Breakout Session 3: Track A

Beyond Class Balance: Dataset Diversity and Model Performance in Deep-Learning Classification Tasks

Dr. Josiah Couch

Postdoctoral Research Fellow, Beth Israel Deaconess Medical Center

Beyond Class Balance:

Dataset Diversity and Model Performance in Deep-Learning Classification Tasks

Award Title: ENRICHing NIH Imaging Datasets to Prepare them for Machine Learning

Josiah Couch, Ph.D.
PIs: Rima Arnaout, M.D. and Ramy Arnaout, M.D., D.Phil.

Beth Israel Deaconess Medical Center

27 March 2024

Motivation

Motivation

- We want to understand dataset quality

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?

Motivation

- We want to understand dataset quality
  - high quality dataset $\rightarrow$ high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size
- Just look at these two datasets →

Figure: Dataset 1: wasps vs grasshoppers (more diverse)

Figure: Dataset 2: wasps vs grasshoppers (less diverse)

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size
- Just look at these two datasets →
  - Same class balance

Figure: Dataset 1: wasps vs grasshoppers (more diverse)

Figure: Dataset 2: wasps vs grasshoppers (less diverse)

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size
- Just look at these two datasets →
  - Same class balance
  - Same number of images

Figure: Dataset 1: wasps vs grasshoppers (more diverse)

Figure: Dataset 2: wasps vs grasshoppers (less diverse)

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size
- Just look at these two datasets →
  - Same class balance
  - Same number of images
  - But dataset 1 clearly has higher diversity

Figure: Dataset 1: wasps vs grasshoppers (more diverse)

Figure: Dataset 2: wasps vs grasshoppers (less diverse)

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size
- Just look at these two datasets →
  - Same class balance
  - Same number of images
  - But dataset 1 clearly has higher diversity
  - And thus perhaps a higher quality?

Figure: Dataset 1: wasps vs grasshoppers (more diverse)

Figure: Dataset 2: wasps vs grasshoppers (less diverse)

Motivation

- We want to understand dataset quality
  - high quality dataset → high performance model
- In particular, what indicators diagnose a dataset as high quality?
  - Class balance
  - Dataset size
  - Other things?
- There must be more to quality than class balance and size
- Just look at these two datasets →
  - Same class balance
  - Same number of images
  - But dataset 1 clearly has higher diversity
  - And thus perhaps a higher quality?
- Our starting hypothesis is that diversity contributes to quality independently of class balance (and of dataset size)

Figure: Dataset 1: wasps vs grasshoppers (more diverse)

Figure: Dataset 2: wasps vs grasshoppers (less diverse)

Diversity Framework

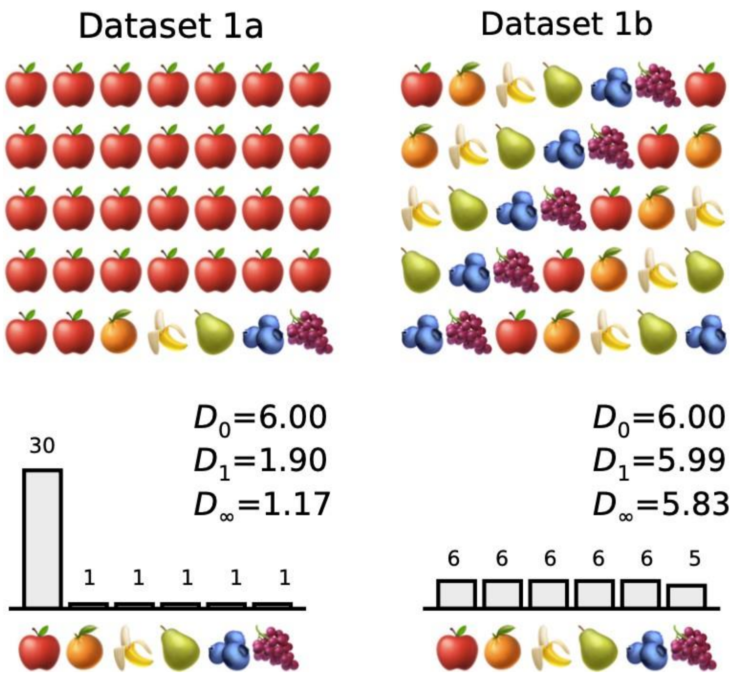

Figure: Diversity depends on frequency, image taken from [1]

- How do we measure **diversity**?

Diversity Framework

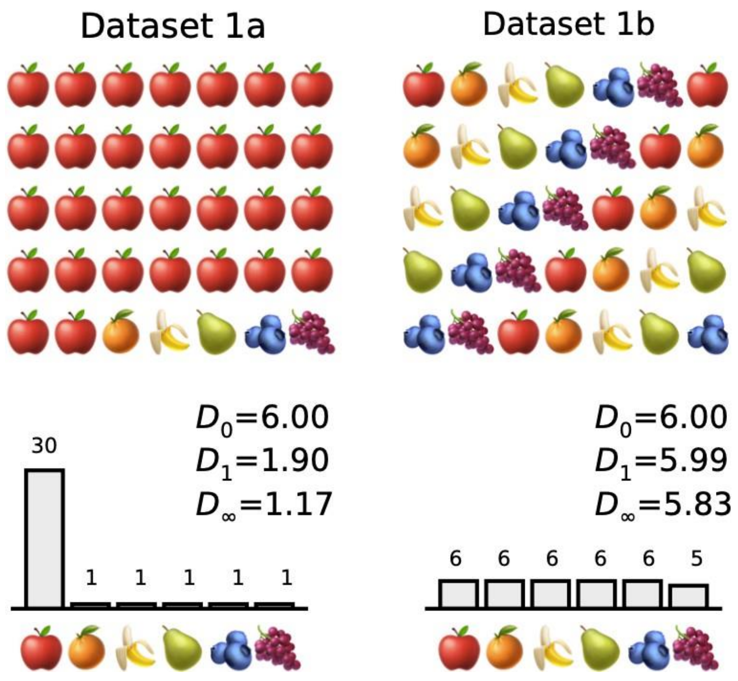

Figure: Diversity depends on frequency, image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]

Diversity Framework

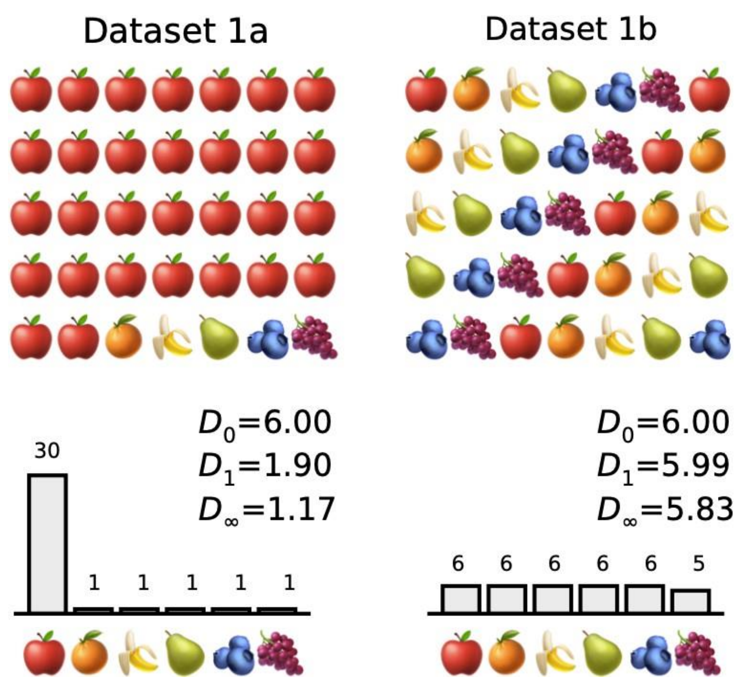

Figure: Diversity depends on frequency, image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]

