Vendi Information Gain for Active Learning and its Application to Ecology

Quan Nguyen[1, 2] and Adji Bousso Dieng[1, 2]

[1]Department of Computer Science, Princeton University
[2]Vertaix

September 12, 2025

Abstract

While monitoring biodiversity through camera traps has become an important endeavor for ecological research, identifying species in the captured image data remains a major bottleneck due to limited labeling resources. Active learning—a machine learning paradigm that selects the most informative data to label and train a predictive model—offers a promising solution, but typically focuses on uncertainty in the individual predictions without considering uncertainty across the entire dataset. We introduce a new active learning policy, Vendi information gain (VIG), that selects images based on their impact on dataset-wide prediction uncertainty, capturing both informativeness and diversity. Applied to the Snapshot Serengeti dataset, VIG achieves impressive predictive accuracy close to full supervision using less than 10% of the labels. It consistently outperforms standard baselines across metrics and batch sizes, collecting more diverse data in the feature space. VIG has broad applicability beyond ecology, and our results highlight its value for biodiversity monitoring in data-limited environments.

Keywords: Active Learning, Information Gain, Diversity, Experimental Design, Ecosystem Monitoring, Information Theory, Ecology, Vendi Scoring.

1 Introduction

The ability to monitor biodiversity at scale is critical for understanding ecosystem health and informing conservation efforts. Camera traps—remotely activated cameras triggered by motion or heat—have become a key tool for ecological data collection, enabling large-scale, non-invasive monitoring of wildlife in their natural habitats (Trolliet et al., 2014; Delisle et al., 2021; Tuia et al., 2022). These devices generate vast volumes of image data, often spanning multiple times of day and geographies. However, the subsequent task of identifying and labeling the species in these images remains a significant bottleneck. Manual annotation is labor-intensive, costly, time-consuming, and may require expert knowledge, especially when dealing with rare species or poor image quality.

Recent advances in machine learning, specifically deep learning for image classification, offer a promising direction for automating species identification (Norouzzadeh

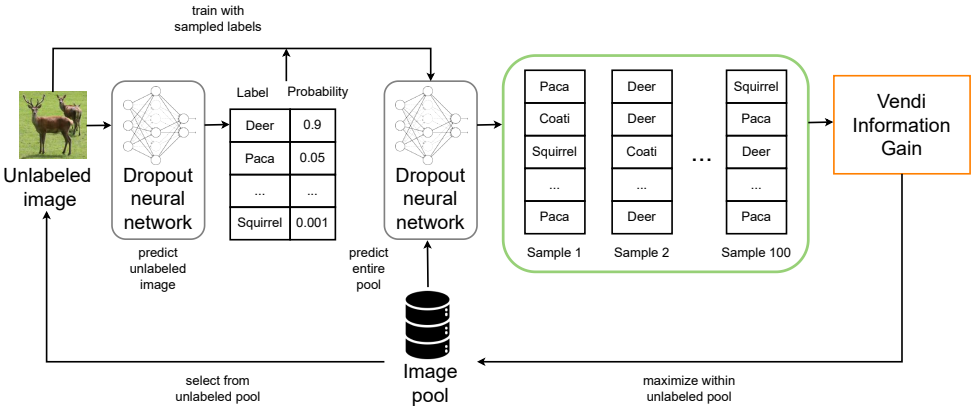

Figure 1: Overview of Vendi information gain (VIG) for active learning. We use a trained dropout neural network to sample labels for a candidate datapoint. The neural network is then retrained on this fantasized data to sample labels of the entire pool. Uncertainty in these predictions is captured by VIG, and we select the candidate that yields the highest information gain (i.e., lowest uncertainty) in the predictions to label. The result is then added to the training dataset, and the process repeats until the labeling budget is exhausted.

et al., 2018; Beery et al., 2018). Yet the performance of these models crucially depends on the availability of large amounts of high-quality labeled training data. In many ecological applications, however, labels are scarce and labeling is costly. These challenges motivate the need for intelligent sampling strategies that maximize model performance while minimizing labeling effort.

Active learning (Settles, 2009; Bothmann et al., 2023) offers a principled solution to this problem. By iteratively selecting the most informative examples to label, active learning algorithms can achieve high accuracy with fewer labeled instances than naïve approaches. Existing active learning solutions reason about the level of informativeness of candidates for labeling on an individual basis—targeting datapoints that the model is most uncertain about—without accounting for the effects of those datapoints post-labeling. This perspective neglects the overall structure of the entire image pool.

