Supplementary Materials: Training Optimizations, Additional CNN Ensemble Architecture, Sensitivity/Specificity, Heatmaps, and Alternative TCAV Metric

Kaveri A. Thakoor, *Student Member, IEEE*, Sharath C. Koorathota, *Student Member, IEEE*, Donald C. Hood, and Paul Sajda, *Fellow, IEEE*

I. Methods for Training Optimization

In addition to the new architectures developed and described in our main paper, we used the 'Field' test set to evaluate two model training optimizations that did not require changing model architectures: data augmentation and multimodal image input [1], [2]. The following sections describe the methods and results based on implementation of these training-based optimizations.

A. Varying CNN Training Factors: Data Augmentation and Input Image Modality

For the best-performing hybrid deep learning/machine learning (DL/ML) CNN on both 'Lab' and 'Field' datasets (from left half of Table 1 in main paper), the impact on performance of incorporating data augmentation during training and varying input (only RNFL probability maps vs. RNFL and RGCP maps together, red and violet boxes, respectively, in Figure 1) was assessed.

1) Varying Training Via Data Augmentation: Since we know that the 'Field' dataset is obtained from a machine that induced an 8% scale change in collected images [1], [2], we attempted to mimic this modulation in our training process by augmenting the training dataset using Keras ImageDataGenerator 'zoom' setting at 0.1 (augmenting with images 10% larger or smaller in size than original images). Training with data augmentation was carried out on the 737-image dataset from previous work [3] using Monte Carlo cross-validation splits: the data used for training vs. for validation was split randomly in a ratio of 80%:20% on each run. Note that the 737-image dataset contains the 'Lab' test set, so due to data augmentation and randomized (Monte Carlo) cross validation used during training, a small randomized portion of this set may be present in the training set for the end-to-end DL models. Impact of data augmentation on test accuracy for the best performing hybrid DL/ML model for the 'Field' dataset is reported in the 'Results of Training Optimization' section below.

2) Varying Input Image Modality to the CNNs: In addition to evaluating performance with and without data augmentation for the top-performing hybrid DL/ML model, we assessed impact on CNN accuracy as a result of varying input image modality: features were extracted and classified from RNFL map input alone or were extracted, concatenated, and then classified from RNFL maps and RGCP maps combined during CNN training and testing. In both cases, classification of single modality or dual modality concatenated features was carried out by a Random Forest classifier [1], [2].

B. Results of Training Optimization

We present results of the non-architectural (training-based) improvements that enhanced robustness of the best hybrid DL/ML model on the new 'Field' dataset.

Since the ResNet-18 + Random Forest hybrid DL/ML model exhibited best performance accuracy on both 'Lab' and 'Field' datasets (left half of Table 1 from main paper, in bold, replicated from main paper below for ready reference), data augmentation and varied input image modality training enhancements were performed only for the ResNet-18 + Random Forest model.

The ResNet-18 + Random Forest model with data augmentation and with RNFL probability map input alone served to most improve model performance on the 'Field' dataset, enabling up to 85.9% accuracy when transferred to this new test set. In addition to 10% scale change, we observed that by increasing data augmentation variants to include vertical flips and horizontal flips (analogous to providing more left eyes or right eyes for each patient, as only one eye from each patient was originally present in any of the training, validation, or test sets) [1], [2], we were able to increase

This work was supported in part by NIH Grant RO1-EY02115 awarded to D.C.H. The NSF Graduate Research Fellowship (Grant DGE-1644869) awarded to K.A.T. is gratefully acknowledged.

K. Thakoor is with the Department of Biomedical Engineering, Columbia University, New York, NY 10027, USA (e-mail: k.thakoor@columbia.edu).

S. Koorathota is with the Department of Biomedical Engineering, Columbia University, New York, NY 10027, USA (e-mail: sharath.k@columbia.edu).

D. Hood is with the Departments of Psychology and Ophthalmology, Columbia University, New York, NY 10027, USA (e-mail: dch3@columbia.edu).

P. Sajda is with the Departments of Biomedical Engineering, Electrical Engineering, and Radiology, Columbia University, New York, NY 10027, USA (e-mail: ps629@columbia.edu).

