Can semi-supervised learning reduce the amount of manual labelling required for effective radio galaxy morphology classification?

Inigo V. Slijepcevic
Department of Physics and Astronomy
University of Manchester, UK
inigo.slijepcevic@postgrad.manchester.ac.uk

Anna M. M. Scaife[*]
Department of Physics and Astronomy
University of Manchester, UK
anna.scaife@manchester.ac.uk

Abstract

In this work, we examine the robustness of state-of-the-art semi-supervised learning (SSL) algorithms when applied to morphological classification in modern radio astronomy. We test whether SSL can achieve performance comparable to the current supervised state of the art when using many fewer labelled data points and if these results generalise to using truly unlabelled data. We find that an artifical SSL scenario provides additional regularisation and improves baseline test set accuracy for a range of labelled data volumes. However, using truly unlabelled data degrades accuracy, which we show may be due to class imbalance in the unlabelled data.

1 Introduction

The Fanaroff-Riley (FR) binary classification of radio galaxies has been widely adopted and utilised since its inception over four decades ago [10]. However, in spite of progress in relating the two classes to the dynamics and energetics of the sources [12, 21], our understanding of the causal relationship between the source physical properties/environment and FR classification is far from complete. To improve our inferences about the physics of these objects, we need to accurately identify and classify more radio galaxies while also detecting anomalous/rare examples [16].

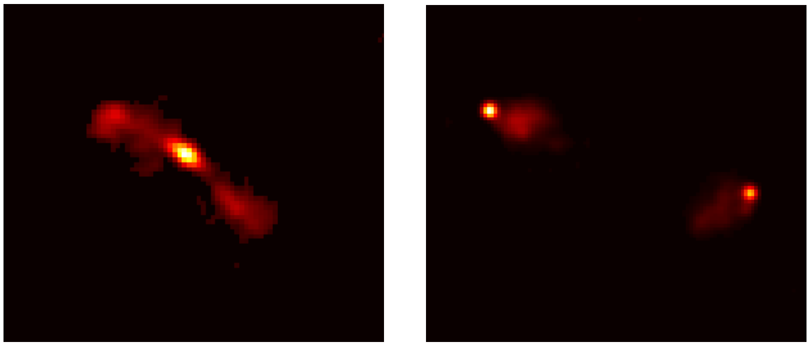
In particular for new sky surveys, such as those anticipated for the Square Kilometre Array (SKA) radio telescope, automated classification algorithms are increasingly being developed to replace the manual *by eye* approaches used historically. CNNs have been used with success for image-based classification of radio galaxies ([25, 1]), including attempts to use more novel techniques such as capsule networks [15] and attention gating [5] to help improve performance and interpretability. [2] provides a more comprehensive survey of current techniques.

Currently, archival data-sets for training radio galaxy classifiers are of comparable size to many of those used in computer vision (e.g. CIFAR [13]), with around 10^5 samples available. However, a fundamental difference is the sparsity of labels in radio galaxy data-sets: only a small fraction of data-points are labelled. This is largely due to the domain knowledge required for labelling, which incurs a high cost per label compared to typical machine learning data-sets. Currently, the largest labelled machine learning data-set of radio galaxies is MiraBest [17, 20] which has 1256 samples, orders of magnitude lower than the number of unlabelled images in its originating sky survey.

This work: In this work, we focus on morphological classification of radio-galaxies within the FR classification scheme, see Figure 1. We aim to test whether we can achieve performance comparable to the current supervised state of the art on the MiraBest dataset with many fewer labels by using

[*] The Alan Turing Institute, 96 Euston Rd, London, UK a.scaife@turing.ac.uk

Preprint. Under review.



(a) FR Class I

(b) FR Class II

Figure 1: Example radio galaxies from the MiraBest dataset (cherry-picked for prototypicity) [17].

semi-supervised learning (SSL). This work will guide decision making on how to construct pipelines for upcoming radio surveys such as those for the SKA.

