AMERICAN ACADEMY OF OPHTHALMOLOGY®

Comparative Analysis of Vision Transformers and Conventional Convolutional Neural Networks in Detecting Referable Diabetic Retinopathy

Jocelyn Hui Lin Goh, BEng,[1,*] *Elroy Ang, BEng,*[2,*] *Sahana Srinivasan, BEng,*[1] *Xiaofeng Lei, MSc,*[3] *Johnathan Loh, MEng,*[1] *Ten Cheer Quek, BEng,*[1] *Cancan Xue, PhD,*[1] *Xinxing Xu, PhD,*[3] *Yong Liu, PhD,*[3] *Ching-Yu Cheng, PhD,*[1,4,5,6] *Jagath C. Rajapakse, PhD,*[2] *Yih-Chung Tham, PhD*[1,4,5,6]

Objective: Vision transformers (ViTs) have shown promising performance in various classification tasks previously dominated by convolutional neural networks (CNNs). However, the performance of ViTs in referable diabetic retinopathy (DR) detection is relatively underexplored. In this study, using retinal photographs, we evaluated the comparative performances of ViTs and CNNs on detection of referable DR.

Design: Retrospective study.

Participants: A total of 48 269 retinal images from the open-source Kaggle DR detection dataset, the Messidor-1 dataset and the Singapore Epidemiology of Eye Diseases (SEED) study were included.

Methods: Using 41 614 retinal photographs from the Kaggle dataset, we developed 5 CNN (Visual Geometry Group 19, ResNet50, InceptionV3, DenseNet201, and EfficientNetV2S) and 4 ViTs models (VAN_small, CrossViT_small, ViT_small, and Hierarchical Vision transformer using Shifted Windows [SWIN]_tiny) for the detection of referable DR. We defined the presence of referable DR as eyes with moderate or worse DR. The comparative performance of all 9 models was evaluated in the Kaggle internal test dataset (with 1045 study eyes), and in 2 external test sets, the SEED study (5455 study eyes) and the Messidor-1 (1200 study eyes).

Main Outcome Measures: Area under operating characteristics curve (AUC), specificity, and sensitivity.

Results: Among all models, the SWIN transformer displayed the highest AUC of 95.7% on the internal test set, significantly outperforming the CNN models (all $P < 0.001$). The same observation was confirmed in the external test sets, with the SWIN transformer achieving AUC of 97.3% in SEED and 96.3% in Messidor-1. When specificity level was fixed at 80% for the internal test, the SWIN transformer achieved the highest sensitivity of 94.4%, significantly better than all the CNN models (sensitivity levels ranging between 76.3% and 83.8%; all $P < 0.001$). This trend was also consistently observed in both external test sets.

Conclusions: Our findings demonstrate that ViTs provide superior performance over CNNs in detecting referable DR from retinal photographs. These results point to the potential of utilizing ViT models to improve and optimize retinal photo-based deep learning for referable DR detection.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100552*

Supplemental material available at www.ophthalmologyscience.org.

In recent years, artificial intelligence (AI) in health care has experienced a rapid expansion of its applications, particularly in the domain of medical imaging. Specifically, deep learning algorithms such as convolutional neural networks (CNNs) have shown promising results in numerous medical image classification, segmentation, and diagnosis tasks.[1–4] The popularity of CNNs stems from their ability to learn high-level representations from vast datasets,[5] and it became the predominant approach in the detection of diabetic retinopathy (DR), where identifying pathological features of DR on the retina is crucial.[6–9]

Meanwhile, the growing availability of big data and continuous technological advancements paved the way for the emergence of newer deep learning techniques. New types of neural networks such as transformers have evolved from their original use in natural language processing tasks[10] to vision transformers (ViTs) for visual recognition and image classification tasks.[11] Vision transformers have

https://doi.org/10.1016/j.xops.2024.100552
ISSN 2666-9145/24

demonstrated superior performance over conventional CNNs in various large-scale image classification tasks with ImageNet.[12] However, the potential of the ViT for detection of DR is yet to be fully explored.

Recent literature evaluating CNNs and ViTs in DR detection predominantly points to the superior performance of ViTs.[13–15] Nevertheless, it is important to note that these earlier studies often rely on relatively small and imbalanced datasets. In a separate study conducted by Wu et al, the authors utilized a larger training dataset comprising 30 000 retinal images from the Kaggle DR detection competition.[15] However, data augmentation was performed on the training dataset with large proportion of poor quality retinal images. The inclusion of a significant number of low-quality images through data augmentation potentially compromised the model's performance. Notably, none of these past studies conducted external validation, making it challenging to gauge the relative efficacy of ViTs compared with CNNs in referable DR.

