Grading of diabetic retinopathy using a pre-segmenting deep learning classification model

Validation of an automated algorithm

Similié, Dyllan Edson; Andersen, Jakob K.H.; Dinesen, Sebastian; Savarimuthu, Thiusius R.; Grauslund, Jakob

Published in:
Acta Ophthalmologica

DOI:
10.1111/aos.16781

Publication date:
2025

Document version:
Final published version

Received: 28 February 2024 | Accepted: 5 October 2024

DOI: 10.1111/aos.16781

Acta Ophthalmologica

ORIGINAL ARTICLE

Grading of diabetic retinopathy using a pre-segmenting deep learning classification model: Validation of an automated algorithm

Dyllan Edson Similié[1] | **Jakob K. H. Andersen**[2,3] | **Sebastian Dinesen**[1] | **Thiusius R. Savarimuthu**[2] | **Jakob Grauslund**[1,3,4,5]

[1]Department of Ophthalmology, Odense University Hospital, Odense, Denmark

[2]The Maersk Mc-Kinney Moeller Institute, SDU Robotics, University of Southern Denmark, Odense, Denmark

[3]Steno Diabetes Center Odense, Odense University Hospital, Odense, Denmark

[4]Department of Clinical Research, University of Southern Denmark, Odense, Denmark

[5]Department of Ophthalmology, Vestfold Hospital Trust, Tønsberg, Norway

Correspondence
Jakob Grauslund, Department of Ophthalmology, Odense University Hospital, Odense, Denmark.
Email: jakob.grauslund@rsyd.dk

Abstract

Purpose: To validate the performance of autonomous diabetic retinopathy (DR) grading by comparing a human grader and a self-developed deep-learning (DL) algorithm with gold-standard evaluation.

Methods: We included 500, 6-field retinal images graded by an expert ophthalmologist (gold standard) according to the International Clinical Diabetic Retinopathy Disease Severity Scale as represented with DR levels 0–4 (97, 100, 100, 103, 100, respectively). Weighted kappa was calculated to measure the DR classification agreement for (1) a certified human grader without, and (2) with assistance from a DL algorithm and (3) the DL operating autonomously. Using any DR (level 0 vs. 1–4) as a cutoff, we calculated sensitivity, specificity, as well as positive and negative predictive values (PPV and NPV). Finally, we assessed lesion discrepancies between Model 3 and the gold standard.

Results: As compared to the gold standard, weighted kappa for Models 1–3 was 0.88, 0.89 and 0.72, sensitivities were 95%, 94% and 78% and specificities were 82%, 84% and 81%. Extrapolating to a real-world DR prevalence of 23.8%, the PPV were 63%, 64% and 57% and the NPV were 98%, 98% and 92%. Discrepancies between the gold standard and Model 3 were mainly incorrect detection of artefacts ($n=49$), missed microaneurysms ($n=26$) and inconsistencies between the segmentation and classification ($n=51$).

Conclusion: While the autonomous DL algorithm for DR classification only performed on par with a human grader for some measures in a high-risk population, extrapolations to a real-world population demonstrated an excellent 92% NPV, which could make it clinically feasible to use autonomously to identify non-DR patients.

KEYWORDS
automated classification, decision support, deep-learning, diabetic retinopathy, validation

1 | INTRODUCTION

Diabetic retinopathy (DR) is a leading cause of vision impairment globally but can often be treated before irreversible visual loss occurs when diagnosed timely (Blindness and Vision Impairment Collaborators & Vision Loss Expert Group of the Global Burden of Disease Study, 2021). The process of diagnosing and classifying DR involves manual intervention by ophthalmologists, which is time-consuming and prone to considerable variability among different observers (Krause et al., 2018). No formal education and certification for healthcare professionals regarding DR screening exists in Denmark as well as in many other countries (Andersen, Hubel, Rasmussen, et al., 2022; Andersen, Hubel, Savarimuthu, et al., 2022).

As the fourth industrial revolution continues to reshape many aspects of modern society, it seems inevitable that a broader implementation of artificial intelligence (AI) will take place in a medical setting. Deep learning (DL) with convolutional neural networks is a new state-of-the-art branch of AI, which has substantially

increased the performance in automated image recognition (LeCun et al., 2015). In 2018, IDx-DR was the first DL algorithm to be approved by the Food and Drug Administration with the ability to detect more than mild DR (Abràmoff et al., 2018). EyeART is another algorithm that subsequently received regulatory approval with the potential to detect detection of vision-threatening DR (Ipp et al., 2021).