Many SSL algorithms are successful on benchmarking datasets such as CIFAR-10 [13], MNIST [14] and SVHN [18] in the regime of low data ($\sim [10, 10^3]$ data points). However, less work has been done in assessing the robustness to various real world problems, such as unclean or covariate shifted data, new classes appearing in the unlabelled data or even simply varying sizes of labelled/unlabelled data: the shortcomings of the SSL literature have previously been documented in this context [19].

Of particular interest in astronomy is the problem of biased sampling for our training data. We observe phenomena driven by complex natural processes which are selected for labelling in a biased manner due to instrumental, observational and intrinsic effects which favour, for example, particular flux density (brightness) and redshift (distance) ranges. This makes it difficult to know how representative our data catalogues are of all observed data, and the consequent effects on model performance are therefore not always clear a priori. Specific examples of how this is being addressed for machine learning applications in astronomy include the use of Gaussian process modelling to augment training data and make it more representative of test data in photometric classification [4] and in galaxy merger classification where domain adaptation techniques have also been explored [8]. In both of these cases the solutions are tackling covariate shift between the labelled and test data.

2 FixMatch

Semi-supervised learning leverages unlabelled data when only a small labelled data-set is available. The model is fed a set of image-label pairs $(\boldsymbol{x}_l, y_l) \in \boldsymbol{X}_l$ along with a (usually larger) set of unlabelled images $\boldsymbol{x}_u \in \boldsymbol{X}_u$. The goal is to predict labels for held out samples $\boldsymbol{x}_{test} \in \boldsymbol{X}_{test}$. An overview of semi-supervised learning can be found in [7]

We choose the FixMatch algorithm [23] from the pool of SSL techniques as it achieves state of the art performance on benchmarking datasets, has few hyperparameters, and is computationally cheap [23]. FixMatch makes use of the unlabelled data through consistency regularisation and pseudo-labelling by adding a loss term computed on two different augmentations of the same image:

$$\mathcal{L} = \lambda \sum_{u=0}^{\mu B} \underbrace{\mathbb{1}(\max(p_m(\alpha(\boldsymbol{x}_u))) \geq \tau)}_{\text{threshold mask}} \times \underbrace{H(\hat{q}_u, p_m(y|\mathcal{A}(\boldsymbol{x}_u)))}_{\text{pseudo-label cross entropy}} + \sum_{l=0}^{B} \underbrace{H(y_l, p_m(y|\alpha(\boldsymbol{x}_l)))}_{\text{supervised loss}} . \tag{1}$$

Here the "weak" augmentation, denoted $\alpha(\cdot)$, retains the semantic meaning of the image and simply uses the standard rotation/flipping augmentations. The "strong" augmentation, denoted by $\mathcal{A}(\cdot)$, uses RandAugment [9] to apply a sequence of augmentations that can significantly alter the image.

If $\max(p_m(y|\alpha(\boldsymbol{x}_u))) > \tau$, the prediction $\hat{q}_u = \text{argmax}(p_m(y|\alpha(\boldsymbol{x}_u))$ is used as a pseudo-label for $\mathcal{A}(\boldsymbol{x}_u)$. Figure 2 shows how an unlabelled data point flows through the model. B is the labelled batch size, μ, τ are hyperparameters controlling unlabelled batch size and confidence threshold respectively. λ controls the weighting of the unlabelled loss term and is set to 1, following [23].

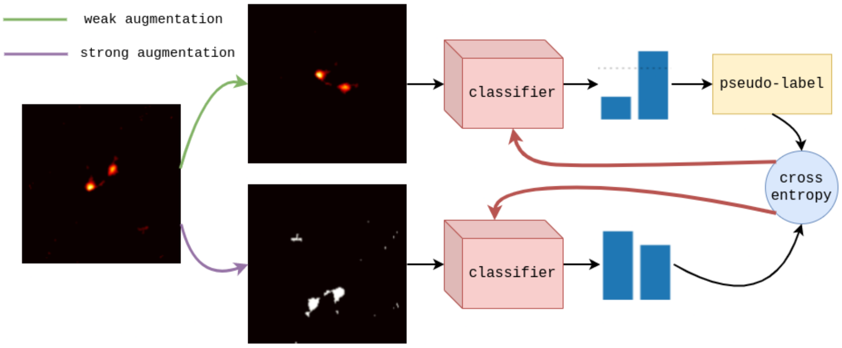


Figure 2: An unlabelled data point flowing through the FixMatch algorithm [23]