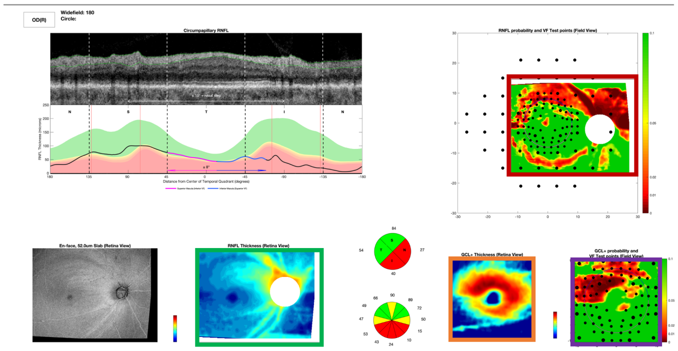
OD(R)
Widefield: 180
Circle:
Circumpapillary RNFL
± 15° + nasal step
N
S
T
I
N
RNFL Thickness (microns)
±8°
Distance from Center of Temporal Quadrant (degrees)
Superior Macula (Inferior VF)
Inferior Macula (Superior VF)
RNFL probability and VF Test points (Field View)
En-face, 52.0um Slab (Retina View)
RNFL Thickness (Retina View)
GCL+ Thickness (Retina View)
GCL+ probability and VF Test points (Field View)

Fig. 1. Full OCT Report used by OCT expert to detect glaucoma. Red box indicates an RNFL probability map. Violet box shows an RGCP probability map. Orange box contains RGCP thickness map, and green box contains RNFL thickness map. (Replicated from main paper for ready reference.)

TABLE I

ACCURACY RATES (%) FOR HYBRID DL/ML MODELS (LEFT HALF) AND FOR END-TO-END DL MODELS (RIGHT HALF). PERCENT REDUCTION FROM 'LAB' TO 'FIELD' TEST SETS SHOWN IN FOURTH COLUMN AND EIGHTH COLUMN FOR EACH MODEL TYPE, RESPECTIVELY.

Hybrid DL/ML Models	Lab	Field	% Reduction	End-to-End DL Models	Lab	Field	% Reduction
Conv+FC	95.7	79.3	17.1	DenseNet-121+FC	95.9	85.2	11.2
Conv+RF	94.0	80.7	14.1	CNN Ensemble	94.4	88.9	5.83
VGG16+RF	95.0	74.1	22.0	VGG16+FC	97.0	85.2	12.2
ResNet18+RF	**94.8**	**80.7**	14.9	ResNet18+FC	94.9	83.0	12.5
InceptionV3+RF	94.2	69.6	26.1	**InceptionV3+FC**	**90.4**	**91.1**	-0.774

'Field' set accuracy, reducing deterioration from the 'Lab' set to 9.39% (compared to the reduction of 14.9% prior to incorporation of data augmentation for ResNet-18 + Random Forest from Table 1, fourth column). Comparison of impact of data augmentation and varying multimodal image input is shown in the Receiver Operating Characteristic (ROC) curves in Fig. 2. ROC Area under the curve (AUC) is highest for RNFL input alone with data augmentation. RNFL input alone without data augmentation has second-highest AUC, followed by RNFL + RGCP maps with data augmentation and finally RNFL + RGCP maps alone without data augmentation. While input image modality and data augmentation can improve model invariance and thus robustness to changes in scale, orientation, and even image content, these methods alone are not sufficient to exploit the full generalization power of deep learning models. Training optimizations therefore go hand-in-hand with architectural enhancements discussed in the main paper.

C. Applying Cross-Validation and Data Augmentation to Hybrid DL/ML Models and End-to-End DL Models

As described in the main paper, to confirm that the significantly higher performance of the end-to-end DL models was not due to training optimizations alone (not used during original hybrid DL/ML model development [3]), we trained both types of models with cross-validation and data augmentation using the 797-image dataset from past work [3]. Even with these additional training optimizations applied to both model types, the end-to-end DL models exhibited significantly less reduction in performance than the hybrid DL/ML models upon evaluation of transfer performance on the 'Field' dataset, as shown in Table 2.

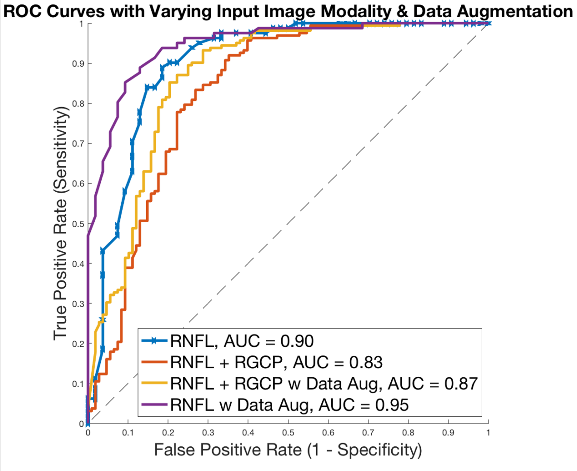

Fig. 2. ROC Curve showing impact of varying input image format and incorporating data augmentation during training. Highest (most robust) performance on 'Field' Dataset was achieved by hybrid DL/ML ResNet-18 + Random Forest model with RNFL input alone and with data augmentation.