Creating an algorithm in silico based on datasets using high-resolution retinal images might be useful in the developmental phase, but in a clinical context, it may not translate into an overall benefit (Gulshan et al., 2016; Khan et al., 2021). It is of paramount importance that DL algorithms are trained and validated on real-world datasets from relevant populations, representing a range of different image qualities, levels of disease and lesions in order to increase the performance of the classification as well as the clinical usefulness (Gulshan et al., 2016; Khan et al., 2021). For the creation of tailored screening and treatment plans for individual patients, distinguishing between the different levels of DR is important and the addition of a lesion-specific segmentation mask to detect individual DR lesions may further improve the clinical usefulness. In 2022, we developed a model with pre-segmentation of DR-related retinal abnormalities, which compared to a model developed on raw image features increased the mean per class accuracy from 54.3% to 70.4% and performed on par with a human expert grader to detect microaneurysms, haemorrhages, hard exudates, cotton wool spots, intraretinal microvascular abnormalities, retinal neovascularization and photocoagulation scars (Andersen, Hubel, Rasmussen, et al., 2022; Andersen, Hubel, Savarimuthu, et al., 2022).

The aim of the study was to validate the accuracy of DR grading by comparing the performance of three different models against a gold standard. In specific, we aimed to evaluate the grading results achieved by (1) a certified grader working independently, (2) the same independent grader with the assistance of a DL algorithm and (3) the DL algorithm working autonomously.

2 | METHODS

2.1 | Data collection

We included 500 real-world mydricatic 6-field retinal images from the regional DR screening center at Steno Diabetes Center Odense, Denmark (Grauslund, 2022). Retinal images were captured using Topcon TRC-NW8 (before March 1, 2019) and Topcon DRI OCT Triton (after March 1, 2019), Topcon, Tokyo, Japan, respectively. To mimic a real-life scenario with different retinal image qualities and equipment, the resolution and image quality of the retinal images varied with a factor of 5.2 from 2906 × 2824 to a maximum of 6528 × 6528 (the resolution on which the algorithm was trained; Andersen, Hubel, Rasmussen, et al., 2022; Andersen, Hubel, Savarimuthu, et al., 2022) with a mean resolution of 3836 × 3595. Of these, 19% of the images had the highest resolution. All images were already graded by an expert grader (JG with 17 years of DR grading experience) prior to the study, which served as a gold standard. To validate the accuracy of the ground truth grading, a second grader (SD with 7 years of experience in retinal imaging and DR grading) regraded the entire dataset and there was a full and weighted agreement of DR classification between the two graders of 84.4% (κ 0.81) and 95.4% (κ 0.89). This was higher than the exact-level kappa agreement of 0.52–0.78, which has previously been reported between retinal experts (Grzybowski et al., 2022). The grading was performed using the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR), consisting of five levels: no DR, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR and proliferative DR (PDR) (Wilkinson et al., 2003). The gold-standard evaluation was not known for the certified grader nor the DL algorithm. To test the robustness of the DL algorithm in a high-risk population, each level was represented approximately equally (n = 97, 100, 100, 103 and 100 for DR levels 0–4, respectively), but the exact number in each level was not known prior to completion of grading.

The AI model used for the study has previously been developed, as presented by Andersen, Hubel, Rasmussen, et al. (2022); Andersen, Hubel, Savarimuthu, et al. (2022). In brief, this was an automated DR classification algorithm based on lesion-specific pre-segmentation, which was trained to present a full-scale ICDR scale DR classification. At first, 34 075 different DR lesions were used to construct the segmentation model, and, second, 31 325 expert-annotated retinal 6-field images was used to construct the final classification model. The pre-segmentation-model performed on par with a retinal expert, and increased the grading performance of the final model by 29.7% for average per class accuracy, which resulted in a final multiclass macro area under the curve of 0.92. The model used was a variation of the U-net encoder-decoder architecture equipped with an Inception-v3 encoder. To build the AI algorithm, we used retinal images from 5127 patients with diabetes from Odense University Hospital (mean age 54.7 years, 40.0% with type 1 diabetes), and the data set was divided into graining, tuning and test using a 75%, 10% and 15% split.

