Training Same Size Effect on Deep Learning Models for Geographic Atrophy

Robert Slater[2], Jacob Bogost[1], Rohit Balaji[1], Gelique Ayala[1], Jeong Pak[1], Rick Voland[1], Barbara A. Blodi[1], Roomasa Channa[1], Donald Fong[3], Amitha Domalpally[1,2]

[1]WISCONSIN READING CENTER
[2]A-EYE Unit
[3]Annexon Biosciences
Department of Ophthalmology and Visual Sciences
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH
Research to Prevent Blindness

224-C0063
ARVO 2023

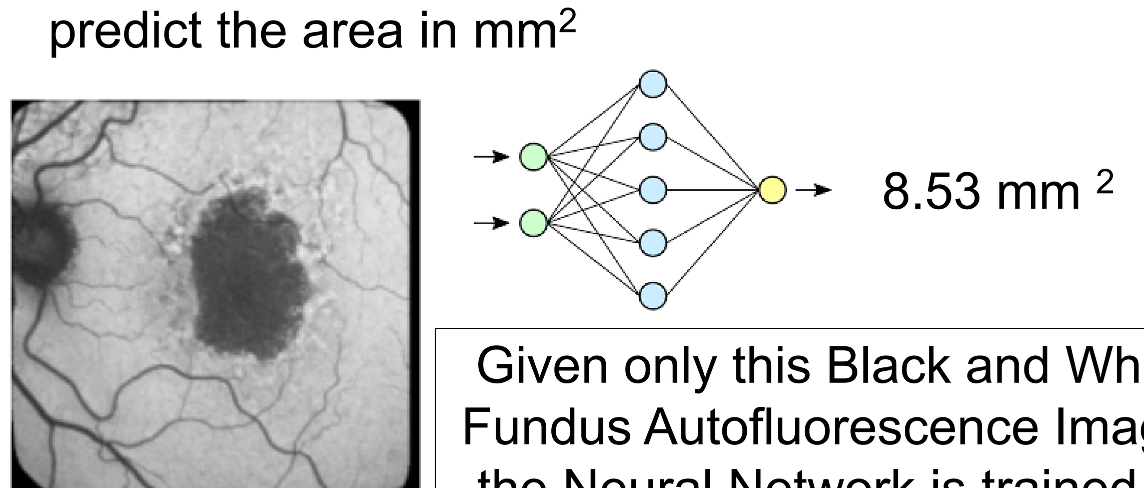
Background and Purpose

- Data is expensive and time consuming to explain and annotate
- One of the most important questions for experiments is “How much data do we need”
  - The answer is often “As much as we can get”
  - More is better
- Geographic Atrophy (GA) is a debilitating eye disease with only a single approved treatment
- The standard modality to measure GA is Fundus Autofluorescence, which is not widely available, limiting the number of examples that can be used for Deep Learning (DL) or Artificial Intelligence (AI) Models.

- **What effect does sample size have on the performance of AI models?**

Methods

- 1515 Autofluorescence images from Age-Related Eye Disease Study 2 were used in the training set
- An independent set of 511 images were used for validation
- At each percent level an EffcientNetB0 was trained on either the full training set (p=100) or a random subset of the training set (p<100).
- The target of the model was the area of GA as measured by human graders. The AI was trained to predict the area in mm^2

