tvst

Artificial Intelligence

Deep Learning Model for Static Ocular Torsion Detection Using Synthetically Generated Fundus Images

Chen Wang[1], Yunong Bai[1], Ashley Tsang[1], Yuhan Bian[1], Yifan Gou[1], Yan X. Lin[1], Matthew Zhao[1], Tony Y. Wei[1], Jacob M. Desman[1], Casey Overby Taylor[1], Joseph L. Greenstein[1], Jorge Otero-Millan[2,3], Tin Yan Alvin Liu[4], Amir Kheradmand[2], David S. Zee[2], and Kemar E. Green[2]

[1] Johns Hopkins University Department of Biomedical Engineering, Baltimore, MD, USA
[2] Johns Hopkins University School of Medicine, Department of Neurology, Baltimore, MD, USA
[3] University of California Berkeley, Herbert Wertheim School of Optometry and Vision Science, Berkeley, CA, USA
[4] Johns Hopkins University School of Medicine, Department of Ophthalmology, Baltimore, MD, USA

Correspondence: Kemar E. Green, Johns Hopkins Hospital, 600 N. Wolfe St., Meyer 6-113, Baltimore, MD 21287, USA. e-mail: kgreen66@jhmi.edu

Received: August 12, 2022
Accepted: December 18, 2022
Published: January 11, 2023

Keywords: ocular torsion; fundus photography; deep learning; ocular motility; artificial intelligence; diplopia; dizziness; skew deviation; fourth nerve palsy

Citation: Wang C, Bai Y, Tsang A, Bian Y, Gou Y, Lin YX, Zhao M, Wei TY, Desman JM, Taylor CO, Greenstein JL, Otero-Millan J, Liu TYA, Kheradmand A, Zee DS, Green KE. Deep learning model for static ocular torsion detection using synthetically generated fundus images. Transl Vis Sci Technol. 2023;12(1):17, https://doi.org/10.1167/tvst.12.1.17

Purpose: The objective of the study is to develop deep learning models using synthetic fundus images to assess the direction (intorsion versus extorsion) and amount (physiologic versus pathologic) of static ocular torsion. Static ocular torsion assessment is an important clinical tool for classifying vertical ocular misalignment; however, current methods are time-intensive with steep learning curves for frontline providers.

Methods: We used a dataset ($n = 276$) of right eye fundus images. The disc-foveal angle was calculated using ImageJ to generate synthetic images via image rotation. Using synthetic datasets ($n = 12{,}740$ images per model) and transfer learning (the reuse of a pretrained deep learning model on a new task), we developed a binary classifier (intorsion versus extorsion) and a multiclass classifier (physiologic versus pathologic intorsion and extorsion). Model performance was evaluated on unseen synthetic and nonsynthetic data.

Results: On the synthetic dataset, the binary classifier had an accuracy and area under the receiver operating characteristic curve (AUROC) of 0.92 and 0.98, respectively, whereas the multiclass classifier had an accuracy and AUROC of 0.77 and 0.94, respectively. The binary classifier generalized well on the nonsynthetic data (accuracy = 0.94; AUROC = 1.00).

Conclusions: The direction of static ocular torsion can be detected from synthetic fundus images using deep learning methods, which is key to differentiate between vestibular misalignment (skew deviation) and ocular muscle misalignment (superior oblique palsies).

Translational Relevance: Given the robust performance of our models on real fundus images, similar strategies can be adopted for deep learning research in rare neuro-ophthalmologic diseases with limited datasets.

Introduction

Ocular torsion consists of a static and dynamic component and is defined as a rotation of the eye around the line of sight in response to head tilt in the roll (ear to shoulder) plane. This response is called ocular counter roll (OCR) and occurs under both physiologic and pathologic conditions. The response of the vestibular system to both dynamic (during) and static (after) head tilt must ensure the eyes remain aligned. The dynamic OCR is mediated by both utricular (linear acceleration receptors) and semicircular canal (angular acceleration receptors) inputs, whereas the static component is primarily driven by the utricle. The motion and gravity information from the labyrinth directly affects the tonic level of activity within the vestibular and ocular motor nuclei. An imbalance

Figure 1. Examples of DFA measurement for right eye showing intorsion and extorsion. A line is drawn manually by the examiner from the center of the optic disc (*large yellow circle*) to the center of the fovea (*small yellow circle*). Another *horizontal line* is drawn through the center of the optic disc. The angle between the two lines is the DFA.

between these nuclei can lead to (1) torsional nystagmus beating toward the side of the head tilt (dynamic OCR), and (2) a torsional position offset opposite the direction of the head tilt (static OCR).[1–10]

Pathologic static torsion results from central and peripheral utriculo-ocular pathway lesions. The pattern of pathologic static torsion distinguishes the two main causes of vertical misalignment of the eyes: "skew deviation," caused by an imbalance in the vestibulo-ocular motor pathways, and vertical strabismus from either a fourth cranial nerve (trochlear)/superior oblique palsy (SOP) or a partial third cranial nerve palsy. A skew deviation can identify vertiginous patients at risk for having a stroke.[11–18] A skew deviation of the eyes is often accompanied by a pathologic head tilt and change in torsion (OCR)[11,13,14]; the triad (i.e. skew deviation, head tilt, and pathologic OCR) comprises the ocular tilt reaction (OTR). A compensatory head tilt away from the higher (hypertropic) eye occurs with an SOP. With a skew deviation, the hypertrophic eye intorts, whereas the lower (hypotropic) eye extorts. Whereas in an SOP, the hypertropic eye extorts, whereas the hypotropic eye exhibits no pathologic torsion.[16]

Currently, there are no simple reliable bedside methods of differentiating SOP from skews in patients with an acute onset of vertigo or double vision. The Parks-Bielschowsky three-step test identifies paretic muscles (e.g. superior oblique) in vertical diplopia[19,20]; however, no torsional information is available to help distinguish between skews and SOPs. Even though subjective torsion is often assessed, there are pitfalls.[21] The supine-upright test distinguishes skews from SOP without assessing torsion,[16] but lacks sensitivity in patients who are acutely ill.[22]

There are several methods of assessing objective static torsion.[23–25] Fundus photography is most commonly used and can distinguish skews from SOPs[26] by measuring the disc-fovea angle (DFA).[24] The DFA is an inclination of the line connecting the optic nerve and foveal centers (Fig. 1).[27–30] Digital fundus photography is well suited for objective torsional assessment given its easy to use and accessible[24,26,30]; however, processing of images manually is labor-intensive, time-consuming, and prone to error.[31]

Deep learning methods may be useful in rapid and automated screening for static ocular torsion. In neuro-ophthalmology, deep learning models have been used to detect papilledema[32–34] and other optic neuropathies.[35] Successfully trained models require large datasets to avoid overfitting.[36] Access to fundus images for research has become difficult, both because pathologic fundus torsion datasets are sparse and heightened concern for patient privacy, as fundus photographs can be used for biometric identification.[37] Three possible solutions to address model training when data are scarce include (1) data augmentation (introduce more variations),[36] (2) transfer learning (enhance training by using previously learned features for a new task),[38,39] and (3) synthetic image generation (increase dataset size).[40–42] Using these strategies, two deep learning-based static torsional classifiers to differentiate the direction (intorsion versus extorsion) and amount (physiologic versus pathologic) of static ocular torsion from a small digital fundus image dataset were developed.

