# Deep Learning Model for Static Ocular Torsion Detection Using Synthetic 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]; Kemar E. Green[2]**

**Affiliations:**

[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 School of Optometry, Berkeley, CA, USA
[4]Johns Hopkins University School of Medicine, Department of Ophthalmology, Baltimore, MD, USA

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

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

## 1. Abstract

**Purpose:** The objective of the study is to develop deep learning models using synthetic fundus images to assess the direction (intorsion vs. extorsion) and amount (physiologic vs. pathologic) of static ocular torsion. Static ocular torsion assessment is an important clinical tool for identifying abnormalities in the vestibular-ocular-motor pathway, but current methods are time-intensive with steep learning curves. Advanced deep learning techniques are promising strategies to detect ocular torsion rapidly and accurately and can be applied to distinguish vestibular causes of vertical misalignment from cranial nerve palsies.

**Methods:** We curated a dataset (n=276) of right eye fundus images from the Johns Hopkins Hospital. The disc-foveal angle was calculated and used to generate synthetic images using image rotation. Using the synthetic datasets and transfer learning, we developed a binary classifier (intorsion vs. extorsion) and a multiclass classifier (physiologic vs. pathologic intorsion and extorsion). The models performance (accuracy, sensitivity, specificity, precision, F1 score, and the ROC curve) were evaluated on synthetic and non-synthetic testing sets.

**Results:** On the synthetic dataset, the binary classifier had an accuracy and AUROC of 0.92 and 0.98 respectively, while the multiclass classifier had an accuracy and AUROC of 0.77 and 0.94 respectively. The binary classifier generalized well on the holdout testing set (accuracy=0.94; AUROC=1.00); however, the multiclass classifier did not (accuracy = 0.32; AUROC = 0.65).

**Conclusion:** Static ocular torsion can be detected from synthetic fundus images using deep learning methods. Future models to differentiate vestibular and ocular misalignment from cranial nerve palsies can be more accurately adopted and tested in the future.

## 2. Introduction

Ocular torsion, which consists of a static and dynamic component, 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 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 andchange 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 hypertropic eye intorts while the lower (hypotropic) eye extorts. Whereas in an SOP, the hypertropic eye extorts, while 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 torsional [23–28]. Fundus photography is most commonly used, and can distinguish skews from SOPs[29] by measuring the disc-fovea angle (DFA)[30]. The DFA is an inclination of the line connecting the optic nerve and foveal centers **(Fig. 1)**[31–33]. Digital fundus photography is well suited for objective torsional assessment given its easy to use and accessible 24[26]; however, processing of images manually is labor-intensive, time-consuming, and prone to error[34].

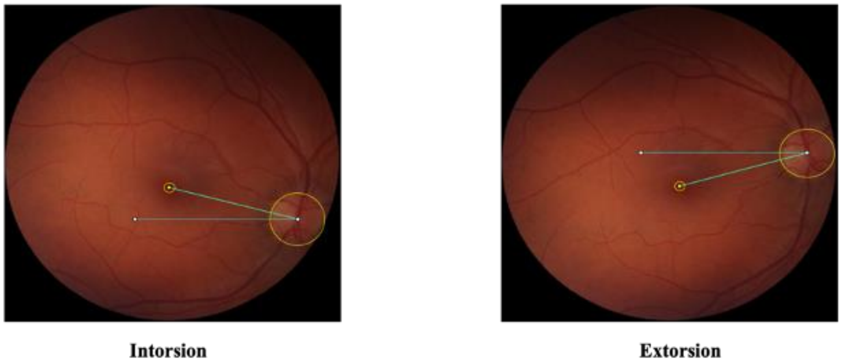

**Fig. 1.** Examples of DFA measurement for right eye showing intorsion and extorsion. A line is drawn manually by the examiner from the center of the optic disc (large yellow circle) to the center of the fovea (small yellow circle). Another horizontal line is drawn through the center of the optic disc.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

Deep learning methods may be useful in rapid and automated screening for static ocular torsion. In neuro-ophthalmology, deep learning models have been successfully used to detect papilledema[35–37] and other optic neuropathies[38]. Successfully trained models require large datasets to avoid overfitting[39]. Access to fundus images for research has become difficult, both due to sparsity of pathologic fundus torsion datasets and heightened concern for patient privacy, as fundus photographs can be used for biometric identification[40]. Three possible solutions to address model training when data is scarce include: 1) data augmentation (introduce more variations)[39], 2) transfer learning (enhance training by transferring already learned features)[41,42], and 3) synthetic image generation (increase dataset size)[43–45]. Using these strategies, we developed two deep learning-based static torsional classifiers to differentiate the direction (intorsion vs. extorsion) and amount (physiologic vs. pathologic) of static ocular torsion from a small digital fundus image dataset.

## 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 two image classifiers. All images were taken by the same technician using the same non-mydriatic fundus camera (Zeiss Visucam 224) with a 45° field-of-view. Only images that showed a clear optic disc, and fovea were used. DFA values were measured by one author (K.E.G.) using ImageJ [26]. The DFA of each image is determined as shown in **Fig. 1**. The images were divided using 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).

### *Data Preprocessing*

The image acquisition process produces a protruding notch at the corner of each photo **(Fig. 2)**. During generation of synthetic images, the notch rotates with respect to the degree of 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.

