Data-Driven Decision-Making for Classification of Diabetic Retinopathy Using Convolutional Neural Network (CNN) in a Clinical Decision Support System

Ahmad Faiz Ghazali[1*] Nuriyah Mohamad Zakaria[2] and Azliza Mohd Ali[3]

[1,2,3] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
faizghazali@uitm.edu.my

Abstract. Data-driven decision-making is getting more attention since its modelling core depends on data availability, rather than long and costly processes of getting data and requirements from domain experts, patients, or decision-makers. The current practice for detecting diabetic retinopathy is conducted manually by ophthalmologists. Diabetes may cause a serious eye complication issue called diabetic retinopathy. The procedure is time-consuming in getting eye examinations where only a limited number of patients can be examined. Therefore, deep learning techniques such as convolutional neural networks (CNN) can be employed to help with faster detection of diabetic retinopathy. The architecture of CNN including ResNet-50 can classify the fundus images of diabetic retinopathy. ResNet50 leverages transfer learning, which helps with better generalisation. The APTOS 2019 dataset is acquired and undergoes pre-processing to increase the quality of fundus images such as cropping, contrast enhancement, performing oversampling and data augmentation. Training and testing are necessary to improve the accuracy of using ResNet-50 by different attributes. The model's performance is then evaluated using evaluation metrics and has achieved a high accuracy of 87.0%. A clinical decision support system (CDSS) using the model is developed as the proof-of-concept for this research.

Keywords: Diabetic Retinopathy, Convolutional Neural Network (CNN), Res Net-50, Clinical Decision Support System (CDSS).

1 Introduction

Decision support system involves data acquisition and analysis in supporting decision makers or domain experts in specific problem domains targeted. In this study, a clinical decision support system (CDSS) is modelled for ophthalmologists' usage in diagnosing diabetes mellitus, which is a chronic metabolic disease distinguished by high blood glucose levels. One of the diabetic conditions which could lead to vision im-

N. A. S. Abdullah et al. (eds.), *Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024)*, Advances in Computer Science Research 117,
https://doi.org/10.2991/978-94-6463-589-8_33

pairment is known as diabetic retinopathy (DR). It is a condition where blood vessels inside the retina deteriorate due to high glucose levels. In developed nations like Malaysia, DR is also recognized as the primary cause of blindness among adults [1].

Manual DR detection necessitates the expertise of a qualified opthalmologist. However, a study in 2020 revealed a significant shortage of ophthalmologists worldwide, estimating only 7.16 ophthalmologists for every 1000 vision-threatened DR patients [2]. This leads to the use of computer-aided diagnosis (CAD) systems for automatic DR diagnosis over the past three decades [8]. A variety of machine learning (ML) approaches have been applied to detect common features of DR in CAD. However, it is time-consuming due to the manual feature extraction that needs to be done by human experts. Therefore, this study employed a deep learning model called Convolutional Neural Network (CNN), which is tailored to improve the effectiveness of detecting DR.

In the medical field, CNN is widely used for its excellent feature extractor in medical image recognition [3]. Using CNN in clinical applications such as CDSS can lead to faster detection of DR and its stages based on the presented fundus images. Features of DR can be extracted from the fundus images accordingly. Ophthalmologists detect DR by examining coloured fundus retinal images which allows them to determine its clinical stages based on the observed features [4]. DR can be classified into two types which are non-proliferative DR, separated into mild, moderate, or severe and proliferative DR based on features present in the fundus images [5]. Each stage has distinct characteristics, but as DR worsens, later stages may show features from earlier stages [4, 6].

The position of optic disc, macular region and other retinal layer are visible in a healthy retina. However, this is not the case for DR retinal, where the presence of microaneurysms, soft and hard exudates, haemorrhages and neovascularization are seen. Microaneurysms are the common initial sign of diabetic retinopathy, appearing as tiny red circular spots on the retina due to the enlargement of thin blood vessel walls. Meanwhile haemorrhages appear as bigger red circular spots, hence the sizes can be up to 125 μm with an erratic fringe [7, 8]. Hard exudates appear on the retina as sharp-edged yellow patches while soft exudates are a paler yellow with rounded edges [9, 10]. The last feature, neovascularization is characterized by networks of vessels that extend into the vitreous cavity from the surface of the retina.

The motivation for conducting this study is to explore the possibility of employing artificial intelligence (AI) engine to diagnose diseases by experimenting with suitable datasets [11]. The utilization of AI in CDSS was emphasizes by the reviews of CDSS by [12] but the importance of CDSS has been stressed much earlier in reviews by [13]. The challenges faced by domains experts which are the ophthalmologists, signifies the needs of this study. It was reported in 2023 that an estimated 500,000 Malaysians suffer from DR.