Methods

Data Source

Digital color fundus photographs of the right eye (n = 276) from the Johns Hopkins Hospital (JHH) collected between June 2020 and March 2022 were used for training and evaluating 2 image classifiers. All photographs were collected from routine clinic patients presenting with vestibular and ocular motor symptoms, and they were taken by the same technician using the same non-mydriatic fundus camera (Zeiss Visucam 224) with a 45 degrees field-of-view. We selected images based on the following criteria: (1) fundus photograph of the right eye; (2) clear visualization optic disc and fovea; and (3) photographs showing an intact (uncropped) retina. All images had a resolution of 1280 × 935 pixels. DFA values were measured by one author (K.E.G.) using ImageJ.[24] The DFA of each image is determined as shown in Figure 1. The images were divided by a ratio of 8:1 into model development data (n = 245) and holdout testing data (n = 31). The study was approved by the Institutional Review Board (IRB). Our IRB was approved for de-identified fundus images only, therefore patient demographics and diagnosis were not available.

Data Preprocessing

The image acquisition process produces a protruding notch at the corner of each photograph (Fig. 2). During generation of synthetic images, the notch rotates with respect to image rotation. To eliminate potential bias, we developed a novel algorithm to remove the artifactual notch before rotation. First, each image was converted to grayscale and represented as a matrix of different values (0 = black and 255 = white). A black filter was then applied to detect the margin of the retina – forming a square. Using the center and the diagonal line of the square (diameter of circle), a circular mask was then generated and overlaid on the image to remove the notch.

Data Synthesis

Static torsional data was artificially synthesized by first rotating preprocessed images (n = 245) by its measured DFA – yielding images with DFA = 0 degrees; all rotated images retained their original resolution We defined DFA >0 degrees as extorsion and DFA <0 degrees as intorsion. Two torsional datasets were synthesized: a binary (intorsion [IN] and extorsion [EX]) and a multiclass (physiologic intorsion [PHYSIIN], pathologic intorsion [PATHOIN], physiologic extorsion [PHYSIEX], and pathologic extorsion [PATHOEX]).

Previous studies suggest a mean physiologic DFA for extorsion of 7.76 degrees ± 3.63 degrees in adults.[43] To generate the 2 EX classes in the multiclass dataset, we defined physiologic and pathologic DFA ranges as 1 degrees to 7 degrees and 8 degrees to 20 degrees, respectively.[24] We assumed similar ranges for IN (i.e. physiologic = −7 degrees to −1 degrees; and pathologic =

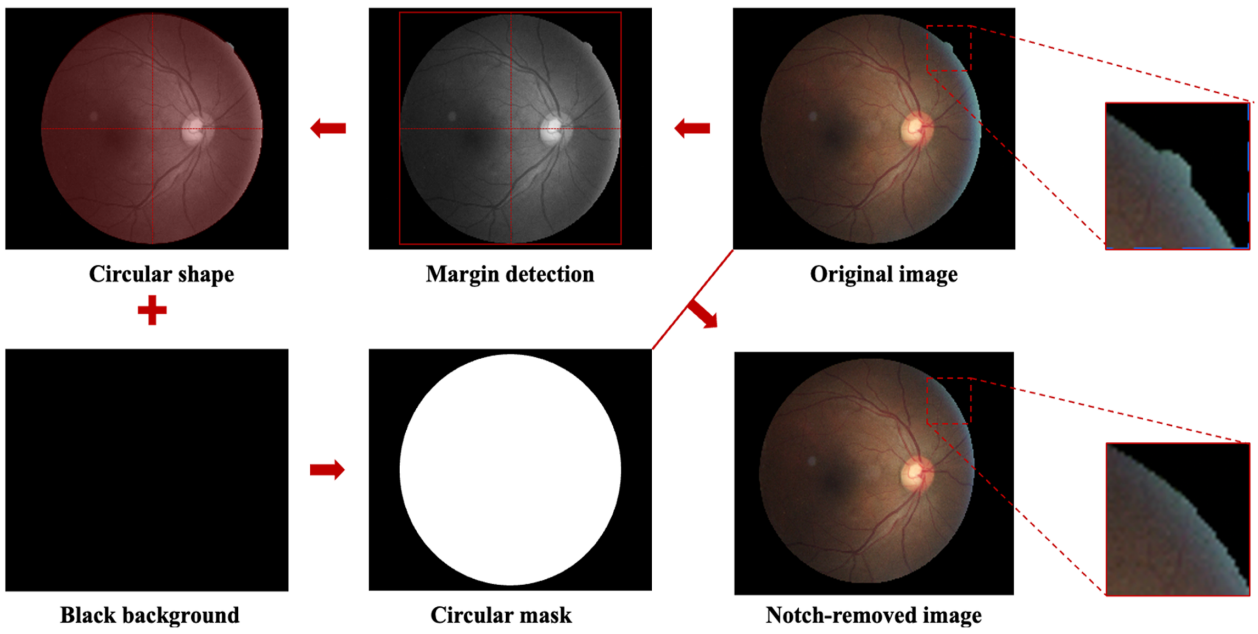

Figure 2. Overview of novel algorithm for fundus image notch removal.

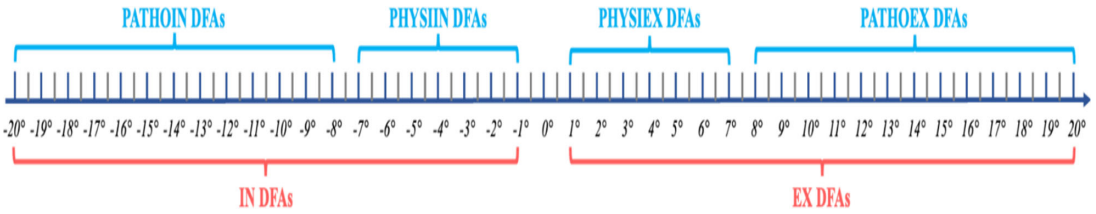

Figure 3. Predefined DFA range for each class. DFA, disc-fovea angle; IN, intorsion; EX, extorsion; PATHOEX, pathologic extorsion; PATHOIN, pathologic intorsion; PHYSIEX, physiologic extorsion; PHYSIIN, physiologic intorsion.

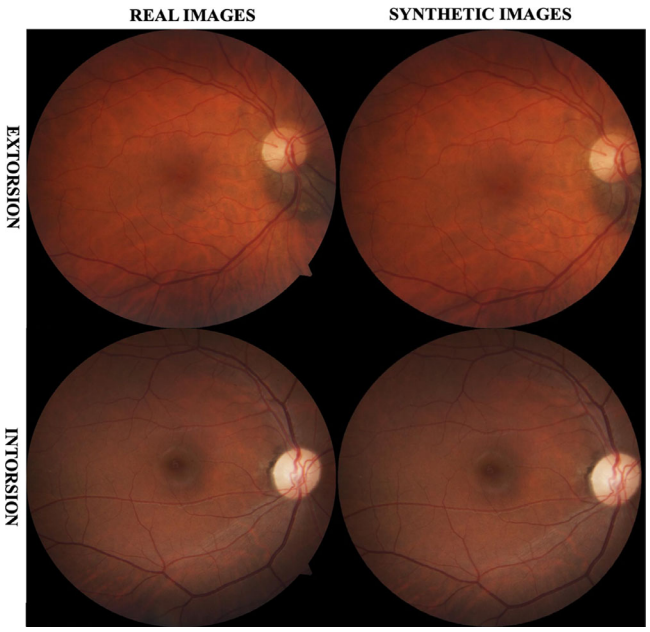

Figure 4. Comparison of real versus synthetic torsional data for right eye intorsion and extorsion examples.

−20 degrees to −8 degrees) because there are limited studies reporting DFA ranges for intorsion[44] (Fig. 3). Preprocessed images were rotated by the defined DFAs for all four classes (Figs. 4, 5). Specifically, physiologic DFAs were rotated by an increment of 0.5 degrees, and pathologic DFAs by an increment of 1 degree. This resulted in 3185 images per class. To compile the binary dataset, we assigned all images in the multiclass dataset with DFAs <0 degrees to the IN class ($n = 6370$), and those with DFAs >0 degrees to the EX class ($n = 6370$). Both datasets were divided into training, validation, and testing sets by a ratio of 5:1:1.

