Generalizable AI-based glaucoma prediction via a stable model selection method

Homa Rashidisabet[1,3], Abhishek Sethi[2,3], Ponpawee Jindarak[2], James Edmonds[2,3], R.V. Paul Chan[2,3], Yannek I. Leiderman[1,2,3], Thasarat Sutabutr Vajaranant[2,3], Darvin Yi[1,2,3]

[1] Department of Biomedical Engineering, University of Illinois Chicago, 851 S Morgan St, Chicago, IL 60607

[2] Illinois Eye and Ear Infirmary, Department of Ophthalmology and Visual Sciences, University of Illinois Chicago

[3] Artificial Intelligence in Ophthalmology (Ai-O) Center, University of Illinois Chicago

INTRODUCTION

BACKGROUND

In recent years, deep learning models in computer vision and natural language processing have become more complex, with a vast number of parameters. These models have remarkable ability to learn complex relationships in datasets, leading to unprecedented performances. However, a major disadvantage of these models is their tendency to overfit, which occurs when the model learns from noise in the data during training, negatively affecting its ability to generalize. This can cause the model to perform poorly on validation and test data compared to the training data. Additionally, overfitting can cause high variance, which measures the variability in the model's predictions if a different dataset or slight changes in data are used. As a result, the model can be significantly impacted by even minor variations in data, leading to reduced reliability.

CHALLENGES

In the medical domain, obtaining data for training deep learning models is challenging and expensive. This is because the data requires professional domain knowledge to label, making it difficult to obtain sufficient quantities of labeled data. Furthermore, the scarcity of data in combination with the high complexity of deep learning networks increases the likelihood of overfitting, which makes it difficult to use conventional approaches to mitigate overfitting and high variance.

EXISTING METHODS

To address overfitting and reduce model variance, there are several methods and techniques that can be utilized. One approach is to obtain more data, which can help increase the generalizability of the model. Another method is to incorporate data augmentation, regularization (such as L1 or L2), dropout, smaller input dimensionality, smaller models, or decreasing batch size. Ensemble modeling is another technique that can mitigate overfitting and reduce variance. This involves creating multiple models with diverse hyperparameters and random initializations and then aggregating the predictions of each model to produce a final prediction. The majority voting for the models can help decrease the generalization error of the model's prediction.

OUR PURPOSE

Given the limited amounts of labeled data in the Ophthalmology domain, we propose a novel model selection method for the DL-based classification task that avoids overfitting in small data regimes. We evaluated our method on the glaucoma prediction application.

HYPOTHESES

(1) The baseline classification (BCL) method, selecting the best model solely based on maximum validation accuracy, overfits the data when the validation and train sets are small.

(2) Selecting the best model based on both high accuracy and low variability could improve the generalization and reduce overfitting.

METHODS

We used 1340 fundus images from the Illinois Eye and Ear Infirmary with 683 glaucoma and 657 non-glaucoma patients. Table 1 summarizes the number of images in the train, validation, and test sets.

CLASS	TRAIN SET	VALIDATION SETS					TEST SET
Glaucoma	450	1	5	15	50	150	83
Non-glaucoma	450	1	5	15	50	150	57
Total	**900**	**2**	**10**	**30**	**100**	**300**	**140**

TABLE 1
Data breakdown per class in train, validation and test sets.

Our proposed method selects the best model based on both high validation accuracy and low validation variability, shown in Figure 1. The red star shows the model with max validation accuracy, picked by BCL. The green star shows the best model selected by our proposed method.

We summarized our method in the following steps.

1. We employed 15 random hyperparameter (HP) searches. Per HP-search, we:
   - Slide a window over the validation accuracy curve.
   - Scored each window using equations 1 and 2
   - Selected the window with the highest score (S^*) using equation 2
   - Within the optimal window scored S^* for HP_i, we selected the model with max accuracy (M_i^*) using equation 3
2. We use the majority vote of the 15 obtained M_i^* to predict the labels

$$CV_j = \frac{\sigma_j}{\mu_j} \qquad (1)$$

$$S^* = \underset{S_j}{\operatorname{argmax}}\left(\mu_j - CV_j\right) \qquad (2)$$

$$M_i^* = \underset{HP_i}{\operatorname{argmax}}\left(M_{iS^*}\right) \qquad (3)$$