Given this backdrop, our study aims to provide a comprehensive comparison between ViTs and CNNs for referable DR detection. We utilized a single, well-curated, and standardized training dataset and further validated these models using external datasets.

Methods

An overview of the methodology is depicted in a flowchart in Figure 1.

Datasets

We used a total of 41 614 good quality macular-centered retinal images obtained from the Kaggle DR detection competition in 2015 to develop all our models. The original full Kaggle dataset consists of 88 702 images of which a large proportion are unclean or of poor quality. Hence, we performed automated quality check to filter and exclude retinal images with insufficient quality. Among the 41 614 good quality retinal images included, they were split into training, validation, and internal test set, of 39 531, 1038, and 1045 images respectively. To avoid overfitting the models, the dataset split was done at individual level so that there was no overlapping of individuals in the training and internal test sets. We further ensured consistent ratio of eyes with nonreferable DR and eyes with referable DR in all training, validation, and internal test sets. The number of eyes used in this study is summarized in Table 1.

Additionally, we evaluated the models in 2 other external datasets, the Singapore Epidemiology of Eye Diseases (SEED) dataset and Messidor-1 dataset.[16,17] The SEED study comprises of 3 Singaporean ethnic population-based cohort studies, which are the Singapore Malay Eye study, the Singapore Chinese Eye study, and the Singapore Indian Eye study. Participants' written informed consent was obtained in each study. All studies adhered to the tenets of the Declaration of Helsinki and had local ethical committee approval (SingHealth Centralised Instituitional Review Board, R1107/9/2014 and R498/47/2006). From the SEED dataset, we included a set of 5455 macula-centered retinal images from 2855 participants across all 3 cohorts; 55.8% of the individuals included in our external test set were of ≥60 years old. Of the 2855 participants, 993 participants were ethnic Malays, 577 individuals were ethnic Chinese, and 1287 individuals were ethnic Indians (Table S2, available at www.ophthalmologyscience.org). From Messidor-1, we included 1200 macula-centered retinal images. The SEED study used the Canon CR-1 Mark-II nonmydriatic digital retinal camera, the Messidor-1 study used Topcon TRC NW6 nonmydriatic retinal camera, and the Kaggle dataset primarily consists of retinal images captured using various retinal cameras.

DR Grading

In the training, validation and internal test sets, the retinal images were graded for presence of DR into 5 severity classes: normal, mild, moderate, severe, and proliferate. In

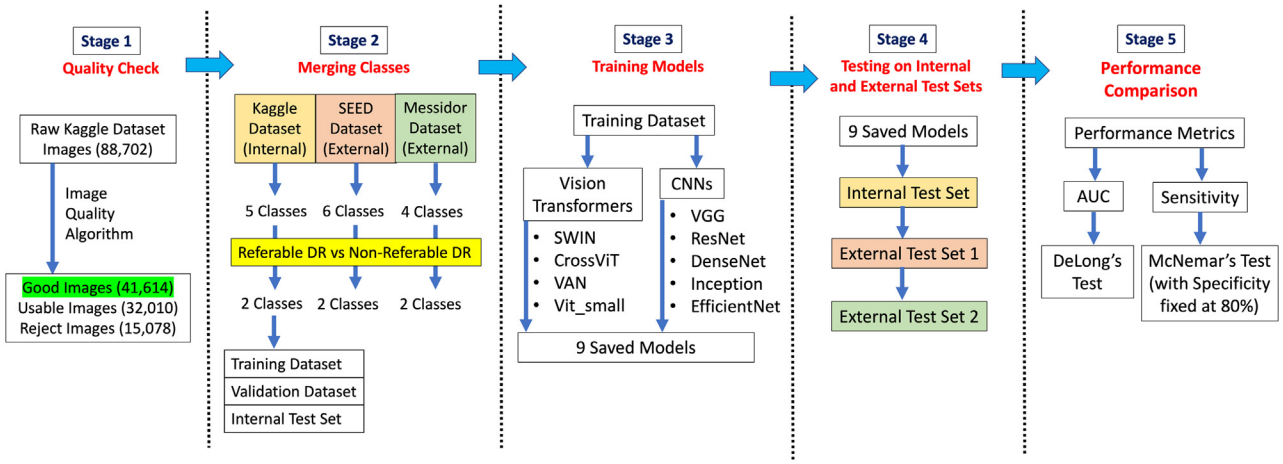

Figure 1. Overview of methods. AUC = area under curve; CNNs = convolutional neural networks; DR = diabetic retinopathy; SEED = Singapore Epidemiology of Eye Diseases; SWIN = Hierarchical Vision transformer using Shifted Windows; VAN = Visual Attention Network; VGG = Visual Geometry Group; ViT = vision transformer.

