

Explaining deep learning of galaxy morphology with saliency mapping

Prabh Bhambra,^{1*} Benjamin Joachimi,¹ Ofer Lahav¹

¹*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We successfully demonstrate the use of explainable artificial intelligence (XAI) techniques on astronomical datasets in the context of measuring galactic bar lengths. The method consists of training convolutional neural networks on human classified data from Galaxy Zoo in order to predict general galaxy morphologies, and then using SmoothGrad (a saliency mapping technique) to extract the bar for measurement by a bespoke algorithm. We contrast this to another method of using a convolutional neural network to directly predict galaxy bar lengths. These methods achieved correlation coefficients of 0.76 and 0.59, and root mean squared errors of 1.69 and 2.10 respective to human measurements. We conclude that XAI methods outperform conventional deep learning in this case, which could be reasonably explained by the larger datasets available when training the models. We suggest that our XAI method can be used to extract other galactic features (such as the bulge-to-disk ratio) without needing to collect new datasets or train new models. We also suggest that these techniques can be used to refine deep learning models as well as identify and eliminate bias within training datasets.

Key words: methods: data analysis – techniques: image processing – galaxies: general – galaxies: bar

1 INTRODUCTION

From autonomous vehicles to targeted advertising, machine learning (ML) algorithms have become more and more ubiquitous in society and a greater emphasis is being put on understanding the decisions they make. We call an algorithm interpretable if it can be understood by humans on its own without much analysis. Algorithms themselves vary in their interpretability, with the likes of decision trees tending towards being more interpretable, while neural networks tend towards being some of the least interpretable (Došilović et al. 2018). Despite their opacity, neural networks are widely used due to their relatively good performance compared to other ML algorithms. The field of explainable artificial intelligence (XAI) seeks to help explain the behaviour of difficult to interpret algorithms (Adadi & Berrada 2018).

ML has been widely adopted in the field of astronomy, although to date relatively little thought has been given to the field of XAI. This paper aims to use XAI methods to generalise ML models trained on astronomical datasets. Due to astronomy's reliance on telescopic imaging, we considered datasets primarily consisting of images. Due to the richness of the data, the Galaxy Zoo project was chosen (Willett et al. 2013b). To our knowledge this is the first attempt at quantitatively using saliency mapping for image processing in astronomy, although there are recent cases of them being used qualitatively for spectral analysis (Peruzzi et al. 2021) and image processing (Villanueva-Domingo & Villaescusa-Navarro 2021). This paper intends to act as a proof of concept that these techniques are effective when applied to this field, and that they can be used to infer quantitative measurements.

Galaxy Zoo is a citizen science project in which volunteer participants are shown images of galaxies, and then asked a series of ques-

tions regarding their morphologies. Human classification of galaxy morphologies is extremely time consuming and requires expert level domain knowledge. However the use of citizen science platforms helps to reduce the reliance on these limited resources. Showing the same image to multiple participants can be used to reach a crowd-sourced consensus of the galaxy morphology without requiring extreme time commitment from any single participant (Lintott et al. 2008).

There have been early applications of ML methods to aid in the classifications of galaxies (e.g. Lahav et al. 1995, 1996; Banerji et al. 2010). Recently, the Galaxy Zoo project has provided a rich data source for the training of these ML models. Many implementations can be seen on the Kaggle competition “Galaxy Zoo - The Galaxy Challenge” in which participants were tasked to train an ML model on a subset of Galaxy Zoo data in order to predict galaxy morphologies. Dieleman et al. (2015) provided the winning submission to the contest by using convolutional neural networks (CNNs), a model architecture which is widely used in image processing due to their excellent performance (Aloysius & Geetha 2017). More recently, Domínguez Sánchez et al. (2019) used Galaxy Zoo data to train ML models which were then used to classify galaxies from the Dark Energy Survey. Walmsley et al. (2020) used Bayesian CNNs, which can more effectively learn from examples with uncertain labels, to predict posteriors for the morphologies of galaxies.

XAI methods can be used to produce saliency maps to show which areas in an image are most important to a CNN's classification of an image. This paper aims to use SmoothGrad, an XAI method developed to give visual explanations of CNNs, to produce saliency maps explaining which parts of an image are important in the image's classification as a barred galaxy. From these saliency maps the length of the bars can then be measured and compared to other methods of measurement. Galactic bars are of particular interest

* E-mail: prabh.bhambra.12@ucl.ac.uk

due to their ability to contribute to the redistribution of matter in a galaxy through transferring angular momentum between the disc, bulge, and any spiral arms or rings present (Kormendy & Kennicutt 2004; Athanassoula et al. 2009). Bar lengths are an especially good application of XAI due to the simplicity of their geometric structure. Galaxies are seen in 2D projection, and the highly linear structure of bars makes it straightforward to quantify the output of XAI methods and compare this quantification to other methods of measurement. This comparison can be used to demonstrate the predictive power of CNN classifications. We aim to show that XAI methods can be used to extract new information, which has never been presented to the CNN, from its classifications.

This paper is structured as follows: Section 2 describes the data used throughout this paper, while Sections 3 and 4 describe the XAI methods used and the modelling process respectively. Section 5 describes the results of comparing our methods to human measurements. Finally, Section 6 contains our conclusion.

2 DATA

The data used to train the models was taken from Kaggle’s “Galaxy Zoo - The Galaxy Challenge” dataset¹. This dataset consists of galaxy images each with dimensions $424 \times 424 \times 3$ (i.e. a height and width of 424 pixels, and 3 (red, green, and blue) colour channels). The images were partitioned into a training set and a test set. The training set consisted of 61,578 images with associated galaxy classification probability distributions. The probability distributions represent the proportion of Galaxy Zoo volunteers who associated the galaxy into each of the 37 classes. The test set consisted of 79,975 galaxy images without labels (as the testing of the models is handled solely by Kaggle’s submission system). The galaxy images are composites of images taken by the Sloan Digital Sky Survey (SDSS) camera through its u, g, r, i, and z filters (Gunn et al. 1998).

The Hoyle bar length catalogue (Hoyle et al. 2011) was collected as an extension to the Galaxy Zoo project. Galaxy Zoo participants were asked to measure the length of the bars of 3,150 barred galaxies by drawing a line on the images using a web interface. With multiple participants measuring the bars on each galaxy, participants were able to reproduce each other’s measurements with a scatter of 17 per cent (Hoyle et al. 2011). The bar length measurements in the Hoyle catalogue as well as the Galaxy Zoo images were scaled by the `petror90_r` value (the radius containing 90 per cent of the r-band Petrosian aperture flux) for the respective galaxy (Willett et al. 2013b; Hoyle et al. 2011), hence this data was required to undo this scaling. These values were taken from the Galaxy Zoo metadata file hosted on the project’s data page². The use of the bar length catalogue also required a full set of Galaxy Zoo 2 images (Willett et al. 2013a).