Diversity Framework

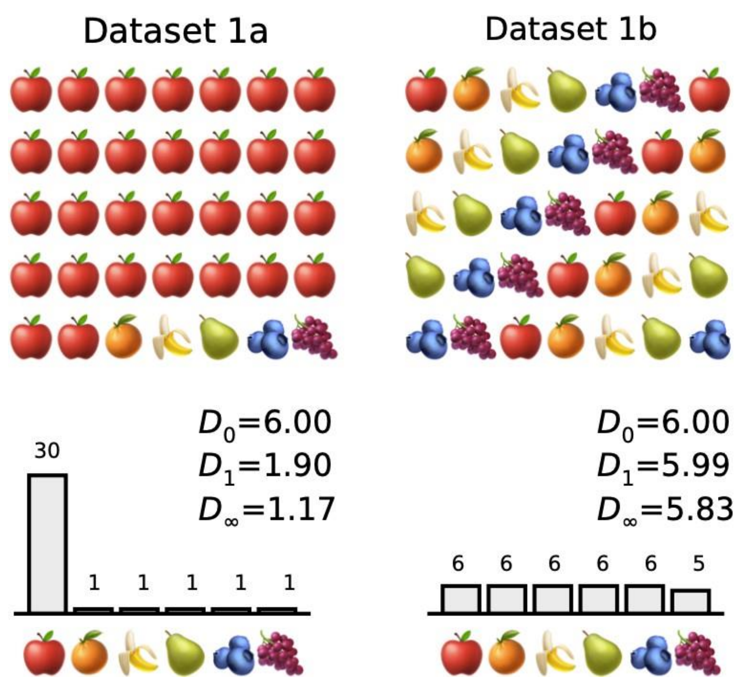

Figure: Diversity depends on frequency, image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]
  - Like Hill numbers, this framework accounts for the frequency with which different types of things occur

Diversity Framework

Figure: Diversity depends on similarity, , image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]
  - Like Hill numbers, this framework accounts for the frequency with which different types of things occur
  - Unlike Hill numbers, it also accounts for the similarities between types. Higher similarities lead to lower diversities

Diversity Framework

Figure: Diversity depends on similarity, , image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]
  - Like Hill numbers, this framework accounts for the frequency with which different types of things occur
  - Unlike Hill numbers, it also accounts for the similarities between types. Higher similarities lead to lower diversities
- In the context of an image classification training set . . .

Diversity Framework

Figure: Diversity depends on similarity, , image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]
  - Like Hill numbers, this framework accounts for the frequency with which different types of things occur
  - Unlike Hill numbers, it also accounts for the similarities between types. Higher similarities lead to lower diversities
- In the context of an image classification training set . . .
  - We will consider each image to be a unique type

Diversity Framework

Figure: Diversity depends on similarity, , image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]
  - Like Hill numbers, this framework accounts for the frequency with which different types of things occur
  - Unlike Hill numbers, it also accounts for the similarities between types. Higher similarities lead to lower diversities
- In the context of an image classification training set . . .
  - We will consider each image to be a unique type
  - The similarity between images will be based on their euclidean distance in pixel space (smaller distance $\leftrightarrow$ higher similarity)

Diversity Framework

Figure: Diversity depends on similarity, , image taken from [1]

- How do we measure **diversity**?
  - We will use the framework of Leinster and Cobbold [2] and Reeve et al. [3]
  - This framework generalizes the Hill numbers [4]
  - Like Hill numbers, this framework accounts for the frequency with which different types of things occur
  - Unlike Hill numbers, it also accounts for the similarities between types. Higher similarities lead to lower diversities
- In the context of an image classification training set . . .
  - We will consider each image to be a unique type
  - The similarity between images will be based on their euclidean distance in pixel space (smaller distance $\leftrightarrow$ higher similarity)
- We can also treat class balance in this framework by using a similarity matrix based on sharing the same class label

Methodology

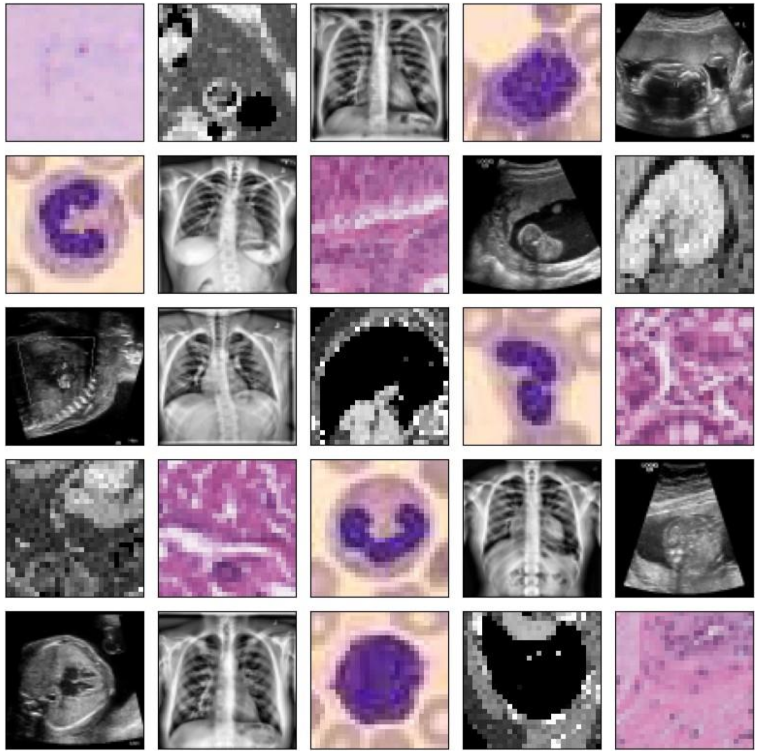

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets

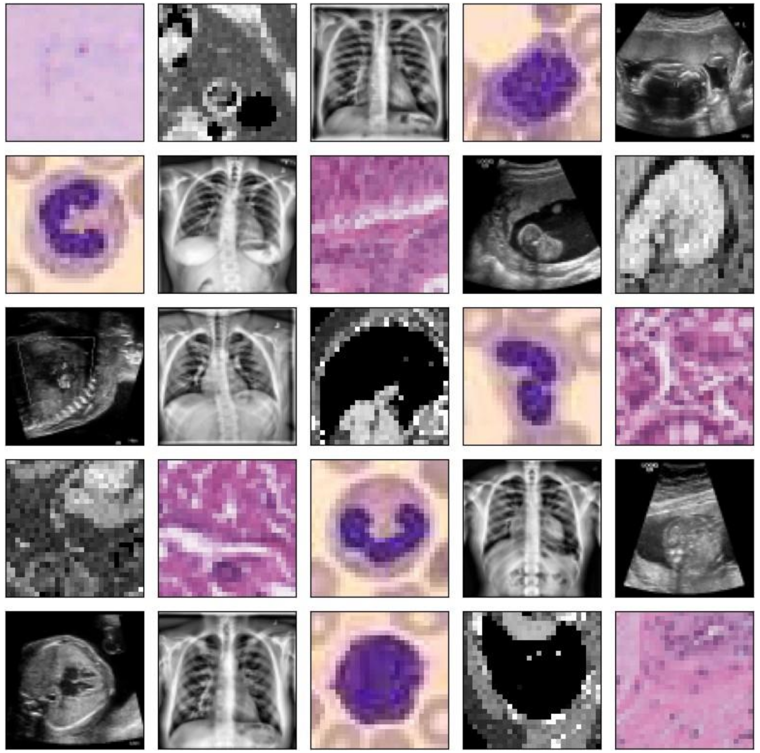

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [5, 6]

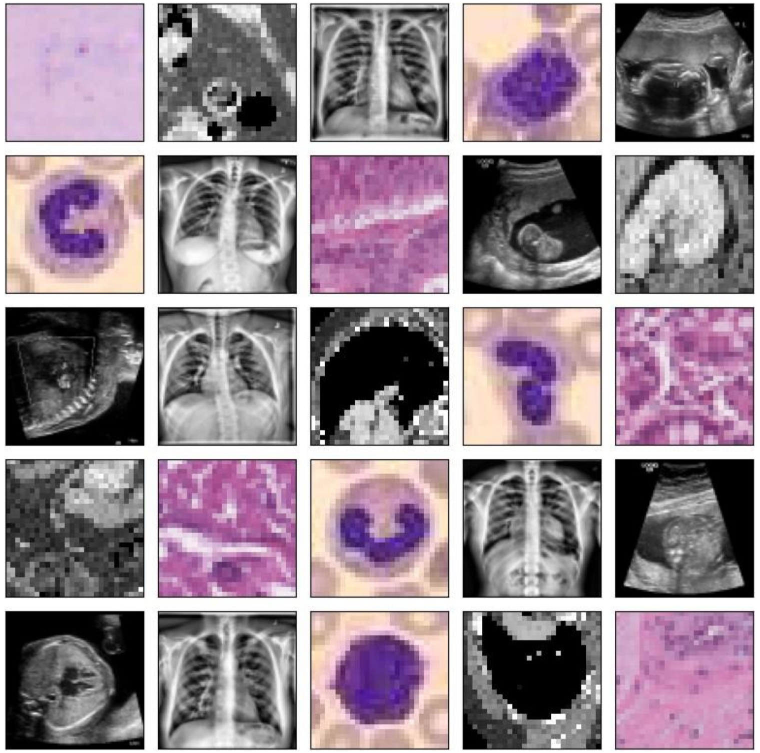

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [5, 6]
   - Several additional datasets, including those used in Madani et al. [7] and Chinn et al. [8]

