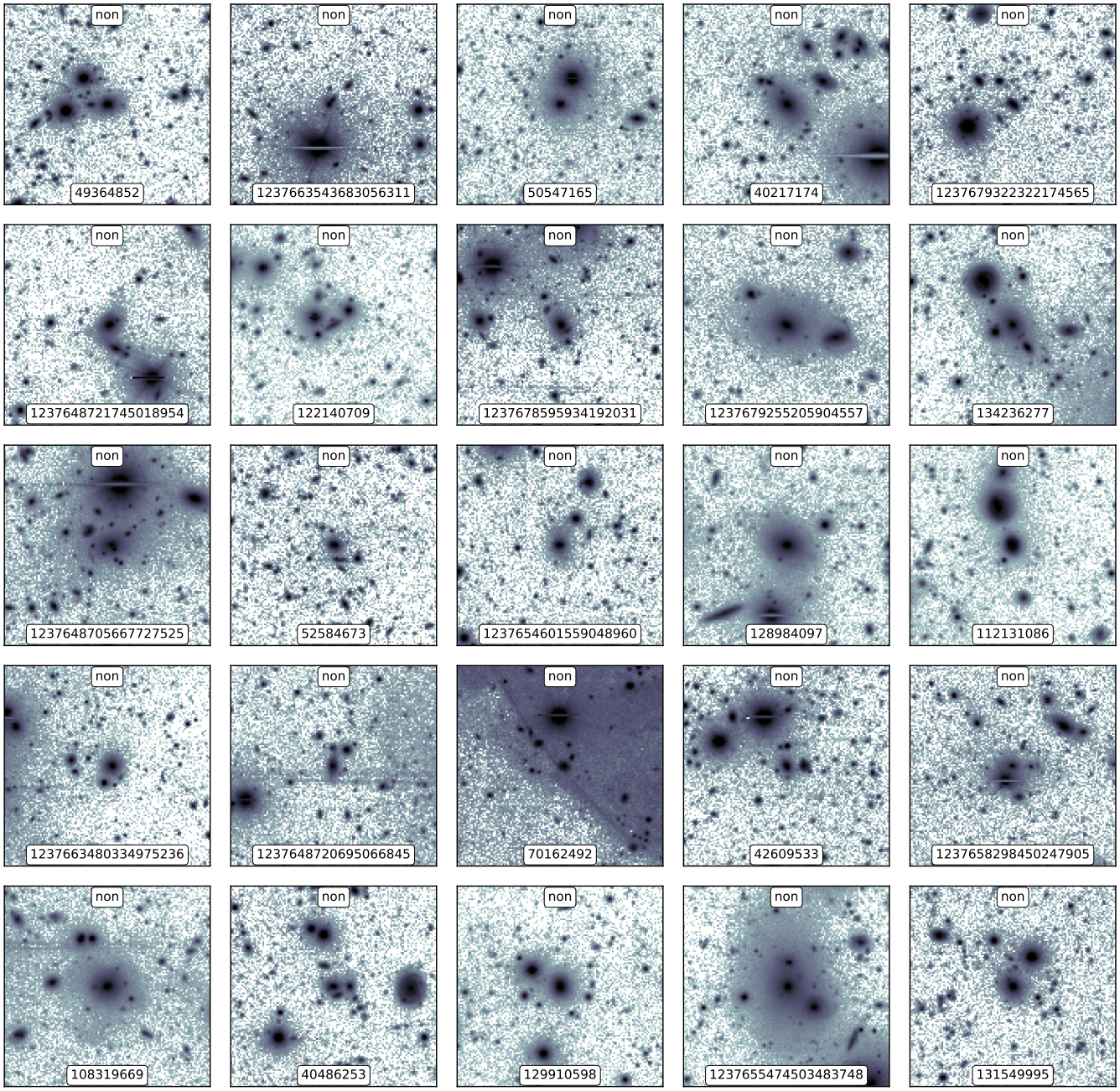

Figure 7. Correctly classified non-detections by the CNN for which the confusion matrix is shown in Figure 6. The images that are classified as non-detections by the network appear to be genuine non-detections despite the possibility of misclassifications in the test data target classes due to the heavy crowding that comes with the HSC-SSP imaging depth. It should be noted that these are the test images *as the CNN sees them* and have consequently degraded to (128,128) pixels for their 87 arcsec field of view.