Model Architecture

ResNet has been reported as a state-of-the-art image classification model that resolves gradient vanishing and overfitting problems.[45] Two classification models were developed using the ResNet architecture and adoption of transfer learning.[38,39] We initially developed a binary classifier (model 1) using the binary dataset (intorsion versus extorsion). The architecture incorporates the ResNet50 model. This is a 50 layers deep convolutional neural network (CNN) with ImageNet (a large dataset of 1000 generic object classes and about 1.2 million color images) pretrained model weights loaded from the Keras library.[45] We removed the last fully connected layers containing 1000 neurons (corresponding to ImageNet object classes) from the original model. Two additional layers with 128 and 2 neurons, respectively, were stacked to the modified network – each neuron corresponding to the 2 distinct classes (i.e. IN and EX).

As the model was pretrained using ImageNet, preceding layers only extract universal features (e.g. edges and curves). To avoid overfitting and reduce training time, all but the last two layers were frozen; model weights at frozen layers were not updated during the training process. Softmax activation[46] was used in the last layer to normalize the values into a probability distribution over predicted output classes, as shown in Equation 1 (where y represents the input vector from fully connected layer, $exp()$ the standard exponential function, and n the number of classes).

$$softmax(y)_i = \frac{exp(y_i)}{\sum_j^n exp(y_i)} \quad (1)$$

The multiclass classifier (model 2) used the multiclass dataset to classify images into physiologic and pathologic DFA ranges for both intorsion and extortion. The architecture was identical to model 1's except for the output layer where the number of neurons was increased to 4 – corresponding to the 4 output classes (PATHOIN, PHYSIIN, PHYSIEX, and PATHOEX).

Model Training

Model 1 was trained for 20 epochs (the number of complete iterations the algorithm makes through the training dataset) on 9100 images and validated on 1820 images (see Fig. 5). Model 2 was trained

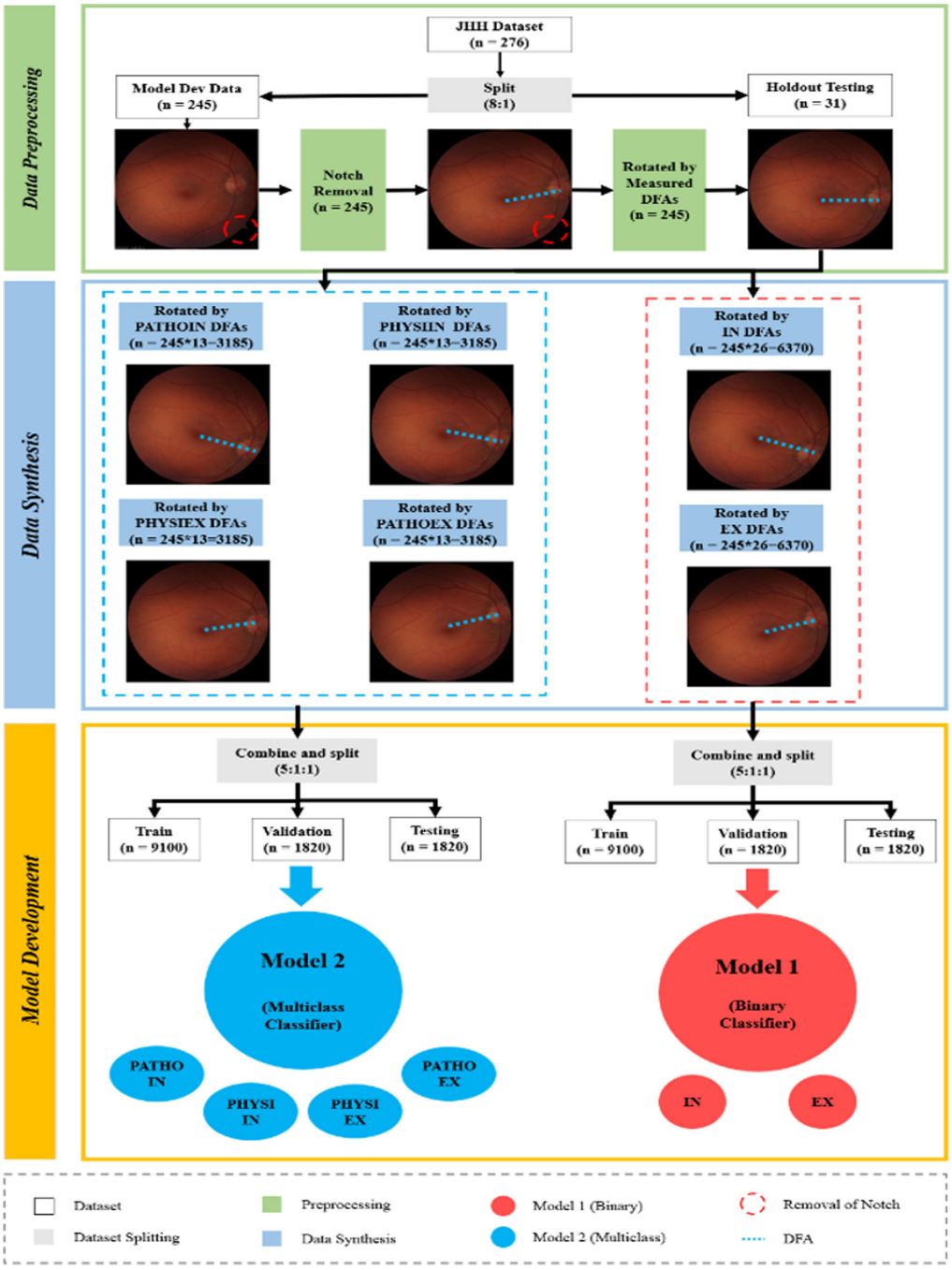

Figure 5. Pipeline for data preparation and model development. *Preprocessing stage*: The JHH dataset was split into a holdout testing and model development datasets. Each photograph had the notch removed and DFA set at 0 degrees using rotation. *Data synthesis stage*: Synthetic torsion photographs were generated using different predefined DFA ranges. *Model development stage:* The dataset was further divided for the training, validation, and testing of both binary and multiclass classifiers. JHH, Johns Hopkins Hospital; DFA, disc-fovea angle; IN, intorsion; EX, extorsion; PATHOEX, pathologic extorsion; PATHOIN, pathologic intorsion; PHYSIEX, physiologic extorsion; PHYSIIN, physiologic intorsion; model 1, binary classifier; model 2, multiclass classifier.

and validated using 2275 images per class (total = 9100) and 455 images per class (total = 1820), respectively. The categorical cross-entropy loss function (CE Loss)[46] was used for both models to quantify the difference between probability distributions of predicted probabilities and ground-truth labels, as explained in Equation 2 (where n, y_i and $\hat{y}_i$ represents the number of classes, corresponding true label [0 or 1 for the

current class], and probability for the current class from the model output).

$$CE\ Loss\ =\ -\sum_{i=1}^{n} y_i log\hat{y}_i \quad (2)$$

We adopted Adam optimization,[37] an extended version of the stochastic gradient descent algorithm with better computational efficiency, for model weights decay with the default learning rate of 0.001.

