International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Risk-Controlled Release and Audit of Machine-Generated Galaxy Morphology Catalogues

Amogh Shrivastava

South City International School, Kolkata 700068, India
Corresponding Author Email: *amoghshrivastava[at]gmail.com*

Abstract: ***Galaxy morphology catalogues for the Vera C. Rubin Observatory's Legacy Survey of Space and Time will require machine-generated labels at scale, but a model score alone cannot justify releasing a label. Existing studies have already combined machine learning, citizen science, active learning, uncertainty estimation, and human review. This paper addresses the downstream catalogue decision: when may a catalogue release a machine label, when should reviewers inspect it, and when should the catalogue abstain? We define a classifier-independent policy based on predeclared observational strata, random audits, one-sided exact binomial bounds, post-release monitoring, catalogue rollback, and label provenance. The policy converts a stated unsafe-label ceiling into an explicit audit requirement. With no unsafe outcomes in a fixed certification sample for one stratum, the 95% one-sided Clopper-Pearson rule requires 59, 149, or 299 reviewed entries to support ceilings of 5%, 2%, or 1%. We train no classifier and analyse no survey data. This analysis supplies an analytical catalogue-release policy for later testing with Rubin or precursor-survey data.***

Keywords: astronomical catalogues; citizen science; galaxy morphology; human-in-the-loop learning; machine learning; selective classification; Vera C. Rubin Observatory

1. Introduction

Ivezić et al. [1] project that the Legacy Survey of Space and Time (LSST) will detect about 20 billion galaxies over ten years. Even if only a fraction of those galaxies support resolved morphological measurements, exhaustive visual classification will not scale to the resulting sample. Galaxy Zoo has shown that volunteers can produce detailed morphological labels, and later studies trained deep models on volunteer vote distributions to extend those measurements to millions of galaxies [2]-[4]. Zoobot provides adaptable pretrained models for morphology tasks [5].

Prior work already implements the human-machine workflow that motivated this study. Beck et al. [6] integrated visual and automated classifications. Walmsley et al. [7] used Bayesian convolutional neural networks and active learning to reduce the number of newly labelled galaxies required for Galaxy Zoo tasks. Fu et al. [8] incorporated a human-in-the-loop module into a large vision model for galaxy-image analysis. Recent studies have examined uncertainty estimation and out-of-distribution detection for galaxy morphology [9], [10]. Human review of uncertain model outputs is not the contribution of this paper.

The unresolved problem begins after a classifier produces its predictions. Catalogue teams must decide which labels to release, how much independent evidence each release claim requires, how to allocate a finite review budget, and how to respond when later audits breach the declared error limit. A nominal model score of 0.95 does not answer those questions. Modern neural networks can be miscalibrated [11], and out-of-distribution inputs can still receive confident predictions [10]. A release decision should rest on a probability sample from the proposed catalogue subset, not on the score alone.

We formulate that decision for galaxy morphology catalogues. The policy places selective classification and exact binomial risk bounds inside a versioned release process. Predeclared observational strata and independent audits govern release, review, and abstention. Separate probability samples estimate the unsafe-label rate among released entries and the prevalence of missed morphologies among withheld entries. A provenance firewall prevents machine-only catalogue labels from becoming training truth without separate validation. The policy applies to any morphology model that supplies a documented acceptance score.

2. Previous Work and the Remaining Catalogue Problem

2.1 Human and machine classification

Galaxy Zoo 2 collected detailed decision-tree classifications for more than 300,000 Sloan Digital Sky Survey galaxies [2]. Galaxy Zoo DECaLS later combined deeper imaging, volunteer classifications, and deep learning for 314,000 galaxies [3]. Galaxy Zoo DESI then released automated morphology measurements for 8.67 million galaxies, using models that learned from Galaxy Zoo vote distributions [4]. These projects establish survey-scale combinations of human labels and machine predictions.

These projects leave the release decision unresolved. Active learning already directs volunteer effort towards examples that inform model training [7]. Human-in-the-loop systems already allow people to correct or supplement model output [6], [8]. What evidence is sufficient to publish the resulting machine label as a scientific measurement?

2.2 Calibration, abstention, and out-of-distribution inputs

A classifier's numerical score need not match its empirical probability of correctness. Guo et al. [11] documented poor calibration in modern neural networks and found that post-hoc calibration improved their probability estimates.

