bioengineering

MDPI

Brief Report

Multi-Dataset Comparison of Vision Transformers and Convolutional Neural Networks for Detecting Glaucomatous Optic Neuropathy from Fundus Photographs

Elizabeth E. Hwang [1,2,†], Dake Chen [1,†], Ying Han [1], Lin Jia [3,*] and Jing Shan [1,*]

[1] Department of Ophthalmology, University of California, San Francisco, San Francisco, CA 94143, USA
[2] Medical Scientist Training Program, University of California, San Francisco, San Francisco, CA 94143, USA
[3] Digillect LLC, San Francisco, CA 94158, USA
* Correspondence: linjia@digillect.xyz (L.J.); jing.shan@ucsf.edu (J.S.)
† These authors contributed equally to this work.

Abstract: Glaucomatous optic neuropathy (GON) can be diagnosed and monitored using fundus photography, a widely available and low-cost approach already adopted for automated screening of ophthalmic diseases such as diabetic retinopathy. Despite this, the lack of validated early screening approaches remains a major obstacle in the prevention of glaucoma-related blindness. Deep learning models have gained significant interest as potential solutions, as these models offer objective and high-throughput methods for processing image-based medical data. While convolutional neural networks (CNN) have been widely utilized for these purposes, more recent advances in the application of Transformer architectures have led to new models, including Vision Transformer (ViT,) that have shown promise in many domains of image analysis. However, previous comparisons of these two architectures have not sufficiently compared models side-by-side with more than a single dataset, making it unclear which model is more generalizable or performs better in different clinical contexts. Our purpose is to investigate comparable ViT and CNN models tasked with GON detection from fundus photos and highlight their respective strengths and weaknesses. We train CNN and ViT models on six unrelated, publicly available databases and compare their performance using well-established statistics including AUC, sensitivity, and specificity. Our results indicate that ViT models often show superior performance when compared with a similarly trained CNN model, particularly when non-glaucomatous images are over-represented in a given dataset. We discuss the clinical implications of these findings and suggest that ViT can further the development of accurate and scalable GON detection for this leading cause of irreversible blindness worldwide.

Keywords: glaucoma; deep learning; vision transformer; fundus photography

Citation: Hwang, E.E.; Chen, D.; Han, Y.; Jia, L.; Shan, J. Multi-Dataset Comparison of Vision Transformers and Convolutional Neural Networks for Detecting Glaucomatous Optic Neuropathy from Fundus Photographs. *Bioengineering* **2023**, *10*, 1266. https://doi.org/10.3390/bioengineering10111266

Academic Editors: Karanjit S. Kooner and Osamah J. Saeedi

Received: 17 October 2023
Revised: 26 October 2023
Accepted: 27 October 2023
Published: 30 October 2023

1. Introduction

Glaucoma is a group of chronic, progressive optic neuropathies are a leading cause of vision loss worldwide [1]. Primary open-angle glaucoma (POAG) is the most common type of glaucoma, with cases estimated to rise from 2.7 million in 2011 to 7.3 million by 2050 in the United States alone [2]. While most often associated with increased intraocular pressure (IOP), the disease process can also occur with normal or low IOP and is often referred to as the "silent thief of sight" because it typically progresses slowly and without noticeable symptoms in its early stages. Thus, early detection, close monitoring, and timely interventions are key to preserving vision in glaucoma patients, especially among minority populations such as Hispanics/Latinos and African Americans, who are disproportionately affected relative to non-Hispanic Whites [3]. However, currently the United States Preventive Services Task Force (USPSTF) does not recommend screening for primary open-angle glaucoma in asymptomatic adults 40 years or older. In their updated 2022 review, the USPSTF cited the need for targeted screening among high-risk populations

(such as individuals with a family history of glaucoma or from disproportionately affected minority groups), optimizing contemporary screening approaches and modalities to improve both efficiency and cost-effectiveness, and clinical trials demonstrating the utility of such screening approaches in vision-related patient outcomes [4].

Deep learning-aided diagnostic interpretation has received significant interest for its potential to improve the accuracy of diagnosing glaucoma and deliver high-throughput screening tools optimized for early diagnosis in at-risk patients [5]. Glaucoma diagnosis often requires complex medical imaging of the optic nerve and retina in a specialist setting, and even then, is subject to inter-observer variability. Deep learning models have the potential to detect subtle structural changes missed by the eye, provide consistent results, and improve efficiency by reducing the burden on glaucoma specialists. Application of deep learning models to glaucoma diagnosis would also allow for high-throughput screening to identify asymptomatic disease and improve patient outreach, particularly in resource-limited settings.

Among the imaging modalities, fundus photography is a widely available, relatively low-cost approach already employed for clinical use in diabetic retinopathy tele-screening. In glaucoma, fundus photos provide the vertical optic nerve cup-to-disc ratio (vCDR), which quantifies the relationship between the cup (the central depression on the optic nerve head) and the disc (the entire optic nerve head) which enlarges as the disease progresses. Interpretation of these photos, however, can be difficult to reproduce among even expert specialists, and exhibit high rates of inter-observer variability [6–8], as well as being subject to observer bias (e.g., the tendency to under-call optic neuropathy in small optic discs while overcalling disease in physiologically large discs [9]). Therefore, the development and application of an AI tool to classify GON could greatly enhance fundus photography's utility as a population-based screening tool.

Previous studies have shown that deep learning models individually trained on color fundus photos [10], visual field analysis [11–14], and optical coherence tomography (OCT) [15–19] are able to identify glaucomatous optic neuropathy (GON) with robust performance (comparisons of specific deep learning models developed for glaucoma diagnosis and discussions of the different approaches are thoroughly covered in excellent reviews from Thompson et al. [5] and Yousefi [20]). Indeed, a recent meta-analysis of 17 deep-learning models trained on diagnosing GON from fundus photographs reported an overall AUC of 0.93 (95% CI 0.92–0.94), slightly lower than the AUC reported for studies using OCT (overall AUC 0.96, 95% CI 0.94–0.98) [21]. Several of these studies included external validation sets of up to six cohorts, suggesting that their models may generalize to unseen outside data. However, because these models are large and require intensive computational resources to train, they have been trained on datasets that are most often inaccessible to the public, thus making it difficult to compare whether the models themselves show differences during the training process.

To date, many AI models for glaucoma classification have utilized convolutional neural networks (CNNs). CNNs provide a scalable approach to object recognition within images by processing spatial patterns and extracting relevant features [22]. This architecture enables CNNs to automatically learn hierarchical representations of the features in an image. In supervised learning, CNNs are trained on labeled datasets, while in unsupervised learning, unsupervised methods like autoencoders are utilized for feature extraction. Semi-supervised learning, such as transfer learning, are also commonly described, as pre-trained models can be fine-tuned with smaller labeled datasets to improve performance [5,23]. However, a well-known attribute of CNNs is their inherent bias towards translation-invariant object recognition [24] which permits the interpretation of features outside of their spatial context [25], leaving models vulnerable to artifactual errors. Current models attempt to alleviate this by strict standardization of inputs, which, unfortunately, further restricts the ability of CNN algorithms to generalize to new, and even related, tasks without labor-intensive preprocessing.

