Mitigating bias in deep learning: training unbiased models on biased data for the morphological classification of galaxies

Esteban Medina-Rosales [1,2] Guillermo Cabrera-Vives [1,2,3,4★] and Christopher J. Miller [5]

[1] *Department of Computer Science, Universidad de Concepción, Edmundo Larenas 219, 4070409, Concepción, Chile*
[2] *Data Science Unit, Universidad de Concepción, Edmundo Larenas 310, 4070415, Concepción, Chile*
[3] *Millennium Nucleus on Young Exoplanets and their Moons (YEMS), Chile*
[4] *Millenium Institute of Astrophysics (MAS), Av. Vicuña Mackenna 4860, 8331150, Macul, Santiago, Chile*
[5] *Department of Astronomy and Department of Physics, University of Michigan, 1085 S. University, Ann Arbor, MI 48109, USA*

Reporter: Yu Huajian

2024.06.21

Outline

1. Introduction

2. METHODOLOGY

3. RESULTS

4. CONCLUSIONS

1 Introduction

- Galaxies and their properties have been studied for over a century.
- For some time, visual classification played the dominant role in galaxy morphologies, either done by expert astronomers or through crowdsourcing systems such as Galaxy Zoo. Recently, new data sets of morphologically classified galaxies obtained using diverse machine learning methods have been published.
- The issue is that it has been demonstrated that human-generated labels are prone to bias in terms of observable parameters.

2 METHODOLOGY

2.1 Incorporating bias into the likelihood

Data Set and Bias Definition: $D = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$

Establishing the Classification Task: $f_w : X \to Y$

Model Likelihood Expression: Using a bias parameter α (such as image resolution), compute the likelihood of the entire data set.

$p(D|w, \{\alpha_i\}_{i=1}^{N})$ The likelihood is represented as a joint distribution of all data points, depending on the model parameters w and the bias parameter α_i of each data point.

Derivation of the Likelihood Formula: The likelihood of each data point is calculated by marginalizing the true labels y_i taking into account the bias from the true label to the observed label. Specifically, decompose $p(\tilde{y}_i|x_i, w, \alpha_i)$ into $p(\tilde{y}_i|y_i, \alpha_i)p(y_i|x_i, w)$

Where $p(\tilde{y}_i|y_i, \alpha_i)$ represents the probability of observing the biased label given the true label and the bias parameter.

$$p(\mathcal{D}|\mathbf{w}, \{\alpha_i\}_{i=1}^{N}) = \prod_{i=1}^{N} p(\tilde{y}_i|\mathbf{x}_i, \mathbf{w}, \alpha_i), \quad (1)$$

$$= \prod_{i=1}^{N} \sum_{y_i} p(\tilde{y}_i, y_i|\mathbf{x}_i, \mathbf{w}, \alpha_i), \quad (2)$$

$$= \prod_{i=1}^{N} \sum_{y_i} p(\tilde{y}_i|y_i, \alpha_i) p(y_i|\mathbf{x}_i, \mathbf{w}), \quad (3)$$

De-biasing Loss Function: $\mathcal{L} = -\log p(\mathcal{D}|\mathbf{w}, \{\alpha_i\}_{i=1}^{N}).$ (4)

2 METHODOLOGY

2.2 Modelling galaxy morphology labelling bias

Defining Labels: The galaxy classification used in the study is based on data from the Galaxy Zoo 2 (GZ2) project.

Modelling Bias: The author introduces a bias parameter α, which represents the bias in each observation. This parameter specifically addresses the resolution issues in galaxy images observed

Mathematical Expression of Bias: In the model, labelling bias is expressed through a conditional probability formula, specifically:

$$P(\tilde{y} = \text{smooth} \mid y = \text{disc}, \alpha) = e^{-\alpha^2/(2\theta^2)}$$

For smooth galaxies, the model assumes they are always correctly labelled as smooth:

$$P(\tilde{y} = \text{smooth} \mid y = \text{smooth}, \alpha) = 1$$

and:

$$P(\tilde{y} = \text{disc} \mid y = \text{smooth}, \alpha) = 0$$

2 METHODOLOGY

2.3 Implementing the de-biasing

We now consider how to implement the de-biasing technique within two different classification models using data that have already been classified into smooth/disc categories by humans.