II. ADDITIONAL CNN ENSEMBLE ARCHITECTURE

We present here an additional CNN ensemble architecture composed of all four end-to-end DL pre-trained CNNs (ResNet-18, VGG-16, InceptionV3, and DenseNet-121) followed by a fully connected classifier, which achieves accuracy slightly higher (89.6% on the 'Field' dataset and 95.4% on the 'Lab' dataset) than the CNN ensemble composed of 3 pre-trained models (ResNet-18, VGG-16, and InceptionV3) described in the main paper (88.9% on the 'Field' dataset and 94.4% on the 'Lab' dataset). A schematic of this CNN

TABLE II
ACCURACY RATES (%) FOR HYBRID DL/ML MODELS (LEFT HALF) AND FOR END-TO-END DL MODELS (RIGHT HALF) TRAINED WITH CROSS-VALIDATION AND DATA AUGMENTATION.

Hybrid DL/ML Models	Lab	Field	% Reduction	End-to-End DL Models	Lab	Field	% Reduction
Conv+FC	95.4	78.5	17.7	DenseNet-121+FC	95.9	85.2	11.2
Conv+RF	91.9	78.5	14.5	CNN Ensemble	94.4	88.9	5.83
VGG16+RF	95.4	80.7	15.4	VGG16+FC	97.0	85.2	12.2
ResNet18+RF	**95.4**	**82.2**	13.3	ResNet18+FC	94.9	83.0	12.5
InceptionV3+RF	95.4	77.8	18.4	**InceptionV3+FC**	**90.4**	**91.1**	-0.774

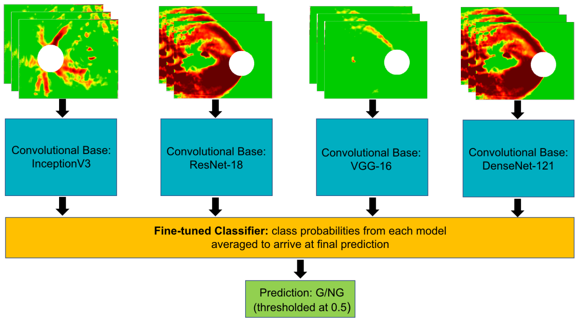
Convolutional Base: InceptionV3
Convolutional Base: ResNet-18
Convolutional Base: VGG-16
Convolutional Base: DenseNet-121
Fine-tuned Classifier: class probabilities from each model averaged to arrive at final prediction
Prediction: G/NG (thresholded at 0.5)

Fig. 3. Schematic of CNN ensemble made up of four end-to-end deep learning models (each separately fine-tuned on RNFL maps, shown as input at top) followed by dense fine-tuned layers which predict if the input image is glaucomatous (G) or not glaucomatous (NG). Predictions were averaged to arrive at the final ensemble prediction between 0 and 1, with 0.5 serving as threshold probability for binary classification.

ensemble architecture is shown in Fig. 3.

III. SENSITIVITY AND SPECIFICITY OF END-TO-END DL MODELS

In addition to accuracy percentages provided in the main paper, Table 3 shows sensitivity and specificity percentages for the end-to-end DL models, to provide a holistic characterization of these models' performance on the 'Lab' and 'Field' datasets. The reduction in sensitivity exhibited on the 'Field' dataset can partially be attributed to the fact that the 'Field' dataset is composed of a higher proportion of early-stage (difficult-to-diagnose) glaucoma patients compared to the 'Lab' dataset, which is predominately composed of later-stage glaucoma patients.

IV. COMPARING GRAD-CAM AND EYE TRACKING HEATMAPS WITH ORIGINAL FULL OCT REPORTS

We provide in Fig. 4 the Grad-CAM heatmap from Fig. 7 of the main paper, showing regions that positively contribute to CNN classification in bright red/yellow, overlayed on the original OCT full report. Directly beside that is the eye tracking heatmap from Fig. 11 of the main paper, showing expert eye fixation position and duration overlayed on the original full OCT report. Beneath each heatmap is its corresponding original OCT full report for comparison. This side-by-side view enables one to see common regions of importance both for classification by CNNs (Grad-CAM heatmap) and for diagnosis by human experts (eye tracking heatmap).

