CITIZEN SCIENCE:
THEORY AND PRACTICE

Understanding Confusion: A Case Study of Training a Machine Model to Predict and Interpret Consensus From Volunteer Labels

COLLECTION:
THE FUTURE
OF ARTIFICIAL
INTELLIGENCE AND
CITIZEN SCIENCE

RESEARCH PAPER

]u[ubiquity press

RAMANAKUMAR SANKAR
KAMESWARA MANTHA
COOPER NESMITH
LUCY FORTSON
SHAWN BRUESHABER
CANDICE HANSEN-KOHARCHECK
GLENN ORTON

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

ABSTRACT

Citizen science has become a valuable and reliable method for interpreting and processing big datasets, and is vital in the era of ever-growing data volumes. However, there are inherent difficulties in the generating labels from citizen scientists, due to the inherent variability between the members of the crowd, leading to variability in the results. Sometimes, this is useful — such as with serendipitous discoveries, which corresponds to rare/unknown classes in the data — but it might also be due to ambiguity between classes. The primary issue is then to distinguish between the intrinsic variability in the dataset and the uncertainty in the citizen scientists' responses, and leveraging that to extract scientifically useful relationships. In this paper, we explore using a neural network to interpret volunteer confusion across the dataset, to increase the purity of the downstream analysis. We focus on the use of learned features from the network to disentangle feature similarity across the classes, and the ability of the machines' "attention" in identifying features that lead to confusion. We use data from Jovian Vortex Hunter, a citizen science project to study vortices in Jupiter's atmosphere, and find that the latent space from the model helps effectively identify different sources of image-level features that lead to low volunteer consensus. Furthermore, the machine's attention highlights features corresponding to specific classes. This provides meaningful image-level feature-class relationships, which is useful in our analysis for identifying vortex-specific features to better understand vortex evolution mechanisms. Finally, we discuss the applicability of this method to other citizen science projects.

CORRESPONDING AUTHOR:
Ramanakumar Sankar
University of California, Berkeley, US
ramanakumar.sankar@berkeley.edu

KEYWORDS:
citizen science; crowdsourcing; auto-encoder; semi-supervised network; machine attention; planetary atmospheres

TO CITE THIS ARTICLE:
Sankar, R, Mantha, K, Nesmith, C, Fortson, L, Brueshaber, S, Hansen-Koharcheck, C and Orton, G. 2024. Understanding Confusion: A Case Study of Training a Machine Model to Predict and Interpret Consensus From Volunteer Labels. *Citizen Science: Theory and Practice,* 9(1): 41, pp. 1–14. DOI: https://doi.org/10.5334/cstp.731

INTRODUCTION

Across various scientific fields, such as astrophysics, improvements in data collection methods (including robotic telescopes, planetary missions, etc.) have led to a growing need to quickly and accurately process the data, and reduce information from raw (usually imagery) to more useful data products (e.g., labels for the dataset, or annotations of specific features) (Lintott et al. 2008; Fortson 2021; Zou et al. 2024). As computing tools such as machine and deep learning pipelines have risen at a commensurate rate, many studies have turned to using AI to process a large chunk of the data (Zou et al. 2024), reducing the overhead on doing scientific research. However, these tools are only as good as the data that they are trained on, and struggle to effectively process data that is dissimilar to the training dataset (Walmsley and Scaife 2023). This is a problem for researchers, since the advancement of scientific knowledge requires the accurate detection and processing of the new and unknown. Particularly, while methods are emerging to detect these new and unknown data (i.e., anomalies) in large datasets (Ishida et al. 2021; Lochner and Basset 2021), understanding why they are interesting and comparing them with known labels is challenging. Unsupervised deep learning methods have attempted to tackle this problem, but more often than not, these architectures have difficulty in differentiating between known image artifacts or noise, and new, scientifically significant data (Mantha et al. 2024; this collection), and have struggled even more in providing interpretation to the model-filtered data. To tackle the ever-growing volume of data, and to improve the scientific returns of these datasets, we require models that are efficiently able to identify and characterize anomalies that provide scientific value to the dataset and provide some measure of interpretability (i.e., relation between the features in the data and associated class labels). Our efforts in this work are to tackle the latter problem, specifically to provide characterization and interpretability of the dataset using machine models.

One potential option for help is through enlisting the general public to crowdsource anomaly detection and characterization in large data sets, commonly known as citizen science or participatory science. However, despite the rise in popularity of the citizen science methodology, there are inherent difficulties in the label generation process: For example, since the processed information is crowd sourced, there is an issue of variability between the members of the crowd, leading to large variability in the processed results, necessitating either sophisticated algorithms to detect and correct mislabeled data (Krivosheev et al. 2020) or restricting complicated labeling to experienced volunteers (Kosmala et al. 2016; Zevin et al. 2024). In some cases, this "confusion" (which we define as variability in the labels between different volunteers for the same data, or low confidence from the machine model) is useful (e.g., high confusion could lead to serendipitous discoveries, since it likely corresponds to rare or unknown classes in the data; Cardamone et al. 2009), but at other times, it might be due to lack of training samples for the citizen science volunteers or ambiguity between classes (Zevin et al. 2024). Therefore, a primary concern is discerning between the intrinsic variability in the dataset and the uncertainty in the citizen scientists' responses (Hunter et al. 2012; Li et al. 2020).

To be clear, both machines and humans are susceptible to confusion arising from the variability in the data set or the ambiguity between classes. However, because we can quantify the machine response in a more statistically meaningful manner, we can leverage this strength of the machine to deconstruct what may be confusing features for humans, whereas humans are good at finding the odd relationships between features because they can quickly learn and retain context (e.g., Cardamone et al 2009). Processing such data with contextual information can provide interpretation and meaning to the intrinsic variability within a dataset. Our motivation here is to provide a mechanism using the combined strengths of humans (in their capturing of context) and machines (in their quantification of the feature relationships) to identify scientifically interesting feature relationships by using the information in spurious labeling coming from human uncertainty.

Work to date has shown that deep neural network–based mechanisms have vastly reduced volunteer efforts by quickly labeling "easy" data, while distributing the more complicated data to volunteers (Richards et al. 2011; Willi et al. 2018; Walmsley et al. 2019; Sankar et al. 2023). Although these techniques have improved the labeling efficiency and scientific throughput of citizen science projects, they do not offer much information about why specific subsets of the data are more complicated, or why they confuse the machine models. Indeed, while machine models (specifically deep neural networks, due to their black-box natures) produce good results for their training data, they tend to fail spectacularly on new datasets, necessitating complicated metrics and training regimens to improve generalizability (Liao et al. 2021; Mantha et al. 2022). Simultaneously, deep neural networks have shown effectiveness in their ability to automatically extract features for classification and clustering (Syarif et al. 2012), which are useful for drawing meaningful relationships from large datasets (e.g., Storey-Fisher et al. 2021; Etsebeth et al. 2024). These dataset features provide basic context between the downstream task that the machine model is trained on (e.g., classification) and what information the model used for their respective tasks. As such, these techniques enable us to provide a measure of interpretability to models, allowing for more sophisticated understanding and control over the machine performance.

Motivated by these advancements, we use deep neural networks in a citizen science project to tackle the problem from both sides: 1) to augment citizen science data, especially those with high volunteer confusion, using a machine model to characterize the feature-class relationships, and 2) to increase trustworthiness of machine models applied to scientific datasets by providing a measure of interpretability to the machine's prediction, specifically when it pertains to anomalous or confusing data. Fundamentally, our overarching goal is to improve the purity of the downstream scientific outputs by having a better characterization of the human confusion and the machine-derived, feature-class relationship. We organize herein by first describing the dataset shown to the citizen scientists, as well as the classes and the labeling strategy. We then describe the results from our machine model and how we use it to infer and interpret scientifically interesting feature relationships. Finally, we discuss the relevance to other citizen science projects, and the generalization of the methodology used here, and discuss the nuances of this work.