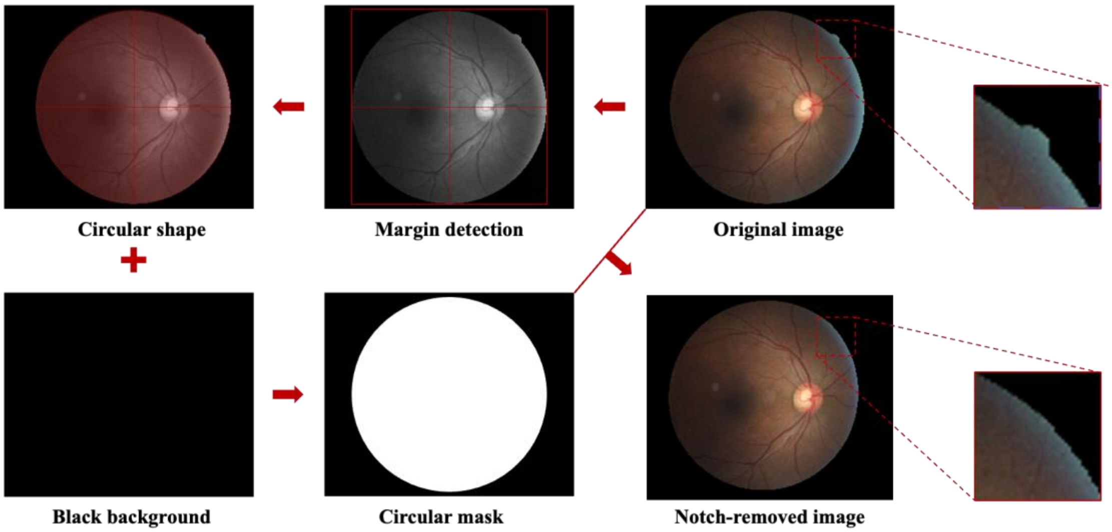

**Fig. 2.** Overview of novel algorithm for fundus image notch removal.

### *Data Synthesis*

Static torsional data was artificially synthesized by first rotating preprocessed images (n=245) by its measured DFA – yielding images with DFA= 0°. We defined DFA ≥ 0° as extorsion and DFA ≤ 0° 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 ± 3.63° in adults[27]. To generate the two *EX* classes in the multiclass dataset, we defined physiologic and pathologic DFA ranges as 1° to 7° and 8° to 20°, respectively[26]. We assumed similar ranges for *IN* (i.e., physiologic: -7° to -1°; and pathologic: -20° to -8°) since there are limited studies reporting DFA ranges for intorsion[46] **(Fig. 3)**. Preprocessed images were rotated by the defined DFAs for all four classes **(Fig. 4-5)**. Specifically, physiologic DFAs were rotated by an increment of 0.5°, and pathologic DFAs by an increment of 1°. This resulted in n=3185 images per class. To compile the binary dataset, all images in the multiclass dataset with DFAs < 0° were assigned to the *IN* class (n=6370), and those with DFAs > 0° to the *EX* class (n=6370). Both datasets were divided into training, validation and testing sets by a ratio of 5:1:1.

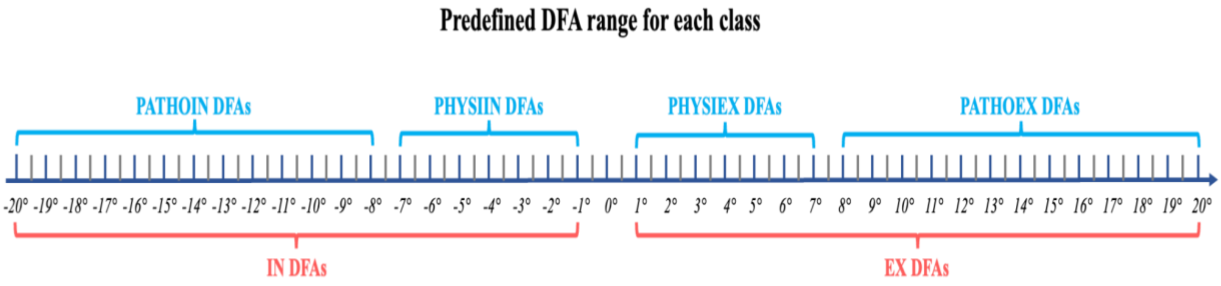

**Fig. 3.** Predefined DFA range for each class. DFA: Disc-Fovea Angle; *IN*: Intorsion; *EX*: Extorsion; *PATHOEX*: Pathologic Extorsion; *PATHOIN*: Pathologic Intorsion; *PHYSIEX*: Physiologic Extorsion; *PHYSIIN*: Physiologic Intorsion.