3 Experiments

3.1 Ensuring fair evaluation

Fair evaluation is a problem in the SSL literature [19], yet is crucial if we wish to apply it to real data. **Hyperparameters and validation set**: we keep the validation set a realistic size by scaling it to 20% of $\boldsymbol{X}_l$, but set a hard lower limit to produce meaningful results when $\boldsymbol{X}_l$ is small. We set the learning rate to 0.005 but scale the batch size, B, with $\boldsymbol{X}_l$. Astronomical data-sets vary in size and content and large validation sets may not be available: our models need to perform well across a range of scenarios without too much hyperparameter sensitivity. For these reasons, we keep hyperparameter tuning to a minimum, opting instead to choose reasonable values close to the optimal values in the computer vision literature [23]. We choose $\mu = 7$, which allows the model to use a large proportion of the unlabelled data in a single epoch, and $\tau = 0.95$, which allows the model's confidence on unlabelled data to pass the threshold before it begins to overfit. **Model architecture**: we use the convolutional architecture from [24], which has good performance for radio galaxy classification and relatively few parameters, giving a realistic baseline that isn't prone to overfitting. **Reproducibility and variance in results and data splits**: each experiment's results are averaged over 10 runs initialised using seeds 0-9. This ensures that the same splits are used during each experiment, while keeping weight initialisations consistent. We use randomly selected, stratified labelled/unlabelled data splits. The same test set is used after choosing the model weights with the best validation set accuracy.

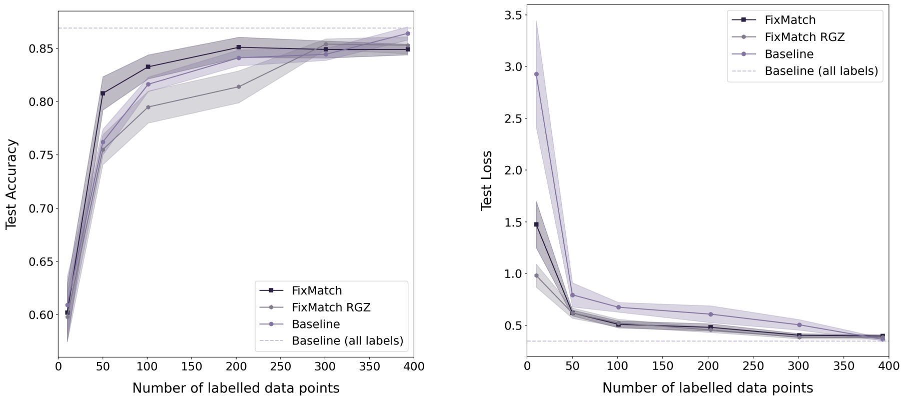
The experiments were performed on a single Nvidia A100 GPU with a total of 2.34 days of runtime. We used Weights & Biases [3] to track experiment results. Code can be found at `https://github.com/inigoval/fixmatch`.

3.2 Does FixMatch outperform the baseline?

Fully supervised baseline. The baseline model is trained in a fully supervised fashion with a cross-entropy loss, shown as the second term in Equation 1. We use the same subset of (weakly augmented) labelled samples as in the FixMatch case, but ignore the unlabelled samples.

Throwing away labels to create an artificial SSL scenario (Case A). We perform multiple experiments by using a small labelled subset of the MiraBest dataset [17] and using the remainder as unlabelled data. In all cases we keep the labelled data stratified (class balance is preserved). In Figure 3 we see that FixMatch achieves a consistently lower loss on the test data and outperforms the baseline in test set accuracy when there is little labelled data available. We are able to recover comparable accuracy ($85.1\% \pm 1.1\%$) to the supervised baseline with all labels ($86.93\% \pm 0.54\%$) using just 20% (203) of the labels. However, FixMatch's performance degrades quickly with very few labels, which is in stark contrast to similar experiments in [23], where good performance is achieved even with only one sample per class. Furthermore, the "sweet spot" where FixMatch has a significant advantage does not cover the full range of labelled data volumes.