The Hoyle bar length catalogue also includes an error that introduces a proportional offset. The values in the catalogue are quoted in units of h^{-1} kpc, but are actually in units of kpc with h assumed to be equal to 0.7. Therefore, in order to use the catalogue in its intended units of h^{-1} kpc, we multiplied through by $h = 0.7$ (Walmsley & Geron, personal communication, 15 November 2021).

3 SALIENCY MAPPING

Saliency mapping is a term used to refer to a family of techniques seeking to explain the behaviour of CNNs. Saliency maps are heatmaps used to demonstrate the relative importance of parts of an image to the classification of that image into a certain class by the CNN in question (Smilkov et al. 2017). These heatmaps are often computed by taking gradients from the output layer of the CNN to various other layers, in this way the resulting saliency map can be seen as visualising the sensitivity of an output to a given input (Simonyan et al. 2013).

Various methods were considered for use. Grad-CAM (Selvaraju et al. 2016) and Grad-CAM++ (Chattopadhyay et al. 2017) take gradients from the output of the CNN to the penultimate convolution layer in order to assign importance to neurons within the CNN. This approach produces saliency maps which are able to locate features well but are unable to pull out finer detail within the image. Conversely, Guided Backpropagation (Springenberg et al. 2014) manages to highlight features well, but is unable to strongly isolate important areas of the image. We ultimately decided to use SmoothGrad (Smilkov et al. 2017), which provides explanations up to the pixel space of the input image. This allows SmoothGrad to pick out the finer detail in the image when explaining the behaviour of the CNN.

SmoothGrad builds on top of the idea of traditional saliency mapping (Simonyan et al. 2013), in which the importance (L) of a pixel’s intensity (x) to the classification of class (c) is given by the gradient of the score of the relevant class (y) with respect to the pixel intensity:

$$L^c(x) = \frac{\partial y^c}{\partial x}. \quad (1)$$

L^c is calculated for each pixel within the image and then a two-dimensional saliency map is constructed by stacking the individual values of L^c .

However, due to modern models combining many linear and non-linear (e.g. Rectified Linear Unit) operations, these gradients are locally unstable. Small fluctuations in the value of a pixel create visually similar images, however these can result in wildly different saliency maps (Smilkov et al. 2017). SmoothGrad aims to address this issue by creating N visually similar samples of the same image by adding Gaussian noise to each, and then averaging the saliency maps for each sample:

$$\hat{L}^c(x) = \frac{1}{N} \sum_{i=1}^N L^c[x + \mathcal{N}(0, \sigma^2)], \quad (2)$$

where $\mathcal{N}(0, \sigma^2)$ is the probability density function for a Gaussian distribution with a mean of 0 and a standard deviation of σ^2 . Both N and σ^2 are free parameters in the SmoothGrad algorithm.

Examples of SmoothGrad saliency maps can be found in Figure 1. The galaxy image was used to generate saliency maps explaining why models classified the galaxy as a barred galaxy with a prominent bulge and spiral arms. The saliency maps were compiled by averaging the individual saliency maps generated from each of the models described in Section 4.2, using SmoothGrad parameters of $N = 256$ and $S_{\text{noise}} = 0.1$ (where S_{noise} is a transformation of σ^2 , see Section 4.5). This example demonstrates SmoothGrad’s ability to pick out features of the galaxy when explaining the CNN’s decisions to classify the image into various classes. It can be seen that SmoothGrad is successfully able to highlight the bar, bulge, and spiral arms in each successive saliency map.

¹ <https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/>

² <https://data.galaxyzoo.org/>

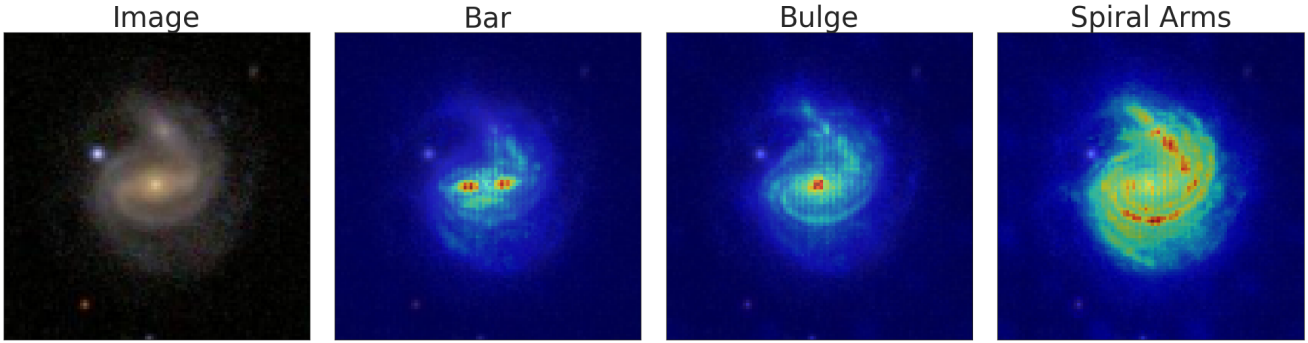


Figure 1. Example SmoothGrad saliency maps explaining the classification of a galaxy image (left) as a barred (centre-left) galaxy with a prominent bulge (centre-right) and spiral arms (right). Red pixels represent pixels with the highest importance \hat{L}^c , while blue represents those with the least importance. The saliency maps were generated using the models described in Section 4.2. All images were cropped to the centre 100×100 pixels for this figure. The galaxy shown is identified by the SDSS Objid of 587722983363248237, which was chosen due to its clear demonstration of multiple features. Similar saliency maps for a random selection of galaxies can be found in Appendix A.

4 MACHINE LEARNING MODEL

4.1 Data augmentation

Initial attempts at training models directly on unaugmented images resulted in poor convergence, therefore image augmentation was used. The images were down-sampled to dimensions of $224 \times 224 \times 3$ in order to reduce the impact of noise, and were also subjected to random augmentations during each epoch in order to reduce overfitting:

- (i) random rotations in the range $0^\circ - 360^\circ$,
- (ii) random horizontal and vertical flips,
- (iii) random horizontal and vertical shifts in the range of 0-5 per cent of the image dimensions,
- (iv) random brightness scaling in the range of 70-130 per cent.

4.2 Model architecture

In total, three different models were trained to predict galaxy morphologies, each based on a common image classification model (the VGG16 (Simonyan & Zisserman 2014), ResNet50v2 (He et al. 2016), and Xception (Chollet 2016) architectures). Each model was created with a number of input nodes to match the size of the down-sampled images, and 37 (i.e. the number of Galaxy Zoo classes) nodes in the output layer. All models were implemented with the “imagenet” pre-trained weights from the Keras API for TensorFlow (Abadi et al. 2016).