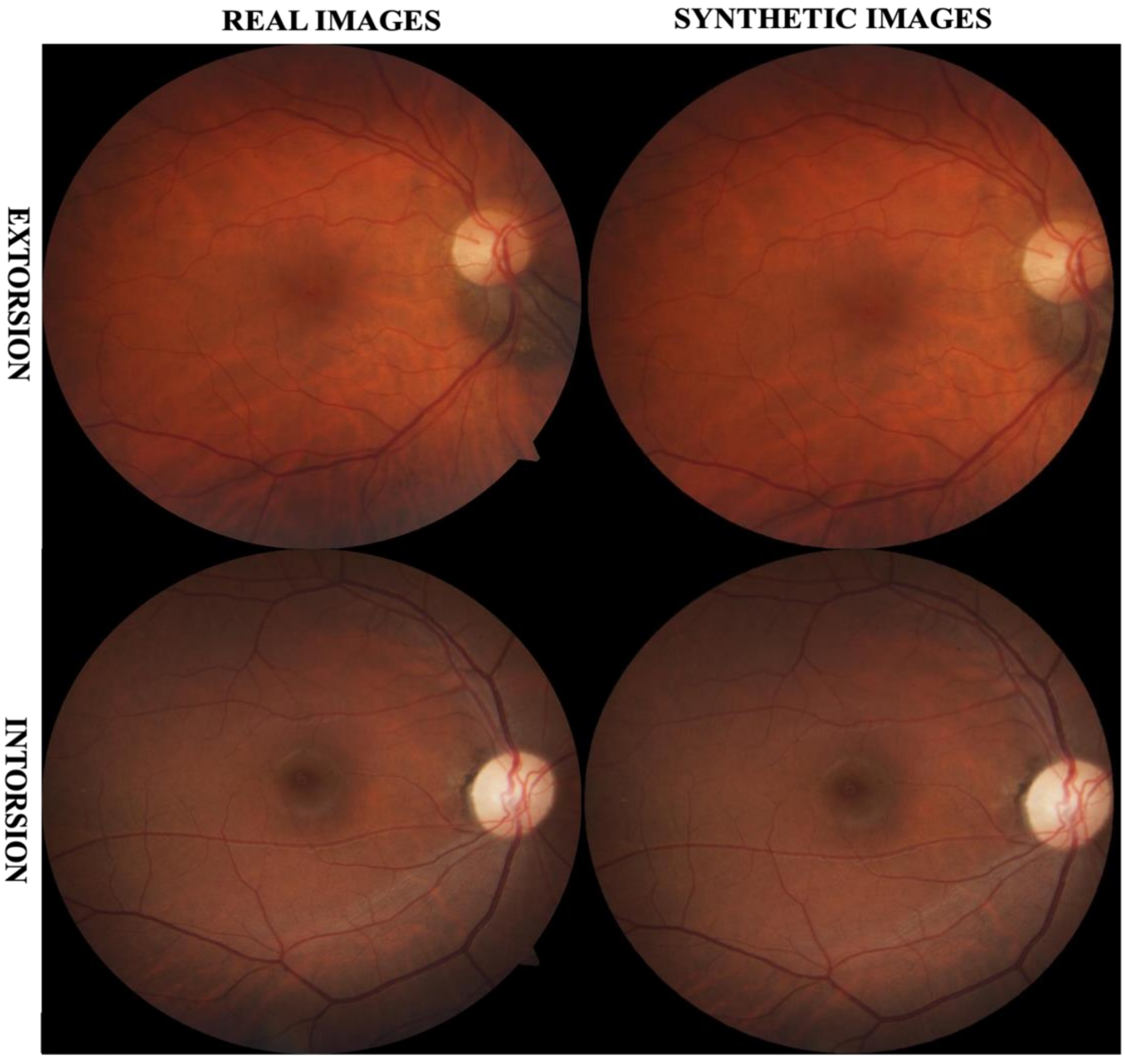

**Fig. 4**. Comparison of real versus synthetic torsional data for right eye intorsion and extorsion examples.

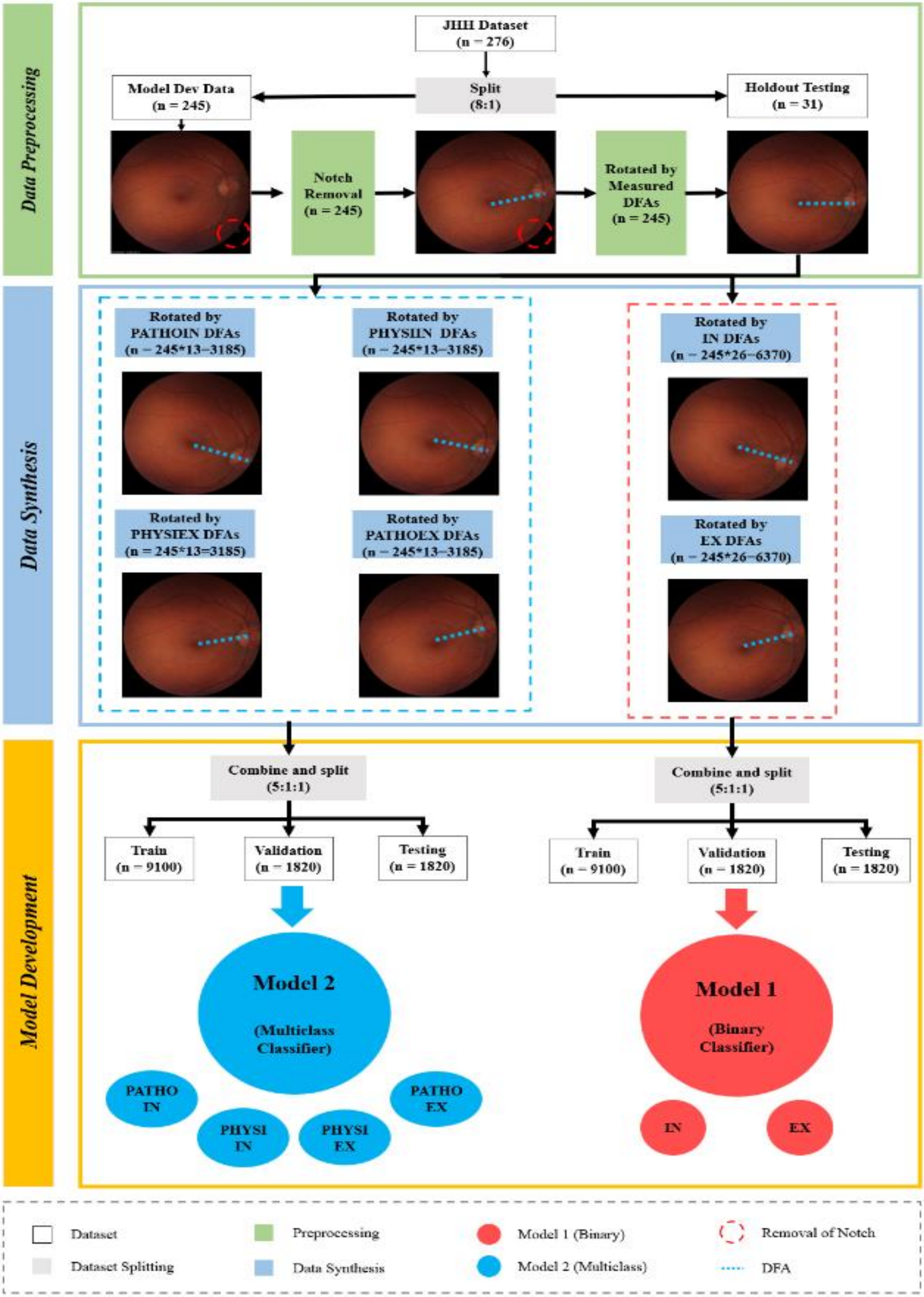
Data Preprocessing
JHH Dataset
(n = 276)
Model Dev Data
(n = 245)
Split
(8:1)
Holdout Testing
(n = 31)
Notch
Removal
(n = 245)
Rotated by
Measured
DFAs
(n = 245)
Data Synthesis
Rotated by
PATHOIN DFAs
(n = 245*13=3185)
Rotated by
PHYSIIN DFAs
(n = 245*13=3185)
Rotated by
IN DFAs
(n = 245*26=6370)
Rotated by
PHYSIEX DFAs
(n = 245*13=3185)
Rotated by
PATHOEX DFAs
(n = 245*13=3185)
Rotated by
EX DFAs
(n = 245*26=6370)
Model Development
Combine and split
(5:1:1)
Train
(n = 9100)
Validation
(n = 1820)
Testing
(n = 1820)
Combine and split
(5:1:1)
Train
(n = 9100)
Validation
(n = 1820)
Testing
(n = 1820)
Model 2
(Multiclass
Classifier)
PATHO
IN
PHYSI
IN
PHYSI
EX
PATHO
EX
Model 1
(Binary
Classifier)
IN
EX
Dataset
Dataset Splitting
Preprocessing
Data Synthesis
Model 1 (Binary)
Model 2 (Multiclass)
Removal of Notch
DFA

**Fig. 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° 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.

### *Model Architecture*

ResNet has been reported as a state-of-the-art image classification model that resolves gradient vanishing and overfitting problems[47]. Two classification models were developed using the ResNet architecture and adoption of transfer learning[41,42]. We initially developed a binary classifier (Model 1) using the binary dataset (intorsion vs. 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) pre-trained model weights loaded from the Keras library[47]. The last fully connected layers containing 1000 neurons (corresponding to ImageNet object classes) were removed. Two additional layers with 128 and 2 neurons, respectively, were stacked to the modified network; each neuron corresponding to the two distinct classes (i.e., *IN* and *EX*).

As the model was pre-trained 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[48] 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).

$$\boldsymbol{softmax(y)_i = \frac{exp(y_i)}{\sum_j^n \quad exp(y_i)}} \qquad \textbf{Eq. 1}$$

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 n=9100 images and validated on n=1820 images (**Fig. 5)**. Model 2 was trained and validated using n=2275 images per class (total = 9100) and 455 images per class (total = 1820), respectively. The categorical cross-entropy loss function (***CE Loss)***[48] 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 $\boldsymbol{n}, \boldsymbol{y_i}$ and $\boldsymbol{\widehat{y}_i}$ 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).

$$\boldsymbol{CE\ Loss} = -\sum_{i=1}^{n} \quad \boldsymbol{y_i log \widehat{y}_i} \qquad \textbf{Eq. 2}$$

Adam optimization[37], an extended version of the stochastic gradient descent algorithm with better computational efficiency, was adopted 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; 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 (Grad-CAM)*

To interpret the deep learning model and better understand its predictions, heatmaps were generated to show important regions at different convolutional layers for each image according to Grad-CAM[49]. 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° to 19.8° **(Fig. 6)**, with a mean and median of 5.29° and 5.30°, respectively. Extorsion represented 89.1% (n=246), while only 3.62% (n=10) of the images were intorsion; the remaining 7.25% (n=20) had no measurable torsion (DFA = 0°).