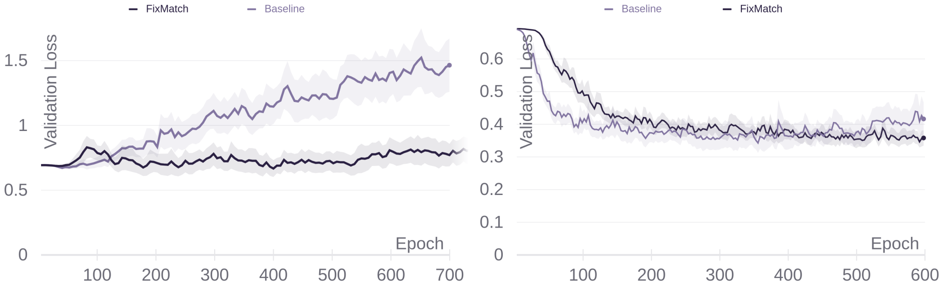
We hypothesise that there is a strong regularisation effect from the strongly augmented samples, as well as new information being learned from the unlabelled data. This is illustrated in Figure 4(a), where FixMatch avoids overfitting at low data volumes, as demonstrated by the well behaved



(a) Accuracy on the test set.

(b) Loss on the test set

Figure 3: Performance as a function of $\boldsymbol{X}_l$ size.



(a) Validation loss with 50 labelled samples.

(b) Validation loss with 393 labelled samples.

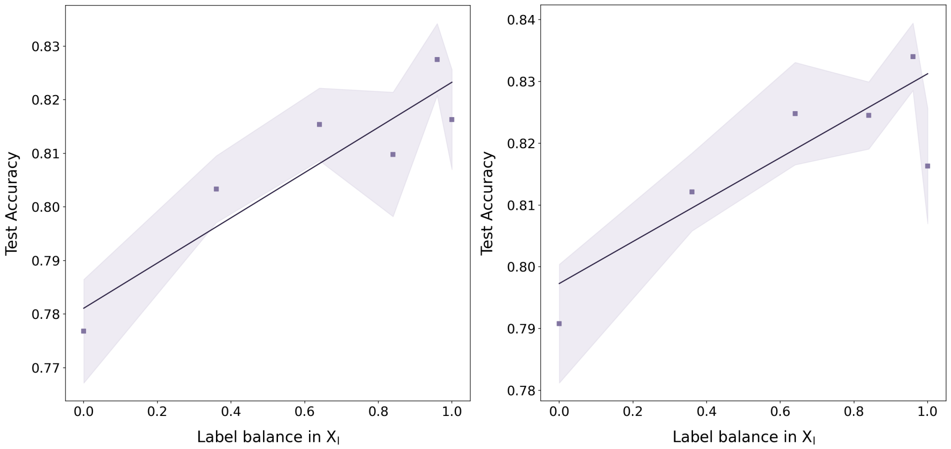
Figure 4: FixMatch (Case A) regularisation effect on validation loss. The shaded area shows the standard error over 10 runs. We use exponential moving average smoothing with a weight of 0.3.

validation loss. We also observe that this effect is not present with 393 labelled samples, which is reflected in the equal test loss at high label volumes in Figure 4(b).

Testing FixMatch on real unlabelled data (Case B). We test FixMatch by using a pool of 20,000 unlabelled data from the Radio Galaxy Zoo Data Release 1 (RGZ DR1) catalogue (Wong et al. in prep) with labelled samples from MiraBest. While these data originate from the same radio survey and are pre-processed in the same manner as MiraBest, the class balance of the RGZ DR1 catalogue is unknown and the choice of filters for choosing data-points is wider, resulting in a dataset of $\sim 10^5$ datapoints. We find that the RGZ DR1 data-set contains many unresolved sources, which are uninformative to our model. To remove these sources, we enforce a lower limit on source extension by calculating the Frechet Distance (FD; [11]) between the labelled data-set and the RGZ DR1 dataset (in feature space) for a range of lower limits. We take the limit with the lowest score (28 arcsec), after which a total of 8848 unlabelled data samples remain.