4.3 Training

The models were trained with the augmented images as the input and the Galaxy Zoo class probability distributions as the targeted output. 5 per cent of the training dataset was used for validation to monitor overfitting.

The models were trained in order to minimise the root mean squared error (RMSE) between the model’s predictions and the Galaxy Zoo class probability distributions of galaxy morphologies. This was done by using an Adam optimiser (Kingma & Ba 2014). The learning rate was also set to reduce by a factor of 0.5 if the loss on the validation dataset did not decrease over the course of five epochs.

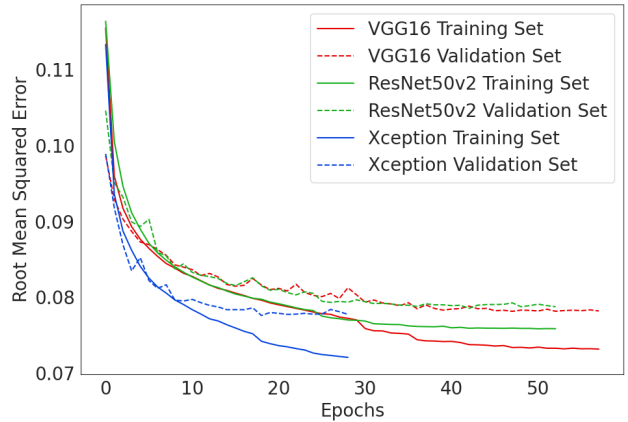


Figure 2. The learning curves for the three CNNs architectures trained. The losses for the VGG16 (red), ResNet50v2 (green), and Xception (blue) based models are shown for both the training (solid) and validation (dashed) sets.

The models were set to be trained for up to 100 epochs; however training was halted in the case that the RMSE of the validation set did not decrease over the course of 10 epochs.

The learning curves for the models can be found in Figure 2. The final metrics for each model can be found in Table 1 which shows the number of epochs each model trained for before stopping, as well as the RMSE on the training, validation, and test sets. The models were also ensembled by averaging the predictions of each individual model. The evaluation of the models on the test set was handled by Kaggle’s submission system. Table 1 shows little variation in the performance of the models, which demonstrates the robustness of our training methods to model architecture choices. These models performed slightly worse than those trained by Dieleman et al. (2015) due to not using as intensive data augmentation practices.

Table 1. RMSE on the training, validation, and test sets for each model architecture, including an ensemble model created by averaging the predictions of the other three models.

Model	Epochs	Train RMSE	Val RMSE	Test RMSE
VGG16	58	0.07315	0.07818	0.07939
ResNet50v2	53	0.07584	0.07874	0.08110
Xception	29	0.07206	0.07774	0.07933
Ensemble	N/A	N/A	N/A	0.07747

4.4 Bar length measurement

We wanted to find a way to quantify how well SmoothGrad was able to explain the behaviour of the CNNs and how much information could be extracted from the CNN’s classifications. Attempting to measure bar lengths from saliency maps was decided upon due to the availability of human measured bar lengths which could act as “ground truth” measurements. SmoothGrad was used to generate saliency maps highlighting the pixels within each galaxy in the Hoyle Bar Length catalogue that had the most importance (i.e. highest \hat{L}^c) in classifying the images as barred galaxies by the models. SmoothGrad has two hyperparameters (Smilkov et al. 2017):

- (i) N - the number of samples to consider,
- (ii) σ^2 - the standard distribution of the Gaussian noise to be added to each sample.

N was set to 256. A greater number of samples would lead to less noisy saliency maps, but this would increase the required computing time significantly. The optimal value of σ^2 is dependent on the dataset being considered, therefore this was considered as a hyperparameter to be optimised.

In addition, elementwise multiplication of the saliency map with the input image is considered. This hyperparameter will be denoted as X_{input} and acts as a boolean flag. As we expect the bar of the galaxy to be brighter than the average brightness of the image, this should bias the saliency maps towards highlighting features of the galaxy rather than mistakenly highlighting any other artefacts or parts of the background.

Algorithm 1 was used to measure the bar length of a galaxy from its respective saliency map. The algorithm isolates the bar of the galaxy by zeroing out pixels in the saliency map below a specific threshold. This threshold is incremented until either the non-zero pixels are linearly distributed over the image with a Pearson correlation coefficient greater than a user-specified value, or the number of non-zero pixels is less than a user-specified required number of pixels. This algorithm requires two hyperparameters to be optimised:

- (i) T_c the Pearson correlation threshold over which the non-zero pixels are considered linearly distributed;
- (ii) T_l - the lower bound on the number of non-zero pixels required.

The bar lengths returned by Algorithm 1 are measured in units of pixels. However, Galaxy Zoo images are scaled (Willett et al. 2013b) by $(\text{petroR90}_r \times 0.02 \times \text{pixel_scale})$ for each galaxy, where pixel_scale is equal to $0.396 \text{ arcsec pixel}^{-1}$ (Gunn et al. 1998). Therefore our measured bar lengths were later multiplied by this value to convert the measurements to units of $h^{-1} \text{ kpc}$. The measured lengths also had to be multiplied by a factor of $(424/224)$ in order to undo the image down-sampling described in Section 4.1.

³ The `corr()` function calculates the Pearson correlation coefficient between

Algorithm 1: Measuring the bar length from a SmoothGrad saliency map. ³

```

Input: saliencyMap,  $T_c$ ,  $T_l$ 
Output: barLength
for  $T_h = 0.5$  to  $1.0$  by  $0.0001$  do
     $x, y = \text{where}(\text{saliencyMap} < T_h)$ 
     $\text{corr} = \text{corr}(x, y)$ 
     $\text{size} = \text{len}(x)$ 
    if  $\text{corr} > T_c$  or  $\text{size} < T_l$  then
        | break
    end
end
 $x, y = \text{where}(\text{saliencyMap} > T_h)$ 
 $\text{barLength} = 0$ 
for  $i = 0$  to  $\text{len}(x)$  do
    for  $j = 0$  to  $\text{len}(x)$  do
        |  $\text{length} = \sqrt{(x[i] - x[j])^2 + (y[i] - y[j])^2}$ 
        | if  $\text{length} > \text{barLength}$  then
        | |  $\text{barLength} = \text{length}$ 
        | end
    end
end
return barLength

```

4.5 Hyperparameter optimisation

Overall, optimising the method of measuring bar lengths requires searching the space of four hyperparameters: X_{input} , σ^2 , T_c , and T_l . The values of hyperparameters were chosen as follows:

- (i) $X_{\text{input}} \in \{\text{True}, \text{False}\}$;
- (ii) $S_{\text{noise}} = \sigma / (x_{\text{max}} - x_{\text{min}}) \in \{0.1, 0.15, 0.2, 0.25\}$;
- (iii) $T_c \in \{0.7, 0.75, 0.8, 0.85, 0.9\}$;
- (iv) $T_l \in \{10, 20, 30, 40, 50\}$,

where S_{noise} is a transformation of σ^2 , and x_{max} and x_{min} are the values of the most and least intense pixels in the image being evaluated.