In the last decade, Vision Transformer (ViT) [26], among other transformer architectures [27], has taken advantage of the self-attention mechanisms used in natural language processing to improve upon these limitations of CNNs. In contrast to CNNs, ViT models process the entire image as a sequence of patches, thus allowing for the capture of global relationships. ViTs have also been shown to generalize from smaller datasets than CNNs, which are heavily reliant upon pre-training and fine-tuning for optimal performance [26]. ViT models have now been applied to the analysis and interpretation of a wide range of clinical data ranging from electrocardiograms [28] to intraoperative surgical techniques [29]. In ophthalmology, there are increasing reports of ViT models trained to classify retinal pathologies from fundus photography [30–32] and OCT imaging [33–36], including several assessing their performance relative to CNNs [31,34,35]. Given that glaucoma diagnosis often requires multimodal imaging that correlates structural and functional data, it has been theorized that the global attention mechanisms utilized by ViTs offer an advantage over CNNs' dependency upon local features. However, few such reports in the ophthalmic literature benchmark one model against the other, and even fewer compare the outcomes from more than one training dataset. This represents a knowledge gap for AI-guided GON detection, since an optimal architecture should be able to generalize across variables that vary by clinical setting, such as patient population, image format, and disease prevalence.

In this report, we describe the training of ViT and CNN models on six publicly available, independent datasets, compare the two models' accuracies, and discuss the potential clinical applications for each type of model. We propose that the choice between these two model architectures may depend upon the specific clinical setting, labeled data availability, and computational resources. Ultimately, we hope that our results provide insight into model selection for specific clinical tasks as well as effective database construction.

2. Materials and Methods

2.1. Datasets

A total of six public datasets were included for analysis in this paper (Figure 1, Table 1) [37–42]. Complete dataset sizes varied between 101 images (Drishti-GS1) to 720 images (REFUGE2). Though representation of non-glaucomatous (control) and glaucomatous classes varied between the datasets, no obvious correlation existed between total dataset size and class ratios (Table 1). When provided by the original authors, the patient selection criteria and instrument cameras are also noted in Table 1.

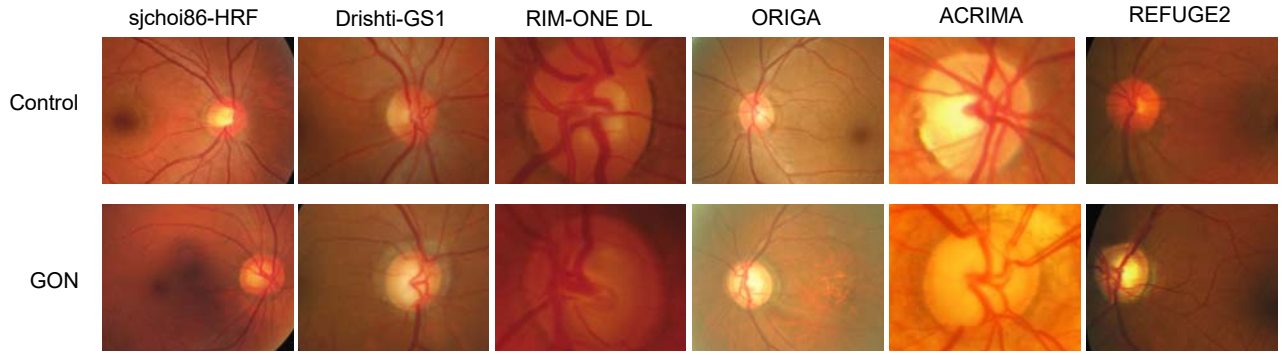


Figure 1. Representative fundus photographs from the datasets used in this study. GON: glaucomatous optic neuropathy.

All datasets included ground truth labels indicating GON or control. Most datasets derived ground truth from expert labeling and clinical annotations with the exceptions of ORIGA (algorithm-based) and sjchoi86-HRF (unknown). Three datasets (sjchoi86-HRF, ORIGA, and REFUGE2) provided whole fundus images, one provided OD-centered images (Drishti-GS1), and two provided OD-cropped images (RIM-ONE DL and ACRIMA) (Figure 1). Sources accessed for each of the datasets are provided in the references. No photographs were excluded from our analysis.

Table 1. Characteristics of publicly available datasets used in this study, as ordered by total set size.

Study	Patient Selection	Instrument	Ground Truth	Images			
				Non-GC (Control)	GC	Total Size	Class Ratio *
Drishti-GS1 [37]	Glaucomatous and routine refraction images selected by experts from patients between ages 40 and 80 at Aravind Eye Hospital in India	Not noted	4 experts	31	70	101	0.44
sfchoi86-HRF [38]	Unknown	Unknown	Unknown	300	101	401	2.97
RIM-ONE DL [39]	Curated extraction from RIM-ONE V1, V2, and V3 of glaucomatous and healthy patients from 3 hospitals in Spain	V1/V2: Nidek AFC-210 camera V3: Kowa WX 3D non-stereo camera	2 experts with tiebreaker	312	173	485	1.80
ORIGA [40]	Glaucomatous and randomly selected non-GC images from cross-sectional population Singaporean study (SiMES) of Malay adults between ages 40 and 80	Canon CR-DGi	ORIGA-GT	482	168	650	2.87
ACRIMA [41]	Glaucomatous and normal images selected by experts in Spain based on clinical findings	Topcon TRC retinal camera	2 experts	309	396	705	0.78
REFUGE2 [42]	Random selection from glaucoma and myopia study cohorts in China (Zongshan Ophthalmic Center)	KOWA, TOPCON	7 experts	720	80	800	9.00

* Calculated as a ratio of non-GC: GC images. GC = glaucomatous.

2.2. Image Preprocessing

To minimize the presence of redundant information which could potentially impact deep learning model performance, we conducted pre-processing of all images to extract the region around the optic nerve head from each fundus image as shown in the depicted model (Figure 2). This was achieved using deeplabv3plus [43], a semantic segmentation model. Once the region of interest was extracted, we automatically cropped a square area centered around the disc. These extracted images were then utilized to train the CNN and ViT models described below. By focusing on specific areas, we aimed to improve the model's ability to identify glaucoma-related features and enhance the accuracy of the automated detection system.

2.3. Vision Transformer (ViT) and ResNet Training and Evaluation

Each one of the public databases was split into a training set (80%) and a testing set (20%) (Figure 2). This ensured a consistent and fair evaluation of both models using identical testing datasets. An overview of our method is shown in Figure 2. For the CNN model, we leveraged the standard ResNet-50 which has 50 layers with incorporated residual connections with no further tuning [43]. For the ViT architecture [25], we used 12 attention layers and a patch size of 16, hidden size of 768, and 12 heads. Following the practices established by [44] and [25], we pretrained the ViT on the ImageNet dataset [45]. The images were resized to a uniform size of 224 × 224 pixels. Additionally, we normalized the pixel values to a range between 0 and 1. During training, we used a batch size of 16 and employed the AdamW optimizer with a learning rate of $6e^{-4}$ and regularization of $6e^{-2}$. These hyperparameters (included in Figure 2) were chosen to optimize the model's convergence and performance. To compute the loss during training, we employed cross-entropy loss with 0.1 label smoothing.