Table 1. Training and Test Dataset Characteristics

	Kaggle Training and Internal Test Dataset	SEED External Test Dataset	Messidor-1 External Test Dataset
No. of study eye	41 614	5455	1200
No. of unique individuals	25 536	2855	NA
Referable diabetic retinopathy	6436	545	501
Nonreferable diabetic retinopathy	35 178	4910	699

NA = not available; SEED = Singapore Epidemiology of Eye Diseases.

our study, the images were reclassified into 2 classes, nonreferable and referable DR, for binary classification. The nonreferable DR images consist of images with normal and mild DR grades while referable DR images consist of images with moderate, severe, and proliferate DR grades.

Similarly, we also reclassified the images in both external validation datasets based on their respective DR grading into 2 classes. The SEED dataset has 6 classes for severity of DR: normal, minimal, mild, moderate, severe, and proliferate. In the nonreferable class, we included images of normal, minimal, and mild DR grades, while for the referable DR class, we included images of moderate, severe, and proliferate DR grades. On the other hand, the Messidor-1 dataset has 4 classes for severity of DR: 0, 1, 2, and 3 (with grade 0 being normal and grade 3 being the most severe). We included images with grade 0 and 1 as nonreferable DR and images with grade 2 and 3 as referable DR in the Messidor-1 external test set. The proportion of images by DR severity class for all training and test datasets is summarized in Table S3 (available at www.ophthalmologyscience.org).

Model Development

In total, there were 9 models included for performance comparison in our evaluations. We have chosen 5 CNN and 4 ViT models. The CNN models used were Visual Geometry Group (VGG)-19,[18] ResNet50,[12] InceptionV3,[19] DenseNet201,[20] and EfficientNetV2S.[21] The ViT models used were Visual Attention Network (VAN_small),[22] CrossViT_small,[23] ViT_small,[11] and SWIN_tiny transformer.[24]

All the models chosen were loaded with weights that were pretrained on ImageNet for a fair and robust comparison. This is because training from scratch may not obtain a well-tuned performance for each architecture, thus compromising the results used for comparison. All the model scales are also similar in the number of parameters they have, approximately around 20 million parameters. To prevent overfitting our models, we employed early stopping at 20 epochs when no further improvements in models' test performance were observed. We also adopted checkpoints to save the models when the models achieve highest area under receiver operating characteristics curve (AUC) in the internal validation dataset. Furthermore, to minimize the effects of class imbalance in our training dataset on the models' performance, we assigned class weights to both the nonreferable DR and referable DR classes.

Statistical Analysis

For each image, the models generate 2 sets of continuous probability output values (from 0 to 1) corresponding to the probability of the image having nonreferable DR and referable DR. Based on the probability scores, we adopted different classification thresholds to determine the predicted binary class for each retinal image, with 0 representing nonreferable DR and 1 for referable DR. For our evaluation, we used AUC to compare the performance between each of our models. We performed the DeLong test between the models' AUC performances to determine significant differences and applied the Bonferroni correction to account for the multiple comparisons.

We also evaluated the sensitivity values of the model after setting a unique threshold for all models based on 0.8 specificity. With specificity fixed at 80%, we performed the McNemar test to determine significant difference in sensitivity values between the models.

Results

Using 9 different models comprising 5 CNN and 4 ViT architectures, we evaluated the performances of models on an internal and 2 independent external test sets.

In the Kaggle internal test set, the SWIN transformer model achieved the highest AUC of 95.7%. The SWIN transformer model performed significantly better than all 5 CNN models, with AUC ranging from 86.1% (VGG model) to 89.4% (ResNet model), and better than the VAN_small transformer model (AUC of 90.7%). There was no significant difference in AUC performance of SWIN transformer model and the other 2 transformer models, ViT_small model (AUC of 94.5%) and CrossViT model (AUC of 93.8%).

Upon validation of our models in 2 external test sets, we observed similar findings as from the internal test set. The SWIN transformer model achieved the highest AUC in both SEED (AUC of 97.3%) and Messidor-1 (AUC of 96.3%). In the SEED test set, the SWIN transformer performed significantly better than all the 5 CNN models which had AUC ranging from 91.7% (Inception model) to 94.1% (EfficientNet model). The SWIN transformer model also achieved significantly higher AUC compared with the other 3 transformer models, CrossViT model (AUC of 94.5%), VAN_small model (AUC of 95.2%), and ViT_small model (AUC of 95.5%). Similarly, in the Messidor-1 test set, the SWIN transformer model significantly outperformed all 5

