Through the Citizen Scientists' Eyes: Insights into Using Citizen Science with Machine Learning for Effective Identification of Unknown-Unknowns in Big Data

CITIZEN SCIENCE:
THEORY AND PRACTICE

COLLECTION:
THE FUTURE
OF ARTIFICIAL
INTELLIGENCE AND
CITIZEN SCIENCE

CASE STUDIES

]u[ubiquity press

KAMESWARA BHARADWAJ MANTHA
HAYLEY ROBERTS
LUCY FORTSON
CHRIS LINTOTT
HUGH DICKINSON
WILLIAM KEEL
RAMANAKUMAR SANKAR
COLEMAN KRAWCZYK
BROOKE SIMMONS
MIKE WALMSLEY
IZZY GARLAND
JASON SHINGIRAI MAKECHEMU
LAURA TROUILLE
CLIFFORD JOHNSON

*Author affiliations can be found in the back matter of this article

ABSTRACT

In the era of rapidly growing astronomical data, the gap between data collection and analysis is a significant barrier, especially for teams searching for rare scientific objects. Although machine learning (ML) can quickly parse large data sets, it struggles to robustly identify scientifically interesting objects, a task at which humans excel. Human-in-the-loop (HITL) strategies that combine the strengths of citizen science (CS) and ML offer a promising solution, but first, we need to better understand the relationship between human- and machine-identified samples. In this work, we present a case study from the Galaxy Zoo: Weird & Wonderful project, where volunteers inspected ~200,000 astronomical images—processed by an ML-based anomaly detection model—to identify those with unusual or interesting characteristics. Volunteer-selected images with common astrophysical characteristics had higher consensus, while rarer or more complex ones had lower consensus. This suggests low-consensus choices shouldn't be dismissed in further explorations. Additionally, volunteers were better at filtering out uninteresting anomalies, such as image artifacts, which the machine struggled with. We also found that a higher ML-generated anomaly score that indicates images' low-level feature anomalousness was a better predictor of the volunteers' consensus choice. Combining a locus of high volunteer-consensus images within the ML learnt feature space and anomaly score, we demonstrated a decision boundary that can effectively isolate images with unusual and potentially scientifically interesting characteristics. Using this case study, we lay important guidelines for future research studies looking to adapt and operationalize human-machine collaborative frameworks for efficient anomaly detection in big data.

CORRESPONDING AUTHOR:
Kameswara Bharadwaj Mantha
University of Minnesota-Twin Cities, US
manth145@umn.edu

KEYWORDS:
Deep learning; anomaly detection; human-machine optimization; unsupervised learning; astronomy imaging

TO CITE THIS ARTICLE:
Mantha, KB, Roberts, H, Fortson, L, Lintott, C, Dickinson, H, Keel, W, Sankar, R, Krawczyk, C, Simmons, B, Walmsley, M, Garland, I, Makechemu, JS, Trouille, L and Johnson, C. 2024. Through the Citizen Scientists' Eyes: Insights into Using Citizen Science with Machine Learning for Effective Identification of Unknown-Unknowns in Big Data. *Citizen Science: Theory and Practice,* 9(1): 40, pp. 1–15. DOI: https://doi.org/10.5334/cstp.740

MOTIVATION

The classification of several million galaxies with the Galaxy Zoo (GZ) project has been one of the cornerstones of citizen science in astronomy over the past 15 years. A key success of the GZ project is the unexpected discovery of previously-unknown objects such as the Green Pea galaxies (Cardamone et al. 2009) and phenomena like Hanny's Voorwerp (Lintott et al. 2009), where volunteer discussions on project forum boards about "odd things" they found played a critical role in the discovery process (see Straub 2016). Although such efforts worked when the largest data products were about a million images (e.g., Sloan Digital Sky Survey; York et al. 2000), it becomes nearly impossible to have eyes on each image in the upcoming big-data era anticipated to produce many millions of images total per night.

While machine learning (ML) algorithms can now quickly sift the data for rare objects (Margalef-Bentabol et al. 2020), these objects are not necessarily scientifically interesting (e.g., image artifacts, see Storey-Fisher et al. 2021). Human-in-the-loop (HITL) strategies (e.g., Lai et al. 2020) enabled by citizen science offer a synergistic middle ground where the relative strengths of humans and machines can be combined to identify scientifically interesting unknown-unknowns (e.g., Lochner and Bassett 2021; Sharifi et al. 2022; Walmsley et al. 2022). However, implementing HITL-based anomaly detection pipelines requires a thorough exploration of the overlap between anomalies found by ML versus human-driven frameworks. Specifically, on the same data set, we need to investigate correlations between which images a machine-trained anomaly detector determines are anomalous versus which images humans determine contain "scientifically interesting" anomalies. Note that data from other modalities such as spectroscopy (1D representation of collected light as a function of wavelength) have also been considered for the purpose of anomaly detection (e.g., Hoyle et al. 2015 and Liang et al. 2023) and citizen science efforts (Coffin et al. 2023). Nevertheless, with the very large data sets involved, research teams will not have capacity to make these "human" determinations. Therefore, to test at scale whether the combination of human and machine methods provides a list of scientifically interesting anomalies, we need to determine whether people with minimal domain knowledge (aka citizen scientists) can reliably supply information on which images contain unusually interesting features. Furthermore, we need to show that the information supplied by the citizen scientists complements the machine-driven anomaly finder such that the combination provides the optimal set of scientifically interesting anomalies. A research team could then develop well-informed selection criteria for the data set to reduce to a tractable number the images that need to be vetted by the research team.

To carry out the investigations detailed above as a case study and provide insights into how research teams could apply this novel approach of a combined human-machine anomaly detection pipeline, we designed a deep learning anomaly detection framework and ran a citizen science project on the Zooniverse (www.zooniverse.org) platform called Galaxy Zoo: Weird and Wonderful (GZ:W&W; https://www.zooniverse.org/projects/zookeeper/galaxy-zoo-weird-and-wonderful). In this work, we describe our methodology and provide insights into the correlation between machine- and GZ:W&W-based anomalies, and comment on promising next steps for applying our methods for much larger future datasets. This paper is structured as follows: First, we describe the imaging-based data used in this work, followed by a brief overview of our deep learning–based anomaly detection model and methods, we describe our citizen science project, GZ:W&W. Next, we comprehensively show various quantitative results from the GZ:W&W project alongside the anomaly detection–based metrics and assess the correlation between these quantities. Based on the insights from our results, we then briefly discuss our recommendations for future research teams towards applying our approach to new datasets and potentially fine-tuning it for specific purposes. Finally, we provide our concluding statements in Section 6.

IMAGING DATA USED IN THIS WORK

In this work, we use the data taken from the Subaru Hyper-Suprime Cam (HSC) survey's public data release 2 (PDR2), which imaged a large portion of the sky in multiple optical wavelengths with the Subaru Telescope and serves as one of the notable modern-era resources for images containing nearby-to-distant galaxies. Specifically, we chose approximately 1.5 million images, among which we used a random selection of 250,000 images for our deep learning model training and a subsequent randomly chosen collection of 200,000 images for visual inspection. The selection process of the images used in this work is explained in more detail in Supplemental File 1: Appendix A.

DEEP LEARNING–BASED ANOMALY DETECTION MODEL

Our anomaly detection framework is based on a generative, convolutional neural network deep learning model. Specifically, our model is based on a framework described in Storey-Fisher et al. (2021) involving astronomical images and the training strategy employed by the fast-AnoGAN model (Schlegl et al. 2019), a generative adversarial network (GAN) (Goodfellow et al. 2014) framework applied to medical imaging. This model comprises two separate design and training steps: 1) a Wasserstein GAN with gradient penalty (wGAN-GP); and 2) an encoder. Next, we briefly describe our model architecture, training methods, and the corresponding model outputs. We reserve our detailed model descriptions and its involved training hyper-parameters to our discussion in Supplemental File 1: Appendix B. We show a schematic of our framework and associated model outputs in Figure 1 in Supplemental File 1: Appendix B.