Selective classification treats abstention as an explicit option and studies the trade-off between the fraction of samples classified and the error among accepted predictions. Geifman and El-Yaniv [12] constructed selective neural classifiers with user-specified risk levels and high-probability guarantees.

Recent work has addressed uncertainty in galaxy morphology. Cheng et al. [9] proposed a post-hoc method that separates several sources of morphological uncertainty. Prakash et al. [10] evaluated calibration and out-of-distribution detection with Galaxy Zoo DECaLS images. These studies ask whether a model recognises ambiguity or unfamiliar inputs. Our analysis starts one step later: how should a survey team translate model scores and audited outcomes into catalogue-release decisions, post-release checks, and version control?

2.3 Bias and feedback

Morphological labels can vary with image resolution and other observational properties. Medina-Rosales et al. [13] showed how biases in human-labelled morphology data can propagate into deep models. A global accuracy value can hide poor performance in faint, small, distant, or crowded subsets. We therefore treat an observational stratum, rather than the catalogue as a whole, as the unit of audit.

Retraining creates a second risk. If a team inserts machine predictions into later training sets as verified labels, the next model learns from any errors in those predictions. Data feedback loops can amplify model-driven biases over repeated cycles [14]. Our policy distinguishes a label that may appear in a catalogue from a label that may serve as training truth.

3. Formal Problem Statement

3.1 Objects, attributes, and candidate labels

Let object i denote a galaxy image or image stack, and let attribute k denote a morphological question such as disk presence, bar presence, spiral structure, tidal disturbance, or multiple nuclei. Treating morphology as a set of attributes avoids forcing every galaxy into one exclusive Hubble class. An existing classifier produces a proposed binary label $\hat{y}_{ik} \in \{0,1\}$, a score s_{ik}, and, where available, a vector z_{ik} of uncertainty or out-of-distribution diagnostics. Before the audit, the model team maps those outputs to a documented scalar acceptance score $a_{ik} = g(s_{ik}, z_{ik})$.

Each object-attribute pair belongs to a predeclared observational stratum h. The team defines each stratum by the predicted label and by quantities that affect visibility, such as signal-to-noise ratio, apparent size, surface brightness, redshift, point-spread function, or crowding. It fixes the score function, stratum definitions, and thresholds before the release audit. Moving those choices after observing audit failures would invalidate the stated error control.

For threshold t_h, the candidate release set is

$$A_h(t_h) = \{(i,k): h(i,k) = h,\ a_{ik} \geq t_h\}.$$

The acceptance score may be a calibrated probability, an entropy-based ranking, a distance-based out-of-distribution score, or a fixed combination of these quantities. The catalogue policy does not require a particular neural architecture.

3.2 Audit outcomes

The team draws a simple random audit sample from $A_h(t_h)$. A predeclared reference protocol assesses each item, for example through several independent volunteer votes and expert adjudication when disagreement remains. The protocol assigns one of three audit outcomes:

$$r_{ik} \in \{\text{confirmed,contradicted,unresolved}\}.$$

For the release bound, define an unsafe outcome as either contradiction or unresolved status:

$$u_{ik} = \begin{cases} 0, & r_{ik} = \text{confirmed}, \\ 1, & r_{ik} \in \{\text{contradicted,unresolved}\}. \end{cases}$$

Classifying unresolved cases as unsafe makes the release test conservative. It prevents a catalogue from treating the limits of the image or the reference process as evidence that a machine label is correct. A project may also report contradiction and unresolved rates separately.

3.3 Release objective

Let q_h be the unsafe-label rate among machine-accepted entries in stratum h:

$$q_h = \Pr\big(u = 1 \mid (i,k) \in A_h(t_h)\big).$$

The catalogue team sets the largest unsafe-label rate it will tolerate, ϵ_h, and a failure probability, α_h. Let X_h denote the audit data and let $U_h(X_h)$ denote a one-sided upper confidence bound with coverage

$$\Pr_{q_h}\big(q_h \leq U_h(X_h)\big) \geq 1 - \alpha_h$$

for every admissible value of q_h. The team releases stratum h only when $U_h(X_h) \leq \epsilon_h$. Under the stated sampling assumptions, this rule limits the probability of releasing a stratum whose true unsafe-label rate exceeds ϵ_h to at most α_h. The claim concerns catalogue risk, not the interpretation of the model score.