In this work, we propose Vendi information gain (VIG), a novel active learning policy designed to optimize the global informativeness of the training data. VIG builds on recent advances in information-theoretic metrics and quantifies the reduction in predictive uncertainty across the entire image pool when a candidate image is labeled. This approach selects datapoints not only because they yield high uncertainty, but because they are likely to inform the model's predictions across the board. Figure 1 shows the schematics of VIG consisting of the following steps. First, we sample candidate labels for each unlabeled image using a dropout neural network predictor. We then retrain the model on these fantasized labels and sample predictions for the entire unlabeled pool. These sampled predictions quantify the reduction in Vendi entropy (Friedman and Dieng, 2023; Nguyen and Dieng, 2025) across the unknown labels, which guides the search for the candidate with the highest information gain. This process repeats iteratively, expanding the labeled set until the labeling budget

is exhausted. The use of a dropout neural network for active learning is described in Section 2.2, and Section 2.4 includes the computational details of VIG.

Applied to the Snapshot Serengeti dataset (Swanson et al., 2015)—a benchmark for camera trap classification—VIG consistently outperforms standard active learning baselines in terms of label efficiency and predictive accuracy. We show that VIG collects more diverse data in feature space, leading to better generalization with fewer labels. Our results suggest that VIG can serve as a general-purpose method for data-efficient ecological monitoring.

2 Method

We first discuss the active learning framework and the use of a dropout neural network as the predictor for this task. We then provide background on VIG as a metric of information gain and present its adoption to active learning.

2.1 Active Learning Policies

Active learning targets the common setting in machine learning where labeling data is costly (in terms of time, money, or some safety-critical conditions). The goal is to design an active learning policy that selects a small amount of data to label, so that the predictive model trained on the labeled data achieves good generalization performance. In our setting, we have access to a large database of unlabeled images $\mathscr{X} = \{x_i\}_{i=1}^n$, where each x_i denotes a particular datapoint (image) within the database. These images are classified into a predetermined number of classes $[C] = \{1, 2, \ldots, C\}$, and the unknown label y_i of datapoint x_i denotes the membership of that point. Active learning proceeds in an iterative manner where at each step, the active learning policy selects a batch of images to label, adding them to the training data. The process repeats such that we accumulate a training set of increasing size until our labeling budget is depleted.

The main focus of active learning is the design of the policy that selects which data to label. Increasing information (or decreasing uncertainty) in the knowledge of the trained model serves as a popular heuristic for this task. Formally, assume that we have a probabilistic model that produces the posterior probability that a point $x \in \mathscr{X}$ belongs to class $c \in [C]$, denoted as $p(y = c \mid x)$. (We omit the dependence on the labeled data $\mathscr{D}$ that the model is trained on for conciseness.) Many active learning policies seek to minimize model uncertainty, quantified by various statistics from the predictive distribution $p(y \mid x)$. For instance, the *Max entropy* policy finds the data that have the highest predictive entropy H (Shannon, 1948) to quantify uncertainty in the predictions:

$$H(y \mid x) = -\sum_{c \in [C]} p(y = c \mid x) \log p(y = c \mid x). \tag{1}$$

Other policies target alternative ways to quantify predictive uncertainty. This includes the *Mean STD* policy targeting the average standard deviation in the predictions:

$$\sigma(x) = \frac{1}{C} \sum_{c \in [C]} \sqrt{\mathbb{E}[p(y = c \mid x)^2] - \mathbb{E}[p(y = c \mid x)]^2}, \tag{2}$$

which corresponds to the standard deviation statistic in the regression setting, but has been recently adopted in classification as well (Kampffmeyer et al., 2016; Kendall et al., 2017). Another popular active learning policy, BALD, maximizes the amount of information gained about the predictive model's parameters $\boldsymbol{\omega}$, which is equivalent to maximizing the mutual information I between predictions and model posterior (Houlsby et al., 2011):

$$I(\boldsymbol{\omega}, y \mid x) = H(\boldsymbol{\omega}) - \mathbb{E}_{p(y|x)}[H(\boldsymbol{\omega} \mid x, y)]. \tag{3}$$

Finally, Kirsch et al. (2019) proposed BatchBALD that extends BALD to account for interactions between datapoints within a batch. We use these active learning policies as baselines to compare VIG against.