Model Evaluation

Both models were evaluated on synthetic testing data (n = 1820) with equal class representations. We then externally validated the models on the holdout testing set (n = 31: PATHOEX = 10; PATHOIN = 2; PHYSIEX = 11; and PHYSIIN = 8). The class associated with the maximum probability after the softmax activation layer was defined as the predicted class. Predicted results were then compared with ground-truth class labels. We calculated the model performance metrics, including the overall classification accuracy, precision, sensitivity, specificity, and F1 score. All the overall values except for accuracy were calculated based on the macro-average metric for each class (i.e. the sum of class-specific values divided by the number of classes). The overall accuracy was calculated as the number of correctly classified images divided by the total number of images. The receiver operating characteristic curves for each class were plotted along with the corresponding area under the curve (AUC) values.

Gradient-Weighted Class Activation Mapping

To interpret the deep-learning model and better understand its predictions, we generated heatmaps to show important regions at different convolutional layers for each image according to gradient-weighted class activation mapping (Grad-CAM).[44] Grad-CAM uses the gradients of a certain target class to generate heatmaps highlighting important regions for the predicted class. The heatmaps are then resized and overlaid on the original image; warmer colors represent regions with the greatest contribution to a class prediction.

Results

JHH Dataset

In the JHH dataset, DFAs ranged from −25.2 degrees to 19.8 degrees (Fig. 6A), with a mean and median of 5.29 degrees and 5.30 degrees, respectively. Extorsion represented 89.1% (n = 246), whereas only 3.62% (n = 10) of the images were intorsion; the remaining 7.25% (n = 20) had no measurable torsion (DFA = 0 degrees). For the holdout testing test (n = 31), the DFA ranged from −11 degrees to 20 degrees

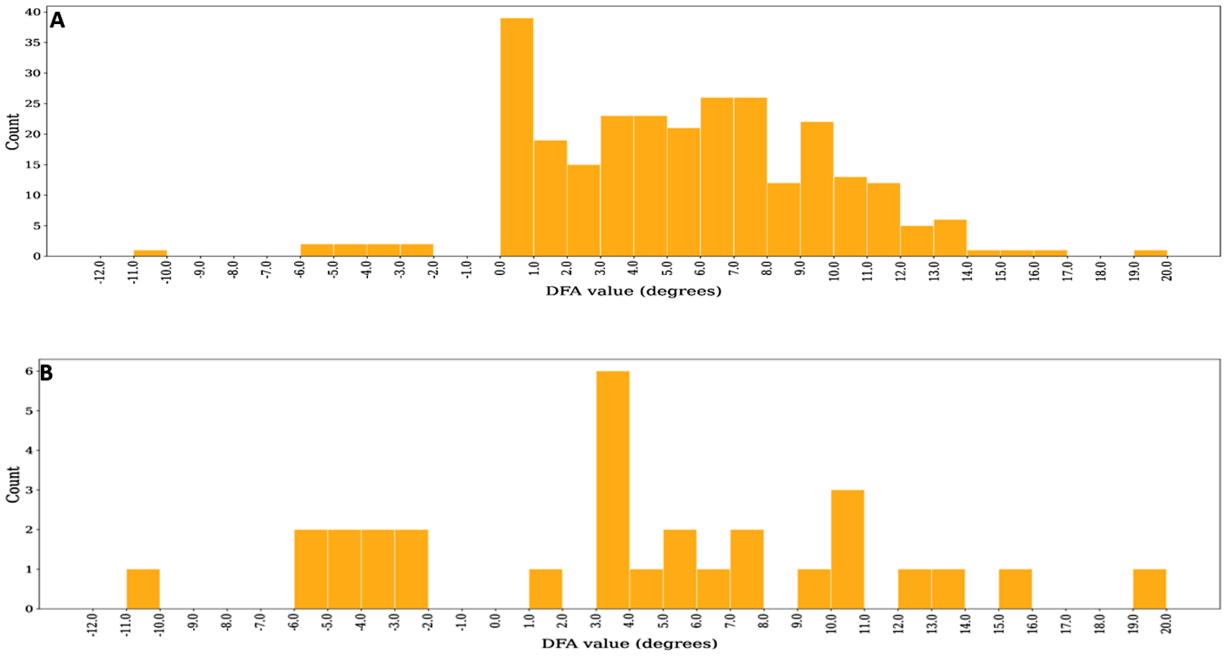

Figure 6. DFA Distribution of (**A**) JHH Dataset (excluding DFA = −25.3 degrees) and (**B**) holdout testing set. DFA, disc-fovea angle; JHH, Johns Hopkins Hospital.

Table. Classification Performance of Both Models on The Different Testing Datasets (Synthetic and Holdout). AUROC: Area Under the Receiver Operating Characteristic Curve

Model 1 (Binary Classifier): Tested on Synthetic Testing Set

Class	Sensitivity	Specificity	Precision	F1 Score	AUROC	No. of Images
EX	0.92	0.97	0.93	0.92	0.98	910
IN	0.93	0.96	0.92	0.92	0.98	910
Overall	0.93	0.97	0.93	0.92	0.98	1820
Overall accuracy	0.92					

Model 1 (Binary Classifier): Tested on Holdout Testing Set

Class	Sensitivity	Specificity	Precision	F1 Score	AUROC	No. of images
EX	0.90	1.00	1.00	0.95	1.00	21
IN	1.00	0.94	0.83	0.91	1.00	10
Overall	0.95	0.97	0.92	0.93	1.00	31
Overall accuracy	0.94					

Model 2 (Multiclass Classifier): Tested on Synthetic Testing Set

Class	Sensitivity	Specificity	Precision	F1 Score	AUROC	No. of images
PATHOEX	0.74	0.99	0.94	0.83	0.98	455
PATHOIN	0.79	0.98	0.90	0.84	0.97	455
PHYSIEX	0.73	0.91	0.67	0.70	0.91	455
PHYSIIN	0.82	0.90	0.65	0.72	0.92	455
Overall	0.77	0.95	0.79	0.77	0.94	1820
Overall accuracy	0.77					

Model 2 (Multiclass Classifier): Tested on Holdout Testing Set

Class	Sensitivity	Specificity	Precision	F1 Score	AUROC	No. of images
PATHOEX	0.70	0.82	0.54	0.61	0.86	10
PATHOIN	1.00	0.74	0.15	0.27	0.79	2
PHYSIEX	0.09	0.91	0.33	0.14	0.70	11
PHYSIIN	0.00	0.92	0.00	0.00	0.18	8
Overall	0.45	0.85	0.26	0.26	0.65	31
Overall accuracy	0.32					

(Fig. 6B). The intorsion class ($n = 10$) represented 32.2% of the total holdout testing set (PATHOIN = 2 and PHYSIIN = 8), whereas extorsion ($n = 21$) accounted for the remaining 78.8% (PATHOEX = 10; PHYSIEX = 11).

Model Performance

Key performance metrics are summarized in the Table and Figures 7 and 8. Model 1 achieved excellent classification performance with balanced sensitivity and specificity on the synthetic testing set and comparable performance on holdout testing set, demonstrating generalizability. Model 2 achieved high specificities and area under the receiver operating characteristic curve (AUROC; 0.94) but relatively lower sensitivities on all classes when tested on the synthetic dataset. Low sensitivity and precision were observed when we tested Model 2 on the holdout data, indicating poor generalizability.

As shown in Figure 9, both models demonstrated high classification accuracy at large DFAs. Lower classification accuracies were observed at DFAs close to its adjacent classes (i.e. DFA between 1 degree and −1 degree for model 1, and between −8 degrees and 8 degrees for model 2), indicating relatively weak classification performance at smaller DFAs.

