Seeing Potential: Deep Learning of Retinal Imaging

Yiyang (Jessie) Yu
Computer Science and Software Engineering
University of Canterbury
Christchurch, New Zealand
yyu69@uclive.ac.nz

Andrew Bainbridge-Smith
Computer Science and Software Engineering
University of Canterbury
Christchurch, New Zealand
andrew.bainbridge-smith@canterbury.ac.nz

***Abstract*—With the increasing prevalence of diabetes worldwide, the need for timely and accurate retinal screenings is crucial. Manual grading processes can be time-consuming and prone to errors, leading to delays in patient treatment. To address this problem, a deep learning model will be trained on a dataset of retinal images to support medical experts in classifying Diabetic Retinopathy (DR) grades.**

To label the training data, an existing web application is extended, allowing medical professionals to assign retinal grades. However, inconsistencies in manual grading among medical professionals have been observed, and incorrect training data can lead to incorrect model predictions. The study aims to assess the reliability of medical professionals and deep learning models in DR grading.

The objective of this research is to compare the performance of the deep learning model with that of human experts in grading retinal images for DR. By comparing the performance of medical professionals with the deep learning model, this study quantifies sensitivity, specificity, and accuracy. This research contributes to advancing automated DR grading systems and exploring a deep learning model's capacity to provide accurate diabetic retinal screenings.

***Index Terms*—Retinal Imaging, Computer Vision, Image Classification, Deep Learning, Diabetic Retinopathy, Fundus.**

I. Introduction

Diabetic retinopathy (DR) is a leading global cause of vision impairment [1]. It can be prevented through early detection, periodic retinal screenings, and medical treatments [2]. As diabetes prevalence continues to rise globally, it is critical to ensure that the future of our health system can keep providing prompt access to diabetic retinal screenings. By 2040, it is projected that more than 600 million people between the ages of 20 and 79 will live with diabetes, requiring regular screenings [3]. Currently, the screening process in New Zealand involves capturing fundus images using a digital camera. These images need to be examined and graded by trained professionals, which introduces delays and the potential for human errors. The dependency on professionals for accurate image readings poses a bottleneck in managing the increasing number of patients the system needs to expand to manage.

Automating the screening process for patients can free up resources and professionals' time to attend to other medical areas needing their care. Predictive technology, such as deep learning, holds promise in aiding the detection and grading of the severity of DR [4]. The objective of this research is to investigate the potential of deep learning solutions in automating DR screening and make comparisons with the current manual process done by professionals.

The next sections will give context to the background and objectives in Section II, followed by the related work in this field in Section III, the proposed solution for this project in Section IV, the plan to achieve this in Section V, and conclusion of current findings in Section VI.

II. Background and Objectives

Technology-aided screening can provide valuable information for developing treatment plans to mitigate the effects of DR. One crucial aspect of diagnosis is grading the severity of the disease, ranging from level 0 (no DR) to level 4 (proliferative DR). Deep learning classifiers can extract features from fundus images to automate the grading process.

Deep learning models require a large dataset for training. Fortunately, this study has access to an extensive retinal image database (CDHB-DR) through collaboration with the University of Otago and the Canterbury District Health Board. However, the images in the database still need to be labelled with retinal grades. To address this, an existing web application will be extended to collect labelled data. Medical professionals will be provided with an interface to assess and assign retinal grades to the images.

The reliability of assessments made by professionals is crucial for training the deep learning model. Reliability in this context refers to consistent results when the same image is repeatedly assessed by the same professional or by different professionals. Studies have shown inconsistencies in manual grading among medical professionals [5]. These inconsistencies further emphasise the accumulating evidence that a deep-learning model needs to be developed to provide additional support to medical professionals with the grading process. This study will explore the reliability of medical professionals in DR grading through the context of sensitivity, specificity, and accuracy.

In a study by Abràmoff et al. [5], DR experts graded one image per eye, and the sensitivity and specificity were estimated by comparing individual gradings with the average of other experts. This comparison helped quantify the performance of medical professionals compared to a deep learning model in terms of sensitivity, specificity, and accuracy. Sensitivity represents the rate of correctly identifying patients with DR. In contrast, specificity represents the rate of correctly identifying

patients without DR. Higher sensitivity reduces the chances of missing a DR diagnosis, while higher specificity reduces false alarms and unnecessary referrals.