(column 2, row 3) of the random 25 images selected from the non-detections sample begs the question of how many other targets are similarly misclassified. We explore misclassifications in the target classes by examining the true positives and false positives from our CNNs. Specifically, we focus on the predictions of one of our CNNs chosen at random (the ones for which the results are shown in Figure 6).

First, Figure 7 shows 25 random correctly classified non-detection images (TN). The true negatives shown here all appear to be genuine negatives – with a possible exception in the case of (column 1, row 5) with a shell-like extending from the lower left of the central galaxy. While some galaxies do appear to be in on-going mergers, our goal is not the detection of early-stage ongoing mergers but stream/shell detection for a central host. These are two very different problems. Streams and shells allow us to infer the history/scenario of a past *or* ongoing merger. Merger scenarios will be much more difficult to infer without these features given that the morphologies and extent of these features are tell-tale signatures of the merger scenario.

Figure 8 shows the true positives in the test data for the CNN

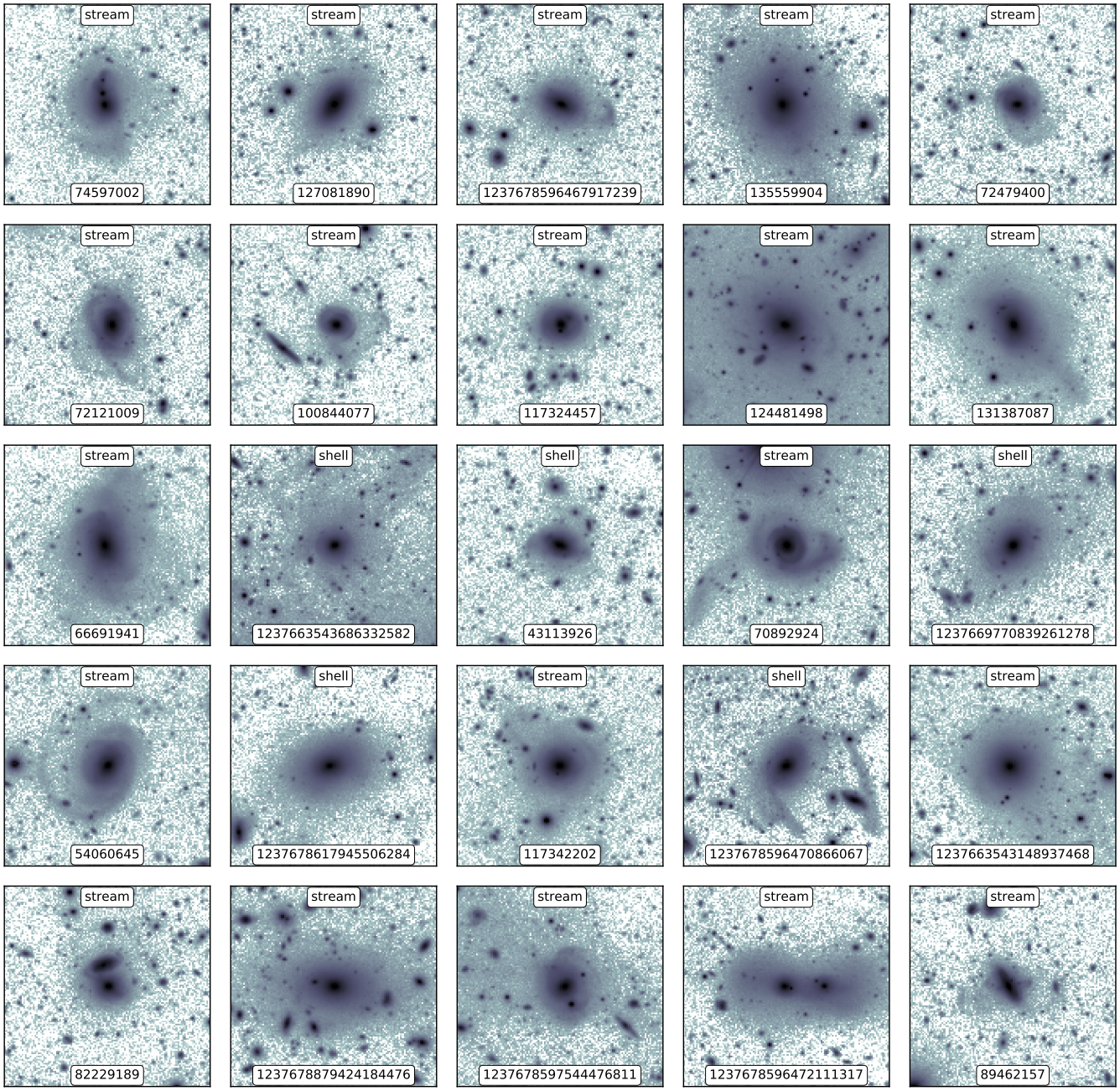

Figure 8. Correctly classified detections by the CNN for which the confusion matrix is shown in Figure 6.

from Figure 6 (correctly classified detection images). These true positive images largely exhibit the desired features. However, there is an exception in (column 2, row 4) where no visible shell can be seen. It is possible that our pre-processing steps have hidden the shell (at least visually). Nevertheless, this galaxy appears to contain no shell in the preprocessed image and yet was classified by the network as a detection. So either the CNN is picking up more subtle features than are visually accessible or the contamination rate is likely to be high. Since Figure 6 showed that the FPR or contamination is already 26% (26% of images whose target classes are non-detection are misclassified as detection), the latter is likely the main culprit for this galaxy being classified as positive despite visually lacking positive features.

In Figure 1 we highlighted a particular galaxy in the non-detection images which exhibits clear stream and shell features. Given the high false positive rate in the data, the prevalence of similar misclassifications in the Kado-Fong et al. (2018) catalogue is a central concern[2]. If a sufficiently large fraction of images with the non-detection target class exhibit stream and shell features, then the CNN's learning will be stifled. One natural consequence is overfitting. During training a CNN will continue to optimize its

[2] In future work, we will contact the authors of Kado-Fong et al. (2018) and collaboratively investigate the factors which may have driven these misclassifications.