4. Risk-Controlled Release and Audit Policy

4.1 Independent threshold selection and audit

The team must not choose a score threshold and certify its error bound through the same unrestricted search. A simple implementation reserves two disjoint datasets. A calibration set selects t_h and fits any probability recalibration; an independent release-audit set estimates the unsafe-label rate above that threshold. If the team evaluates several thresholds on one audit sample, it must use a selection-adjusted procedure, such as the high-probability threshold search developed for selective classification [12].

The team draws the audit sample at random from the candidate release set. Targeted review of unusual or low-

confidence objects can aid discovery and model improvement, but it cannot estimate release-set error unless the analysis records and uses the inclusion probabilities. The policy keeps targeted review separate from random audit.

4.2 Exact upper bound

Suppose an audit examines n_h entries and records $e_h = \sum u_{ik}$ unsafe outcomes. Under an exchangeable Bernoulli model within the stratum, the one-sided Clopper-Pearson upper confidence bound equals

$$U_h = \text{Beta}^{-1}(1 - \alpha_h; e_h + 1, n_h - e_h),$$

for $e_h < n_h$, with $U_h = 1$ when every audited entry is unsafe [15]. The team releases the stratum only when

$$U_h \leq \epsilon_h.$$

When the audit records no unsafe outcome, the bound reduces to

$$U_h = 1 - \alpha_h^{1/n_h},$$

and the following expression gives the minimum audit size:

$$n_h \geq \lceil \frac{\log \alpha_h}{\log(1 - \epsilon_h)} \rceil.$$

Because binomial outcomes are discrete, the exact interval often covers more than its nominal level. A team may choose another bound, but its release documentation must identify the method and state or cite the relevant coverage guarantee. When the audit samples a substantial fraction of a finite stratum without replacement, an exact hypergeometric bound can replace the binomial bound.

4.3 Simultaneous control across strata

A catalogue may release labels from many classes and observing conditions. Assign each stratum a failure budget α_h such that

$$\sum_{h=1}^{H} \alpha_h \leq \alpha_{\text{total}}.$$

The union bound then controls all passing strata simultaneously at level $1 - \alpha_{\text{total}}$. Equal allocation, $\alpha_h = \alpha_{\text{total}}/H$, is simple but not mandatory. A catalogue may assign smaller failure probabilities to strata whose errors carry greater downstream cost, provided that the team fixes the allocation before examining the audit outcomes.

Teams should not release labels from a sparse stratum merely because neighbouring strata perform well. They can merge strata under a predeclared rule, collect more audits, send the entries for review, or abstain.

4.4 Three catalogue decisions

Every candidate attribute receives one of three decisions:

$$d_{ik} \in \{\text{release}, \text{review}, \text{abstain}\}.$$

Release means that the stratum satisfies its risk bound and no current drift or image-quality flag overrides the decision. **Review** means that the image appears interpretable but the machine label lacks enough support for release. **Abstain** means that the available image does not support a defensible morphological statement, or that reviewers cannot resolve the reference assessment within the annotation limit.

The abstention state forms part of the scientific output. Requiring a label for every object converts limited resolution, blending, or low surface brightness into false certainty. A catalogue should retain the reason for abstention rather than collapse it into an "other" morphology.

4.5 Two audit streams

An audit of released labels estimates the unsafe-label rate among accepted entries, but it cannot measure how many true examples the system missed. A second random sample from the pool that the policy sends to review or abstention addresses that gap. Human reviewers use this sample to estimate the prevalence of missed positive attributes and to support class-specific completeness calculations.

The two streams serve different purposes:

1) **Release audit:** a random sample from machine-accepted labels estimates the unsafe-label rate among published entries.
2) **Omission audit:** a random sample from entries that the policy sends to review or abstention estimates missed morphologies and the cost of withholding.

Targeted active-learning samples may run alongside both streams, but teams should not use them as substitutes. Without a probability sample, a project can discover interesting failures yet remain unable to state catalogue-wide unsafe-label rates or completeness.

4.6 Audit-budget allocation

Let N_h be the number of candidate catalogue entries in stratum h, and let B be the number of reference assessments available for the release cycle. The team may use preliminary evidence to allocate the audit budget, but it must keep the final certification sample independent of the evidence that determined its size.