ARCHITECTURE OVERVIEW

Our wGAN-GP model contains two learnable modules: 1) a convolutional generator (*G*) that takes in an N-dimensional "latent space" vector (often represented by *z*) as input and learns to generate realistic images with respect to the input dataset; and 2) a convolutional discriminator (*D;* sometimes called a "critic" network) that learns to predict the realism of the generated images. Conceptually, the *z*-vector serves as a compressed, lower-dimensional encoding of the image-level information (e.g., whiskers for an image containing a cat versus striped pattern of a tiger) and can serve as a landscape in which images with specific (or different) features populate deterministic and distinct locations within the *z* space. It is also important to note that a wGAN-GP model is a variant of the traditional GANs, which optimizes the Wasserstein distance (Rubner et al. 2000) metric and is known for its stability during training.

Although the wGAN-GP framework is set up to learn the realistic generalization of the input dataset such that it can randomly generate representative image samples, it is not equipped to provide the exact feature space representation corresponding to an input image. To do so requires an additional model/module that learns to behave as an inverse of the trained generator. Drawing inspiration from the setup of the fast-AnoGAN framework, we thus define an encoder (*E*) model, which outputs a feature representation vector (*z*) for an input image that has the same dimensions as the input vector used as input by the generator network. In our work, we use D = 128 dimensions for our feature representation vector *z*. This framework is illustrated in Figure 1 in Supplemental File 1: Appendix B.

TRAINING STRATEGY

As mentioned previously, there are two steps in our training strategy. First, we train our wGAN-GP model on the previously described 250,000-image dataset with a batch size of 1,024 and an Adam optimizer with 10^{-4} learning rate for a total of 500 epochs. This model is optimized by jointly minimizing specific loss parameterizations (see Loss parameterizations during training section in Supplemental File 1: Appendix B) of the generator (*G*) and discriminator (*D*). Next, while holding the *G* and *D* models fixed, we train our encoder *(E)* on the same set of 250,000 images with a batch size of 256 and an Adam optimizer with 10^{-4} learning rate for 500 epochs. The encoder model is optimized by minimizing the following two-component loss function (Equation 1):

$$Loss_{enc} = 0.8 \times Loss_{image} + 0.2 \times Loss_{feature}$$

Here, the image loss ($Loss_{img}$) corresponds to the pixel-level difference between the true and generated images and serves as a quantitative metric of how unusual that image is in a spatial context (i.e., high-level features). On the other hand, the feature loss ($Loss_{feature}$) is a difference between the low-level features extracted from the true and generated images, and as such quantifies how unusual are two images in terms of low-level features. The joint optimization of the $Loss_{img}$ and $Loss_{feature}$ conceptually ensures that the encoder learns good image-to-*z* mapping. This in turn yields a generated image (by *G*) that is similar to the input *real* image, while simultaneously ensuring that the set of discriminator features of the *real* and *generated* images are also similar.

We followed common practices used in the literature to gauge the convergence of our wGAN-GP and encoder models by reserving 10% of our entire dataset for validation purposes during the training phase. We assessed the training and validation loss profiles and found that they both reached stagnation around 500 epochs (i.e., no further improvement in loss) while yielding similar loss values.

ANOMALY SCORES AND LATENT SPACE FEATURE REPRESENTATION VECTORS

After our training procedure is complete, we are left with three trained modules *G, D,* and *E*. Conceptually, for each input image, the *G* model provides a $Loss_{image}$ value that encompasses how unusual that image is from a spatial (high-level feature) context, the *D* model provides $Loss_{feature}$ value that captures how unusual the input image is from a low-level feature standpoint. Hereafter, we treat and refer to the $Loss_{image}$ and $Loss_{feature}$ as the Image Score (S_{image}) and Feature Score ($S_{feature}$), respectively, the weighted sum of which make up the "anomaly score" (S_{anom}; see Equation 1). Simultaneously, the *E* model enables us to compute a feature representation for each input image. By inferring our trained *G, D,* and *E* models on a sample of 1.5 million images, we compute their corresponding anomaly scores and the latent space feature representations (*z*). Conceptually, a poor generalization by the Generator and Discriminator directly translates to a poor representation of that particular type of image in the dataset. As such, a high anomaly score would be expected for an image that is rarely occurring in the dataset as it would yield a poorly-matched *G* output and resultant *D* features (i.e., high image and feature losses). For context on the general distribution of S_{image} and $S_{feature}$, see Figure 3 in Supplemental File 1: Appendix B.

THE GALAXY ZOO WEIRD & WONDERFUL PROJECT: SAMPLE CONSTRUCTION AND STATISTICS

With the main aim to understand the relationship between machine-based and human-driven anomaly detection, we designed and launched the citizen science project GZ:W&W, in which volunteers were given a tailored set of images inter-mixed with our ML-based anomalous images for inspection. In this section, we describe the construction of the sample used in the GZ:W&W project and briefly outline the project completion statistics.

We start with the anomaly scores computed on the 1.5 million images and choose the top 1% from their distribution to compile ~15,000 images that have the highest anomaly scores within the entire dataset. Next, we randomly select from the remaining sample of images (i.e., <99% of the anomaly scores) to pool a set of ~185,000 images. Together, these amount to ~200,000 images with ~1:12 ratio of highly anomalous versus the remaining images, which we use for our GZ:W&W project. For visual illustration, we show an example collage of non-anomalous and anomalous images in Supplemental Figure 12 in Supplemental File 1: Appendix G.

We designed and launched our GZ:W&W project on the Zooniverse platform (www.zooniverse.org), where each volunteer was shown a random pool of images from the 200,000 on a 4 × 4 grid and were simply asked to "Click on any galaxies which are particularly interesting to you." Additionally, volunteers could further discuss any images within the discussion boards (called "Talk pages" on Zooniverse) by posting comment threads and providing hash (#) tags. Overall, the project took approximately 2 months to finish with ~2,000 participating volunteers. Each image was taken out of circulation (i.e., "retired") when it had been seen by at least 10 volunteers. Among the 200,000 images, approximately 3,000 were discussed in Talk, for which #tags were indicated.

ANALYSIS AND RESULTS

In this section, we derive simple statistical metrics from the GZ:W&W project and assess their correlations with various anomaly detection model metrics such as the anomaly score and feature space representation.

VOLUNTEER CHOSEN FRACTION AND EXPERIENCE WEIGHTING

For each image used in our GZ:W&W sample, we quantified the "volunteer chosen fraction" as the ratio of the number of times volunteers selected that particular subject to the total number of volunteers who have seen it. To provide them with some context to what might be considered "usual" galaxy images, we encourage volunteers to classify on the standard Galaxy Zoo (GZ) project before participating on GZ:W&W; however, no stringent gating was employed. As such, 305 volunteers who participated in GZ:W&W have also participated in the GZ project with at least 100 classifications each (Supplemental Figure 4 in Supplemental File 1: Appendix C). Hereafter, these 305 volunteers are referred to as GZ-participated, and the classifications from them amount to ~40% of the total GZ:W&W classifications. To take into account the prior domain experience of volunteers who participated in the GZ project, we also compute a "weighted volunteer chosen fraction" which increases the weight of votes from previous GZ participants by a factor of two compared with novice volunteers. See Supplemental File 1: Appendix C for more details on the weighting scheme.

This weighting scheme has the highest impact on images that have low agreement between volunteers. Since volunteers have only a binary choice (i.e., a source is interesting or not), this results in chosen fractions of ~0.5 varying the most between the two weighting schemes. This is because a strong agreement (i.e., a value that falls close to a maximal value of 0 or 1) won't have a strong effect when taking volunteer experience into account because the vote values do not vary significantly. A comparison between weighted and unweighted chosen fraction measurements along with relative differences are shown in Supplemental Figure 5 in Supplemental File 1: Appendix D.

