Nonlinear Analysis: Modelling and Control, Online First, 1–22
https://doi.org/10.15388/namc.2026.31.47651

Vilnius University Press

Large-scale performance evaluation of galaxy image classification models trained on small datasets

Tomas Mūžas, Andrius Vytautas Misiukas Misiūnas, Tadas Meškauskas

Vilnius University, Faculty of Mathematics and Informatics,
Institute of Computer Science
Didlaukio 47, Vilnius LT-08303, Lithuania
tomas.muzas@mif.vu.lt

Received: February 22, 2026 / **Revised:** June 25, 2026 / **Published online:** July 7, 2026

Abstract. Many studies have utilized convolutional neural networks to classify galaxies into spiral and elliptical, achieving near-perfect results. However, the authors used relatively small datasets (up to 104,787) that often consist of higher-quality images, making it difficult to extrapolate model large-scale performance. In contrast, we combine data from Galaxy Zoo projects (1, 2, Hubble, CANDELS, DECaLS), preserving as many galaxies as possible. To achieve that, we propose a novel methodology – spiral certainty index (SCI), which allows to extend binary classification by introducing uncertain classes for galaxies where volunteers do not fully agree on the class. Using SCI, we bin galaxies into 2, 3, and 5 classes, resulting in three large datasets of 719,133–800,448 galaxies. We then quantitatively evaluate large-scale performance of Cavanagh, ResNet50, and EfficientNetV2S architecture models by training them using small fractions (0.4–46.5%) of the datasets and validating them on the rest. Experiments were repeated with and without standard augmentations (rotation, flipping, zoom, noise). We conclude that for future research using Galaxy Zoo data, the training dataset of at least 71,913 images (10% of 5-class dataset) is recommended to reach stable accuracy and F1 metrics for all three architectures studied. Below this threshold, Cavanagh, and ResNet50 models either overfit or display high variability of metrics. Moreover, at this threshold or above, the usage of augmentations is equivalent to doubling or even quadrupling (in case of ResNet50 for all classes) the training dataset size. The most stable architecture was EfficientNetV2S, which only overfitted with 2-class dataset, when trained on 3,715 images (0.4% of the dataset), without augmentations.

Keywords: large-scale model evaluation, galaxy morphology classification, Galaxy Zoo, image augmentation, convolutional neural networks, spiral certainty index.

1 Introduction

In astronomical and cosmological research, the classification of galaxies by morphological type is crucial to understanding their evolutionary stage. Notably, the fundamental disparity between spiral and elliptical galaxies lies in their star formation activity. The

spiral galaxies feature active star-forming regions within their spiral arms, whereas elliptical galaxies lack such regions.

For a long time, galaxy morphological classification was performed by manually inspecting the photographs obtained from the telescopes [9,21,23,24]. However, with the automation of observation process, the amount of data grew exponentially, rendering such an approach impractical. For example, one of the larger datasets, compiled by Nair and Abraham consists of 14,000 classifications [21] while the data from the Sloan Digital Sky Survey (SDSS) contains over 1,000,000 galaxies [35]. Even further, SDSS data volume is dwarfed by the 300 million galaxies from the Dark Energy Survey [29].

To tackle such a vast amount of data, the project called Galaxy Zoo was developed by Lintott et al. [16] (hereinafter, Galaxy Zoo projects, in general, are referred to as GZ, and this specific project is referred to as GZ1). It aimed to obtain a large number of classifications by presenting volunteers (who did not necessarily have any prior education in astrophysics) with an image of a galaxy and asking them to answer a visual questionnaire. Based on their answers, the galaxy was assigned the most likely class of either spiral, elliptical or unknown. The volunteers managed to classify a total of 667,944 galaxies [16].

Following the success of the GZ1 project, four more projects covering different surveys were later developed: Galaxy Zoo 2 (GZ2) [33], CANDELS (GZC) [25], DECaLS (GZD) [31], and Hubble (GZH) [32]. In total, these projects provide more than 1 million classifications (see Section 3 for more information regarding the datasets).

However, such projects have a limitation – to obtain a reliable classification, the image of a galaxy must be shown to a significant number of volunteers. For instance, the GZ2 project had a median of 44 votes for a galaxy [33]. Therefore, with an exponentially increasing amount of data, this approach is becoming unfeasible as well.

Nonetheless, with recent advancements in machine learning, it is possible to fully automate galaxy image classification. Researchers have successfully applied convolutional neural networks to achieve great results – Barchi et al. achieved 99.5% accuracy on the brightest galaxies from the GZ1 dataset [2], and Cavanagh et al. achieved 98% accuracy [6]. Such results show that after training the neural network once, previously unseen galaxies can be classified automatically with high accuracy.

However, as discussed in Section 3, a wide range of dataset sizes was used to train and evaluate models. Moreover, the data is often handpicked to contain only the brightest galaxies (see Section 2 for more information). This makes it difficult to compare models to each other and evaluate their ability to classify diverse unseen images at a large scale.

Therefore, contrary to our previous study [20], this study does not aim to outperform the results of other authors, but rather aims to quantitatively evaluate how models of Cavanagh, ResNet50, and EfficientNetV2S architectures (see Section 4.1), which were trained using small datasets, would perform on a large scale. Another aim of the paper is to quantify the threshold of the training dataset size, using which models of all three architectures would not overfit and display stable results, as per criteria defined in Section 5.1.

To evaluate the large-scale performance of the models, we combine classification data from 5 GZ projects. To address the inherent uncertainty of the GZ volunteer data, we

apply a novel classification methodology. First of all, we calculate the Spiral Certainty Index (SCI, see Section 3.2), which indicates how certain the volunteers are that the galaxy is spiral. Afterwards, in contrast to other studies, we extend the traditional binary classification problem to 2, 3, and 5 classes by binning galaxies based on their SCI value.

Using the aforementioned SCI bins, three large datasets are created, consisting of 800,448 unique galaxies (see Section 3) for 2 classes, 762,620 galaxies for 3 classes, and 719,133 galaxies for 5 classes, which is at least twice as large as dataset used in our other large-scale study [19]. It is ensured that the SCI metric of a galaxy is in the same bucket across all projects that contain that galaxy, hence why the dataset sizes are different. The galaxies whose SCI values differ between projects are excluded.

Finally, models are trained on increasingly larger fractions of the datasets, ranging from 0.4% up to 46.5% (see Table 3), and evaluated on the remaining data. Moreover, the training process was repeated both with and without the standard augmentations (see Section 4.3) to evaluate their impact on model stability. While our previous study on augmentations [19] evaluated the effect of different techniques on metrics separately, this one focuses on their combined impact when using varying amounts of training data.

The main contributions of this study are summarized as follows:

- We conduct a quantitative evaluation of the large-scale performance of Cavanagh, ResNet50, and EfficientNetV2S models by training them using 0.4–46.5% of the aforementioned datasets and validating on the remaining data. The experiments were carried out both with and without augmentations to study their effect on the model stability as well.
- We introduce a Spiral Certainty Index (SCI) – a metric that represents the confidence of Galaxy Zoo volunteers that the galaxy is spiral. We propose a novel classification methodology, which extends the traditional binary classification – the galaxies are binned into 2, 3, and 5 classes based on their SCI value, introducing unknown and uncertain spiral/elliptical classes. Having those classes allows the model to pick a neighbouring class of lesser degree of certainty rather than being forced to pick between two.
- Based on the SCI bin, we have combined Galaxy Zoo data into three datasets containing 719,133 (5 classes) to 800,448 (2 classes) unique galaxies, which is at least 7 times larger than the largest dataset used by other authors and more than 2 times larger than in our previous study [19]. Furthermore, as many galaxies were preserved as possible, excluding only the galaxies whose SCI bins differed across multiple GZ projects. This was done to have as diverse data as possible, which allows to estimate the model's large-scale performance even better.

2 Related work

To classify vast amounts of galaxy images, multiple machine learning techniques (such as random forest or decision trees [2, 7, 13, 17], K-nearest neighbours [7, 13], support vector machines [2,7,13] or auto-encoders [13]) have been studied by other authors, although the best results have been achieved using convolutional neural networks (CNN) [2, 7, 13, 17].