V. ALTERNATIVE SCALED TCAV SCORE METRIC

We report a scaled TCAV score, $STCAV_Q$, taking into account the magnitude of $S_{C,k,l}(x)$, suggested as an alternative metric by TCAV authors [4]. The modified equation is described in (1). It utilizes the positive mean of the directional derivatives specific to a concept of interest for the class and layer being probed. Thus, a high proportion of class images with a positive directional gradient, but with a low magnitude of conceptual sensitivity to the concept (i.e. rate of change of class prediction as a function of change in network activation in the direction of the CAV), would have a relatively lower $STCAV_Q$ (compared to $TCAV_Q$, computed as shown in (5) in main paper).

$$STCAV_Q = \frac{|x \in X_k : S_{C,k,l}(x) > 0|}{|X_k|} * \overline{S^+_{C,k,l}} \quad (1)$$

Note that $\overline{S^+_{C,k,l}}$ denotes the positive mean of the directional derivative. Scaled TCAV results indicate that the greatest influence of the RNFL and RGCP probability maps and the RGCP thickness map is in the shallower of the two dense transfer layers we probed here, $dense_1$ (Figure 5). Solid and textured colors' influence on classifications were lower relative to the non-scaled TCAV results and compared to the probability maps and RGCP thickness maps. This finding is novel when interpreting contribution of concepts toward the prediction of glaucoma, because using conceptual sensitivity as a scale reveals that the most positively contributing concepts are RNFL and RGCP probability maps at the $dense_1$ layer, while the standard TCAV score counterpart (from the main paper) reveals that all three layers ($flatten_1$, $dense_1$, and $dense_3$) contribute almost equally for the same input concepts of interest. Furthermore, consistent with TCAV results in the main paper, RNFL probability maps and RGCP probability maps and RGCP thickness maps stand out with higher STCAV scores than RNFL thickness maps.

REFERENCES

[1] K.A. Thakoor, *et al*, "Impact of Reference Standard, Data Augmentation, and OCT Input on Glaucoma Detection Accuracy by CNNs on a New Test Set." *Investigative Ophthalmology and Visual Science*, 61(7), pp.4540-4540, 2020.

[2] K.A. Thakoor, *et al.* "Strategies to Improve Convolutional Neural Network Generalizability and Reference Standards for Glaucoma Detection from OCT Scans", *Under Review*.

[3] K.A. Thakoor, *et al.* "Enhancing the Accuracy of Glaucoma Detection from OCT Probability Maps using Convolutional Neural Networks", In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2036-2040, 2019.