In Supplemental Figure 7 in Supplemental File 1: Appendix G, we visually illustrate the impact of upweighting the contribution from experienced volunteers, where the top row (bottom row) shows images that had the most decrease (increase) in the chosen fraction. By upweighting the chosen fraction based on the experienced volunteer participation, the number of galaxies with interesting, but less unique features are reduced. Likewise, sources with distinct and rarer features are ranked higher with the introduction of domain experience weighting. For example, the bottom row of Supplemental Figure 7 in Supplemental File 1: Appendix G shows images with increased percentages, notably those hosting gravitational lenses – a rarer and scientifically interesting phenomenon. This demonstrates that introducing a classification weight based on volunteer experience can highlight rarer and more anomalous features than those identified by the general volunteer population.

VOLUNTEER CHOSEN FRACTION DISTRIBUTIONS FOR TALK DISCUSSED IMAGES

For the subset of images that were discussed in the GZ:W&W project Talk discussion boards, we compile information on their corresponding #tags and number of comments made. We process the #tags to be more uniform by taking into account any typographical errors and singular or plural mentions. We then manually group the #tags into broader categories if a particular tag has been used at least 10 times. Some example images along with their corresponding #tags are shown in Figure 1, which also highlights the diversity of characteristics identified by the volunteers.

To further explore potential relationships between the volunteer chosen fraction and the variety of images discussed in Talk, we analyzed the distribution of the weighted volunteer chosen fraction(s) for a subset of subjects that are discussed in the GZ:W&W Talk boards and were marked with different tags (For example, see Supplemental Figure 11 in Supplemental File: Appendix G). The weighted volunteer chosen fraction distribution (as shown in Figure 1) among the entire Talk-discussed sample is a bimodal structure with peaks at ~0.25 and ~0.6. When assessing this distribution, separated into different tag-based subsets, we found that the distribution of chosen fractions for each category showed distinct behavior. Notably, we found that subjects tagged with more abundant or "easily noticeable" characteristics (e.g., #merger or #ring) tend to have higher median chosen fractions, whereas the images containing more subtle categories (e.g., #gravitational lens or #arc) have a lower median chosen fraction. This insight highlights that the observed range of chosen fractions encodes the diversity of different categories of interesting characteristics as well as the "knowledge model" of the volunteer pool who participated in the project. This emphasizes the fact that we should not discount low chosen fraction images if we want to determine which images are of scientific interest; all it takes is one person who knows what a gravitational lens looks like (or thinks it just looks interesting).

Additionally, we assessed the frequency of Talk comments for images tagged with different #tags (*right panel* of Supplemental Figure 11 in Supplemental File: 1, Appendix G). We found that images containing more subtle features or phenomena (e.g., #gravitational lens) tend to be discussed more extensively with higher numbers of Talk comments compared with #mergers that are relatively "easier" to comprehend and have a high median chosen fraction. This is evident especially from the transition between the N = 3–4 comments bins.

VOLUNTEER CHOSEN FRACTION VERSUS MACHINE ANOMALY SCORES

In Figure 2, we assess the correlation between the volunteer chosen fraction and the anomaly detection model scores – anomaly score (S_{anom}), and its two components: image score (S_{image}) and feature score ($S_{feature}$).

Generally speaking, we note that there is no appreciable correlation between the S_{anom} and the weighted volunteer chosen fraction; this is true even if we assessed the subsample of subjects discussed in Talk (see Figure 2a). We also notice a "ridge" in the S_{anom} space (Figure 2a) where it seems to follow a bimodal distribution. Upon investigation, we find that this is an artificial effect as a consequence of the relative weighting values of 0.8 and 0.2 used to combine the S_{image} and $S_{feature}$ components (see the aforementioned equation). Motivated by these observations and taking inspiration from the exploration by Storey-Fisher et al. (2021), we assessed the correlation between the chosen fraction individually for S_{image} and $S_{feature}$.

It is worth noting that the S_{image} encodes information about the generalizability on a spatial-level that can be dominated by features such as galaxy morphology or other image-level signatures, whereas the $S_{feature}$ is indicative

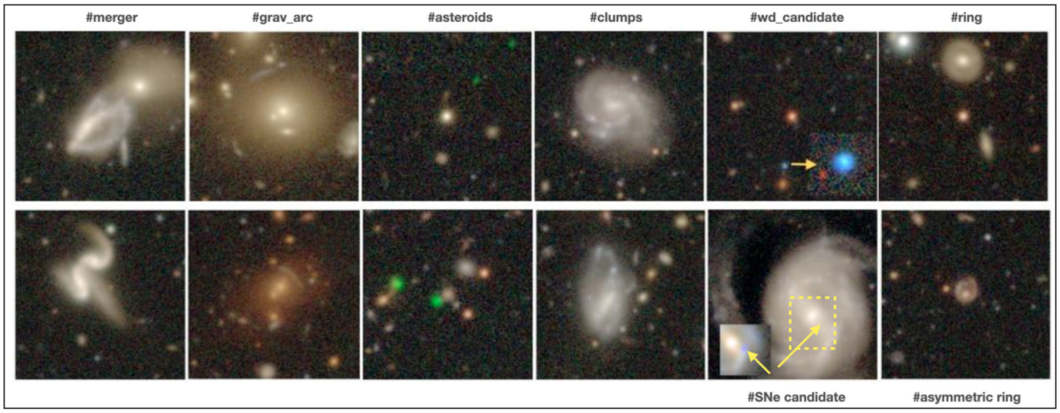

Figure 1 An example collage of images from the Galaxy Zoo: Weird & Wonderful (GZ:W&W) project that have been discussed in the Talk boards and their corresponding volunteer-provided tags.

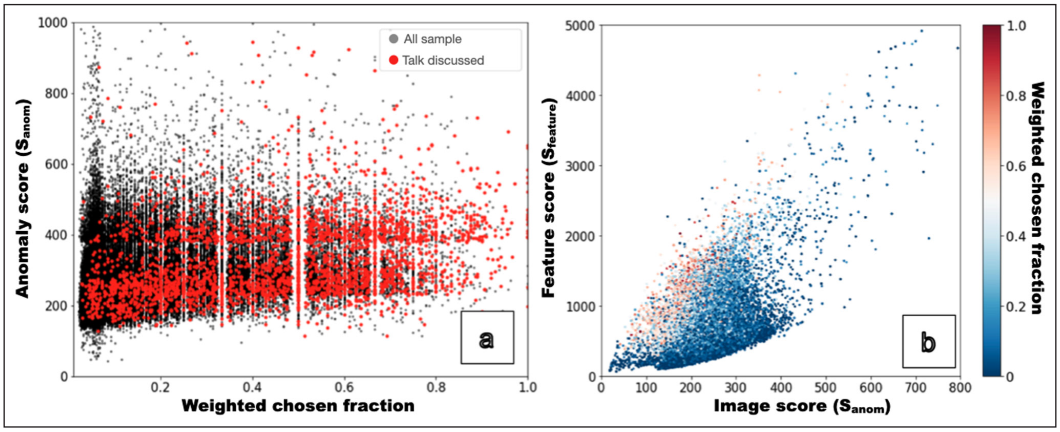

Figure 2 *Left panel:* The anomaly score versus the fraction of times a volunteer identified a subject as interesting in the Galaxy Zoo: Weird & Wonderful (GZ:W&W) (volunteer chosen fraction), with an upweighting of selections by experienced volunteers who have substantial participation in GZ:W&W, with the subset of those subjects discussed in Talk boards (*red points*). *Right panel*: The feature score versus image score for our entire ~200,000 GZ:W&W sample color-coded by the GZ experienced volunteer response weighted chosen fractions, respectively.

of generalizability across both high-level and low-level features that a particular image bears. When comparing the image score versus the feature score in the context of the volunteer chosen fraction (Figure 2b), an interesting trend emerges. We find that weighted (also unweighted) volunteer chosen fraction values qualitatively have higher correlation with $S_{feature}$ rather than the S_{image}. For example, at a fixed S_{image}, images with higher $S_{feature}$ have preferentially higher chosen fraction values. Especially, a majority of the images with chosen fractions >0.5 range between S_{image} ~100–300 and $S_{feature}$ ~1000–2000 (i.e., *top left* region of Figure 2b). This indicates that volunteers mimic the conceptual task of a discriminator module of the anomaly detection model towards finding interesting/rare subjects within the dataset. One should note that, despite this interesting relationship, there is still no quantitative correlation between $S_{feature}$ (or S_{image} for that matter) and volunteer chosen fraction itself. That is, a higher value of chosen fraction doesn't necessarily correlate with the quantitative value of $S_{feature}$ or $S_{image.}$ Nonetheless, our insights motivate the use of a different weighting factor for the feature and image losses (than 0.8 and 0.2) while training the machine model. This is to preferentially prioritize the feature score while considering the information from S_{image} with a smaller weightage.