2.2 Dropout Neural Networks

The previously described active learning policies depend on a probabilistic model producing predictions of the form $p(y = c \mid x)$, and as such have been limited to kernel-based methods such as Gaussian processes (Li and Guo, 2013). In the context of image classification, these methods require a kernel to operate on images, which do not scale well to high-dimensional data or capture spatial information within the input images. On the other hand, convolutional neural networks (Rumelhart et al., 1985; LeCun et al., 1989) have proven to be effective at learning from images and achieved human-level performance at image recognition. However, neural networks do not inherently produce probabilistic predictions with calibrated uncertainty quantification.

Initially developed to regularize neural networks, dropout (Hinton et al., 2012; Srivastava et al., 2014) dictates that random nodes in the hidden layers of a neural network are disabled at each forward pass during training. Gal and Ghahramani (2016) further showed that using dropout during inference produces Monte Carlo samples from the predictive distribution of the corresponding Bayesian neural network trained with variational inference, naming the technique *MC dropout*. Finally, Gal et al. (2017) used MC dropout as the predictive model to perform active learning on high-dimensional image data, showing that combined with MC dropout, the policies previously described outperform kernel-based active learning methods as well as their counterparts that use the predictions of a non-dropout neural network. We use this neural network model with MC dropout as the probabilistic classifier in our experiments.

2.3 Vendi Information Gain

VIG was based on the Vendi Score (VS), a flexible diversity metric. First proposed by Friedman and Dieng (2023) and later extended by Pasarkar and Dieng (2024), the VS operates on a set of datapoints $D = \{\theta_i\}_{i=1}^n$ sampled from some domain Θ. To realize the VS, we first require a positive semidefinite kernel function $k : \Theta \times \Theta \to \mathbb{R}$, where $k(\theta, \theta) = 1, \forall \theta \in \Theta$. We then compute the kernel matrix $K \in \mathbb{R}^{n \times n}$, where each entry $K_{i,j} = k(\theta_i, \theta_j)$. Finally, we define the VS as:

$$\mathrm{VS}_q(D; k) = \exp\left(\frac{1}{1-q} \log\left(\sum_{i=1}^{n} (\overline{\lambda}_i)^q\right)\right). \tag{4}$$

where $\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_n$ are the eigenvalues of K, normalized so that they sum to 1, and the order $q \geq 0$ is a hyperparameter. The VS has since been extended and applied to various domains, including evaluating generative models (Hall et al., 2024; Senthilkumar et al., 2024; Jalali et al., 2024), molecular simulations (Pasarkar et al., 2023), Bayesian optimization (Liu et al., 2024) and active search (Nguyen and Dieng, 2024), sequence generative models (Rezaei and Dieng, 2025a), RAG approaches for LLMs (Rezaei and Dieng, 2025b), analysis of large-scale data Pasarkar and Dieng (2025), and reinforcement learning (Lintunen, 2025).

In particular, Nguyen and Dieng (2025) introduced VIG as a metric of information gain, defining it as the difference in the Vendi entropy H_V of a random variable θ before and after conditioning on another variable y:

$$\mathrm{VIG}(\theta, y; q) = H_V(D; q) - \mathbb{E}_y[H_V(D_y; q)], \tag{5}$$

where $D = \{\theta_i\}_{i=1}^n$ is a set of samples of θ, and $D_y = \{\theta_i \mid y\}_{i=1}^n$ is the corresponding set of samples conditioned on a particular value of y. Here, the Vendi entropy is the logarithm or the VS, or the Rényi entropy of the normalized eigenvalues of the kernel matrix computed from a set of samples:

$$H_V(D; q) = \frac{1}{1-q} \log\left(\sum_{i=1}^{n} (\overline{\lambda}_i)^q\right). \tag{6}$$

Nguyen and Dieng (2025) demonstrated many of VIG's advantages over mutual information, the default measure of information gain in the scientific literature (Shannon, 1948; Cover, 1999). Namely, VIG works well with only samples of the random variable of interest and offers a more principled quantification of information gain that accounts for sample similarity. The authors showcased VIG's superior performance in a wide range of tasks, including experimental design problems and level-set estimation.