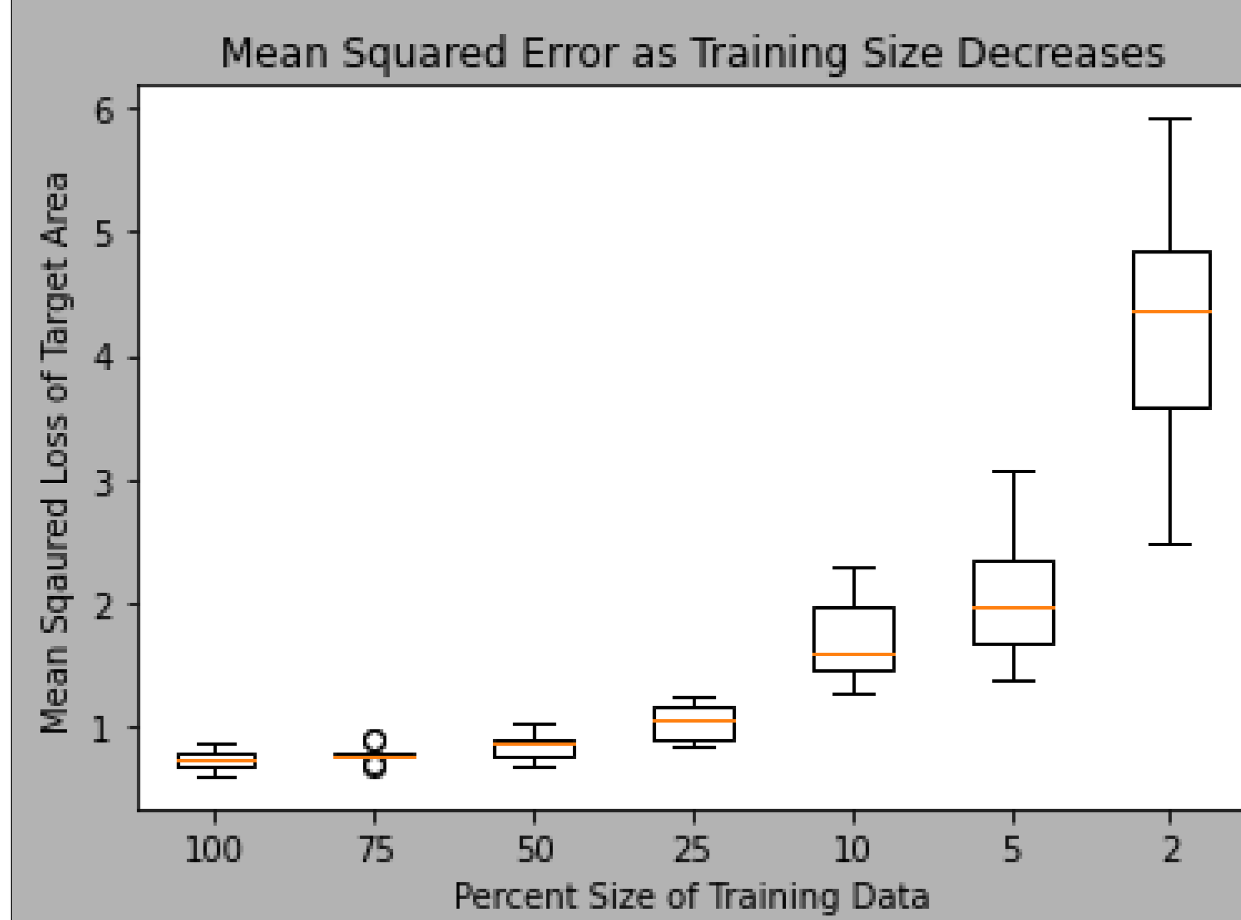


Given only this Black and White Fundus Autofluorescence Image, the Neural Network is trained to predict an area (number only)

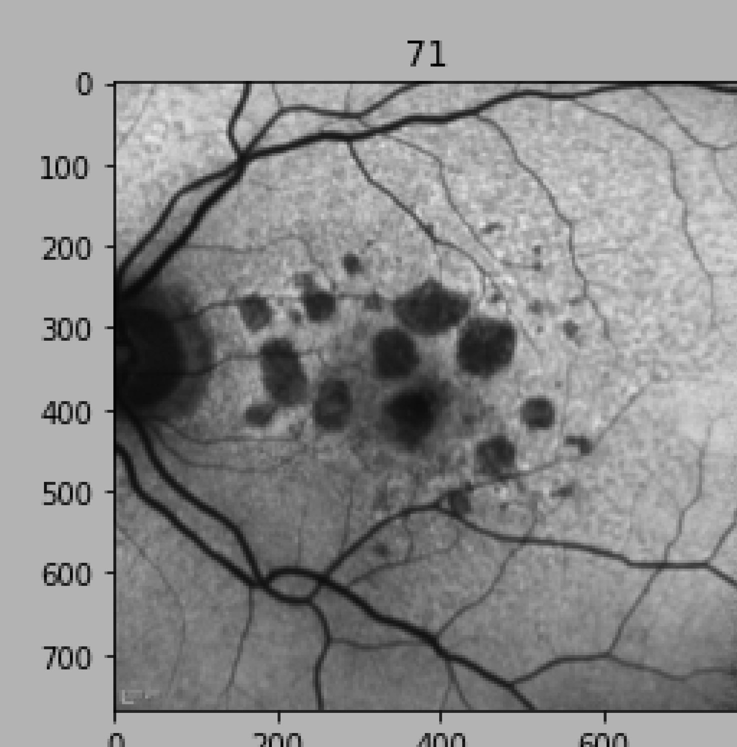
- At p = (100,75,50,25,10,5,2) a p percent random subset of the training data was selected and used to train the model, while the validation set was held constant at n=511
- Analysis:
  - The lowest MSE validation loss for each run
  - The number of training steps; the batch size was held constant so each training step represents a forward and backward pass of the network and is a metric for the time to train.