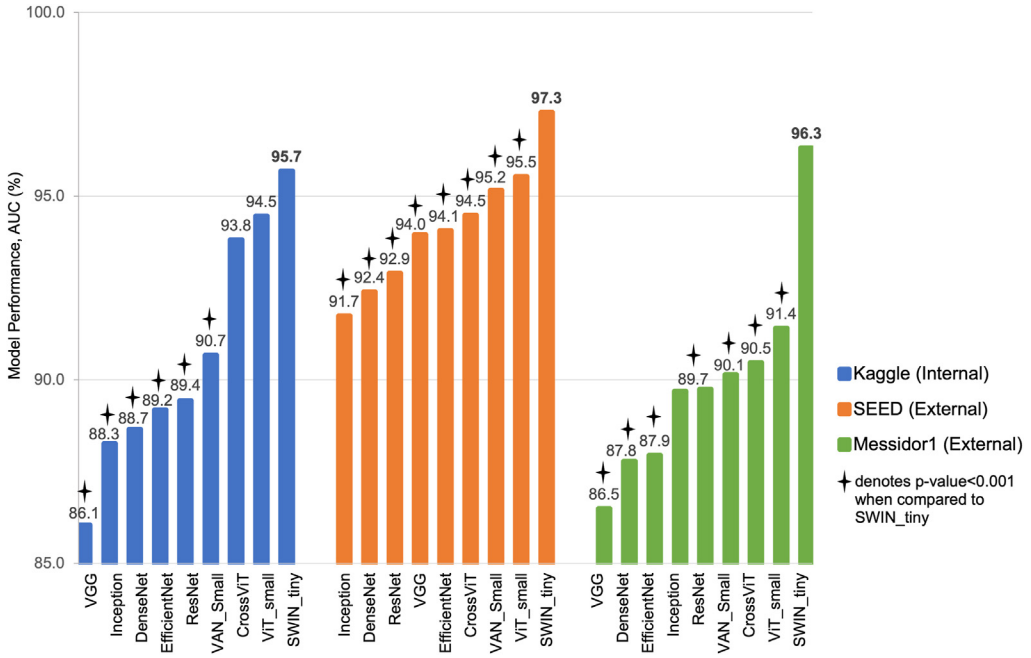

Figure 2. Comparison of AUC between ViT models and CNN models in all test sets. AUC = area under curve; CNN = convolutional neural network; SEED = Singapore Epidemiology of Eye Diseases; SWIN = Hierarchical Vision transformer using Shifted Windows; VAN = Visual Attention Network; VGG = Visual Geometry Group; ViT = vision transformer.

Table 4. Maximum F1 Score for all 9 Models Internal and External Test Datasets

Test Dataset	Model	AUC (%)	$F1_{max}$	Recall (%)	Precision (%)	TP	TN	FP	FN	Threshold*
Kaggle (internal)	VGG19	86.1	0.68	63.8	72.3	102	846	39	58	0.651
	ResNet50	89.4	0.69	66.9	71.3	107	842	43	53	0.434
	DenseNet201	88.7	0.70	61.9	81.2	99	862	23	61	0.706
	InceptionV3	88.3	0.68	71.9	63.9	115	820	65	45	0.203
	EfficientNetV25	89.2	0.69	60.6	78.9	97	859	26	63	0.775
	VAN_Small	90.7	0.68	68.1	68.6	109	835	50	51	0.114
	SWIN_Tiny	95.7	**0.80**	83.1	76.4	133	844	41	27	0.690
	CrossViT_Small	93.8	0.74	74.4	73.0	119	841	41	41	0.704
	ViT_Small	94.5	0.76	72.5	80.0	116	856	29	44	0.840
SEED (External)	VGG19	94.0	0.69	67.5	70.1	368	4753	157	177	0.992
	ResNet50	92.4	0.63	62.9	62.1	343	4701	209	202	0.986
	DenseNet201	92.9	0.63	60.7	66.1	331	4740	170	214	0.984
	InceptionV3	91.7	0.65	61.1	70.1	333	4768	142	212	0.974
	EfficientNetV25	94.1	0.67	66.7	68.4	363	4742	168	182	0.991
	VAN_Small	95.2	0.71	71.4	69.7	389	4741	169	156	0.946
	SWIN_Tiny	97.3	**0.79**	79.5	78.2	433	4789	121	112	0.990
	CrossViT_Small	94.5	0.70	69.2	71.7	377	4761	149	168	0.976
	ViT_Small	95.5	0.74	73.4	74.1	400	4770	140	145	0.926
Messidor (External)	VGG19	86.5	0.77	71.5	82.7	358	624	75	143	0.547
	ResNet50	89.7	0.80	75.1	85.1	376	633	66	125	0.244
	DenseNet201	87.8	0.77	78.8	75.5	395	571	128	106	0.143
	InceptionV3	89.7	0.80	73.9	87.1	370	644	55	131	0.646
	EfficientNetV25	87.9	0.78	76.5	80.0	383	603	96	118	0.198
	VAN_Small	90.1	0.79	80.0	78.8	401	591	108	100	0.084
	SWIN_Tiny	96.3	**0.90**	87.8	91.5	440	658	41	61	0.810
	CrossViT_Small	90.5	0.81	81.0	81.2	406	605	94	95	0.623
	ViT_Small	91.4	0.82	79.6	84.9	399	628	71	102	0.691