2.4 Vendi Information Gain for Active Learning

We adopt the VIG criterion for active learning, proposing a policy that minimizes the Vendi entropy of the posterior predictions across the entire database of images, conditioned on a candidate datapoint. Formally, denote $\boldsymbol{\theta}$ as the vector that concatenates the unknown labels of the images within the database, the VIG policy finds the datapoint x that minimizes the posterior Vendi entropy in $\boldsymbol{\theta}$:

$$\mathrm{VIG}(\boldsymbol{\theta}, x) = H_V(D) - \mathbb{E}_{y|x}\left[H_V(D_y \mid x)\right], \tag{7}$$

where D is a set of samples of the label vector $\boldsymbol{\theta}$, and D_y is the corresponding set of samples conditioned on a particular label y of image x. These samples can be generated using the MC dropout neural network when predicting on the images in the database.

The computation of Vendi entropy requires a kernel that compares two given sample label vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. We compute the Hamming distance d_H between these vectors and subtract the normalized distance from 1 to produce a similarity measure:

$$k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 1 - \frac{d_H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{N}, \tag{8}$$

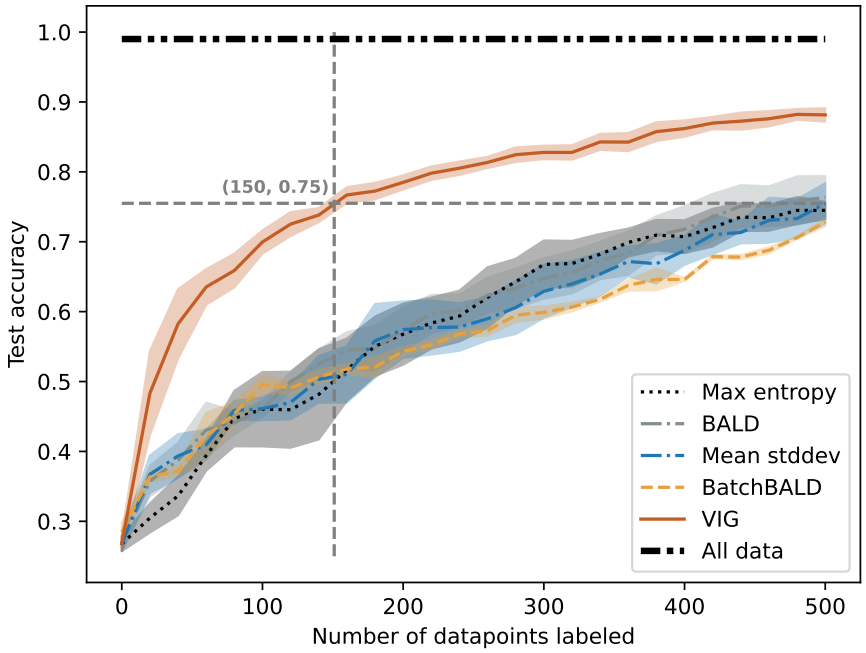

Figure 2: Average test accuracy (±1 standard error) by various active learning policies. VIG obtains a large gain right from the start and maintains its lead throughout the active learning loop. It takes VIG only 150 datapoints to achieve the accuracy of 75% that other methods need 500 points to achieve. Meanwhile, at 500 points, VIG achieves close to 90% accuracy. In comparison, training on all available training data (5000+ images) yields an accuracy of 99%.

where N is the length of the label vectors. Note that this is not the same kernels in the kernel-based active learning policies, which seek to operate on the images themselves.

This choice of kernel is natural, as two labels are similar to each other only if they belong to the same class. When there is only one datapoint in the pool, the Vendi entropy induced by this kernel coincides with the Shannon entropy of the datapoint's class distribution—a reassuring feature.

Overall, while traditional active learning policies target individual predictive uncertainty measures, VIG selects datapoints expected to reduce uncertainty in predictions over the entire unlabeled pool, accounting for global informativeness. To compute the VIG score for a candidate image x, we sample possible labels y, retrain the model with the labeled x, then sample predictions $\boldsymbol{\theta}$ for the full pool. The candidate with the highest VIG score is selected.