Additionally, we also note that there is an anti-correlation between S_{image} and chosen fractions, especially at high S_{image} values. Upon visual inspection of these images, they tend to be dominated by various image artefacts and were not selected by the volunteers. We also explore this aspect further while assessing the feature space learnt by our anomaly detection model. We also show the image versus feature score with both weighted and unweighted chosen fraction in Supplemental Figure 8 in Supplemental File 1: Appendix G.

Furthermore, in Figure 3, we show the distributions of S_{anom}, S_{image}, and $S_{feature}$ for weighted volunteer chosen fraction >0.5 images in comparison with our entire sample. We find that $S_{feature}$ is more predictive in terms of determining if an image has a high volunteer chosen fraction versus the S_{image}. This reinforces our previous comments on the observed correlations between volunteer chosen consensus and $S_{feature}$. More explicitly, this observation addresses our motivation question on which machine-derived metrics hold optimal potential to be combined with human consensus for finding rare and interesting objects.

CORRELATIONS BETWEEN CHOSEN FRACTIONS WITHIN THE FEATURE SPACE

As described in a previous section, our anomaly detection model (specifically the encoder *E*) yields a latent space feature vector *z* (dimensionality D = 128) for each input image. Conceptually, this vector encodes the important feature-level information that describes the overall semantic meaning carried by each image. As such, this can provide important insights into the landscape of images containing different physical properties. In this section, we discuss our assessment of the GZ:W&W chosen fraction metric alongside *z* and the S_{image} and $S_{feature}$ scores.

Although we encode the feature vector with D = 128, not all individual components of the feature vector contribute equal importance towards the semantic information captured within the images. A common practice in computer vision literature is to reduce the feature space's dimensionality so that individual features are sorted in decreasing order of importance and use them for any downstream quantitative and qualitative assessment. Following this approach, we first process the raw D = 128 feature vectors for all our 200,000 images used in the GZ:W&W project to reduce their dimensionality to D = 3

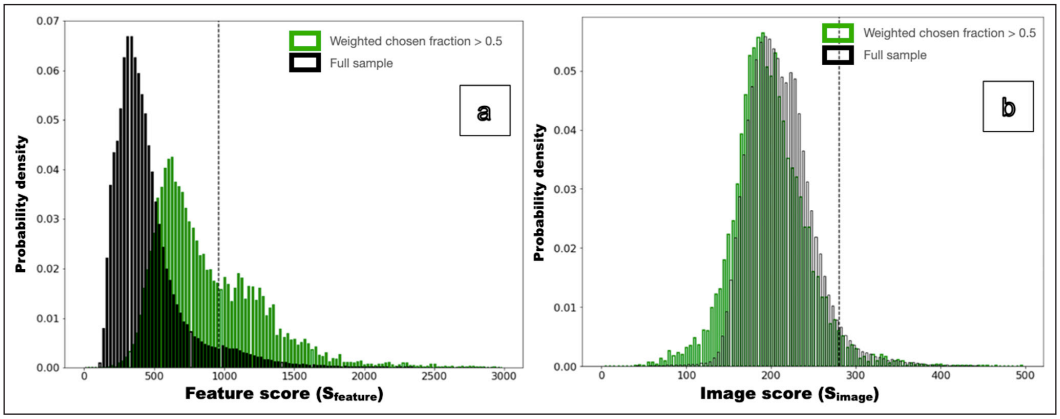

Figure 3 The probability distribution of the feature scores (*left*) and image score (*right*) for our entire sample (*black bars*, 99 percentile value in *black dashed line*) along with the subset that have weighted chosen fraction >0.5 (*green bars*).

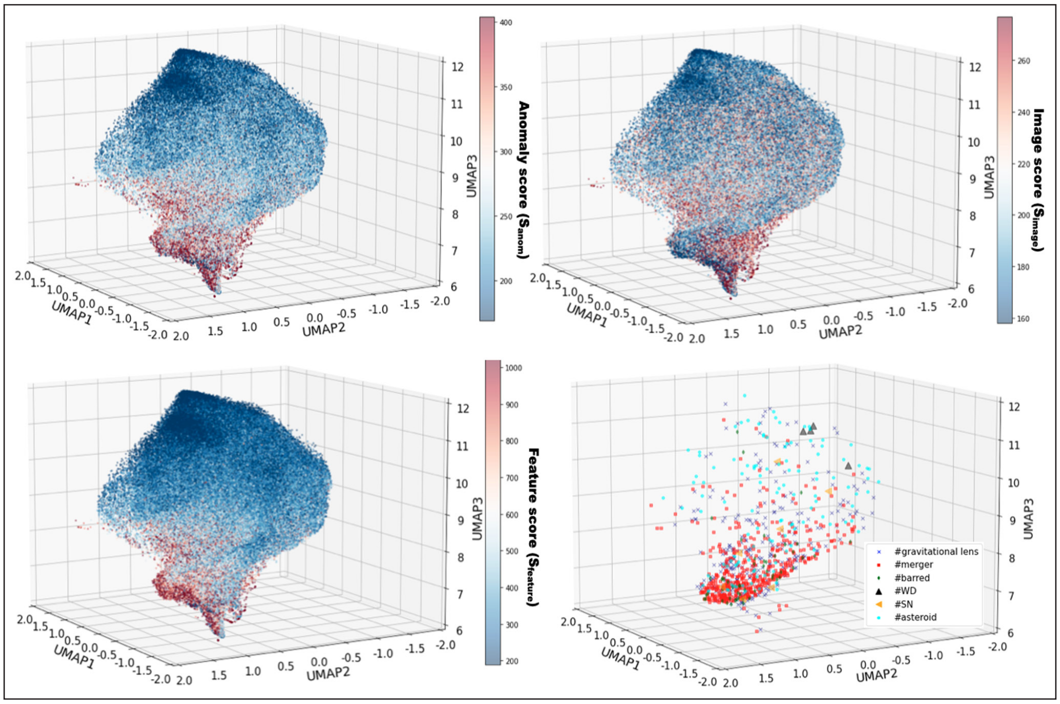

Figure 4 A visualization of the GZ:W&W: Galaxy Zoo: Weird & Wonderful subjects in the three prominent UMAP: Uniform Manifold Approximation and Projection dimensions, color coded by different quantitative metrics: anomaly score (panel a), image score (panel b), feature score (panel c). We also show those subjects that were #tagged in the "Talk" discussion boards (see legend; panel d; WC: #white_dwarf, SN: #supernova_candidates).

using the Uniform Manifold Approximation and Projection (UMAP) technique (McInnes et al. 2018). In Figure 4a,b,c, we show our images' feature space in three UMAP axes, color-coded with different metrics: S_{anom}, S_{image} and $S_{feature}$. We notice that the subjects with relatively high anomaly scores preferentially populate the lower portion of the UMAP space. However, it is quite interesting to note that subjects with high S_{image} and $S_{feature}$ populate different parts of the UMAP space. It is also worth noting that while the majority of the high S_{image} subjects span the lower portion of the UMAP space, a substantial number of them also span a broad range across the UMAP 3 dimension. However, this is not the case for the images with high $S_{feature}$, where they predominantly span only a localized region in the lower portion of the UMAP space.