To ensure the reliability of individual assessments by medical professionals and to compare their performance with deep learning models, the following research questions will be addressed:

RQ1: What is the sensitivity, specificity, and accuracy of retinal grade assessments made by medical professionals compared to their own assessments?

RQ2: What is the sensitivity, specificity, and accuracy of retinal grade assessments made by medical professionals compared to each other?

RQ3: What is the sensitivity, specificity, and accuracy of a deep learning model's DR grade assessments compared to those made by medical professionals?

By answering these research questions, we aim to evaluate the performance of deep learning models and their capacity to support medical professionals with DR grading, ensuring reliable and accurate diagnoses.

III. Related Work

A. Retinal fundus images

Retinal fundus images play a crucial role in diabetic retinopathy (DR) grading and diagnosis. Table I outlines popular publicly accessible databases that have been created to provide labelled fundus images for DR research and open opportunities for quantitative comparisons between different approaches. In this section, we will review some of these databases and discuss their limitations.

TABLE I
DATABASES OF RETINAL FUNDUS IMAGE.

Database name	Year	Number of Images
STARE [6]	1975	400
DIARETDB0 [7]	2006	130
DIARETDB1 [8]	2007	89
Messidor [5] [9]	2013	1748
UoA-DR [10]	2017	200

The STARE (STructured Analysis of the Retina) database [6] is one of the oldest and most widely used databases. It was created in 1975 through scanning and filtering of fundus images, resulting in lower image quality compared to modern fundus photography techniques [11].

DIARETDB0 [7] and DIARETDB1 [8] databases offer additional value by also including manual annotations on the fundus images. DIARETDB0 contains 130 images, while DIARETDB1 consists of 89 images. However, both databases suffer from the same limitation as STARE in terms of outdated images [9]. Modern fundus photography offers higher-quality fundus images with improved resolutions and sensitivities.

The Messidor database [5] is a more recent and comprehensive database for DR grading. It contains 1,748 fundus images from 874 patients. However, it only includes a single eye per patient, which reduces the surface area available for DR grading, as the full picture of both eyes per patient could provide valuable information [12].

The UoA-DR database [10] is another recent database created for DR grading. However, it has a small dataset size with only 200 images, which increases the risk of overfitting and limits the generalizability of the trained model [13]. Fortunately, there are techniques to artificially enlarge the dataset size [14] [15], making the UoA-DR a safe backup dataset in the project's contingency plan (for R3 described in Table V of Appendix C).

To address the limitations of existing public databases, the research project has access to a private DR database (CDHB-DR), which contains 3,369 fundus images collected from 2014 to 2020. The images in CDHB-DR have been verified for adequate quality for DR grading by the previous year's researcher working on the project [16]. Having access to this private database provides a more up-to-date and larger dataset, making it a suitable choice for training and evaluating the deep learning model for predicting DR grades.

While existing public databases have contributed to DR research, they have certain limitations, such as outdated images, limited dataset size, and single-eye representation. The CDHB-DR database, with its larger size and verified image quality, offers an advantage in terms of more recent and comprehensive data for the present research project.

B. Deep learning model

Previous studies have explored technology-aided screening processes for various medical conditions, but few have specifically focused on using fundus images for diabetic retinopathy (DR) diagnosis [17]. However, recent advancements in deep learning and classification techniques have shown promise in supporting DR grading using fundus images [4] [18].

Deep learning models developed for DR diagnosis have faced limitations in terms of dataset quality, which can impact their accuracy [19]. Fortunately, the dataset used in this project (CDHB-DR) has undergone quality assurance filtering to ensure that only appropriate fundus images are included for training, thereby mitigating the quality challenge [16].

Concept Activation Vectors (CAVs), as introduced by Kim et al. [20], offer a method to understand how deep learning models classify DR grades. CAVs highlight the high-scoring features (concepts) that contribute to the model's predictions. This approach allows for a comparison between the decision-making process of deep learning models (quantified concepts) and expert knowledge from medical professionals. In cases of disagreement, it becomes possible to trace back the reasons behind the model's decision and enable medical experts to provide input and potentially correct the model's decision-making.