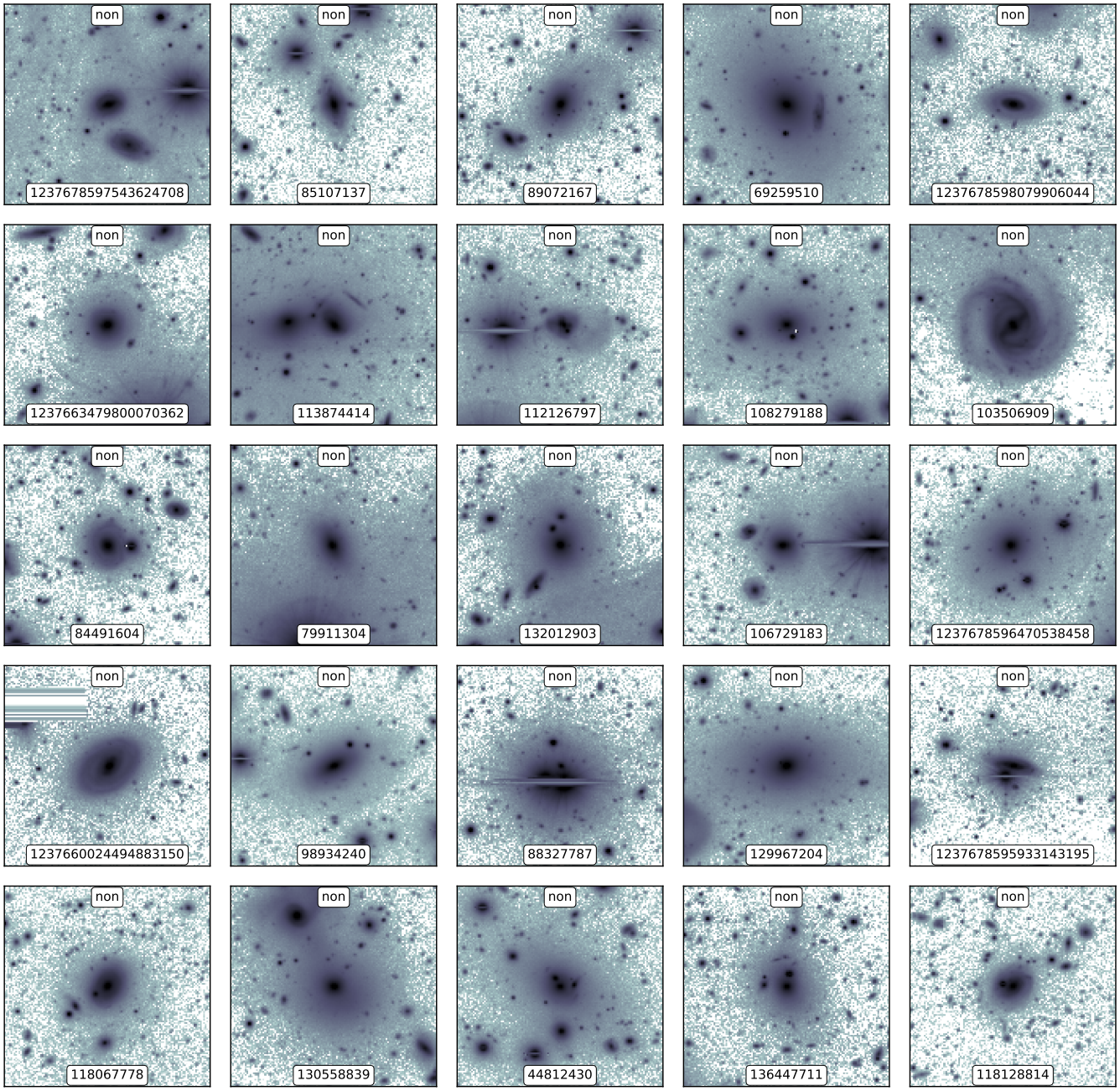

Figure 9. Images with the non-detection target class (shown in the label at the top of each image) which are misclassified as detections by the CNN for which the confusion matrix is shown in Figure 6. While some galaxies in this set simply exhibit spiral arms which may be difficult to distinguish from streams, many visually exhibit genuine stream and shell structures and yet have non-detection target classes.

classifications regardless of whether there are misclassifications in the target data. For example, for the training non-detection and detection images, a CNN may learn to associate specific stream/shell morphologies to each set (even where the non-detection set *should* not contain these morphologies). While this allows the training accuracy to improve, it loses generalizability to new data (the validation and test set). Therefore, if the non-detection dataset contains a large amount of images which host stream and shell features, then we have a possible explanation for the rapid and significantly decoupling of the long-term training and validation accuracies and losses in Figure 3.