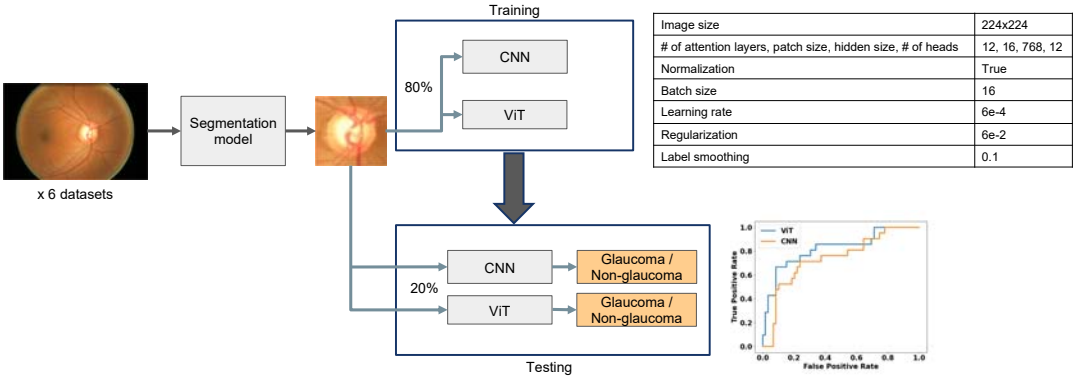


Image size	224x224
# of attention layers, patch size, hidden size, # of heads	12, 16, 768, 12
Normalization	True
Batch size	16
Learning rate	6e-4
Regularization	6e-2
Label smoothing	0.1

Figure 2. Workflow of ViT vs. CNN model training (with hyperparameters) and validation.

Performance metrics, including area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy, F1 score, and mAP (mean Average Precision), were calculated from models evaluated on the held-out test sets (Table 2). Specificities were calculated at a fixed sensitivity threshold of 50%. Confidence intervals (CI) were determined by bootstrap resampling of the test sets with replacement (n = 10 times) while the training sets and models remained fixed.

Table 2. Performance statistics for ViT and CNN models evaluated on held-out test sets. Confidence intervals of 95% are reported in parentheses.

	ViT			CNN		
	AUC	***Sensitivity***	***Specificity***	***AUC***	***Sensitivity***	***Specificity***
Drishti-GS1	0.67 (0.44, 0.97)	0.93 (0.79, 1.00)	0.40 (0.00, 1.00)	0.67 (0.38, 0.91)	0.73 (0.50, 0.93)	0.91 (0.00, 1.00)
sjchoi86-HRF	0.79 (0.67, 90)	0.67 (0.46, 0.82)	0.92 (0.84, 0.98)	0.71 (0.59, 82)	0.52 (0.31, 0.75)	0.90 (0.81, 0.97)
RIM-ONE DL	0.88 (0.81, 0.94)	0.91 (0.79, 1.00)	0.85 (0.76, 0.93)	0.86 (0.78, 0.93)	0.81 (0.67, 0.94)	0.91 (0.83, 0.97)
ORIGA	0.69 (0.60, 0.77)	0.52 (0.37, 0.67)	0.85 (0.77, 0.92)	0.62 (0.54, 0.70)	0.36 (0.21, 0.52)	0.88 (0.81, 0.95)
ACRIMA	0.94 (0.90, 0.97)	1.00 (1.00, 1.00)	0.88 (0.79, 0.95)	0.92 (0.88, 0.96)	0.84 (0.76, 0.92)	1.00 (1.00, 1.00)
REFUGE2	0.95 (0.88, 1.00)	0.94 (0.80, 1.00)	0.97 (0.94, 0.99)	0.89 (0.78, 0.99)	0.81 (0.60, 1.00)	0.97 (0.94, 0.99)
	ViT			**CNN**		
	Accuracy	***F1 Score***	***mAP***	***Accuracy***	***F1 Score***	***mAP***
Drishti-GS1	0.80 (0.60, 0.95)	0.87 (0.73, 0.97)	0.82 (0.63, 0.99)	0.70 (0.50, 0.90)	0.79 (0.58, 0.93)	0.82 (0.61, 0.98)
sjchoi86-HRF	0.81 (0.72, 0.89)	0.68 (0.51, 0.82)	0.53 (0.35, 0.72)	0.80 (0.71, 0.89)	0.58 (0.37, 0.76)	0.46 (0.28, 0.65)
RIM-ONE DL	0.87 (0.79, 0.93)	0.82 (0.71, 0.90)	0.70 (0.56, 0.85)	0.88 (0.80, 0.94)	0.81 (0.69, 0.91)	0.72 (0.56, 0.86)
ORIGA	0.74 (0.66, 0.82)	0.57 (0.44, 0.69)	0.50 (0.37, 0.63)	0.71 (0.63, 0.78)	0.46 (0.31, 0.60)	0.44 (0.32, 0.56)
ACRIMA	0.94 (0.91, 0.98)	0.95 (0.91, 0.98)	0.91 (0.84, 0.96)	0.91 (0.87, 0.96)	0.92 (0.86, 0.96)	0.93 (0.89, 0.97)
REFUGE2	0.97 (0.94, 0.99)	0.83 (0.64, 0.95)	0.72 (0.47, 0.92)	0.96 (0.93, 0.99)	0.79 (0.60, 0.93)	0.64 (0.39, 0.86)

3. Results

In Table 2 and Figure 3, we present the performance statistics and contingency tables of the CNN and ViT models trained to classify non-glaucomatous (non-GC) from glaucomatous (GC) eyes on each of the six public datasets. When compared using relative AUC, the ViT models were non-inferior to the CNN models and appeared to outperform the CNN models on five of the six datasets, though this was not statistically significant given the overlapping confidence intervals. The greatest differences were observed among the sjchoi86-HRF (0.79 ViT vs. 0.71 CNN), ORIGA (0.69 ViT vs. 0.62 CNN) and REFUGE2 (0.95 ViT vs. 0.89 CNN) datasets. No difference in mean AUC was observed for only one dataset, Drishti-GS1 (both 0.67). The performance on the remaining two datasets were also consistently, if marginally, higher for the ViT models (RIM-ONE: 0.88 ViT vs. 0.86 CNN; ACRIMA 0.94 ViT vs. 0.92 CNN). Similar observations were made for the accuracies, F1 scores, mAP, as reflected in the average statistic and the 95% confidence intervals.