Table 1. Comparison of the dataset sizes of various authors that performed binary classification.

Author	Dataset size
Cheng et al. [7]	2,862
Bom et al. [4]	4,232
Cavanagh et al. [6]	14,304
Jimenez et al. [13]	41,424
Gupta et al. [10]	61,578
Barchi et al. [2]	104,787
Mūžas et al. [19]	315,942
This study	**719,133−800,448**

In terms of binary galaxy classification (spiral versus elliptical), near-perfect results have been achieved – Barchi et al. can boast of their 99.5% accuracy on the brightest galaxies and 98.7% on a wider range of galaxies the GZ1 project [2]. Cheng at al. have achieved impressive results as well – 99.4% accuracy on GZ1 and Dark Energy Survey data [7]. Other prominent results include 98.5% accuracy by Bom et al. [4], 98% accuracy of Cavanagh et al. [6], and 96.43% by Jimenez et al. [13]. Such great results are usually achieved by applying augmentation techniques, such as image rotation [6–8, 14, 17, 36], flipping [6, 8, 14, 17, 36] zooming in [8, 14, 36] or random noise [5, 7].

Nonetheless, while the results achieved by the aforementioned authors are impressive, their models were trained and validated on relatively small datasets. The dataset sizes used for binary classification are shown in Table 1.

Considering the fact that all GZ projects combined contain more than 1 million classifications (see Section 3 for more details), this raises a question of whether datasets as small as 2,862 used by Cheng et al. [7] are enough to train a model that reliably classifies millions of galaxies. In addition, the datasets used are often comprised of the best quality images – either the closer, brighter galaxies are selected (such as the K-metric used by Barch et al. [2]), or the ones that have the higher classification confidence (such as the expert-classified dataset chosen by Jimenez et al. [13]), which may yield better results, though the results may not be reproducible with images of lower quality.

In this study, the performance of the models is evaluated using three large datasets – the largest one consisting of 800,448 unique galaxies, and the smallest has 719,133 galaxies, which is at least 7 times larger than the one used by Barchi et al. [2] and at least twice as large and containing data from more GZ datasets compared to our previous study [19], To evaluate the threshold of how much data is enough to train a stable model (as defined in Section 5.1) for large-scale classification, 0.4% to 46.5% (see Table 3) of the aforementioned datasets are dedicated for training, while the rest is left for validation. To ensure that this threshold is applicable to as wide range of models as possible, models of three distinct CNN architectures of different complexity were studied – Cavanagh et al. [6], ResNet50 [11], and EfficientNetV2S [27]. Additionally, the impact of the standard augmentation techniques (see Section 4.3) on the model stability and overfitting were tested as well.

Finally, the three aforementioned datasets are created by calculating the Spiral Certainty Index (SCI, see Section 3.2), which denotes volunteers' certainty that the galaxy is

spiral. The galaxies are grouped into 2, 3, and 5 classes based on this metric. The partition into 3 and 5 classes introduces additional, uncertain classes, which allows the model to assign a less certain, neighbouring class rather than being forced to choose between just two options.

3 Data

This section describes the data filtering pipeline, image gathering process and resulting datasets' distribution. The initial data was combined from five Galaxy Zoo projects – GZ1 [16], GZ2 [33], GZC [25], GZD [31], and GZH [32]. Those datasets combined result in a dataset of more than 1.2 million galaxies that were classified by humans.

However, the approach of the human-classified Galaxy Zoo projects introduces a set of biases. First of all, galaxies are classified by volunteers who answer a visual questionnaire, without the requirement of any prior education in astrophysics [16, 33]. This bias is combated by acquiring a large amount of classifications for a single galaxy (for example, a median of 44 votes per galaxy in GZ2 [33]), as well as reducing the weight of the volunteers who consistently disagree with others [33]. Additionally, since the galaxy morphology can be ambiguous, in some cases, the volunteers might not reach a consistent conclusion on the class (see more information in Section 3.2). Finally, the same galaxy might appear in multiple GZ projects (for example, GZ2 is a subset of GZ1 galaxies [16, 33]), and due to the differences in image quality between projects, the classification may not agree between the projects.

This study aims to address the volunteer uncertainty and possible class mismatch between the datasets by performing a data filtering process. First of all, the galaxies are cross-matched by their coordinates, as all projects, except GZ1 and GZ2, use different galaxy identifiers. This process is outlined in Section 3.1. After that, we introduce a metric called Spiral Certainty Index (SCI), described in Section 3.2, which measures how certain the volunteers are that the galaxy is spiral. Then, based on the SCI value, a galaxy is assigned into a bin of 2, 3, and 5 classes. Finally, it is ensured that the galaxy belongs to the same SCI bin across all Galaxy Zoo datasets (more information is provided in Section 3.3). The following sections describe each step in detail.

3.1 Cross-match by coordinates

The first step was to cross-match galaxies by their coordinates, as a single galaxy could appear in multiple datasets, and their dataset IDs were different depending on the project. The process was performed within the radius of 0.1 arcsec, as measured by a straight line. After this operation, a database of galaxies was built, where each entry contained all the different dataset IDs it was in.

3.2 Spiral certainty index (SCI)

The second step was to assign galaxies their classifications. After analysing the voting data, it was discovered that the volunteer votes are sometimes distributed very similarly between spiral and elliptical, signifying an ambiguous shape of the galaxy. Hence, to

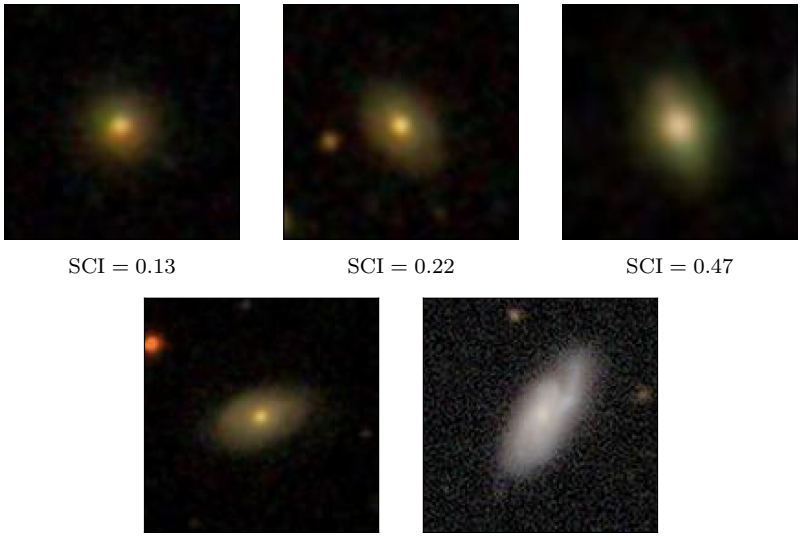

Figure 1. Examples of different Spiral Certainty Index (SCI) ranges.

address the arising classification uncertainty, a metric referred to as the Spiral Certainty Index (hereinafter SCI) was introduced, which evaluates the certainty of the volunteers that the galaxy is spiral from 0 (no certainty, hence elliptical) to 1 (certain spiral).

The SCI was derived from both debiased spiral and elliptical vote fractions. The debiasing process, used by the authors, aimed to address the fact that for the more distant galaxies, their features may be more difficult to resolve [33].

In most of the projects, the debiased vote fraction had a value between 0 and 1. However, in the GZH project, a different debiasing process was performed, which in rare cases resulted in negative values or values greater than 1 [32]. As only 20,847 galaxies were impacted, it was decided to clamp all values to the range of 0 to 1 using the following expression:

$$\mathrm{CP} = \min\bigl(\max(P, 0), 1\bigr).$$

Here P stands for the debiased vote fraction, and CP stands for clamped probability. This operation ensured that all negative values were clamped to 0, while all values above 1 were clamped to 1. The SCI was subsequently calculated from the clamped probabilities as

$$\mathrm{SCI} = \frac{\mathrm{CP}_{\mathrm{Spiral}}}{\mathrm{CP}_{\mathrm{Spiral}} + \mathrm{CP}_{\mathrm{Elliptical}}}$$

The examples of different SCI values are shown in Fig. 1.