AUC = area under curve; $F1_{max}$ = maximum F1 score; FN = false negative; FP = false positive; SEED = Singapore Epidemiology of Eye Diseases; TN = true negative; TP = true positive; SWIN = Hierarchical Vision transformer using Shifted Windows; VAN = Visual Attention Network; VGG = Visual Geometry Group; ViT = vision transformer.
Bold values represent the model with the highest $F1_{max}$ value in the respective test dataset.
*Threshold based on maximum F1 score.

CNN models with AUC ranging between 86.5% (VGG model) and 89.7% (Inception and ResNet models), and also 3 other transformer models with AUC between 90.1% (VAN_small model) and 91.4% (ViT_small model). Figure 2 shows the comparison of AUC values between the models across all test sets. Table 4 summarized the performance of the models in terms of F1 scores, and the corresponding precision and recall values were computed using threshold based on F1 score maximization. The SWIN transformer model obtained the highest maximum F1 score of 0.80 in the Kaggle internal test set, 0.79 in SEED, and 0.90 in Messidor-1.

Additionally, we evaluated the sensitivity values of the models with specificity level fixed at 80%. The best performing model in the Kaggle internal test set was the SWIN transformer, achieving a sensitivity of 94.4%. It performed significantly better than all 5 CNN models, with sensitivity ranging from 76.3% (VGG model) to 83.8% (ResNet model). The SWIN transformer is also significantly different from the VAN_small transformer model (sensitivity of 83.8%), but there was no significant difference in between the other 2 transformer models, ViT_small model (sensitivity of 91.9%) and CrossViT model (sensitivity of 90.0%).

In external validation, the SWIN transformer model achieved the highest sensitivity in both SEED (96.9%) and Messidor-1 (94.8%). In the SEED test set, the SWIN transformer performed significantly better than all 5 CNN models with sensitivity values ranging from 87.2% (Inception model) to 91.7% (EfficientNet model), and better than the CrossViT model (92.3%). There was no significant difference in sensitivity of SWIN transformer model and the other 2 transformer models, ViT_small model (93.9%) and VAN_small model (93.6%) for SEED test set. In the Messidor-1 test set, the SWIN transformer model significantly outperformed all other models, including the 5 CNN models and the other 3 transformer models. The sensitivity level ranged from 77.4% (VGG model) to 84.2% (ViT_small). Figure 3 shows the comparison of sensitivity levels between the models across all test sets.

Discussion

In this study, we designed and rigorously evaluated the performance of 9 distinct models, namely, 5 CNNs and 4 transformers, in detecting referable DR using only retinal photographs. Across both internal (Kaggle) and external test datasets (SEED and Messidor-1), we consistently observed that the SWIN transformer model had superior AUC and sensitivity performance compared with the CNNs. Our finding builds on the existing body of evidence demonstrating that transformer models excel better than traditional CNN models in medical image classifications but specifically underscoring their potential to perform optimally in detecting referable DR. As the field of automated DR detection continues to evolve, our findings could offer valuable insights for the refinement and development of more effective, AI-powered screening and diagnostic tools for DR.

In a post hoc sensitivity analysis, to evaluate the impact of image quality on the models' performance, we added back 200 poor quality images into the internal test set (approximately 20% of the 1045 retinal images in the internal test set) and assessed the models' AUC and sensitivity levels when specificity level was fixed at 80% (see Table S5,

