RESEARCH ARTICLE

Encoding Retina Image to Words using Ensemble of Vision Transformers for Diabetic Retinopathy Grading [version 1; peer review: 1 not approved]

Nouar AlDahoul ^{1,2}, Hezerul Abdul Karim ¹, Myles Joshua Toledo Tan^{2,3}, Mhd Adel Momo ², Jamie Ledesma Fermin^{2,3}

¹Engineering, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

²Artificial Intelligence, YO-VIVO corporation, Bacolod City, 6100, Philippines

³Engineering and Technology, University of St. La Salle, Bacolod City, 6100, Philippines

V1 First published: 21 Sep 2021, 10:948
<https://doi.org/10.12688/f1000research.73082.1>
Latest published: 21 Sep 2021, 10:948
<https://doi.org/10.12688/f1000research.73082.1>

Abstract

Diabetes is one of the top ten causes of death among adults worldwide. People with diabetes are prone to suffer from eye disease such as diabetic retinopathy (DR). DR damages the blood vessels in the retina and can result in vision loss. DR grading is an essential step to take to help in the early diagnosis and in the effective treatment thereof, and also to slow down its progression to vision impairment. Existing automatic solutions are mostly based on traditional image processing and machine learning techniques. Hence, there is a big gap when it comes to more generic detection and grading of DR. Various deep learning models such as convolutional neural networks (CNNs) have been previously utilized for this purpose. To enhance DR grading, this paper proposes a novel solution based on an ensemble of state-of-the-art deep learning models called vision transformers. A challenging public DR dataset proposed in a 2015 Kaggle challenge was used for training and evaluation of the proposed method. This dataset includes highly imbalanced data with five levels of severity: No DR, Mild, Moderate, Severe, and Proliferative DR. The experiments conducted showed that the proposed solution outperforms existing methods in terms of precision (47%), recall (45%), F1 score (42%), and Quadratic Weighted Kappa (QWK) (60.2%). Finally, it was able to run with low inference time (1.12 seconds). For this reason, the proposed solution can help examiners grade DR more accurately than manual means.

Keywords

Diabetic Retinopathy Grading, Ensemble Learning, Imbalanced Data, Vision Transformer, Self-attention Mechanism

Open Peer Review

Approval Status

1

version 1

21 Sep 2021

[view](#)

1. **Shruti Jain** , Jaypee University of Information Technology, Solan, India

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Research Synergy**
Foundation gateway.

Corresponding author: Nouar AlDahoul (nouar.aldahoul@live.iium.edu.my)

Author roles: **AlDahoul N:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Abdul Karim H:** Formal Analysis, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Joshua Toledo Tan M:** Validation, Writing – Review & Editing; **Momo MA:** Conceptualization, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Ledesma Fermin J:** Investigation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research project was funded by Multimedia University, Malaysia.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 AlDahoul N *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: AlDahoul N, Abdul Karim H, Joshua Toledo Tan M *et al.* **Encoding Retina Image to Words using Ensemble of Vision Transformers for Diabetic Retinopathy Grading [version 1; peer review: 1 not approved]** F1000Research 2021, 10:948 <https://doi.org/10.12688/f1000research.73082.1>

First published: 21 Sep 2021, 10:948 <https://doi.org/10.12688/f1000research.73082.1>

Introduction

Background

Diabetes mellitus (DM) is a group of metabolic disorders that are characterized by high levels of blood glucose and are caused by either the deficient secretion of the hormone insulin, its inaction, or both. Chronically high levels of glucose in the blood that come with DM may bring about long-term damage to several different organs, such as the eyes.^{1,2} DM is a pandemic of great concern³⁻⁶ as approximately 463 million adults were living with DM in 2019. This number is expected to rise to about 700 million by the year 2045.⁴

High levels of glucose in the blood damage the capillaries of the retina (diabetic retinopathy [DR]) or the optic nerve (glaucoma), cloud the lens (cataract), or cause fluid to build up in the macula (diabetic macular edema), thereby causing diabetic eye disease.⁶⁻¹¹ DR is the leading cause of blindness among adults in the working age¹² and has brought about several personal and socioeconomic consequences,¹³ and a greater risk of developing other complications of DM and of dying.¹⁴ According to a meta-analysis that reviewed 35 studies worldwide from 1980 to 2008, 34.6% of all patients with DM globally have DR of some form, while 10.2% of all patients with DM have vision-threatening DR.¹⁵

A study found that screening for DR and the early treatment thereof could lower the risk of vision loss by about 56%,¹⁶ proving that blindness due to DR is highly preventable. Moreover, the [World Health Organization \(WHO\) Universal Eye Health: A Global Action Plan 2014–2019](#) advocated for efforts to reduce the prevalence of preventable visual impairments and blindness including those that arise as complications of DM.