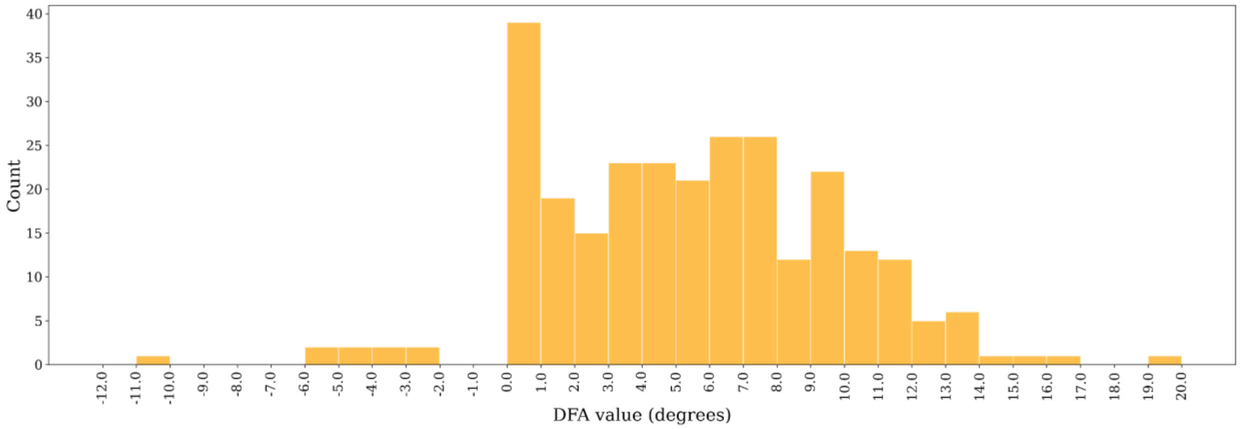

**Fig. 6.** DFA Distribution of JHH Dataset (excluding DFA = -25.3°). DFA: Disc-Fovea Angle; JHH: Johns Hopkins Hospital.

### *Model performance*

Key performance metrics are summarized in **Table 1** and **Fig. 7-8**. 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 AUROC (0.94) but relatively lower sensitivities on all classes when tested on the synthetic dataset. Low sensitivity and precision were observed when Model 2 was tested on the holdout data, indicating poor generalizability.

As shown in **Fig. 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° and -1° for Model 1, and between -8° and 8° for Model 2), indicating relatively weak classification performance at smaller DFAs.

| **Model 1 (Binary Classifier): Tested on synthetic testing set** | | | | | | |
|---|---|---|---|---|---|---|
| **Class** | **Sensitivity** | **Specificity** | **Precision** | **F1 Score** | **AUROC** | **# 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** | **# 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** | **# 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** | **# 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 | | | | | |

**Table 1**. Classification performance of both models on the different testing datasets (synthetic and holdout). AUROC: Area Under the Receiver Operating Characteristic curve.

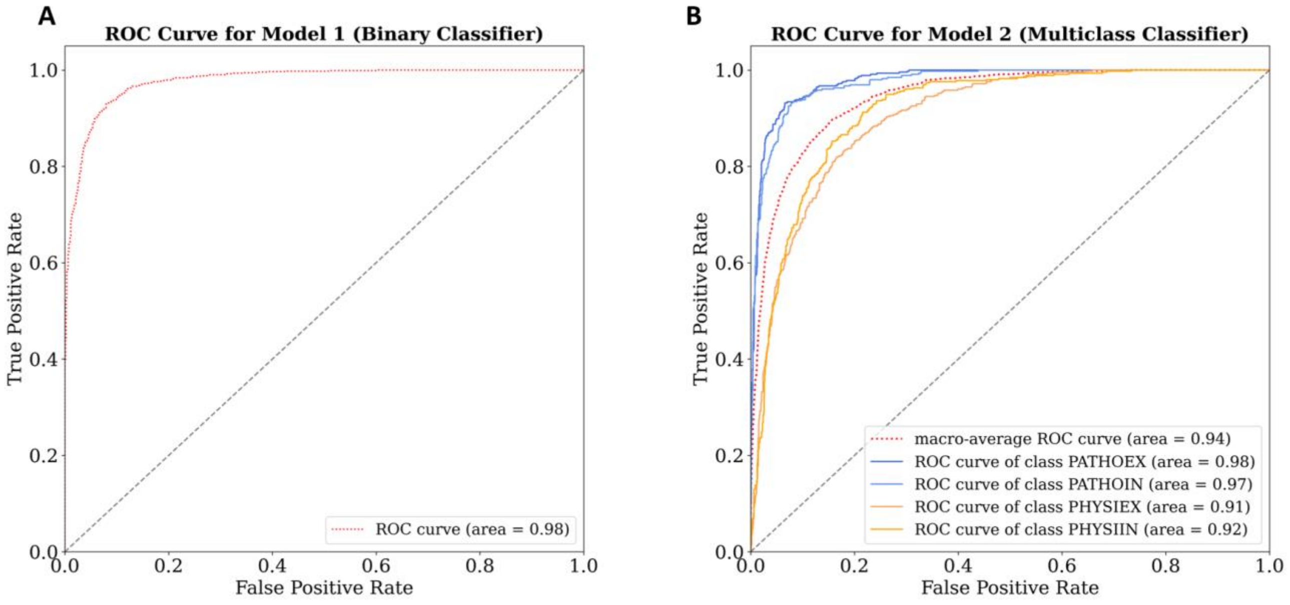

**Fig 7**. ROC curves showing the classification performance for A) Model 1, and B) Model 2 on the synthetic testing set. *IN*: Intorsion; *EX*: Extorsion; *PATHOEX*: Pathologic Extorsion; *PATHOIN*: Pathologic Intorsion; *PHYSIEX*: Physiologic Extorsion; *PHYSIIN*: Physiologic Intorsion; ROC: Receiver Operating Characteristic.