2.2 | Comparison

Prior to analysis, the human grader was certified by a validated virtual ocular learning platform for DR grading (VIOLA) (Andersen, Hubel, Rasmussen, et al., 2022; Andersen, Hubel, Savarimuthu, et al., 2022). VIOLA is a virtual interactive platform for education and certification of DR healthcare professionals in the Region of Southern Denmark, consisting of lectures and exercises corresponding to the different levels of DR and the related lesions, concluded with a final certification exam.

Model 1 consisted of the certified human grader alone classifying the DR level for each of the 500 retinal images. During analysis, two 27-inch, 4k monitors were utilized to ensure an adequate level of detail.

This was followed by Model 2, which consisted of the same human grader with assistance from the DL

algorithm grading the same 500 retinal images that were now presented in a randomized order to avoid recall bias from Model 1. Finally, an analysis was performed by Model 3 with the DL algorithm classifying the retinal images autonomously without human input.

For randomization, we used Python with the Mersenne Twister random number generator. A 'seed' was set to 0, ensuring the same order of the retinal images each time the DL algorithm was started. The assistance by the DL algorithm in Model 2 consisted of segmentation data with the different lesions characterizing each level as well as the proposed classification (Andersen, Hubel, Rasmussen, et al., 2022; Andersen, Hubel, Savarimuthu, et al., 2022). In addition, the DL algorithm presented a certainty (as represented with a *p*-value) on the proposed level of DR.

In order to calculate sensitivity and specificity, a cutoff was set to level 0 (no DR) vs. level 1–4 (any DR) for the comparison of Models 1–3 and the gold-standard reference. Likewise, we calculated the positive predictive value (PPV) and negative predictive value (NPV) using both the DR prevalence of the present dataset (80%) and the real-world prevalence of 23.8% among patients with diabetes in the Region of Southern Denmark (Larsen et al., 2017).

For the comparison of Model 2 against the gold standard, we only used the most prevalent lesion for stratification based on lesions when multiple lesions were detected. If the lesion characterizing the level assigned by the reference standard was detected in the segmentation feature but for some reason did not reflect the assigned level, 'caught in segmentation' was noted. Due to the nature of DL with convolutional neural networks, some instances where a level was assigned by the algorithm without any lesions characterizing the level in the segmentation mask occurred. This was noted as 'unknown'.

2.3 | Statistical analysis

All statistical analyses were performed using StataBE 17 (StataCorp, College Station, TX, USA). Intergrader agreement was assessed by a weighted kappa, considering an interrater deviation of one DR level as a perfect agreement between the respective models. Sensitivity and specificity were calculated with the above-mentioned cutoff of no DR versus any DR.

3 | RESULTS

The weighted kappa was 0.88, 0.89 and 0.72 for Models 1–3 when compared to the gold standard (Figure 1). When comparing Models 1 and 2 with each other, differences in DR classification were detected for five or fewer cases per level (κ 0.96) (Figure 1). Comparing Models 1 and 2 with Model 3, the weighted kappa was 0.72 and 0.76, respectively (Table 1).

The sensitivity for Models 1–3 was 95%, 94% and 78%, and specificities were 83%, 84% and 81% (Table 2). The PPV for the respective models using a high DR prevalence of 80% in this study were 96% for both Models 1 and 2, and 95% for Model 3, and the corresponding NPV were 80%, 77% and 48%. Using the real-world DR prevalence of 23.8% for Funen Island, Denmark (Larsen et al., 2017), the PPV were 63%, 64% and 57%, and the NPV was 98% for Models 1 and 2, and 92% for Model 3 (Table 2).