Many tests can be used for the screening of DR. While sensitivity and specificity are certainly important, the data about performance of tests for DR are different. Researchers employ different outcomes to measure sensitivity, *e.g.*, the ability of a screening test to detect any form of retinopathy, and the ability to detect vision-threatening DR. Additionally, some tests may detect diabetic macular edema better than the different grades of DR according to [World Health Organization. Diabetic retinopathy screening: a short guide. Copenhagen: WHO Regional Office for Europe](#). The examiner's skill is also a source of variation in the test results. A systematic review found that the sensitivity of direct ophthalmoscopy (DO) varies greatly when performed by general practitioners (25%–66%) and by ophthalmologists (43%–79%).¹⁷

DR grading is an essential step in the early diagnosis and effective treatment of the disease. Manual grading is based on high-resolution retinal images examined by a clinician. However, the process is time-consuming and is prone to misdiagnosis. This paper aims to address the matter by developing a fast and accurate automated DR grading system. Here, a novel solution that is based on an ensemble of vision transformers was proposed to enhance grading. Moreover, a public DR dataset proposed in a 2015 [Kaggle challenge](#) was used for the training and evaluation.

Related work

Traditional machine learning (ML) methods have been used to detect DR. Typically, these ML methods require hand-tuned features extracted from small datasets to aid in classification. These traditional methods may involve ensemble learning¹⁸; the calculation of the mean, standard deviation, and edge strength¹⁹; and the segmentation of hard macular exudates.^{20,21} However, these methods require tedious and time-consuming feature engineering steps that are sensitive to the chosen set of features. Work that employs traditional ML methods to detect DR usually yield favorable results using one dataset but fail to obtain a similar success when another dataset is used.^{18,19} This is a common limitation of hand-crafted features.

Deep neural networks, such as CNNs, with much larger datasets have also been used for classification tasks in the diagnosis and grading of DR. These methods involve CNNs developed from scratch to grade the disease using images of the retinal fundus²²; transfer learning based on Inception-v3 neural network to perform multiple binary classification (moderate *versus* worse DR, and severe or worse DR)²³; segmentation prior to detection by pixel classification²⁴ or patch classification.²⁵ A deep learning (DL)-based framework that uses advanced image processing and a boosting algorithm for grading of DR was also proposed by.²⁶ This is one of only a handful of works that have effectively employed transfer learning to train large neural networks for this purpose. Recently, ResNet, a deep CNN, was proposed to address the problem brought about by imbalanced datasets in DR grading.²⁷ Additionally, a bagging ensemble of three CNNs: a shallow CNN, VGG16, and InceptionV3, was used to classify images as DR, glaucoma, myopia and normal.²⁸

Previously, a transformer was also proposed by Vaswani *et al.*²⁹ for natural language processing tasks especially for machine translation. Inspired by the successes of the transformers in NLP, transformers were transferred to computer vision tasks *e.g.* image classification.

Methods

In this section, the DR detection dataset is explored. Additionally, the vision transformer, a DL model that was used on these data, is discussed in detail.

Dataset overview

The **DR detection dataset** is highly imbalanced and consists of high-resolution images with five levels of severity including No_DR, Mild, Moderate, Severe, and Proliferative_DR. It has significantly more samples for the negative (No_DR) category than for the four positive categories. **Table 1** shows the class distribution of the training and testing sets. **Figure 1**, on the other hand, shows a few samples from each class. The images come with different conditions and were labeled with subject IDs. The left and right fields were provided for every subject. The images were captured by different cameras, thus affecting the visual appearance of the images of left and right eyes.