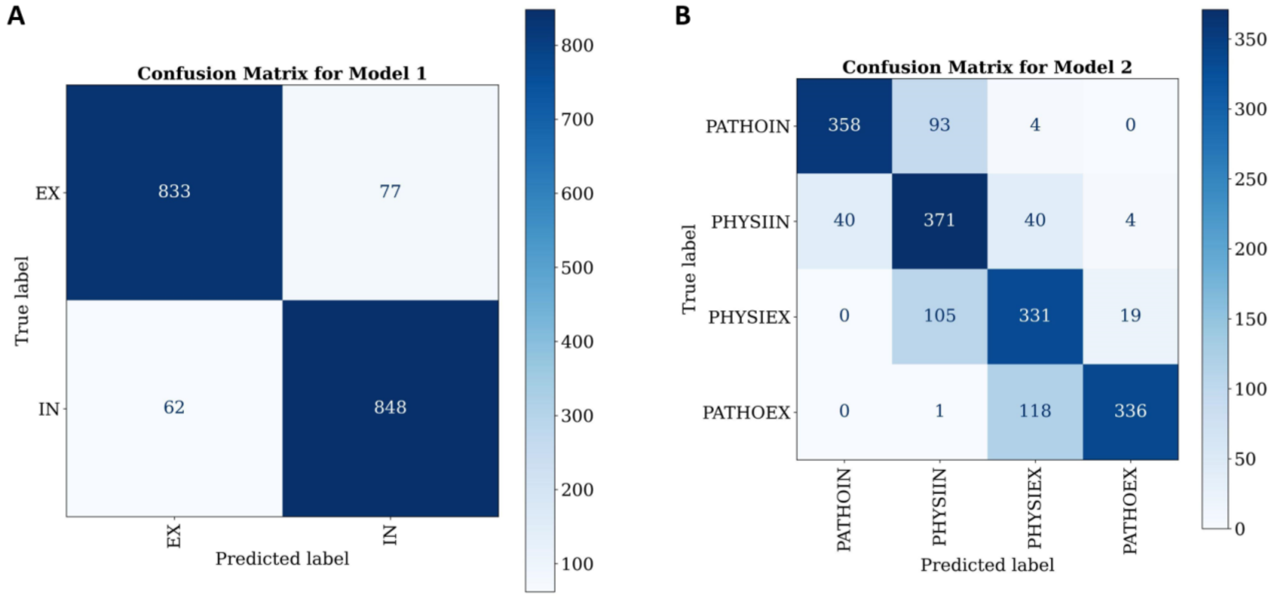

**Fig 8**. Confusion matrices showing the classification results for A) Model 1, and B) Model 2 on synthetic testing sets. *IN*: Intorsion; *EX*: Extorsion; *PATHOEX*: Pathologic Extorsion; *PATHOIN*: Pathologic Intorsion; *PHYSIEX*: Physiologic Extorsion; *PHYSIIN*: Physiologic Intorsion.

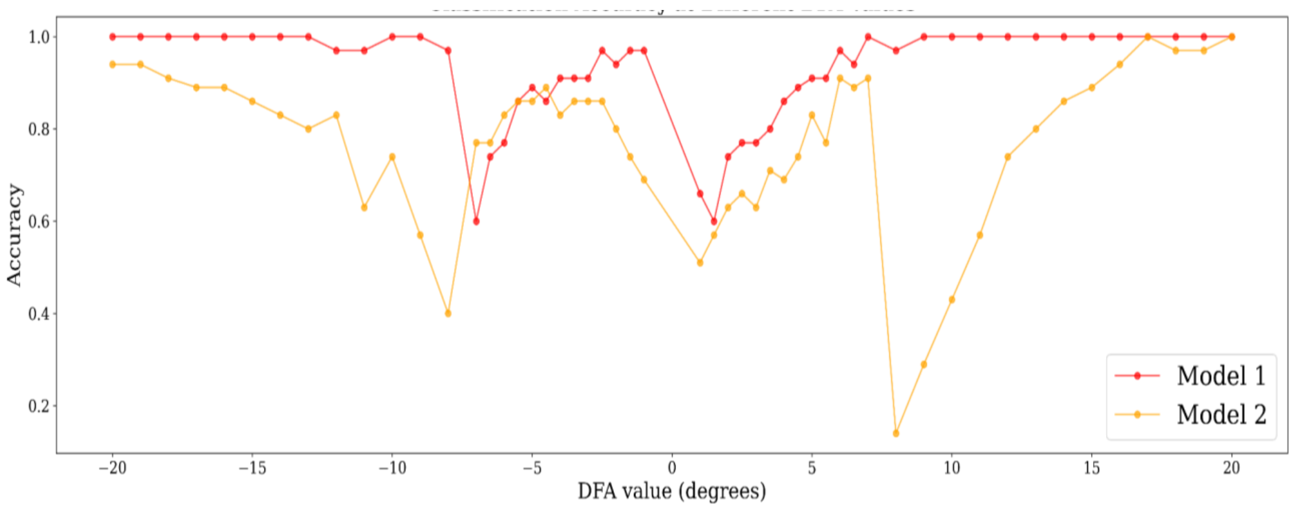

**Fig 9.** Classification accuracy of both models at different DFAs s when tested on e synthetic data. DFA: Disc-Fovea Angle.

### *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 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.

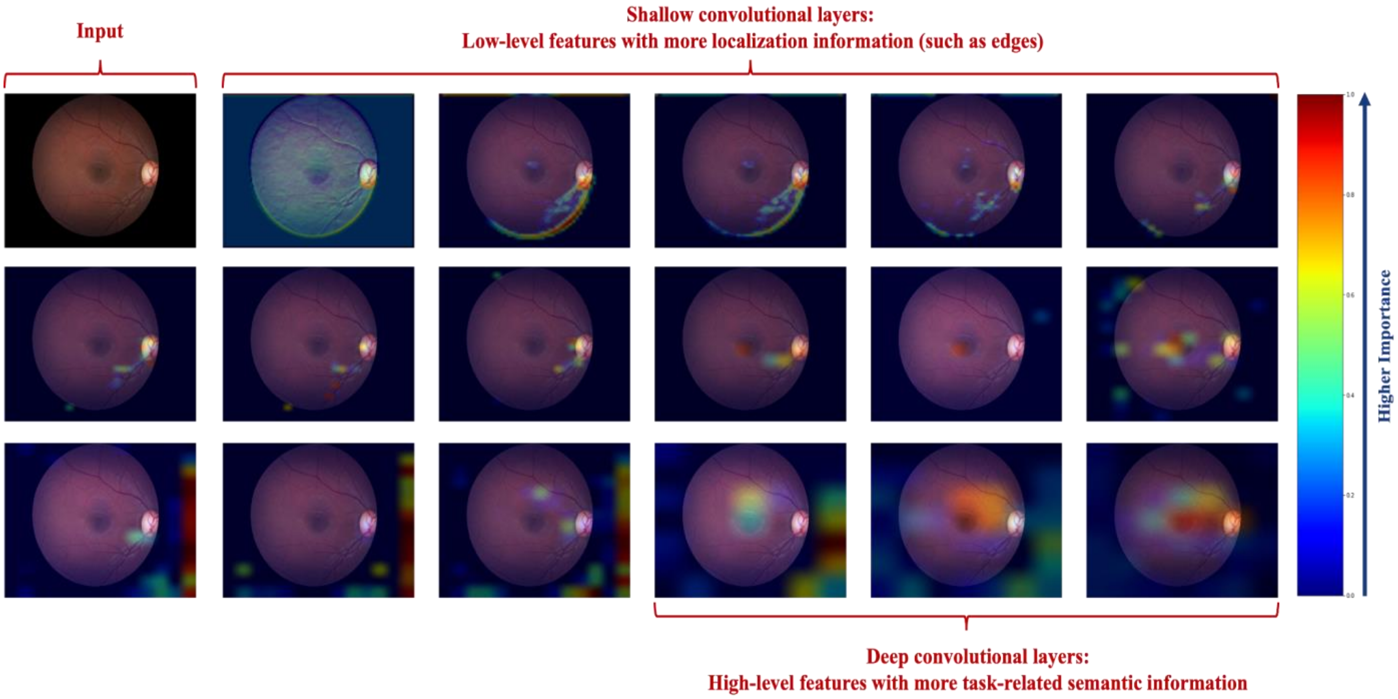
Input
Shallow convolutional layers:
Low-level features with more localization information (such as edges)
1.0
0.8
0.6
0.4
0.2
0.0
Higher Importance
Deep convolutional layers:
High-level features with more task-related semantic information