The recall or sensitivity of the ViT models surpassed those of the CNN models among the six datasets by an average of 0.14, with the largest difference observed in the Dristhi-GS1 (0.93 ViT vs. 0.73 CNN) and the smallest in REFUGE2 (0.94 ViT vs. 0.81 CNN). By contrast, the specificities were more varied between the two methods, ranging from comparable (sjchoi86-HRF: ViT 0.92 vs. CNN 0.90; ORIGA: 0.85 ViT vs. 0.88 CNN; REFUGE2: ViT 0.97 vs. CNN 0.97) to favoring the CNN model (RIM-ONE: 0.85 ViT vs. 0.91 CNN; ACRIMA: 0.88 ViT vs. 1.00 CNN). For Drishti-GS1, the small sample size of non-GC images in the held-out test set (n = 5 images) resulted in inconclusive specificity statistics as reflected by the 95% confidence intervals of (0,1) for both models.

We noted that ViT tended call more false positives (i.e., label control images as GON) than CNN models in several datasets, including RIM-ONE (10 ViT false positives (FP) vs. 6 CNN FP), ORIGA (13 ViT FP vs. 10 CNN FP), and ACRIMA (8 ViT FP vs. 0 CNN FP). Accordingly, two of these ViT models demonstrated lower specificities than their CNN equivalents (i.e., RIM-ONE: 0.74 ViT vs. 0.81 CNN; ACRIMA: 0.91 ViT vs. 1.00 CNN).

Interestingly, ViT outperformed CNN on datasets with higher ratios of non-GC to GC photos (Figure 4a), though not with total dataset size (Figure 4b). This was most clearly evidenced by the delta AUC of the Drishti-GS1 and REFUGE2 models, whose datasets harbored the lowest (0.44) and highest (9.0) ratios of non-GC to GC images, respectively. Furthermore, the differences in specificity between the ViT and CNN models diminished as the ratio of non-GC to GC images increased (Figure 4c): when trained on the REFUGE2 dataset, the specificity of the ViT model overlapped with that of the CNN model (0.97, CI 0.94–0.99).

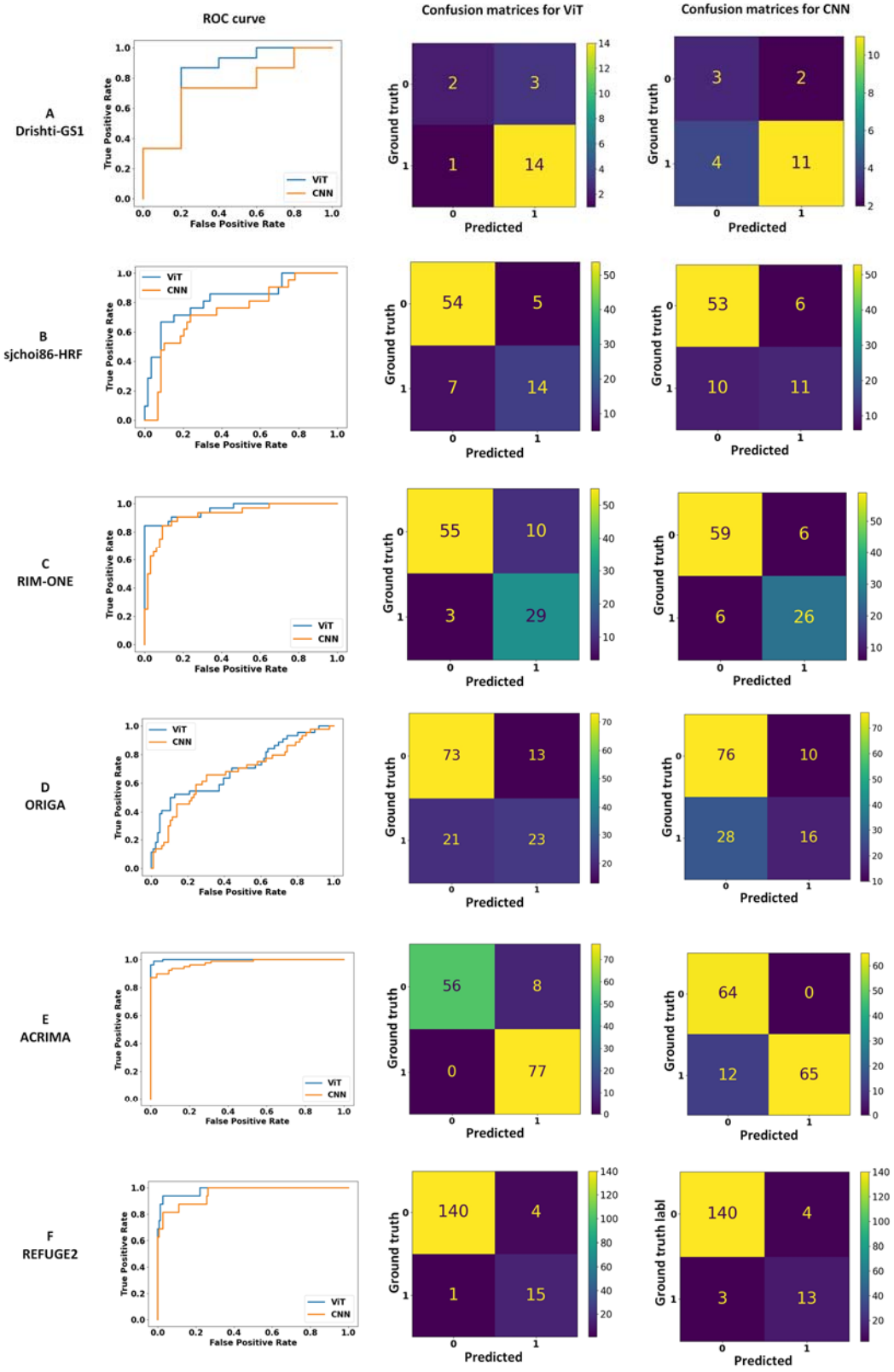
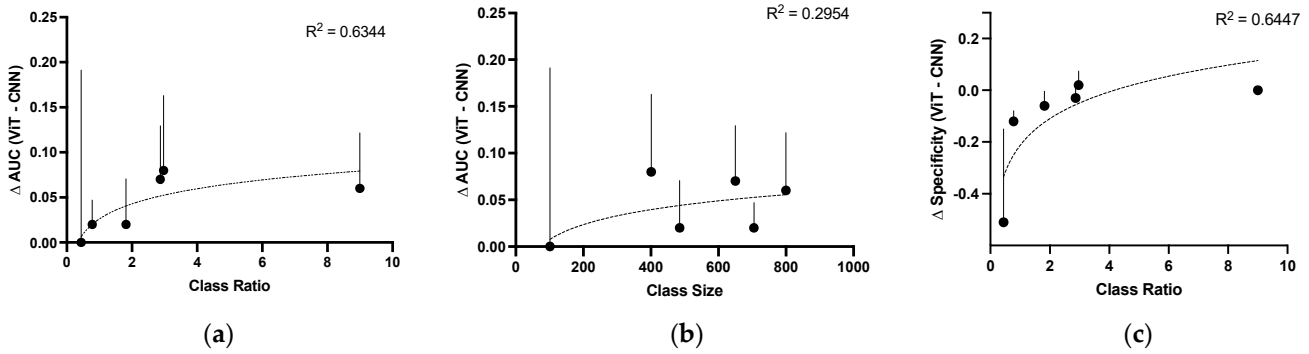


Figure 3. ROC curves and confusion matrices for ViT and CNN models trained on individual datasets (**A**–**F**). For the confusion matrices, a classification of 0 refers to control/non-glaucomatous, whereas a classification of 1 refers to glaucomatous. Ground truth labels were used as provided by the original datasets (ref. Table 1).