The first method uses 1D catalogue data for each galaxy to enable a classification. While these catalogue data were measured from the 2D imaging data, the pixels themselves are not directly used in the classification.

The second classification model is based on Deep Learning techniques and requires only the 2D galaxy imaging data (the pixels).

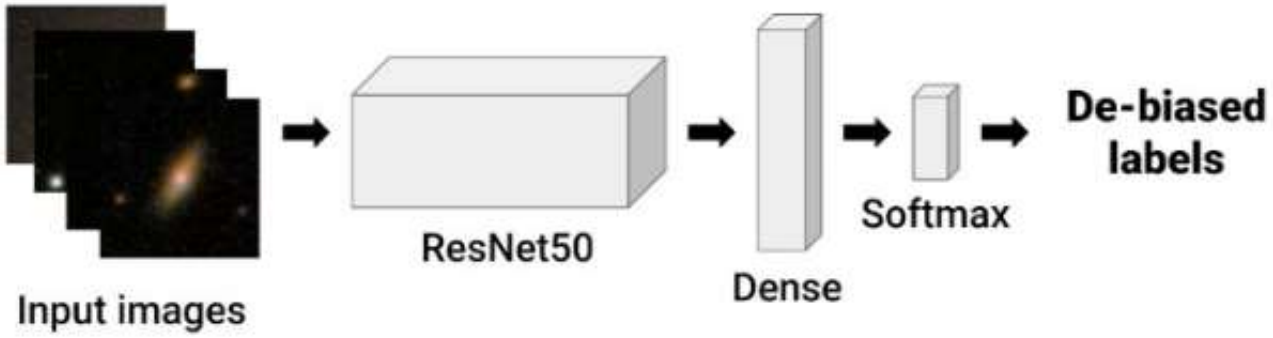


Figure 2. Architecture of our deep learning de-biasing model. We used a ResNet50 with an extra dense layer and JPEG images as input. We used the negative log-likelihood described in Section 2.1 as our loss function.

2 METHODOLOGY
2.4 Quantifying the bias

For the purposes of this work, that is, a binary classification problem with only one observable property, the deviation of the class fraction for a specific class can be expressed as

$$\sigma_k = \sqrt{\frac{1}{N_A} \sum_{l=1}^{N_A} (\tilde{f}_{l,k} - f_k)^2}$$

Then, the deviation of the class fractions over the entire data set can be quantified by

$$V_{14} = \sqrt{\frac{1}{N_K} \sum_{k} \sigma_k^2}$$

2 METHODOLOGY
2.5 The data

Label Assignment:
For all our experiments we used galaxies contained in the GZ2 catalogue (Willett et al. 2013).
We used majority voting according to the weighted vote fractions to define galaxies as smooth or disc, corresponding to our previously defined labels, and discarded objects that the majority of participants classified as ‘star or artefact’. To ensure a well-balanced data set, we conducted a random sampling of 121 984 galaxies from which 62 225 corresponded to smooth and 59 759 corresponded to disc.

Dataset Division：
In order to train the models, we arranged the galaxies into different sets: a training set consisting of 97 600 galaxies, a validation set with 12 192 galaxies, and a test set containing 12 192 galaxies. As input for the models, we used the original JPEG images labelled by the annotators. Following Dieleman et al. (2015), we cropped the original images from 424 × 424 to 207 × 207 pixels.