The samples of the training set were rescaled between $[0, 1]$, cropped to remove their black borders, and augmented by randomly flipping the samples horizontally and vertically, and by randomly rotating the samples by 360° . The samples of the test set were only cropped and rescaled. **Figure 2** shows a few augmented samples from the training set.

Vision transformer

A vision transformer is a state-of-the-art DL model that is used for image classification and was inspired by Dosovitskiy *et al.*³⁰ **Figure 3** shows the architecture of the vision transformer. In this paper, a retinal image that has a sequence of patches encoded as a set of words was applied to the transformer encoder as shown in **Figure 3**. The original image's patches $N = (H \times W)/P^2$ were extracted with a fixed patch size (P, P) where $P = 16$, W is the image width, H is the image height, and N is the number of patches. The extracted patches were flattened and each patch x_p belonged to $\mathbb{R}^{P^2 \cdot C}$, where C is the number of channels.

As a result, the 2D image was converted into a sequence of patches $x \in \mathbb{R}^{N \times (P^2 \cdot C)}$. Each patch in the sequence x was mapped to a latent vector with hidden size $D = 768$. A learnable class embedding $z_0^0 = x_{class}$ was prepended for the embedded patches, whose state at the output of the transformer's encoder (z_L^0) serves as the representation y of the image. After that, a classifier was attached to the image representation y . Additionally, a position embedding E_{pos} was added to the patch embeddings to capture the order of patches that were fed into the transformer encoder. **Figure 4** illustrates the architecture of transformer's encoder with L blocks, each block containing alternating layers of multi-head self-attention (MSA)²⁹ and multi-layer perceptron (MLP) blocks. The layer normalization (LN)³¹ was applied before every block, while residual connections were applied after every block.³⁰

Ensemble learning of vision transformers

Ensemble learning is a ML ensemble meta-algorithm. Bagging (Bootstrap Aggregating) is a type of ensemble learning that uses "majority voting" to combine the output of different base models to produce one optimal predictive model and improve the stability and accuracy.³²

The advantage of ensemble bagging several transformers is that aggregation of several transformers, each trained on a subset of the dataset, outperforms a single transformer trained over the entire set. In other words, it leads to less overfit by removing variance in high-variance low-bias datasets. To increase the speed of training, the training can be done in parallel by running each transformer on its own data prior to result aggregation, as shown in **Figure 5**.

Experimental setup and protocol

The images available in this dataset were resized to $H = 256, W = 256$, the latent vector hidden size was set to $D = 768$, the number of layers of the transformer to $L = 12$, the MLP_{size} to 3072, the MSA_{heads} to 12, and the default value of the patch size to $P = 16$. Thus, the sequence's number N was 256.

Table 1. Training and Testing Class Distribution in dataset of EyePACS, Diabetic Retinopathy Detection.

Class	Training	Testing
No_DR	25810	39533
Mild	2443	3762
Moderate	5292	7861
Severe	873	1214
Proliferative_DR	708	1206

Figure 1. A few samples of each class in dataset of EyePACS, Diabetic Retinopathy Detection: "No_DR" (red borders), "Mild" (blue), "Moderate" (green), "Severe" (yellow), "Proliferative_DR" (violet). The images have various sizes but were resized uniformly.

Figure 2. A few samples cropped and augmented randomly.

Figure 3. The vision transformer architecture.³⁰

Figure 4. Encoder Architecture of the Transformer.³⁰

Figure 5. Ensemble learning of vision transformers.

In the experiments conducted, 20% from each class in the training set were selected for validation. All transformers were fine-tuned using the weights of the transformer pre-trained on ImageNet-21K.³³

For optimization, the ADAM algorithm³⁴ was utilized with a batch size of 8. Furthermore, the mean squared error loss function was used. The training process for each transformer consists of two stages:

- 1) All layers in the transformer backbone were frozen and the regression head that was initialized randomly was unfrozen. Then, the regression head was trained for five epochs.
- 2) The entire model (transformer backbone + regression head) which was trained for 40 epochs was unfrozen.

Data augmentation, early stopping, dropout, and learning rate schedules were used to prevent overfitting and loss divergence. **Figure 6** shows the attention map of a few samples extracted from the transformer.

The classification heads of all transformers were removed and replaced by a regression head with one node instead of logits. The regression output of a transformer was interpreted as shown in **Table 2** to be converted into a category.