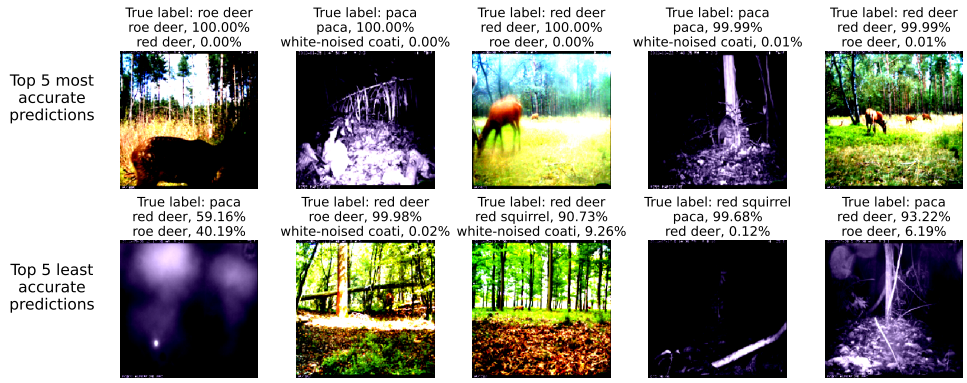

Figure 3: The 5 most (top row) and 5 least (bottom row) accurate predictions by the model trained with data collected by VIG. In the bottom row, the model understandably makes mistakes on instances where the animal is barely visible.

Table 1: Average test statistics by various active learning policies at 500 labeled datapoints. (Recall is omitted as it coincides with accuracy by definition.) VIG consistently outperforms the baselines across the different metrics.

	Max entropy	BALD	Mean stddev	BatchBALD	VIG
Precision ↑	0.780	0.799	0.780	0.764	**0.888**
F1 score ↑	0.755	0.775	0.765	0.738	**0.883**
Cross-entropy loss ↓	0.705	0.635	0.625	0.707	**0.402**

3 Experiments

We benchmark our method VIG against existing baselines in active learning described in Section 2. At each iteration of the active learning loop, each policy obtains a batch of 20 images to label, and the process repeats until 500 images are collected.

Figure 2 shows the accuracy on a hold-out test set of the model trained on data collected by various active learning policies, averaged across repeated, as a function of the number of datapoints labeled. VIG significantly outperforms the baselines, achieving a higher accuracy with fewer labels. After obtaining 500 labeled datapoints, VIG yields a test accuracy close to 90%, while other policies reach 75%. To achieve the same performance, VIG needs only 150 labels. In comparison, assuming unlimited labeling resources, the model trained on all available training data (5585 images) yields a test accuracy of 99%.

To inspect the learned behavior of VIG's model, Figure 3 visualizes the top five most and least accurate predictions on the test set by VIG. On the top row that the trained model reassuringly makes accurate predictions with high confidence on instances where the animal is visibly in the middle of the camera trap images. On the bottom row, the model understandably makes mistakes on instances with low visibility, including those taken in the dark—situations even humans find challenging.

Table 1 lists other metrics of classification performance than accuracy in Figure 2. This includes the cross-entropy loss, which accounts for the model's predictive

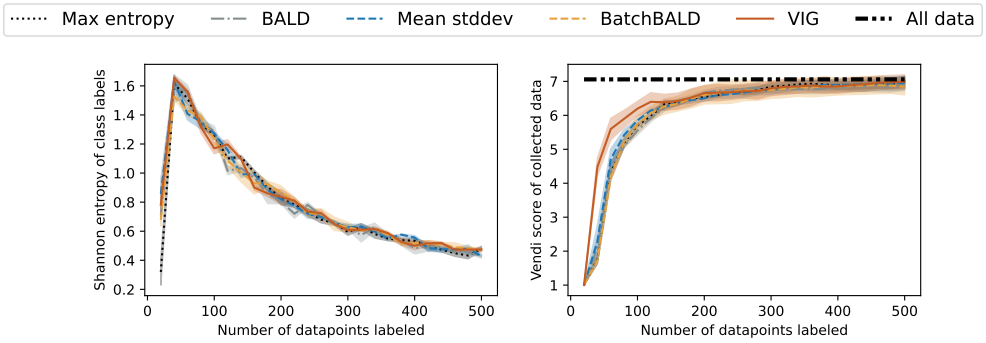

Figure 4: Diversity of the collected data by various active learning policies. **Left**: The Shannon entropy of the class distribution of the collected data. Here, all methods are comparable. **Right**: The Vendi score of the collected data using the embedding in the second-to-last layer of the neural network classifier trained on all available data. VIG selects more diverse data right from the beginning.

confidence, rewarding confident correct predictions and punishing confident incorrect ones. Overall, VIG consistently achieves the best performance across the metrics.