TCAV was applied in an example of DR grading. Hanif et al. [18] evaluated the weights of features such as microaneurysms and pan-retinal laser scars in fundus images. Their findings revealed that the model tended to "overestimate DR severity by assigning a high TCAV score to aneurysms". This

allowed experts to identify and correct the model's prediction behaviour in making those undesirably weighted predictions.

Wu et al. [14] explored a different approach to understanding DR gradings in fundus images intuitively. They used Convolutional Neural Networks (CNN) to generate image captions that support the diagnostic process of DR.

Abràmoff et al. [5] conducted one of the first studies on the computer-aided diagnosis of DR, applying the International Clinical Diabetic Retinopathy (ICDR) [21].

Previous research has made strides in utilizing deep learning models and classification techniques for DR grading that inspires techniques in approaching this research such as the use of CAV [20] [18], generative captions [14], and the ICDR [5]. However, there are still opportunities for comparative analysis. There has been limited exploration of comparing the sensitivity, specificity, and accuracy of medical professionals' DR gradings among themselves and against deep learning models using a recent database of fundus images. This project seeks to distinguish itself by conducting a comprehensive evaluation of the performance of medical professionals and deep learning models in DR grading using the CDHB-DR database.

IV. Proposed Solution

A. Design and Implementation

The end product of this project is to develop a deep learning model capable of accurately classifying DR grades. To achieve this, the proposed solution consists of two main processes, as depicted in Figure 1. Process 1 involves gathering labelled retinal imaging for evaluation, while process 2 focuses on developing the deep learning model using the labelled dataset. Both processes contribute to achieving outcomes to addressing RQ1, RQ2, and RQ3.

1) Gather labelled data: Process 1 is currently underway and involves recruiting medical professionals to evaluate DR grades and extending a web application[1].

Ethical approval and the expertise of medical professionals are necessary to label the CDHB-DR retinal image database. The CDHB-DR dataset offers recent and high-quality images in large quantities, making it suitable for training the model in Process 2. However, before training can begin, the DR grades of the CDHB-DR retinal images need to be labelled. This label-gathering process is essential to establish the expected DR grades and will provide a benchmark for evaluating the manual grading performance to address RQ1, RQ2, and RQ3.

To facilitate the annotation process for medical professionals, an existing proof-of-concept web application (Figure 2 in Appendix A) needs to be extended. This extension involves creating a new web page dedicated to labelling DR grades following the International Clinical Diabetic Retinopathy (ICDR) severity scale [21]. The web application, accessible through the GitLab repository[2], provides a convenient platform for professionals to input their fundus image annotations. The necessary technology stack for this extension includes Java, JavaScript, and Vue.js.