Summarily, we use a variational auto-encoder (VAE) with convolutional layers to learn spatial morphologies from our dataset, which consists of images of Jupiter's atmospheric features from the JunoCam instrument on board the Juno spacecraft (Hansen et al. 2014). This VAE is combined with a classification layer (which we refer to as a cVAE) and trained on the labels provided by citizen science volunteers from the Jovian Vortex Hunter (JVH) citizen science project on the Zooniverse platform (Sankar et al. 2024). Zooniverse is a web-based citizen science platform where research teams can upload their data and create simple interfaces for citizen scientists to interact with, for classification, annotation, or other tasks. The volunteers were shown the dataset and asked to label the atmospheric features within. The combination of the VAE architecture and the machine classification layer results in a "semi-supervised" framework that can be used to map and relate class confusion from the volunteer annotations with the morphological characteristics in the images. Beyond the initial task of simply improving the reliability of the characterization of machine and volunteer confusion, we further use the learned feature representation and network architecture to simplify the confusion space in the dataset to quickly extract meaningful scientific insights from the labeled dataset.

DATA

JOVIAN VORTEX HUNTER CITIZEN SCIENCE PROJECT

The Jovian Vortex Hunter (JVH) is a citizen science project hosted on Zooniverse.org (Sankar et al. 2024). The goal of the project is to identify atmospheric vortices on Jupiter in images taken by the JunoCam instrument (Hansen et al. 2014) on board the Juno spacecraft, in order to better understand the jovian atmospheric structure and dynamics. JunoCam is itself a citizen science project where the raw data are processed mainly by amateurs, and has led to amazing results (see https://www.missionjuno.swri.edu/junocam/processing?source=public). The JVH project launched in June 2022, and after more than a million classifications, successfully completed its first round of data in December 2023. We presented 68,322 image crops from 23 orbits of the Juno spacecraft (13 through 36). Specifically, we stack and mosaic the individual JunoCam images onto a global map. We then make random crops, using an equal-area projection, such that each crop measures 7000 km × 7000 km, and is at least 1500 km from a neighboring crop. In this way, the same atmospheric feature can be seen across multiple crops, ensuring that we are accurately sampling the region around a feature of interest. The project consisted of two workflows:

1. Is there a vortex?: In this first workflow, volunteers were shown an image from JunoCam and asked to identify the features in the image from a list of: vortex; turbulent features (or folded filamentary region [FFR]); cloud bands; pixelated; or blurry. The first three correspond to atmospheric classes of features, whereas the last two correspond to (poor) quality of the image. The volunteers could select multiple options per image since features can coexist in the same image (i.e., a vortex within a cloud band). We required at least 10 classifications from independent classifiers before retiring the image from the data pool.
2. Circle the vortex: this workflow was seeded by images that had high confidence of "vortex" from the first workflow. Here, we asked volunteers to circle the vortices in the image based on their color. The color represents unique chemistry and cloud microphysics within that region and presents an opportunity to understand the link between the fluid dynamics and the cloud chemistry (a currently poorly understood facet of the jovian atmosphere). Here, we required at least 12 independent annotations per image before retiring the image from the subject pool.

For this study, we primarily focus on the data from the first workflow because that gives us a better understanding of class confusion. Particularly, we are interested in seeing whether machine learning techniques can be used to augment the data provided by the volunteers with a generative model that can also learn the intrinsic features within the image. Ideally, we would like to determine whether specific correlations exist between the confusion derived from volunteer votes for a given class and type of image-level features (for example, is it harder to identify vortices if the images have a certain color, or have a high fraction of turbulence?). Examples of the workflow and the dataset are shown in Supplemental File 1: Model Description and Additional Results.

DATA CLASSES AND DETERMINING CONSENSUS

The dataset primarily consists of three main classes that are used in downstream analyses: vortex, FFR, and cloud bands. An example of each class is shown in Figure 1. The vortices are characterized by circular features, with radial color gradients (Ingersoll et al. 2007). They come in a wide spectrum of sizes from between 100–200 km (about 5–10 pixels) to a few thousand kilometers (covering nearly the full image). Vortices on Jupiter are driven primarily by shear instability in the fluid, and the colors of the clouds within the vortex are dictated both by the local temperature structure and by chemistry. Vorticity dynamics is a very useful proxy to understand the local fluid dynamics, which is inherently difficult to achieve without local sounding data or in-situ probe data.

FFRs share some similarity with vortices, since they also contain circular structures (Orton et al. 2017). However, as their name suggests, they are primarily filamentary features, which are characteristic of local turbulent mixing in the atmosphere. The folded filamentary structures sometimes contain small vortices, but these are mostly short-lived. Owing to the lack of high-resolution observations of these features, very little is understood about their dynamics and longevity (Orton et al. 2017; Hueso et al. 2022).

Cloud bands are much more prominent and well known in the jovian atmosphere. They are long-lived features, mostly near the equatorial regions, where sharp transitions in the east-west jet streams result in temperature variations in the atmosphere and lead to different cloud chemistry and color variations. While observations of the global cloud bands have existed for a long time, perturbations and formation of instabilities are poorly characterized (Hueso et al. 2022), but are likely vital for global energy balance on Jupiter (Ingersoll et al. 2000). As such, while the locations of the cloud bands are well known, characterizing the variability in cloud band–like features is vital in determining and characterizing turbulent eddy formation in the jovian atmosphere.

For these three classes in the dataset, we define volunteer consensus as the ratio of the number of volunteers who selected the classes to the number of volunteers who saw the image. Since each volunteer can select multiple classes per image (as each image could contain one or more classes, such as a vortex within a cloud band), the consensus for each class within an image is independently determined (e.g., the vortex consensus for an image is the number of volunteers who selected the vortex class, divided by the total number of volunteers who classified that image).

One of the primary issues with this classification task is that images might contain features that reside in the boundary between classes or share common features across one or more classes (for example, vortices and FFRs both feature sharp color gradients and turbulent patterns, making it difficult to distinguish the two features). In these scenarios, it is very difficult to disentangle and consistently

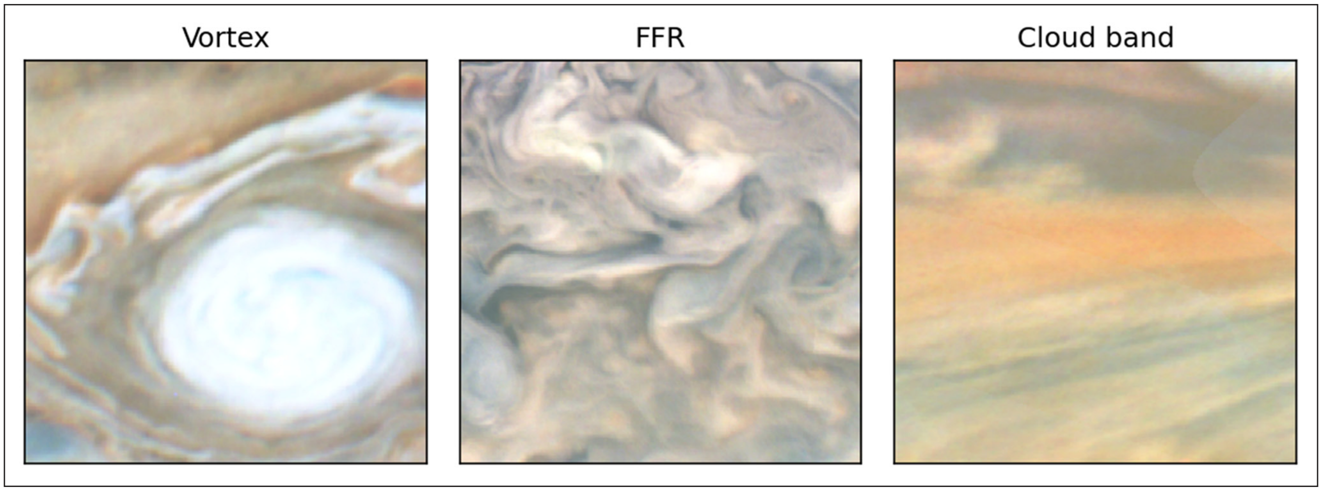

Figure 1 Examples of each class in the dataset: Vortex, Folded Filamentary Region (FFR) and cloud bands. These represent images which contain stereotypical features corresponding to each class.