Assessing the differences in the lesions assigned by Model 3 and the reference standard, the most common disagreement was due to Model 3 missing microaneurysms ($n=26$), detecting noise as haemorrhages ($n=24$) and simply detecting noise (artefacts) as

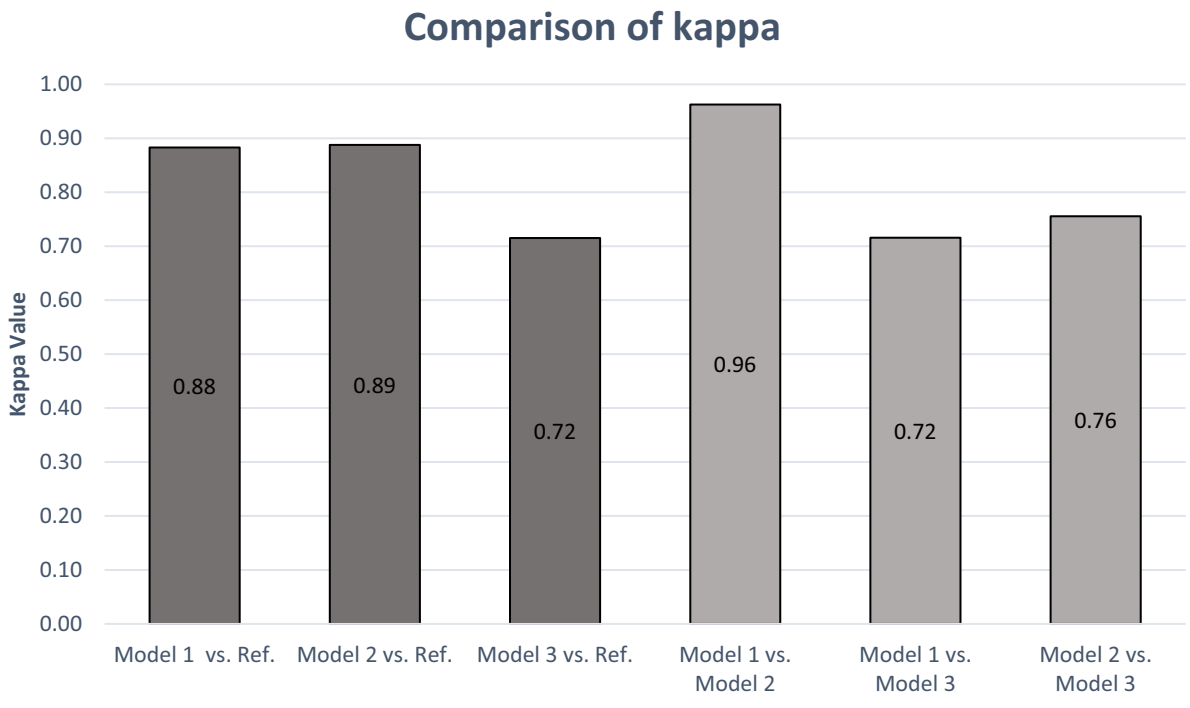

FIGURE 1 Comparison of the weighted kappa considering a deviation of one level as a full agreement between Models (1) a certified human grader without, and (2) with assistance from a deep-learning (DL) model and (3) the DL algorithm acting alone, compared to a human expert (JG) that served as the reference standard ('Ref.') and to each other, respectively.

TABLE 1 Agreements in diabetic retinopathy grading in 500 6-field retinal images between three independent models: (1) a certified human grader without, and (2) with assistance from a deep-learning (DL) algorithm and (3) the DL algorithm acting alone, compared to a human expert (JG), which served as the reference standard, respectively.

	Reference standard					
	Level 0	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Total**
Model 1						
Level 0	80	16	4	0	0	100
Level 1	5	65	23	5	1	99
Level 2	8	13	58	22	0	101
Level 3	3	5	15	74	3	100
Level 4	1	1	0	2	96	100
Total	97	100	100	103	100	500
Model 2						
Level 0	81	17	7	0	0	105
Level 1	5	70	20	5	1	101
Level 2	8	10	59	22	0	99
Level 3	2	2	14	73	3	94
Level 4	1	1	0	3	96	101
Total	97	100	100	103	100	500
Model 3						
Level 0	79	47	34	5	1	166
Level 1	5	22	11	6	0	44
Level 2	8	25	34	26	0	93
Level 3	0	0	13	61	5	79
Level 4	5	6	8	5	94	118
Total	97	100	100	103	100	500

Note: The respective levels correspond to the five levels of the International Clinical Diabetic Retinopathy Disease Severity Scale (Wilkinson et al., 2003). The varying shades of grey correspond to the different levels of agreement between the respective models 1-3 and the reference standard, with darker gradients representing a higher levels of agreement and lighter gradients representing lower levels of agreement.