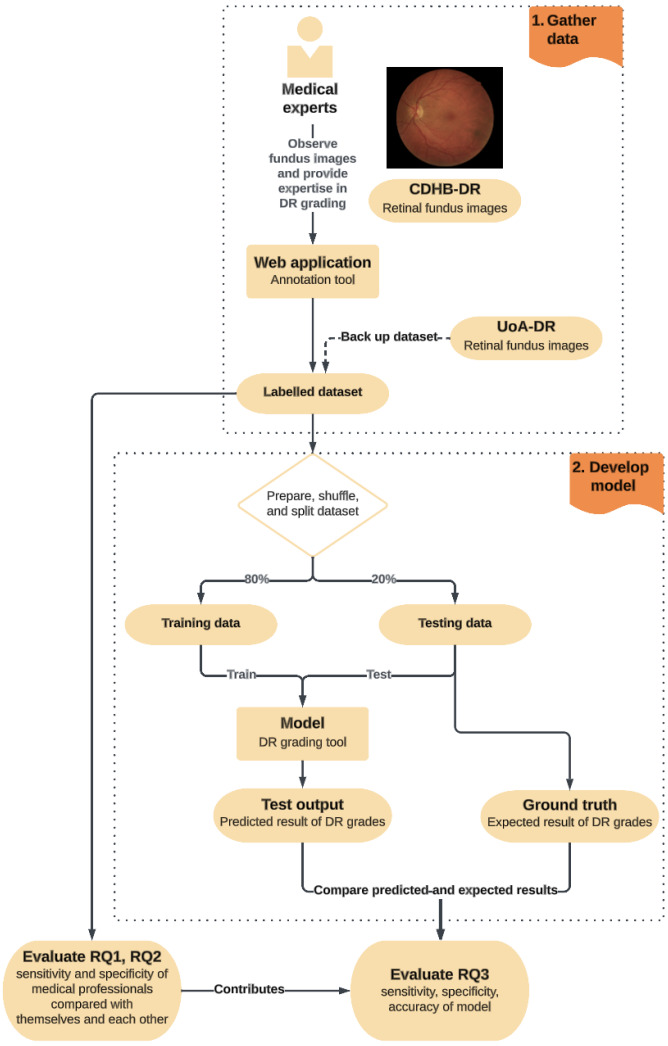

Fig. 1. Process flow of the project.

2) Develop deep learning model: Process 2 involves the development of a deep learning model using the labelled retinal images as inputs. The inputs will be prepared, shuffled, and then split into 80:20 ratios for training and testing data. Shuffling is needed to ensure the data is not ordered in a specific way that will introduce bias.

Detailed planning for this stage will occur in future milestones (M6, M7 in Table IV from Appendix C) once additional knowledge on the topic is acquired through the COSC440 course on "Deep Learning." The development process will explore the incorporation of image classification or segmentation techniques as auxiliary tasks within the deep learning model. The candidate solution will be a deep learning model capable of predicting DR grades, and the result of the model performance will be used to address RQ3.

[1]Existing web application available here: https://diabetic-retinopathy.csse.canterbury.ac.nz/Labeling

[2]GitLab repository for the proof-of-concept available here: https://eng-git.canterbury.ac.nz/yyu69/seng402-deploy

B. Method and Project Management

This project combines research and software development components to explore deep-learning approaches for DR grading and implement a web application. Therefore the design method for this project follows the subsequent processes as explained in this section.

1) Research Process: An academic literature review was conducted using IEEE Xplore, Google Scholar, and the ACM digital library. The search string used was: ('diabetic retinopathy grading' OR 'retinal imaging grading') AND ('computer vision' OR 'deep learning'). Unlike other digital libraries, Google Scholar can contain "grey" literature or research papers that have not yet undergone peer review. Therefore the search results from Google Scholar were filtered in an extra step to include only peer-reviewed papers. From the search results, papers were selected based on the relevancy of their title, abstract, and conclusions to the research scope. The snowballing technique [22] was applied to identify additional papers that mentioned interesting approaches. Papers not in English or not freely accessible were excluded.

2) Evaluation Process: To answer RQ1, RQ2, and RQ3, the evaluation process involves measuring the performance of medical experts and the deep learning model in DR grading. Various evaluation metrics will be employed, including accuracy, sensitivity, and specificity [23]. The metric formulas and their descriptions are provided in Table II, and the metric variables and their descriptions are defined in Table III.

For RQ1 and RQ2, which assess the performance of medical experts, the metric variables will be obtained from the grade-labelled dataset, as illustrated in Process 1 of Figure 1. The interpretation of the metric variables will be adapted for this research, following the approach used by Abràmoff et al. [5] for quantifying comparisons between medical professionals. The predicted cases will represent an individual medical expert's DR grade assessment, while the true cases will be calculated as follows:

1) The average of an individual medical expert's DR grade assessment repeated with the same image three times. This will be the true case for addressing RQ1.
2) The average of all medical experts' DR grade assessments for the same image. This will be the true case for addressing RQ2.

RQ3 will measure the performance of the deep learning model's DR grade assessments compared to those made by medical professionals. The test outputs from the model will serve as the predicted cases, and the ground truth of the test will serve as the true cases, as shown in Figure 1.

For deep learning models, achieving higher sensitivity and specificity is generally desirable. However, this choice is complex and depends on socio-economic considerations (e.g. the motivation for higher specificity can reduce costs and resources associated with unnecessary referrals [5]). This research will not establish a specific goal for sensitivity and specificity but focuses on comparing deep learning solutions with manual grading by asking RQ1, RQ2, and RQ3.

TABLE II
EVALUATION METRIC FORMULAS AND DESCRIPTION

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall performance of the model, indicating the proportion of correctly predicted cases.
Sensitivity	$\frac{TP}{TP+FN}$	Coverage of actual positive samples, indicating the proportion of correctly predicted positive cases.
Specificity	$\frac{TN}{TN+FP}$	Coverage of actual negative samples, indicating the proportion of correctly predicted negative cases.