To understand what drives VIG's performance, we inspect the diversity of the data collected by each method in Figure 4. The left panel shows diversity in the labels of the collected data, measured by the Shannon entropy of the class distribution. Here, all policies are comparable. In the right panel we show diversity in the features of the images, quantified by the Vendi score (VS) (Friedman and Dieng, 2023) of the labeled images. The VS is a flexible diversity metric whose output has the natural interpretation of the effective number of unique elements in a set. The VS requires a kernel function to compute the similarity of two given datapoints. Following previous works (Friedman and Dieng, 2023; Pasarkar and Dieng, 2024; Askari Hemmat et al., 2024), we choose the cosine kernel operating on the image embedding. To have a consistent embedding across different active learning policies, we train a neural network classifier on all available data and use the features in the second-to-last layer. Right from the start of the active learning loop, VIG collects more diverse data (feature-wise), a behavior previous works have demonstrated to be beneficial for active learning (Yang et al., 2015; Du et al., 2015; Buchert et al., 2022).

Figure 2 shows the performance of the active learning policies when the batch size (the number of images selected to be labeled at each step of the learning loop) is set to 20. We repeat these experiments while varying this batch size to investigate the effect of this parameter. The left panel of Figure 5 shows the same results under batch size 10, representing a low-throughput setting, while the right panel gives batch size 50 (a high-throughput setting). We see the reasonable trend that policies tend to perform better when the batch size is small, as they get more frequent feedback from the labels and thus can be more adaptive in their selections. Further, VIG stays competitive across the different batch sizes, illustrating the benefits of our method.

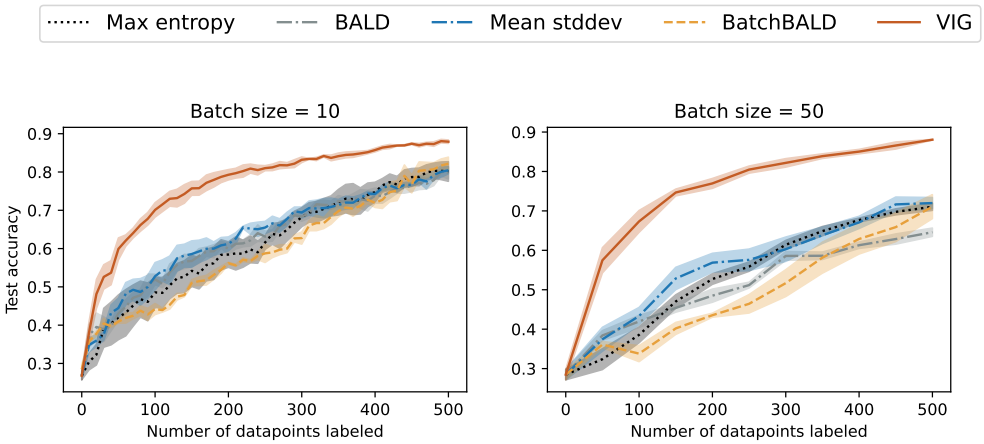

Figure 5: Test accuracy by various active learning policies under different batch sizes. VIG's superior performance stays consistent in both low- and high-throughput settings, underscoring its robustness to selection frequency.

These results collectively show that VIG's reasoning allows it to extract more information from fewer labels, making it suitable for ecological settings with limited annotation budgets.

4 Discussion

We study a new active learning policy, Vendi information gain (VIG), and demonstrate its effectiveness in image-based biodiversity monitoring. By selecting images that maximize information gain over the entire unlabeled pool, VIG prioritizes examples that not only have high uncertainty but are also informative and diverse. With camera trap data from the Snapshot Serengeti dataset, VIG achieves substantial gains in label efficiency and predictive performance compared to established baselines.

Though we focus on species classification from camera trap images, VIG is general-purpose and model-agnostic. The method only requires a probabilistic predictor capable of generating samples, such as a dropout neural network like ours or a Gaussian process. This makes VIG generalizable to a broad range of machine learning tasks beyond ecological applications.

VIG requires retraining the predictive model for each evaluation of a candidate, which leads to high computational complexity. To make the method more efficient, we employ early stopping for this retraining step, terminating the training process early if the training loss converges. Our justification is that during VIG's computation, as we add one single sampled label to the training set, the model trained at the previous step is already close to an optimum. Further, due to the difficulty in obtaining labels in active learning settings, the size of the training set is often limited, which allows for faster training. In our experiments, VIG takes about 4 seconds per evaluation—an acceptable speed given the boost in performance from the method.