TABLE 2 Sensitivity and specificity for the respective models (1) a certified human grader without, and (2) with assistance from a deep-learning (DL) algorithm and (3) a DL algorithm acting alone, compared to a human expert (JG), which served as the reference standard.

	Sensitivity	**Specificity**	**PPV (prevalence 80%)**	**NPV (prevalence 80%)**	**PPV (prevalence 23.8%)**	**NPV (prevalence 23.8%)**
Model 1	95.0	82.5	95.8	80.0	62.9	98.2
Model 2	94.0	83.5	96.0	77.1	64.0	97.8
Model 3	78.4	81.4	94.6	47.6	56.9	92.4

Note: A cutoff was set between level 0 (no diabetic retinopathy [DR]) and levels 1–4 (any DR), using the International Clinical Diabetic Retinopathy Disease Severity Scale (Wilkinson et al., 2003). The positive predictive value (PPV) and negative predictive value (NPV) were calculated using both the prevalence in this study (80%), and the prevalence among patients with diabetes in the Region of Southern Denmark (23.8%) (Larsen et al., 2017).

photocoagulation scars ($n=20$) ('noise as photocoagulation scars') where no DR lesions were present (Figure 2). Furthermore, a category of 'unknown' discrepancies was necessary to construct due to the aforementioned nature of DL ($n=20$) (Figure 2).

4 | DISCUSSION

In this validation study of an autonomous pre-segmenting DR classification DL algorithm tested in a high-risk real-world population, the findings revealed that the stand-alone AI model had a comparable specificity but a lower sensitivity when compared to a certified human grader either acting with or without assistance from the DL algorithm. However, projecting the results to a representative DR screening population, comparable PPV and NPV between all three models were found, indicating that the stand-alone DL algorithm is acting on par with a certified human grader for the identification of any DR in a real-life setting.

Models 1 and 2 achieved an almost identical performance in the grading DR on retinal images across all levels of DR, with a weighted kappa of 0.88 and 0.89, respectively. The intragrader agreement between Models 1 and 2 was very high (weighted κ: 0.96) (Figure 2), demonstrating a high uniformity in the grading process, but also underlining the fact that Model 3 seldom caused a change in the level assigned by Model 1. However, when a change was made, it was generally in agreement with the reference standard (Table 1).

Regarding the sensitivity and specificity, Models 1 and 2 performed very similar, with Model 3 performing slightly lower, achieving sensitivities of 95%, 94% and