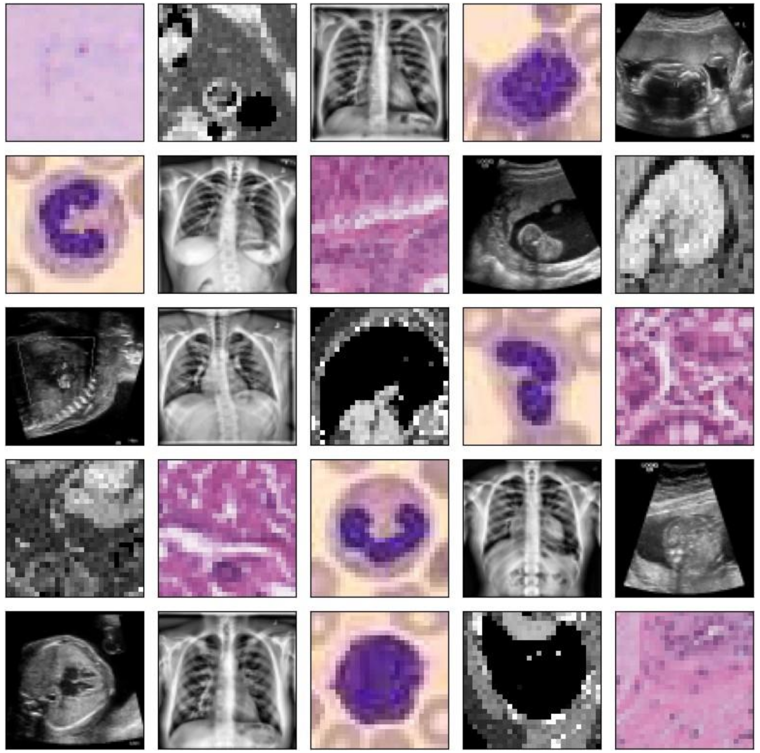

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [5, 6]
   - Several additional datasets, including those used in Madani et al. [7] and Chinn et al. [8]
2. From each dataset, sample the training set to create many subsets

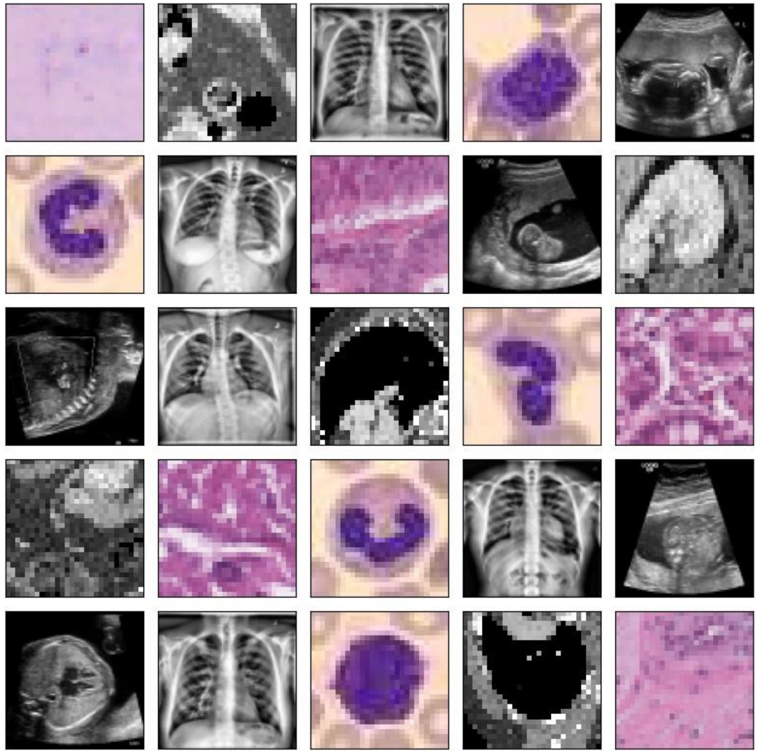

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [5, 6]
   - Several additional datasets, including those used in Madani et al. [7] and Chinn et al. [8]
2. From each dataset, sample the training set to create many subsets
3. Train a neural network classifier on each subset, and measure the performance of this classifier against a test set (common to subsets from the same parent dataset)

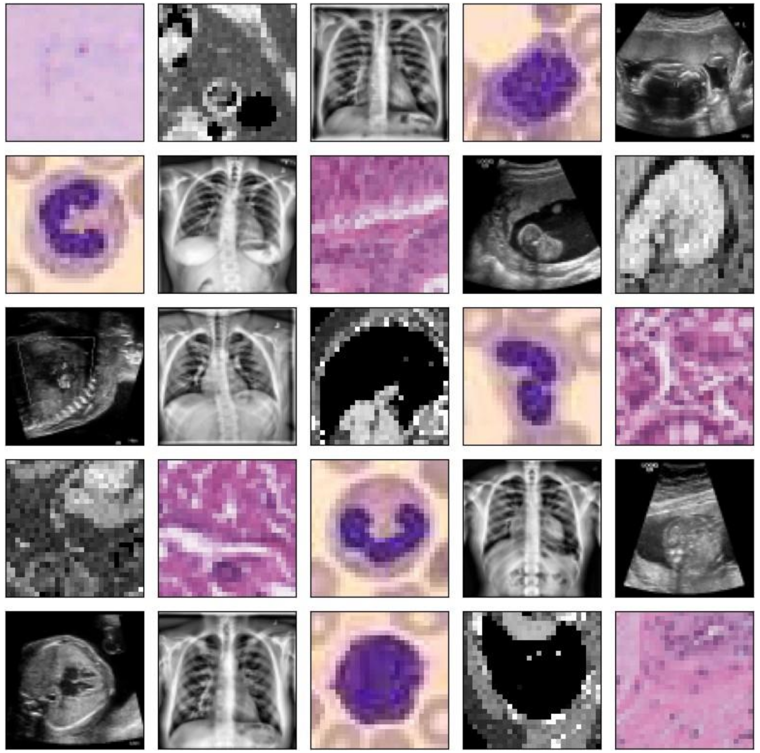

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [5, 6]
   - Several additional datasets, including those used in Madani et al. [7] and Chinn et al. [8]
2. From each dataset, sample the training set to create many subsets
3. Train a neural network classifier on each subset, and measure the performance of this classifier against a test set (common to subsets from the same parent dataset)
4. Measure an assortment of diversity indices for each subset (including class balance)

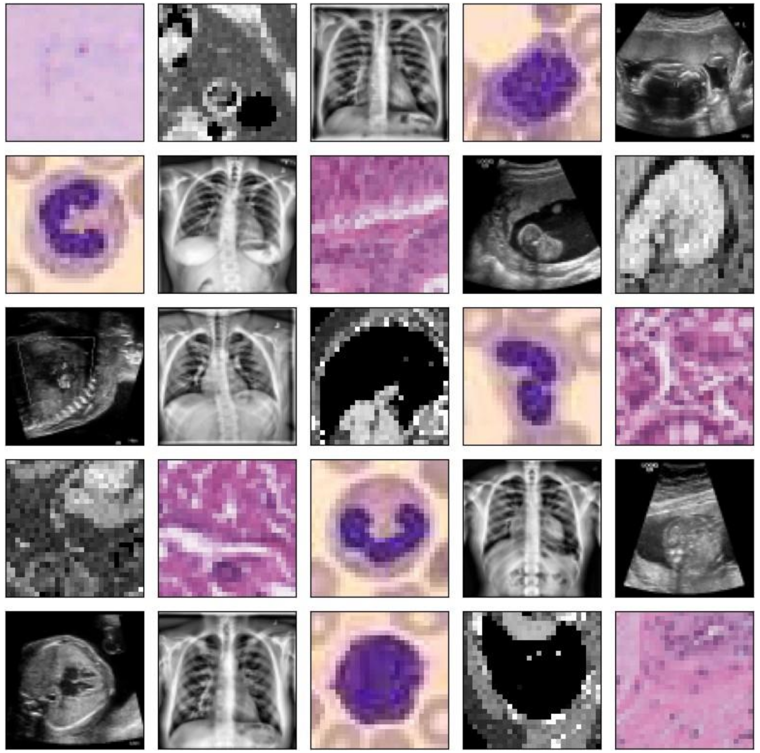

Figure: Images from some of the selected datasets

Methodology

1. Collect a number datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [5, 6]
   - Several additional datasets, including those used in Madani et al. [7] and Chinn et al. [8]
2. From each dataset, sample the training set to create many subsets
3. Train a neural network classifier on each subset, and measure the performance of this classifier against a test set (common to subsets from the same parent dataset)
4. Measure an assortment of diversity indices for each subset (including class balance)
5. Use linear regression to measure how much variation in model performance is explained by different sets of diversity indices.