Class Activation Mapping

Class activation mapping analysis showed a gradual shift in the activation loci with increased convolutional layer depth (Fig. 10). The first few convolutional

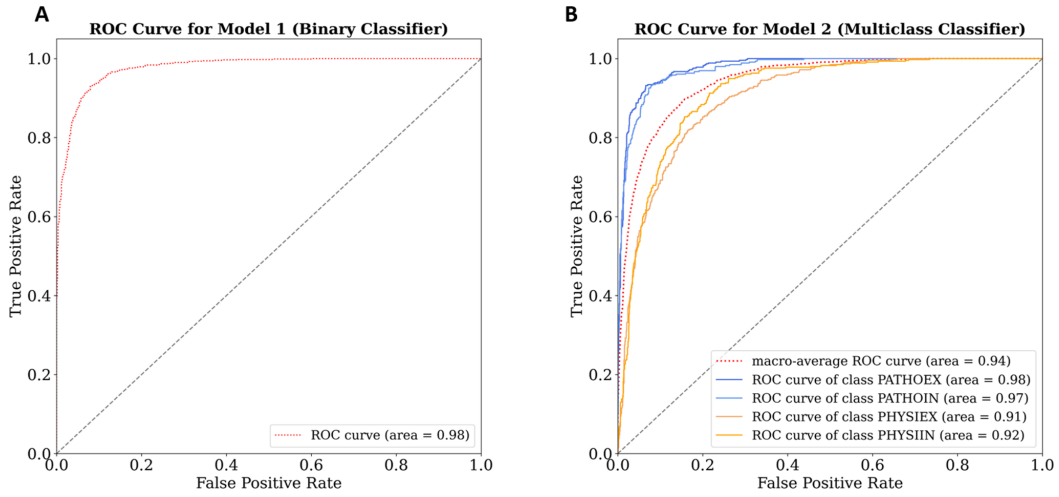

Figure 7. Receiver operating characteristic (ROC) curves showing the classification performance for (**A**) model 1, and (**B**) model 2 on the synthetic testing set. IN, intorsion; EX, extorsion; PATHOEX, pathologic extorsion; PATHOIN, pathologic intorsion; PHYSIEX, physiologic extorsion; PHYSIIN, physiologic intorsion.

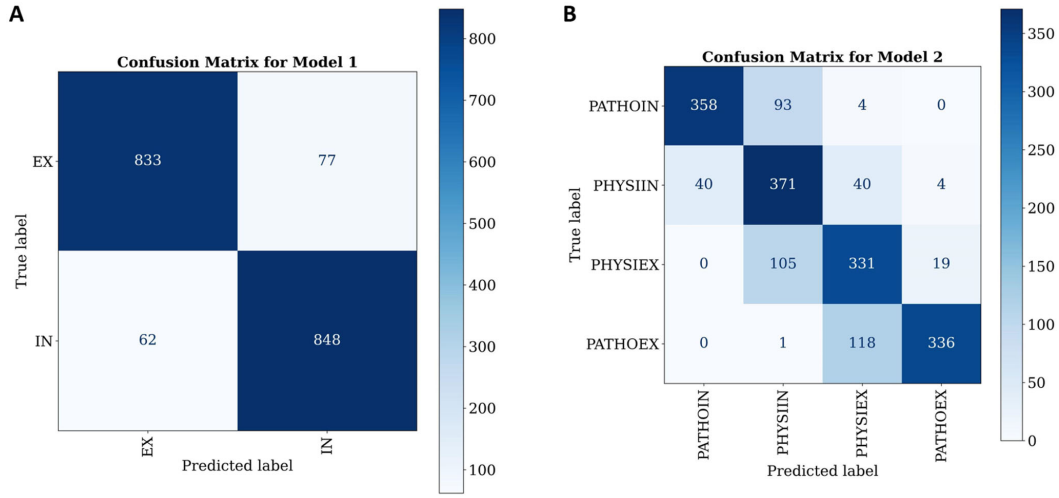

Figure 8. Confusion matrices showing the classification results for (**A**) model 1, and (**B**) model 2 on synthetic testing sets. IN, intorsion; EX, extorsion; PATHOEX, pathologic extorsion; PATHOIN, pathologic intorsion; PHYSIEX, physiologic extorsion; PHYSIIN, physiologic intorsion.

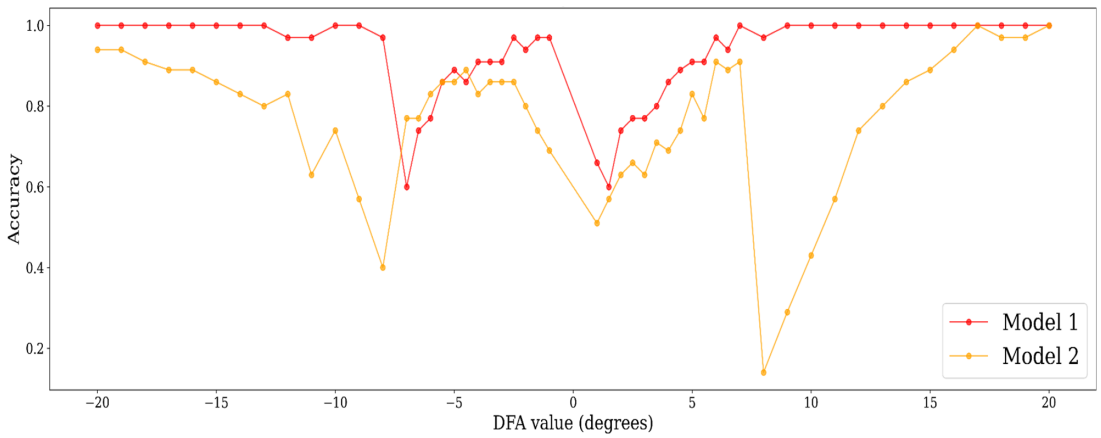

Figure 9. Classification accuracy of both models at different DFAs when tested on synthetic data. DFA, disc-fovea angle.

translational vision science & technology

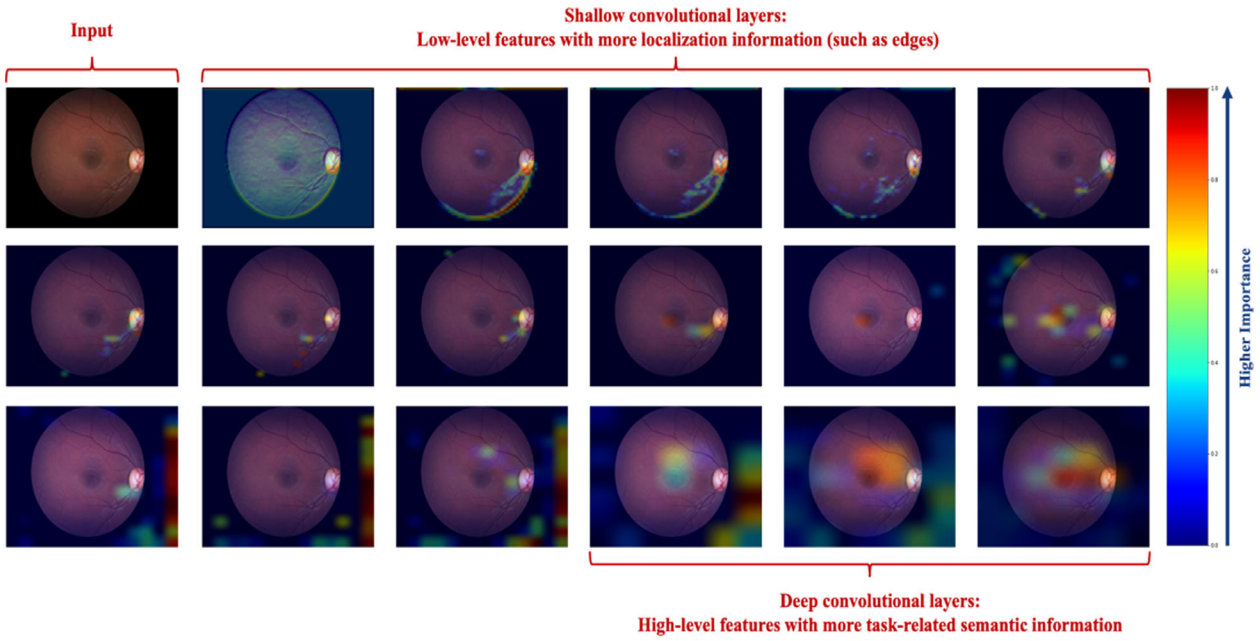

Figure 10. Original image (*top left*) and class activation mappings at different convolutional layers (from shallow to deep convolutional layers) for an example image labeled as physiologic intorsion. Shallow layers showing low-level feature importance such as edges, and deeper convolutional layers showing high-level feature importance (e.g. fovea and optic nerve).

layers capture local features (e.g. edges and repetitive patterns) followed by large blood vessels in the next few layers. Ensuing layers showed activation in the optic disc and the fovea. The final layers had simultaneous activation in the optic disc, fovea, and retinal region in between. This is analogous to the attention of human experts while assessing DFAs in fundus photographs.