A two-stage design satisfies that requirement. The team first draws a planning sample from each candidate stratum and computes a preliminary upper bound $\widetilde{U}_h$. It uses those planning results to freeze the certification sample sizes before opening any certification outcomes. One transparent priority score is

$$P_h = w_h N_h \widetilde{U}_h,$$

where w_h is a predeclared scientific weight. The team assigns a minimum certification floor to every candidate stratum and distributes the remaining budget in proportion to P_h, subject to the total budget B. This heuristic directs review towards strata that could contribute the largest weighted number of potentially unsafe labels. It does not claim statistical optimality.

The team then draws a fresh certification sample and computes the release bound from those outcomes alone. Planning outcomes do not enter the final Clopper-Pearson calculation. If the team expands or stops the certification sample after inspecting its failures, a fixed-sample interval no longer supplies the stated guarantee. Such adaptation requires a time-uniform confidence sequence or another fresh holdout sample [16]. The release documentation must report planning and certification sample sizes, scientific weights, sampling probabilities, and any finite-population correction.

4.7 Post-release monitoring and rollback

A pre-release audit certifies only the catalogue version and data distribution that it sampled. Survey coadds deepen, observing conditions change, and retraining can alter the score distribution. Each materially changed model, threshold, or data distribution therefore needs fresh monitoring. If a project reports one guarantee across several versions or repeated checks, it must preallocate an error budget or use a time-uniform method [16]; otherwise the stated guarantee applies only to the individual audit.

If a post-release audit produces $U_h > \epsilon_h$, the team should suspend automatic release for that stratum. The catalogue must retain the affected records, attach a quality flag, and place them in a correction queue while preserving the model and threshold versions that produced them. A later release may replace the labels, but it must link each replacement to the superseded value. This policy turns audit failure into a versioned correction rather than an undocumented model update.

4.8 Provenance firewall

Store the origin of every label:

$$L_{ik} \in \{M, H, E, U\},$$

where M denotes machine-only, H human-reviewed, E expert-adjudicated, and U unresolved. Machine-only labels may enter the released catalogue after the relevant stratum passes the audit rule. By default, the supervised training set contains only records with human-reviewed or expert-adjudicated provenance:

$$\mathrm{D}_{\mathrm{train}} \subseteq \{(x_{ik}, y_{ik}) : L_{ik} \in \{H, E\}\}.$$

A project may study pseudo-labelling as a separate experiment, but it should not silently convert a catalogue-release decision into a ground-truth decision. This distinction limits the direct pathway by which an early model error can become supervision for later models, a concern raised by research on data feedback loops [14].

5. Analytical Results

5.1 Audit size with zero unsafe outcomes

Table 1 gives the minimum audit size required when a stratum records zero unsafe outcomes, using a one-sided 95% Clopper-Pearson bound for one stratum.

Table 1: Minimum audit size with zero unsafe outcomes ($\alpha = 0.05$).

Maximum unsafe-label rate ϵ	Minimum audit size n	Achieved upper bound
10%	29	9.81%
5%	59	4.95%
2%	149	1.99%
1%	299	1.00%
0.50%	598	0.50%

The table quantifies the sample size behind each claim. “We observed no errors” has little meaning without the denominator. Zero failures among 30 audited labels cannot support a 2% upper limit; under the stated assumptions, zero failures among 149 can support it.

5.2 Cost of multiple strata

Simultaneous control increases the audit requirement because each stratum receives a smaller failure budget. Table 2 assumes a family-wise failure probability of 5%, equal allocation across H strata, a 2% unsafe-rate ceiling, and zero unsafe outcomes in every stratum.

Table 2: Audit cost for simultaneous control at $\epsilon = 0.02$ and $\alpha_{\mathrm{total}} = 0.05$

Number of strata H	Per-stratum α_h	Audits per stratum	Total audits
1	0.0500	149	149
5	0.0100	228	1,140
10	0.0050	263	2,630
20	0.0025	297	5,940
1	0.0500	149	149

These totals argue against unnecessary binning. Strata should represent plausible performance changes, not every available metadata combination.

5.3 Worked six-stratum example

Consider a release that places one million machine-generated attribute labels in six predeclared observational strata. The team sets $\epsilon_h = 0.02$ for every stratum and $\alpha_{\mathrm{total}} = 0.05$. Equal allocation gives $\alpha_h = 0.00833$. If the audit records zero unsafe labels, each stratum requires 237 reviewed entries, or 1,422 audits in total, before all six strata can pass simultaneously.