bin datasets into separate classes, which share feature similarity. As such, we define volunteer confusion as a circumstance in which the volunteer consensus is not at the extremes (i.e., either one or zero), but somewhere in between. Our work involves disentangling two ways to obtain low consensus (volunteer confusion): The first is image-level features that do not conform to representatives of a class; the second is feature-class confusion, where there may be strong similarities between the features of two separate classes. For example, Figure 2 shows images with confusion (i.e., low volunteer consensus) in the vortex class, but with varying confusion across the other two classes, showing how the features in each image can be specifically tied to individual classes or confused between classes. In the far left, we get examples of image-level confusion for a given class (i.e., the features in the image do not necessarily define vortices well, but there are no signatures of the other two classes). In the far right, we get images that clearly represent FFRs or cloud bands but they also contain some signatures of vortices (e.g., FFR spirals can be easily mistaken for vortices), resulting in low consensus for the vortex class. This is an example of the second type of confusion illustrated above, showing that for a given image, the features that represent vortices may be shared by other classes (e.g., vortices and FFRs both contain bright white swirls). Alternatively, features in the same image may be characterized with high confidence, for example, cloud bands might not share much feature confusion with the other classes, and so are easier to classify, but there are still features pertaining to vortices within that image. In each of these cases, the characteristic features within the image are markedly different, which makes it difficult to interpret why the consensus was low or interpret the diversity of features that lead to confusion. Therefore, because we can generate independent per-class consensus for each image, we can use that information with machine learning to obtain a better understanding of which feature combinations for a given class lead to stronger image-level confusion or signify strong feature similarity between classes.

Through this case study on the JVH data, we aim to enable future research teams using citizen science data to better identify feature-class confusion, as well as image-level features that lead to volunteer confusion. By using a machine model to better identify the source of the volunteer confusion, we believe that this will ultimately accelerate the process by which research teams explore the dataset in search of those objects with high scientific value.

RESULTS

Once the volunteer labels were obtained through JVH, we trained our machine model (cVAE) on the JVH dataset. The details of the model, training process, and classification performance are provided in the Supplemental File 1: Model Description and Additional Results.

USING THE cVAE DATA TO IDENTIFY FEATURE-CLASS RELATIONSHIP

The cVAE model learns and maps information between the image-level features in the dataset and the corresponding classification labels. Specifically, we can use the information mapped by the cVAE to better understand what features in the image correspond to specific classes. More importantly, the network implicitly learns how these features share similarities between classes, driving the ability to understand feature confusion. We have two modalities of information from the cVAE: firstly, the latent space, which maps feature similarity, and secondly, the feature-classification relationship. We detail both components below.

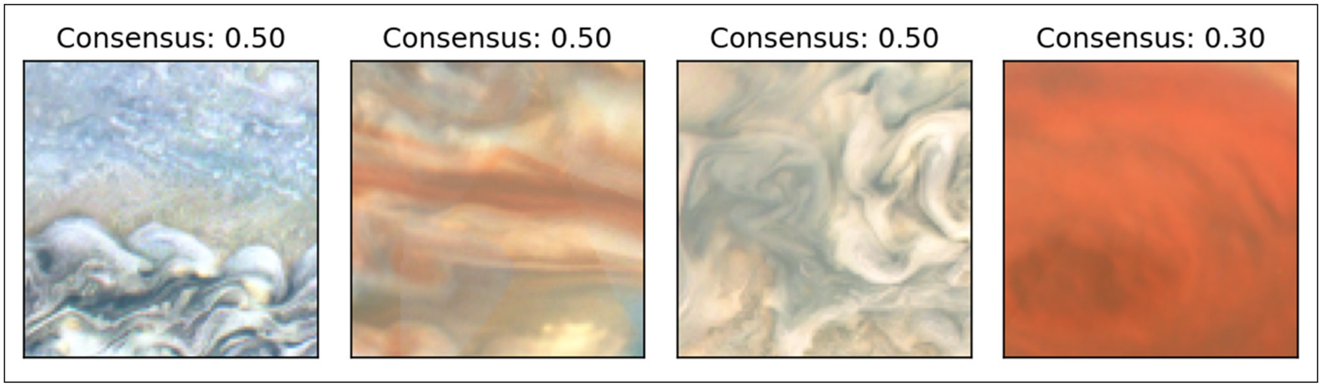

Figure 2 From left to right, we increase the Folded Filamentary Region (FFR, *top row*) and cloud band (*bottom row*) consensus showing images where volunteers were confused about vortices (consensus between 0.3–0.7), while varying in their confidence of the other two classes in the image. Note that from left to right we get clearer signatures of the other two classes (e.g., more swirls for FFRs and clear north-south color gradients for cloud bands).

Model latent space

Figure 3 shows the latent space from the model, compressed to 2 dimensions using the uniform manifold approximation and projection (UMAP). Each point corresponds to an image and is colored with the model-predicted class probability on the top and the volunteer consensus on the bottom. The variation in the latent space is directly related to the intrinsic classes within the image (since the latent vectors contain information about the spatial distribution and morphologies of the atmospheric features in the image). Certain locations are tied to images that contain only one class (e.g., vortex or cloud band), other locations contain a mixture of the classes (e.g., vortices in cloud bands, etc.). Therefore, the latent space is vital to understanding how image-level features can contribute to feature confusion (i.e., class overlap in the latent space signifies feature overlap between the classes), and it provides a method to identify common features for a given class (i.e., regions in the latent space containing only high consensus of cloud bands contain features corresponding only to cloud bands). A description of the feature variation within the latent space with relation to the classes is provided in the Supplemental File 1: Model Description and Additional Results.

Feature localization

The cVAE also has the ability to identify defining characteristics in the image that lead to specific classification. Several techniques exist for defining this "attention," including, for example, GradCAM and GradCAM++(Selvaraju et al. 2016; Chattopadhyay et al. 2017), which rely on the propagation of the neural network's gradient from the classification layer back through the input layer. In this study, we instead use the implementation of ScoreCAM (Wang et al. 2019), which removes the dependence on the gradients by modifying the activations of individual layers and inferring the corresponding effect on the classification. In this way, ScoreCAM provides a more robust representation of the individual morphological features in the input image that led to a given classification.