TABLE III
EVALUATION METRIC VARIABLES AND DESCRIPTIONS

Metric variable	Description
TP	True positive, representing correctly predicted positive cases.
TN	True negative, representing correctly predicted negative cases.
FP	False positive, representing incorrectly predicted positive cases when the true class is negative.
FN	False negative, representing incorrectly predicted negative cases when the true class is positive.

3) Software Process: An adapted Kanban methodology [24] was used to manage project tasks and project capacity. The Kanban board consisted of five boards: "Backlog," "Next 7 Days," "Blocked," "Progress," and "Done" (Figure 3 found in Appendix B).

Tasks were created by breaking down the work increments required to meet the project milestones in (Table IV in Appendix C). New tasks are added to the "Backlog" board, progressing to completion and moving to the "Done" board. Adaption to the Kanban methodology is as follows:

- New tasks are added to the "Backlog" board before being moved to the "Next 7 days" board for prioritization.
- Task picked up for work will be moved to the "Progress" board
- Blocked tasks were moved to the "Blocked" board if they were waiting on external dependencies to enable it to progress.
- The "Progress" board was limited to two tasks at a time to help identify work that was having challenges progressing forward.

To push tasks forward in the Kanban board and ensure sufficient time was allocated to the research, ten hours of work per week were blocked out on the researcher's personal calendar. Clockify was a tool used to track and motivate meeting the weekly ten-hour goal.

4) Development Tools: The research utilized various tools for different deliverables identified from the outcomes of the

milestones in Section V. Overleaf was used for managing reports in IEEE style, while Zotero was used for reference management and organizing research papers. Software artefacts were managed using version control through a GitHub repository, and dataset artefacts were stored in the University of Canterbury's research database. Posters were created using Canva, and other documentation was recorded in Google Drive. These online resources allowed for easy access and were backed up in the cloud for data security.

5) Feedback Process: Fortnightly check-in meetings were conducted with the research supervisor, and weekly catch-up meetings were scheduled with academic professionals in Deep Learning and Computer Vision. These meetings provided an opportunity to communicate progress, discuss any challenges, and receive feedback on potential research approaches.

Feedback from medical professionals using the web application will be collected after their labelling sessions through the application. Their experiences will be shared with the research supervisor during the weekly catch-up meetings to address any potential risk in scope creeps (R4 in Table V).

V. Plan

Table IV and Table V outline the milestones and associated risks, respectively, and can be found in Appendix C. Milestones M1, M2, and M3 have been completed, and M4 is currently underway. The risk of M4 being delayed has been addressed by R3, which will be utilizing the UoA-DR database as a backup resource. The outcomes from M5 onwards will be included in the final report, with contingency plans in place to mitigate potential risks (Table V). Please refer to the tables in Appendix C for more details.

VI. Conclusion

With the global rise in diabetes prevalence, ensuring prompt access to diabetic retinal screenings is crucial. The current system relies on trained professionals to interpret fundus images manually, which introduces delays and potential human errors. The dependency on professionals to read images accurately is a stressful bottleneck in the system's capacity to handle a growing number of patients.

This research project strives to contribute to the field of computer-aided diagnostics for DR and address the challenges associated with the current screening process. This report has presented the related work, proposed solution, and plan for producing a deep learning model for classifying Diabetic Retinopathy (DR) grades. The project has involved two parallel processes: the development of the model and the gathering of retinal imaging annotations. The details of Process 1, which focuses on gathering grade-labelled data, have been outlined, and the ongoing progress of Process 2, involving the development of the deep learning model, has been discussed. The evaluation of the model's performance will be based on metrics such as accuracy, precision, recall, and F1-score.

The primary objective of this research is to answer the research questions related to the classification of DR grades and compare the sensitivity, specificity and accuracy of the deep learning model against that of a medical expert. The results obtained from the evaluation of the deep learning model will provide valuable insights into its performance and capacity to support experts in this DR grading process.

References

[1] W. H. Organization, "Blindness and vision impairment," Available at https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (2022/10/13).

[2] N. E. I. National Institutes of Health, "Diabetic retinopathy," Available at https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy#section-id-10543 (2022/07/08).