2 Pre-processing of Data for Implementation of the Model

This section explains the data acquisition, employed data preprocessing techniques, implementation of the ResNet-50 model and the evaluation metrics used.

2.1 Dataset

The dataset used is from the APTOS 2019 Blindness Detection dataset on Kaggle collected by the Asia Pacific Tele-Ophthalmology Society (APTOS) [11]. It comprises 3,662 high-resolution retinal images showing varying degrees of DR from rural Indian subjects as seen in Figure 1. Experienced ophthalmologists labeled each image into five classes: no DR (Class '0'), mild DR (Class '1'), moderate DR (Class '2'), severe DR (Class '3'), and proliferative DR (Class '4').

Fig. 1. Sample fundus images of APTOS 2019 dataset

2.2 Data preprocessing

Since this research uses secondary data, the fundus images inevitably come in variations, such as being overexposed or underexposed. Therefore, it was necessary to preprocess to enhance the quality and standardize these fundus images. The preprocessing steps commenced by resizing all fundus images to 224x244 pixels. Next steps involved converting the images into grayscale and cropping their dark areas to ensure that the images would retain only the retina region. Image enhancement such as weighted blending and Gaussian blur were further applied to the fundus images to produce a smoother aspect. Figure 2 shows a sample of an image significantly helped to improve the classification model's performance, as it now can better learn the enhanced features.

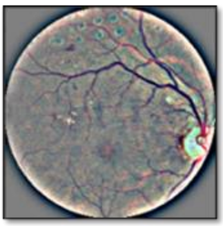

Fig. 2. Sample of pre-processed image

The fundus images of APTOS 2019 present a formidable challenge in the form of a significant imbalance in class distribution, as seen in Table 1. The high discrepancy can result in an overfitting problem where the model solely replicate's the training data, leading to inaccurate predictions on unseen data.

Table 1. Class distribution of APTOS 2019 dataset

Label	Number of Images
0	1805
1	999
2	370
3	295
4	193

Synthetic Minority Oversampling Technique (SMOTE) was used to address this problem by creating artificial samples from the minority class. Oversampling on the minority classes that is Class '1', Class '2', Class '3', and Class '4' are done to increase the number of images so that it matches the majority class, Class '0'. This significantly increased the overall number of fundus images from 3,662 images to 9,025 images in total. Before implementing the ResNet-50, two image augmentation techniques which are zooming, and horizontal flipping were applied to help improve the robustness of the trained ResNet-50 classification model.

2.3 Implementation of ResNet-50

A ResNet variation called ResNet-50 successfully tackles a deep network problem known as the vanishing gradient problem. The key idea behind ResNet-50 is transfer learning, in which the network learns the difference between the intended output and the input rather than the mapping itself.

2.4 Evaluation metrics

Model evaluation is a crucial step that helps to evaluate how accurate the model is in classifying diabetic retinopathy into severity stages based on the fundus images. As a result, accuracy, F1 score, precision, and recall were used to evaluate the proposed model.

Precision measures the rate of instances that were classified as positive are actually positives,

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

Recall measures the rate of actual positive instances accurately detected by the classfier,

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

F1-score measures how well the classifer perform multi-class classifying,

$$F1 = \frac{2 \times Precision \times Recall}{(Precision+Recall)} \tag{3}$$

Accuracy measures the rate of instances that were correctly classified,

$$Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)} \tag{4}$$

2.5 Hyperparameters of ResNet-50 model

Before training the ResNet-50 model, hyperparameters must be determined. The training of the ResNet-50 model was done where its runtime type is Python 3 with V100 GPU as the hardware accelerator to help speed up the training process of the classification model. Table 2 shows the hyperparameters and their best values used.

Table 2. Hyperparameters for the classification model

Hyper-parameters	Values
Number of classes	5
Batch size	64
Number of epochs	100
Early stop patience	10
Learning rate	0.0001
Optimizer	Adam
Activation function	Softmax
Loss	Sparse Categorical Crossentropy

3 Results and Discussion

The model was trained and tested with two different ratios, which are 80:20 and 90:10 to determine the ratio with the best result for the classification system. The model was trained using the hyperparameters in Table 2. For early stop patience, the training will stop after 10 consecutive epochs if there is no improvement in validation loss monitored. The callback will restore the model's weight to its best state once training stops due to early stopping to prevent the model from overfitting.

3.1 Training and Testing with 80:20 ratio

80% of the dataset was set aside for training and the remaining 20% for testing, and trained for 100 epochs, stopping early at epoch 73. Figure 3 shows that validation loss declined faster than training loss, but around epoch 60, training loss stabilized, indicating possible overfitting. Meanwhile, Figure 4 shows that both training and validation accuracy increased over time. However, around epoch 50, training accuracy steadily increased while validation accuracy fluctuated, ending lower than training accuracy, suggesting potential overfitting.