The Hoyle bar length catalogue was separated into two sets, a validation set of 1,050 galaxies used for hyperparameter selection, and a test set of 2,100 galaxies used for final evaluation. The measured bar lengths were then compared to the `length_scaled` values from the Hoyle bar length catalogue. The catalogue values had been scaled by the `petroR90_r` value for each galaxy (Hoyle et al. 2011), therefore this scaling was undone to ensure the values being used were in units of $h^{-1} \text{ kpc}$.

Bar lengths were measured for each combination of hyperparameters using the following methods:

- (i) saliency maps generated from the VGG16 based model;
- (ii) saliency maps generated from the ResNet50v2 based model;
- (iii) saliency maps generated from the Xception based model;
- (iv) saliency maps generated by elementwise averaging of the saliency maps from (i), (ii), and (iii);
- (v) averaging the bar lengths found by methods (i), (ii), and (iii).

Table 2 shows the Pearson correlation coefficients (ρ) between the measured bar lengths and the catalogue values for the optimal set of hyperparameters of each method. The hyperparameters were

two variables, and the `where()` function returns the indices of an array at which the condition is evaluated as True.

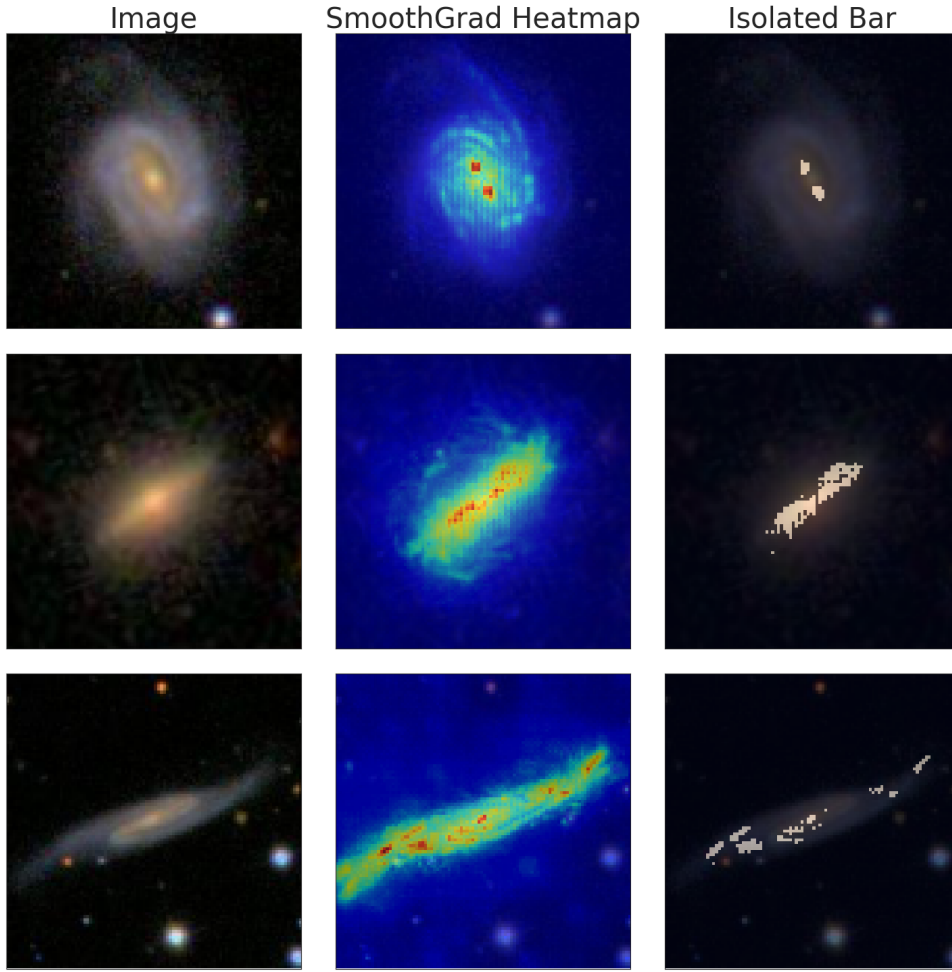


Figure 3. Examples of isolating the bar from galaxy images for three different galaxies. The leftmost column shows the initial image cropped to the centre 100×100 pixels. The middle column shows a saliency map for each galaxy highlighting the pixels most important in classifying the images as barred galaxies. The rightmost column shows the use of Algorithm 1 in isolating the bar from the relevant saliency maps. The first two examples were chosen at random, however the bottom example was chosen to present a case where this method fails (discussed in Section 5.1). The galaxies shown are identified by their SDSS Objids 588848901007933476, 588017977830342894, and 588016878291124448 from top to bottom. Similar saliency maps for a random selection of galaxies can be found in Appendix B.

Table 2. Best Pearson correlation coefficients (ρ) found for the validation set, and required hyperparameters for each method (for all, the hyperparameters X_{input} and S_{noise} were found to be False and 0.1 respectively).

Method	T_c	T_l	ρ
VGG16	0.9	40	0.68
ResNet50v2	0.85	30	0.71
Xception	0.8	40	0.67
Average Saliency Maps	0.85	30	0.75
Average Measurements	0.85	40	0.76

optimised in order to maximise ρ , which is used as a measure of predictiveness.

4.6 Control model

We also trained a CNN in order to directly predict the bar lengths of galaxies in the Hoyle catalogue from their respective galaxy images.

Section 4.3 showed the the Xception architecture performed best at classifying galaxies, hence this architecture was chosen again to directly predict bar lengths. The same image augmentation methods were used as in Section 4.1, however the number of output nodes in the model was changed to one. The model was trained to minimise the RMSE of its predictions with respect to the `length_scaled` value from the catalogue. The model was trained for 100 epochs with early stopping implemented such that training was ceased if there was no improvement over the course of 10 epochs.

For the purpose of training this model, the 3,150 galaxies in the catalogue were randomly split into training, validation, and test sets containing 64, 16, and 20 per cent of the images, respectively. Training was halted after 36 epochs at which the RMSE on the validation set was 0.10203.