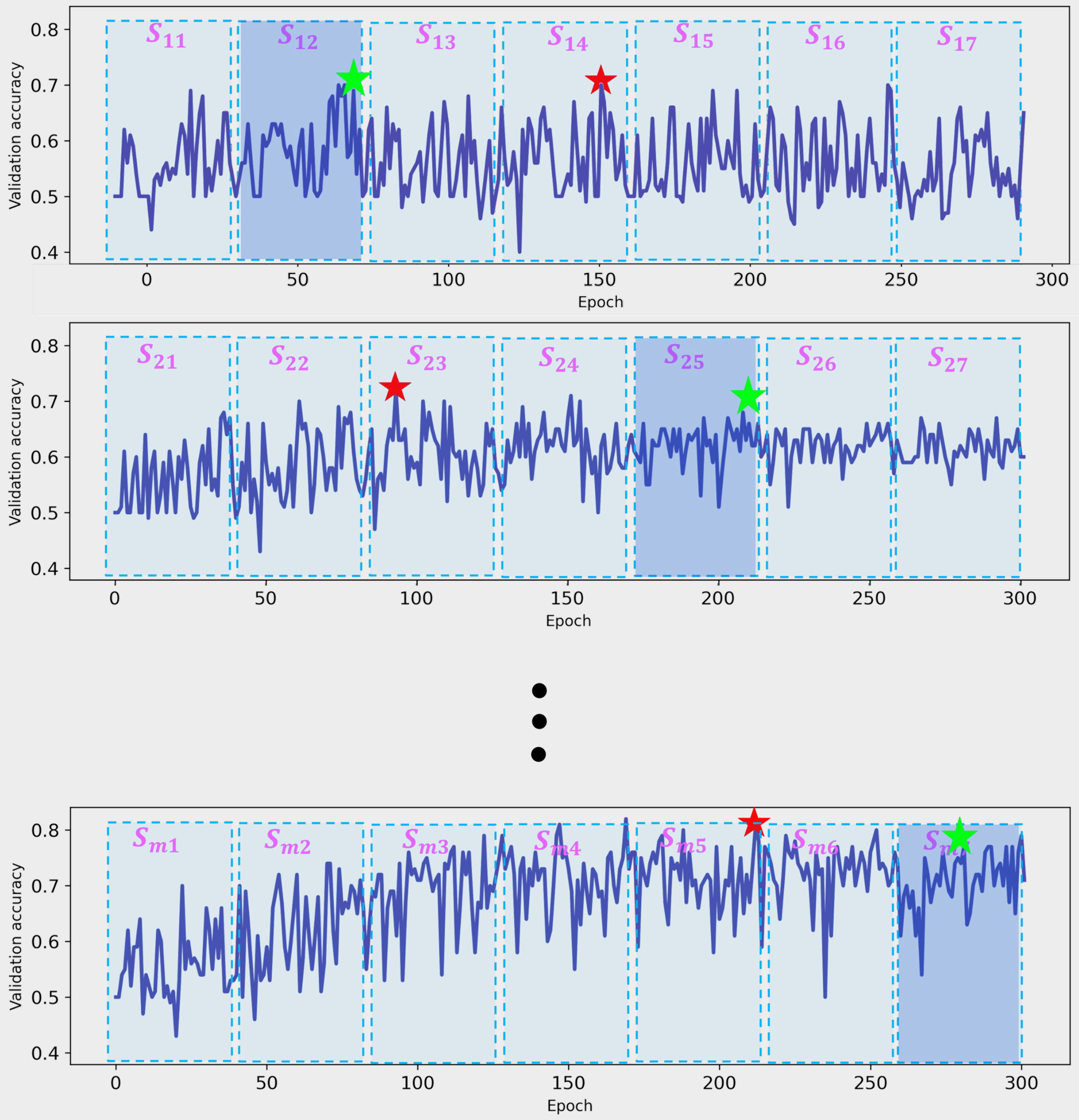

FIGURE 1
Our proposed model selection method for deep learning-based classification in small data regimes.

RESULTS

- Our proposed method outperforms the baseline classification (BCL) in classifying glaucoma using different validation set sizes.
- For validation sizes of 2, 10, 30, and 300, our method improves the test accuracy by 17%, 11%, 2%, and 9%, respectively, compared to BCL.
- BCL outperforms our method by only 1% when the validation size is 100.
- Our proposed method predicts glaucoma with high accuracies of 79% and 85% when there are only 2 and 10 images in the validation set, respectively, whereas BCL cannot predict accurately with limited train and validation data.
- Our proposed method improves the performance of BCL for predicting glaucoma and generalizes well to unseen test data across various validation set sizes.

DATASET	MODEL	SAMPLE SIZES PER CLASS IN THE VALIDATION SET				
		2	5	15	50	150
Fundus	BCL	62 (64)	74 (84)	80 (88)	**85** (92)	79 (89)
	Proposal	**79** (88)	**85** (90)	**82** (88)	84 (94)	**88** (92)

TABLE 2
Test sets' performance is shown as accuracy % (AUC %). Bold values show the max accuracy

CONCLUSION

1) Our proposed method was able to improve deep learning generalization, indicating that the selected model could perform well on unseen data.
2) Our method was not reliant on the validation set size, making it effective even in situations where the amount of available data is limited.
3) Our findings highlighted our method's potential for improving the accuracy and stability of DL models for medical image analysis, particularly in the context of glaucoma diagnosis.

SUPPORT

All work done for this project was supported by the Research to Prevent Blindness (RPB) Foundation.