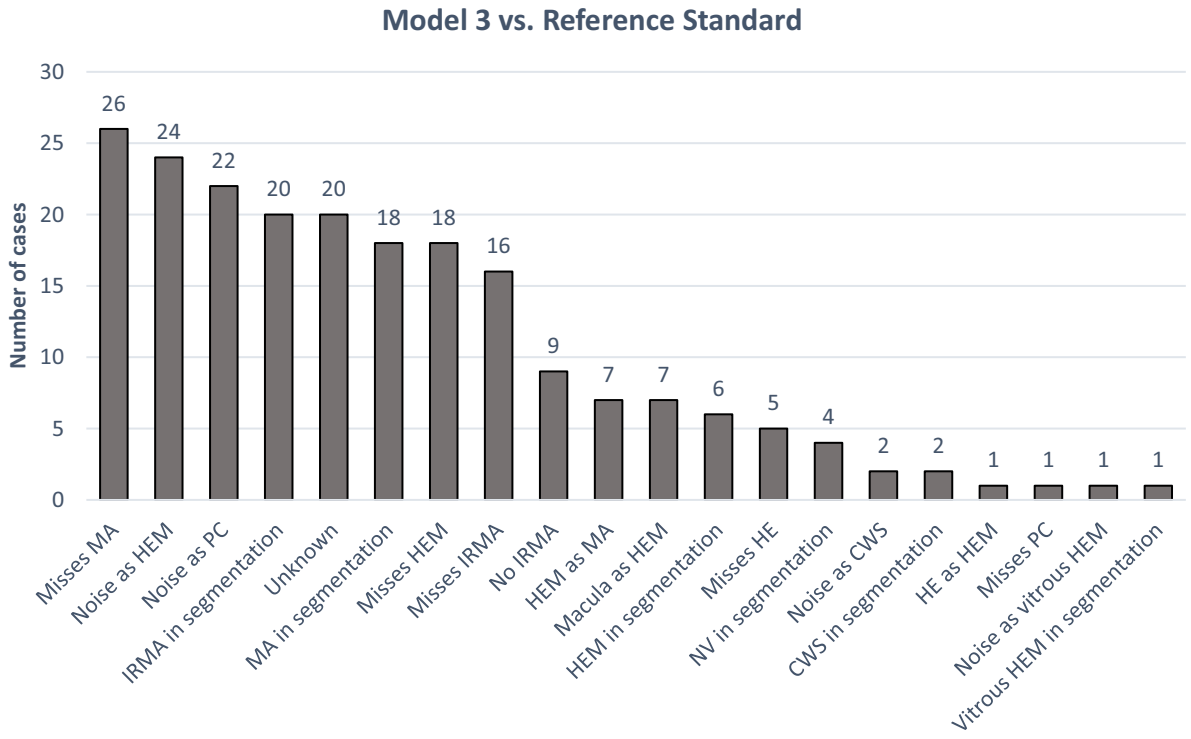

FIGURE 2 Comparison of disagreement between a deep-learning model (Model 3) and a human expert (JG) (reference standard), in the lesions assigned to the respective retinal images during the grading process. CWS, cotton wool spot; HE, hard exudate; HEM, haemorrhage; IRMA, intraretinal microvascular abnormality; MA, microaneurysm; Noise as PC, artefact (e.g. dust on the camera lens being perceived as photocoagulation scars [PC]); NV, new vessels.

78%, respectively (Table 2). On the contrary, the specificities for all the three models were very similar (82%, 84% and 81%, respectively) (Table 2), which indicated that even in a high-risk population the autonomous DL algorithm performed just as well as a human grader in the identification of retinal images without DR. This was supported for the extrapolating analysis of a real-world representative DR prevalence of 23.8% (Larsen et al., 2017), where Model 3 achieved a comparable PPV and NPV to that of Models 1 and 2 (PPV 63%, 64% and 57%, respectively) (NPV 98%, 98% and 93%, respectively) (Table 2), indicating the possibility of using Model 3 as a stand-alone tool for DR screening in a real-world setting.

When comparing the difference in lesions assigned by Model 3 to that of the reference standard leading to a different classification, the majority of the cases were regarding the detection of noise, most likely originating from artefacts, which occurred in 49/210 cases of disagreement (Table S1). Since artefacts are likely to be present in a real-life scenario, the retinal images were still analysed to examine the robustness of the respective models, but nevertheless underlines the importance of assessing the gradeability of a photo before being analysed to avoid unnecessary and unreliable grading.

Another frequent reason for disagreement in the levels assigned by Model 3 and the reference standard was the detection of MA, which was completely missed in 26 cases, but caught in the segmentation feature in 18 cases (Figure 2). This lesion type represents less than 0.5% of the total amount of pixels (Andersen et al., 2020). For real-world imaging that is often of suboptimal quality due to artefacts and blurred ocular media, the detection thereof is made difficult, with a performance of Model 3 in a prior study of 57% (Andersen et al., 2020). A study published in Tavakoli and Nazar (2020) reached a sensitivity between 80% and 87% for different vessel segmentation methods using convolutional neural networks in the detection of MA. Furthermore, in our study, the assistance provided by Model 3 to Model 2 aided in correctly identifying an additional 5 level 1 DR images (i.e. microaneurysms) (Table 1).

Moreover, in 51 cases where disagreement occurred, the segmentation mask of Model 3 caught lesions representing the level corresponding to the reference standard but did not assign the correct classification. Just because a lesion was caught in the segmentation mask does not mean it was correct in either presence or location, but rather would have contributed to a correct classification. In some instances, the proposed classification by Model 3 did not correspond to the lesions assigned, wherefore a category of 'unknown' was created (Table S1). A potential solution to the above-mentioned causes of disagreement in lesion detection would be the implementation of an error correction feature in the DL algorithm, giving the clinician the opportunity to not just override but simultaneously correct and improve the algorithm, preferably done in real time.