[3] K. Ogurtsova, J. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N. Cho, D. Cavan, J. Shaw, and L. Makaroff, "Idf diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," *Diabetes Research and Clinical Practice*, vol. 128, pp. 40–50, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168822717303753

[4] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D. I. Fotiadis, and K. Marias, "Deep learning for diabetic retinopathy detection and classification based on fundus images: A review," *Computers in Biology and Medicine*, vol. 135, p. 104599, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521003930

[5] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D. C. Moga, G. Quellec, and M. Niemeijer, "Automated Analysis of Retinal Images for Detection of Referable Diabetic Retinopathy," *JAMA Ophthalmology*, vol. 131, no. 3, pp. 351–357, 03 2013. [Online]. Available: https://doi.org/10.1001/jamaophthalmol.2013.1743

[6] J. Tetazoo, "Stare (structured analysis of the retina) project," Available at http://www.ces.clemson.edu/~ahoover/stare (2023/05/15).

[7] T. Kauppi, V. Kalesnykiene, J.-K. Kämäräinen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "Diaretdb 0 : Evaluation database and methodology for diabetic retinopathy algorithms," 2007.

[8] T. Kauppi, V. Kalesnykiene, J. Kämäräinen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "The diaretdb1 diabetic retinopathy database and evaluation protocol," in *British Machine Vision Conference*, 2007.

[9] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The messidor database," *Image Analysis Stereology*, vol. 33, no. 3, pp. 231–234, 2014. [Online]. Available: https://www.ias-iss.org/ojs/IAS/article/view/1155

[10] R. J. Chalakkal, W. H. Abdulla, and S. Sinumol, "Comparative analysis of university of auckland diabetic retinopathy database," in *Proceedings of the 9th International Conference on Signal Processing Systems*, ser. ICSPS 2017. New York, NY, USA: Association for Computing Machinery, p. 235–239. [Online]. Available: https://doi.org/10.1145/3163080.3163087

[11] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[12] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning," *Investigative Ophthalmology Visual Science*, vol. 57, no. 13, pp. 5200–5206, 10 2016. [Online]. Available: https://doi.org/10.1167/iovs.16-19964

[13] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?" *PLoS Computational Biology*, vol. 4, p. e27, 01/2008 2008. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0040027

[14] L. Wu, C. Wan, Y. Wu, and J. Liu, "Generative caption for diabetic retinopathy images," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2017, pp. 515–519.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017. [Online]. Available: https://doi.org/10.1145/3065386

[16] N. Huynh, "Seng402 final report," 2022, report from previous year's iteration of this project - proposes an automatic system that can effectively determine whether an image is adequate for further diabetic retinopathy screening.
[17] M. D. Abramoff and M. S. Suttorp-Schulten, "Web-based screening for diabetic retinopathy in a primary care population: The eyecheck project," *Telemedicine and e-Health*, vol. 11, no. 6, pp. 668–674, 2005, pMID: 16430386. [Online]. Available: https://doi.org/10.1089/tmj.2005.11.668
[18] A. M. Hanif, S. Beqiri, P. A. Keane, and J. Campbell, "Applications of interpretability in deep learning models for ophthalmology," *Current Opinion in Ophthalmology*, vol. 32, pp. 452 – 458, 2021.
[19] J. Rio, P. Nderitu, C. Bergeles, S. Sivaprasad, G. Tan, and R. Raman, "Evaluating a deep learning diabetic retinopathy grading system developed on mydriatic retinal images when applied to non-mydriatic community screening," *Journal of Clinical Medicine*, 01 2022.
[20] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." in *ICML*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 2673–2682. [Online]. Available: http://dblp.uni-trier.de/db/conf/icml/icml2018.html#KimWGCWVS18
[21] C. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642003004755
[22] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '14. New York, NY, USA: Association for Computing Machinery, 2014. [Online]. Available: https://doi.org/10.1145/2601248.2601268
[23] A. Baratloo, M. Hosseini, A. Negida, and G. E. Ashal, "Part 1: Simple definition and calculation of accuracy, sensitivity and specificity," *Emergency*, vol. 3, pp. 48 – 49, 2015.
[24] D. J. Anderson, *Successful Evolutionary Change for Your Technology Business*. Blue Hole Press, 2010.