We use ScoreCAM to identify local features within the image that lead to class confusion by virtue of being common to multiple classes. Simply, the model consensus and the volunteer confusion provide a method to subset images where feature confusion could exist, while ScoreCAM is used to identify the features used in the image for making these classifications. Overlap in the feature "attention" (i.e., the model using those image-level features for the classification of the corresponding label) between different classes for confusing images indicates that the feature is commonly shared between these classes. For example, Figure 4 shows several input images, along with their corresponding attention map, showing the locations of the features most directly related to the vortex and the FFR classes. Here, we can see how the network attends

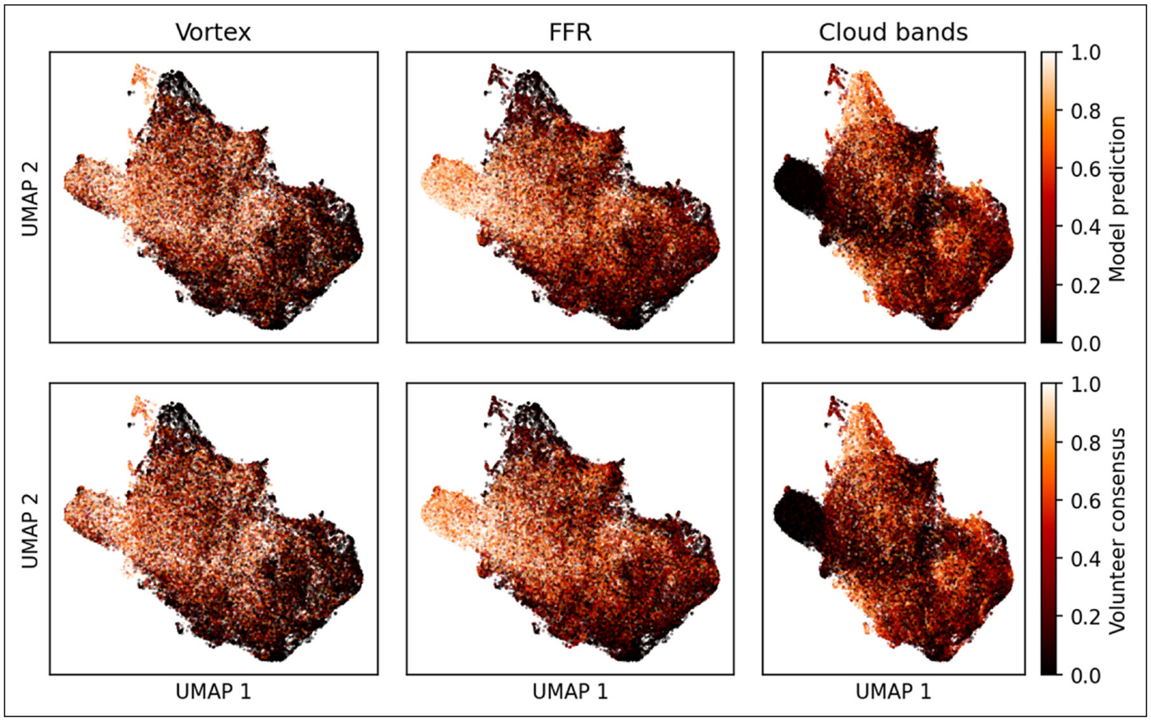

Figure 3 Latent space from the conditional variational autoencoder (cVAE) encoding colored using the model class probability (*top row*) and volunteer consensus (*bottom row*). Each point corresponds to an image, where two images that share similar characteristics are close together while those that are markedly different are farther apart. Note how the location on the latent space is strongly correlated with the class, showing that there is a strong relationship between the feature morphology and the corresponding class.

to features that have circular signatures for making the vortex prediction, and specifically looks at locations where the color gradient is high. These activations change for different classification targets where, for example, for FFRs, the network attends to bright regions within the image.

We leverage these attention maps to identify and diagnose class-wise confusion in our dataset. Figure 5 shows the attention maps for vortex, FFR, and cloud bands for a selection of images that have low vortex consensus. Here, we see that the attention maps between the vortex and other classes have significant overlap, showing that these features share similarities across classes. Furthermore, the attention points to locations within the image where the network has learned to look for vortices. Therefore, it is likely that within the features attended to by the network, there are other images in the dataset where vortices exist. For the images shown in Figure 5, we can sample the 5 closest images in the latent space (which would have similar morphological characteristics) that have high vortex consensus from volunteers, as shown in Figure 6. Here, observe that the features corresponding to where attention is high in the reference image have vortices

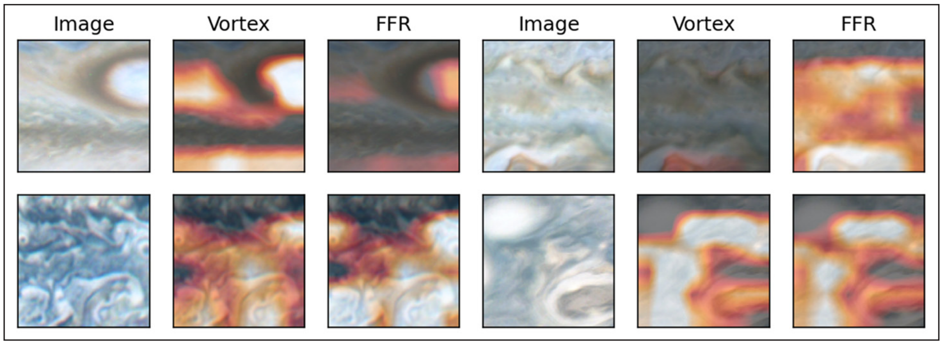

Figure 4 There are four sets of example images with each set having at left the original image used in model training, followed by the model's attention for the vortex and Folded Filamentary Region (FFR) classification, middle and right, respectively. Lighter colors denote higher attention by the model to those features, while darker regions are attended to less. Notice how the network attends to the vortex itself, or regions of color gradients, which are characteristic of either vortices themselves, or locations where vortices exist, to make the vortex prediction, but attends to different features for the FFR classification. In the bottom two panels, the regions attended to by the network contains signatures of both vortices (i.e., sharp color gradients and circular features) and FFRs (i.e., bright white, turbulence).

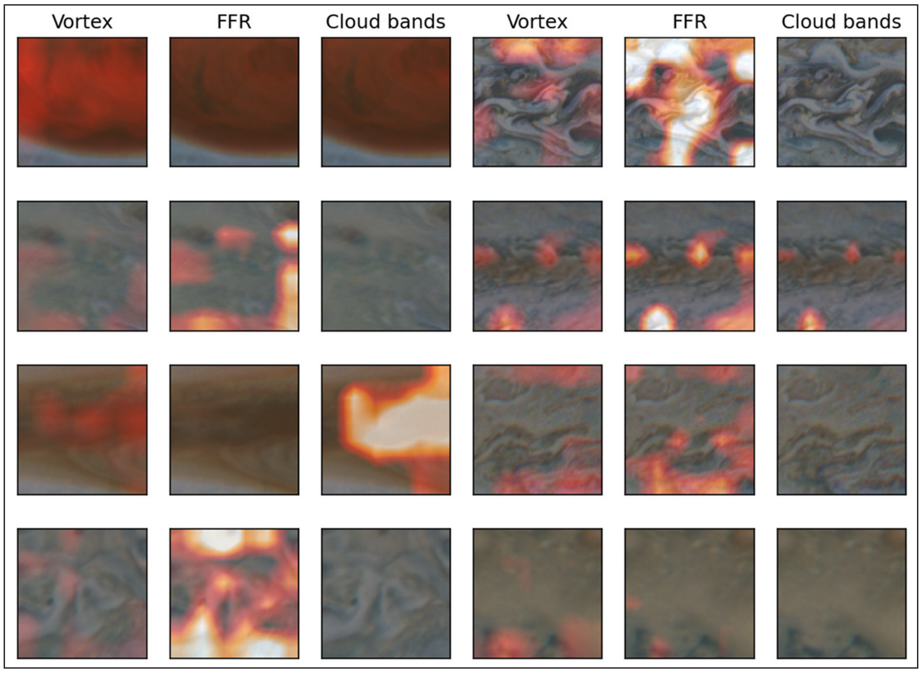

Figure 5 Example of images with vortex confusion with their attention for each of the three classes. Each row represents two sets of images with each set of three images corresponding to the three classes (Vortex, Folded Filamentary Region, and cloud bands). Notice how the attention for the vortex shifts to other classes based on their morphologies.