It was decided not to exclude galaxies with an ambiguous shape (SCI ≈ 0.5), but instead to keep all galaxies to maximise image diversity, allowing for a better estimation of the model's real-life performance. Instead, the SCI value was used to extend the traditional binary classification – the galaxies were grouped into 2, 3, and 5 bins, respectively.

3.2.1 Interpretation and significance of SCI

The resulting SCI values can be interpreted in an astrophysical sense. For 2 classes, there are only elliptical and spiral galaxies, which corresponds to the traditional binary classification. For 3 classes, there are elliptical, unknown, and spiral galaxies. For 5 classes, there are certain elliptical, uncertain elliptical, unknown, uncertain spiral, and certain spiral galaxies.

These newly introduced uncertain classes serve a purpose in the reliability of model predictions. If a model encounters a galaxy with an ambiguous shape, rather than picking between only spiral and elliptical, it now has a choice to assign either a class of lower certainty, or pick an unknown class. The latter signifies that the shape of the galaxy is too ambiguous or there is not enough data to classify it properly. Such galaxies can then be either inspected manually or given to another classifier to further inspect them.

3.3 Cross-match by SCI bins

The last step was to remove galaxies whose SCI bins did not agree between the different GZ datasets. The SCI bins could have differed, if, for example, the images of the same galaxy were better in another project. Therefore, only those galaxies, where the SCI bin remains the same across all projects they are in, are kept.

This process was performed in the following way:

(i) Based on the cross-matching by coordinates described in Section 3.1, it was determined in which of the original GZ datasets the galaxy is contained. If at least one dataset's SCI bin, calculated in Section 3.2, did not match with the other ones, the galaxy was filtered out and treated as ambiguously classified.

(ii) The remaining dataset was further cleaned from the duplicate galaxies – the galaxies that have agreeing classifications, though based on distance cross-match, appear in the dataset more than once. This left only a unique entry for each galaxy.

It is important to emphasize that this process was performed separately for each class, since SCI values may agree, for example in 2-class dataset, however they may differ in 3-class dataset. This resulted in different dataset sizes. The resulting sizes are as follows: 2 classes – 820,535; 3 classes – 781,974; 5 classes – 737,561 unique and unambiguously classified galaxies.

3.4 Galaxy images

Finally, after filtering and merging the datasets, the corresponding images of the remaining galaxies had to be gathered. The initial images were taken from the `galaxy-datasets` package[1]. Additionally, since the package does not contain the images from the GZ1 project, the GZ1 images were collected using the SkyServer's ImgCutout service[2] [22] and SciServer script[3], preserving the same scaling parameters as described by

[1] `https://pypi.org/project/galaxy-datasets/`

[2] `http://skyservice.pha.jhu.edu/DR7/ImgCutout/getjpeg.aspx`

[3] `https://github.com/sciserver/SciScript-Python/blob/master/py3/SciServer/SkyServer.py`

Table 2. Experiment setup: possible choices for architecture, training split size, and the usage of augmentations.

	Architecture	Training dataset split	Augmentations
Choices	{Cavanagh, ResNet50, EfficientNetV2S} (Section 4.1)	{0.4%, 1%, 2.2%, 4.6%, 10%, 21.5%, 46.4%} (Section 4.2)	{yes, no} (Section 4.3)

the authors [16]. Finally, it is worth mentioning that some subsets (such as GZD1 and GZD2) were not present in the package. However, since all projects were represented, it was decided not to gather missing images.

It is important to note that even though the information to which different datasets a particular galaxy belonged was preserved, it was decided to take only a single image, from the galaxy's original dataset. Therefore, this fact, combined with the omission of some subsets discussed above, resulted in some additional galaxies being filtered out from the three datasets obtained in Sections 3.1–3.3.

The final dataset sizes, after the filtering process and the omission of galaxies without images were: 2 classes – 800,448, 3 classes – 762,620, 5 classes – 719,133. The distribution of classes is as follows:

- 2 classes – $\mathrm{SCI} \in \mathbf{[0, 0.5)}$: 338,786 (42%), $\mathbf{[0.5, 1]}$: 461,662 (58%);
- 3 classes – $\mathrm{SCI} \in \mathbf{[0, 0.33)}$: 246,188 (32%), $\mathbf{[0.33, 0.66)}$: 151,702 (20%), $\mathbf{[0.66, 1]}$: 364,730 (48%);
- 5 classes – $\mathrm{SCI} \in \mathbf{[0, 0.2)}$: 154,072 (21%), $\mathbf{[0.2, 0.4)}$: 105,415 (15%), $\mathbf{[0.4, 0.6)}$: 84,801 (12%), $\mathbf{[0.6, 0.8)}$: 107,088 (15%), $\mathbf{[0.8, 1]}$: 267,757 (37%).

For all of the remaining galaxies, images with an initial size of 424×424 pixels were cropped to the size of 212×212 pixels and finally downscaled to the final training size of 128×128 pixels.

As a final remark on the data, it is important to mention that recently the authors of the Galaxy Zoo projects produced an even larger dataset GZ DESI [30], containing 8.7 million galaxies, though the classifications were obtained by a machine learning model. However, since the model predicted the vote distribution of volunteers [30], this would have introduced an additional bias into the study, and would not fit into the calculation of SCI (see Section 3.2), as it represents the volunteer certainty based on their actual voting results. Therefore, this dataset was not included into this study.

4 Experiment setup

This section describes the setup of the experiments. Overall, an experiment is defined by a combination of three key parameters – architecture, training dataset split size, and usage of data augmentation. All parameter values are summarised in Table 2.

The average model performance was evaluated using Monte Carlo cross-validation with 10 folds, further described in Section 4.4.1. In order to ensure that only the best results of each experiment were used for comparison, early stopping was applied as described in Section 4.4.3. The models were trained using fixed training parameters (see

Section 4.4.2) and a batch size of 1,024, however, some restrictions were applied as defined in Section 4.4.4.

4.1 Architectures

In total, three distinct architectures were chosen to carry out the experiments. The first one was used by one of the researchers who have achieved high classification accuracy – C2 architecture by Cavanagh et al. [6]. This architecture is a slight modification of the one initially proposed by Dieleman et al. [8]. It is simple and is comprised only of the basic blocks – convolution, batch normalization, dropout, and dense layers [6].

The second architecture is the well-known ResNet50 [11], which was used by Jimenez et al. [13] or Gupta et al. [10]. This architecture includes a more sophisticated residual layer [11].

The final and the most sophisticated architecture studied in this paper is EfficientNetV2S [27], the first iteration of which was used by Urechiatu and Frincu [28], though not yet thoroughly studied in the field of galaxy classification. It includes new layers such as MBConv and Fused-MBConv [27]. The smallest version was chosen due to the memory and computational resource constraints.

4.1.1 Architecture modifications

Each architecture was implemented as described in its original paper, though modified in such a way that it accepts a batch of images of size 128×128, and the last layer produces an output of 2, 3 or 5 nodes in the range of $(0, 1)$ using softmax activation function. The final classification was determined by selecting the node with the highest value (argmax).

To avoid randomness as much as possible, before conducting any experiments, the initial weights for each architecture and each dataset (as there is a different number of output nodes, hence – different number of parameters) combination were generated randomly and saved. In each experiment, the models were initialized with the corresponding initial weights and then trained independently.

4.2 Training dataset splits

To evaluate how much data is required for stable model performance, the 2-, 3-, and 5-class datasets obtained in Section 3 were divided into training and validation datasets. It was decided to use a logarithmic scale with an emphasis on the small numbers, so that as much data remains for large-scale model evaluation. These were the training split fractions: 0.4%, 1.0%, 2.2%, 4.6%, 10.0%, 21.5%, and 46.4%. The dataset split sizes are shown in Table 3.