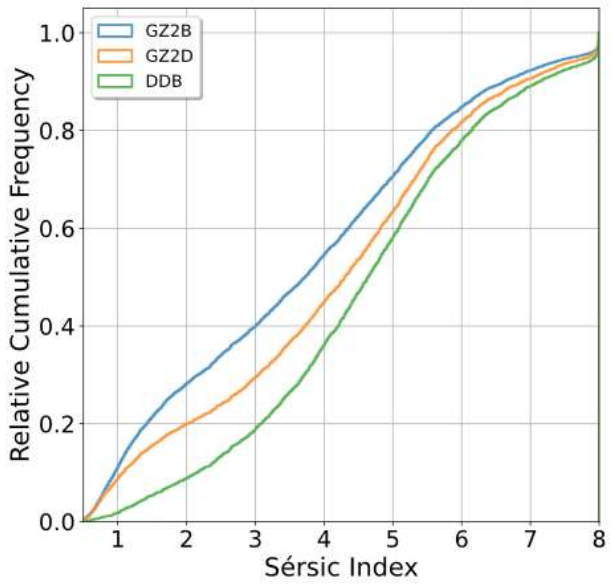
3 RESULTS
3.1 Bias measurement

In Table 1 we present the quantified bias as described in Section 2.4. We first measure the bias of the labels from GZ2, both the original crowdsourced labels (GZ2B) and the de-biased labels from Hart et al. (2016) (GZ2D).

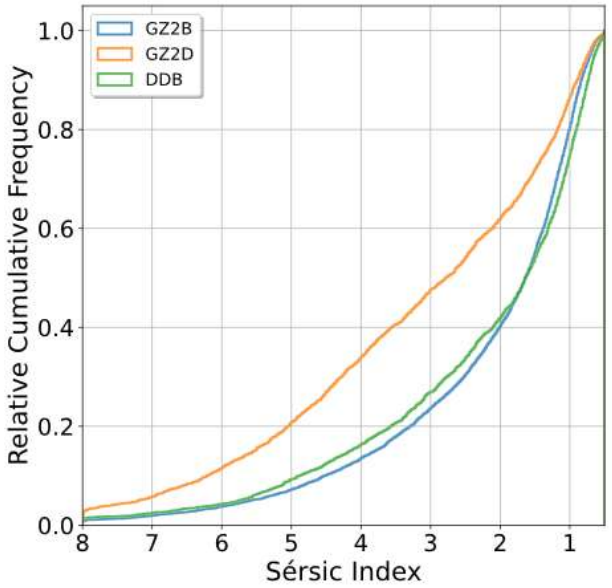
Table 1. Biases for the different data sets and models implemented.

Data set/Method	Bias CV18
(a) GZ2B (Willett et al. 2013)	0.3696 ± 0.0095
(b) GZ2D (Hart et al. 2016)	0.3106 ± 0.0108
(c) ResNet50 over GZ2B	0.3781 ± 0.0091
(d) ResNet50 over GZ2D	0.3258 ± 0.0107
(e) DCV14 over GZ2B	0.2994 ± 0.0102
(f) DDB (ours) over GZ2B	$\mathbf{0.2866 \pm 0.0112}$