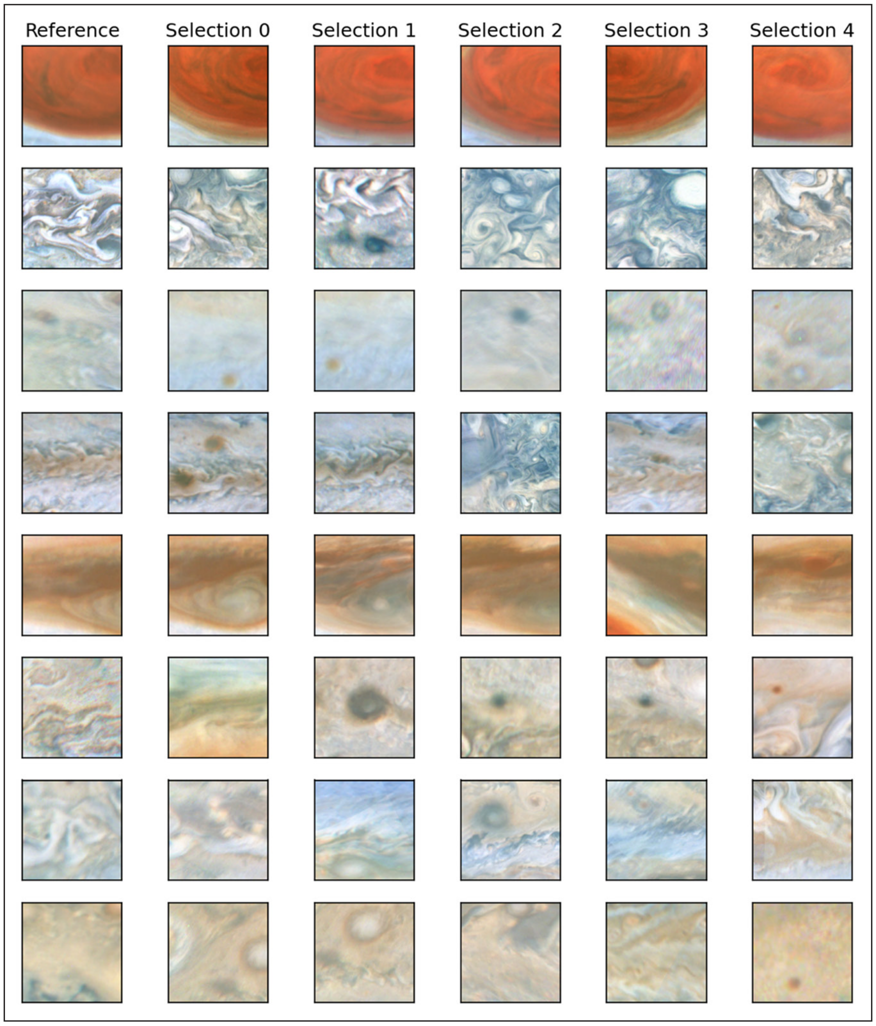

Figure 6 Examples of images with high vortex consensus based on their closeness to the reference image (see Figure 5 for reference image attention). Each row corresponds to a different image in Figure 5, and columns 2–6 correspond to different neighboring images in the model latent space with high vortex confusion. Vortices in the image form in features that the model attends to (i.e., bright values in the attention overlay) in Figure 5.

in the sampled images. This is particularly true within FFRs (row 2) and high shear between cloud bands (rows 4 and 5). However, surprisingly, this is also true of locations where the color gradient is relatively flat (row 3).

Therefore, the ScoreCAM-based attention maps highlight important information that relates the image-level features with the classification labels. These are particularly useful when discerning between feature similarity for multiple classes (thereby helping us understand the class confusion).

INTERPRETING CONFUSION FROM VAE-ADDED DATA

The cVAE, therefore, provides a multi-dimensional view into the dataset, using the latent space to identify feature similarity and the attention to highlight feature importance towards a specific class. These two modalities are used to understand feature relationship in the dataset and possibly disentangle class confusion due to feature similarity. Given the large volume of data (>68,000 images), it is important to identify methods to simplify the process of characterizing confusing samples. Specifically, we highlight the use of these two modalities as a method to quantify and explore the diversity in the confusion, which simplifies the process of characterizing the source of confusion in the data and their downstream scientific value. Here, we use the latent space as a method to subset data and identify feature similarity, and the ScoreCAM attention to identify class-feature relationships. Both are vital for understanding class confusion in the dataset and contain key scientific value, as detailed below.

Feature diversity for confusing targets

The latent space generated by the cVAE is useful for disentangling different subclasses of confusing features.

For instance, Figure 7 shows the distribution of the latent space for confusing vortices (volunteer consensus for vortices between 0.3 and 0.7) along with the corresponding consensus for the other classes. This allows us to distinguish between images that show poor vortex characteristics as a function of the morphological features within the image. Since the latent space encodes the morphological features, we can investigate the latent space for the confusing features to investigate qualitative correlations between the vortex features and other classes.

Figure 8 shows the filtered latent space from various samples overlaid. We see that the latent space shows a distribution of morphological characteristics in the image. As stated above, the use of the latent space simplifies the exploration of the diversity of confusing samples in the dataset, which makes it easier for us to characterize the different sources of confusion. Here, we find that the confusion in the dataset (specifically for vortices as highlighted in Figure 8), appears to be from the diversity of features that correspond to vortices. For example, the images

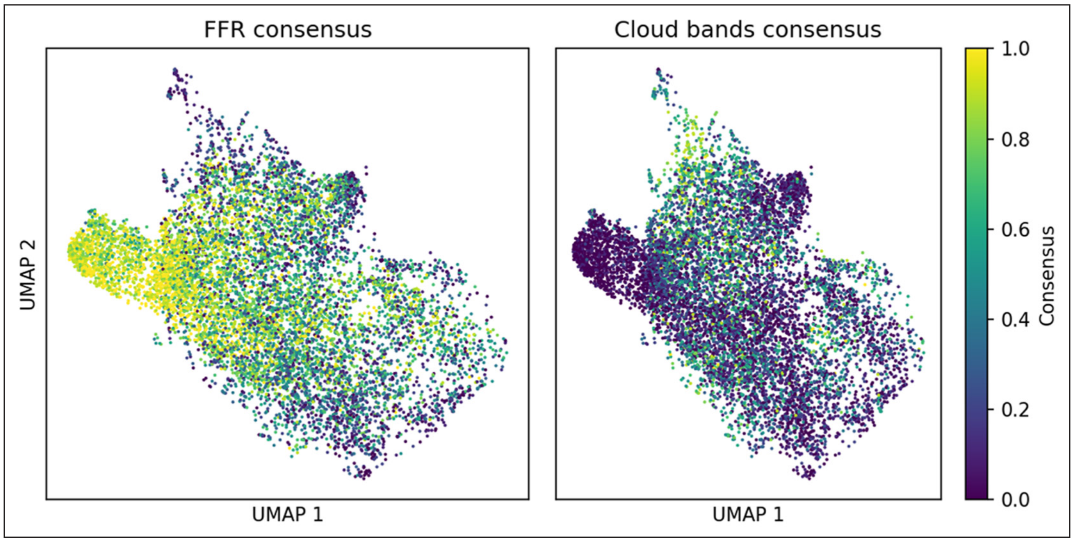

Figure 7 Distribution of the latent space for the confusing vortex sample, colored by Folded Filamentary Region (FFR) and cloud band consensus. Note how there are separate regimes of features that contain high/low consensus and are located in different latent space locations.