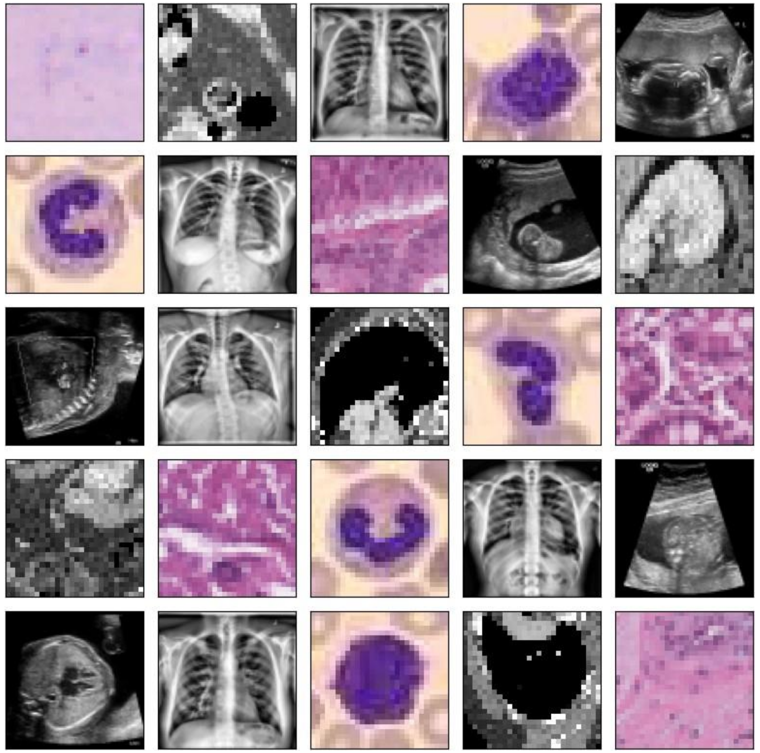

Figure: Images from some of the selected datasets

Preliminary Results

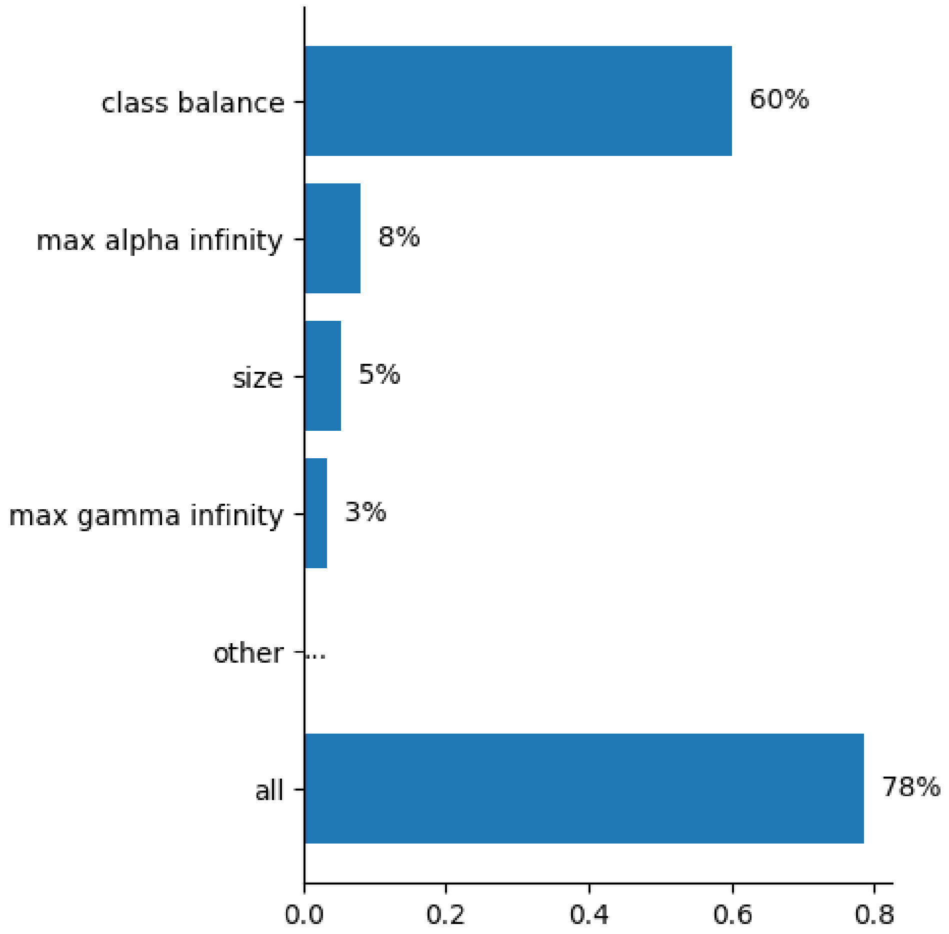

Figure: Additional variance of performance explained by feature

Preliminary Results

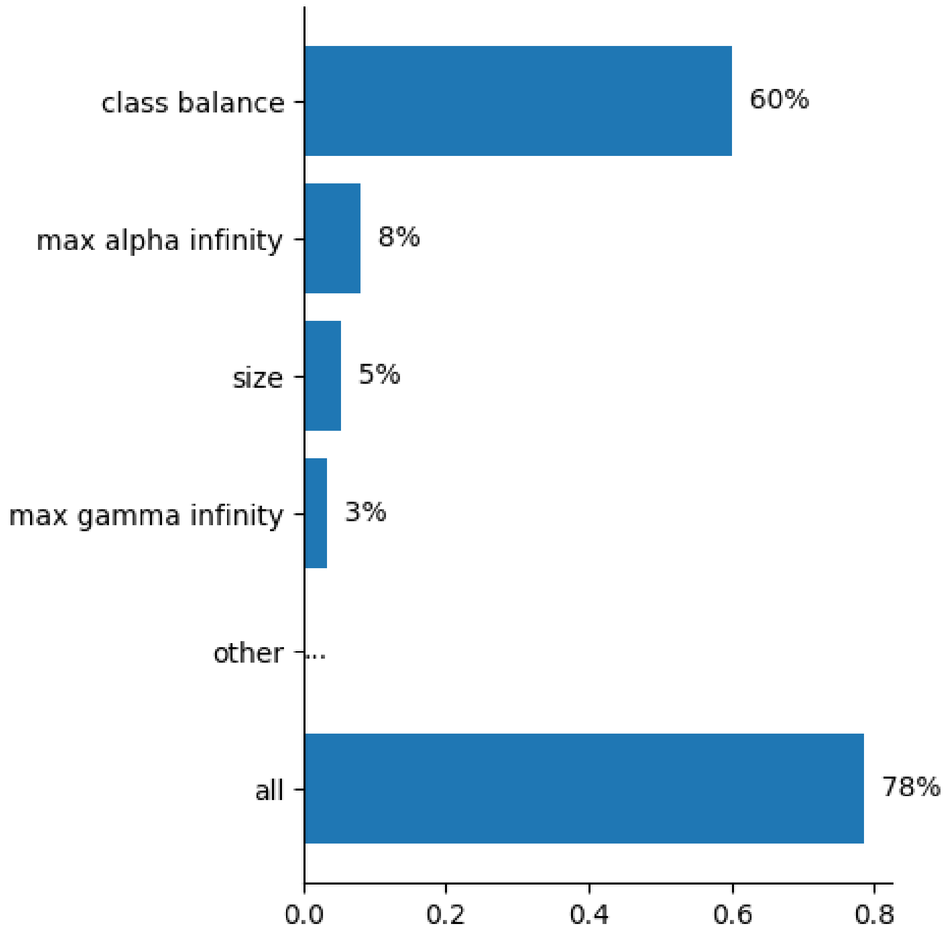

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%

Preliminary Results