5 RESULTS

5.1 Bar length measurements

Table 2 shows that the highest ρ on the validation set were gained from the “Average Measurements” method. Therefore, this method was chosen to be evaluated on the test set of the catalogue (and will hereafter referred to as the “Explainable Deep Learning” method). Evaluation of this method on the test set was performed twice to reduce and quantify any uncertainty due to the Gaussian noise added as part of the SmoothGrad algorithm. The measured bar lengths are computed as the mean of the bar lengths found during each run. Performance was shown to be similar between the validation and test sets, with measured bar lengths having $\rho = 0.76$, and an RMSE of 1.69 relative to human measured values (using hyperparameters: $X_{\text{input}} = \text{False}$, $S_{\text{noise}} = 0.1$, $T_c = 0.85$, and $T_1 = 40$).

In comparison, the CNN described in Section 4.6 (hereafter referred to as the “Direct Deep Learning” method) only managed to achieve $\rho = 0.59$, and an RMSE of 2.10 between the measured and catalogue bar length values on the model’s test set. This relatively poor performance could reasonably be explained by the much smaller size of the bar length catalogue used to train the model, in comparison to the size of the full Galaxy Zoo data used to train the other models. Although we could decouple the size of the dataset from the method, we feel like this would be intentionally handicapping XAI methods as one of their distinct advantages is that they do not require specialised data which is often harder to collect.

In most cases XAI methods are able to correctly highlight the bar of a galaxy image and measure it accordingly, as can be seen in the middle row of Figure 3. XAI methods can result in a few extreme outliers as seen in the bottom row of Figure 3. In this case the CNN mistakenly sees the entire galaxy as a bar due to the galaxy’s inclination in the image, resulting in a vastly overestimated bar length measurement. The “double blob” structure in the saliency maps for the first example in Figure 3 can be explained by the model not requiring the use of the prominent bulge in classifying these galaxies as barred, the bulge is not unique to barred galaxies and hence holds little importance in the model’s classification.

We did not expect machine learning methods to be able to reproduce absolute values of human bar length measurements. Defining the start and end points of the bar is a difficult task for both humans and machines, so we expect differences between the two methods. However, a strong correlation between the two methods would be considered a success. Linear regression fits were performed by Scikit-learn (Pedregosa et al. 2011) between the Hoyle bar length catalogue and the measured lengths for both the “Explainable Deep Learning” and “Direct Deep Learning” methods. The regression fits were forced to go through the origin, and were weighted by the uncertainty associated with each data point. For both fits the error on the Hoyle bar lengths were set to 17 per cent as specified in Hoyle et al. (2011). Errors for the bar lengths measured by the “Explainable Deep Learning” method were estimated using the spread of the measured bar lengths from each run over the test set. The spread for each galaxy (δ^s) was computed as $\delta^s = |l_1 - l_2|$, where l_1 and l_2 are the first and second measurements taken. The galaxies were then binned into deciles by their measured bar lengths. δ^s was averaged over all galaxies in each decile, and this average was used as the error for all the galaxies in the decile. Errors for the predictions of the “Direct Deep Learning” method were not quantified.

Figure 4 shows that the “Explainable Deep Learning” method not only correlates strongly with the catalogue values but also matches the values extremely well, as demonstrated by the close relation between the regression fit and the one-to-one relation. Although the

regression fit for the “Direct Deep Learning” method is also close to unity, it is important to remember that the most important metric for evaluating the performance of these methods is ρ . The distribution of the data points on the “Direct Deep Learning” plot has more scatter than for “Explainable Deep Learning” plot, and is thus it is a less predictive method.

5.2 Misclassification

Throughout this study, we came across many instances where the XAI model predictions differed from the Galaxy Zoo consensus (i.e. the model’s predicted answer to a Galaxy Zoo question was different to the answer with the highest share of votes amongst Galaxy Zoo participants). For the entire Kaggle dataset we observed that the XAI model predictions on whether a galaxy has any discernible features differed from the Galaxy Zoo consensus 8 per cent of the time. We present two examples in Figure 5 in which our ensemble models predict different morphologies to the consensus. These examples were chosen in order to present cases both where the model predicted features where Galaxy Zoo participants did not, and vice versa.

The top row shows a galaxy which our XAI model classifies as smooth (i.e. no bar, spiral arms, or other structure present) with 72 per cent confidence. Contrarily, only 39 per cent of Galaxy Zoo participants classified it as a smooth galaxy, with most classifying it as a having features (and some even classifying it as a galaxy with spiral arms). We can only speculate that these classifications are due to the neighbouring objects in the image (although Galaxy Zoo participants are instructed to only focus on the object in the centre of the image). The related saliency map shows the models being unable to pick out structure in this galaxy apart from a slight overdensity in the centre of the image. The bottom row shows a galaxy where 40 per cent of Galaxy Zoo participants have classified it as a smooth galaxy. Again, our model disagrees with this with a confidence of 53 per cent, and we can see from the saliency map that the model confidently picks out what looks like either a ring structure or dust lane in the galaxy image.

Looking at these examples, we agree more with the XAI model predictions than the Galaxy Zoo consensus. The SmoothGrad saliency maps shown in Figure 5 can help us understand why the models make their decisions, however we can only speculate as to why the Galaxy Zoo participants classified these examples in the way they did. It is possible to ask participants for their reasoning via the Galaxy Zoo Forum⁴ This brings up an interesting scenario in which the citizen science method of classifying galaxies is less easily explained than deep learning methods.

6 CONCLUSIONS

Although machine learning (ML) is widely used in astronomy, we believe that we are the first to quantitatively demonstrate the effectiveness of saliency mapping within the field. We used SmoothGrad to extract bar length measurements from convolutional neural networks (CNN) trained on Galaxy Zoo data. This method of measuring bar lengths correlated with human measurements with a Pearson correlation coefficient of 0.76 and achieved a root mean squared error

⁴ <https://www.galaxyzooforum.org/>, although this would be extremely time consuming to do for a large number of galaxies and does not guarantee responses from the participants who provided the responses in the dataset.

of 1.69. This vastly outperformed a CNN trained specifically to predict bar lengths, which was only able to achieve a correlation of 0.59 and root mean squared error of 2.10, with respect to human measurements.

The superiority of explainable artificial intelligence (XAI) methods could reasonably be explained due to the larger dataset that was used to train the ML models. The Galaxy Zoo dataset used contained 61,578 galaxy images for the training of models which can predict galaxy morphologies. In comparison, only 2,520 galaxy images were used in the training of the model tasked with directly predicting galactic bar lengths. The larger size of the Galaxy Zoo dataset allowed more general learning of galactic features and helped increase the predictiveness of the XAI based methods. A smaller dataset would reduce the performance of the initial model (Henghes et al. 2021) and thus would reduce the ability of XAI methods to extract galactic features.