An ensemble of ten transformers with similar architectures and hyperparameters was used. The samples were divided randomly into ten sets and each transformer was trained on each one. After interpreting the regression output from each transformer, the predicted classes from ten transformers were aggregated with “majority voting” to predict the final class.

Training, validation, and testing were carried out using the TensorFlow framework on an NVIDIA Tesla T4 GPU.

Figure 6. The Attention Map of samples A) No DR, B) Mild, C) Moderate, D) Severe, E) Proliferative DR.

Table 2. Pseudocode for the transformer regression output interpretation.

```

Algorithm: Regression output interpretation
function classify( $x_{sample}$ ) {
  inputs:  $x_{sample}$  output float number from the transformer
  outputs:  $y_{sample}$  which represents the class of the presented sample
  If  $x_{sample} < 0.8$  then
    return  $y_{sample} = \text{No DR}$ 
  Else If  $x_{sample} < 1.5$  then
    return  $y_{sample} = \text{Mild}$ 
  Else If  $x_{sample} < 2.5$  then
    return  $y_{sample} = \text{Moderate}$ 
  Else If  $x_{sample} < 3.5$  then
    return  $y_{sample} = \text{Severe}$ 
  return  $y_{sample} = \text{Proliferative DR}$ 
}

```


Results and discussion

Performance metrics

In this section, the results of the proposed ensemble of transformers are discussed. The performance metrics, such as precision, recall, and F1 score were calculated. Additionally, the **quadratic weighted kappa** (QWK) metric was utilized in this dataset because these data needed specialists to label the images manually since the small differences among the classes can only be recognized by specialist physicians. QWK which lies in the range $[-1, +1]$ measures the agreement between two ratings and is calculated between the scores assigned by human raters (doctors) and predicted scores (models) as shown in **Table 3**. The dataset has five ratings: 0, 1, 2, 3, 4.

QWK was calculated as follows:

- 1) The confusion matrix O between predicted and actual ratings was calculated.
- 2) A histogram vector was computed for each rating in the predictions and in the actual.
- 3) The E ($N \times N$) matrix which represents the outer product between the two histogram vectors was calculated.
- 4) The W ($N \times N$) weight matrix was constructed representing the difference between the ratings as shown in **Table 4**.³⁵

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2} \quad (1)$$

Where $1 \leq i \leq 5, 1 \leq j \leq 5$

- 5) QWK was defined as follows³⁵:

$$QWK = 1 - \frac{\sum_i \sum_j W_{ij} \times O_{ij}}{\sum_i \sum_j W_{ij} \times E_{ij}} \quad (2)$$

where N is the number of classes.

Table 3. OWK interpretation.

Kappa	Agreement
< 0	No
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect

Table 4. The Weight Matrix W represents the difference between the classes.

	No_DR	Mild	Moderate	Severe	Proliferative_DR
No_DR	0	0.0625	0.25	0.5625	1
Mild	0.0625	0	0.0625	0.25	0.5625
Moderate	0.25	0.0625	0	0.0625	0.25
Severe	0.5625	0.25	0.0625	0	0.0625
Proliferative_DR	1	0.5625	0.25	0.0625	0

Experimental results

Table 5 shows the performance metrics of ten transformers with each one trained on a subset of data. It is obvious that there is a big difference among the performances of these individual transformers. Transformer_1 was able to yield a Kappa of 55.1%. On the other hand, transformer_10 yielded a Kappa of 30.9%. Ensembles of various numbers of transformers including all ten transformers, four transformers (1,3,8,9), and other configurations were also evaluated. The best model was an ensemble of two transformers (1,3) which yielded a Kappa of 60.2%.

This Kappa is at the boundary between moderate and substantial agreement. The previous results confirm that the performance of the ensemble of transformers (1,3) trained with fewer training images outperformed the ensemble of ten transformers trained with five times the number of images. **Table 6** compares the performance of the ensemble of transformers with the ensemble of ResNet50 CNNs. The ResNet50 CNN was transferred from ImageNet 1K. The top layers were replaced by a support vector machine that was tuned with this dataset. The proposed ensemble of transformers outperformed the ensemble of ResNet50 CNNs significantly by >18% Kappa.