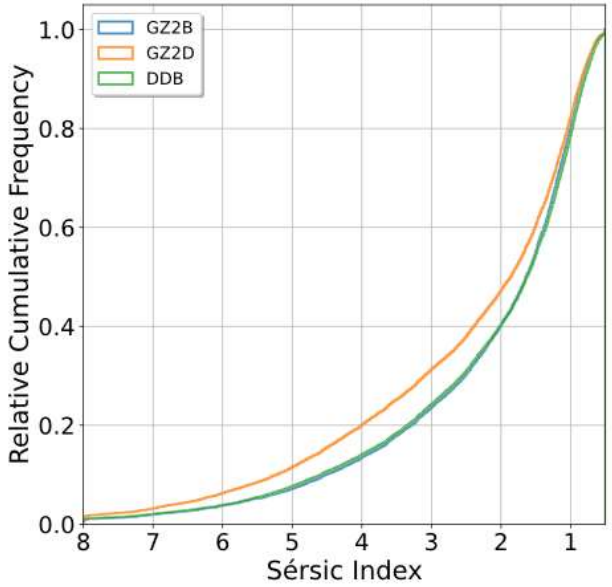
3 RESULTS
3.2 Sérsic profiles analysis



(a) Galaxies classified as smooth



(b) Galaxies changed from smooth to disk



(c) Galaxies classified as disk

Results

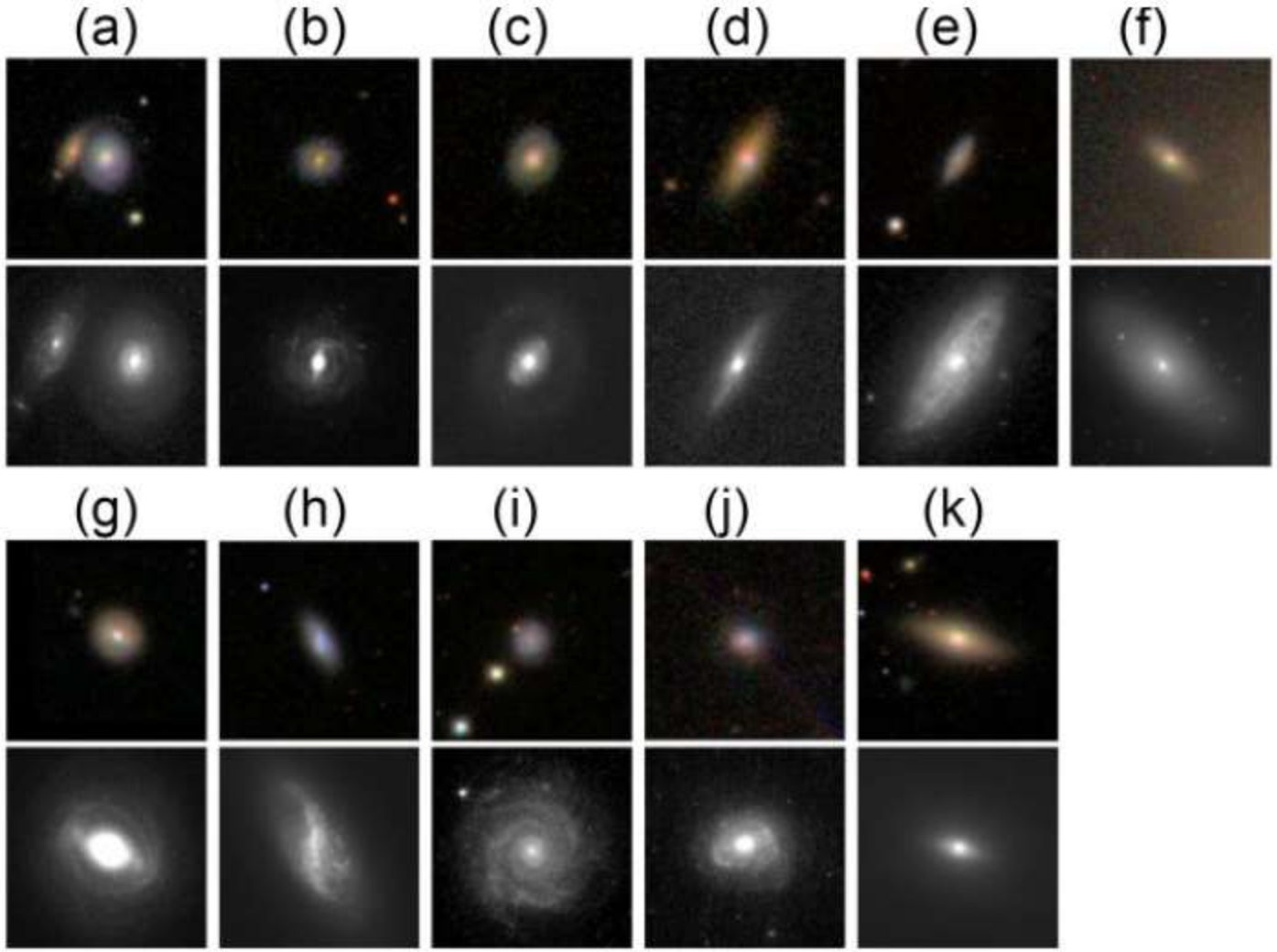
1.Smooth Galaxies:

1. For smooth galaxies, the Sérsic index distribution is more concentrated after de-biasing, indicating that the de-biasing process effectively reduces systematic biases in the original data.
2. The increased concentration of the Sérsic index shows that the de-biased labels better reflect the true light distribution of smooth galaxies.