In Supplemental Figure 13 of Supplemental File 1: Appendix G, we further illustrate the distribution of volunteer chosen fractions within the feature space as a function of the overall S_{anom} scores. First, we note that the images with higher weighted volunteer chosen fractions localize within a specific portion of the UMAP (see *top left panel*), with some overlap between the regions represented by high S_{image} and $S_{feature}$ scores. When assessing the UMAP locality by subsetting our images into those having high (>99%), intermediate (>68% and <99%), and low S_{anom} values (<68%), we find that a large portion of the chosen fraction ~0 subjects do correspond to low S_{anom} images populated in the upper portions of the UMAP. This suggests that a machine-based approach can be well-suited to filter out images that are likely to be unusual, but is less adept at clearly delineating which ones are interesting versus uninteresting. As will be discussed later, this suggestion comes with the caveat that the ability of the machine filtering is also dependent on the categories of interesting images. Of special note is a region in the lower part of the UMAP where the volunteer chosen fraction is low (e.g., 6 < UMAP3 < 7), but has images with high image or feature scores (as seen in Figure 4). We randomly sampled groups of 50 images that span this UMAP region and visually inspected by at least 3 domain experts. We find that a preponderance (>95%) of these images are predominantly image artifacts (colored streaks, large patches of noise, or saturated images) that the volunteers didn't select as being unusual or interesting. We show some examples of these in the Supplemental Figure 9 of Supplemental File 1: Appendix G. This observation particularly underscores the value that humans bring towards filtering out unusual but uninteresting objects within the data and highlights the potential for combining human and machine learning approaches.

Taking the above for context, we also assess the distribution of images in the UMAP space with a weighted volunteer chosen fraction >0.5 (shown in Supplemental Figure 14 in Supplemental File 1: Appendix G). This again highlights the specific region of localization of high volunteer-consensus images in the feature space. It also shows that when color-coding each data point with the S_{image}, and $S_{feature}$, a more apparent correlation with $S_{feature}$ can be seen where the majority of the images have higher $S_{feature}$ scores as indicated by the red color. On the other hand, no correlation is seen with S_{image}, where most of the images have low S_{image} scores. For more context on the overall ranges of the feature and image scores, see the Supplementary Figure 3 in Supplementary File 1: Appendix B, Figure 5 containing the distribution of S_{image} and $S_{feature}$.

FEATURE SPACE CORRELATIONS WITH TALK-BASED CHARACTERIZATIONS

Extending our previous exploration of the feature space correlations with the volunteer chosen fraction and anomaly scores, we analyze the feature space distribution of the subset of images (N = 3043) that were discussed in the Talk boards with their corresponding #tags. In Figure 4d, we show the images associated with a select subset of #tags in the UMAP space. Broadly speaking, it becomes evident that certain categories such as #mergers and #barred are highly grouped towards the lower portion of the UMAP space. While this is also generally true for the #gravitational_lens, #asteroid category and #supernova_candidates, a substantial portion of images with these tags (esp. #gravitational_lens and #asteroids) also are dispersed diffusely away from the general locus of points (see example images in Supplemental Figures 10 and 12 in Supplemental File 1: Appendix G). However, it is worth noting that images tagged with #white_dwarf or #white_dwarf_candidate (tags provided by a single volunteer to six images) solely resides away from the general locus of other categories. We show some examples of the images tagged with #gravitational_lens in Supplemental Figure 10 of Supplemental File 1: Appendix G.

While a thorough follow-up of various images containing interesting characteristics with expert verification is beyond the scope of this work, we note some preliminary advanced explorations done by engaged volunteers. For example, images tagged by #white_dwarf_candidate have been searched across existing astronomical catalogs and one of them was identified (by a volunteer) as known white dwarf (e.g., https://www.zooniverse.org/projects/zookeeper/galaxy-zoo-weird-and-wonderful/talk/subjects/87936472). Simultaneously, lists of images tagged as #gravitational lens have also been explored by volunteers and lists of objects identified (https://www.zooniverse.org/projects/zookeeper/galaxy-zoo-weird-and-wonderful/talk/4513/2899132) not identified (https://www.zooniverse.org/projects/zookeeper/galaxy-zoo-weird-and-wonderful/talk/4513/2912339) as being part of any published papers have been assembled. Our future work will focus on expert verification of these images and deriving important scientific outcomes. An example collage of images tagged by #gravitational lens are shown in Supplemental Figure 10 of Supplemental File 1: Appendix G and they indeed show gravitational lensing arcs.

It is worth refreshing that one of our main motivations is to enable research teams to make a well-informed selection that would reduce a large dataset down to a tractable number of scientifically interesting samples for further investigation/vetting. With this motivation in context, our above exploration highlights the caveat that if one were to focus only on selecting images within the general locus (see Figure 4d), it would cover a wide range of images containing a variety of characteristics; however, some rarer categories might be excluded. This also highlights the need to have additional modalities such as tagging (or potentially even semantic descriptions) as a way to help further interpret the feature space and volunteer consensus.

CLASSIFICATION BOUNDARY BASED ON LOGISTIC REGRESSION

With the aim of defining an efficient method for preselection of images from an unseen dataset, we leverage various correlations and insights discussed so far between the GZ:W&W consensus and anomaly detection model metrics to define a classification boundary within the latent space that can maximize the retrieval of volunteer-selected, interesting subjects. For this, we follow standard practice of applying Principal Component Analysis to extract the prominent features from D = 128 feature space vectors for all our images in the dataset. We find that a D = 25 account for a substantial (75%) of the total explained variance in the feature vectors, and as such we use these features for our next steps. Then, we split our 200,000 Σ_{CF} are given a class label = 1 (i.e., selected by volunteers) and vice-versa.

As a demonstrative example exercise, in Figure 5, we showcase the logistic regression decision boundary for a case of Σ_{CF} = 0.5. Generally speaking, when applied to our validation sample, the logistic regression selected images align with the general locus of images with a high volunteer chosen fraction. These selected images amount to ~20% of the total validation sample. We also assess the precision (fraction of correct positive predictions), recall (fraction of positive predictions that are correct) and the F1 score (measures the prediction accuracy by taking both precision and recall into account; F1 = precision x recall/precision+recall) of the decision boundary (see Figure 5c). Owing to the correlation between the feature score and the chosen fraction in the UMAP space, we are motivated to explore the precision, recall, and the F1 scores as a function of $S_{feature}$ score by iteratively limiting the sample to higher $S_{feature}$ values. We found that the F1 score (and as such the precision and recall) started around 0.2 and improved as we limited to images with higher $S_{feature}$. This behavior is reflected in the relative differences in the $S_{feature}$ distributions between the overall sample and the logistic regression selected images (see Figure 3).

While a threshold value of Σ_{CF} = 0.5 yields a sample of images that have a high chance of being chosen by volunteers, it can be seen that the UMAP feature space distribution of these images is quite restrictive. As discussed in the previous section, some specific images containing certain kinds of characteristics (e.g., #white_dwarf_candidates) populate in the part of the feature space that is not captured by the Σ_{CF} = 0.5 threshold. As such, the definition and derivation of a decision boundary becomes an optimization problem between effectively isolating the interesting images while minimizing the contaminants, all while ensuring that a maximal number of unwanted samples are taken out of consideration. This translates to understanding "what is the ideal value to choose for Σ_{CF}?".