The confusion matrix of each configuration including ensembles of transformers with ten, four, and two transformers, and the ensemble of two ResNet50

CNNs were shown in **Figure 7**. The confusion matrix (c) which represents the best Kappa of 60.2% shows that the model was able to recognize the categories of severe and proliferative DR from one side, and NO DR and mild DR from the second side.

Table 5. Performance metrics for various ensemble models.

Model	Precision %	Recall %	F1 Score %	QWK %
Transformer_1	43	47	40	55.1
Transformer_2	37	46	31	40.2
Transformer_3	39	46	37	52.0
Transformer_4	45	41	30	43.3
Transformer_5	41	45	35	48.6
Transformer_6	39	45	33	44.9
Transformer_7	44	42	33	43.3
Transformer_8	40	47	39	51.3
Transformer_9	40	46	38	51.0
Transformer_10	35	36	26	30.9
Ensemble of ten transformers	45	47	38	53.0
Ensemble of four transformers	44	47	41	57.5
Ensemble of two transformers	47	45	42	60.2

Table 6. Comparison between the ensemble of transformers and the ensemble of ResNet50 CNNs.

Model	Precision %	Recall %	F1 Score %	QWK%
Ensemble of two transformers	47	45	42	60.2
Ensemble of ten ResNet50	32	44	32	36.97
Ensemble of two ResNet50	35	40	35	41.52

Figure 7. The Confusion Matrix for the ensemble bagging of A) ten transformers, B) four transformers, C) two transformers, D) two ResNet50 CNN.

Conclusion

This study is a new attempt to demonstrate the capability of the ensemble bagging of vision transformers applied on retinal image classification for the grading of DR into five levels of severity. The experiments conducted showed that even when the dataset was challenging, the proposed method was able to yield promising performance measures in terms of precision (47%), recall (45%), F1 score (42%), and QWK (60.2%). Furthermore, the inference time was low at 1.12 seconds. Hence, we intend to enhance the performance by utilizing a collection of various DR datasets. This can increase the size and variety of training data to train the proposed model from scratch instead of starting from the weights of the ImageNet 21K-based model. By doing so, we can ultimately enhance performance.

Author contributions

Conceptualization by N.A., M.A.M.; Data Curation by N.A.; Formal Analysis by N.A., H.A.K., M.A.M.; Funding Acquisition by H.A.K.; Investigation by N.A., J.L.F.; Methodology by N.A., H.A.K., M.A.M.; Project Administration by H.A.K.; Software by N.A., M.A.M.; Validation by N.A., M.J.T.T.; Visualization by N.A.; Writing – Original Draft Preparation by N.A., M.A.M., J.L.F.; Writing – Review & Editing by N.A., H.A.K., M.J.T.T., M.A.M., J.L.F.

Ethics and consent

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

The Retinal images are public third part dataset provided by [EyePACS](#), a free platform for retinopathy screening.

Competing interests

None of the authors declare any competing interests.

Grant information

This research project was funded by Multimedia University, Malaysia.

Data availability

The [dataset](#) used in this work is accessible to the public on the Kaggle website. It was created in 2015 for the Kaggle Diabetic Retinopathy Detection competition. This competition is sponsored by the [California Healthcare Foundation](#). Retinal images were provided by [EyePACS](#), a free platform for retinopathy screening.