2.Disk Galaxies:

1. For disk galaxies, the Sérsic index distribution also shows significant improvement after de-biasing.
2. After de-biasing, the Sérsic index distribution for disk galaxies is more consistent, reducing the biases present in the original data.

3 RESULTS
3.3 Visual inspection of high-resolution images



3 RESULTS
3.4 GradCAM

Aiming to understand the difference in the ResNet50 results before and after including the de-biasing model, we employ GradCAM (Selvaraju et al. 2020), a technique for visualizing the decision-making process of a convolutional neural net like ResNet50. GradCam generates a localization map that accentuates the relevant regions of the image used when predicting a specific class.

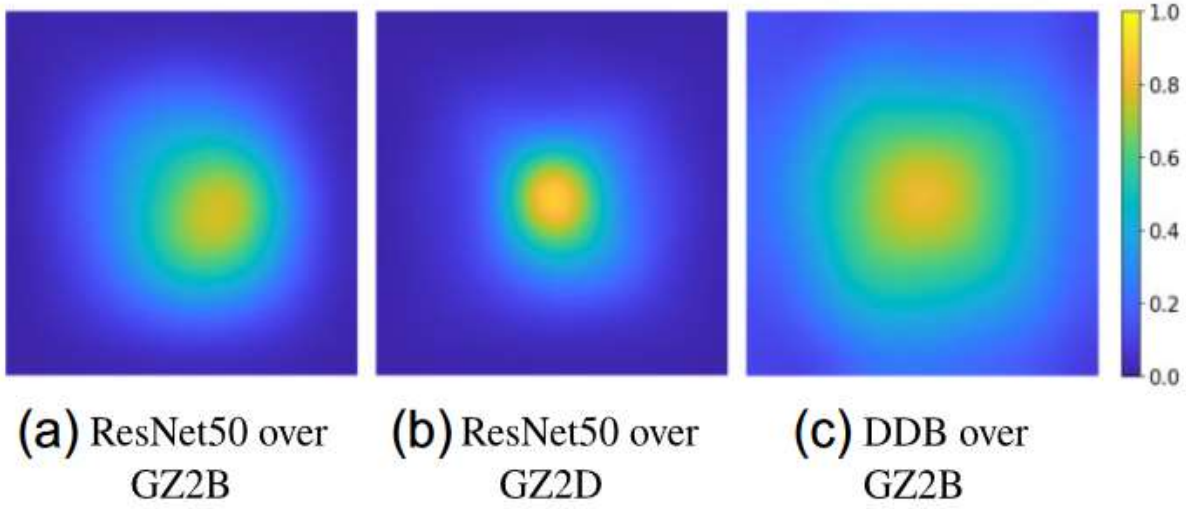


Figure 5. Average GradCAM localization maps for the disc class. Highlighted in yellow are the class-discriminative regions. Notice how (a) and (b) focus on the galaxy bulge, while (c) considers the entire galaxy.