Some of the earlier DL algorithms in this field use a binary classification system with DR being either 'not referable' (mild or no DR) or 'referable' (moderate or worse DR) (Li et al., 2022) or did not use a segmentation mask (Abràmoff et al., 2018). This, however, leaves out information regarding the severity of the disease and is not suitable for all settings. In Denmark, referral is only needed for vision-threatening DR. Therefore, assigning a photo with a single level 2 lesion (i.e. a haemorrhage) to the same category as severe PDR would lead to a 90% false positive rate in referrals (Grauslund et al., 2020).

Therefore, in this study, a cutoff was set between level 0 (no DR) and levels 1–4 (any DR), thereby examining the ability of the AI model to correctly filter out level 0 DR retinal images. This again highlights the importance of exact classification and thereby the implementation of segmentation into the grading process, which in addition would also contribute to reducing the opaqueness generally associated with deep learning by increasing the interpretability.

4.1 | Strengths and limitations

The overall strength of our algorithm is the ability to grade six-field retinal images with exact five-level classification according to the ICDR scale as well as automated full segmentation of all DR lesions. From a clinical point of view, this would decrease the workload of ophthalmologists in DR grading and might also improve performance. To exemplify, automated identification of images with no DR (which was the principal endpoint of this study) would enable ophthalmologists to prioritize their effort in evaluating high-level images. Second, we have previously reported that even trained ophthalmologists have a tendency to underestimate the level of DR by as much as 78.6% (Thykjær et al., 2023). Automated identification of all DR lesions in a given retinal image, would lead to a better decision support, as it will automatically identify high-level lesions (e.g. intraretinal microvascular abnormalities and new vessels), thereby decreasing the risk of missing these.

The images used in this study represented all DR levels equally, contributing to the increased robustness of the models in detecting uncommon levels/lesions. However, because Europe has the highest number of ophthalmologists per patient with vision-threatening DR (18/1000) (Teo et al., 2020), few individuals might reach high-risk PDR in Denmark without undergoing screening/treatment. In the present study, photocoagulation scars were the main lesions characterizing level 4 DR, and the DL algorithm correctly classified 94/100 of these images. Photocoagulation scars were only one image, and the correct lesion(s) were caught in the segmentation mask of the remaining five images (Table S1). Furthermore, new vessels, representing active PDR, was the least frequent lesion in a previous study for the training set of the segmentation mask used by Model 3, which contained 157 occurrences (compared to 11 024 microaneurysms) (Andersen, Hubel, Rasmussen, et al., 2022; Andersen, Hubel, Savarimuthu, et al., 2022). Like microaneurysms, new vessels represents very few pixels, highlighting the necessity of high-resolution retinal images. In this study, the image resolution varied with a factor of 5.2, which contributed to the difficulty in the detection of small lesions. However, new vessel detection was performed by Tang et al. (2021) on a Malaysian dataset with a sensitivity and specificity of 99% and 88%, respectively, proving the possibility of a high performance despite several difficulties.

Although the certified human grader in Models 1 and 2 had little prior experience in the grading of DR images, his performance was comparable to that in Krause et al. with three ophthalmologists reaching an agreement with their reference standard, measured by the quadratic weighted kappa, between 0.80 and 0.84, with a majority decision of 0.87, compared to that of Models 1 and 3 in this study of 0.88 (Figure 1). Nevertheless, we cannot entirely rule out potential recall bias for the human grader, although we tried to prevent this by randomizing the order of the retinal images to be evaluated in Models 1 and 2.

5 | CONCLUSION

The application of the deep-learning model has shown promising results with comparable performance to a human grader in a real-world prevalence of DR, providing a possibility of the potential use in a clinical setting, in particular, to rule of DR. We are currently transferring the DL algorithm into clinically implementable software, which can be used in real-world DR grading. In parallel, we would be happy to set up collaborative efforts with international partners in order to provide external validation of the algorithm.

ORCID

Jakob K. H. Andersen https://orcid.org/0000-0003-3238-868X
Jakob Grauslund https://orcid.org/0000-0001-5019-0736

REFERENCES

Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N. & Folk, J.C. (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1, 39.