Analysis Model 2's Misclassification

Analysis of model 2's misclassifications (false positives and negatives), as shown in Figure 11, reveals that the model failed mostly when images had DFAs close to the upper and lower limits of the physiologic and pathologic ranges for all four classes (normal or pathological, intorsion, or extorsion).

Discussion

Evaluation of torsional eye movements is clinically valuable to differentiate between central and peripheral pathologies that can affect alignment such as in patients with skew deviation or SOP. In the clinical setting, fundus imaging is a reliable method for objective measurement of ocular torsion using the macula and optic disk as the major anatomical landmarks. Currently, such evaluation must be performed manually on the fundus photograph and automated methods are not available to improve screening for pathologies that can affect ocular alignment in each eye. Such automated methods can be clinically valuable to improve diagnosis and aid clinicians who are not trained in their specialty to detect ocular motor abnormalities. Similar approaches with using machine learning have been shown to be valuable clinically in the field of neuro-ophthalmology.[32–35]

Here, we developed a binary classifier (differentiates intorsion from extorsion), and a multiclass classifier (characterizes torsion in pathologic and physiologic DFA ranges) with deep learning using only synthetic images generated from a small dataset ($n = 245$). In our study, model 1 produced robust, reproducible results (see the Table) that would make it suitable for clinical practice as a screening tool for distinguishing skew deviations from fourth nerve palsies. It can also help localize lesions to the vestibulo-ocular pathways in patients with acute vertigo without clear skew deviation but possessing other features of an OTR (partial OTR). Our multiclass classifier had a lower sensitivity and poor generalization compared to the binary classifier. Given the high false positive rate, model 2 might be less helpful in screening patients with acute vertigo and vertical diplopia.

Studies quantifying the DFA in vertical strabismus have mainly involved patients with superior oblique palsies, and the physiologic and pathologic ranges often

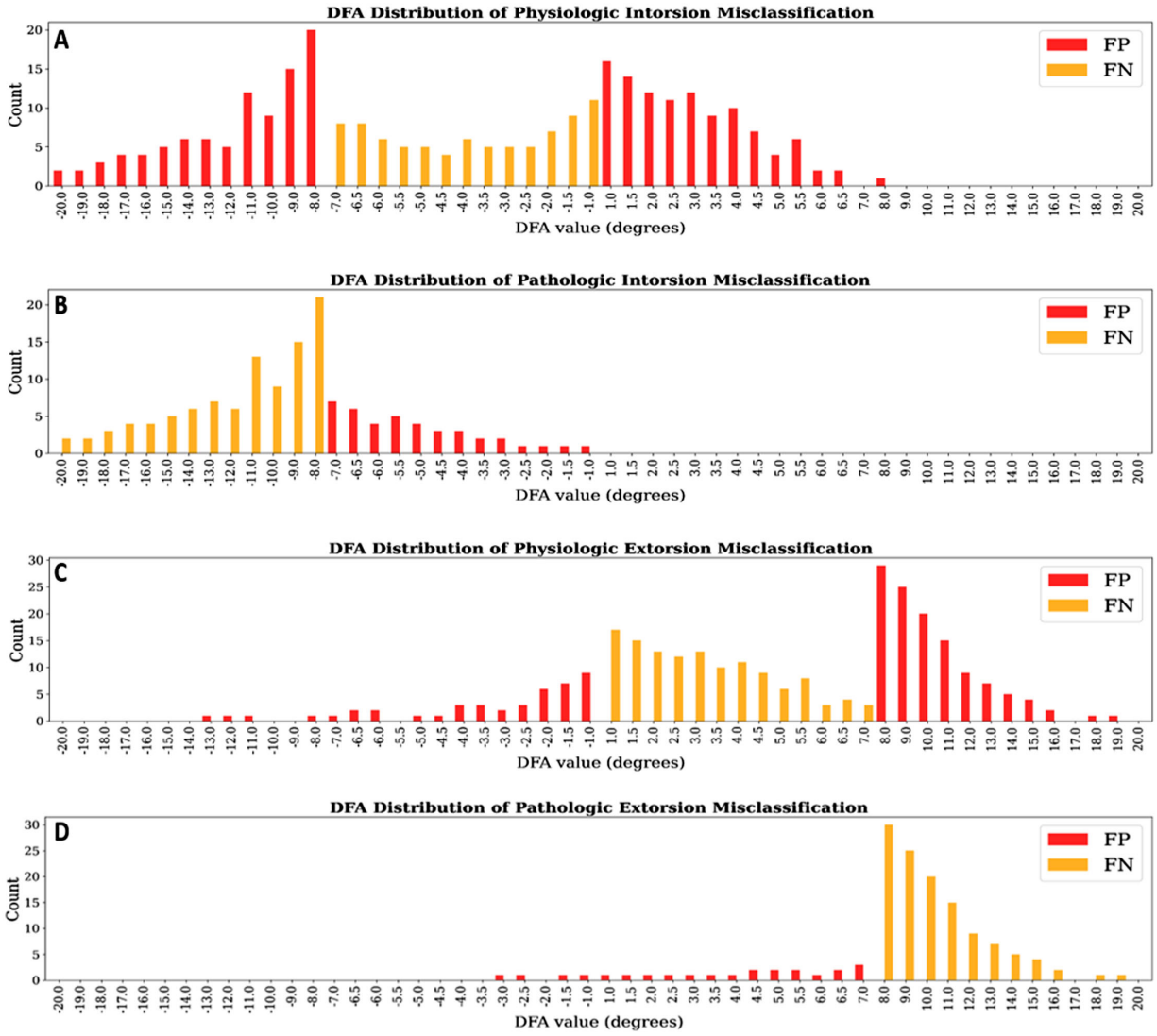

Figure 11. Model 2's misclassification DFA distributions for (**A**) physiologic intorsion, (**B**) pathologic intorsion, (**C**) physiologic extorsion, and (**D**) pathologic extorsion. DFA, disc-fovea angle; FP, false positive; FN, false negative.

overlapped.[24,47,48] Additionally, changes in vergence, and the position of the eye relative to head, known to influence the degree of torsion,[49] were not accounted for during fundus photography. Tight control of the position of the eye in the orbit during fundus photography might create a better distinction between physiologic and pathologic ranges of the DFA. Furthermore, data on the ranges of the DFA in patients with abnormal static torsion have not been correlated with the degree of vertical misalignments. Therefore, differentiating physiologic from pathologic static torsion using current ranges of DFA is not useful clinically for distinguishing skews from SOPs in the setting of acute vertigo or vertical diplopia. More real-world data are needed to establish better ranges of what is normal for machine learning diagnosis.