**Fig 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).

## Discussion

A binary classifier (differentiates intorsion from extorsion), and a multiclass classifier (characterizes torsion in pathologic and physiologic DFA ranges) were successfully developed with deep learning using only synthetic images generated from a small dataset (n=245). When training small datasets, artificial data synthesis and transfer learning are necessary. With advances in generative adversarial networks (GAN)-based models in artificial intelligence (AI) research, synthetic data is increasingly used to overcome the scarcity of annotated medical datasets[43–45,56]. Since skew deviation and SOP fundus datasets are rare, we generated synthetic fundus torsional images to train our models. 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 validated by the authors to accurately mimic real fundus images and torsion cases. Therefore, given the robust performance of Model 1 **(Table 1 and Fig. 7-8)** and its generalizability on non-synthetic data, we can conclude that the generated photographs were comparable to real fundus torsion **(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 better[57,58].

The training of CNN involves parallel computation and a massive number of floating-point operations such as matrix and vector operations[48]. Such computing patterns are well suitable for graphics processing units (GPU). As such, GPUs are more preferred than central processing units (CPU) for the training of CNN[59]. In this study, however, to protect patient privacy, all the experiments were conducted in an internal computing platform which does not support GPU nodes. 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[41]. This technique facilitates model training within a reasonable timeframe while using less computational power - making it an efficient and effective approach. When applied to our models, the cross-entropy loss decreased gradually during the training process and the training time (~ 400 seconds per training epoch) was acceptable for creating our "robust" model performance.

A 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[60]. 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 **(Table 1)**.

The multiclass classifier (Model 2) may have generalized poorly for various reasons. First, our holdout testing set only contained 31 images, with only 10 images from the intorsion classes (*PATHOIN*=2; *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[44,45,57,58], only artificial ocular torsion was introduced to our dataset. Therefore, there could  still be flaws in the data synthesis pipeline causing invisible differences between synthetic and real data, even though the synthetic and real images seemed identical **(Fig 4)**; making it difficult to accurately explain the model's prediction.

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[49]. The Grad-CAM output from our model **(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. It must be noted that the class activation maps generated were occasionally less reliable. Also, our method of generating synthetic images may have biased the model into detecting only changes in fundus torsion since 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, etc.) to determine whether other fundus-specific factors influence the model's torsional detection accuracy.

Even though ophthalmologists, neurologists, and neuro-otologists can reliably identify static torsion and quantify DFA from fundus photographs, they often find it difficult to distinguish skew deviations from SOPs or partial third nerve palsies when these tools are not readily available. In neuro-otology, the presence of a skew deviation helps to differentiate central and peripheral lesions in patients with acute vertigo. The distinction is especially important for frontline providers who must make rapid triage decisions for acute stroke evaluation and intervention. The three-step bedside HINTS (HI- head impulse, N- Nystagmus, TS – test of skew) battery has been shown to be more sensitive and specific than early brain imaging in identifying strokes in the brainstem and cerebellum in patients with acute vertigo[50–52]. Similar time-sensitive triage decisions are important in the setting of acute vertical diplopia sans vertigo.

For non-specialist frontline providers who have not been trained to detect subtle ocular motor abnormalities, the accuracy of the HINTS battery is much lower[53,54]. Furthermore, small, but usually transient, skew deviations can also be seen in patients with acute peripheral labyrinthine lesions[11,55]. In these cases, the direction of OCR relative to the head tilt and skew deviation may help to differentiate utriculo-ocular pathway lesions rostral to the pontomedullary decussation (always central) from caudal lesions (peripheral or central)[11,14,46]. It is also difficult to differentiate central lesions below the decussation from those affecting the vestibular nerve or labyrinth. In the case of acute vertical double vision without vertigo, the conjugacy of the OCR (direction of rotation of both eyes) helps differentiate skews (conjugate) from SOPs (dysconjugate). When only one eye can be evaluated, accessing static torsion in the hypertropic eye can also readily distinguish skews (intorsion) from SOP (extorsion).

Model 1 produced robust, reproducible results (**Table 1)** 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 the patients with acute vertigo without clear skew deviation but possessing other features of the 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 not help in acute clinical diagnosis/screening of acute vertigo and vertical diplopia.

Analysis of Model 2's misclassifications (false positives and negatives) as shown in **Fig. 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. Studies quantifying the DFA in vertical strabismus have mainly involved patients with superior oblique palsies, and the physiologic and pathologic ranges often overlapped[26,46]. Additionally, changes in vergence, and position of the eye relative to head position, known to influence the degree of torsion, were not accounted for during fundus photography. Tight control of the orbital position of the eye during fundus photography for DFA estimation might create better distinction between physiologic and pathologic ranges. Data on DFA ranges in static torsional pathologies correlated with the degree of vertical misalignments is also lacking. Therefore, differentiating physiologic from pathologic static torsion using current DFA ranges is not very useful clinically for distinguishing skew from SOPs in the setting of acute vertigo and diplopia. Multicenter collaborations will be necessary to obtain enough real-world data for further improvements.