(**a**) (**b**) (**c**)

Figure 4. ViT outperforms CNN models in datasets with greater class imbalance but not class size. (Δ = ViT − CNN, where ViT outperforms CNN when Δ > 0, and CNN outperforms ViT when Δ < 0) Log-linear regression models (dotted lines) are included with coefficients of determination as indicated. (**a**) ΔAUC as a function of class ratio. (**b**) ΔAUC as a function of class size. (**c**) ΔSpecificity as a function of class ratio. See Table 1 for class sizes and ratios.

4. Discussion

Here we focus the performance of ViT and CNN models trained on glaucoma detection from a single imaging modality, fundus photography. We take advantage of ImageNet pre-trained models to test and train each architecture on six publicly available annotated datasets, which were collected from at least four countries (India, Spain, Singapore, and China) and varied in size from 101 to 800 total images. Class imbalances between control and glaucomatous labeled images were present to varying degrees among the datasets, from the most evenly matched (Drishti-GS1, class ratio of 0.44 or 69% glaucoma prevalence) to the least (REFUGE2, class ratio of 9.0 or 10% glaucoma prevalence). A survey of 14 US-based studies found all glaucoma prevalence rates ranging from 2.1% to 25.5%, and POAG prevalence rates between 1.86% and 13.8% [44]. Thus, though these datasets represent selected cohorts rather than a population survey, it is likely that the "lower" prevalence cohorts more accurately reflect the dataset composition that would be expected from a moderate to high-risk screening population. We had two objectives from comparing the models trained upon multiple, rather than pooled, datasets: first, to ask whether the ViT model could perform equal to, or better than, a widely accepted CNN model, ResNet-50, and second, to determine whether ViT or CNN models, when trained on datasets of different sizes and class representations, demonstrated any trends in performance metrics including AUC, sensitivity, and specificity, that might inform future clinical application of the two architectures.

As predicted, we found that the pre-trained ViT model matched or outperformed the equivalent CNN model on all six datasets by AUC and accuracy measures. We also observed that the ViT models increasingly outperformed CNN models, as measured by AUC and specificity, on datasets with greater representations of controls (i.e., higher class ratio). We suggest that this difference may reflect that, when presented with insufficient control representation, ViT struggles with the greater variability present among non-glaucomatous optic nerve discs due to the wider array of potential relationships when using a global attention mechanism. However, as the ViT algorithm is presented with an increasing number of control examples, it can better assign global relationships to a given class, even when it is as varied as "non-glaucoma".

Given our observations, we would recommend CNN models for GON detection in tasks with uniform data collection where high test specificity outweighs other considerations. In contrast, we would nominate ViT models for tasks requiring collaborative data collections (e.g., clinical trials, multi-site tele-screening), whereby different operators, patient demographics, camera models, and data processing standards are likely to result in datasets with levels of heterogeneity beyond that which CNN models can accommodate. ViT performance could be enhanced further by targeted deployment to patients with iden-

tifiable risk factors, such as a family history of glaucoma, advanced age, or predisposing conditions such as steroid use. Such at-risk populations exhibit higher pre-test probability and would thus benefit from ViT's greater sensitivity.

While CNN and ViT models are widely utilized for high-throughput image analysis and classification, their differences in feature detection and training requirements have led to the suggestion that ViT models may improve upon CNN performance. CNN architecture utilizes a sliding window method to extract features in a local fashion and thus has strict input requirements [45]. Previous strategies for improving CNN performance of photographic GON detection have focused on the optimization of pre-processing techniques like data augmentation [46] and feature extraction [47], as well as more clinically motivated strategies such as structure-function correlation between multiple testing modalities [48,49]. More recently, transformer architectures have gained interest for their ability to use global attention mechanisms to identify long-range interactions [50] and their flexibility in allowing for non-uniform inputs [26]. Therefore, while the prevalence of inductive biases in CNNs relative to transformers may enable ResNet models to outperform ViT models when classification relies upon the presence or absence of locally identifiable features (e.g., optic nerve thinning in defined superior-temporal or inferior-nasal patterns) [26,51]. ViT may ultimately offer superior performance when diagnostic features are distributed in a disconnected manner (e.g., identifying glaucomatous features such as bayoneting). This would be particularly applicable in the setting of multimodal imaging datasets that could potentially rely upon global features, such as the correlation of functional visual field testing with structural changes in the OCT, which so far have required a multi-algorithmic approach [52].

Yet, despite the potential for transformers to incorporate long-range feature detection from multimodal datasets, the literature comparing ViT models to CNN models have generally focused on single modalities due to the challenges of multimodal data integration as an input into a single algorithm [31,32,35,51–55]. One outstanding report compares ViT to CNN models trained on Diabetic Retinopathy (DR) detection from multiple independent datasets consisting of either fundus photos or OCT imaging, and finds that ViT models are superior in both cases; however, no multimodal datasets are used [31]. To the best of our knowledge, only two publications so far have compared the performance of published ViT and CNN models on glaucoma detection from fundus photos [51,56]. In one report, the authors found that Data-efficient Image Transformer (DeIT) models outperformed similarly trained ResNet-50 models [51]. They further compared the DeIT attention maps with ResNet-50 average saliency maps to demonstrate that the transformer model more precisely focused upon the borders of the optic disc where glaucomatous features are most often identified, whereas the CNN saliency maps highlighted the entire optic nerve. Intriguingly, the more recent report found that the ViT model underperformed the CNN models (VGG, ResNet, Inception, MobileNet, DenseNet) on an external validation set [56]. While not directly comparable to our results, we note that their training set was also comprised of three nearly equally represented classes (GON, non-GON, and normal optic discs), perhaps resembling our "lower" class ratio datasets, such as ACRIMA.