The major advantage of utilising XAI methods is that information which the model was never presented with can be extracted from the model predictions. This greatly reduces the reliance of collecting data in order to learn features of galaxies. We focused on bar lengths, however we could also extract other galactic features (such as the bulge-to-disk ratio) from the same models. Traditional deep learning methods would require a new and extensive dataset in order to be able to predict these features.

We also briefly discussed the difficulties in reconciling differences between the ML model predictions and the Galaxy Zoo consensus. One possible path for future work is to use models trained only on galaxies where there is an extremely strong consensus between Galaxy Zoo participants (see e.g. Domínguez Sánchez et al. 2018), models designed to handle galaxies with uncertain labels (see e.g. Walmsley et al. 2020), or to use pretrained models (see e.g. Walmsley et al. 2022). This could improve the ML model predictions, and lead to cleaner saliency maps and more accurate bar length measurements. Another possibility would be to use the saliency maps in order to identify and eliminate bias from the training dataset, which would lead to a higher quality dataset for training models.


We conclude that SmoothGrad was able to explain the decisions made by the models in a way that could be quantified and compared favourably to human action. This demonstrates that saliency mapping was successful in explaining the ML models used in this paper within the context of bar lengths, and could see a wide range of applications within the field of astronomy.

ACKNOWLEDGEMENTS

The authors would like to thank A. Bevan for drawing our attention to saliency mapping techniques, and M. Walmsley for their insightful and helpful feedback. PB is supported by the STFC UCL Centre for Doctoral Training in Data Intensive Science. BJ acknowledges support by an STFC Consolidated Grant, grant no. ST/V000780/1. OL acknowledges STFC Consolidated Grant ST/R000476/1 and a Visiting Fellowship at All Souls College, Oxford.

DATA AVAILABILITY

All code required to reproduce this work will be made publicly available on acceptance. The code includes a readme detailing the steps required to reproduce the study, including downloading all data and setting the seeds used when randomly splitting the datasets and augmenting training data. Please note that the version of the

saliency mapping package used does not support the use of random seeds, however the high number of samples used in the SmoothGrad algorithm should minimise the impact of this. [GitHub](#) .

REFERENCES

- Abadi M., et al., 2016, arXiv e-prints, p. [arXiv:1603.04467](#)
- Adadi A., Berrada M., 2018, *IEEE Access*, 6, 52138
- Aloysius N., Geetha M., 2017, in 2017 International Conference on Communication and Signal Processing (ICCSP), pp 0588–0592
- Athanassoula E., Romero-Gómez M., Bosma A., Masdemont J. J., 2009, *MNRAS*, 400, 1706
- Banerji M., et al., 2010, *MNRAS*, 406, 342
- Chattopadhyay A., Sarkar A., Howlader P., Balasubramanian V. N., 2017, arXiv e-prints, p. [arXiv:1710.11063](#)
- Chollet F., 2016, arXiv e-prints, p. [arXiv:1610.02357](#)
- Dieleman S., Willett K. W., Dambre J., 2015, *MNRAS*, 450, 1441
- Domínguez Sánchez H., Huertas-Company M., Bernardi M., Tuccillo D., Fischer J. L., 2018, *MNRAS*, 476, 3661
- Domínguez Sánchez H., et al., 2019, *MNRAS*, 484, 93
- Došilović F. K., Brčić M., Hlupić N., 2018, in 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp 0210–0215, doi:10.23919/MIPRO.2018.8400040
- Gunn J. E., et al., 1998, *AJ*, 116, 3040
- He K., Zhang X., Ren S., Sun J., 2016, arXiv e-prints, p. [arXiv:1603.05027](#)
- Henghes B., Pettitt C., Thiyyagalingam J., Hey T., Lahav O., 2021, *MNRAS*, 505, 4847
- Hoyle B., et al., 2011, *MNRAS*, 415, 3627
- Kingma D. P., Ba J., 2014, arXiv e-prints, p. [arXiv:1412.6980](#)
- Kormendy J., Kennicutt Robert C. J., 2004, *ARA&A*, 42, 603
- Lahav O., et al., 1995, *Science*, 267, 859
- Lahav O., Naïm A., Sodr e L. J., Storrie-Lombardi M. C., 1996, *MNRAS*, 283, 207
- Lintott C. J., et al., 2008, *MNRAS*, 389, 1179
- Pedregosa F., et al., 2011, arXiv e-prints, p. [arXiv:1201.0490](#)
- Peruzzi T., Pasquato M., Ciroti S., Berton M., Marziani P., Nardini E., 2021, *A&A*, 652, A19
- Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., 2016, arXiv e-prints, p. [arXiv:1610.02391](#)
- Simonyan K., Zisserman A., 2014, arXiv e-prints, p. [arXiv:1409.1556](#)
- Simonyan K., Vedaldi A., Zisserman A., 2013, arXiv e-prints, p. [arXiv:1312.6034](#)
- Smilkov D., Thorat N., Kim B., Viégas F., Wattenberg M., 2017, arXiv e-prints, p. [arXiv:1706.03825](#)
- Springenberg J. T., Dosovitskiy A., Brox T., Riedmiller M., 2014, arXiv e-prints, p. [arXiv:1412.6806](#)
- Villanueva-Domingo P., Villaescusa-Navarro F., 2021, *ApJ*, 907, 44
- Walmsley M., et al., 2020, *MNRAS*, 491, 1554
- Walmsley M., et al., 2022, *MNRAS*, 509, 3966
- Willett K. W., et al., 2013a, Galaxy Zoo 2: Images from Original Sample, <https://doi.org/10.5281/zenodo.3565489>
- Willett K. W., et al., 2013b, *MNRAS*, 435, 2835

APPENDIX A: RANDOMISED SALIENCY MAPPING EXAMPLES

Here we present further examples similar to Figure 1, of SmoothGrad saliency maps for three chosen classes on five randomly selected galaxies. The galaxies selected are from the Kaggle dataset used to train the CNNs in Section 4.2, with the condition that the selected galaxy must have at least a 50 per cent of voters classifying it as having a bar, a bulge, and spiral arms.

APPENDIX B: RANDOMISED BAR LENGTH MEASUREMENT EXAMPLES

Here we present further examples similar to Figure 3, of SmoothGrad saliency maps isolating the bar of five randomly selected galaxies. The galaxies selected are from the test set of the Hoyle catalogue used to evaluate the performance of the XAI based methods in Section 5.1.