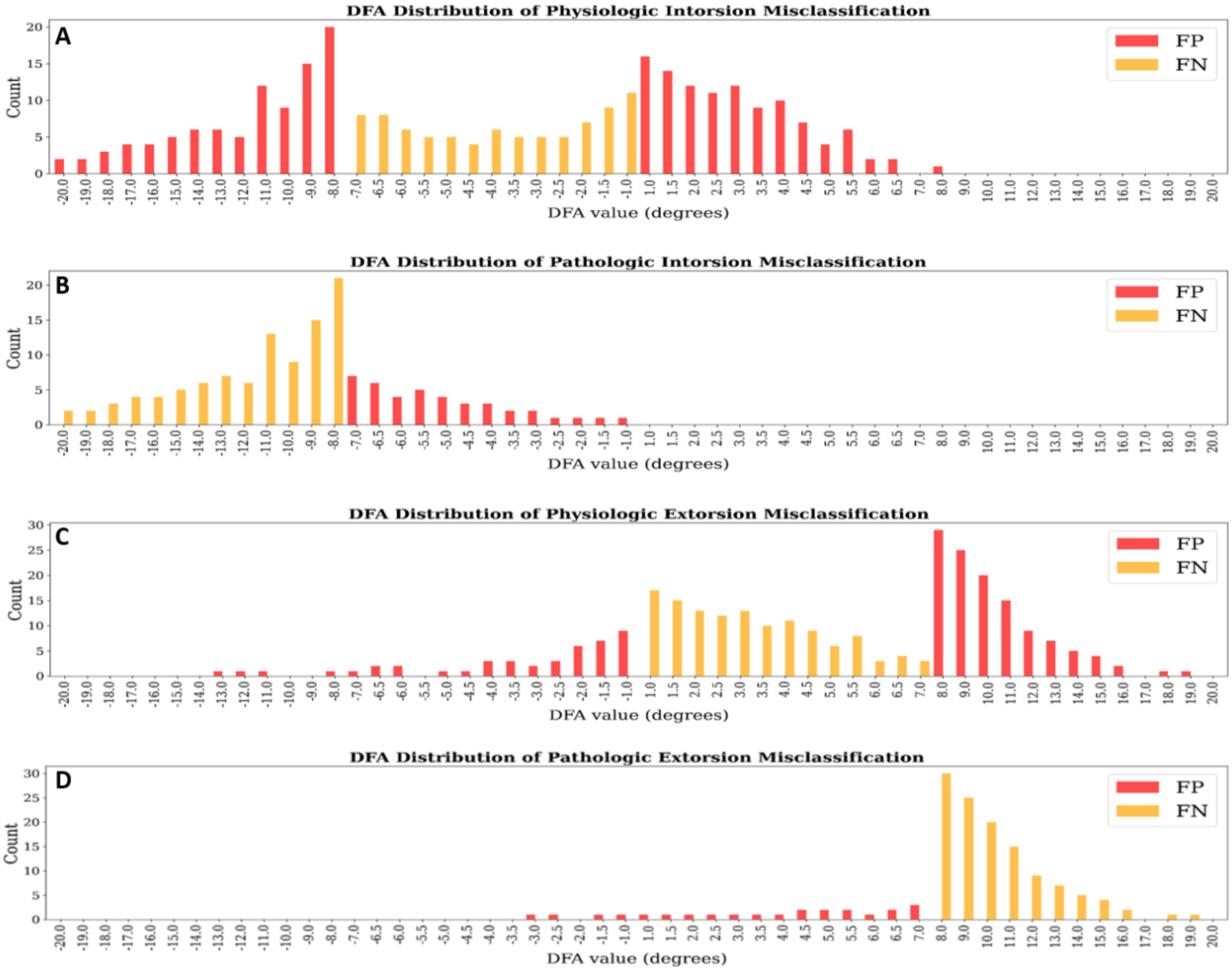

**Fig 11**. Model 2's misclassification DFA distributions for **(A)** physiologic intorsion, **(B)** pathologic intorsion, **(C)** physiologic extorsion, and **(D)** pathologic extorsion. DFA: Disc-Fovea Angle; FP: False Positives; FN: False Negatives.

### *Limitations*

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[38,42,61]. Automated screening for skews and fourth nerve palsies will require assessing both eyes. Third, our holdout testing set is relatively small and not balanced for all classes (more extorsions than intorsions). Finally, we have not verified the model performance using other external physiologic and pathologic datasets. Differences among datasets such as how images were acquired, the resolution of images and the characteristics of patients might affect the model's generalizability

## 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, though 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 much modifications, and 2) monitor treatment responses in neuro-ophthalmic, strabismic and neuro-vestibular diseases.

## 3. 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. doi:10.1111/j.1749-6632.1992.tb25234.x

2. Diamond SG, Markham CH. Ocular counterrolling as an indicator of vestibular otolith function. *Neurology*. 1983;33(11):1460-1460. doi:10.1212/WNL.33.11.1460

3. Kingma H, Stegeman P, Vogels R. Ocular torsion induced by static and dynamic visual stimulation and static whole body roll. *European Archives of Oto-Rhino-Laryngology*. 1997;254(S1):S61-S63. doi:10.1007/BF02439726

4. Leigh RJ, Zee DS. *The Neurology of Eye Movements*. 5th edition. Oxford University Press; 2015.

5. Raps EC, Solomon D, Galetta SL, Liu GT, Volpe NJ. Cyclodeviation in Skew Deviation. *American Journal of Ophthalmology*. 1994;118(4):509-514. doi:10.1016/S0002-9394(14)75804-0

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. doi:10.1001/jamaoto.2021.0176

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. doi:10.1080/00016489950180379

8. Schworm HD, Ygge J, Pansell T, Lennerstrand G. Assessment of Ocular Counterroll during Head Tilt Using Binocular Video Oculography. *Investigative Ophthalmology & Visual Science*. 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. doi:10.1097/00001756-200604240-00011

10. Dieterich M, Brandt T. Perception of Verticality and Vestibular Disorders of Balance and Falls. *Front Neurol*. 2019;10:172. doi:10.3389/fneur.2019.00172

11. Green KE, Gold DR. HINTS Examination in Acute Vestibular Neuritis: Do Not Look Too Hard for the Skew. *Journal of Neuro-Ophthalmology*. 2020;Publish Ahead of Print. doi:10.1097/WNO.0000000000001013

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. doi:10.1002/ana.410330518

13. Brodsky MC, Donahue SP, Vaphiades M, Brandt T. Skew deviation revisited. *Surv Ophthalmol*. 2006;51(2):105-128. doi:10.1016/j.survophthal.2005.12.008

14. Halmagyi GM, Gresty MA, Gibson WPR. Ocular tilt reaction with peripheral vestibular lesion. *Annals of Neurology*. 1979;6(1):80-83. doi:10.1002/ana.410060122

15. Hotson JR, Baloh RW. Acute Vestibular Syndrome. *New England Journal of Medicine*. 1998;339(10):680-685. doi:10.1056/NEJM199809033391007

16. Wong AMF. Understanding skew deviation and a new clinical test to differentiate it from trochlear nerve palsy. *Journal of American Association for Pediatric Ophthalmology and Strabismus*. 2010;14(1):61-67. doi:10.1016/j.jaapos.2009.11.019

17. Gold DR, Shin RK, Galetta S. Pearls and Oy-sters: Central fourth nerve palsies. *Neurology*. 2012;79(23):e193-e196. doi:10.1212/WNL.0b013e3182768998

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, 2022:10.1212/WNL.0000000000200835. doi:10.1212/WNL.0000000000200835

19. Bielschowsky A. LECTURES ON MOTOR ANOMALIES OF THE EYES: II. PARALYSIS OF INDIVIDUAL EYE MUSCLES. *Archives of Ophthalmology*. 1935;13(1):33. doi:10.1001/archopht.1935.00840010043006

20. Bielschowsky A. DISTURBANCES OF THE VERTICAL MOTOR MUSCLES OF THE EYES. *Archives of Ophthalmology*. 1938;20(2):175-200. doi:10.1001/archopht.1938.00850200013001

21. Yoo HS, Park E, Rhiu S, et al. A computerized red glass test for quantifying diplopia. *BMC Ophthalmology*. 2017;17(1):71. doi:10.1186/s12886-017-0465-8