Our work builds upon these studies by incorporating the use of independent training sets similar to [31] as well as avoiding the use of fine-tuning between datasets, thus allowing for observations on the baseline performance of the two architectures in multiple settings. Within the constraints of the public datasets utilized by our models, our results suggest that simply switching to ViT-based architecture alone will not significantly improve model performance. This is for two reasons: First, though there is an appreciable trend of higher mean AUCs across the ViT models, the differences between the individual ViT and CNN models were not statistically significant. Second, while ViT models uniformly demonstrated greater sensitivities than the CNN models, we observed that the under-representation of non-GC images during training may have led to lower model specificities. This implies that one trade-off of ViT's global attention mechanism may result in increased dependence upon sufficient class representation during training, which aligns with previous

observations that, for smaller datasets, ViT-based architectures are more dependent upon training set representation than CNN-based architectures [26]. Thus, improving model performance may not rely only upon optimizing the model itself, but also the training data and processes involved. Here we utilized pre-trained models, but other techniques to improve model performance have included transfer learning [57], artifact-tolerant feature representation [10], cross-teaching between CNN and transformer models [58], and hybrid CNN-ViT architectures which extract local features in a patch-based manner [55]. While not addressed here, many of these strategies appear promising and merit further investigation.

We acknowledge a couple of limitations in our study. First, our comparisons of the two models were limited to datasets containing only fundus photography, while in practice, the gold standard diagnosis of glaucomatous optic neuropathy requires the correlation of structural findings (optic nerve thinning) with functional ones (visual field defects) [45]. Secondly, we pre-processed the fundus photos with optic nerve head segmentation to avoid biasing the models with non-disc-related information. Given that ViT uses a global mechanism, we anticipate that the performance of the ViT models may have been disproportionately affected relative to the CNN models. However, given that real-world application of these models often incorporates similar pre-processing for a variety of reasons [59,60], we suggest that this approach remains relevant to clinical practice.

Future works based on these findings may benefit from comparisons of CNN-based vs. ViT-based models on larger cohorts, more rigorous investigations of whether non-glaucomatous representation impacts model performance, and ideally, side-by-side comparisons of both models in various clinical contexts (i.e., screening vs. specialty visits) to determine their efficacies and practicalities in different settings.

5. Conclusions

Overall, our findings suggest that ViT-based algorithms show excellent results regarding glaucoma detection in line with previous studies. However, our results indicate that, despite recent publication trends, CNN models may offer advantages over Transformer models for training datasets with more equal representation of both non-glaucomatous and glaucomatous images. For high-risk populations or other situations where the importance of detecting any disease outweighs the risk of false positives, we propose that ViT models should be considered superior to the more widely utilized CNN-based architectures established within the field.

Automated image processing algorithms for the detection of glaucomatous optic neuropathy can empower population-based screening towards preventing irreversible vision loss. We hope our findings here can further the development of accurate and scalable high-throughput methods for this leading cause of blindness worldwide.

Author Contributions: Conceptualization, Methodology, and Software, J.S., L.J. and D.C.; Investigation, Formal Analysis, and Visualization, E.E.H. and D.C.; Writing—Methods, D.C.; Writing—Original Draft Preparation, E.E.H.; Writing—Review and Editing, E.E.H., J.S., D.C., Y.H. and L.J.; Supervision, Project Administration, and Funding Acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Computational Innovator Faculty Research Award to Dr. Jing Shan from the UCSF Initiative for Digital Transformation in Computational Biology & Health, All May See Foundation and Think Forward Foundation to Dr. Jing Shan, UCSF Irene Perstein Award to Dr. Jing Shan, and National Institutes of Health under NCI Award Number F30CA250157 and NIGMS T32GM007618 MSTP training grant to Elizabeth Hwang.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tham, Y.-C.; Li, X.; Wong, T.Y.; Quigley, H.A.; Aung, T.; Cheng, C.-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology* **2014**, *121*, 2081–2090. [CrossRef] [PubMed]
2. Vajaranant, T.S.; Wu, S.; Torres, M.; Varma, R. The changing face of primary open-angle glaucoma in the United States: Demographic and geographic changes from 2011 to 2050. *Arch. Ophthalmol.* **2012**, *154*, 303–314.e3. [CrossRef] [PubMed]
3. Stein, J.D.; Khawaja, A.P.; Weizer, J.S. Glaucoma in Adults—Screening, Diagnosis, and Management: A Review. *JAMA* **2021**, *325*, 164–174. [CrossRef] [PubMed]
4. Chou, R.; Selph, S.; Blazina, I.; Bougatsos, C.; Jungbauer, R.; Fu, R.; Grusing, S.; Jonas, D.E.; Tehrani, S. Screening for Glaucoma in Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* **2022**, *327*, 1998–2012. [CrossRef]
5. Thompson, A.C.; Jammal, A.A.; Medeiros, F.A. A Review of Deep Learning for Screening, Diagnosis, and Detection of Glaucoma Progression. *Transl. Vis. Sci. Technol.* **2020**, *9*, 42. [CrossRef]
6. Chan, H.H.; Ong, D.N.; Kong, Y.X.; O'Neill, E.C.; Pandav, S.S.; Coote, M.A.; Crowston, J.G. Glaucomatous optic neuropathy evaluation (gone) project: The effect of monoscopic versus stereoscopic viewing conditions on optic nerve evaluation. *Am. J. Ophthalmol.* **2014**, *157*, 936–944. [CrossRef]
7. Denniss, J.; Echendu, D.; Henson, D.B.; Artes, P.H. Discus: Investigating subjective judgment of optic disc damage. *Optom. Vis. Sci.* **2011**, *88*, E93–E101. [CrossRef]
8. Jampel, H.D.; Friedman, D.; Quigley, H.; Vitale, S.; Miller, R.; Knezevich, F.; Ding, Y. Agreement Among Glaucoma Specialists in Assessing Progressive Disc Changes From Photographs in Open-Angle Glaucoma Patients. *Arch. Ophthalmol.* **2009**, *147*, 39–44.e1. [CrossRef]
9. Nixon, G.J.; Watanabe, R.K.; Sullivan-Mee, M.; DeWilde, A.; Young, L.; Mitchell, G.L. Influence of Optic Disc Size on Identifying Glaucomatous Optic Neuropathy. *Optom. Vis. Sci.* **2017**, *94*, 654–663. [CrossRef]
10. Shi, M.; Lokhande, A.; Fazli, M.S.; Sharma, V.; Tian, Y.; Luo, Y.; Pasquale, L.R.; Elze, T.; Boland, M.V.; Zebardast, N.; et al. Artifact-Tolerant Clustering-Guided Contrastive Embedding Learning for Ophthalmic Images in Glaucoma. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4329–4340. [CrossRef]
11. Datta, S.; Mariottoni, E.B.; Dov, D.; Jammal, A.A.; Carin, L.; Medeiros, F.A. RetiNerveNet: Using recursive deep learning to estimate pointwise 24-2 visual field data based on retinal structure. *Sci. Rep.* **2021**, *11*, 12562. [CrossRef] [PubMed]
12. Kang, J.H.; Wang, M.; Frueh, L.; Rosner, B.; Wiggs, J.L.; Elze, T.; Pasquale, L.R. Cohort Study of Race/Ethnicity and Incident Primary Open-Angle Glaucoma Characterized by Autonomously Determined Visual Field Loss Patterns. *Transl. Vis. Sci. Technol.* **2022**, *11*, 21. [CrossRef]
13. Saini, C.; Shen, L.Q.; Pasquale, L.R.; Boland, M.V.; Friedman, D.S.; Zebardast, N.; Fazli, M.; Li, Y.; Eslami, M.; Elze, T.; et al. Assessing Surface Shapes of the Optic Nerve Head and Peripapillary Retinal Nerve Fiber Layer in Glaucoma with Artificial Intelligence. *Ophthalmol. Sci.* **2022**, *2*, 100161. [CrossRef]
14. Yousefi, S.; Pasquale, L.R.; Boland, M.V.; Johnson, C.A. Machine-Identified Patterns of Visual Field Loss and an Association with Rapid Progression in the Ocular Hypertension Treatment Study. *Ophthalmology* **2022**, *129*, 1402–1411. [CrossRef] [PubMed]
15. Mariottoni, E.B.; Datta, S.; Shigueoka, L.S.; Jammal, A.A.; Tavares, I.M.; Henao, R.; Carin, L.; Medeiros, F.A. Deep Learning–Assisted Detection of Glaucoma Progression in Spectral-Domain OCT. *Ophthalmol. Glaucoma* **2023**, *6*, 228–238. [CrossRef] [PubMed]
16. Mariottoni, E.B.; Jammal, A.A.; Urata, C.N.; Berchuck, S.I.; Thompson, A.C.; Estrela, T.; Medeiros, F.A. Quantification of Retinal Nerve Fibre Layer Thickness on Optical Coherence Tomography with a Deep Learning Segmentation-Free Approach. *Sci. Rep.* **2020**, *10*, 402. [CrossRef]
17. Medeiros, F.A.; Jammal, A.A.; Mariottoni, E.B. Detection of Progressive Glaucomatous Optic Nerve Damage on Fundus Photographs with Deep Learning. *Ophthalmology* **2021**, *128*, 383–392. [CrossRef]
18. Shigueoka, L.S.; Mariottoni, E.B.; Thompson, A.C.; Jammal, A.A.; Costa, V.P.; Medeiros, F.A. Predicting Age From Optical Coherence Tomography Scans with Deep Learning. *Transl. Vis. Sci. Technol.* **2021**, *10*, 12. [CrossRef]
19. Xiong, J.; Li, F.; Song, D.; Tang, G.; He, J.; Gao, K.; Zhang, H.; Cheng, W.; Song, Y.; Lin, F.; et al. Multimodal Machine Learning Using Visual Fields and Peripapillary Circular OCT Scans in Detection of Glaucomatous Optic Neuropathy. *Ophthalmology* **2022**, *129*, 171–180. [CrossRef]
20. Yousefi, S. Clinical Applications of Artificial Intelligence in Glaucoma. *J. Ophthalmic. Vis. Res.* **2023**, *18*, 97–112. [CrossRef]
21. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *npj Digit. Med.* **2021**, *4*, 65. [CrossRef]
22. Krichen, M. Convolutional Neural Networks: A Survey. *Computers* **2023**, *12*, 151. [CrossRef]
23. Shan, J.; Li, Z.; Ma, P.; Tun, T.A.; Yonamine, S.; Wu, Y.; Baskaran, M.; Nongpiur, M.E.; Chen, D.; Aung, T.; et al. Deep Learning Classification of Angle Closure based on Anterior Segment OCT. *Ophthalmol. Glaucoma* **2023**. [CrossRef] [PubMed]
24. Myburgh, J.C.; Mouton, C.; Davel, M.H. Tracking Translation Invariance in CNNs. In *Southern African Conference for Artificial Intelligence Research*; Springer International Publishing: Cham, Switzerland, 2020; pp. 282–295.
25. Sadeghzadeh, H.; Koohi, S. Translation-invariant optical neural network for image classification. *Sci. Rep.* **2022**, *12*, 17232. [CrossRef] [PubMed]