4.3 Data augmentation

Data augmentations allow to artificially increase the training dataset size without having to collect new samples. In this study, 5 augmentation techniques were used – rotation, horizontal and vertical flipping, zoom, and random noise. While more augmentation techniques, such as targeted noise inside or outside the galaxy, were explored in our

Table 3. Training and validation split sizes of all three datasets. More information on dataset sizes can be found in Section 3.4.

Split, %	2 Classes	3 Classes	5 Classes
0.4	3,715/796,733	3,540/759,080	3,338/715,795
1.0	8,004/792,444	7,626/754,994	7,191/711,942
2.2	16,009/784,439	15,252/747,368	14,383/704,750
4.6	40,022/760,426	38,131/724,489	35,957/683,176
10.0	80,045/720,403	76,262/686,358	71,913/647,220
21.5	176,099/624,349	167,776/594,844	158,209/560,924
46.4	368,206/432,242	350,805/411,815	330,801/388,332

previous study [19], it was shown that the combination of these 5 techniques resulted in the most stable metric values out of all the combinations studied. Additionally, these techniques are frequently used in other galaxy image classification studies [5–8,14,17,36].

It is important to mention that it would have been beneficial to explore different combinations and parameters of the augmentations. However, the aim was not to study the impact of each augmentation technique individually, but rather to assess an overall effect of augmentations on the model stability. Therefore, due to limited computational resources, it was deemed out of the scope of this study.

The implementation details of the augmentation techniques are described below:

(i) *Rotation* used by multiple authors [6–8, 14, 17, 36]. An image is rotated by either 90, 180, or 270 degrees.

(ii) *Horizontal flip* used by [6,8,14,17,36]. The image pixels are flipped horizontally.

(iii) *Vertical flip* used by [6, 8, 14, 17, 36]. The image pixels are flipped vertically.

(iv) *Zoom*. This technique, or its variations, was used by [8, 14, 36]. In this paper, 10 pixels are cropped from each side of an image. The cropped image is then rescaled back to the size of 128×128 pixels.

(v) *Random noise*. This technique was used by [5,7]. In this study, the random noise was generated by sampling values from -0.2 to 0.2 from the uniform distribution. The value of each pixel was later clipped to be between 0 and 1. While other distributions and value ranges could have been tested, due to limited computational resources, only the values explored in our previous study [19]were used.

During the training phase, each augmentation technique could have been applied independently with a probability of 0.5 (none of the augmentations were applied during the validation phase). This means either none, some combination, or in rare cases, all augmentations could have been applied to an image, theoretically increasing the dataset size by up to 48 times. However, all of the dataset sizes reported in this paper count only the original images.

4.4 Training setup

4.4.1 Monte Carlo cross–validation

Since many different splits of the training dataset are studied, using the standard k-fold cross-validation [3] method would not result in an equal number of folds for each split,

hence it was decided to use a variation of cross-validation called Monte Carlo cross-validation [34] or k-fold random subsampling [3]. In this case, for each dataset split, training data is randomly sampled 10 times, and the remaining data is left for validation. It is important to note that for each split, all 10 folds were pre-computed, remained the same for all the models during the training process. This ensured that the resultswere comparable.

4.4.2 Optimizer

The loss function to optimize was chosen to be categorical cross-entropy. This was done to have the same setup for the experiments of all classes. Additionally, the categorical cross-entropy loss function simplifies to a binary cross-entropy function for two classes.

The optimizer for the loss function was chosen to be Adam [15], as it was used by other related studies [6,7]. The default parameters, as defined by the authors Kingma and Ba [15], of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \cdot 10^{-8}$ were used. In order to prevent overfitting, a learning rate of $1 \cdot 10^{-4}$ was chosen, at the cost of potentially requiring more epochs for the model to converge.

4.4.3 Early stopping

In this study, the training process was stopped early when the model failed to decrease the validation loss for 10 consecutive epochs, with a maximum limit of 500 epochs (though none of the models has reached this threshold). All the results reported are taken from the epoch where the model achieved the lowest loss value before it started to overfit.

4.4.4 Batch sizes and constraints

Since all experiments were carried out using Google Tensor Processing Unit (TPU) acceleration, some dataset size restrictions were imposed. First of all, since TPU splits calculations across 8 cores, the batch size must be divisible by 8. To optimize training and validation speeds, it was decided to use a batch size of 1,024.

The TPU training process required batches to be full – this meant that the last batch must have been of size 1,024, even if there was no more data left. This restriction was overcome using two different strategies. For the training split, the last batch was artificially filled in by randomly sampling the same galaxies that were already selected. Regarding the validation split, it was decided to drop the last batch if it was not full. This meant that at most 1,023 galaxies were removed, making it at most approximately 0.3% of the split in the case of 46 – 54% split.

However, the last underfull validation split was discarded only during training. The final metrics reported in Section 5 were evaluated using the full validation split, without excluding any galaxies.

4.4.5 Environment

All experiments were performed using Tensorflow [1], version 2.12.0 environment, running on Google's tensor processing units (TPU), either V2-8 or V3-8, depending on the resource availability at the training time.

5 Results and discussion

After evaluating model prediction results, it was noticed that there were significant outliers in some folds, hence, instead of using the average, the median was chosen, as it is less susceptible to outliers. Afterwards, as an alternative to the standard deviation, the median absolute deviation (MAD) was used. Given a set X, where x_i are its elements, $\mathrm{MAD}(X)$ is calculated as follows:

$$\tilde{X} = \mathrm{median}(x_i), \qquad \mathrm{MAD}(X) = \mathrm{median}\big(|x_i - \tilde{X}|\big).$$

Since the classes were not evenly balanced (see Section 3.4), two main metrics will be studied – accuracy, which is the common metric, and F1, which is less susceptible to class imbalance.

The overall accuracy distribution of all experiments is shown in Fig. 2.

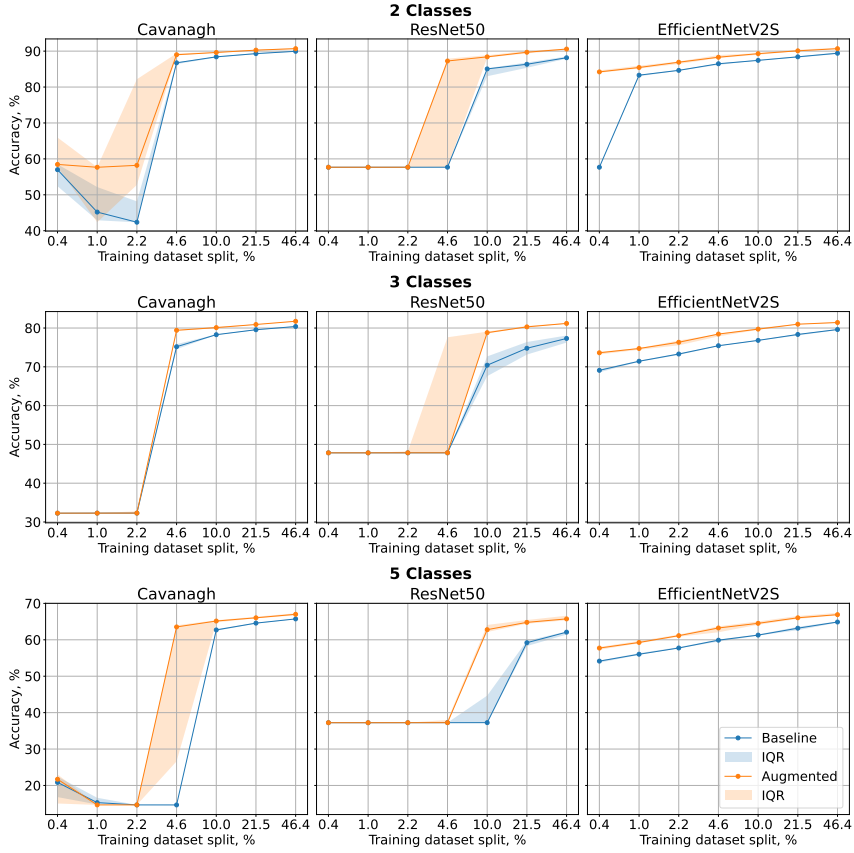

Figure 2. Overview of the accuracy distribution of all experiments. The shaded areas of the graph represent the interquartile range.