References

- American Diabetes Association: **Diagnosis and classification of diabetes mellitus**. *Diabetes Care*. Jan. 2010; **33**(Suppl 1): S62–S69. [PubMed Abstract](#) | [Publisher Full Text](#)
- Fowler MJ: **Microvascular and Macrovascular Complications of Diabetes**. *Clin Diab*. Apr. 2008; **26**(2): 77. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhou B, et al.: **Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants**. *Lancet*. Apr. 2016; **387**(10027): 1513–1530. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- International Diabetes Federation: *IDF Diabetes Atlas*. 9th ed. Brussels, Belgium: International Diabetes Federation; 2019, pp. 32–61.
- Narayan KMV: **The Diabetes Pandemic: Looking for the Silver Lining**. *Clinical Diabetes*. Apr. 2005; **23**(2): 51–52. [Publisher Full Text](#)
- Cheloni R, Gandolfi SA, Signorelli C, et al.: **Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis**. *BMJ Open*. Mar. 2019; **9**(3): e022188. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wu L, Fernandez-Loaiza P, Sauma J, et al.: **Classification of diabetic retinopathy and diabetic macular edema**. *World J Diabetes*. Dec. 2013; **4**(6): 290–294. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang W, Lo ACY: **Diabetic Retinopathy: Pathophysiology and Treatments**. *Int J Mol Sci*. Jun. 2018; **19**(6): 1816. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Song BJ, Aiello LP, Pasquale LR: **Presence and Risk Factors for Glaucoma in Patients with Diabetes**. *Curr Diab Rep*. Dec. 2016; **16**(12): 124–124. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pollreis A, Schmidt-Erfurth U: **Diabetic cataract-pathogenesis, epidemiology and treatment**. *J Ophthalmol*. 2010; **2010**: 608751–608751. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Das A, McGuire PG, Rangasamy S: **Diabetic Macular Edema: Pathophysiology and Novel Therapeutic Targets**. *Ophthalmology*. 2015; **122**(7): 1375–1394. [PubMed Abstract](#) | [Publisher Full Text](#)
- Cheung N, Mitchell P, Wong TY: **Diabetic retinopathy**. *Lancet*. Jul. 2010; **376**(9735): 124–136. [Publisher Full Text](#)
- Rees G, et al.: **Association Between Diabetes-Related Eye Complications and Symptoms of Anxiety and Depression**. *JAMA Ophthalmol*. Sep. 2016; **134**(9): 1007–1014. [PubMed Abstract](#) | [Publisher Full Text](#)
- Kramer CK, Rodrigues TC, Canani LH, et al.: **Diabetic Retinopathy Predicts All-Cause Mortality and Cardiovascular Events in Both Type 1 and 2 Diabetes**. *Diabetes Care*. May 2011; **34**(5): 1238. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yau JWY, et al.: **Global Prevalence and Major Risk Factors of Diabetic Retinopathy**. *Diabetes Care*. Mar. 2012; **35**(3): 556–564. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rohan TE, Frost CD, Wald NJ: **Prevention of blindness by screening for diabetic retinopathy: a quantitative assessment**. *BMJ*. Nov. 1989; **299**(6709): 1198–1201. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hutchinson A, et al.: **Effectiveness of screening and monitoring tests for diabetic retinopathy—a systematic review**. *Diabet Med*. Jul. 2000; **17**(7): 495–506. [PubMed Abstract](#) | [Publisher Full Text](#)
- Bhatia K, Arora S, Tomar R: **Diagnosis of diabetic retinopathy using machine learning classification algorithm**. *2016 2nd Int Conf Next Generation Computing Technologies (NGCT)*. 2016, pp. 347–351. [Publisher Full Text](#)
- Asha PR, Karpagavalli S: **Diabetic Retinal Exudates Detection Using Extreme Learning Machine**. *Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, Cham*. 2015; pp. 573–578. [Publisher Full Text](#)
- Soparak A, Uyyanonvara B: **Automatic exudates detection from diabetic retinopathy retinal image using fuzzy c-means and morphological methods**. *Proceedings of the 3rd IASTED International Conference of Advances in Computer Science and Technology*. 2007; pp. 359–364. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Osareh A, Mirmehdi M, Thomas B, et al.: **Automatic recognition of exudative maculopathy using fuzzy C-means clustering and neural networks**. *Proc Medical Image Understanding Analysis Conference*. 2001, vol. **3**, pp. 49–52.
- Pratt H, Coenen F, Broadbent DM, et al.: **Convolutional Neural Networks for Diabetic Retinopathy**. *Procedia Computer Sci*. Jan. 2016; **90**: 200–205. [Publisher Full Text](#)
- Gulshan V, et al.: **Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs**. *JAMA*. Dec. 2016; **316**(22): 2402–2410. [PubMed Abstract](#) | [Publisher Full Text](#)
- Prentašić P, Lončarić S: **Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion**. *Comput Methods Programs Biomed*. Dec. 2016; **137**: 281–292. [PubMed Abstract](#) | [Publisher Full Text](#)
- van Grinsven MJJP, van Ginneken B, Hoyng CB, et al.: **Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images**. *IEEE Transactions Medical Imaging*. May 2016; **35**(5): 1273–1284. [Publisher Full Text](#)
- Wang Y: **A Deep Learning Based Pipeline for Image Grading of Diabetic Retinopathy**. *Master of Science: Virginia Polytechnic Institute and State University*. 2018.
- Sallam MS, Asnawi AL, Olanrewaju RF: **Diabetic Retinopathy Grading Using ResNet Convolutional Neural Network**. *2020 IEEE Conference on Big Data and Analytics (ICBDA)*. 2020, pp. 73–78. [Publisher Full Text](#)