26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
27. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. *Pr. Mach. Learn. Res.* **2021**, *139*, 7358–7367.
28. Vaid, A.; Jiang, J.; Sawant, A.; Lerakis, S.; Argulian, E.; Ahuja, Y.; Lampert, J.; Charney, A.; Greenspan, H.; Narula, J.; et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digit. Med.* **2023**, *6*, 108. [CrossRef]
29. Kiyasseh, D.; Ma, R.; Haque, T.F.; Miles, B.J.; Wagner, C.; Donoho, D.A.; Anandkumar, A.; Hung, A.J. A vision transformer for decoding surgeon activity from surgical videos. *Nat. Biomed. Eng.* **2023**, *7*, 780–796. [CrossRef]
30. Liu, H.; Teng, L.; Fan, L.; Sun, Y.; Li, H. A new ultra-wide-field fundus dataset to diabetic retinopathy grading using hybrid preprocessing methods. *Comput. Biol. Med.* **2023**, *157*, 106750. [CrossRef]
31. Playout, C.; Duval, R.; Boucher, M.C.; Cheriet, F. Focused Attention in Transformers for interpretable classification of retinal images. *Med. Image Anal.* **2022**, *82*, 102608. [CrossRef]
32. Yu, S.; Ma, K.; Bi, Q.; Bian, C.; Ning, M.; He, N.; Li, Y.; Liu, H.; Zheng, Y. MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part VIII 24*; Springer International Publishing: Cham, Switzerland, 2021; pp. 45–54.
33. Kihara, Y.; Shen, M.; Shi, Y.; Jiang, X.; Wang, L.; Laiginhas, R.; Lyu, C.; Yang, J.; Liu, J.; Morin, R.; et al. Detection of Nonexudative Macular Neovascularization on Structural OCT Images Using Vision Transformers. *Ophthalmol. Sci.* **2022**, *2*, 100197. [CrossRef] [PubMed]
34. Li, A.L.; Feng, M.; Wang, Z.; Baxter, S.L.; Huang, L.; Arnett, J.; Bartsch, D.-U.G.; Kuo, D.E.; Saseendrakumar, B.R.; Guo, J.; et al. Automated Detection of Posterior Vitreous Detachment on OCT Using Computer Vision and Deep Learning Algorithms. *Ophthalmol. Sci.* **2023**, *3*, 100254. [CrossRef] [PubMed]
35. Philippi, D.; Rothaus, K.; Castelli, M. A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images. *Sci. Rep.* **2023**, *13*, 517. [CrossRef]
36. Xuan, M.; Wang, W.; Shi, D.; Tong, J.; Zhu, Z.; Jiang, Y.; Ge, Z.; Zhang, J.; Bulloch, G.; Peng, G.; et al. A Deep Learning–Based Fully Automated Program for Choroidal Structure Analysis Within the Region of Interest in Myopic Children. *Transl. Vis. Sci. Technol.* **2023**, *12*, 22. [CrossRef] [PubMed]
37. Sivaswamy, J.; Krishnadas, S.R.; Joshi, G.D.; Jain, M.; Tabish, A.U.S. Drishti-Gs: Retinal Image Dataset for Optic Nerve Head(ONH) Segmentation. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April–2 May 2014; pp. 53–56. Available online: https://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php (accessed on 16 October 2023).
38. sjchoi86. sjchoi86-HRF. Available online: https://github.com/yiweichen04/retina_dataset (accessed on 20 October 2016).
39. Fumero, F.; Diaz-Aleman, T.; Sigut, J.; Alayon, S.; Arnay, R.; Angel-Pereira, D. Rim-One Dl: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning. *Image Anal. Ster.* **2020**, *39*, 161–167. [CrossRef]
40. Zhang, Z.; Yin, F.S.; Liu, J.; Wong, W.K.; Tan, N.M.; Lee, B.H.; Cheng, J.; Wong, T.Y. ORIGA(-light): An online retinal fundus image database for glaucoma analysis and research. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2010**, *2010*, 3065–3068. Available online: https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection (accessed on 16 October 2023). [CrossRef]
41. Diaz-Pinto, A.; Morales, S.; Naranjo, V.; Köhler, T.; Mossi, J.M.; Navea, A. CNNs for automatic glaucoma assessment using fundus images: An extensive validation. *Biomed. Eng. Online* **2019**, *18*, 29. [CrossRef]
42. Fang, H.; Li, F.; Wu, J.; Fu, H.; Sun, X.; Son, J.; Yu, S.; Zhang, M.; Yuan, C.; Bian, C. REFUGE2 Challenge: A Treasure Trove for Multi-Dimension Analysis and Evaluation in Glaucoma Screening. *arXiv* **2022**, arXiv:2202.08994. Available online: https://ai.baidu.com/broad/download (accessed on 16 October 2023).
43. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Cham, Switzerland, 8–14 September 2018; pp. 833–851.
44. *Vision & Eye Health Surveillance System. Review: Glaucoma*. Available online: https://www.cdc.gov/visionhealth/vehss/data/studies/glaucoma.html (accessed on 12 October 2023).
45. Dhillon, A.; Verma, G.K. Convolutional neural network: A review of models, methodologies and applications to object detection. *Prog. Artif. Intell.* **2019**, *9*, 85–112. [CrossRef]
46. Wang, P.; Yuan, M.; He, Y.; Sun, J. 3D augmented fundus images for identifying glaucoma via transferred convolutional neural networks. *Int. Ophthalmol.* **2021**, *41*, 2065–2072. [CrossRef]
47. Rogers, T.W.; Jaccard, N.; Carbonaro, F.; Lemij, H.G.; Vermeer, K.A.; Reus, N.J.; Trikha, S. Evaluation of an AI system for the automated detection of glaucoma from stereoscopic optic disc photographs: The European Optic Disc Assessment Study. *Eye* **2019**, *33*, 1791–1797. [CrossRef] [PubMed]
48. Jammal, A.A.; Thompson, A.C.; Mariottoni, E.B.; Berchuck, S.I.; Urata, C.N.; Estrela, T.; Wakil, S.M.; Costa, V.P.; Medeiros, F.A. Human Versus Machine: Comparing a Deep Learning Algorithm to Human Gradings for Detecting Glaucoma on Fundus Photographs. *Arch. Ophthalmol.* **2020**, *211*, 123–131. [CrossRef] [PubMed]
49. Thompson, A.C.; Jammal, A.A.; Medeiros, F.A. A Deep Learning Algorithm to Quantify Neuroretinal Rim Loss From Optic Disc Photographs. *Arch. Ophthalmol.* **2019**, *201*, 9–18. [CrossRef]