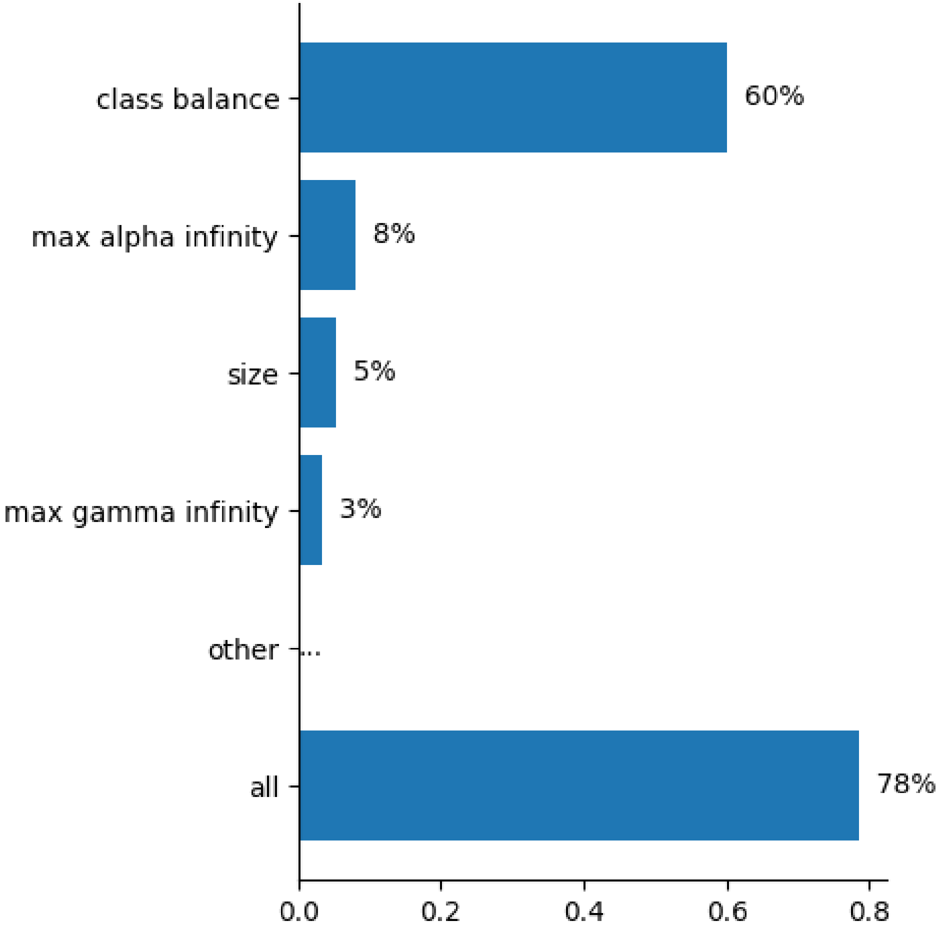

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.

Preliminary Results

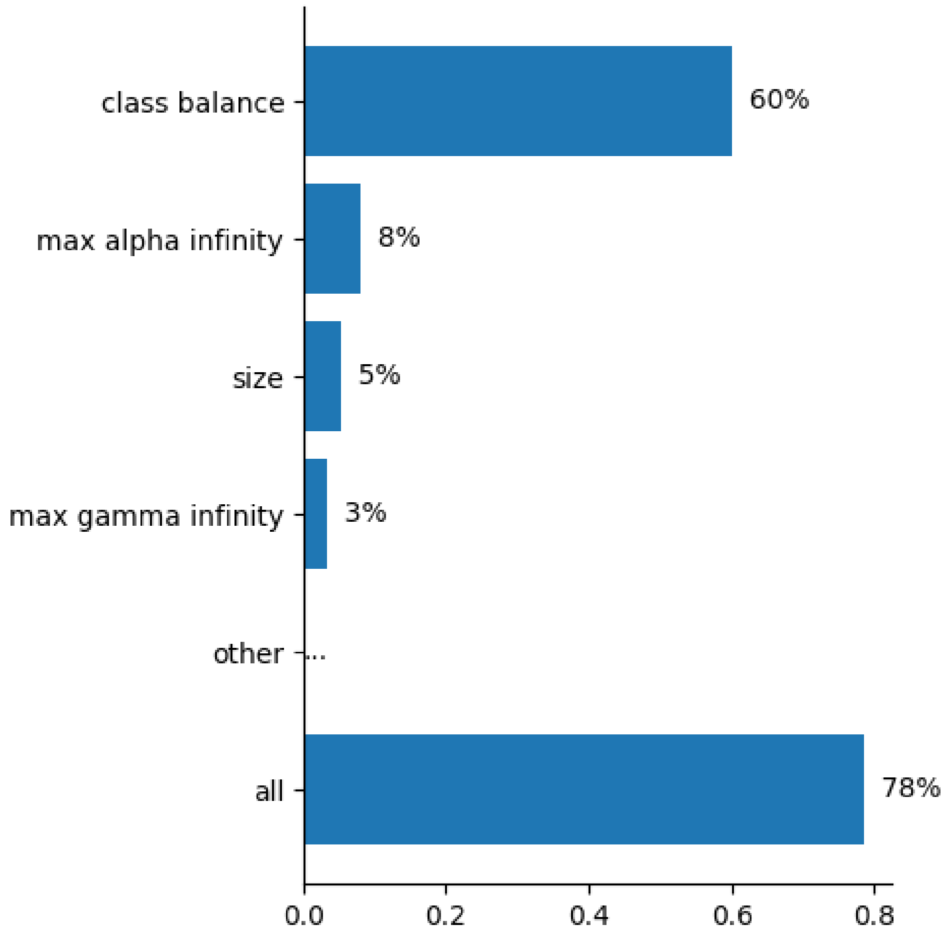

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- $\leftarrow$ Here we have the R^2 by feature

Preliminary Results

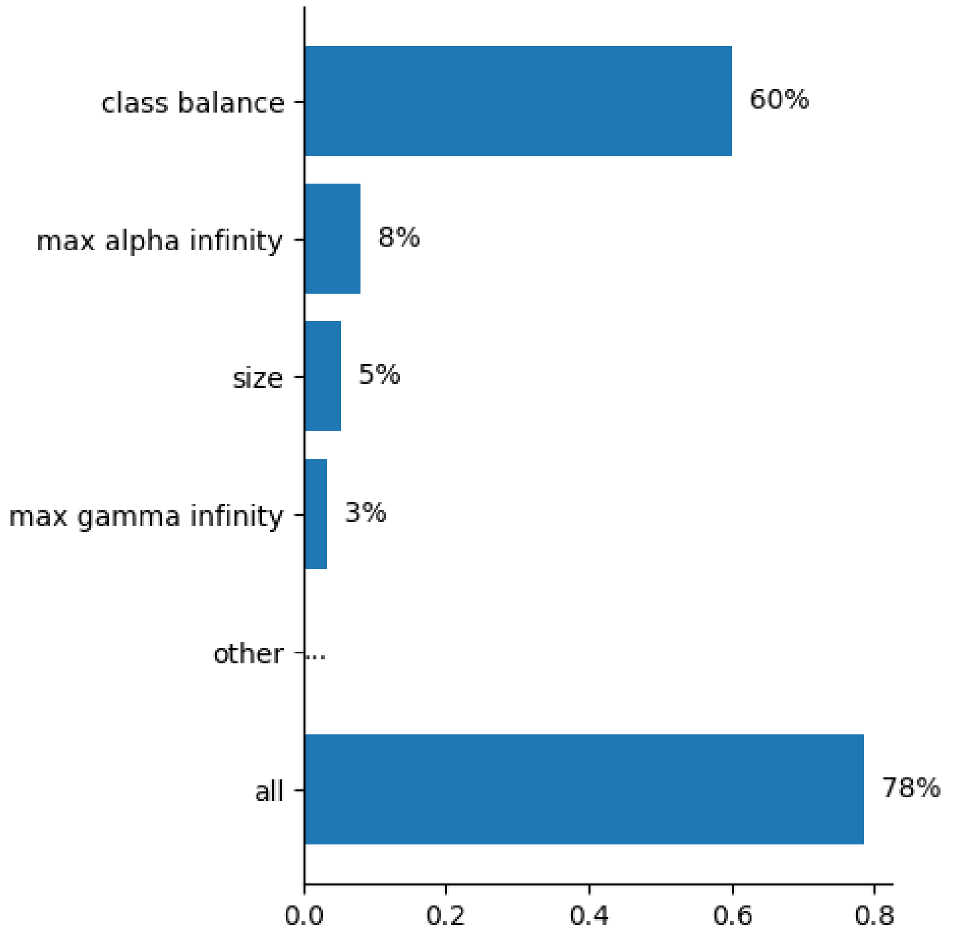

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.