When training small datasets, artificial data synthesis and transfer learning are necessary. With advances in generative adversarial network (GAN)-based models in artificial intelligence (AI) research, synthetic data is increasingly used to overcome the scarcity of annotated medical datasets.[40–42,50] For the detection of static ocular torsion from fundus photographs, the use of synthetic images seems appropriate given the limited datasets available containing the desired pathologies (i.e. SOP and skew deviation). Furthermore, there is evidence that synthetic images augment the performance of medical image analysis using machine learning methods.[40]

For this study, the models were only required to detect the position of the optic disc relative to the fovea. Therefore, only basic image processing techniques

(image rotation) were used to generate the synthetic data, and no other fundus heterogeneity was added. The synthetic data and image processing technique was reviewed by the authors to accurately mimic real fundus images and torsion from human subjects. Therefore, given the robust performance of model 1 (see the Table, Figs. 7, 8) and its generalizability on nonsynthetic data, we conclude that the generated photographs were comparable to real fundus torsion (see Fig. 4). Future synthetic fundus datasets for studying other retinal and optic nerve pathologies will require more fundus heterogeneity, and thus GAN-based synthetic images might be a next step to test models that detect torsion.[51,52]

The training of CNNs involves parallel computation and a massive number of floating-point operations, such as matrix and vector operations.[46] Such computing patterns are suitable for graphics processing units (GPUs). As such, GPUs are more preferred than central processing units (CPU) for the training of CNN.[53] Because fundus photographs contain biometric features, we conducted all the experiments in an internal computing platform which does not support GPU nodes to maintain data integrity. Therefore, to accelerate the training process, transfer learning was used to tune the model weights in the last two layers. Transfer learning increases the efficiency of training models by only training network weights in a few selected layers.[38] This technique facilitates model training within a reasonable timeframe (approximately 400 seconds per training epoch in our case) by using less computational power – making it an efficient and effective approach.

The model is considered "robust" if it generalizes well on external datasets. Generalizability refers to the model's capacity at replicating results on unseen data. In most cases, the training and real-world data are different, and not identically distributed. This leads to a distribution shift problem, which often causes the poor generalizability of machine/deep learning models.[54] In our study, we created a holdout testing set to evaluate the generalizability of our model (trained on synthetic data) on real static ocular torsional data. When tested on the holdout testing sets, model 1 generalized well to real data, whereas model 2 did not (see the Table).

The multiclass classifier (model 2) did not generalize well, probably for several reasons. First, our holdout testing set only contained 31 images, with only 10 images from the intorsion classes (PATHOIN = 2 and PHYSIIN = 8). A larger and less skewed holdout testing set with more examples from each class would be better. Second, we used relatively simple image processing techniques (image rotation and notch removal) to generate synthetic images. Unlike synthetic data generated using generative models, in which several features of the source images are generated,[41,42,51,52] only artificial ocular torsion was introduced to our dataset. Therefore, there could still be flaws in the data synthesis pipeline causing small differences between synthetic and real data, even though the synthetic and real images seemed identical (see Fig. 4); making it difficult to accurately explain the predictions by the model.

Deep learning is a powerful tool for image classification, yet knowing "why" it makes its predictions is often unknown. We applied Grad-CAM to our model to better understand its predictions with the idea that deeper convolutional layers carry more deterministic spatial information for model prediction.[44] The Grad-CAM output from our model (see Fig. 10) demonstrated that the predictions are based on the spatial information of the optic disc, the fovea, and the area in between. This indicates that the model's spatial focus aligns with clinical expectations, suggesting its predictions are trustworthy. One caveat, however, is that the class activation maps were occasionally less reliable. In addition, our method of generating synthetic images may have biased the model into detecting only changes in fundus torsion because most of the other fundus features were relatively homogenous. As such, future work should introduce other differences in the synthetic images (e.g. hemorrhage, disc edema, retinal ischemia, lens opacities, etc.) to determine whether other factors influence the ability of the model to detect torsion.

Limitations

Since skew deviation and SOP fundus datasets are rare, we generated synthetic fundus torsional images to train our models. Although our model performed well on synthetic datasets, it has some limitations. First, model 2 does not generalize well on real data; classification was less accurate with the holdout testing set. Second, we did not address torsion for the left eye or both eyes; however, others have successfully developed models that distinguished between images from the left and right eyes.[35,39,55] Automated screening for skews and fourth nerve palsies will be most useful when the fundi of both eyes are assessed and compared. Third, our holdout testing set was relatively small and not balanced for all classes (more extorsions than intorsions). Finally, we have not verified the model performance using other physiologic and pathologic datasets from other institutions. Differences among datasets, such as how images were acquired, the resolution of images, and the characteristics of patients, might affect the generalizability of the model.

Conclusion

With data synthesis and transfer learning, different types and degrees of ocular torsion can be detected from fundus photographs using deep learning. Our model has promising clinical applicability, although some limitations still exist. In the future, model performance can be further improved when greater and more diverse datasets become available for training and evaluation. Future models can be adopted to (1) aid in the automated diagnosis of acute vertigo or vertical diplopia without many modifications, and (2) monitor treatment responses in neuro-ophthalmic, strabismic, and neuro-vestibular diseases.

Acknowledgments

Disclosure: **C. Wang**, None; **Y. Bai**, None; **A. Tsang**, None; **Y. Bian**, None; **Y. Gou**, None; **Y.X. Lin**, None; **M. Zhao**, None; **T.Y. Wei**, None; **J.M. Desman**, None; **C.O. Taylor**, None; **J.L. Greenstein**, None; **J. Otero-Millan**, None; **T.Y.A. Liu**, None; **A. Kheradmand**, None; **D.S. Zee**, None; **K.E. Green**, None