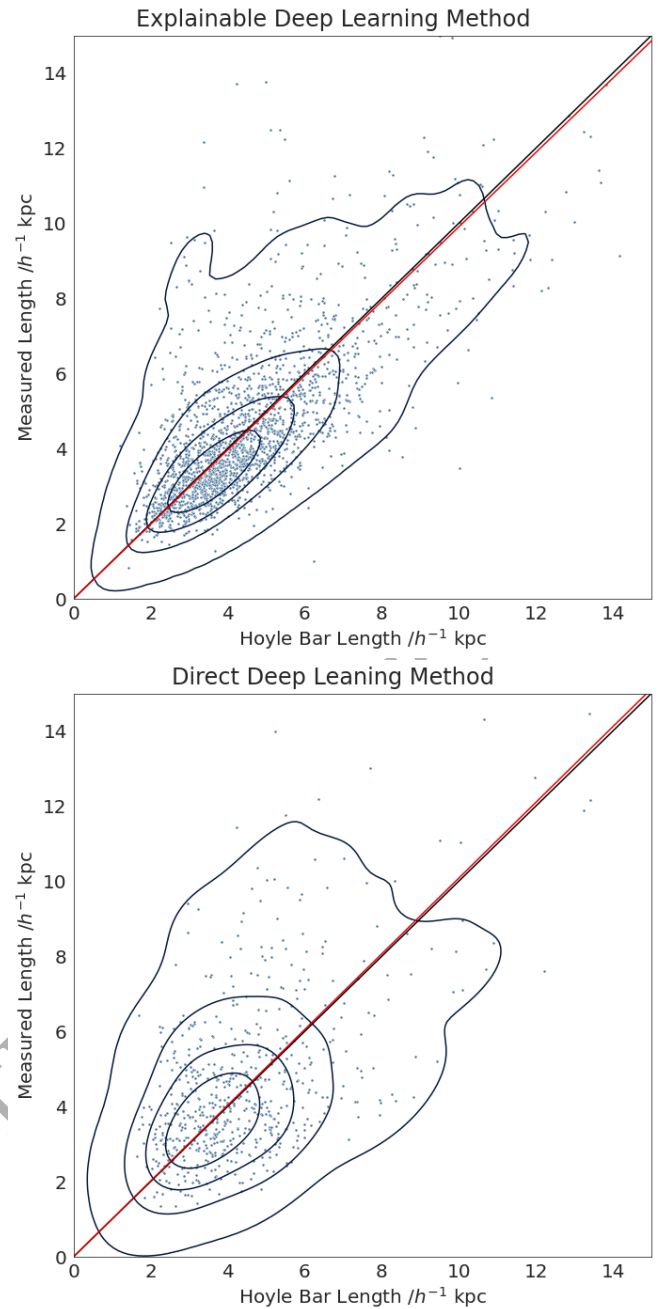


Figure 4. Bar lengths measured by the “Explainable Deep Learning” (top) and “Direct Deep Learning” (bottom) methods vs the Hoyle bar length catalogue for each method’s test dataset. Weighted regression fits are also shown (red lines), as well as the one-to-one relation (black lines). A kernel density estimate overlay (dark blue contours) is provided to help illustrate the distribution of data points in each panel and error bars are not displayed for the sake of clarity. The regression fits for both are close to unity, however the “Explainable Deep Learning” has a smaller scatter than that of the “Direct Deep Learning” method indicating stronger predictive power.

ORIGINAL UNEDITED

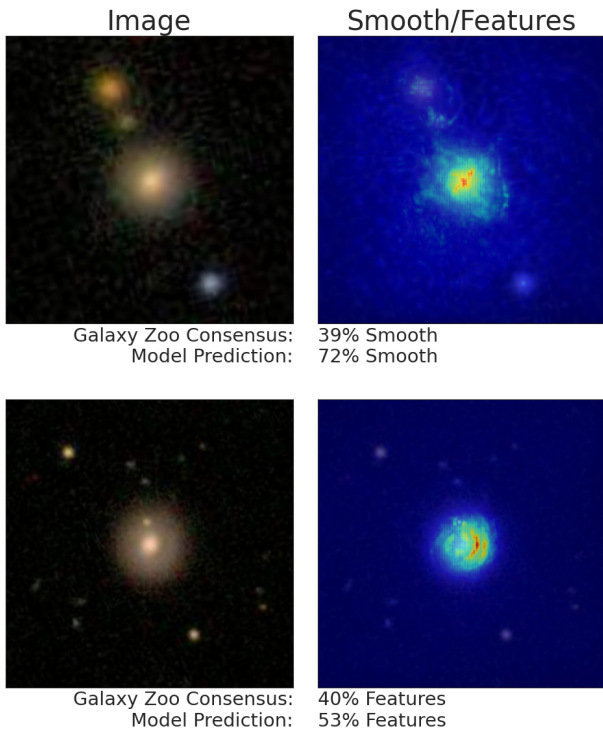


Figure 5. Two examples where the machine learning prediction disagrees with the Galaxy Zoo citizen science consensus. The image (left column) is shown along with SmoothGrad saliency maps explaining the model's classification of this image (right column). The top row shows a galaxy where the Galaxy Zoo consensus is that of a galaxy with features and structure, while the bottom row shows an example where the consensus is that of a smooth galaxy. Our model disagrees with both of these. The Galaxy Zoo consensus as well as the model predictions probabilities are given for both examples. The galaxies shown are identified by the filenames 102271.jpg and 120250.jpg from the Kaggle dataset.