5.1 Model stability criteria

As the aim of the study is to evaluate model performance and stability when training models on a small amount of data, it is important to define the stability criteria. In this study, a model is considered stable if the following criteria are satisfied:

(i) The metric values have a narrow spread, as indicated by the median absolute deviations of the accuracy and F1-score metrics:

$$\mathrm{MAD}(\mathrm{Acc}) \leqslant 0.05 \cdot \mathrm{median}(\mathrm{Acc}), \tag{1}$$

$$\mathrm{MAD}(\mathrm{F1}) \leqslant 0.05 \cdot \mathrm{median}(\mathrm{F1}). \tag{2}$$

Here Acc and $\mathrm{F1}$ stand for the sets of accuracy and $\mathrm{F1}$ values of one experiment, and from all of its 10 cross-validation folds (see Section 4.4.1). A strict 5% tolerance margin is empirically chosen to ensure that the MAD across the 10 folds does not exceed 5% of the median model performance, signaling high model stability.

(ii) The model performs better than just assigning the majority class. This criterion is defined as follows:

$$\mathrm{median}(\mathrm{Acc}) > 1.05 \cdot \max(C_i). \tag{3}$$

Here C_i stands for the percentage of each class in the dataset, described in Section 3.4. Hence, $\max(C_i)$ is the percentage of the majority class: for the 2-class dataset, $\max(C_i) = 57.67\%$, for 3-class – 47.83%, and for 5-class – 37.23%. The criterion is made even stricter by applying an empirically chosen 5% margin to exclude the case when some folds managed to overcome overfitting, yet the majority of folds overfitted to predict the majority class. Such models are not considered stable.

In summary, definitions (1)–(3) mean that for the model to be considered stable, accuracy and $\mathrm{F1}$ metric values must have a narrow spread, and the model must achieve better accuracy than simply assigning the majority class to each galaxy.

The criterion (ii) will be further interpreted as overfitting of the model. Namely, assigning the majority class could be interpreted as one of the signs of overfitting – that the model learns the features of the training dataset too well, though failing to generalize and predict yet unseen data [18]. Furthermore, it could also be argued that it is the worst-case scenario of overfitting, as the model learns training data too much, failing to predict instances of minority classes et all.

5.1.1 Overfitting detection and prevention

Since one of the model stability criteria is overfitting, it was important to take measures to detect and prevent overfitting as much as possible.

Two types of measures were used – generic, applicable for any architecture, and architecture-specific means. One of the generic methods is data augmentation, described in Section 4.3. This is the only generic method studied in this paper which can impact model metrics, therefore, its effect was studied by conducting experiments both with and

without augmentations. The other generic method, applied in all of the experiments, is early stopping, further discussed in Section 4.4.3. It is used to preserve the best model weights before it starts to overfit. Finally, Monte Carlo cross-validation (see Section 4.4.1) allows to evaluate average model performance across multiple folds, which in turn helps detect if models overfit.

The aforementioned generic techniques were applied throughout the whole training process (with the exception of augmentations) to ensure that all models have the same basic setup that helps prevent overfitting. Nonetheless, each architecture could have implemented additional layers, which could further prevent overfitting. The first such layer is dropout, used by the Cavanagh architecture [6]. The dropout layer has one parameter – dropout probability – which describes the probability that incoming input is ignored (multiplied by 0) at random during the training phase [26]. The application of this layer helps to prevent co-adaptation where the model relies on specific connections too heavily [26]. Another layer is the batch normalization, used by Cavanagh and ResNet50 architectures [6, 11]. The use of this layer normalizes outputs of the previous layer, reducing the covariate shift, increasing training speed, and improving generalization [12].

5.2 2-class results

The detailed results of the 2-class dataset are presented in Table 4 and visualised in Fig. 2.

Table 4. Median accuracy and F1 values across 10 folds of each **2-class** dataset split. The values in parentheses denote median absolute deviation (MAD). The best values of both metrics for each split are highlighted in bold.

Model	Split	Accuracy, %		F1, %	
		Base	Augmented	Base	Augmented
Cavanagh	0.4%	56.99 (2.16)	**58.46** (1.29)	51.76 (6.17)	**73.16** (2.70)
	1.0%	45.19 (2.61)	**57.66** (13.82)	12.78 (10.95)	**70.91** (4.38)
	2.2%	42.37 (0.06)	**58.21** (8.56)	0.21 (0.21)	**73.35** (16.18)
	4.6%	86.75 (0.33)	**89.00** (0.35)	88.58 (0.37)	**90.55** (0.28)
	10.0%	88.41 (0.14)	**89.62** (0.05)	90.12 (0.04)	**91.10** (0.08)
	21.5%	89.30 (0.16)	**90.28** (0.09)	90.88 (0.10)	**91.60** (0.10)
	46.4%	89.95 (0.13)	**90.69** (0.14)	91.44 (0.13)	**91.96** (0.14)
ResNet50	0.4%	57.67 (0.00)	57.67 (0.00)	73.16 (0.00)	73.16 (0.00)
	1.0%	57.67 (0.00)	57.67 (0.00)	73.15 (0.00)	73.16 (0.00)
	2.2%	57.67 (0.01)	57.68 (0.01)	73.16 (0.01)	73.16 (0.01)
	4.6%	57.68 (0.01)	**87.25** (0.93)	73.16 (0.01)	**89.22** (0.72)
	10.0%	85.05 (0.56)	**88.41** (0.28)	87.22 (0.46)	**90.06** (0.29)
	21.5%	86.34 (0.54)	**89.69** (0.35)	88.11 (0.84)	**91.07** (0.29)
	46.4%	88.16 (0.31)	**90.58** (0.19)	89.85 (0.28)	**91.90** (0.13)
EffNetV2S	0.4%	57.67 (0.00)	**84.22** (0.35)	73.16 (0.00)	**86.64** (0.21)
	1.0%	83.31 (0.19)	**85.44** (0.48)	85.86 (0.22)	**87.60** (0.35)
	2.2%	84.64 (0.14)	**86.92** (0.32)	86.97 (0.21)	**88.80** (0.22)
	4.6%	86.48 (0.16)	**88.30** (0.60)	88.42 (0.14)	**90.03** (0.42)
	10.0%	87.42 (0.09)	**89.30** (0.31)	89.26 (0.07)	**90.77** (0.26)
	21.5%	88.41 (0.15)	**90.09** (0.16)	90.06 (0.22)	**91.55** (0.14)
	46.4%	89.41 (0.24)	**90.68** (0.24)	90.84 (0.17)	**91.94** (0.18)

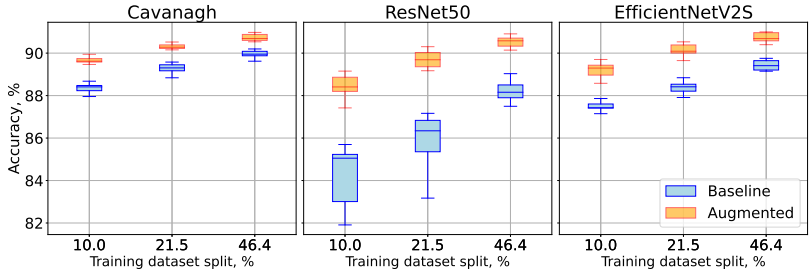

Figure 3. Accuracy distribution with and without augmentations (baseline) for larger splits of the **2-class** dataset.

The first tendency to be observed for the 2-class dataset is that the models of simpler architectures – Cavanagh and ResNet50 – tend to overfit when trained on the splits below 4.6%, as per the definition in Section 5.1, both with and without augmentations. The augmentations begin to help combat the overfitting only when using at least 40,022 images (corresponding to 4.6% of all the data). The most apparent difference is for ResNet50, where at this split it overfits without the use of augmentations, although after applying them, the model begins to meet the stability criteria defined in Section 5.1.