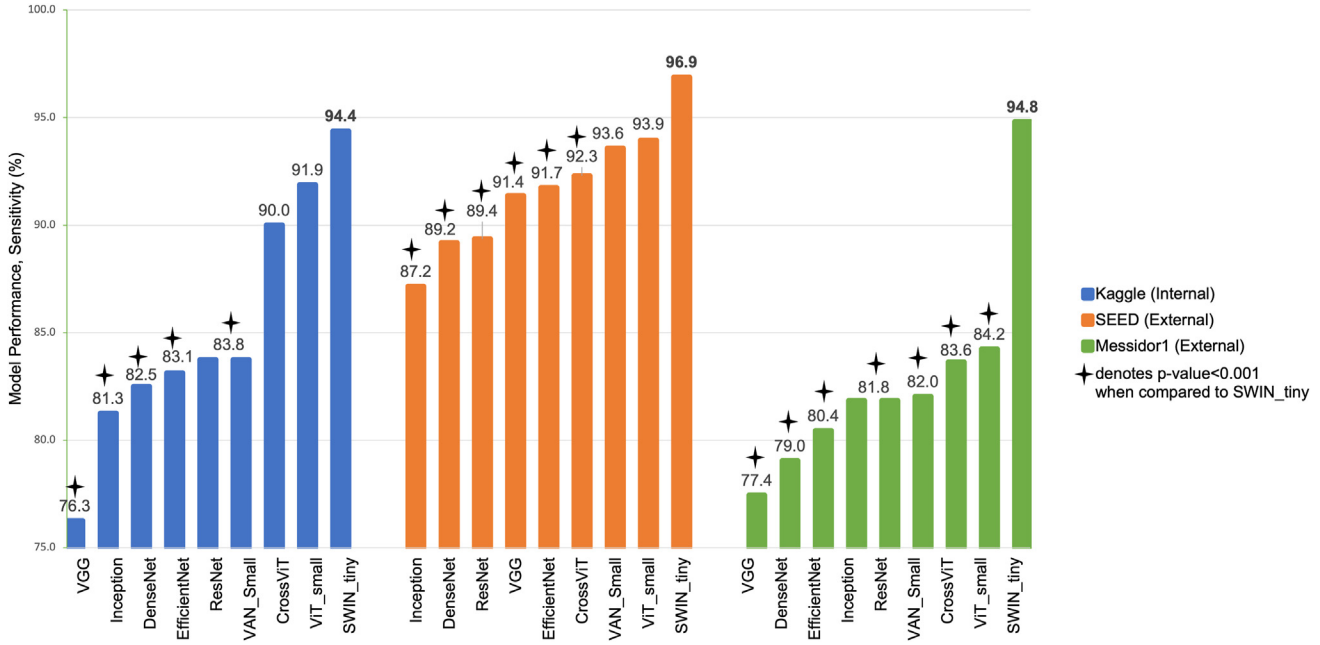

Figure 3. Comparison of sensitivity levels (with specificity fixed at 80%) between ViT models and CNN models in all test sets. CNN = convolutional neural network; SEED = Singapore Epidemiology of Eye Diseases; SWIN = Hierarchical Vision transformer using Shifted Windows; VAN = Visual Attention Network; VGG = Visual Geometry Group; ViT = vision transformer.

available at www.ophthalmologyscience.org). Overall, in this sensitivity analysis, we observed consistent findings that the ViT models performed better than CNN models, especially the SWIN transformer model.

Additionally, when we evaluated the performance of all models in subgroups of aged below and >60 years old in the SEED external test set (Table S6, available at www.ophthalmologyscience.org), it was consistently observed that the ViT models performed better than CNN models, with the SWIN transformer model achieving the higher AUC than the CNN models. This trend was also consistent in 2 other subgroup analyses by gender and ethnicity (Table S7 and S8, available at www.ophthalmologyscience.org).

There are several possible factors contributing to the SWIN transformer model's superior performance over the CNN models. Firstly, the SWIN transformer adopts the method of efficient hierarchical representation learning in the training process. Thus, it is designed to efficiently learn hierarchical representations of images by dividing the input image into smaller patches, which are then processed by the transformer layers in a hierarchical manner.[24] This allows the model to capture both local and global features of the image effectively, which is important for image classification tasks. Secondly, the SWIN transformer also uses the shifted window attention mechanism that allows the model to bridge the windows of the preceding layers. This in turn provides connections among layers, significantly enhancing modeling power.[24] Lastly, the SWIN transformer requires less data than CNNs to achieve state-of-the-art performance in image classification tasks because it is better at leveraging the available data by learning more efficient representations.[24] Since DR detection involves identifying often small pathological changes on retinal images, the SWIN transformer model's ability to capture both local and global features on images effectively likely enabled the model to achieve better performance in DR detection.

While the AI technology continues to advance and new model backbones emerge, there are still many challenges associated with the downstream deployment of AI-based tools in actual clinical settings. One of the barriers to implement and integrate AI-tools with existing clinical workflow is the lack of evidence on the efficacy of AI interventions in prospective clinical trials, especially multicenter trials.[25] Specifically, the generalizability of the models' performance in actual, clinical settings is still not well studied. To this end, application of implementation science frameworks may help to further guide integration of AI tools into existing health care systems and workflows.[25,26]

This study compares the performances of CNN and transformer models for detection of referable DR. Unlike past studies, which often relied on small, imbalanced datasets and lacked external validation, our work employed the Kaggle dataset for training and leveraged 2 robust external datasets for validation (SEED and Messidor-1). These 2 external datasets are particularly valuable because of their well-defined DR grading protocols with detailed assessment of lesions on retinal images.[27–29] Studies in the past have only included small datasets with data imbalance in their evaluation and many lacked external validation. This may have inevitably affected the reliability of their findings. The detailed procedures we have taken to minimize potential model bias were challenging but important to ensure the reliability of the findings.