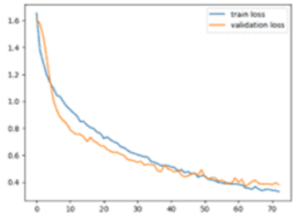

Fig. 3. Train and validation loss graph for 80:20

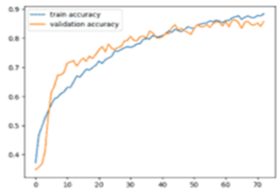

Fig. 4. Train and validation accuracy graph for 80:20

The model achieved an 84.0% accuracy with the test set. As seen from the evaluation metrics in Table 3, Class '0' managed to obtain precision, recall and F1-score for more than 90.0%. This proves that the model is performing significantly well in classifying instances from Class '0'. Although the precision for Class 2 is high with 80.0%, there is a huge difference when compared to its recall of 63.0%. This imbalance indicates that the model can correctly predict instances of Class '2' as Class '2' but unable to correctly capture instances of Class '2'.

Table 3. Evaluation metrics summary for 80:20

Class	Precision (%)	Recall (%)	F1-score (%)
0	93.0	93.0	93.0
1	85.0	87.0	86.0
2	80.0	63.0	70.0
3	82.0	95.0	88.0
4	80.0	83.0	82.0

Based on the confusion matrix in Figure 5, the model was able to most correctly identify instances from Class ‘3’ with 344, proving that the model is best at detecting features of severe DR. The model also obtained a high number of true positives for Class ‘0’, Class ‘1’, and Class ‘4’ where the correctly classified instances for each classes were more than 300 instances.

Fig. 5. Confusion matrix for 80:20

3.2 Training and Testing with 90:10 ratio

90% of the dataset was set aside for training and the remaining 10% for testing, trained for 100 epochs, with early stopping at epoch 91. Figure 6 shows that both training and validation losses declined gradually. Before epoch 50, validation loss decreased at a faster rate than training loss. However, this situation was interchanged after epoch 50 with training loss surpassing the validation loss, indicating overfitting. Figure 7 shows training and validation accuracy increasing over time, but around epoch 60, the train accuracy had increased slightly at a faster rate, surpassing the validation accuracy. Although both lines were still relatively close to one another, this suggests a potential overfitting.

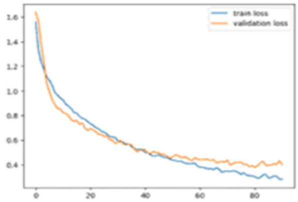

Fig. 6. Train and validation loss graph for 90:10

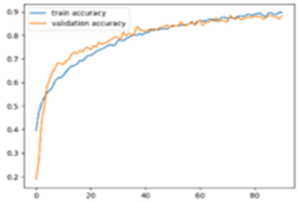

Fig. 7. Train and validation accuracy graph for 90:10

When predicted with the test set, the model achieved an accuracy of 87.0%. As seen in Table 4, the model also shows the highest precision, recall and F1-score result for Class ‘0’, however it was followed by Class ‘1’ instead of Class ‘3’, unlike the 80:20 ratio. This proves that the model is performing significantly well in classifying and detecting instances from these three classes. The F1-score for Class ‘2’ increased by 3.0% compared to ratio 80:20 and had a smaller difference between its precision and recall. From the results overall, a few metrics decrease but majority increase significantly, proving that this ratio produced the best result compared to 80:20 ratio.

Table 4. Evaluation metrics summary for 90:10

Class	Precision (%)	Recall (%)	F1-score (%)
0	93.0	92.0	92.0
1	90.0	92.0	91.0
2	77.0	70.0	73.0
3	87.0	95.0	90.0
4	85.0	88.0	87.0

Based on the confusion matrix in Figure 8, all classes correctly classified more than 100 fundus images out of 181. The number of images misclassified for all classes were at most 20 images, showing that the model's performance is notably solid and practical in accurately identifying most instances across most classes. This shows that the model is unable to correctly capture features of moderate DR, often confusing it with mild and severe DR.

Fig. 8. Confusion matrix for 90:10

Table 5 summarizes the calculated average of precision, recall, and F1-score for train test splits of 80:20 and 90:10. The model with a ratio split of 90:10 gives the highest precision, recall, F1-score, and accuracy, with 87.0%, 87.2%, 87.2%, and 87.0%, respectively. The enchanced performance metrics prove that the model with 90:10 ratio split is able to to detect and classify the stages of DR more accurately compared with 80:20 ratio split. This implies that the classification model is reliable, which can boost confidence in its clinical application for automated DR detection.