Commercial interest disclosures: NONE for RB, RS, JB, GA, JP, RV, BB, RC, AD
DF is an employee of Annexon Biosciences

Training Sample size	Percentage of available training sample	Mean area on training set (n = 511)	Mean difference between ground truth and AI prediction (95% CI)	Intraclass Correlation (ICC)
1515	100%	5.37	1.39 (0.07 – 4.52)	0.94
1135	75%		1.39 (0.06 – 4.74)	0.93
757	50%		1.50 (0.08 – 5.14)	0.93
378	25%		1.61 (0.07 – 5.65)	0.90
151	10%		2.07 (0.10 – 7.39)	0.85
75	5%		2.33 (0.13 – 8.23)	0.81
30	2%		3.35 (0.22 – 11.55)	0.57



- Mean Squared Error was used as a loss function and thus us a proxy for model performance.
- The lower the MSE on the validation set, the better the model performance.
- The MSE is comparable at 100%, 75% and 50% sample size indicating that model accuracy is maintained even with 50% (n= xx) training data.
- At 25% sample size (n =378) , model accuracy starts dropping



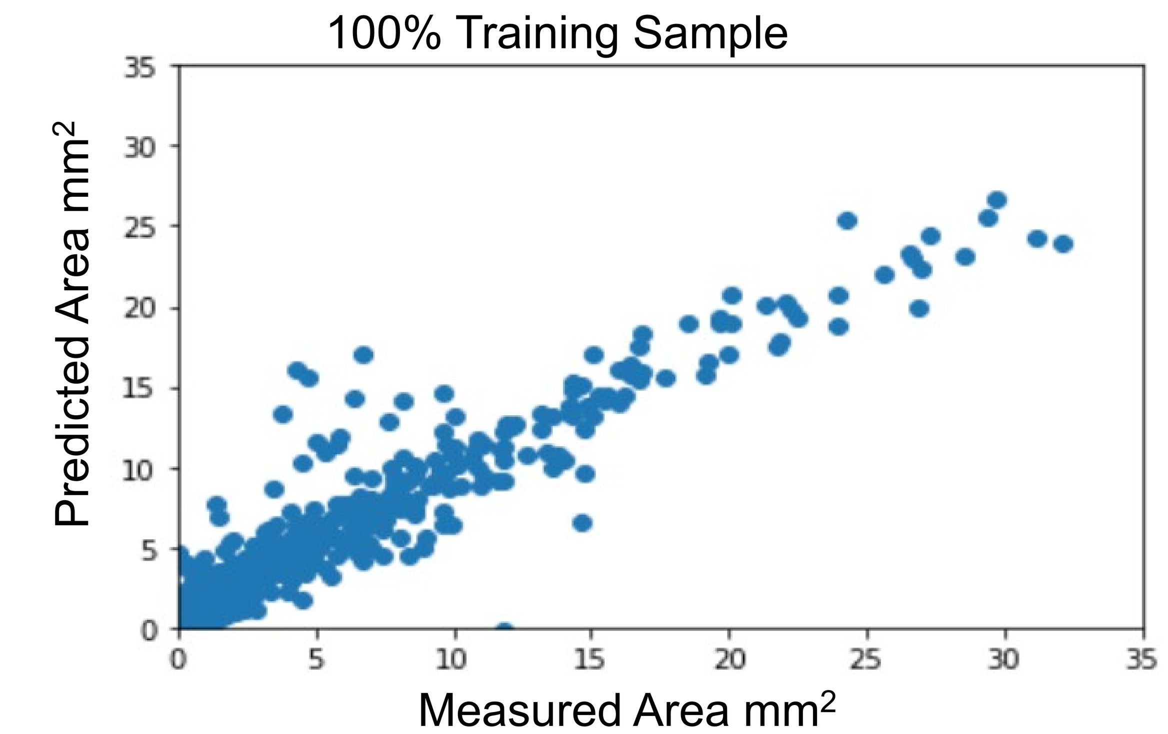
Ground truth : 3.25 mm^2

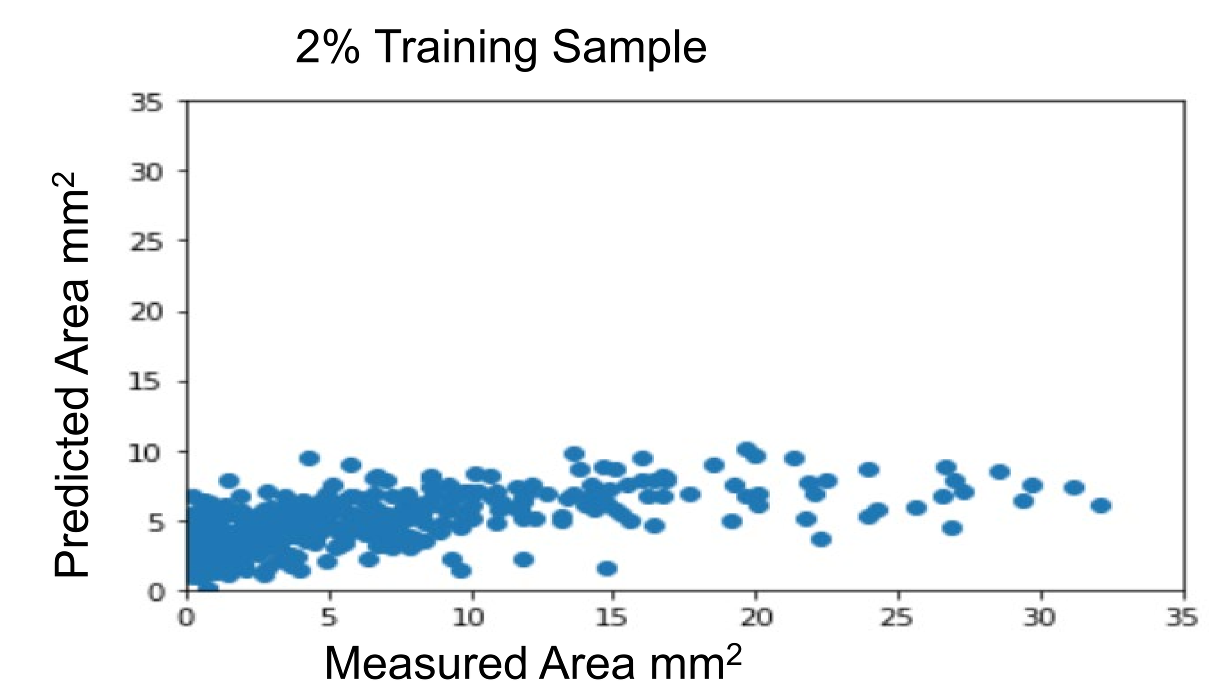
The table below shows AI prediction of area of GA on this image . Each row represents the prediction of an independent model that was trained on sequential reduction of training data

Model training size (percentage of available data used for training	Area prediction	Percentage difference between ground truth and AI prediction
100%	3.07	5.4%
75%	3.19	2.0%
50%	3.00	7.4%
25%	2.76	15.1%
10%	2.97	8.3%
5%	2.98	8.0%
2%	3.17	2.5%

Results

- In this use case of predicting GA area, the algorithm performance metrics were similar with sample size ranging between 500 - 1500 autofluorescence images
- Performance of the algorithm started dropping at 25% data (~375 autofluorescence images and was significantly worse at < 25%.





Decreasing the training size causes the model to severely under preform at larger GA areas

Conclusions

- Training sample size is an important hyperparameter that influences the learning of the algorithm

- Annotations can take 30-60 Minutes, so training with less data can decrease cost, at the risk of decreased performance.

References

1. Bailly A, Blanc C, Francis É, et al. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Computer Methods and Programs in Biomedicine. 2022/01/01/ 2022;213:106504.
2. Domalpally A, Danis R, Agron E, Blodi B, Clemons T, Chew E. Evaluation of Geographic Atrophy from Color Photographs and Fundus Autofluorescence Images: Age-Related Eye Disease Study 2 Report Number 11. Ophthalmology. Jul 19 2016