VIG's superior performance highlights the value of using structured diversity to quantify uncertainty—an approach that aligns well with the complexity and richness of ecological data. Future work may explore its application in regression tasks such as estimating the abundance of species, or integration with crowd-sourced labeling platforms to elicit expert labeling effort when it is most needed.

Dedication

This paper is dedicated to Wangarĩ Muta Maathai.

References

Askari Hemmat, R., Hall, M., Sun, A., Ross, C., Drozdzal, M., and Romero-Soriano, A. (2024). Improving Geo-diversity of Generated Images with Contextualized Vendi Score Guidance. In *European Conference on Computer Vision*, pages 213–229. Springer.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on computer vision (ECCV)*, pages 456–473.

Bothmann, L., Wimmer, L., Charrakh, O., Weber, T., Edelhoff, H., Peters, W., Nguyen, H., Benjamin, C., and Menzel, A. (2023). Automated wildlife image classification: An active learning tool for ecological applications. *Ecological Informatics*, 77:102231.

Buchert, F., Navab, N., and Kim, S. T. (2022). Exploiting Diversity of Unlabeled Data for Label-Efficient Semi-Supervised Active Learning. In *International Conference on Pattern Recognition*, pages 2063–2069. IEEE.

Cover, T. M. (1999). *Elements of Information Theory*. John Wiley & Sons.

Delisle, Z. J., Flaherty, E. A., Nobbe, M. R., Wzientek, C. M., and Swihart, R. K. (2021). Next-generation camera trapping: systematic review of historic trends suggests keys to expanded research applications in ecology and conservation. *Frontiers in Ecology and Evolution*, 9:617996.

Du, B., Wang, Z., Zhang, L., Zhang, L., Liu, W., Shen, J., and Tao, D. (2015). Exploring Representativeness and Informativeness for Active Learning. *IEEE Transactions on Cybernetics*, 47(1):14–26.

Friedman, D. and Dieng, A. B. (2023). The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Gal, Y., Islam, R., and Ghahramani, Z. (2017). Deep Bayesian Active Learning with Image Data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Hall, M., Bell, S. J., Ross, C., Williams, A., Drozdzal, M., and Soriano, A. R. (2024). Towards Geographic Inclusion in the Evaluation of Text-to-Image Models. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 585–601.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint*. arXiv:1207.0580.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *arXiv preprint*. arXiv:1112.5745.

Jalali, M., Ospanov, A., Gohari, A., and Farnia, F. (2024). Conditional Vendi Score: An Information-Theoretic Approach to Diversity Evaluation of Prompt-based Generative Models. *arXiv preprint*. arXiv:2411.02817.

Kampffmeyer, M., Salberg, A.-B., and Jenssen, R. (2016). Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9.

Kendall, A., Badrinarayanan, V., and Cipolla, R. (2017). Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In *The British Machine Vision Conference*. British Machine Vision Association.

Kirsch, A., Van Amersfoort, J., and Gal, Y. (2019). BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Advances in Neural Information Processing Systems*, 32.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.

Li, X. and Guo, Y. (2013). Adaptive active learning for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866.

Lintunen, E. M. (2025). VendiRL: A Framework for Self-Supervised Reinforcement Learning of Diversely Diverse Skills. *arXiv preprint*. arXiv preprint arXiv:2509.02930.

Liu, T.-W., Nguyen, Q., Dieng, A. B., and Gómez-Gualdrón, D. A. (2024). Diversity-driven, efficient exploration of a mof design space to optimize mof properties. *Chemical Science*, 15(45):18903–18919.

Nguyen, Q. and Dieng, A. B. (2024). Quality-Weighted Vendi Scores And Their Application To Diverse Experimental Design. In *International Conference on Machine Learning*.

Nguyen, Q. and Dieng, A. B. (2025). Vendi information gain: An alternative to mutual information for science and machine learning. *arXiv preprint*. arXiv:2505.09007.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *The National Academy of Sciences*, 115(25):E5716–E5725.

Pasarkar, A. P., Bencomo, G. M., Olsson, S., and Dieng, A. B. (2023). Vendi Sampling For Molecular Simulations: Diversity As A Force For Faster Convergence And Better Exploration. *The Journal of Chemical Physics*, 159(14).

Pasarkar, A. P. and Dieng, A. B. (2024). Cousins Of The Vendi Score: A Family Of Similarity-Based Diversity Metrics For Science And Machine Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3808–3816. PMLR.