Only when at least 80,045 images are used, which corresponds to 10% of all available images, do models of all three architectures satisfy all stability criteria, both with and without augmentations. Further increasing the training dataset size improves the results even more, however, with diminishing returns.

Nonetheless, the model of more sophisticated EfficientNetV2S architecture is almost not affected by the overfitting problem. The only split where it overfits is 0.4%, which means that only 3,715 galaxies are used for training.

Another key observation is that once the model does not overfit, the impact of the augmentations becomes noticeable. In fact, starting from 10% dataset splits, the usage of augmentations is equivalent to the dataset size increase by at least two times, or in the case of ResNet50 and EfficientNetV2S – even four times: using 80,045 (10% split) galaxies with augmentations yields the same accuracy and F1 scores (within the margin of error) as using 368,206 (46.4% split) without augmentations. This is highlighted in Fig. 3.

5.3 3- and 5-class results

The same tendencies are observed when classifying galaxies into 3 or 5 classes. The accuracy and F1 scores are shown in Table 5.

First of all, for Cavanagh and ResNet50, the models still overfit when trained on 0.4–2.2% splits, regardless of augmentations. This stands true for both 3- and 5-class datasets. Furthermore, the variability of the 4.6% still persists. For 3 classes, Cavanagh fulfils stability requirements both with and without augmentations, though ResNet50 overfits. In the case of the 5-class dataset, Cavanagh overfits without the augmentations, however, it reaches stability with them. ResNet50 on the other hand, overfits regardless of augmentations. As is the case with the 2-class dataset, only when using 10% of the data results meet stability criteria, except for ResNet50, which still requires augmentations.

Table 5. Median accuracy and F1 values across 10 folds of each **3-class** dataset split (a) and each **5-class** dataset split (b). The values in parentheses denote median absolute deviation (MAD). The best values of both metrics for each split are highlighted in bold.

Model	Split	Accuracy, %		F1, %	
		Base	Augmented	Base	Augmented
			(a)		
Cavanagh	0.4%	32.27 (0.01)	32.28 (0.00)	16.31 (0.02)	16.27 (0.00)
	1.0%	32.29 (0.01)	32.28 (0.00)	16.27 (0.00)	16.27 (0.00)
	2.2%	32.28 (0.01)	32.29 (0.01)	16.27 (0.00)	16.27 (0.00)
	4.6%	74.92 (0.53)	**79.42** (0.17)	69.01 (0.96)	**73.23** (0.25)
	10.0%	78.28 (0.19)	**80.11** (0.22)	71.53 (0.34)	**73.90** (0.23)
	21.5%	79.63 (0.10)	**80.91** (0.03)	72.97 (0.49)	**75.28** (0.18)
	46.4%	80.44 (0.12)	**81.75** (0.05)	74.15 (0.15)	**76.20** (0.03)
ResNet50	0.4%	47.83 (0.00)	47.83 (0.00)	21.57 (0.00)	21.57 (0.00)
	1.0%	47.83 (0.00)	47.83 (0.00)	21.57 (0.00)	21.57 (0.00)
	2.2%	47.82 (0.00)	47.84 (0.02)	21.57 (0.00)	21.57 (0.01)
	4.6%	47.82 (0.02)	47.84 (0.03)	21.57 (0.00)	21.57 (0.01)
	10.0%	70.23 (1.98)	**78.81** (0.26)	64.74 (1.70)	**73.51** (0.22)
	21.5%	72.98 (0.88)	**80.30** (0.26)	67.65 (1.98)	**74.48** (0.83)
	46.4%	76.99 (1.00)	**81.19** (0.14)	69.87 (2.44)	**76.06** (0.32)
EffNetV2S	0.4%	69.09 (0.35)	**73.62** (0.26)	52.51 (1.99)	**64.88** (1.33)
	1.0%	71.44 (0.16)	**74.72** (0.23)	59.90 (1.13)	**67.47** (0.94)
	2.2%	73.30 (0.19)	**76.33** (0.49)	65.11 (0.52)	**69.65** (0.74)
	4.6%	75.48 (0.11)	**78.45** (0.48)	68.98 (0.57)	**72.05** (0.45)
	10.0%	76.80 (0.17)	**79.72** (0.22)	70.71 (0.38)	**73.63** (0.19)
	21.5%	78.38 (0.14)	**80.99** (0.07)	72.28 (0.49)	**75.54** (0.20)
	46.4%	79.62 (0.14)	**81.42** (0.15)	74.20 (0.02)	**75.97** (0.06)
			(b)		
Cavanagh	0.4%	20.85 (2.54)	**21.71** (2.34)	**11.44** (1.80)	9.91 (4.17)
	1.0%	**15.29** (0.63)	14.66 (0.01)	**7.58** (2.09)	5.13 (0.01)
	2.2%	14.66 (0.01)	14.66 (0.01)	5.12 (0.00)	5.11 (0.00)
	4.6%	14.65 (0.01)	**63.53** (0.74)	5.11 (0.00)	**53.29** (0.62)
	10.0%	62.71 (0.22)	**65.16** (0.30)	52.92 (0.34)	**54.84** (0.50)
	21.5%	64.58 (0.14)	**66.04** (0.20)	54.39 (0.45)	**56.31** (0.25)
	46.4%	65.74 (0.06)	**67.00** (0.30)	56.22 (0.24)	**57.92** (0.17)
ResNet50	0.4%	37.23 (0.00)	37.23 (0.00)	10.85 (0.00)	10.85 (0.00)
	1.0%	37.23 (0.01)	37.23 (0.01)	10.85 (0.00)	10.85 (0.00)
	2.2%	37.24 (0.00)	37.24 (0.00)	10.85 (0.00)	10.85 (0.00)
	4.6%	37.24 (0.01)	37.25 (0.03)	10.85 (0.00)	10.86 (0.01)
	10.0%	37.25 (0.05)	**62.78** (1.01)	10.86 (0.01)	**52.40** (1.48)
	21.5%	59.20 (0.79)	**64.79** (0.63)	49.40 (1.28)	**55.75** (0.94)
	46.4%	62.10 (0.64)	**65.74** (0.84)	50.08 (1.89)	**57.25** (0.63)
EffNetV2S	0.4%	54.15 (0.17)	**57.73** (0.36)	36.29 (2.61)	**45.19** (0.49)
	1.0%	56.04 (0.27)	**59.27** (0.40)	41.35 (1.44)	**48.90** (0.49)
	2.2%	57.75 (0.15)	**61.14** (0.30)	46.17 (1.03)	**51.55** (0.49)
	4.6%	59.89 (0.21)	**63.25** (0.56)	49.96 (0.17)	**53.36** (0.68)
	10.0%	61.29 (0.18)	**64.52** (0.48)	51.84 (0.52)	**54.91** (0.87)
	21.5%	63.19 (0.42)	**66.04** (0.44)	54.22 (0.44)	**56.98** (0.36)
	46.4%	64.88 (0.18)	**66.91** (0.40)	55.32 (0.31)	**57.99** (0.36)

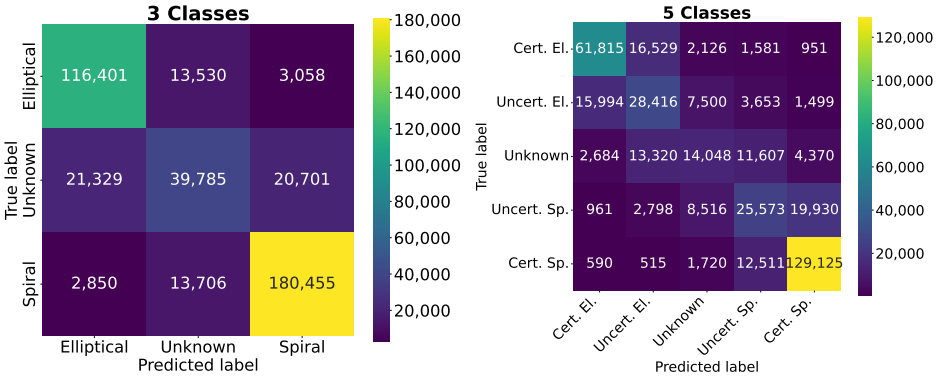

Figure 4. Confusion matrix examples of Cavanagh models, trained on **3-class** and **5-class** datasets, with augmentations and with 46.4% training split.