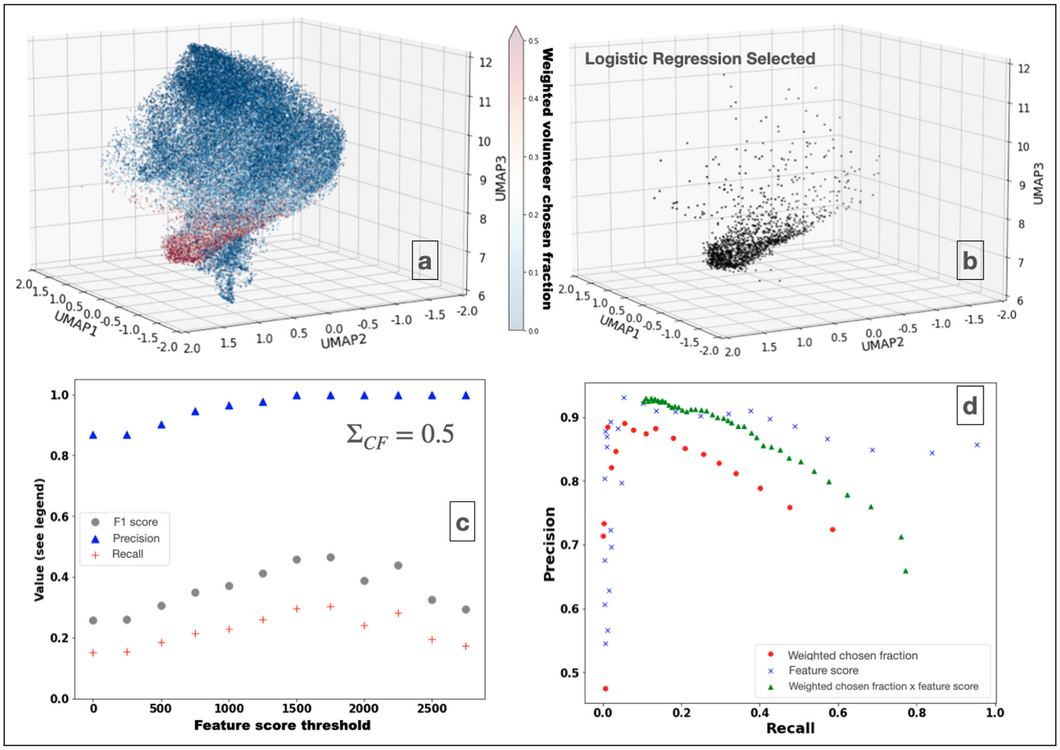

Figure 5 UMAP: Uniform Manifold Approximation and Projection distribution of a subset of images validated with our logistic regression decision boundary (panel a) and those images that were chosen as satisfying the decision criteria (black points; panel b), respectively. Panel c shows the P: precision, R: recall, and the F1 score (2PR/P+R) as a function of an applied lower-limit on the feature score: $S_{feature}$. Panel d shows the precision vs recall of various logistic regression decision boundaries where each of the three parameters are incrementally thresholded: weighted chosen fraction (red points), feature score (blue crosses), and a product of feature score and weighted chosen fraction (green triangles).

A liberal value of Σ_{CF} (e.g., 0.1) would aim at selecting a more complete sampling of a variety of characteristics in the data as opposed to, for example, $\Sigma_{CF} = 0.5$, which could miss certain image characteristics but maintain a higher purity with those that were selected.

Motivated by the previously discussed correlation between images' weighted chosen fraction, feature score, and the locality in the feature representation space (see Figure 4 and Supplemental Figure 10 in Supplemental File 1: Appendix G), we investigated whether a combination of the weighted chosen fraction and the feature score could better localize the interesting samples within the data than either quantity alone. As such, we iteratively ran the logistic regression decision boundary calculation by changing the threshold at which a quantity used is binarized. In Figure 5d, we show precision versus recall curve for the decision boundaries calculated by changing the threshold between $0.1 < \Sigma_{CF} < 1$. In the same figure, we also show the precision vs recall curves for two other scenarios where we used the $S_{feature}$ and varied it (as informed by their histograms) across a range of $300 < \Sigma_{feature} < 3000$, and a "combined" score that is a product of weighted chosen fraction and the $S_{feature}$ and varied it across a range $\Sigma_{feature} < 600$.

We notice that the combined product of weighted chosen fraction and the $S_{feature}$-based decision boundaries are generally more separable than the weighted chosen fraction, as evidenced by the systematically higher precision and recall values of the former when compared with the latter. While it is tempting to say that a simple $S_{feature}$ based approach yields better distinguishing boundaries, it should be noted that such an approach would select specific localities within the feature space that only sometimes overlap with the volunteer chosen consensus (see Figure 4 and Figure 5a for more context), but does not prioritize the selection of preferential selection of images with scientific interest.

In addition to the varying precision and recall values as a function of Σ values discussed above, we also showcase the fractional amount of sample selected by the decision boundary from an overall validation set in the context of the precision and recall values (black data points in Figure 6). We show this information for two scenarios: 1) binarization of weighted chosen fractions and 2) binarization of the product of the feature score and weighted chosen fractions. We interpret these results as the following: Assuming that we have a new dataset of images that haven't been inspected by volunteers and have been processed through our anomaly detection model, if we were to apply a decision boundary classifier within the feature space derived using a value of $\Sigma_{CF} = 0.1$ (see Figure 6a), then we would have a ~50% reduction in the image sample size with ~75% precision and ~58% recall in terms of containing images which would have a weighted chosen fraction >0.1. Applying decision boundaries derived using higher Σ_{CF} values (e.g., 0.5) will yield an ~80% reduction in the image sample size, along with > 80% precision and ~20% recall. However, as evidenced by the feature space distribution of weighted chosen fraction in Figure 4, employing a higher Σ_{CF} value comes with the risk of omitting certain images containing relatively unique (and plausibly rarer) characteristics.

Ongoing and upcoming astronomical surveys will both benefit from these methods and serve to vastly improve them. For example, the Vera Rubin Observatory will obtain 20 billion galaxy images over the time span of a decade. An 80% reduction in this data would still yield 400 million galaxy images that require inspection. However, continued implementation of HITL methods on incoming images will strengthen these methods in terms of reducing the number of images that require follow up, while also expanding the known feature space of typical galaxy images to allow for better characterization of potentially anomalous features.

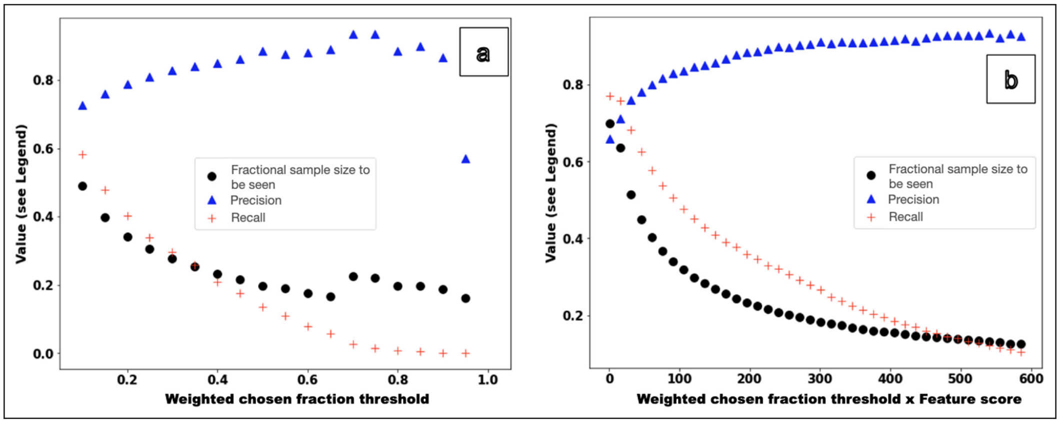

Figure 6 The precision and recall of the logistic regression decision boundary derived by varying the binarizing threshold: Σ_f for two different scores: weighted chosen fraction (left panel; Σ_{CF}) and the product of feature score and weighted chosen fraction (right panel; $\Sigma_{CF \times Feature}$). In each panel, we also show the overall fraction of a new sample of images that requires visual inspection as a function of Σ.

In addition to the above, comparing the trends shown in panels *a* and *b* in Figure 6, for example, we find that at threshold value $\Sigma_{CF \times feature} \sim 200$ for the combined score (product chosen fraction and the feature score) yields a decision boundary with ~85% precision and ~40% recall with ~20% effective remaining sample size of images (i.e., ~80% reduction in samples to be inspected). To achieve a similar recall of ~40% with the chosen fraction only case (panel *a*), it would mean that the effective sample size remaining would be at ~40%. Similarly, choosing a Σ_{CF} value (e.g., ~0.7) to achieve a similar precision of ~85% will yield an effective sample size of ~20%, but comes with a cost of low recall of <5%. Our insights from Figures 5 and 6 indicate that a combination of the chosen fraction and the machine-based feature score yields better discernibility in the feature space for pre-selecting potentially interesting images from a larger dataset.