Figure 9 shows 25 random false positives in the test data for the CNN from Figure 6 (non-detection images misclassified as detections). There is a large number of galaxies in this sample of 25 random images with the non-detection target class that visually exhibit stream and shell features (even in these degraded (128,128) pixel images). On one level, this set of images provides validation that the network is behaving in the desired way. Despite the fact that these images have non-detection target classes, they are being classified as detections by the CNN. The catch is that with random splitting of the training, validation, and test data, these same images may appear in the training set of a CNN. Owing to the similar per-

formances of 10 CNNs trained on unique random selections of the training data shown in Figures 3, 4, 5, it is unlikely that the problem is unique to this particular split. Following the arguments in the last paragraph, having target misclassifications such as these presents at least two crucial problems: (1) overfitting to the training data and consequently stifling/decreasing validation and test accuracy and (2) erroneous F_1 statistics and consequently a poor idea for the completeness or purity of the detections for a given CNN. Owing to these factors, target misclassifications may represent the most significant barrier to improving our classification performance.

However, we emphasize here that Kado-Fong et al. (2018) combined visual inspection with the detections from their filtering method. This combination provides strict criteria for an image to be included in the detection category. Consequently, their sample of galaxies which were classified as exhibiting streams/shells is probably very pure but may suffer from incompleteness – as highlighted by the identification of galaxies with streams and shells as false positives by our CNNs. The incompleteness may originate from the sky-subtraction methods in the HSC-SSP PDR1 which have been updated significantly in PDR2 (Aihara et al. 2019) and are known to affect detection of low surface-brightness features. It may also arise from variation in angular resolution across the HSC-SSP. We will explore the impact of these effects in future work. We emphasize that Kado-Fong et al. (2018) catalogue of streams and shell hosts still offers remarkable scientific opportunities for studying the properties of stream and shell hosts due to the high purity of the detection sample. However, studies which will focus on the enhancement or suppression of physical phenomena or properties of stream/shell hosts *relative* to matched control galaxies which do not exhibit such tidal features must be aware of the contamination in the non-detection sample.

4.5 Future work

Improving the performance of a CNN for this task may require more accurate targets in the non-detections sample as highlighted in Section 4.4. While the Kado-Fong et al. (2018) catalogue of stream and shell hosts and non-detections represents enormous progress towards obtaining a sufficiently large dataset that an automated classification model can be trained, the misclassifications of images shown in Figures 1 and 9, in particular, may hamper the CNNs' performances and limit the meaningfulness of the F_1 statistics we present. Therefore, a re-classification of the images (at least the non-detections from the catalogue) may be necessary. To ensure that the CNN is trained using data with the same standards as the human classifier, the same normalization should be used (optimizing the contrast for the central target as discussed in Section 2.3) but allowing the user to adjust scale and contrast as desired. The method used by Kado-Fong et al. (2018) for enhancing high spatial frequency features can be used to compliment these classifications but should not be the primary basis for the classifications.

In this work we trained 10 CNNs using different random splits of the training, validation and test data to examine the robustness of CNNs' performances to the specific split used. After a more extensive grid search over alternate architectures and hyper-parameters, an ensemble of CNNs trained on *the same* training data can be used to boost performance (e.g. Walmsley et al. 2019). With a model that is able to effectively discriminate between detections and non-detections (and between streams and shells, amongst detections), we can perform automated classifications over the whole of the HSC-SSP imaging campaign to obtain unprecedented sample statistics on streams and shells and their hosts. Another item that remain to be explored is the redshift and magnitude dependence of the classification accuracies and whether streams or shells in particular exhibit stronger dependence on surface brightness and resolution limits. Additionally, a greater exploration of CNN architectures is warranted. In this work, we used only the i-band imaging. An examination of the improvements to our CNNs' performances by including other bands (allowing for colour sensitivity) is of interest. These tests will provide insight into the origin of the limited performance of our classifiers.

Furthermore, one particular advantage of the algorithm developed in Kado-Fong et al. (2018) is that it identifies the *specific pixels* which correspond to a stream or shell feature. We are therefore also planning a parallel project which will employ the recent Morpheus tool (Hausen & Robertson 2019) to perform automatic segmentation and classification of stream and shell features on a pixel-by-pixel level. The pixel-level identification of tidal features will open entirely new opportunities in the characterization of streams and shells along with their host galaxies.