APPENDIX

A. Web Application

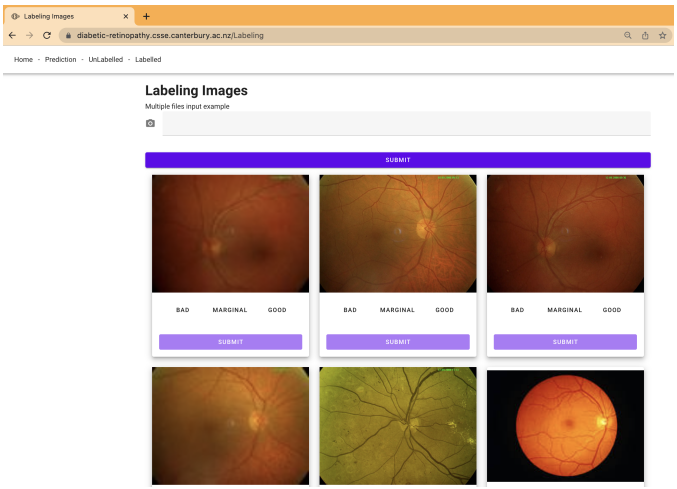

Fig. 2. Web application for labelling retinal images.

B. Kanban board

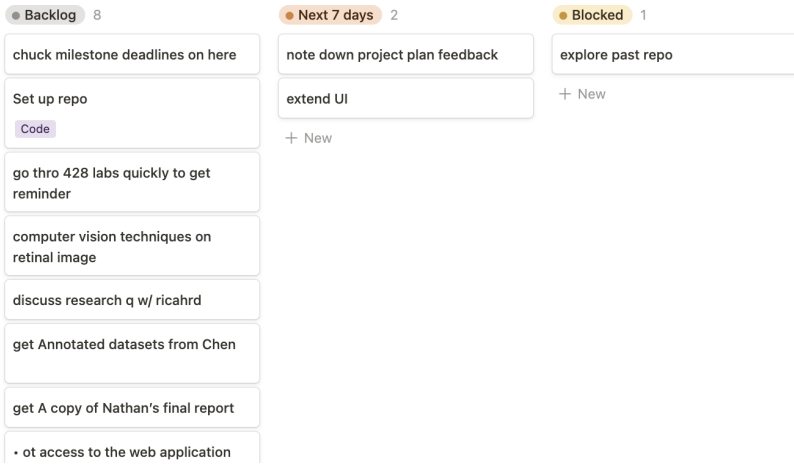

Fig. 3. Kanban board for tracking project tasks.

C. Project Plan's Milestones and Risk

Table IV and Table V outline the milestones and associated risks, respectively.

TABLE IV
TIMELINE OF PROJECT MILESTONES.