50. Ma, J.; Bai, Y.; Zhong, B.; Zhang, W.; Yao, T.; Mei, T. Visualizing and Understanding Patch Interactions in Vision Transformer. In *IEEE Transactions on Neural Networks and Learning Systems*; IEEE: New York, NY, USA, 2023; pp. 1–10. [CrossRef]
51. Fan, R.; Alipour, K.; Bowd, C.; Christopher, M.; Brye, N.; Proudfoot, J.A.; Goldbaum, M.H.; Belghith, A.; Girkin, C.A.; Fazio, M.A.; et al. Detecting Glaucoma from Fundus Photographs Using Deep Learning without Convolutions: Transformer for Improved Generalization. *Ophthalmol. Sci.* **2023**, *3*, 100233. [CrossRef]
52. Song, D.; Fu, B.; Li, F.; Xiong, J.; He, J.; Zhang, X.; Qiao, Y. Deep Relation Transformer for Diagnosing Glaucoma With Optical Coherence Tomography and Visual Field Function. *IEEE Trans. Med. Imaging* **2021**, *40*, 2392–2402. [CrossRef] [PubMed]
53. Hou, K.; Bradley, C.; Herbert, P.; Johnson, C.; Wall, M.; Ramulu, P.Y.; Unberath, M.; Yohannan, J. Predicting Visual Field Worsening with Longitudinal OCT Data Using a Gated Transformer Network. *Ophthalmology* **2023**, *130*, 854–862. [CrossRef]
54. Yi, Y.; Jiang, Y.; Zhou, B.; Zhang, N.; Dai, J.; Huang, X.; Zeng, Q.; Zhou, W. C2FTFNet: Coarse-to-fine transformer network for joint optic disc and cup segmentation. *Comput. Biol. Med.* **2023**, *164*, 107215. [CrossRef]
55. Zhang, Y.; Li, Z.; Nan, N.; Wang, X. TranSegNet: Hybrid CNN-Vision Transformers Encoder for Retina Segmentation of Optical Coherence Tomography. *Life* **2023**, *13*, 976. [CrossRef]
56. Vali, M.; Mohammadi, M.; Zarei, N.; Samadi, M.; Atapour-Abarghouei, A.; Supakontanasan, W.; Suwan, Y.; Subramanian, P.S.; Miller, N.R.; Kafieh, R.; et al. Differentiating Glaucomatous Optic Neuropathy From Non-glaucomatous Optic Neuropathies Using Deep Learning Algorithms. *Arch. Ophthalmol.* **2023**, *252*, 1–8. [CrossRef]
57. Christopher, M.; Belghith, A.; Bowd, C.; Proudfoot, J.A.; Goldbaum, M.H.; Weinreb, R.N.; Girkin, C.A.; Liebmann, J.M.; Zangwill, L.M. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci. Rep.* **2018**, *8*, 16685. [CrossRef]
58. Luo, X.; Hu, M.; Song, T.; Wang, G.; Zhang, S. Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer. In Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022; pp. 820–833.
59. Li, A.; Cheng, J.; Wong, D.W.K.; Liu, J. Integrating holistic and local deep features for glaucoma classification. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 1328–1331.
60. Li, L.; Xu, M.; Liu, H.; Li, Y.; Wang, X.; Jiang, L.; Wang, Z.; Fan, X.; Wang, N. A Large-Scale Database and a CNN Model for Attention-Based Glaucoma Detection. *IEEE Trans. Med. Imaging* **2020**, *39*, 413–424. [CrossRef] [PubMed]