5 CONCLUSIONS

In this work, we trained CNNs to detect stream and shell features in images from the HSC-SSP Data Release 1 using the Kado-Fong et al. (2018) stream and shell search catalogue. The search catalogue covers an unprecedentedly large sample of 21,208 galaxies from HSC-SSP with confirmation spectra and redshifts from the SDSS DR12 of which 1,201 are stream or shell hosts and 20,010 are non-detections. In this pilot study, we focus primarily on detection of stream/shell features rather than distinguishing between them. We evaluate the CNNs' performances in a binary classification of (0) non-detection and (1) detection of stream and shell features. Our results are as follows:

- **The training data in for CNN models are quickly overfitted** (Figures 3, 4, 5). Our characterization of the CNNs' performances in the training, validation, and test data showed that the CNNs quickly and substantially overfit the training data – yielding relatively poorer accuracies in the validation and test data. Our best-fit CNNs achieve accuracies of 77% in both the validation and test data.
- **Non-detection images from the** Kado-Fong et al. (2018) **catalogue contain galaxies which exhibit strong stream and shell features** (Figure 1, column 2, row 3 and most panels from Figure 9). Our cursory characterization of the data used to train the CNN showed that many galaxies from non-detection set are misclassifications.
- **One of our CNNs (selected at random) has a** TPR **or recall of** 80% **and an** FPR **or contamination of** 26% (Section 3.4 and Figure 6). We compare the performance of this CNN with the work of Walmsley et al. (2019) and show that we achieve very similar accuracies. This is particularly notable given the much greater sample size and redshift range in the Kado-Fong et al. (2018) stream/shell search catalogue compared to the (Atkinson et al. 2013) sample used by Walmsley et al. (2019).
- **We argue that misclassifications in the** Kado-Fong et al. (2018) **catalogue (particularly for the non-detection set) may limit improvements to the models' performances** (Section 4.4 and Figure 9). The misclassifications in the targets for each image invite overfitting to the training data and loss of generalizability to the validation and test data. Consequently, the statistics which represent purity, contamination, miss rates, etc. are not likely representative of the true statistics. However, we have yet to explore redshift, magnitude, and angular resolution dependence on the CNNs'

performances. So misclassifications in the targets may not be the only (or dominant) factor which is limiting CNN performance.

Looking forward, an automated method for detecting and distinguishing between stream and shell features in large ground-based imaging programs such as the HSC-SSP is necessary if we want to infer the scenarios of the interactions which produced these features. The Kado-Fong et al. (2018) visual classification sample allows the training/calibration of more automated methods which can then be used to automatically classify other galaxies in the full HSC-SSP imaging campaign. However, we have shown that more accurate classifications of the training data are probably required for this goal to be realized.

ACKNOWLEDGEMENTS

We all acknowledge and thank the Kavli Foundation for making this collaboration possible and Pascale Garaud for organizing this year's Kavli Summer Program in Astrophysics at the University of California Santa Cruz.

This paper uses data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center at National Astronomical Observatory of Japan. The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