Preliminary Results

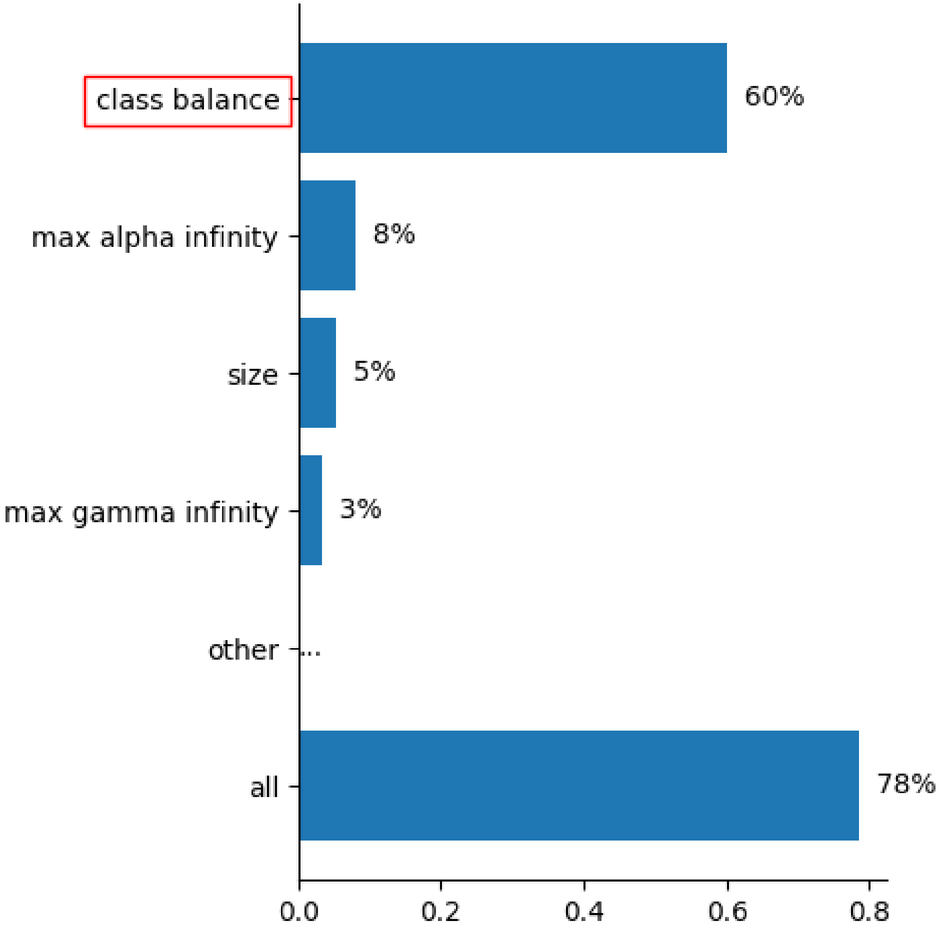

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)

Preliminary Results

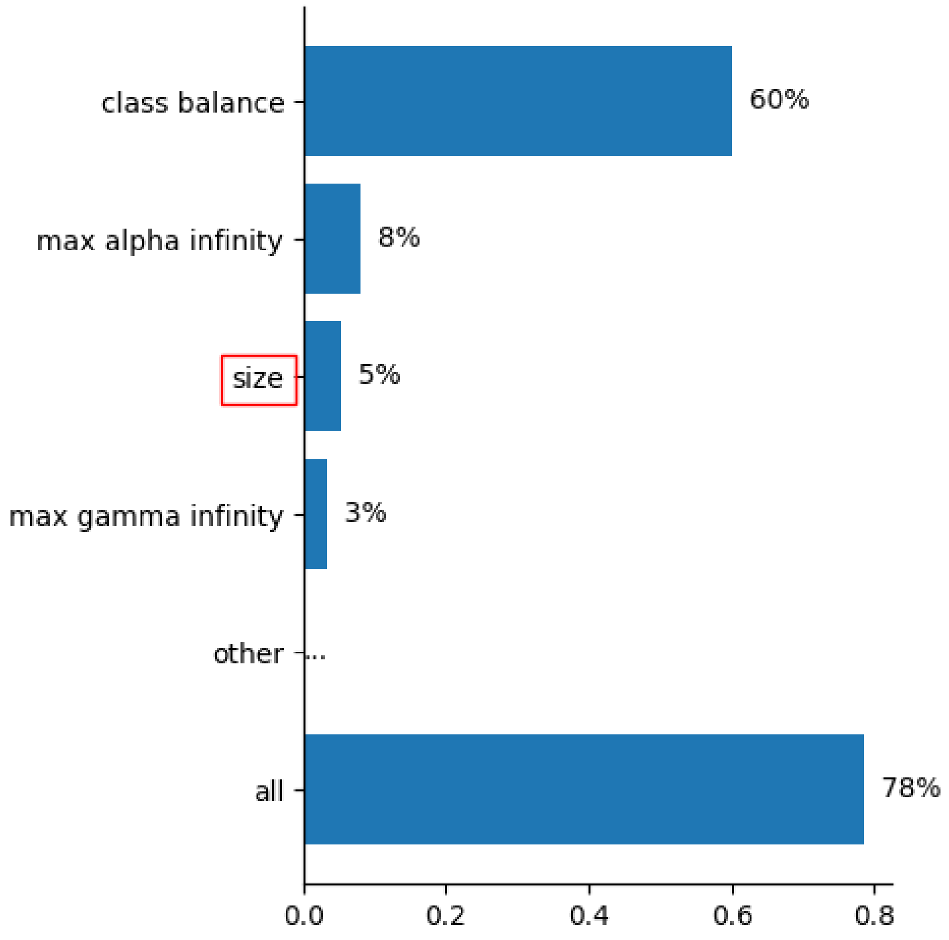

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important

Preliminary Results

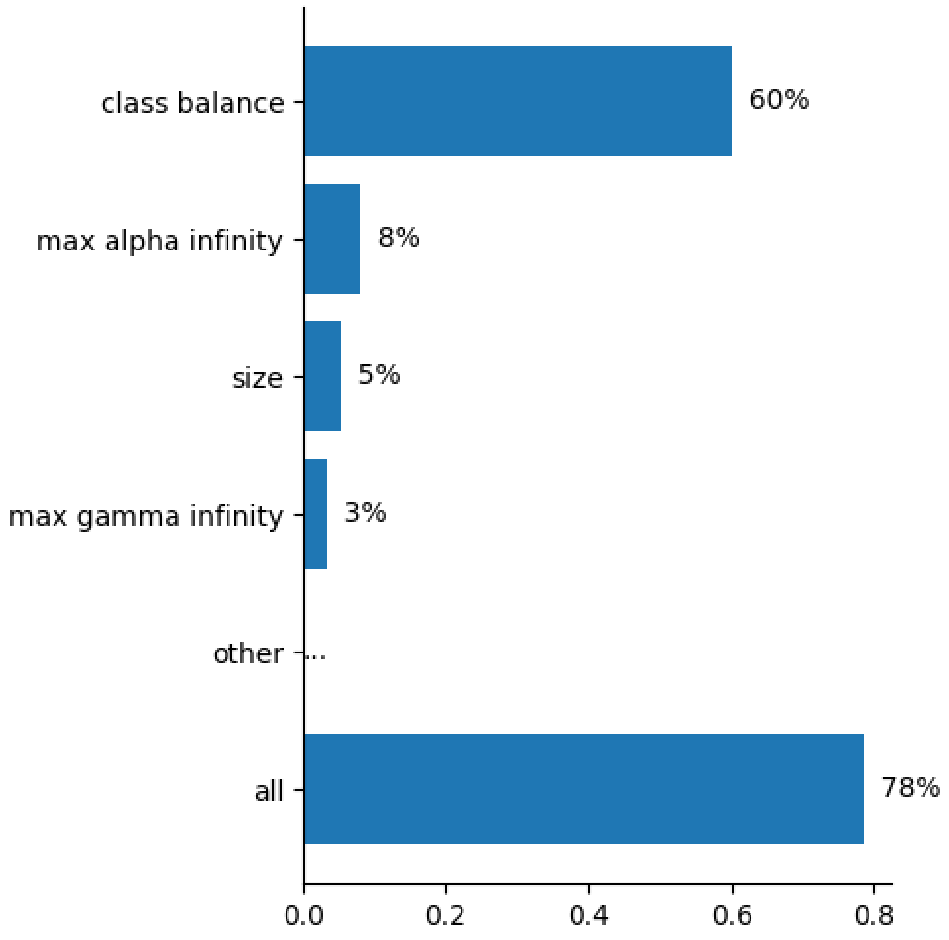

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important
  - A different diversity measure from the diversity framework turns out to be more important than subset size