CONCLUSIONS

Combining citizen science with machine learning methodologies to efficiently parse large astronomical data for finding anomalous and scientifically interesting objects is one of the critical challenges in the era of big-data astronomy. In this work, we explore some of the key questions emerging from the task of operationalizing a human-machine combined anomaly detection framework. We developed and applied a deep learning–based anomaly detection framework on a large dataset of 1.5 million astronomy images, which yields their learnt feature-level representations and an anomaly score metric that indicates how unusual a given image is. Using a subset of these images (~200,000) which also contained interspersed highly anomaly-scored ones (~15,000), we successfully ran a citizen science project (GZ:W&W), in which volunteers were asked to select images that they thought were interesting. By correlating the metrics from the GZ:W&W project (chosen fraction; the fraction of volunteers selecting an image to be interesting) with the anomaly detection based metrics (anomaly score which quantifies the unusuality of an image compared to a general population), in this case study, we present some insights into the relationship(s) between citizen science versus machine learning driven anomaly detection, and offer general recommendations on road-mapping our approach to other (potentially much larger) data domains. Below, we outline some of the main points from our work, split into high-level insights and summary of quantitative findings.

HIGH-LEVEL INSIGHTS FROM OUR CITIZEN SCIENCE PROJECT

1. When the volunteer choices based on their prior participation and experience in working with images containing galaxies were given more weight, we found that the resultant high-consensus images contained more rarer and anomalous features than those identified by the general volunteer population. Simultaneously, images containing more general signatures were preferentially downweighed. As such, volunteer experience weighting can enhance the finding of rarer and more interesting samples within the data.
2. Volunteers selected images containing a wide variety of characteristics. Images containing relatively more common but interesting features (e.g., colliding/merging galaxies) usually had a higher consensus (chosen fractions) compared with those low chosen fraction images containing more subtle, rarer, or challenging characteristics (e.g., gravitational arcs, galaxies with rings). This suggests that the chosen fraction consensus jointly traces a spectrum of volunteer knowledge/experience and the rarity and complexity of information that is being perceived.
3. While it is natural to consider discarding samples with low consensus for any downstream purposes, in the context of finding rare and scientifically interesting objects, one should not discount low consensus images as all it takes is one person to identify a specific characteristic within an image.
4. Deep learning–based anomaly detection methods can help swiftly identify images that are least likely to be unusual. Humans are particularly efficient at filtering out unusual and uninteresting objects, a task that machine learning is particularly less adept at.

SUMMARY OF MACHINE LEARNING CORRELATION FINDINGS

1. The experience-weighted chosen fraction has a stronger correlation with the feature score, a component of the overall anomaly score that is specific to the unusualness of low-level features within a particular image, than the image score. However, this correlation is still quite weak quantitatively.
2. Assessing the feature space representations of the images, those that have high feature scores occupy a notably different distribution (and are more spatially localized) when compared to those that have a high image score (another component of the anomaly score).

3. Images corresponding to higher weighted chosen fraction values formed a notable locus in the feature space with some correlation with the locality of those with high feature scores. Based on a subset of images that were tagged by volunteers in the Talk boards, images with certain relatively-high incidence characteristics (e.g., #merger and #ring) tend to be tightly grouped compared with some rarer categories (e.g., supernova candidates, gravitational arcs) that tend to have subtle characteristics.
4. We used logistic regression to define decision boundary conditions based on the feature representation of images to classify if it would be considered interesting. We explored using the weighted chosen fraction and feature scores as metrics to define if an image is to be treated as interesting or not. We find that the product of the feature score and the chosen fraction serves as a better metric to distinguish images in the feature space, while also ensuring that the selected number of images by the decision boundary is small (~20% of a total sample).

OPEN QUESTIONS AND MOTIVATIONS FOR FUTURE WORK

Herein, we explore the combination of machine learning and citizen science-based anomaly detection within astronomical imaging data. Throughout, we have made several choices, both from a machine learning standpoint and for our citizen science project (GZ:W&W). These choices naturally open various opportunities for further exploration in future works, which we discuss in this section along with potential next steps for consideration for the research community.

Firstly, for our anomaly detection model, we used one specific class of deep learning model (GAN). As elaborated in Supplemental File 1: Appendix B, there are several anomaly detection methods, including more recent (and potentially more robust) architectures, for example, stable diffusion (Rombach et al. 2022) transformer-GANs (Zhang et al. 2022). Exploring the anomalous samples returned by these different approaches is an open avenue for exploration and will illuminate if specific methods are more adept at identifying specific kinds of anomalies within data.

Even within the purview of our wGAN framework, we have made several architectural and hyperparameter assumptions. In fact, our Generator and Discriminator architectures consist of simplistic convolutional layers. A particularly interesting avenue of exploration could involve introducing "ResNet" like layers or the novel attention mechanisms within the architecture and quantifying the enhancements or differences in the feature level representations learned by the wGAN. Additionally, we made an important choice for the relative weightage between the feature and image scores (0.8 and 0.2, respectively) when optimizing the wGAN and encoder frameworks. Understanding the variation in our models' learning by smoothly varying the feature versus image score weighting parameter is an open question.

Another potentially interesting avenue for future exploration is to perform an iterative anomaly detection on a set of identified anomalies. Such a model would learn to re-generalize itself to a focused set of anomalies and will be better at embedding their characteristics into a potentially more separable feature space. This would yield a clustering of anomalies as per their similar characteristics in which deviations from individual clusters can be considered candidates of interest, and truly anomalous objects would stand out as outliers amongst all the clusters. Such an approach might prove valuable if one is trying to identify "rarest of the rare" objects within datasets.

In our citizen science project, GZ:W&W, we have also made a couple of assumptions and analysis choices. For example, we retired an image from further visual inspection if it had been seen by 10 volunteers. While this choice was motivated by various other citizen science projects, it is nevertheless interesting to test the outcomes and consequences of using a larger retirement limit and to gauge whether such a choice provides benefits in detecting certain anomalies (at the cost of longer completion time). Additionally, while quantifying the consensus chosen fraction of the selected images, we applied a simplistic weighting scheme that enhanced the choices of participants who had substantial participation (>100 classification) in the GZ project. More sophisticated weights can be assigned, for example, a continuous weighting scheme that is related to the number of GZ classifications.

Talk tags have played a key role in helping our interpretation of the volunteer-identified interesting images as well as correlating different categories of anomalies within the anomaly detection model learned feature space. It is worth noting that only a small fraction of the overall volunteer identified images have been discussed on Talk and tagged. Future works should aim to include a simultaneous free-form tagging task in addition to selecting the interesting images to procure a more complete set of characteristics. Such a dataset would be instrumental towards applying the latest ML methods such as large language foundation models (e.g., LLaMA, or vision-LLaMA) to unlock new exploration avenues for human-machine combined anomaly detection.

DATA ACCESSIBILITY STATEMENT

Machine learning model development and training were performed using Python's Tensorflow package (Abadi et al. 2015). All the figures were generated using Matplotlib (Hunter 2007). Tabular data analysis was made using Astropy (Robitaille et al. 2013). Computational analysis including machine model training and inference was done using Minnesota Super Computing Institute (MSI). In our work, we analyzed aggregated consensus of overall classification export from the Galaxy Zoo: Weird & Wonderful project across all participants (per image) and their classification/task answers, and any Zooniverse Talk participation information. We used the usernames of the participants to crossmatch between the Galaxy Zoo project and Galaxy Zoo: Weird & Wonderful project. Processed data tables including anomaly scores, volunteer unweighted and weighted chosen fractions along with talk tags (wherever applicable) is available on GitHub along with scripts that are able to parse the data produce key figures from this paper at https://github.com/AgentM-GEG/galaxy-zoo-weird-and-wonderful.