Pasarkar, A. P. and Dieng, A. B. (2025). The Vendiscope: An Algorithmic Microscope For Data Collections. *arXiv preprint*. arXiv:2502.10828.

Rezaei, M. R. and Dieng, A. B. (2025a). The $alpha$-Alternator: Dynamic Adaptation To Varying Noise Levels In Sequences Using The Vendi Score For Improved Robustness and Performance. *arXiv preprint*. arXiv:2502.04593.

Rezaei, M. R. and Dieng, A. B. (2025b). Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. *arXiv preprint*. arXiv:2502.11228.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report.

Senthilkumar, N. K., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A. B., Bhattacharyya, P., and Dave, S. (2024). Beyond Aesthetics: Cultural Competence in Text-to-Image Models. *Advances in Neural Information Processing Systems*, 37:13716–13747.

Settles, B. (2009). Active Learning Literature Survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015). Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2(1):1–14.

Trolliet, F., Vermeulen, C., Huynen, M.-C., and Hambuckers, A. (2014). Use of camera traps for wildlife studies: a review. *Biotechnologie, Agronomie, Société et Environnement*, 18(3).

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., Van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792.

Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. (2015). Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization. *International Journal of Computer Vision*, 113(2):113–127.

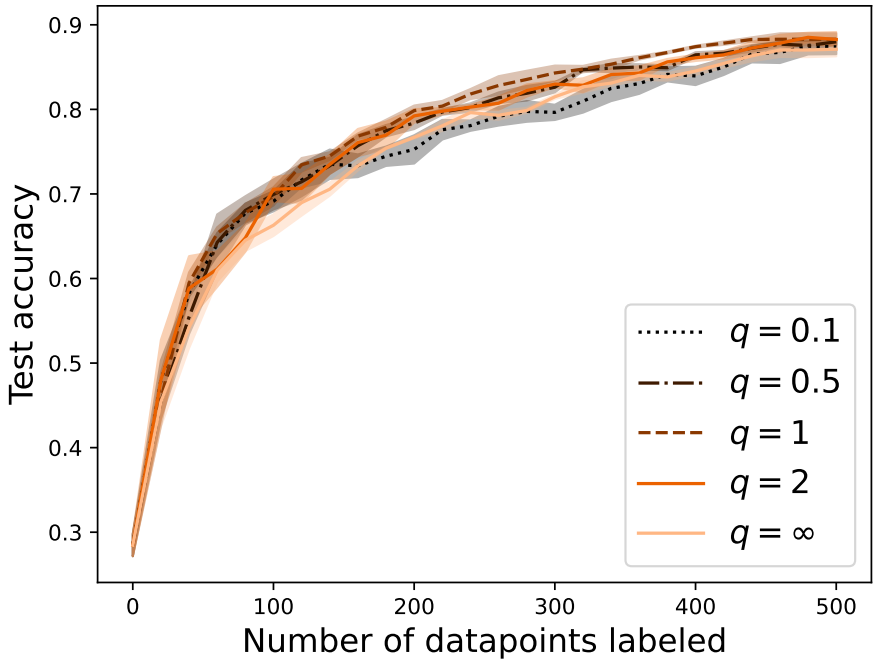

Figure 6: Average test accuracy and one standard error by VIG of different orders q. VIG's performance is robust against the value of q.

A Additional Experiment Results

We now present the result of an ablation study where we investigate the effect of the hyperparameter q in the formulation of VIG. Pasarkar and Dieng (2024) showed that the order q controls the sensitivity of the Vendi score (and thus the Vendi entropy and VIG) to rarity: low values of q lead to more sensitivity to rare features, while high values of q prioritize common features of the samples. By setting this hyperparameter, we can induce a family of VIG policies with different levels of sensitivity to rare samples. Figure 6 shows the results of VIG across a wide range of values for q, where test performance is comparable across the VIG policies. This shows that the performance improvement from existing active learning baselines we obtain is mainly due to VIG's information gain-based reasoning, which is robust against the order q when computing Vendi entropy.

B Data

We use the latest iteration of the Snapshot Serengeti dataset (Swanson et al., 2015) and extract the top five species, making a dataset with the following class breakdown:

- Paca: 1196 images
- Red deer: 2830 images

- Red squirrel: 639 images
- Roe deer: 1271 images
- White-nosed coati: 1295 images

This makes up a 7231-image dataset. In each experiment, we pick a randomly selected 20% of the data as the test set to evaluate the trained models; the other 80% acts as the image pool for active learning.