28. Smaida M, Yaroshchak S: **Bagging of convolutional neural networks for diagnostic of eye diseases.** *CEUR Workshop Proceedings*. 2020; **2604**, 715–729.
29. Vaswani A, *et al.*: **Attention Is All You Need.** *arXiv:1706.03762 [cs]*. Dec. 2017. Accessed: May 30, 2021.
[Reference Source](#)
30. Dosovitskiy A, Lucas B, Alexander K, *et al.*: **An image is worth 16x16 words: Transformers for image recognition at scale.** *ICLR*. 2021.
31. Ba JL, Kiros JR, Hinton GE: **Layer Normalization.** *arXiv:1607.06450 [cs, stat]*. Jul. 2016. Accessed: Jun. 30, 2021.
[Reference Source](#)
32. Fan W, Zhang K: **Bagging** In: *Encyclopedia of Database Systems*. LIU L, ÖZSU MT, Eds. Boston, MA: Springer US; 2009, pp. 206–210.
[Publisher Full Text](#)
33. Ridnik T, Ben-Baruch E, Noy A, *et al.*: **ImageNet-21K Pretraining for the Masses.** *arXiv:2104.10972 [cs]*. Jun. 2021. Accessed: Jun. 30, 2021.
[Reference Source](#)
34. Kingma DP, Ba J: **Adam: A Method for Stochastic Optimization.** *arXiv:1412.6980 [cs]*. Jan. 2017. Accessed: Jun. 30, 2021.
[Reference Source](#)
35. Tymchenko B, Marchenko P, Spodarets D: **Deep Learning Approach to Diabetic Retinopathy Detection.** *arXiv:2003.02261 [cs, stat]*. Mar. 2020. Accessed: Jun. 30, 2021.
[Reference Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 25 November 2021

<https://doi.org/10.5256/f1000research.76706.r94946>

© 2021 Jain S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Shruti Jain

Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Solan, Himachal Pradesh, India

1. Add a section highlighting the main contributions of your methodology, with detailed reference to, and comparison with, existing work.
2. I find it difficult to understand, from the abstract, the proposed methodology by which you seek to solve the problem of your paper. For example, the authors could clarify the following: "To enhance DR grading, this paper proposes a novel solution based on an ensemble of state-of-the-art deep learning models called vision transformers". What are vision transformers? Are authors proposing this or what is new in it? <https://viso.ai/deep-learning/vision-transformer-vit/>

Likewise the authors write "A challenging public DR dataset proposed in a 2015 Kaggle challenge was used for training and evaluation of the proposed method." Yet there are many datasets for DR grading, so why only Kaggle? The author can validate their model with other datasets.
3. Add some latest papers and cite them.
4. Read the whole manuscript for typos and grammatical mistakes.
5. Recheck, not all references that could be given are given. In 'Dataset Overview', in the sentence "the quadratic weighted kappa (QWK) metric was utilized in this dataset because of these data..." the reference is missing. Likewise there are places where authors can give references.
6. Finally, your conclusion needs to be more tailored to your findings. The authors write "Hence, we intend to enhance the performance by utilizing a collection of various DR datasets. This can increase the size and variety of training data to train the proposed model

from scratch instead of starting from the weights of the ImageNet 21K-based model. By doing so, we can ultimately enhance performance”

Yet, why don't authors have tried increasing the datasets utilized already? Furthermore, the ImageNet 21K-based model is mentioned significantly in the conclusion without having been elaborated upon in the main text of the article: why was it introduced, what is it, and what is it for and how does it improve performance? Likewise, there is no mention of ensemble transformers or vision transformers in the conclusion.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Image and Signal Processing, Soft Computing, Internet-of-Things, Pattern Recognition, Bio-inspired Computing and Computer-Aided Design of FPGA and VLSI circuits

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research