SUPPLEMENTAL FILE

The supplemental file for this article can be found as follows:

- **Supplementary File 1.** Appendix A – Details of imaging data used in this work; Appendix B – Choice of anomaly detection framework architecture; Appendix C – Selection of experienced Galaxy Zoo (GZ) project participants; Appendix D – Calculation of the experience weighted chosen fraction; Appendix E – Expert verification of volunteer-selected images; Appendix F – General roadmap for our methodological steps; and Appendix G – Additional Supplemental Visualizations. DOI: https://doi.org/10.5334/cstp.740.s1

ACKNOWLEDGEMENTS

We thank all the amazing volunteers of the Galaxy Zoo and Galaxy Zoo: Weird & Wonderful projects for their valuable participation.

FUNDING INFORMATION

KBM, LF, HR would like to acknowledge partial support for this work from the following grants: NSF IIS 2006894 and NASA Award # 80NSSC20M0057. CJL acknowledges support from the Sloan Foundation. RS acknowledges partial support from NASA Award # 80NSSC22K0804.

COMPETING INTERESTS

The lead author of this work directly collaborates and works with Prof. Lucy Fortson and group and Zooniverse, which includes Dr. Ramanakumar Sankar and Dr. Sarah Huebner.

AUTHOR AFFILIATIONS

Kameswara Bharadwaj Mantha orcid.org/0000-0002-6016-300X
University of Minnesota-Twin Cities, US

Hayley Roberts orcid.org/0000-0003-0046-9848
University of Minnesota-Twin Cities, US

Lucy Fortson orcid.org/0000-0002-1067-8558
University of Minnesota-Twin Cities, US

Chris Lintott orcid.org/0000-0001-5578-359X
University of Oxford, UK

Hugh Dickinson orcid.org/0000-0003-0475-008X
Open University, UK

William Keel orcid.org/0000-0002-6131-9539
University of Alabama, US

Ramanakumar Sankar orcid.org/0000-0002-6794-7587
University of California Berkeley, US

Coleman Krawczyk orcid.org/0000-0001-9233-2341
University of Portsmouth, UK

Brooke Simmons orcid.org/0000-0001-5882-3323
Lancaster University, UK

Mike Walmsley orcid.org/0000-0002-6408-4181
University of Toronto, CA

Izzy Garland orcid.org/0000-0002-3887-6433
Lancaster University, UK

Jason Shingirai Makechemu orcid.org/0009-0009-6545-8710
Lancaster University, UK

Laura Trouille orcid.org/0000-0002-1113-4122
Adler Planetarium, US

Clifford Johnson orcid.org/0000-0002-0511-6737
Adler Planetarium, US

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., et al. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283. *Software available from tensorflow.org*.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S.P., Bennert, N., Urry, C. M., Lintott, C.J., et al. (2009) Galaxy Zoo green peas: discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*.

399, pp. 1191–1205. DOI: https://doi.org/10.1111/j.1365-2966.2009.15383.x

Coffin, S.C., (2023) Redshift Wrangler: Conducting a citizen science study of extragalactic spectroscopy. *Rochester Institute of Technology*.

Goodfellow, I.J., Shlens, J. and **Szegedy, C.** (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Hoyle, B., Rau, M.M., Paech, K., Bonnett, C., Seitz, S. and **Weller, J.,** (2015) Anomaly detection for machine learning redshifts applied to SDSS galaxies. *Monthly Notices of the Royal Astronomical Society*. 452(4), pp. 4183–4194. DOI: https://doi.org/10.1093/mnras/stv1551

Hunter, J.D., (2007) Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(03), pp.90–95. DOI: https://doi.org/10.1109/MCSE.2007.55

Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J. and **Shao, L.** (2020) Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*. 23, pp. 2086–2099. DOI: https://doi.org/10.1109/TMM.2020.3007321

Liang, Y., Melchior, P., Lu, S., Goulding, A. and **Ward, C.,** (2023) Autoencoding Galaxy Spectra. II. Redshift Invariance and Outlier Detection. *The Astronomical Journal*. *166*(2), p.75. DOI: https://doi.org/10.3847/1538-3881/ace100

Lintott, C.J., Schawinski, K., Keel, W., Van Arkel, H., Bennert, N., Edmondson, E., Thomas, D., et al. (2009) Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo. *Monthly Notices of the Royal Astronomical Society*. 399, pp. 129–140. DOI: https://doi.org/10.1111/j.1365-2966.2009.15299.x

Lochner, M. and **Bassett, B.A.** (2021) ASTRONOMALY: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*. 36, p. 100481. DOI: https://doi.org/10.1016/j.ascom.2021.100481

Margalef-Bentabol, B., Huertas-Company, M., Charnock, T., Margalef-Bentabol, C., Bernardi, M., Dubois, Y., Storey-Fisher, K., et al. (2020) Detecting outliers in astronomical images with deep generative networks. *Monthly Notices of the Royal Astronomical Society*. 496, pp. 2346–2361. DOI: https://doi.org/10.1093/mnras/staa1647

McInnes, L., Healy, J. and **Melville, J.,** (2018) UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*. 3(29), p. 861. DOI: https://doi.org/10.21105/joss.00861

Robitaille, T.P., Tollerud, E.J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., et al. (2013) Astropy: A community python package for astronomy. *Astronomy & Astrophysics*. 558, p. A33. DOI: https://doi.org/10.1051/0004-6361/201322068

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & **Ommer, B.,** (2022) High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022*. pp. 10684–10695. DOI: https://doi.org/10.1109/CVPR52688.2022.01042

Rubner, Y., Tomasi, C. and **Guibas, L. J.** (2000) The earth mover's distance as a metric for image retrieval. *International journal of computer vision*. 40, pp. 99–121. DOI: https://doi.org/10.1023/A:1026543900054

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G. and **Schmidt-Erfurth, U.** (2019) f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*. 54. pp. 30–44. DOI: https://doi.org/10.1016/j.media.2019.01.010

Sharifi Noorian, S., Qiu, S., Gadiraju, U., Yang, J. and **Bozzon, A.** (2022) What should you know a human-in-the-loop approach to unknown unknowns characterization in image recognition. *In Proceedings of the ACM Web Conference 2022*. pp. 882–892. DOI: https://doi.org/10.1145/3485447.3512040

Storey-Fisher, K., Huertas-Company, M., Ramachandra, N., Lanusse, F., Leauthaud, A., Luo, Y., Huang, S., et al. (2021) Anomaly detection in hyper suprime-cam galaxy images with generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*. 508, pp. 2946–2963. DOI: https://doi.org/10.1093/mnras/stab2589

Straub, M.C.P., (2016) Giving citizen scientists a chance: a study of volunteer-led scientific discovery. *Citizen Science: Theory and Practice*, 1(1), p.5. DOI: https://doi.org/10.5334/cstp.40

Walmsley, M., Scaife, A. M., Lintott, C., Lochner, M., Etsebeth, V., Géron, T., Dickinson, H., et al. (2022) Practical galaxy morphology tools from deep supervised representation learning. *Monthly Notices of the Royal Astronomical Society*. 513, 1581–1599. DOI: https://doi.org/10.1093/mnras/stac525

York, D.G., Adelman, J., Anderson Jr, J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J.A., et al. (2000) The sloan digital sky survey: technical summary. *The Astronomical Journal*. 120.3, p. 1579. DOI: https://doi.org/10.1086/301513

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., et al. (2022) Styleswin: Transformer-based gan for high-resolution image generation. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11304–11314. DOI: https://doi.org/10.1109/CVPR52688.2022.01102

TO CITE THIS ARTICLE:
Mantha, KB, Roberts, H, Fortson, L, Lintott, C, Dickinson, H, Keel, W, Sankar, R, Krawczyk, C, Simmons, B, Walmsley, M, Garland, I, Makechemu, JS, Trouille, L and Johnson, C. 2024. Through the Citizen Scientists' Eyes: Insights into Using Citizen Science with Machine Learning for Effective Identification of Unknown-Unknowns in Big Data. *Citizen Science: Theory and Practice,* 9(1): 40, pp. 1–15. DOI: https://doi.org/10.5334/cstp.740

Submitted: 16 February 2024 **Accepted:** 23 September 2024 **Published:** 09 December 2024

Citizen Science: Theory and Practice is a peer-reviewed open access journal published by Ubiquity Press.