Suppose a fixed certification sample contains 237 entries and two unsafe outcomes. Its one-sided upper bound is 3.59%, so the stratum fails the 2% release rule. For comparison, a fixed sample of 429 entries with exactly two unsafe outcomes yields an upper bound of 2.00% after rounding. This comparison does not permit a team to inspect the first 237 outcomes and then extend the same fixed-sample analysis to 429. The team must declare the larger sample size before opening its outcomes, or use a time-uniform method [16]. The raw failure proportion alone does not determine whether the evidence supports the declared catalogue limit.

These calculations do not predict Rubin performance. They state what an audit must observe before a team can make a chosen risk claim.

6. Application to Rubin LSST Galaxy Morphology

A Rubin morphology catalogue could apply the policy to attribute predictions from Zoobot or another documented model. The release rule should remain independent of model architecture. For each attribute, the team would first identify the population whose images support morphological measurement, then define strata that reflect the properties most likely to affect visibility.

Useful variables include apparent half-light radius relative to the point-spread function, mean surface brightness, signal-to-noise ratio, redshift, blending or crowding flags, and the depth or band combination that forms the cutout. Prior knowledge or a separate development sample should determine the number and boundaries of the strata. The team must not alter them to make an audit subset pass.

The catalogue should report performance within those strata rather than force equal morphology rates across stellar-mass or redshift bins. Different galaxy populations may have different true prevalences, so equalisation could remove astrophysical signal. Release documentation should report calibration, the unsafe-label rate against the stated bound, separate contradiction and unresolved rates, completeness, and the selection function across observational conditions. Medina-Rosales et al. [13] show why image-dependent label bias requires this conditional reporting.

Citizen science serves two distinct functions under this policy. Targeted review may improve the model or inspect unusual sources; random audit estimates catalogue quality. Mixing those samples destroys a simple probability interpretation unless the analysis records and uses the inclusion probabilities.

Rare morphologies require additional care. A low-prevalence class may have too few audited positives to support a narrow unsafe-rate bound. The policy allows a larger scientific weight in audit allocation, but the team cannot claim a reliable rare-object catalogue before the sample supports that claim. It may instead publish a candidate list with explicit uncertainty.

7. Limitations and Implementation Requirements

Audit validity first depends on representative sampling within each stratum. Targeted selection, missing images, or unrecorded exclusions can invalidate the bound. The reference protocol adds another limitation. Human reviewers can disagree, and image quality can make the attribute unobservable. Counting unresolved cases as unsafe protects the release claim, but the resulting bound refers to confirmation under the stated protocol, not to the galaxy's underlying physical state.

The binomial model assumes a stable unsafe-label probability within each stratum and independent audited outcomes. Duplicate observations, several attributes from one galaxy, spatially correlated image defects, or a shared processing failure can violate those assumptions. Such dependence requires clustered sampling or a more suitable model. Exact binomial intervals can also cover more than their nominal level, especially after the team divides the failure probability across many strata. Hierarchical models could borrow information across related strata, but they introduce assumptions that require separate validation.

The policy controls the quality of labels that enter the catalogue; it cannot make the catalogue representative of the full galaxy population. Selection cuts, detection incompleteness, and abstention can still bias scientific analyses. Data releases must publish those selection rules and the omission-audit results. This paper provides an analytical protocol rather than an empirical evaluation. Only Rubin or precursor-survey data can establish attainable coverage, volunteer workload, audit cost, and sensitivity to survey drift.

Implementation requires a catalogue schema that records the model version, score definition, threshold version, observational stratum, release decision, label provenance, audit status, and superseded values. The project must also keep threshold-selection data, release-certification data, and post-release monitoring data separate. Without that separation, the numerical bound may look precise while relying on reused evidence.

8. Conclusion

Galaxy-morphology research already uses people to review or supply labels for machine-learning systems. The unresolved operational question concerns evidence: when does a machine-generated label have enough independent support to enter a scientific catalogue, and what should happen when later data invalidate that support? This paper specifies a classifier-independent policy built around predeclared strata, random audits, exact upper risk bounds, a release-review-abstain decision, separate release and omission audits, catalogue rollback, and label provenance.

The analytical results make the evidence requirement explicit. With a one-sided 95% bound and zero observed audit failures, a team must review 149 entries to support a 2% unsafe-rate ceiling for one stratum; simultaneous claims across several strata demand more audits. These figures do not establish the performance of any galaxy classifier. They establish the amount of verification needed before a project can state a chosen catalogue risk limit.