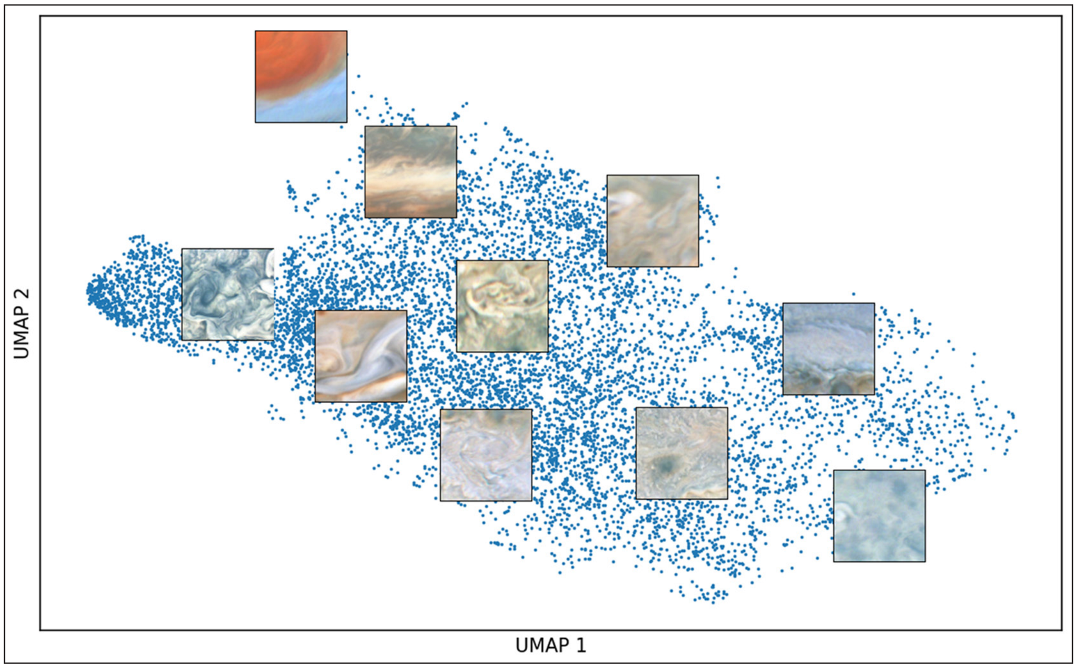

Figure 8 Latent space distribution for the confusing vortex sample but with images overlaid showing the types of phenomenon in each image. Note how specific regions seem to show vortical phenomena while others do not.

with the Great Red Spot (GRS; top left) show poor consensus, particularly when the core of the storm is not visible. This is likely due to the fact that the GRS was easily confused either with a similar (but significantly smaller storm) which was also red, or with cloud bands (Figure 4 shows an increase in cloud band consensus near this cluster).

In other areas, for example, on the far left, FFRs were much more prominent, and volunteers struggled to accurately differentiate between swirls in the FFRs and vortices. Indeed, here the definition of the vortex breaks down without having access to data that showcases the temporal evolution of these features. FFRs are known to have small-scale short-lived vortices (Hueso et al. 2022), but drawing a clear decision boundary between the inner cores of an FFR and vortices is difficult, even for domain experts. Reconciling the feature overlap between these classes is a fundamentally important avenue of research for understanding how vortices form and evolve, and large-number statistics of such features that reside in between these two classes is important for understanding how (or if) vortices transition to FFRs, or vice versa (Iñurrigarro et al. 2022). The use of the latent space to disentangle these classes is vital in providing the samples necessary to study this phenomenon.

Elsewhere, closer to the main cluster, volunteers struggled with identifying small vortices (only 10–15 pixels across in these images). These small vortices, usually embedded in either FFRs (which are in the left half of the latent space) or in the middle of cloud bands (right half of the latent space), are useful for better understanding localized hydrodynamical instabilities in the jovian atmosphere, which is the scientific goal of the JVH project.

Therefore, the use of the latent space, in combination with the filters provided by class consensus by the citizen scientists, allows us to segregate different sub-classes of atmospheric features and study them in isolation. Particularly, it is significantly easier to navigate the confusion space primarily through the use of morphological feature separability afforded by the latent space and simplify the process of identifying scientifically relevant sources of confusion. For instance, the smaller vortices are much more vital for jovian atmospheric studies compared with the Great Red Spot since they are much harder to detect, and therefore, the use of the latent space simplifies their identification. Summarily, by slicing the latent space using class consensus, we can map feature variation and understand the relationship between the features in the image and how they are common across different classes.

Confusion and scientific value

Given that the fundamental goal of the project is to understand and correlate the dynamics of the atmosphere with the resulting features, let us briefly investigate the ability of the cVAE in improving the scientific return of the JVH dataset. With these tools, we are able to significantly reduce the overhead of disentangling the source of confusion for objects in the dataset. For instance, by just looking at the confusing vortex subset (vortex consensus between 0.3 and 0.7), we filter out about 58,672 images from the dataset (~85%). This still leaves 9,650 images to manually characterize, but we can subset this based on different subclasses. Out of these 9,650, 3,962 correspond to images that feature high consensus on FFRs and only 612 correspond to images that feature high consensus on cloud bands. While it is understandable that FFR-like features have a high rate of confusion with vortices, it is less clear why cloud bands share feature similarity.

Figure 9 shows the latent space distribution for the 612 images that have a high consensus on cloud bands (volunteer agreement >0.7) for confusing vortices (volunteer agreement between 0.3 and 0.7). Note how these images have a wide diversity of background features where the lower right region represents very low gradients in color, and the upper left features very sharp color gradients. These variations correspond to wind shear (or background vorticity) in the atmosphere at these regions, since steep wind shear results in a strong temperature gradient, which results in sharp gradients in cloud type and cloud chemistry; whereas low shear generally results in the opposite, with smooth color variations. Therefore, we have a gradient of potential vortices forming in low-shear environments (lower right) and high-shear environments (upper left). This is particularly important, since vortices forming across this spectrum of wind shear produce markedly different features, and present insights into the patterns of fluid dynamical instabilities (which are poorly known on Jupiter). A deeper discussion on these features is in the Supplemental File 1: Model Description and Additional Results, which details the scientific significance of the cVAE in finding arbitrary relationships between the jovian atmospheric dynamics and the learned features.

GENERALIZATION TO CITIZEN SCIENCE PROJECTS

The results presented here are specific to a case study on the Jovian Vortex Hunter project, but our analysis methodology can be easily generalized to other image-classification projects. We present the lessons learned from our analysis:

Firstly, we find the combination of the traditional VAE architecture with the classification head provides additional information using gradient backpropagation about the