Our study is not without limitations. First, because of computational constraints, we selected models of similar parameter scales for performance comparison, approximately 20 million parameters. While this allowed for a more manageable evaluation, it also restricted our ability to determine the optimal model scale for each architectural type in the context of this specific image classification task. As such, comprehensive testing involving parameter optimization and fine-tuning with data from actual settings is necessary to fully assess the capabilities of newer transformer models relative to established CNN models. Second, the Kaggle dataset is open-source, and while this accessibility is an advantage, it also introduces the potential for mislabeling that cannot be entirely ruled out. However, the dataset's considerable size (approximately 41 000 images) may serve to mitigate or offset the effects of any such labeling inaccuracies. Third, the presence of confounding factors and potential biases in our single training dataset could limit the generalizability of our study findings to other external datasets. It is vital to ensure the continuous model validation in larger and more diverse datasets such that the models' applications can be further optimized. Future study in this aspect is warranted.

In conclusion, our findings suggest that ViTs, particularly the SWIN transformer model, offer superior performance in the detection of referable DR from retinal photographs, compared with conventional CNNs. The SWIN transformer model consistently outperformed all evaluated CNN models across internal and external test sets. This finding underscores the potential of ViTs to refine existing DR detection algorithms and serve as a basis for future advancements in the field.

Availability of Data

The SEED dataset contains retinal images and patient information and is not publicly available due to patient privacy. On reasonable request, deidentified individual-participant data from the SEED dataset may be made available for academic research purposes from the principal investigator (Prof. Ching-Yu Cheng), subject to permission from the local institutional review board. The Kaggle and Messidor-1 datasets are publicly available and can be assessed online through the following links: Kaggle (https://www.kaggle.com/c/diabetic-retinopathy-detection/data) and Messidor-1 (https://www.adcis.net/en/third-party/messidor/).

Availability of Codes

The test codes for all the models can be assessed on GitHub [https://github.com/SERI-EPI-DS/CNN_vs_ViT]). Custom codes can be made available for research purpose from the corresponding author (Assistant Professor Yih-Chung

Tham) upon reasonable request. All requests for code will be reviewed by the SingHealth Intellectual Property Unit and the NUS Intellectual Property Office, to verify whether the request is subject to any IP or confidentiality constraints.

Acknowledgments

The Messidor dataset was kindly provided by the Messidor program partners (see https://www.adcis.net/en/third-party/messidor/).

Footnotes and Disclosures

Originally received: January 2, 2024.
Final revision: May 9, 2024.
Accepted: May 13, 2024.
Available online: May 17, 2024. Manuscript no. XOPS-D-24-00002.

[1] Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore.

[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore.

[3] Institute of High-Performance Computing, A*STAR, Singapore, Singapore.

[4] Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.

[5] Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore.

[6] Ophthalmology and Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School Singapore, Singapore, Singapore.

*J.H.L.G. and E.A. are contributed equally to this work.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

Y.C.T. is supported by the National Medical Research Council's HPHSR Clinician Scientist Award (NMRC/MOH/HCSAINV21nov-0001). The sponsor or funding organization had no role in the design or conduct of this research. This project is supported by the Agency for Science, Technology and Research (A*STAR) under its RIE2020 Health and Biomedical Sciences (HBMS) Industry Alignment Fund Pre-Positioning (IAF-PP) grant no. H20c6a0031. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the A*STAR.

HUMAN SUBJECTS: Human subjects were included in this study. All studies adhered to the tenets of the Declaration of Helsinki and had local ethical commitee approval (SingHealth Centralised Instituitional Review Board, R1107/9/2014 and R498/47/2006). Participants' written informed consent was obtained in each study.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Goh, Ang, Lei, Xu, Liu, Cheng, Rajapakse, Tham

Data collection: Ang, Loh, Quek, Xue, Cheng, Tham

Analysis and interpretation: Goh, Ang, Srinivasan, Tham

Obtained funding: Xu, Cheng, Tham

Overall responsibility: Rajapakse, Tham

Abbreviations and Acronyms:
AI = artificial intelligence; **AUC** = area under receiver operating characteristics curve; **CNN** = convolutional neural network; **DR** = diabetic retinopathy; **SEED** = Singapore Epidemiology of Eye Diseases; **SWIN** = Hierarchical Vision transformer using Shifted Windows; **VAN** = Visual Attention Network; **VGG** = Visual Geometry Group; **ViTs** = vision transformers.