REFERENCES

Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, `http://tensorflow.org/`
Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415
Aihara H., et al., 2018a, PASJ, 70, S4
Aihara H., et al., 2018b, PASJ, 70, S8
Aihara H., et al., 2019, arXiv e-prints, p. arXiv:1905.12221
Alam S., et al., 2015, ApJS, 219, 12
Atkinson A. M., Abraham R. G., Ferguson A. M. N., 2013, ApJ, 765, 28
Baes M., Verstappen J., De Looze I., Fritz J., Saftly W., Vidal Pérez E., Stalevski M., Valcke S., 2011, ApJS, 196, 22
Barnes J. E., Hernquist L. E., 1991, ApJ, 370, L65
Blumenthal K. A., Barnes J. E., 2018, MNRAS, 479, 3952
Bosch J., et al., 2018, PASJ, 70, S5
Bottrell C., Torrey P., Simard L., Ellison S. L., 2017, MNRAS, 467, 1033
Camps P., Baes M., 2015, Astronomy and Computing, 9, 20
Chollet F., et al., 2015, Keras, `https://keras.io`
Cox T. J., Jonsson P., Somerville R. S., Primack J. R., Dekel A., 2008, MNRAS, 384, 386
Domínguez Sánchez H., Huertas-Company M., Bernardi M., Tuccillo D., Fischer J. L., 2018, MNRAS, 476, 3661
Ellison S. L., Patton D. R., Simard L., McConnachie A. W., 2008, AJ, 135, 1877
Ellison S. L., Viswanathan A., Patton D. R., Bottrell C., McConnachie A. W., Gwyn S., Cuillandre J.-C., 2019, MNRAS, 487, 2491
Furusawa H., et al., 2018, PASJ, 70, S3
Gwyn S. D. J., 2012, AJ, 143, 38
Hani M. H., Sparre M., Ellison S. L., Torrey P., Vogelsberger M., 2018, MNRAS, 475, 1160
Hausen R., Robertson B., 2019, arXiv e-prints, p. arXiv:1906.11248
Hayward C. C., Hopkins P. F., 2017, MNRAS, 465, 1682
Helmi A., White S. D. M., 1999, MNRAS, 307, 495
Hendel D., Johnston K. V., 2015, MNRAS, 454, 2472
Hernquist L., 1989, Nature, 340, 687
Hernquist L., Quinn P. J., 1989, ApJ, 342, 1
Hood C. E., Kannappan S. J., Stark D. V., Dell'Antonio I. P., Moffett A. J., Eckert K. D., Norris M. A., Hendel D., 2018, ApJ, 857, 144
Huertas-Company M., et al., 2015, ApJS, 221, 8
Huertas-Company M., et al., 2019, arXiv e-prints, p. arXiv:1903.07625
Johnston K. V., Bullock J. S., Sharma S., Font A., Robertson B. E., Leitner S. N., 2008, ApJ, 689, 936
Kado-Fong E., et al., 2018, ApJ, 866, 103
Kawanomoto S., et al., 2018, PASJ, 70, 66
Keel W. C., Kennicutt R. C. J., Hummel E., van der Hulst J. M., 1985, AJ, 90, 708
Komiyama Y., et al., 2018, PASJ, 70, S2
Koss M., Mushotzky R., Veilleux S., Winter L., 2010, ApJ, 716, L125
Lacey C., Cole S., 1993, MNRAS, 262, 627
Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2008, MNRAS, 391, 1137
Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2010, MNRAS, 404, 575
Malin D. F., Carter D., 1983, ApJ, 274, 534
Miyazaki S., et al., 2012, in Proc. SPIE. p. 84460Z, doi:10.1117/12.926844
Miyazaki S., et al., 2018, PASJ, 70, S1
Morales G., Martínez-Delgado D., Grebel E. K., Cooper A. P., Javanmardi B., Miskolczi A., 2018, A&A, 614, A143
Moreno J., et al., 2019, MNRAS, 485, 1320
Nair P. B., Abraham R. G., 2010, ApJS, 186, 427
Oke J. B., Gunn J. E., 1983, ApJ, 266, 713
Patton D. R., Torrey P., Ellison S. L., Mendel J. T., Scudder J. M., 2013, MNRAS, 433, L59
Satyapal S., Ellison S. L., McAlpine W., Hickox R. C., Patton D. R., Mendel J. T., 2014, MNRAS, 441, 1297
Tal T., van Dokkum P. G., Nelan J., Bezanson R., 2009, AJ, 138, 1417
Toomre A., Toomre J., 1972, ApJ, 178, 623
Walmsley M., Ferguson A. M. N., Mann R. G., Lintott C. J., 2019, MNRAS, 483, 2968
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
Zeiler M. D., 2012, arXiv e-prints, p. arXiv:1212.5701

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