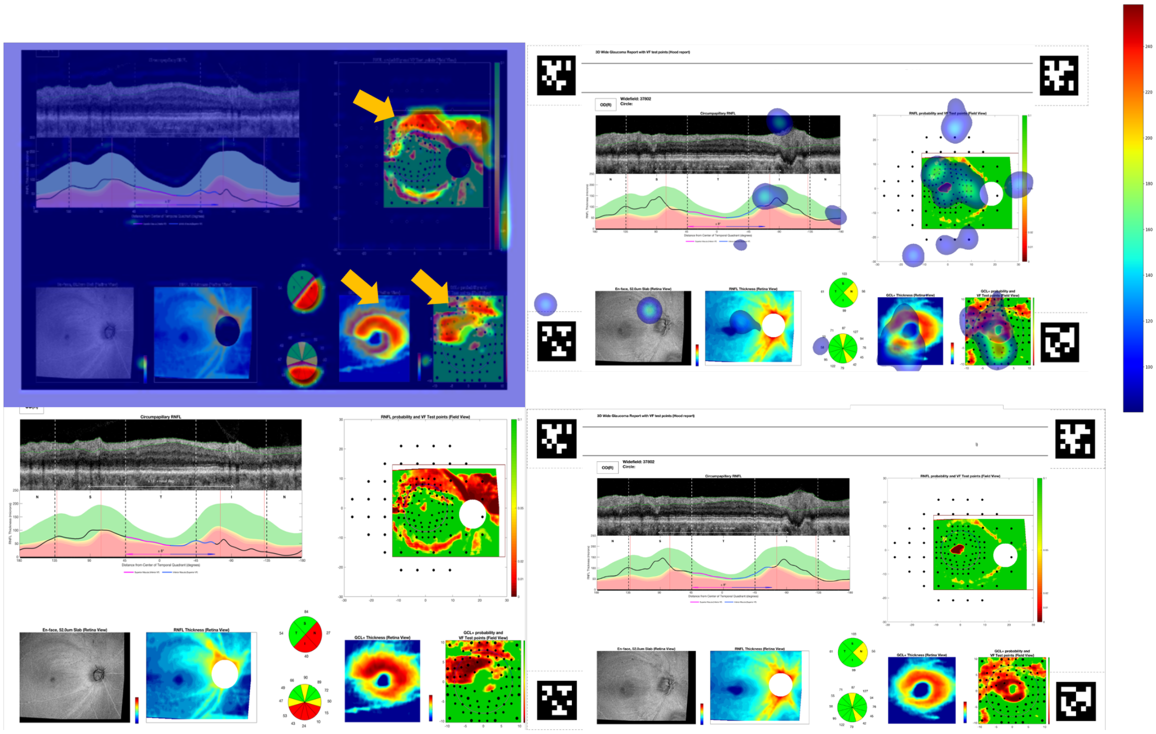

Fig. 4. CNN Grad-CAM heatmap (top left), eye tracking heatmap (top right) and corresponding original full OCT reports beneath each.

TABLE III
ACCURACY RATES, SENSITIVITY, AND SPECIFICITY FOR END-TO-END DEEP LEARNING MODELS ON LAB AND FIELD TEST SETS.

End-to-End DL Models	Lab (Accuracy, %)	Sensitivity (%)	Specificity (%)	Field (Accuracy, %)	Sensitivity (%)	Specificity (%)
DenseNet121+FC	95.9	91.1 (5 FN)	97.9 (3 FP)	85.2	68.5 (17 FN)	96.3 (3 FP)
CNN Ensemble	94.4	96.4 (2 FN)	93.6 (9 FP)	88.9	76.8 (13 FN)	98.6 (2 FP)
VGG16+FC	97.0	94.6 (3 FN)	97.9 (3 FP)	85.2	68.5 (17 FN)	96.3 (3 FP)
ResNet18+FC	94.9	96.4 (2 FN)	94.3 (8 FP)	83.0	68.5 (17 FN)	92.6 (6 FP)
InceptionV3+FC	90.4	100 (0 FN)	86.5 (19 FP)	91.1	87.0 (7 FN)	93.8 (5 FP

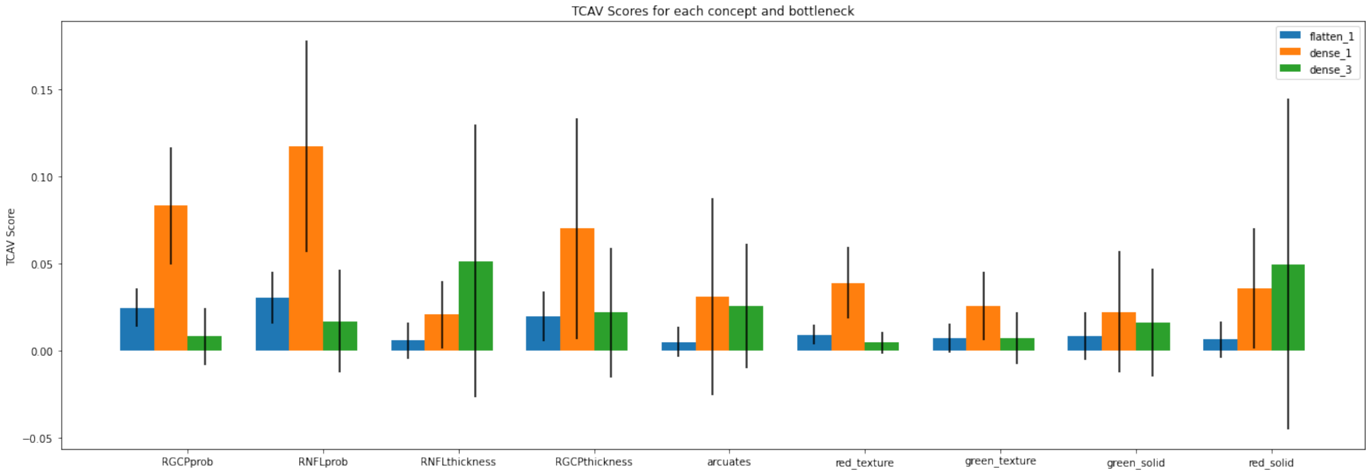

Fig. 5. Scaled TCAV results for selected, relevant OCT concepts: RNFL probability maps, RGCP probability maps, RNFL thickness maps, RGCP thickness maps, arcuates, green solid, green texture, red solid, and red texture.

[4] B. Kim, *et al*. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)”, *International Conference on Machine Learning*, 2017.