We find that in Case B FixMatch still provides regularisation as shown by the comparable test set loss to Case A, see Figure 3(b). However, this does not result in an improvement in accuracy. Indeed we see a decrease in test set accuracy compared to the baseline in Case B, for all labelled data volumes. We hypothesise that this may be due to the unknown class imbalance in the RGZ DR1 data-set.

To test this hypothesis we rerun Case A with an artificially class imbalanced unlabelled data-set, produced by removing either FRI or FRII samples from the unlabelled data. We measure test set accuracy for unlabelled data with varying proportion β of FRIs. Figure 5 shows that class imbalance has a significant negative effect on test set accuracy. This suggests that the performance gap between Cases A and B could be accounted for by class imbalance in RGZ DR1, although it is not conclusive that this is the only factor.



(a) Removing FRI samples

(b) Removing FRII samples.

Figure 5: Test set accuracy as a function of label balance in the unlabelled (MiraBest) data-set. We use only samples qualified as confidently classified in the MiraBest dataset for the labelled subset (69 labels) to reduce the noise in our results, see [22] for more details on this qualification. Label balance is quantified by computing $4(1-\beta)\beta$. Error bars show the standard error after aggregating 20 runs (seeded 0-19).

4 Conclusion

We find that FixMatch provides some regularisation benefits when learning with few labelled samples, mitigating the effect of overfitting, as well as learning from unlabelled data. Furthermore, we are able to achieve better accuracy on the test set than the baseline with fewer labels. While this is relatively promising, the improvement in accuracy is much smaller than for the computer vision data-sets the algorithm was initially tested on.

Poor results using the "real" RGZ DR1 data highlight an important obstacle to applying SSL "in the wild" on scientific observational data: $\boldsymbol{X}_l$ and $\boldsymbol{X}_u$ are unlikely to be drawn from identical distributions. We see that even class imbalance in $\boldsymbol{X}_{\mathrm{u}}$ can have a major effect on SSL performance, noting that in a real scenario we have no way of knowing the class balance of $\boldsymbol{X}_{\mathrm{u}}$.

We believe that a naive application of SSL, although it may outperform the baseline in some cases and provide useful regularisation, requires further domain specific development to give a worthwhile advantage in the case of radio galaxy morphology classification. Future work will consider handcrafting radio galaxy specific augmentations and implementing more sophisticated SSL approaches to help tackle the problem of covariate shift between $\boldsymbol{X}_{\mathrm{u}}$ and $\boldsymbol{X}_{\mathrm{l}}$, such as pretraining and/or incorporating domain adaptation (e.g. [6]).

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge support from the UK Alan Turing Institute under grant reference EP/V030302/1. IVS gratefully acknowledges support from the Frankopan Foundation.

This work has been made possible by the participation of more than 12,000 volunteers in the Radio Galaxy Zoo Project. The data in this paper are the result of the efforts of the Radio Galaxy Zoo volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at `http://rgzauthors.galaxyzoo.org`.

References

[1] A. K. Aniyan and K. Thorat. Classifying Radio Galaxies with the Convolutional Neural Network. *The Astrophysical Journal Supplement Series*, 230(2), 2017.

[2] B. Becker, M. Vaccari, M. Prescott, and T. Grobler. CNN architecture comparison for radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 503(2):1828–1846, 3 2021.

[3] L. Biewald. Experiment Tracking with Weights and Biases, 2020.

[4] K. Boone. Avocado: Photometric Classification of Astronomical Transients with Gaussian Process Augmentation. *The Astronomical Journal*, 158(6):257, 12 2019.

[5] M. Bowles, A. M. Scaife, F. Porter, H. Tang, and D. J. Bastien. Attention-gating for improved radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 501(3):4579–4595, 2021.

[6] T. Cai, R. Gao, J. D. Lee, and Q. Lei. A Theory of Label Propagation for Subpopulation Shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1170–1182, 2021.