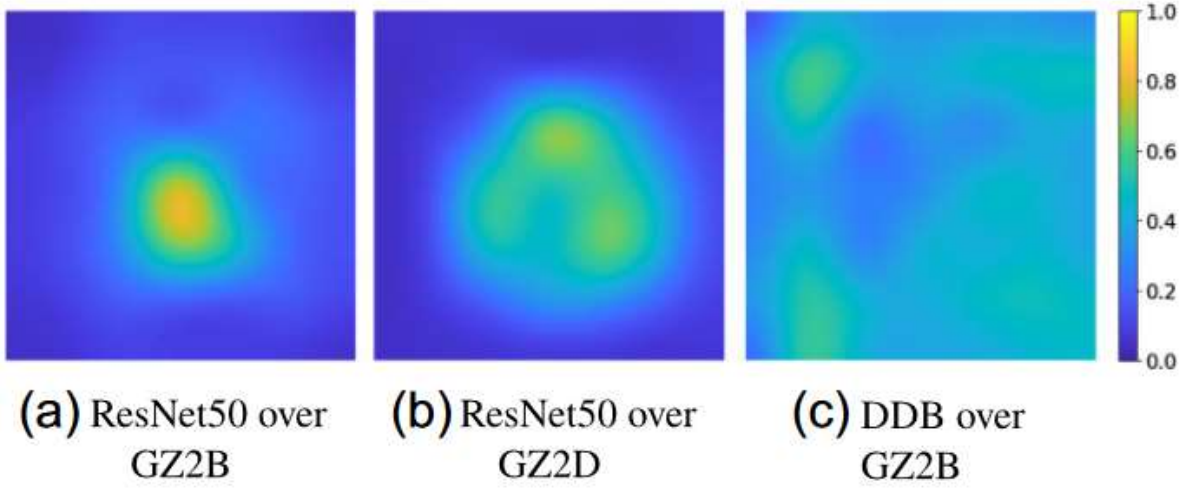
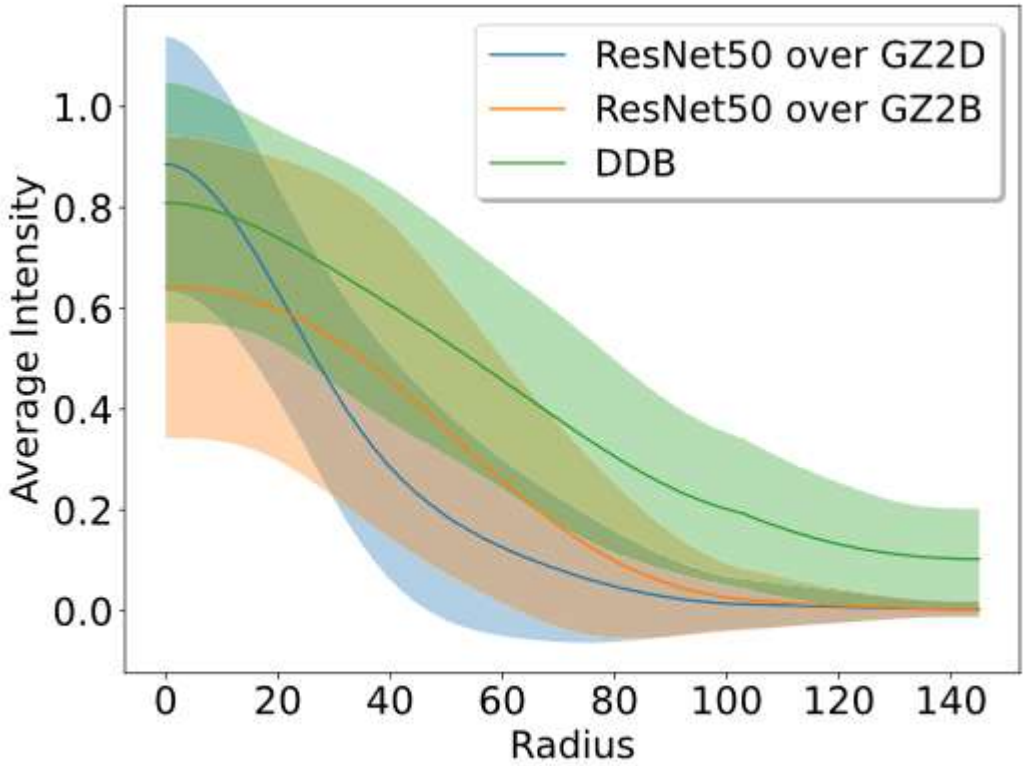
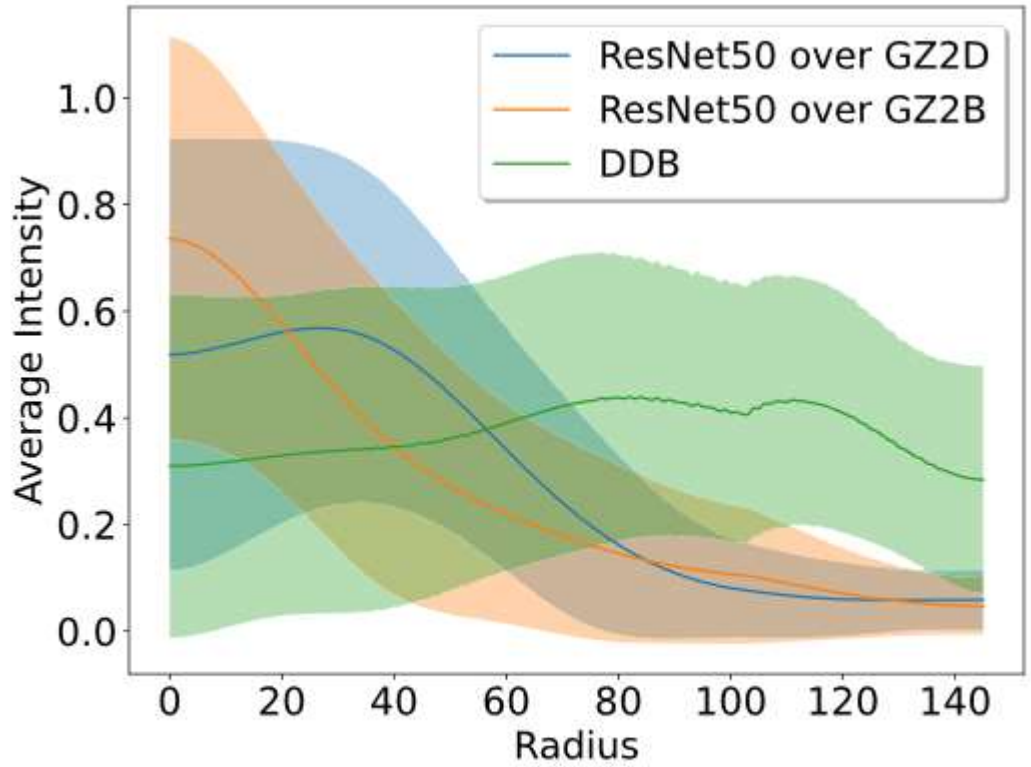


Figure 6. Average GradCAM localization maps for the smooth class. Highlighted in yellow are the class-discriminative regions. Notice how (a) focuses on the bulge of the galaxy, (b) on the regions around the bulge, and (c) on the edges of the galaxy.

3 RESULTS
3.4 GradCAM



(a) Disk



(b) Smooth

Figure 7. Average intensity versus radial distance from the centre of the image. (a) corresponds to the average GradCAM maps for the disc class (Fig. 5) and (b) to the average GradCAM maps for the smooth class (Fig. 6).

4 CONCLUSIONS

1.Data can exhibit biases in terms of observable parameters, such as resolution, which can introduce bias during the human labelling process. This inherent bias, stemming from observable properties rather than annotators, leads to systematic labelling biases in the data.

2. We demonstrate that training these models directly on biased data results in biased models. We introduce a method for training deep learning models that takes into account this labelling bias, to obtain unbiased models even when training with biased data.

3. We employ a visualization technique to comprehend the effect of our method on the deep learning model, comparing it with models trained without utilizing our approach.

4. We conclude that using our method, we can directly train a deep learning model on the biased data and obtain a model capable of both de-biasing existing data sets and labelling new data.

Thanks