The usage of augmentations for 3- and 5-class datasets is also equivalent to doubling or even quadrupling the amount of training data (for example, ResNet50 with both 3- and 5-class datasets, and EfficientNetV2S with 3-class).

Another important observation is that the EfficientNetV2S displays stable results, however, contrary to 2-class dataset, with 3- and 5-class datasets, no overfitting is observed at all, even without the use of augmentations.

Finally, since the 3- and 5-class datasets contain uncertain classes, it was also important to evaluate the prediction distribution. As the Cavanagh model with the use of augmentations achieved the best results with both 3- and 5-class datasets. Examples of one fold from each dataset using augmentations and 46.4% training split are shown in Fig. 4.

In Fig. 4, it can be observed that the majority of misclassifications occur in neighbouring classes. This means that the models, in most cases, do not just assign an opposite incorrect class, as it would happen in the case of binary classification. Instead, models attribute a similar class, just to a lesser degree of certainty, when inspecting 5-class results. In the case of 3-class predictions, a model often attributes an unknown class to the galaxy. While this may seem like a misclassification, however, such galaxies could be further investigated either manually or by some other classifier, which could possibly even rule out these galaxies as being some other cosmological objects, such as a cluster of stars. On a large scale, such technique could prove useful, although more costly, ensuring that as many galaxies as possible are classified correctly.

Combining the results from all of the experiments, it is shown that a significant amount – at least 71,913 galaxies, corresponding to 10% of all available reliable 5-class data – is required to achieve stable prediction results of galaxy classes using a large validation dataset. With such an amount of data, all three architectures of different complexity, used in this paper, were able to meet the stability criteria defined in Section 5.1, except for ResNet50 on 5-class dataset, where augmentations were still required for stability. For the larger splits (21.5% and 46.4%), all models of all architectures showed stable results, regardless of augmentations.

Another important insight is that the usage of augmentations reduces the amount of data needed to achieve the same accuracy by two, or, in some cases, even four times (see Fig. 3 and Tables 4–5). Furthermore, in none of the experiments, augmentations hindered the model performances (within the MAD). Therefore, it is recommended to apply augmentations with any choice of architecture.

It is important to note that the aim of this paper was not to outperform other authors, therefore, the model accuracy and F1 scores were not compared to each other or highlighted. The aim was to prove that to reliably classify large datasets of galaxy images, especially with upcoming datasets of millions or even billions of galaxies, a model has to be trained on a substantial amount of data to achieve stable results.

The results provide a glimpse into the future trend as well – the upcoming projects are expected to contain hundreds of millions of galaxies (e.g., the Dark Energy Survey [29], consisting of 300 million galaxies [29]). The amount of reliably labeled data (whether classified by professionals or by such projects as Galaxy Zoo) will not scale at the same pace as the upcoming data. Therefore, the proportion of training data will approach lower splits that were studied in this paper. This means that at some point, if not enough training data is available at that time, the predictions of the models may become unreliable.

6 Conclusions

In this study, we provide the quantitative evaluation of the large-scale performance of models that were trained using small datasets. To preserve the diversity of galaxy images, we do not discard galaxies with an ambiguous shape, but rather we propose a novel classification methodology based on the SCI metric (see Section 3.2) – extending the standard binary classification, we group galaxies into 2, 3, and 5 classes by their SCI value, introducing either an unknown class in case of 3-class dataset or uncertain spiral/elliptical and unknown classes for 5-class dataset. This allows models to not be forced to pick just between two classes, but rather to pick neighboring classes of lesser certainty level or to assign an unknown class for further investigation. Based on the results, the following conclusions can be drawn:

(i) According to Tables 4–5 and Fig. 2, the recommended training dataset size threshold when using Galaxy Zoo data is at least 71,913 images for 5 classes, 76,262 for 3 classes, and 80,045 for 2 classes, all corresponding to 10% of the respective dataset. With this amount of data, all models of the three architectures studied achieve stability (as defined in Section 5.1), both with and without the use of augmentations. Even though below this threshold (specifically at 4.6% splits) models can still achieve similar accuracy and F1 results, the models are highly unstable, or they overfit (by definition explained in Section 5.1), hence they are considered unsuitable for large-scale classification.

(ii) According to Tables 4–5 and Fig. 3, for all of the models that were trained with the dataset size at or above the aforementioned threshold, applying standard augmentation techniques is equivalent to increasing the dataset size at least twice, or in the case of ResNet50, even up to 4 times. This property is observed in all 2-, 3-, and

5-class datasets. Additionally, in all of the experiments, augmentations did not have any adverse effect on the metrics (within the median absolute deviation range). This stands true even for the lower training splits. Therefore, it is recommended to use augmentation techniques, regardless of the chosen architecture.

(iii) Out of the three architectures studied, the EfficientNetV2 displayed the best performance (see Fig. 2). It remained stable with all datasets, except that it had overfitted only once, when using the smallest split – 0.4% (or 3,715 images) – of the 2-class dataset.

The aforementioned conclusions provide guidelines for research working with Galaxy Zoo data on how much data is required to train stable models that could be used for large-scale galaxy image classification.

In addition, with an ever-increasing amount of data from the new telescopes, the amount of labelled data will lag behind the vast amount of new photographs of galaxies. The results of this study provide an insight that as the gap between labelled (training) and unlabelled (validation) data grows, the overall reliability of the current models is likely to decrease, and that the more lenient classification methodology based on SCI could be used to minimise misclassification.

As for future research, it would be beneficial to study the impact of each augmentation technique separately, further expanding on our previous study [19]. Additionally, different parameters of the augmentations (such as noise distribution function and scale, zoom factor, and rotation angles) or colour and saturation-based augmentations could be explored as well. Finally, model performance could be evaluated on an even larger scale if the omitted data from GZ DESI [30] would be included using the same data filtering process, although, adjustments would have to be made in terms of how SCI is calculated.

Author contributions. All authors (To.M., A.V.M.M., and Ta.M.) have contributed as follows: methodology, To.M., A.V.M.M., and Ta.M.; formal analysis, To.M., A.V.M.M., and Ta.M.; software, To.M.; validation, To.M., A.V.M.M., and Ta.M.; writing – original draft preparation, To.M.; writing – review & editing, A.V.M.M. and Ta.M. All authors have read and approved the published version of the manuscript.

Conflicts of interest. The authors declare no conflicts of interest.

Acknowledgment. Experiments conducted in this paper were done using Google Cloud and funded by research credits provided by Google LLC, credit ID 360427050. The credits were used to perform model tuning using their Cloud TPU V2-8 and V3-8 resources.

References

1. M. Abadi, P. Barham, J. Chen, and others, TensorFlow: A system for large-scale machine learning, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, USENIX Assoc., Berkeley, CA, 2016, pp. 265–283, https://doi.org/10.48550/arXiv.1605.08695.

2. P.H. Barchi, R.R. de Carvalho, R.R. Rosa, and others, Machine and deep learning applied to galaxy morphology – A comparative study, *Astron. Comput.*, **30**, 2020, ISSN 2213-1337, https://doi.org/10.1016/j.ascom.2019.100334.

3. D. Berrar, Cross-validation, in *Encyclopedia of Bioinformatics and Computational Biology*, Academic Press, Oxford, 2019, pp. 542–545, https://doi.org/10.1016/B978-0-12-809633-8.20349-X, https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

4. C.R. Bom, A. Cortesi, U. Ribeiro, and others, An extended catalogue of galaxy morphology using deep learning in southern photometric local universe survey data release 3, *Mon. Not. R. Astron. Soc.*, **528**(3):4188–4208, 2023, https://doi.org/10.1093/mnras/stad3956.