Preliminary Results

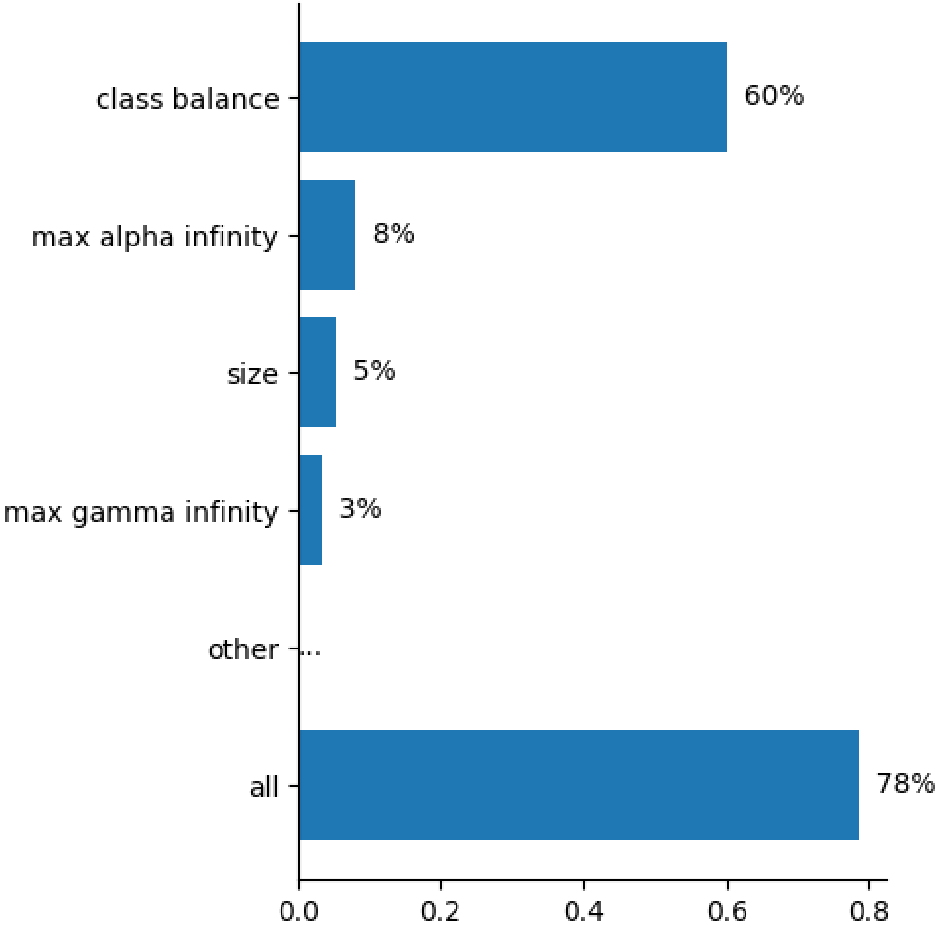

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important
  - A different diversity measure from the diversity framework turns out to be more important than subset size
  - A second diversity measures from the diversity framework turns out to be similarly important

Preliminary Results

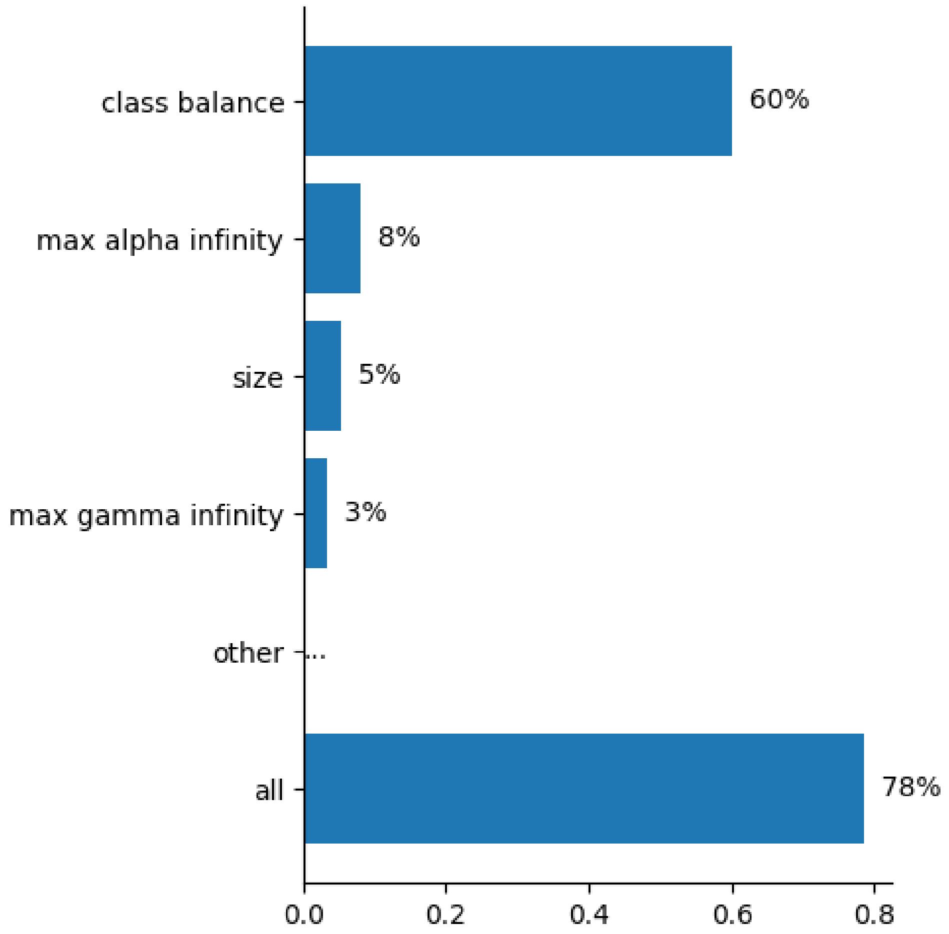

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important
  - A different diversity measure from the diversity framework turns out to be more important than subset size
  - A second diversity measures from the diversity framework turns out to be similarly important
  - The top four features explain about 77% of the variance in performance

Preliminary Results

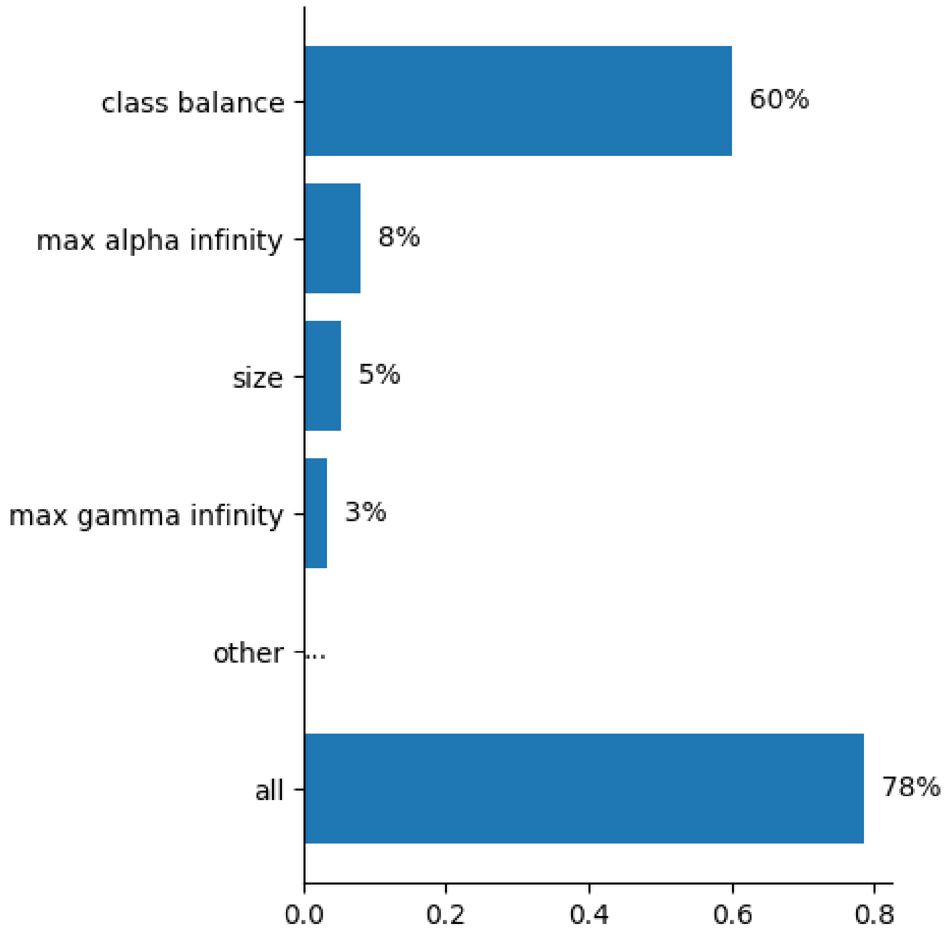

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important
  - A different diversity measure from the diversity framework turns out to be more important than subset size
  - A second diversity measures from the diversity framework turns out to be similarly important
  - The top four features explain about 77% of the variance in performance
  - The remaining features add only an additional $\approx$ 1%