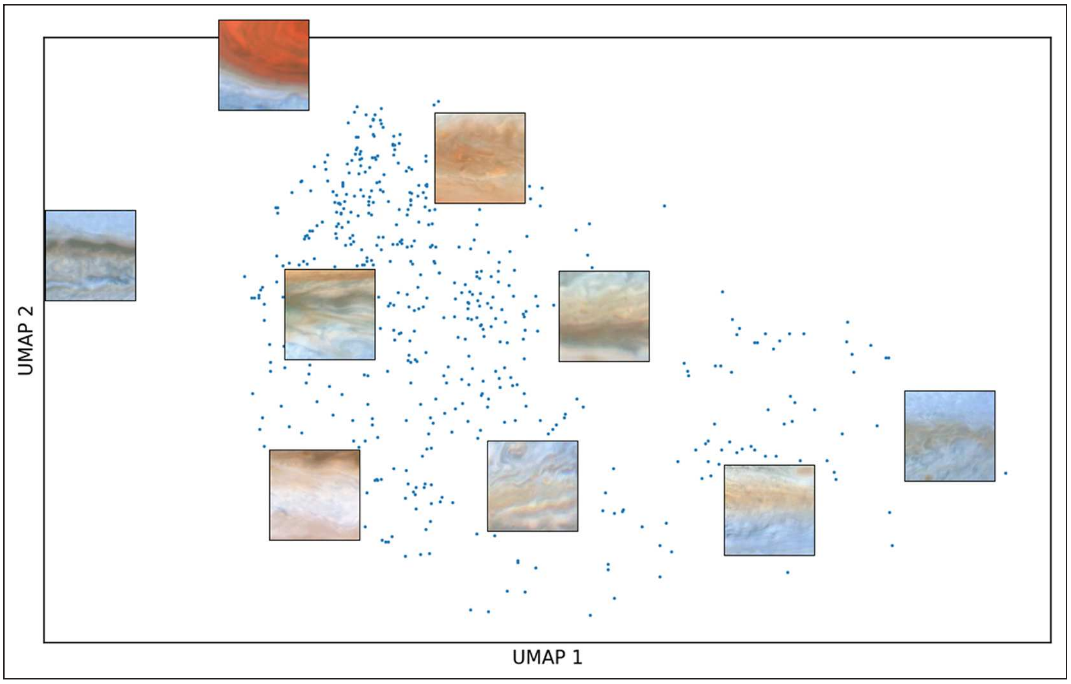

Figure 9 Distribution of latent space for samples with vortex confusion but high consensus on cloud bands, with several images overlaid at their latent space location.

types of features associated with each class in the image. In this way, images with poor volunteer consensus can provide further context about the features in the image that cause confusion. This is significant because obtaining such interpretative information from volunteers is challenging, and even if such information is requested, it would result in a large increase in the volume of volunteer data to process (e.g., through forum boards) (see Cardamone et al. 2009; Oesterlund et al. 2014, etc.).

Secondly, we find that the use of the latent space helps separate the different image-level features that correspond to different sub-classes. Traditional unsupervised methods have produced good clustering performance (e.g., Syarif et al. 2012), but combining the latent space with volunteer consensus offers efficient ways to subsample data and identify interesting ways to relate classes and underlying scientific value. For example, we efficiently determined vortex-like signatures that exist in high-shear versus low-shear environments on Jupiter. In the latent space, these were separable because the wind shear results in different color gradients in the image. If there is a strong correlation between the classes and the image-level features, then it is easy to disentangle different sub-classes of features that result in volunteer confusion, and study each sub-class individually.

Finally, we have shown that our cVAE does not strongly overfit the training data and has instead learned generalizable image-level features across the dataset. As such, it is possible to use this framework to predict consensus on future datasets, and quickly process images that show clear class distinction. In this way, we can choose to show volunteers only those data that show large feature confusion, which would greatly reduce their effort (Walmsley et al. 2019).

CONCLUSIONS AND FUTURE WORK

In conclusion, we have shown that the use of semi-supervised machine learning techniques can add great value to the citizen science–labeled dataset. Using the additional information provided by the distribution of the learned latent variables and the use of layer attention, we can autonomously sub-classify features within the dataset. This is particularly useful when interpreting confusion, where confusion in a classification label can be due to a multitude of factors. Intrinsically, most confusion is due to feature similarity with other classes, and the use of the cVAE helps us disentangle confusion due to the different classes. In particular, in our dataset, we found that we are able to successfully separate confusing vortices between vortex-like structures forming alongside the cloud bands and those within FFRs, and correlate them with fluid dynamical properties. Using these relationships and the latent space distribution, it was easier to disentangle true positive vortex classifications in the confusing sample.

The model presented here features a simple CNN-based cVAE. Current improvements in deep neural networks offer much more sophisticated methods to learn attention, such as using the Transformer architectures (Vaswani et al. 2017; Dosovitskiy et al. 2020), which provide better learned representations of the latent space and much better explainability. Transformers also offer methods to learn features more efficiently across spatial scales (Dosovitskiy et al. 2020), which will improve our model performance on the smaller vortices. We will investigate the improvements offered by these models in a future study.

Additionally, while the model has learned characteristics of the volunteer consensus distribution, it has not provided a way to autonomously binarize the distribution (i.e., learn better representations of the data in order to remove the volunteer confusion). While the confusing sample identified from the volunteer agreement scores has proven valuable, there are still implicit variables not related to the features in the image that cause confusion, such as volunteer skill and prior knowledge. Leveraging the network to disentangle these implicit volunteer variability parameters with true data variability is a much more difficult problem for the network but possibly an avenue of future study. In this fashion, the network can become much more autonomous in flagging and identifying confusing subjects within the dataset, which will alleviate significant burden from the research teams.

Finally, due to the simplicity of the model and the fact that no other information is needed apart from the volunteer labels and input image, this method essentially offers "free information" for citizen science projects where confusion is a significant burden. The use of the neural network to simplify the intrinsic data variability and relate the image-level features to the volunteer agreement provides a great benefit at very little overhead.

DATA ACCESSIBILITY STATEMENT

The classification data is available on Zenodo at https://doi.org/10.5281/zenodo.11659728 and imaging data used for this work will be shared on reasonable request to the author. The machine model and training script is available at https://github.com/ramanakumars/cvae.

SUPPLEMENTAL FILE

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

- **Supplemental File 1.** Model Description and Additional Results. DOI: https://doi.org/10.5334/cstp.731.s1

ACKNOWLEDGEMENTS

The authors would like to thank the Zooniverse volunteers on the Jovian Vortex Hunter project for their valuable time in labeling the JunoCam data. We would also like to thank the reviewers and editors who provided constructive comments that improved the clarity of the manuscript.

FUNDING INFORMATION

RS, KBM, and LF would like to acknowledge partial support for this work from the following grants: NSF IIS 2006894 and NASA Award # 80 NSSC20M0057. RS would also like to acknowledge partial support from NASA Award # 80 NSSC22K0804. Some of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80 NM0018D0004).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

RS, KM and CN designed and trained the neural network model used in this study. RS conceptualized the idea for the project and conducted the post-model analysis. LF, SB, CH and GO helped provide interpretation of the results and revise the manuscript draft.

AUTHOR AFFILIATIONS

Ramanakumar Sankar orcid.org/0000-0002-6794-7587
University of California, Berkeley, US
Kameswara Mantha orcid.org/0000-0002-6016-300X
University of Minnesota, Twin Cities, US
Cooper Nesmith
University of Minnesota, Twin Cities, US
Lucy Fortson orcid.org/0000-0002-1067-8558
University of Minnesota, Twin Cities, US
Shawn Brueshaber orcid.org/0000-0002-3669-0539
Michigan Technological University, US
Candice Hansen-Koharcheck
Planetary Science Institute, US
Glenn Orton orcid.org/0000-0001-7871-2823
Jet Propulsion Laboratory/California Institute of Technology, US

REFERENCES

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S.P., Bennert, N., Urry, C.M., Lintott, C., 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

Chattopadhyay, A., Sarkar, A., Howlader, P., and **Balasubramanian, V.N.** (2017) Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. DOI: https://doi.org/10.1109/WACV.2018.00097

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., et al. (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. DOI: https://doi.org/10.48550/ARXIV.2010.11929

Etsebeth, V., Lochner, M., Walmsley, M., and **Grespan, M.** (2024) Astronomaly at scale: searching for anomalies amongst 4 million galaxies. *Monthly Notices of the Royal Astronomical Society*. 529, pp. 732–747. DOI: https://doi.org/10.1093/mnras/stae496

Fortson, L. (2021) From Green Peas to STEVE: Citizen Science Engagement in Space Science. DOI: https://doi.org/10.1016/b978-0-12-817390-9.00009-9