5. J. Cao, T. Xu, Y. Deng, L. Deng, M. Yang, Z. Liu, W. Zhou, Galaxy morphology classification based on Convolutional vision Transformer (CvT), *Astron. Astrophys.*, **683**:A42, 2024, https://doi.org/10.1051/0004-6361/202348544.

6. M.K. Cavanagh, K. Bekki, B.A. Groves, Morphological classification of galaxies with deep learning: Comparing 3-way and 4-way CNNs, *Mon. Not. R. Astron. Soc.*, **506**(1):659–676, 2021, https://doi.org/10.1093/mnras/stab1552.

7. T.-Y. Cheng, C. J. Conselice, A. Aragón-Salamanca, and others, Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging, *Mon. Not. R. Astron. Soc.*, **493**(3):4209–4228, 2020, https://doi.org/10.1093/mnras/staa501.

8. S. Dieleman, K.W. Willett, J. Dambre, Rotation-invariant convolutional neural networks for galaxy morphology prediction, *Mon. Not. R. Astron. Soc.*, **450**(2):1441–1459, 2015, https://doi.org/10.1093/mnras/stv632.

9. M. Fukugita, S. Okamura O. Nakamura, and others, A catalog of morphologically classified galaxies from the sloan digital sky survey: North Equatorial Region, *Astron. J.*, **134**(2):579–593, 2007, https://doi.org/10.1086/518962.

10. A. Gupta, K. Kaur, N. Jindal, Predicting galaxy morphology using attention-enhanced ResNets, *Earth Sci. Inf.*, **17**:5335—5346, 2024, https://doi.org/10.1007/s12145-024-01449-6.

11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Piscataway, NJ, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

12. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on Machine Learning*, Proc. Mach. Learn. Res., Vol. 37, JMLR.org, 2015, p. 448–456, https://doi.org/10.48550/arXiv.1502.03167.

13. M. Jiménez, M. Torres Torres, R. John, I. Triguero, Galaxy image classification based on citizen science data: A comparative study, *IEEE Access*, **8**:47232–47246, 2020, https://doi.org/10.1109/ACCESS.2020.2978804.

14. N.E. Khalifa, M. Hamed Taha, A.E. Hassanien, I. Selim, Deep galaxy V2: Robust deep convolutional neural networks for galaxy morphology classifications, in *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, IEEE, Piscataway, NJ, 2018, pp. 1–6, https://doi.org/10.1109/ICCSE1.2018.8374210.

15. D. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, https://doi.org/10.48550/arXiv.1412.6980.

16. C.J. Lintott, K. Schawinski, A. Slosar, and others, Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey, *Mon. Not. R. Astron. Soc.*, **389**(3):1179–1189, 2008, https://doi.org/10.1111/j.1365-2966.2008.13689.x.

17. A. Mittal, A. Soorya, P. Nagrath, D. Jude Hemanth, Data augmentation based morphological classification of galaxies using deep convolutional neural network, *Earth Sci. Inf.*, **13**:601–617, 2020, ISSN 1865-0473, https://doi.org/10.1007/s12145-019-00434-8.

18. O.A. Montesinos López, A. Montesinos López, J. Crossa, Overfitting, model tuning, and evaluation of prediction performance, in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer, Cham, 2022, pp. 109–139, https://doi.org/10.1007/978-3-030-89010-0_4.

19. T. Mūžas, A.V. Misiukas Misiūnas, T. Meškauskas, Large scale study of binary galaxy image classification and the impact of image augmentation techniques, in *Computational Science and Its Applications – ICCSA 2023*, Volume 13957, Springer, Berlin, Heidelberg, 2023, pp. 402–412, https://doi.org/10.1007/978-3-031-36808-0_27.

20. T. Mūžas, A.V. Misiukas Misiūnas, T. Meškauskas, Evaluation of domain-specific convolutional neural network layers for galaxy classification tasks, *Nonlinear Anal. Model. Control*, **31**(3): 544–561, 2026, https://doi.org/10.15388/namc.2026.31.45923.

21. P.B. Nair, R.G. Abraham, A catalog of detailed visual morphological classifications for 14,034 galaxies in the Sloan Digital Sky Survey, *Astrophys. J. Suppl. Ser.*, **186**(2):427–456, 2010, https://doi.org/10.1088/0067-0049/186/2/427.

22. M. Nieto-Santisteban, A. Szalay, J. Gray, ImgCutout, an engine of instantaneous astronomical discovery, in *Astronomical Data Analysis Software and Systems (ADASS) XIII*, Astronomical Society of the Pacific, San Francisco, CA, 2004, pp. 666–670.

23. A. Sandage, *The Hubble Atlas of Galaxies*, Carnegie Inst. Washington Publ., No. 619, Carnegie Institution of Washington, Washington, 1961.

24. A. Sandage, J. Bedke, *The Carnegie Atlas of Galaxies, Vol. 638*, Carnegie Institution of Washington, Washington, 1994.

25. B.D. Simmons, C. Lintott, K.W. Willett, and others, Galaxy Zoo: Quantitative visual morphological classifications for 48 000 galaxies from CANDELS, *Mon. Not. R. Astron. Soc.*, **464**(4): 4420–4447, 2016, https://doi.org/10.1093/mnras/stw2587.

26. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov;, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, **15**(56):1929–1958, 2014, http://jmlr.org/papers/v15/srivastava14a.html.

27. M. Tan, Q.V. Le, EfficientNetV2: Smaller models and faster training, in *Proceedings of the 38th International Conference on Machine Learning*, Proc. Mach. Learn. Res., Vol. 139, JMLR.org, 2021, pp. 10096–10106, https://doi.org/10.48550/arXiv.2104.00298.

28. R. Urechiatu, M. Frincu, Improved galaxy morphology classification with convolutional neural networks, *Universe*, **10**:230, 2024, https://doi.org/10.3390/universe10060230.

29. T.N. Varga, D. Gruen, S. Seitz, and others, Synthetic galaxy clusters and observations based on Dark Energy Survey Year 3 Data, *Mon. Not. R. Astron. Soc.*, **509**(4):4865–4885, 2021, https://doi.org/10.1093/mnras/stab3269.

30. M. Walmsley, T. Géron, S. Kruk, and others, Galaxy Zoo DESI: Detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys, *Mon. Not. R. Astron. Soc.*, **526**(3):4768–4786, 2023, https://doi.org/10.1093/mnras/stad2919.

31. M. Walmsley, C. Lintott, T. Géron, and others, Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies, *Mon. Not. R. Astron. Soc.*, **509**(3):3966–3988, 2021, https://doi.org/10.1093/mnras/stab2093.

32. K.W. Willett, M.A. Galloway, S.P. Bamford, and others, Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging, *Mon. Not. R. Astron. Soc.*, **464**(4):4176–4203, 2016, https://doi.org/10.1093/mnras/stw2568.

33. K.W. Willett, C.J. Lintott, SP. Bamford, and others, Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey, *Mon. Not. R. Astron. Soc.*, **435**(4):2835–2860, 2013, https://doi.org/10.1093/mnras/stt1458.

34. Q.-S. Xu, Y.-Z. Liang, Monte Carlo cross validation, *Chemom. Intell. Lab. Syst.*, **56**(1):1–11, 2001, https://doi.org/10.1016/S0169-7439(00)00122-2.

35. D.G. York, J. Adelman, J.E. Anderson, and others, The Sloan Digital Sky Survey: Technical summary, *Astron. J.*, **120**(3):1579–1587, 2000, https://doi.org/10.1086/301513.

36. X.-P. Zhu, J.-M. Dai, C.-J. Bian, Y. Chen, S. Chen, C. Hu, Galaxy morphology classification with deep convolutional neural networks, *Astrophys. Space Sci.*, **364**(4):55, 2019, https://doi.org/10.1007/s10509-019-3540-1.