Preliminary Results

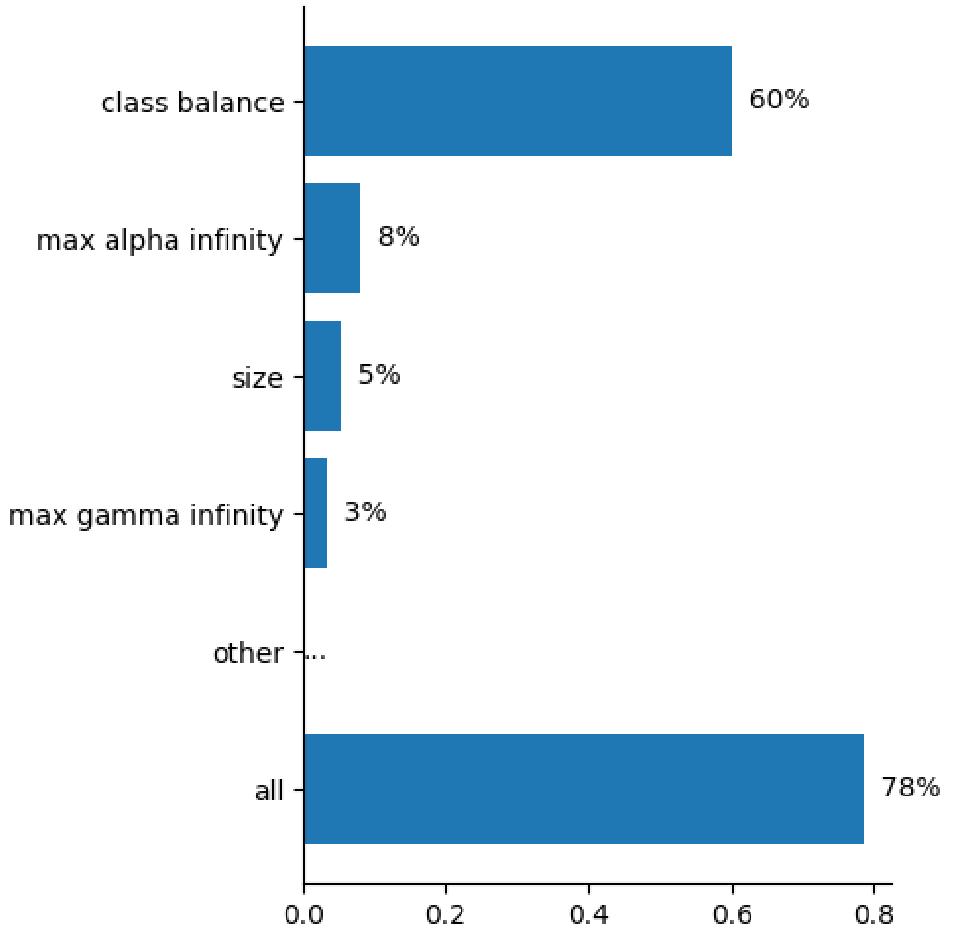

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important
  - A different diversity measure from the diversity framework turns out to be more important than subset size
  - A second diversity measures from the diversity framework turns out to be similarly important
  - The top four features explain about 77% of the variance in performance
  - The remaining features add only an additional ≈ 1%
- In summary, we have quantified the importance of class balance, and demonstrated that other diversity indices contribute to dataset quality

Preliminary Results

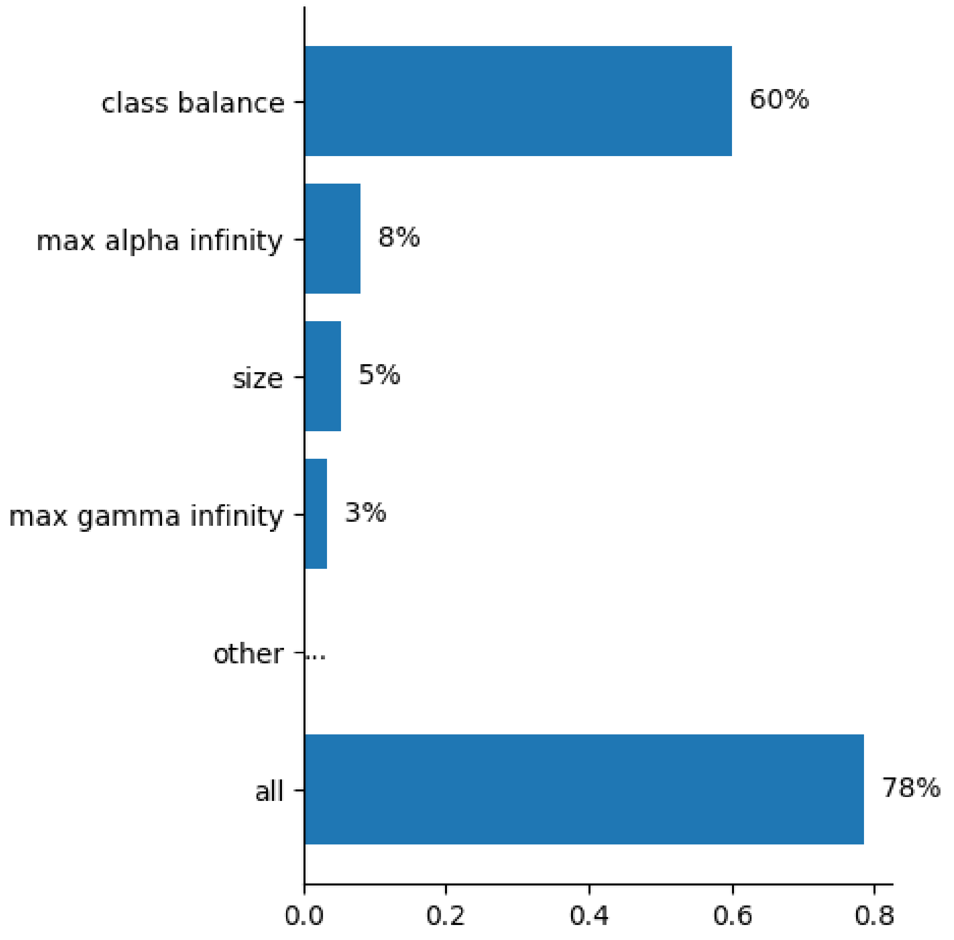

Figure: Additional variance of performance explained by feature

- Using only class balance and size, we achieve an R^2 of approximately 68%
  - I.e., these two features jointly explain about 2/3 of the variance across all datasets in the performance of models trained on those datasets.
- ← Here we have the R^2 by feature
  - These features were found using a greedy search. Reported R^2 contribution is the difference in R^2 before and after that features is included.
  - Class balance is the most important ($R^2 \approx 0.60$)
  - Subset size becomes the third most important
  - A different diversity measure from the diversity framework turns out to be more important than subset size
  - A second diversity measures from the diversity framework turns out to be similarly important
  - The top four features explain about 77% of the variance in performance
  - The remaining features add only an additional ≈ 1%
- In summary, we have quantified the importance of class balance, and demonstrated that other diversity indices contribute to dataset quality
- **We are testing on multiple datasets, and would love to test on more. If you have data that might benefit from this approach, we would love to collaborate, just reach out!**

Thank you for attending my talk!

Websites:

- arnaoutlab.org
- arnaoutlab.ucsf.edu

Contact Info:

- JC: jcouch1@bidmc.harvard.edu
- Ramy Arnaout (PI): rarnaout@bidmc.harvard.edu
- Rima Arnaout (PI): rima.arnaout@ucsf.edu

References

- P. Nguyen, R. Arora, E. D. Hill, J. Braun, A. Morgan, L. M. Quintana et al., *greylock: A python package for measuring the composition of complex datasets*, 2023.
- T. Leinster and C. A. Cobbold, *Measuring diversity: the importance of species similarity*, Ecology **93** 477.
- R. Reeve, T. Leinster, C. A. Cobbold, J. Thompson, N. Brummitt, S. N. Mitchell et al., *How to partition diversity*, 2016.
- M. O. Hill, *Diversity and evenness: A unifying notation and its consequences*, .
- J. Yang, R. Shi and B. Ni, *Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis*, in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.
- J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke et al., *Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification*, Scientific Data **10** (2023) 41 [2110.14795].
- A. Madani, R. Arnaout, M. Mofrad and R. Arnaout, *Fast and accurate view classification of echocardiograms using deep learning*, .
- E. Chinn, R. Arora, R. Arnaout and R. Arnaout, *Enrich: Exploiting image similarity to maximize efficient machine learning in medical imaging*, Journal of the American Medical Informatics Association **30** (2023) 1079 [https://www.medrxiv.org/content/early/2021/05/25/2021.05.22.21257645.full.pdf].