Researchers can test the policy around an existing model with public precursor-survey data or future Rubin products. A valid test must preserve the separation among threshold selection, certification, targeted review, and post-release monitoring. It must also retain abstentions, version corrections, and label provenance.

Statements and Declarations

Funding: The author received no specific funding for this work.

Competing Interests: The author declares no relevant financial or non-financial competing interests.

Acknowledgements: I'd like to acknowledge the proofreading of Izum Yasar Anique, Kripesh V, Adwitya Kumar, Aryamann Gupta, Nayomi Sengupta, and Satvik Jha of the Mathematics and Astronomy Society of South City International School.

Ethics Approval: Not applicable. The work presents a methodological and analytical proposal and did not involve human participants, animals, or private data.

Data Availability: The authors generated and analysed no astronomical data. The numerical values in Tables 1 and 2 and the worked example follow directly from the stated one-sided exact binomial formulas.

References

[1] Ž. Ivezić, S. M. Kahn, J. A. Tyson, et al., "LSST: From science drivers to reference design and anticipated data products," *The Astrophysical Journal*, vol. 873, no. 2, art. 111, 2019. doi:10.3847/1538-4357/ab042c

[2] K. W. Willett, C. J. Lintott, S. P. Bamford, et al., "Galaxy Zoo 2: Detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 435, no. 4, pp. 2835-2860, 2013. doi:10.1093/mnras/stt1458

[3] M. Walmsley, C. J. Lintott, T. Géron, et al., "Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies," *Monthly Notices of the Royal Astronomical Society*, vol. 509, no. 3, pp. 3966-3988, 2022. doi:10.1093/mnras/stab2093

[4] M. Walmsley, T. Géron, S. Kruk, et al., "Galaxy Zoo DESI: Detailed morphology measurements for 8.7 million galaxies in the DESI Legacy Imaging Surveys," *Monthly Notices of the Royal Astronomical Society*, vol. 526, no. 3, pp. 4768-4786, 2023. doi:10.1093/mnras/stad2919

[5] M. Walmsley, C. Allen, B. Aussel, et al., "Zoobot: Adaptable deep learning models for galaxy morphology," *Journal of Open Source Software*, vol. 8, no. 85, art. 5312, 2023. doi:10.21105/joss.05312

[6] M. R. Beck, C. Scarlata, L. F. Fortson, et al., "Integrating human and machine intelligence in galaxy morphology classification tasks," *Monthly Notices of the Royal Astronomical Society*, vol. 476, no. 4, pp. 5516-5534, 2018. doi:10.1093/mnras/sty503

[7] M. Walmsley, L. Smith, C. Lintott, et al., "Galaxy Zoo: Probabilistic morphology through Bayesian CNNs and active learning," *Monthly Notices of the Royal Astronomical Society*, vol. 491, no. 2, pp. 1554-1574, 2020. doi:10.1093/mnras/stz2816

[8] M.-X. Fu, Y. Song, J.-M. Lv, et al., "A versatile framework for analyzing galaxy image data by implanting human-in-the- loop on a large vision model," *Chinese Physics C*, vol. 48, no. 9, art. 095001, 2024. doi:10.1088/1674-1137/ad50ab

[9] K. Cheng, R. Wang, and Q. Luo, "Estimating uncertainty in galaxy morphology classification," arXiv:2608.08398, 2026. doi:10.48550/arXiv.2608.08398

[10] P. Prakash, S. Desai, and P. K. Srijith, "Galaxy morphology classification: Uncertainty modeling and out-of-distribution detection," arXiv:2608.16654, 2026. doi:10.48550/arXiv.2608.16654

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1321-1330, 2017.

[12] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems 30*, pp. 4878-4887, 2017.

[13] E. Medina-Rosales, G. Cabrera-Vives, and C. J. Miller, "Mitigating bias in deep learning: Training unbiased models on biased data for the morphological classification of galaxies," *Monthly Notices of the Royal Astronomical Society*, vol. 531, no. 1, pp. 52-60, 2024. doi:10.1093/mnras/stae1088

[14] R. Taori and T. B. Hashimoto, "Data feedback loops: Model-driven amplification of dataset biases," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pp. 33883-33920, 2023.

[15] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404-413, 1934. doi:10.1093/biomet/26.4.404

[16] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon, "Time-uniform, nonparametric, nonasymptotic confidence sequences," *The Annals of Statistics*, vol. 49, no. 2, pp. 1055-1080, 2021. doi:10.1214/20-AOS1991