[7] O. Chapelle, B. Scholkopf, and A. Zien, Eds. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE Transactions on Neural Networks*, 20(3), 2009.

[8] A. Ćiprijanović, D. Kafkes, S. Jenkins, K. Downey, G. N. Perdue, S. Madireddy, T. Johnston, and B. Nord. Domain adaptation techniques for improved cross-domain study of galaxy mergers. In *Machine Learning and the Physical Sciences - Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. RandAugment: Practical data augmentation with no separate search. *CoRR*, abs/1909.1, 2019.

[10] B. L. Fanaroff and J. M. Riley. The Morphology of Extragalactic Radio Sources of High and Low Luminosity. *Monthly Notices of the Royal Astronomical Society*, 1974.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 6627–6638, 2017.

[12] J. Ineson, J. H. Croston, M. J. Hardcastle, and B. Mingo. A representative survey of the dynamics and energetics of FR II radio galaxies. *Monthly Notices of the Royal Astronomical Society*, 2017.

[13] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. … *Science Department, University of Toronto, Tech. …*, 2009.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[15] V. Lukic, M. Bruggen, J. K. Banfield, O. I. Wong, L. Rudnick, R. P. Norris, and B. Simmons. Radio Galaxy Zoo: Compact and extended radio source classification with deep learning. *Monthly Notices of the Royal Astronomical Society*, 2018.

[16] B. Mingo, J. H. Croston, M. J. Hardcastle, P. N. Best, K. J. Duncan, R. Morganti, H. J. Rottgering, J. Sabater, T. W. Shimwell, W. L. Williams, M. Brienza, G. Gurkan, V. H. Mahatma, L. K. Morabito, I. Prandoni, M. Bondi, J. Ineson, and S. Mooney. Revisiting the fanaroff–riley dichotomy and radio-galaxy morphology with the LOFAR two-metre sky survey (LoTSS). *Monthly Notices of the Royal Astronomical Society*, 2019.

[17] H. Miraghaei and P. N. Best. The nuclear properties and extended morphologies of powerful radio galaxies: The roles of host galaxy and environment. *Monthly Notices of the Royal Astronomical Society*, 466(4):4346–4363, 2017.

[18] Y. Netzer and T. Wang. Reading digits in natural images with unsupervised feature learning. *Neural Information Processing Systems*, pages 1–9, 2011.

[19] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow. Realistic evaluation of semi-supervised learning algorithms. In *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*, volume 2018-Decem, 2018.

[20] F. A. M. Porter. MiraBest Batched Dataset. *https://zenodo.org/record/4288837*, 11 2020.

[21] L. Saripalli. Understanding the fanaroffriley radio galaxy classification. *Astronomical Journal*, 144(3), 2012.

[22] A. M. M. Scaife and F. Porter. Fanaroff–Riley classification of radio galaxies using group-equivariant convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 503(2):2369–2379, 2021.

[23] K. Sohn, D. Berthelot, C. L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020.

[24] H. Tang, A. M. Scaife, and J. P. Leahy. Transfer learning for radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 488(3):3358–3375, 2019.

[25] C. Wu, O. I. Wong, L. Rudnick, S. S. Shabala, M. J. Alger, J. K. Banfield, C. S. Ong, S. V. White, A. F. Garon, R. P. Norris, H. Andernach, J. Tate, V. Lukic, H. Tang, K. Schawinski, and F. I. Diakogiannis. Radio Galaxy Zoo: CLARAN - A deep learning classifier for radio morphologies. *Monthly Notices of the Royal Astronomical Society*, 482(1):1211–1230, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section XX
   (c) Did you discuss any potential negative societal impacts of your work? [No] Data and impacts are not societal.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No] No theoretical results
   (b) Did you include complete proofs of all theoretical results? [No] No theoretical results
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 2
   (b) Did you mention the license of the assets? [Yes] See Section 2
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] No new assets
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] No personal data used; only public astronomy data
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] No personal data used; only public astronomy data
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] No human data used; only public astronomy data
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] No human data used; only astronomy data
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] No human data used; only astronomy data