Table 5. Evaluation metrics summary for all ratios

Ratio	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
80:20	83.0	83.2	83.0	83.0

90:10	87.0	87.2	87.2	87.0

4 Outcomes

As seen in Figure 9, the outcome is a web application prototype function for the proposed clinical decision support system (CDSS) for DR. The prototype can produce output that determines the severity stage of the diabetic retinopathy for the inputted fundus images using the classification model. The prototype was the core function for the engine of the proposed CDSS and will be further improved to best suit the function of a DR detection system in clinical settings.

Fig.9. Screenshot of the DR web application

5 Conclusion

Automated detection of diabetic retinopathy (DR) can help reduce the labour-intensive burden on ophthalmologists, leading to shorter delays in diagnosing a large number of diabetic patients. This study proposed an automated detection of DR by employing a state-of-art convolutional neural network (CNN) which is ResNet-50. The model was trained and tested on APTOS 2019 dataset and was divided into two sets of ratios; 80:20 and 90:10. The classification model achieved the highest accuracy at 87.0% with ratio of 90:10 and have shown excellent results for detecting features of no DR, moderate DR, and severe DR.

Despite the relatively good overall performance achieved by the model, there are a few limitations that can be addressed. First of all, the model could not classify moderate DR, Class ‘2’ for both ratios. This issue might be caused by insufficient feature representation for Class ‘2’, making it harder for the model to learn the features for this class effectively. The CDSS prototype successfully produced an output that determines the DR severity stage for the inputted fundus image. In the future, the work can be enhanced by using a larger dataset and other algorithms to boost the accuracy of DR classification in a fully functional user-friendly CDSS.

6 Acknowledgement

This study acknowledges College of Computing, Informatics and Mathematics (KPPIM), Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia, for funding this research outcomes.

References

1. Simó-Servat, O., Hernández, C., & Simó, R. (2019). Diabetic Retinopathy in the Context of Patients with Diabetes. In Ophthalmic Research (Vol. 62, Issue 4, pp. 211–217). S. Karger AG. https://doi.org/10.1159/000499541
2. Teo, Z. L., Tham, Y.-C., Yu, M., Cheng, C.-Y., Wong, T. Y., & Sabanayagam, C. (2020). Do we have enough ophthalmologists to manage vision-threatening diabetic retinopathy? A global perspective. Eye (Lond), 34(7), 1255–1261.
3. Asia, A.-O., Zhu, C.-Z., Althubiti, S. A., Al-Alimi, D., Xiao, Y.-L., Ouyang, P.-B., & Al-Qaness, M. A. A. (2022). Detection of Diabetic Retinopathy in Retinal Fundus Images Using CNN Classification Models. Electronics, 11(17). https://doi.org/10.3390/electronics11172740
4. Welikala, R. A. (2014). Automated detection of proliferative diabetic retinopathy from retinal images. In PQDT - UK & Ireland.
5. Raj, A., Noor, N. M. M., Mohemad, R., Mata, N. A. C., & Hussain, S. (2024). An Intelligent Mathematically Modified Fuzzy C-Means Clustering Technique for Fundus Image Segmentation for Diabetic Retinopathy Identification. International Journal of Intelligent Systems and Applications in Engineering, 12(7), 603–612.
6. Wilkinson, C. P., Ferris 3rd, F. L., Klein Ronald E and Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J. T., & Global Diabetic Retinopathy Project Group. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology, 110(9), 1677–1682.
7. Aziz, T., Charoenlarpnopparut, C., & Mahapakulchai, S. (2023). Deep learning-based hemorrhage detection for diabetic retinopathy screening. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-28680-3
8. Patra, P., & Singh, T. (2022). Diabetic Retinopathy Detection using an Improved ResNet50-InceptionV3 Structure. 2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022. https://doi.org/10.1109/ICCCNT54827.2022.9984253
9. McDonagh, J., & Keating, D. (2005). Development of an automated screening tool for diabetic retinopathy using artificial intelligence. In PQDT - Global.
10. Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2020b). Diabetic retinopathy detection through deep learning techniques: A review. In Informatics in Medicine Unlocked (Vol. 20). Elsevier Ltd. https://doi.org/10.1016/j.imu.2020.100377
11. Karthik, Maggie, Sohier Dane. (2019). APTOS 2019 Blindness Detection. Kaggle. https://kaggle.com/competitions/aptos2019-blindness-detection
12. Sutton, R. T., David, P., Daniel, C. B., Richard, N. F., & Karen, I. K. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. In npj Digital Medicine (www.nature.com/npjdigitalmed).
13. Omididan, Z., & Hadianfar, A. (2011). The role of clinical decision support systems in healthcare (1980-2010): a systematic review study. Jentashapir Sci.-Res Q (2) 125-134.