Keywords:
Convolutional neural network, Referable diabetic retinopathy, Retinal photographs, Vision transformer.

Correspondence:
Yih-Chung Tham, Yong Loo Lin School of Medicine, National University of Singapore, Level 13, MD1 Tahir Foundation Building, 12 Science Drive 2, Singapore 117549. E-mail: thamyc@nus.edu.sg; and Jagath C. Rajapakse, PhD, School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore. E-mail: ASJagath@ntu.edu.sg.

References

1. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60−88.
2. Burlina PM, Joshi N, Pekala M, et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135(11):1170−1176.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature*. 2017;542(7639):115−118.
4. Cheng J-Z, Ni D, Chou Y-H, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci Rep*. 2016;6(1):24454.
5. Fast r-cnn. In: Girshick R, ed. *Proceedings of the IEEE international conference on computer vision*. Washington, DC: Institute of Electrical and Electronics Engineers (IEEE) COmputer Society; 2015.
6. Alyoubi WL, Shalash WM, Abulkhair MF. Diabetic retinopathy detection through deep learning techniques: a review. *Inform Med Unlocked*. 2020;20:100377.
7. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131(3):351−357.
8. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402−2410.
9. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211−2223.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.
11. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2010.11929.
12. He K, Zhang X, Ren S, Sun J, eds. *Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV: Institute of Electrical and Electronics Engineers (IEEE); 2016.

13. Wu J, Hu R, Xiao Z, et al. Vision Transformer-based recognition of diabetic retinopathy grade. *Med Phys*. 2021;48(12):7850−7863.
14. Diabetic retinopathy detection using CNN, transformer and MLP based architectures. In: Kumar NS, Karthikeyan BR, eds. *2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. Hualien City, Taiwan: IEEE; 2021.
15. Gu Z, Li Y, Wang Z, et al. Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention. *Comput Intell Neurosci*. 2023;12:1305583.
16. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Anal Stereol*. 2014;33(3):231−234.
17. Majithia S, Tham Y-C, Chee M-L, et al. Cohort profile: the Singapore Epidemiology of eye diseases study (SEED). *Int J Epidemiol*. 2021;50:41−52.
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*. 2014. https://doi.org/10.48550/arXiv.1409.1556.
19. Rethinking the inception architecture for computer vision. In: Szegedy C, Vanhoucke V, Ioffe S, et al., eds. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV: Institute of Electrical and Electronics Engineers (IEEE); 2016.
20. Densely connected convolutional networks. In: Huang G, Liu Z, Van Der Maaten L, Weinberger KQ, eds. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, Hawaii: Institute of Electrical and Electronics Engineers (IEEE); 2017.
21. Tan M, Le Q, eds. *Efficientnetv2: Smaller models and faster training. International conference on machine learning*. PMLR, Virtual; 2021.
22. Guo M-H, Lu C-Z, Liu Z-N, et al. Visual attention network. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2202.09741.
23. Crossvit: cross-attention multi-scale vision transformer for image classification. In: Chen C-FR, Fan Q, Panda R, eds. *Proceedings of the IEEE/CVF international conference on computer vision*. Institute of Electrical and Electronics Engineers (IEEE), Virtual conference; 2021.
24. Swin transformer: hierarchical vision transformer using shifted windows. In: Liu Z, Lin Y, Cao Y, et al., eds. *Proceedings of the IEEE/CVF international conference on computer vision*. Institute of Electrical and Electronics Engineers (IEEE), Virtual conference; 2021.
25. Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull*. 2021;139(1):4−15.
26. Tseng RMWW, Gunasekeran DV, Tan SSH, et al. Considerations for artificial intelligence real-world implementation in ophthalmology: providers' and patients' perspectives. *Asia Pac J Ophthalmol*. 2021;10(3):299−306.
27. Ramachandran N, Hong SC, Sime MJ, Wilson GA. Diabetic retinopathy screening using deep neural network. *Clin Exp Ophthalmol*. 2018;46(4):412−416.
28. Baget-Bernaldiz M, Pedro RA, Santos-Blanco E, et al. Testing a deep learning algorithm for detection of diabetic retinopathy in a Spanish diabetic population and with MESSIDOR database. *Diagnostics*. 2021;11(8):1385.
29. Tan GS, Gan A, Sabanayagam C, et al. Ethnic differences in the prevalence and risk factors of diabetic retinopathy: the Singapore Epidemiology of eye diseases study. *Ophthalmology*. 2018;125(4):529−536.