Andersen, J.K.H., Grauslund, J. & Savarimuthu, T.R. (2020) Comparing objective functions for segmentation and detection of microaneurysms in retinal images. In: *Proceedings of the third conference on medical imaging with deep learning. Proceedings of machine learning research*, Vol. 121, pp. 19–32. PMLR. [Virtual conference].

Andersen, J.K.H., Hubel, M.S., Rasmussen, M.L., Grauslund, J. & Savarimuthu, T.R. (2022) Automatic detection of abnormalities and grading of diabetic retinopathy in 6-field retinal images: integration of segmentation into classification. *Translational Vision Science & Technology*, 11, 19.

Andersen, J.K.H., Hubel, M.S., Savarimuthu, T.R., Rasmussen, M.L., Sørensen, S.L.B. & Grauslund, J. (2022) A digital online platform for education and certification of diabetic retinopathy health care professionals in the region of southern Denmark. *Acta Ophthalmologica*, 100, 589–595.

Blindness and Vision Impairment Collaborators & Vision Loss Expert Group of the Global Burden of Disease Study. (2021) Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the global burden of disease study. *The Lancet Global Health*, 9, e144–e160.

Grauslund, J. (2022) Diabetic retinopathy screening in the emerging era of artificial intelligence. *Diabetologia*, 65, 1415–1423.

Grauslund, J., Andersen, N., Andresen, J., Flesner, P., Haamann, P., Heegaard, S. et al. (2020) Reply: is automated screening for DR indeed not yet ready as stated by Grauslund et al? *Acta Ophthalmologica*, 98, e258.

Grzybowski, A., Brona, P., Krzywicki, T., Gaca-Wysocka, M., Berlińska, A. & Święch, A. (2022) Variability of grading DR screening images among non-trained reinal specialists. *Journal of Clinical Medicine*, 11, 3125.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A. et al. (2016) Development and validation of

a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316, 2402–2410.

Ipp, E., Liljenquist, D., Bode, B., Shah, V.N., Silverstein, S., Regillo, C.D. et al. (2021) Pivotal evaluation of an artificial intelligence system for autonomous detection of Referrable and vision-threatening diabetic retinopathy. *JAMA Network Open*, 4, e2134254.

Khan, S.M., Liu, X., Nath, S., Korot, E., Faes, L., Wagner, S.K. et al. (2021) A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3, e51–e66.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S. et al. (2018) Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125, 1264–1272.

Larsen, M.B., Henriksen, J.E., Grauslund, J. & Peto, T. (2017) Prevalence and risk factors for diabetic retinopathy in 17,152 patients from the Island of Funen, Denmark. *Acta Ophthalmologica*, 95, 778–786.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, 521, 436–444.

Li, F., Pan, J., Yang, D., Wu, J., Ou, Y., Li, H. et al. (2022) A multicenter clinical study of the automated fundus screening algorithm. *Translational Vision Science & Technology*, 11, 22.

Tang, M.C.S., Teoh, S.S., Ibrahim, H. & Embong, Z. (2021) Neovascularization detection and localization in fundus images using deep learning. *Sensors*, 21, 5327.

Tavakoli, M. & Nazar, M. (2020) Comparison different vessel segmentation methods in automated microaneurysms detection in retinal images using convolutional neural networks. In: *Proceedings volume 11317, medical imaging 2020: biomedical applications in molecular, structural, and functional imaging*, vol. 11317, Houston, TX: SPIE, pp. 430–439.

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*, 34, 1255–1261.

Thykjær, A.S., Andresen, J., Andersen, N., Bek, T., Heegaard, S., Hajari, J. et al. (2023) Inter-grader reliability in the Danish screening programme for diabetic retinopathy. *Acta Ophthalmologica*, 101, 783–788.

Wilkinson, C.P., Ferris, F.L., 3rd, Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M. et al. (2003) Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110, 1677–1682.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Similié, D.E., Andersen, J.K.H., Dinesen, S., Savarimuthu, T.R. & Grauslund, J. (2025) Grading of diabetic retinopathy using a pre-segmenting deep learning classification model: Validation of an automated algorithm. *Acta Ophthalmologica*, 103, 215–221. Available from: https://doi.org/10.1111/aos.16781