Hansen, C.J., Caplinger, M.A., Ingersoll, A., Ravine, M.A., Jensen, E., Bolton, S., and **Orton, G.** (2014) Junocam: Juno's Outreach Camera. *Space Science Reviews*. 213, pp. 475–506. DOI: https://doi.org/10.1007/s11214-014-0079-x

Hueso, R., Iñurrigarro, P., Sánchez-Lavega, A., Foster, C.R., Rogers, J.H., Orton, G.S., Hansen, C., et al. (2022) Convective storms in closed cyclones in Jupiter's South Temperate Belt: (I) observations. *Icarus*. 380, pp. 114994. DOI: https://doi.org/10.1016/j.icarus.2022.114994

Hunter, J., Alabri, A., and **van Ingen, C.** (2012) Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience*. 25, pp. 454–466. DOI: https://doi.org/10.1002/cpe.2923

Ingersoll, A.P., Dowling, T.E., Gierasch, P.J., Orton, G.S., Read, P.L., Sanchez-Lavega, A., Showan, et al. (2007) Dynamics of Jupiter's Atmosphere. In *Jupiter: The Planet, Satellites, and Magnetosphere, Cambridge: Cambridge University Press.* pp 105–128

Ingersoll, A.P., Gierasch, P.J., Banfield, D., Vasavada, A.R., and **Galileo Imaging Team.** (2000) Moist convection as an energy source for the large-scale motions in Jupiter's atmosphere. *Nature*. 403(6770), pp. 630–632. DOI: https://doi.org/10.1038/35001021

Iñurrigarro, P., Hueso, R., Sánchez-Lavega, A., and **Legarreta, J.** (2022) Convective storms in closed cyclones in Jupiter: (II) numerical modeling. *Icarus*. 386, pp. 115169. DOI: https://doi.org/10.1016/j.icarus.2022.115169

Ishida, E.E.O., Kornilov, M.V., Malanchev, K.L., Pruzhinskaya, M.V., Volnova, A.A., Korolev, V.S., Mondon, F., et al. (2021) Active anomaly detection for time-domain discoveries. *Astronomy & Astrophysics*. 650, pp. A195. DOI: https://doi.org/10.1051/0004-6361/202037709

Kosmala, M., Wiggins, A., Swanson, A. and **Simmons, B.** (2016) Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*. 14, pp. 551–560. DOI: https://doi.org/10.1002/fee.1436

Krivosheev, E., Bykau, S., Casati, F. and **Prabhakar, S.** (2020) Detecting and preventing confused labels in crowdsourced data. *Proc. VLDB Endow*. 13, 12 (August 2020), pp. 2522–2535. DOI: https://doi.org/10.14778/3407790.3407842

Li, J.S., Hamann, A. and **Beaubien, E.** (2020) Outlier detection methods to improve the quality of citizen science data. *International Journal of Biometeorology*. 64, pp. 1825–1833. DOI: https://doi.org/10.1007/s00484-020-01968-z

Liao, T., Taori, R., Raji, I.D., and **Schmidt, L. Are we learning yet a meta review of evaluation failures across machine learning.** (2021) In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).*

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., et al. (2008) Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey★. *Monthly Notices of the Royal Astronomical Society*. 389, pp. 1179–1189. DOI: https://doi.org/10.1111/j.1365-2966.2008.13689.x

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

Mantha, K.B., Sankar, R., Zheng, Y., Fortson, L., Pengo, T., Mashek, D., Sanders, M., et al. (2022) From fat droplets to floating forests: cross-domain transfer learning using a PatchGAN-based segmentation model. DOI: https://doi.org/10.48550/ARXIV.2211.03937

Orton, G.S., Hansen, C., Caplinger, M., Ravine, M., Atreya, S., Ingersoll, A.P., Jensen, E., et al. (2017) The first close-up images of Jupiter's polar regions: Results from the Juno mission JunoCam instrument. *Geophysical Research Letters*. 44, pp. 4599–4606. DOI: https://doi.org/10.1002/2016GL072443

Oesterlund, C., Mugar, G., Jackson, C. B., Hassman, K. D. and **Crowston, K.** (2014) Socializing the Crowd: Learning to talk in citizen science. *Academy of Management Annual Meeting, OCIS Division, Philadelphia, PA.* DOI: https://doi.org/10.5465/ambpp.2014.16799abstract

Richards, J.W., Starr, D.L., Brink, H., Miller, A.A., Bloom, J.S., Butler, N.R., Berian James, J., Long, J.P. and **Rice, J.** (2011) ACTIVE LEARNING TO OVERCOME SAMPLE SELECTION BIAS: APPLICATION TO PHOTOMETRIC VARIABLE STAR CLASSIFICATION. *The Astrophysical Journal*. 744, pp. 192. DOI: https://doi.org/10.1088/0004-637X/744/2/192

Sankar, R., Brueshaber, S., Fortson, L., Hansen-Koharcheck, C., Lintott, C., Mantha, K., Nesmith, C., et al. (2024) Jovian Vortex Hunter: A Citizen Science Project to Study Jupiter's Vortices. *Planetary Science Journal*. 5, pp. 203. DOI: https://doi.org/10.3847/PSJ/ad6e75

Sankar, R., Mantha, K., Fortson, L., Spiers, H., Pengo, T., Mashek, D., Mo, M., et al. (2023) TCuPGAN: A novel framework developed for optimizing human-machine interactions in citizen science. DOI: https://doi.org/10.48550/ARXIV.2311.14177

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and **Batra, D.** (2016) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. DOI: https://doi.org/10.1109/ICCV.2017.74

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

Syarif, I., Prugel-Bennett, A. and **Wills, G.** (2012) Unsupervised Clustering Approach for Network Anomaly Detection. DOI: https://doi.org/10.1007/978-3-642-30507-8_13

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., et al. (2017) Attention Is All You Need. DOI: https://doi.org/10.48550/ARXIV.1706.03762

Walmsley, M., Smith, L., Lintott, C., Gal, Y., Bamford, S., Dickinson, H., Fortson, L., et al. (2019) Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*. 491, pp. 1554–1574. DOI: https://doi.org/10.1093/mnras/stz2816

Walmsley, M. and **Scaife, A.M.M.** (2023) Rare Galaxy Classes Identified In Foundation Model Representations. DOI: https://doi.org/10.48550/ARXIV.2312.02910

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., et al. (2019) Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. DOI: https://doi.org/10.1109/CVPRW50498.2020.00020

Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., et al. (2018) Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*. 10, pp. 80–91. DOI: https://doi.org/10.1111/2041-210X.13099

Zevin, M., Jackson, C.B., Doctor, Z., Wu, Y., Østerlund, C., Johnson, L.C., Berry, C.P.L., et al. (2024) Gravity Spy: lessons learned and a path forward. DOI: https://doi.org/10.1140/epjp/s13360-023-04795-4

Zou, C., Lai, J., Liu, Y., Cui, F., Xu, Y. and **Qiao, L.** (2024) Small lunar crater identification and age estimation in Chang'e-5 landing area based on improved Faster R-CNN. *Icarus*. 410, pp. 115909. DOI: https://doi.org/10.1016/j.icarus.2023.115909

TO CITE THIS ARTICLE:
Sankar, R, Mantha, K, Nesmith, C, Fortson, L, Brueshaber, S, Hansen-Koharcheck, C and Orton, G. 2024. Understanding Confusion: A Case Study of Training a Machine Model to Predict and Interpret Consensus From Volunteer Labels. *Citizen Science: Theory and Practice,* 9(1): 41, pp. 1–14. DOI: https://doi.org/10.5334/cstp.731

Submitted: 15 February 2024 **Accepted:** 26 September 2024 **Published:** 09 December 2024

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