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 Ophthalmology*. 2018;136(4):322. doi:10.1001/jamaophthalmol.2017.6796

23. Brodsky MC, Klaehn L, Goddard SM, Link TP. Heidelberg Spectralis infrared video imaging: a clinical tool for diagnosing ocular torsional instability. *Journal of American Association for Pediatric Ophthalmology and Strabismus*. 2014;18(3):306-307. doi:10.1016/j.jaapos.2014.01.009

24. 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. doi:10.1007/s004170100347

25. Sophocleous S. Use of optical coherence topography for objective assessment of fundus torsion. *BMJ Case Reports*. Published online February 23, 2017:bcr2016216867. doi:10.1136/bcr-2016-216867

26. Kang H, Lee SJ, Shin HJ, Lee AG. Measuring ocular torsion and its variations using different nonmydriatic fundus photographic methods. Madigan M, ed. *PLOS ONE*. 2020;15(12):e0244230. doi:10.1371/journal.pone.0244230

27. 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. doi:10.1371/journal.pone.0141771

28. 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. doi:10.1007/s00415-009-5341-x

29. Lemos J, Eggenberger E. Clinical utility and assessment of cyclodeviation: *Current Opinion in Ophthalmology*. 2013;24(6):558-565. doi:10.1097/ICU.0000000000000003

30. 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 Journal of Ophthalmology*. 2010;58(5):417. doi:10.4103/0301-4738.67060

31. Guyton DL. Ocular torsion: Sensorimotor principles. *Graefe's Archive for Clinical and Experimental Ophthalmology*. 1988;226(3):241-245. doi:10.1007/BF02181189

32. Guyton DL. Ocular Torsion Reveals the Mechanisms of Cyclovertical Strabismus The Weisenfeld Lecture. *Investigative Opthalmology & Visual Science*. 2008;49(3):847. doi:10.1167/iovs.07-0739

33. 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. doi:10.1371/journal.pone.0191007

34. Fleming C. Screening for Primary Open-Angle Glaucoma in the Primary Care Setting: An Update for the US Preventive Services Task Force. *The Annals of Family Medicine*. 2005;3(2):167-170. doi:10.1370/afm.293

35. Biousse V, Newman NJ, Najjar RP, et al. Optic Disc Classification by Deep Learning versus Expert Neuro-Ophthalmologists. *Annals of Neurology*. 2020;88(4):785-795. doi:10.1002/ana.25839

36. 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. doi:10.1212/WNL.0000000000012226

37. Milea D, Najjar RP, Jiang Z, et al. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *New England Journal of Medicine*. 2020;382(18):1687-1695. doi:10.1056/NEJMoa1917130

38. 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 Ophthalmology*. 2019;137(12):1353. doi:10.1001/jamaophthalmol.2019.3501

39. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*. 2019;6(1):60. doi:10.1186/s40537-019-0197-0

40. Akram MU, Abdul Salam A, Khawaja SG, Naqvi SGH, Khan SA. RIDB: A Dataset of fundus images for retina based person identification. *Data in Brief*. 2020;33:106433. doi:10.1016/j.dib.2020.106433

41. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *Journal of Big Data*. 2016;3(1):9. doi:10.1186/s40537-016-0043-6

42. 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. *Journal of Neuro-Ophthalmology*. 2020;40(2):178-184. doi:10.1097/WNO.0000000000000827

43. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*. 2021;5(6):493-497. doi:10.1038/s41551-021-00751-8

44. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification. Published online 2018. doi:10.48550/ARXIV.1801.02385

45. Torfi A, Fox EA, Reddy CK. Differentially Private Synthetic Medical Data Generation using Convolutional GANs. Published online 2020. doi:10.48550/ARXIV.2012.11774

46. 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. doi:10.1007/s00221-019-05529-8

47. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Published online 2015. doi:10.48550/ARXIV.1512.03385

48. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2017.

49. 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. doi:10.1007/s11263-019-01228-7

50. Kattah JC, Talkad AV, Wang DZ, Hsieh YH, Newman-Toker DE. HINTS to diagnose stroke in the acute vestibular syndrome: three-step bedside oculomotor examination more sensitive than early MRI diffusion-weighted imaging. *Stroke*. 2009;40(11):3504-3510. doi:10.1161/STROKEAHA.109.551234

51. Newman-Toker DE, Kerber KA, Hsieh YH, et al. HINTS outperforms ABCD2 to screen for stroke in acute continuous vertigo and dizziness. *Acad Emerg Med*. 2013;20(10):986-996. doi:10.1111/acem.12223

52. Newman-Toker DE, Edlow JA. TiTrATE: A Novel, Evidence-Based Approach to Diagnosing Acute Dizziness and Vertigo. *Neurol Clin*. 2015;33(3):577-599, viii. doi:10.1016/j.ncl.2015.04.011

53. Dmitriew C, Regis A, Bodunde O, et al. Diagnostic Accuracy of the HINTS Exam in an Emergency Department: A Retrospective Chart Review. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*. 2021;28(4):387-393. doi:10.1111/acem.14171

54. Ohle R, Montpellier RA, Marchadier V, et al. Can Emergency Physicians Accurately Rule Out a Central Cause of Vertigo Using the HINTS Examination? A Systematic Review and Meta-analysis. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*. 2020;27(9):887-896. doi:10.1111/acem.13960

55. Korda A, Zamaro E, Wagner F, et al. Acute vestibular syndrome: is skew deviation a central sign? *Journal of Neurology*. 2022;269(3):1396-1403. doi:10.1007/s00415-021-10692-6

56. 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

57. 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. doi:10.1001/jamaophthalmol.2018.6156

58. Guo J, Pang Z, Yang F, Shen J, Zhang J. Study on the Method of Fundus Image Generation Based on Improved GAN. *Mathematical Problems in Engineering*. 2020;2020:1-13. doi:10.1155/2020/6309596

59. 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. doi:10.1109/ICPP.2016.15

60. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain Generalization: A Survey. Published online 2021. doi:10.48550/ARXIV.2103.02503

61. 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. doi:10.1371/journal.pone.0222025

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.

**Fig 2.** Overview of novel algorithm for fundus image notch removal.

**Fig 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.

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

**Fig 5.** Entire pipeline of data preparation and development of models. In the preprocessing stage, the JHH dataset was split into a holdout testing set and a data set for model development. Each photograph had the notch removed and DFA to be 0 degree using rotation. In the data synthesis stage, synthetic torsion photographs were generated using different predefined DFA ranges. In the 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.

**Fig 6.** DFA Distribution of JHH Dataset (excluding DFA = -25.3). DFA: Disc-Fovea Angle; JHH: Johns Hopkins Hospital.

**Fig 7**. 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; ROC: Receiver Operating Characteristic.

**Fig 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.

**Fig 9.** Classification accuracy of both models at different DFAs of photographs in the synthetic testing set. DFA: Disc-Fovea Angle.

**Fig 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).

**Fig 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 Positives; FN: False Negatives.