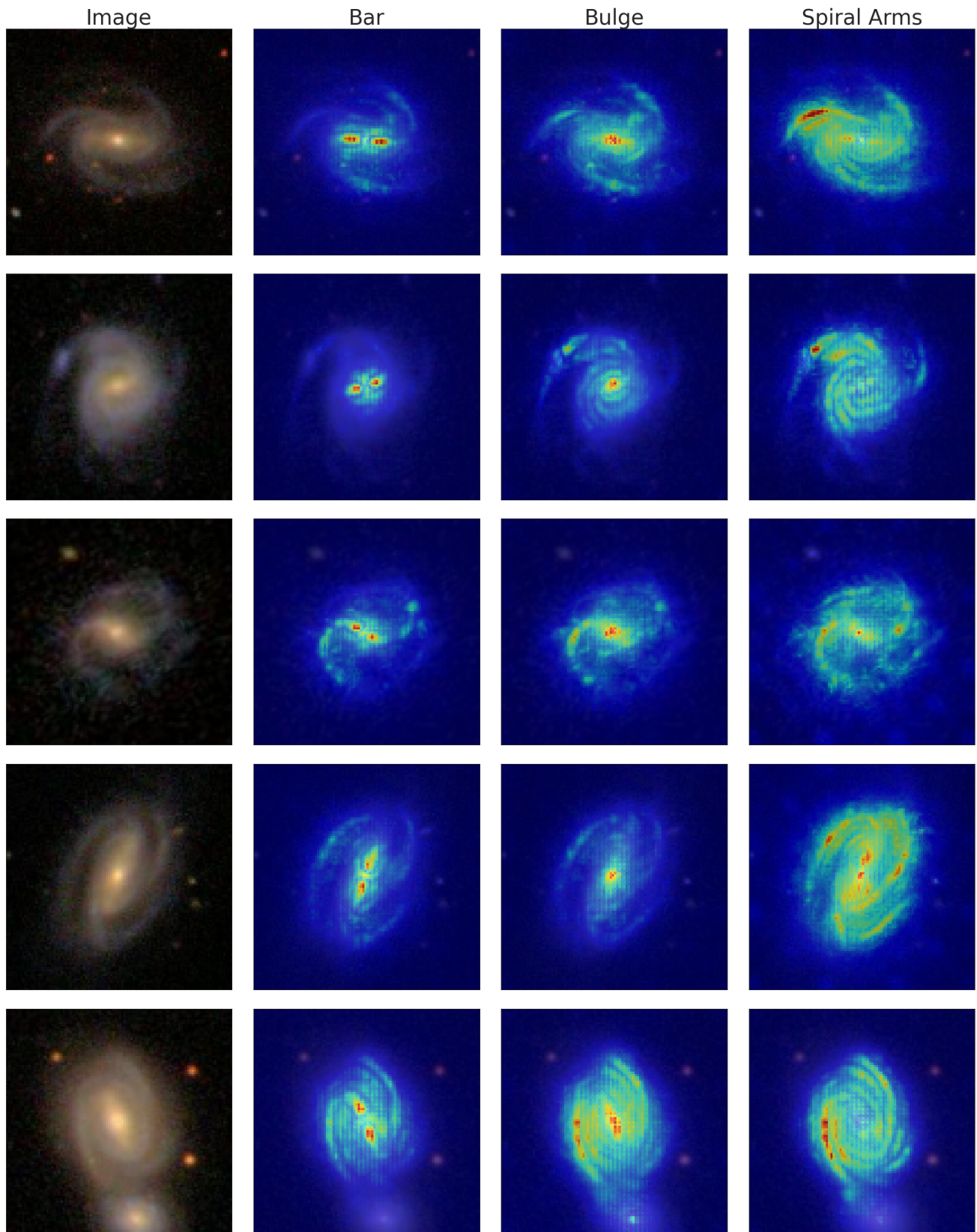


Figure A1. Similar to Figure 1, but with examples randomly selected from the Kaggle dataset, with the conditions that the selections must have at least 50 per cent of voters classifying it as having a bar, a bulge, and spiral arms.

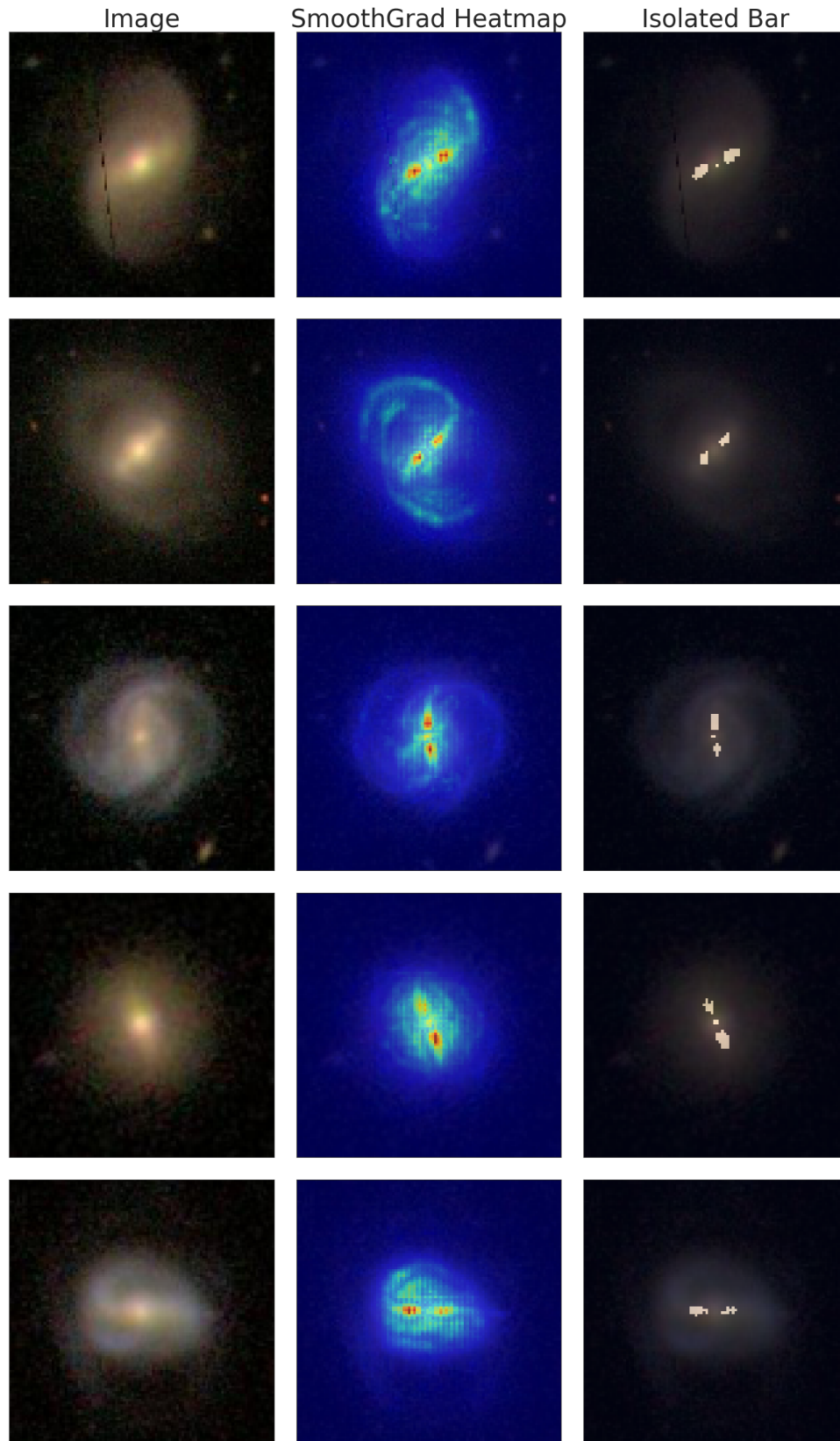


Figure B1. Similar to Figure 3, but with examples randomly selected from the Hoyle catalogue.