References

1. Brandt Th, Dieterich M. Cyclorotation of the Eyes and Subjective Visual Vertical in Vestibular Brain Stem Lesions. *Ann NY Acad Sci*. 1992;656(1 Sensing and C):537–549.
2. Diamond SG, Markham CH. Ocular counterrolling as an indicator of vestibular otolith function. *Neurology*. 1983;33(11):1460–1460.
3. Kingma H, Stegeman P, Vogels R. Ocular torsion induced by static and dynamic visual stimulation and static whole body roll. *Europ Archives Oto-Rhino-Laryngol*. 1997;254(S1):S61–S63.
4. Leigh RJ, Zee DS. *The Neurology of Eye Movements*. 5th edition. Cary, NC: Oxford University Press; 2015.
5. Raps EC, Solomon D, Galetta SL, Liu GT, Volpe NJ. Cyclodeviation in Skew Deviation. *Am J Ophthalmol*. 1994;118(4):509–514.
6. Sadeghpour S, Fornasari F, Otero-Millan J, Carey JP, Zee DS, Kheradmand A. Evaluation of the Video Ocular Counter-Roll (vOCR) as a New Clinical Test of Otolith Function in Peripheral Vestibulopathy. *JAMA Otolaryngol Head Neck Surg*. 2021;147(6):518.
7. Schmid-Priscoveanu A, Böhmer DS.A Vestibulo-Ocular Responses During Static Head Roll and Three-Dimensional Head Impulses After Vestibular Neuritis. *Acta Oto-Laryngologica*. 1999;119(7):750–757.
8. Schworm HD, Ygge J, Pansell T, Lennerstrand G. Assessment of Ocular Counterroll during Head Tilt Using Binocular Video Oculography. *Investig Ophthalmol Visual Sci*. 2002;43(3):662–667.
9. Zingler VC, Kryvoshey D, Schneider E, Glasauer S, Brandt T, Strupp M. A clinical test of otolith function: static ocular counterroll with passive head tilt: *NeuroReport*. 2006;17(6):611–615.
10. Dieterich M, Brandt T. Perception of Verticality and Vestibular Disorders of Balance and Falls. *Front Neurol*. 2019;10:172.
11. Green KE, Gold DR. HINTS Examination in Acute Vestibular Neuritis: Do Not Look Too Hard for the Skew. *J Neuro-Ophthalmol*. 2021;41(4):e672–e678.
12. Brandt T, Dieterich M. Skew deviation with ocular torsion: A vestibular brainstem sign of topographic diagnostic value. *Ann Neurol*. 1993;33(5):528–534.
13. Brodsky MC, Donahue SP, Vaphiades M, Brandt T. Skew deviation revisited. *Surv Ophthalmol*. 2006;51(2):105–128.
14. Halmagyi GM, Gresty MA, Gibson WPR. Ocular tilt reaction with peripheral vestibular lesion. *Annals Neurol*. 1979;6(1):80–83.
15. Hotson JR, Baloh RW. Acute Vestibular Syndrome. *New England J Med*. 1998;339(10):680–685.
16. Wong AMF. Understanding skew deviation and a new clinical test to differentiate it from trochlear nerve palsy. *J Am Assoc Pediatric Ophthalmol Strabismus*. 2010;14(1):61–67.
17. Gold DR, Shin RK, Galetta S. Pearls and Oysters: Central fourth nerve palsies. *Neurology*. 2012;79(23):e193–e196.
18. Shah M, Primiani CT, Kheradmand A, Green KE. Pearls & Oy-sters: Vertical Diplopia and Ocular Torsion: Peripheral vs Central Localization. *Neurology*. Published online June 6, 2022, https://doi.org/10.1212/WNL.0000000000200835.
19. Bielschowsky A. Lectures on motor anomalies of the eyes: II. Paralysis of individual eye muscles. *Archives Ophthalmol*. 1935;13(1):33.
20. Bielschowsky A. Disturbances of the vertical motor muscles of the eyes. *Archives Ophthalmol*. 1938;20(2):175–200.
21. Yoo HS, Park E, Rhiu S, et al. A computerized red glass test for quantifying diplopia. *BMC Ophthalmol*. 2017;17(1):71.
22. Lemos J, Subei A, Sousa M, et al. Differentiating Acute and Subacute Vertical Strabismus Using Different Head Positions During the Upright-Supine Test. *JAMA Ophthalmol*. 2018;136(4):322.

23. Versino M, Newman-Toker DE. Blind spot heterotopia by automated static perimetry to assess static ocular torsion: centro-cecal axis rotation in normals. *J Neurol*. 2010;257(2):291–293.
24. Kang H, Lee SJ, Shin HJ, Lee AG. Measuring ocular torsion and its variations using different non-mydriatic fundus photographic methods. Madigan M, ed. *PLoS One*. 2020;15(12):e0244230.
25. Ehrt O, Boergen KP. Scanning laser ophthalmoscope fundus cyclometry in near-natural viewing conditions. *Graefe's Arch Clin Exp Ophthalmol*. 2001;239(9):678–682.
26. Lemos J, Eggenberger E. Clinical utility and assessment of cyclodeviation. *Curr Opinion Ophthalmol*. 2013;24(6):558–565.
27. Jethani J, Seethapathy G, Purohit J, Shah D. Measuring normal ocular torsion and its variation by fundus photography in children between 5-15 years of age. *Indian J Ophthalmol*. 2010;58(5):417–419.
28. Guyton DL. Ocular torsion: Sensorimotor principles. *Graefe's Archive Clin Experim Ophthalmol*. 1988;226(3):241–245.
29. Guyton DL. Ocular Torsion Reveals the Mechanisms of Cyclovertical Strabismus The Weisenfeld Lecture. *Investig Opthalmol Visual Sci*. 2008;49(3):847.
30. Le Jeune C, Chebli F, Leon L, et al. Reliability and reproducibility of disc-foveal angle measurements by non-mydriatic fundus photography. Andley UP, ed. *PLoS One*. 2018;13(1):e0191007.
31. Fleming C. Screening for Primary Open-Angle Glaucoma in the Primary Care Setting: An Update for the US Preventive Services Task Force. *Ann Family Med*. 2005;3(2):167–170.
32. Biousse V, Newman NJ, Najjar RP, et al. Optic Disc Classification by Deep Learning versus Expert Neuro-Ophthalmologists. *Annals Neurol*. 2020;88(4):785–795.
33. Vasseneix C, Najjar RP, Xu X, et al. Accuracy of a Deep Learning System for Classification of Papilledema Severity on Ocular Fundus Photographs. *Neurology*. 2021;97(4):e369–e377.
34. Milea D, Najjar RP, Jiang Z, et al. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *New England J Med*. 2020;382(18):1687–1695.
35. Liu H, Li L, Wormstone IM, et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA Ophthalmol*. 2019;137(12):1353.
36. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *J Big Data*. 2019;6(1):60.
37. Akram MU, Abdul Salam A, Khawaja SG, Naqvi SGH, Khan SA. RIDB: A Dataset of fundus images for retina based person identification. *Data Brief*. 2020;33:106433.
38. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3(1):9.
39. Liu TYA, Ting DSW, Yi PH, et al. Deep Learning and Transfer Learning for Optic Disc Laterality Detection: Implications for Machine Learning in Neuro-Ophthalmology. *J Neuro-Ophthalmol*. 2020;40(2):178–184.
40. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Engin*. 2021;5(6):493–497.
41. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification. Published online January 8, 2018, https://doi.org/10.48550/ARXIV.1801.02385.
42. Torfi A, Fox EA, Reddy CK. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences*. 2022;586:485–500.
43. Jonas RA, Wang YX, Yang H, et al. Optic Disc - Fovea Angle: The Beijing Eye Study 2011. Frishman L, ed. *PLoS One*. 2015;10(11):e0141771.
44. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336–359.
45. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Published online December 10, 2015. ARXIV preprint, https://doi.org/10.48550/ARXIV.1512.03385.
46. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2017.
47. Korda A, Zamaro E, Wagner F, et al. Acute vestibular syndrome: is skew deviation a central sign? *J Neurol*. 2022;269(3):1396–1403.
48. Cherchi M. Utricular function in vestibular neuritis: a pilot study of concordance/discordance between ocular vestibular evoked myogenic potentials and ocular cycloposition. *Exp Brain Res*. 2019;237(6):1531–1538.
49. Porrill J, Ivins JP, Frisby JP. The variation of torsion with vergence and elevation. *Vision Res*. 1999;39(23):3934–3950.
50. Pakhomov D, Hira S, Wagle N, Green KE, Navab N. Segmentation in Style: Unsupervised Semantic Image Segmentation with Stylegan and CLIP. *arXiv:210712518 [cs]*. Published online July 2021. Accessed October 20, 2021, http://arxiv.org/abs/2107.12518.

51. Burlina PM, Joshi N, Pacheco KD, Liu TYA, Bressler NM. Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration. *JAMA Ophthalmol*. 2019;137(3): 258–264.
52. Guo J, Pang Z, Yang F, Shen J, Zhang J. Study on the Method of Fundus Image Generation Based on Improved GAN. *Mathemat Problems Engin*. 2020;2020:1–13.
53. Li X, Zhang G, Huang HH, Wang Z, Zheng W. Performance Analysis of GPU-Based Convolutional Neural Networks. In: *2016 45th International Conference on Parallel Processing (ICPP)*. IEEE; 2016:67–76. Available at https://www2.seas.gwu.edu/~howie/publications/GPU-CNN-ICPP16.pdf.
54. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain Generalization: A Survey. Published online March 3, 2021; last revised August 12, 2022, https://doi.org/10.48550/ARXIV.2103.02503.
55. Liu C, Han X, Li Z, et al. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. Paranhos A, ed. *PLoS One*. 2019;14(9): e0222025.