	Milestone	Approach	Tangible outcome(s)	Risk	Due
M1.	Project plan submission.	Research the background surrounding my project to make plans for approaching a goal.	Deliver a project plan to guide the research.	R1.	24 MAR
M2.	Resource set up.	Seek access to research data, retinal images, code repositories, and OpenCV.	A local development environment is ready with the resources necessary for further research work.	R2.	31 MAR
M3.	Interim report submission.	Research at least ten other papers to explore the current progress in computer-aided DR diagnostics and with those insights, develop a solution for this project. Include an updated project plan.	Delivery of the interim report as a starting point for the final report. Section IV outlines the proposed solution to be used when achieving M4, M5, M6, and M7.	R1.	02 JUN
M4.	Gather grade-labelled data.	Complete Process 1 described in Section IV-A1. This would involve sourcing ethic approval, recruiting medical professionals, and extending the current web application's capabilities for users to grade-label retina images.	Dataset of label data that feed into answering RQ1, RQ2, and RQ3 as shown in Figure 1 and for training deep learning model.	R3, R4.	20 JUL
M5.	Manual grading comparison.	Make comparisons of professional DR grading as described in the evaluation process in Section IV-B2.	Results that answer RQ1 and RQ2 as artefacts for the final report. Results also feed into answering RQ3.	R2, R5.	30 JUL
M6.	Initial deep learning model.	Demo to the research supervisor to evaluate the solution and get feedback. Evaluate the performance of the model as described in Section IV-B2.	Document improvement steps to enhance the deep learning model when answering RQ3. Code artefact to refine in the final deep learning model. Produce initial results to answer RQ3 as a benchmark to improve against in M7.	R2, R5.	10 AUG
M7.	Final deep learning model.	Enhance the initial model from M6's demo feedback. Evaluate the performance of the model as described in Section IV-B2.	Code artefact of the deep learning model. Results that answer RQ3 as an artefact for the final report.	R2, R5.	05 SEP
M8.	Abstract submission.	Use the abstract template on LEARN and submit it as a YAML file with a text extension.	Delivery of an abstract for content in the showcase booklets.	R1.	15 SEP
M9.	Poster submission.	Note down target audiences to prepare an A1-size poster with the main results.	Delivery of a project poster for the University of Canterbury's College of Engineering Showcase.	R1.	02 OCT
M10.	Presentation submission.	Prepare and practise a pepaha and project presentation.	Delivery of slide pack for the project presentation.	R1.	12 OCT
M11.	Final report submission.	Implement the proposed solution (Section IV) and evaluation process (Section IV-B2) suggested from the interim report to analyze and discuss all results to RQ1, RQ2, and RQ3 results into a scientific report. Update sections (e.g. Method, Design & implementation etc.) as necessary to reflect the most current findings from this research.	Delivery of a final scientific report.	R1.	20 OCT
M12.	Knowledge transfer material.	Prepare and proofread the code repository, documentation, and resources to ensure the continuation of the research's knowledge.	Produce an outline of the project to transfer knowledge. Code artefacts (web application and deep learning model) will be shared in a research demonstration.	R1.	20 OCT
M13.	Demonstration submission.	Four weeks prior, arrange a meeting with the research supervisor and a second marker to inspect the software-related artefacts.	Completed a feedback form.	R1.	20 OCT

TABLE V
RISK ANALYSIS FOR THIS STUDY

	Risk	Associated milestone(s)	Contingency plan
R1.	Time stress to finish submissions.	M1, M3, M8, M9, M10, M11, M12, M13. Assignment deadlines from other university courses, sicknesses and poor time planning can bring difficulties in finishing submissions.	A large submission is divided into smaller tasks and represented as iterative deadlines in my calendar. Blocks of time are also scheduled for working towards these small task deadlines and hold me accountable for the time I'd need to block out to achieve the milestones on time.
R2.	Technical difficulties.	M2, M4, M5, M6, M7. Needing more technical knowledge can lead to feeling stuck and hinder progress.	Set up weekly catch-up meetings with the research supervisor and communicate honest progress and blockage to get the best support and guidance on potential methods to approach the research.
R3.	Delay in gathering labelled data.	M4. Difficulty in getting the appropriate paperwork and process ready for ethics approval, developing the web application, or recruiting medical experts. There is a risk in the time medical experts can take to label large datasets manually. Figure 1 shows the first data gathering is essential in obtaining a labelled dataset as the future milestones and processes in the figure are dependent on it.	Noted as a high priority and urgent task item to complete compared to other tasks. Communication updates with the research supervisor cc'd into email so progress can be pushed forward. Figure 1 incorporates the use of the UoA-DR [10], which will act as an optional backup dataset of labelled retinal images if the CDHB-DR do not get labelled in time. Other labelled datasets in Table I can be used as a backup too; however, they will inflict limitations as described in Section III-A, so this will require an immediate discussion with the supervisor for potential project scope change. Therefore contingency plan A would be to use UoA-DR as backup, and plan B would be to use the datasets in Table I.
R4.	Scope creep in the data collection method.	M4. The web interface for collecting labelled retinal grading can be difficult and disengaging for professionals to input their assessments. Users making frequent decisions via mouse-clicking of drop downs, radio buttons, or input fields may experience decision and user experience fatigue. Therefore the web interface can be a medium influencing the results of the data labelling, changing the scope of the research to improve upon the user interface.	Ask the users after their labelling how their experience was. Their answer will be shared with the research supervisor during the weekly catch-ups. If any influences to the data collection method were identified, discussions would be made on the scope to keep progressing the overall motivation.
R5.	Output different than expected.	M5, M6. M7. Due to novice experience anticipating the unexpected in the research domain.	Prepare for flexibility and potential adaptations to the project's plan. Work on appropriate response plans through honest communication during weekly catch-ups and stand-ups with the research supervisor and computer vision experts.