Forecasting the diabetic retinopathy progression using generative adversarial networks

Check for updates

Huiyu Qiao[1,8], Feilong Tang[2,3,8], Huanfen Zhou[4], Yun Cai[5], Kairou Guo[6], Jin Wang[5], Tong Ma[2], Lie Ju[2,3], Wei Feng[2,3], Zhiqiang Ma[7], Juan Chen[7], Yuan Luo[7], Bin Wang[2], Zongyuan Ge [2,3,9] ✉ & Qiansu Yang [5,9] ✉

Abstract

Background Diabetic retinopathy (DR) is the leading cause of blindness worldwide, making early prediction of DR progression crucial for effectively preventing visual loss. This study introduces a prediction framework DRForecastGAN (Diabetic Retinopathy Forecast Generative Adversarial Network), and investigates its clinical value in predicting DR development.

Methods DRForecastGAN model, consisting of a generator, discriminator, and registration network, was trained, validated, and tested in training (12,852 images), internal validation (2734 images), and external test (8523 images) datasets. A pre-trained ResNet50 classification model identified the DR severity on synthetic images. The performance of the proposed DRForecastGAN model was compared with the CycleGAN and Pix2Pix models in image reality and DR severity of the synthesized fundus images by calculating Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and area under the curve (AUC).

Results DRForecastGAN model has the lowest FID, highest PSNR and highest SSIM on internal validation (FID: 27.3 vs. 32.8 vs. 34.4; PSNR: 25.3 vs. 17.0 vs. 16.9; SSIM: 0.93 vs. 0.79 vs. 0.65) and external test (FID: 37.6 vs.45.1 vs.48.4; PSNR: 20.7 vs.15.2 vs.14.7; SSIM: 0.86 vs.0.69 vs.0.63) datasets compared with Pix2Pix and CycleGAN models. In the prediction of DR severity, our DRForecastGAN model outperforms both Pix2Pix and CycleGAN models, achieving the highest AUC values on both internal validation (0.87 vs. 0.76 vs. 0.75) and external test (0.85 vs. 0.70 vs. 0.69) datasets.

Conclusions The proposed DRForecastGAN model can effectively visualize DR development by synthesizing future fundus images, offering potential utility for both treatment and ongoing monitoring of DR.

Plain language summary

Diabetic retinopathy (DR) is a serious eye condition that can occur in people with diabetes. It requires diagnosis by ophthalmologists (eye doctors) through detailed eye exams involving screening and basic imaging and then tracking of disease progression. This study presents a new prediction tool called DRForecastGAN designed to forecast the progression of DR. The tool was tested in training, internal validation, and external test datasets of eye images of patients. In predicting DR severity, the DRForecastGAN model outperformed several existing tools used for this task. Additionally, DRForecastGAN demonstrated the ability to effectively visualize the progression of DR, offering potential benefits for the treatment planning and continuous monitoring of DR by ophthalmologists.

According to recent studies, diabetic retinopathy (DR) is a predominant cause of visual impairment among the global middle-aged workforce and the elderly[1,2]. The prevalence of DR is increasing along with the growing number of diabetes cases[3,4]. Hence, it is urgent to predict and treat DR early to avoid permanent damage to vision, emphasizing the necessity for reliable forecasting approaches.

In clinical practice, to track the progression of DR, ophthalmologists need to perform detailed retinal exams. These assessments typically include a vision screening and fundamental imaging techniques, such as retinal photography or ophthalmoscopy[5]. Given the limitations of current screening techniques in imaging the peripheral retina, ultra-widefield imaging and confocal scanning laser ophthalmoscopy provide effective

[1]School of Biomedical Engineering, Capital Medical University, Beijing, China. [2]Beijing Airdoc Technology Co., Ltd, Beijing, China. [3]Monash Medical AI Group, Monash University, Clayton, VIC, Australia. [4]Senior Department of Ophthalmology, The Third Medical Center of Chinese PLA General Hospital, Beijing, China. [5]Center of Medicine Clinical Research, Department of Pharmacy, Medical Supplies Center of Chinese PLA General Hospital, Beijing, China. [6]Department of Biomedical Engineering, Medical Supplies Center of Chinese PLA General Hospital, Beijing, China. [7]iKang Guobin Healthcare Group Co., Ltd, Beijing, China. [8]These authors contributed equally: Huiyu Qiao, Feilong Tang. [9]These authors jointly supervised this work: Zongyuan Ge, Qiansu Yang. ✉ e-mail: zongyuan.ge@monash.edu; yqs456@126.com

solutions[6]. Optical coherence tomography/angiography (OCT/OCTA), known for its ultra-high scan speed and resolution, has also been a useful tool for detecting and managing retinal abnormalities. OCT/OCTA, with the aid of artificial intelligence (AI) algorithms, has been used to assess retinal thickness, blood flow, and neovascularization in DR patients[6,7]. Advanced imaging techniques, such as adaptive optics, retinal function assessment, retinal oximetry, metabolic imaging, and fundus auto-fluorescence, have also been used in the evaluation of DR[6]. Currently, fundus imaging remains a critical tool in the clinical screening of DR. However, the interpretation of fundus images relies on skilled ophthalmologists[8]. Additionally, it is challenging for ophthalmologists to accurately evaluate the degree of DR lesions based on the limited diagnostic information from just fundus images[9]. Therefore, developing a predictive model that incorporates the generation and automated analysis of fundus images is crucial for mitigating the risk of irreversible vision loss in patients with DR.

The development of a predictive model is also beneficial when considering the varying recommendations for follow-up examinations. According to guidelines by the International Council of Ophthalmology and the American Diabetes Association, the interval for follow-up examinations can vary from <1 month to 1–2 years, depending on individual patient risk factors[5,10]. Several European studies suggest that for patients with diabetes who show no signs of DR during their initial examination, extending the screening interval to 2 or 3 years may be a cost-effective strategy[11–14]. Therefore, an individualized predictive model could serve to tailor the frequency of retinal examinations, optimizing care for patients with DR.

In developing personalized predictive models for medical conditions like DR, AI emerges as a powerful tool. Currently, AI has been applied to the automated detection and classification of DR, with notable advancements in convolutional neural networks (CNNs) such as ResNet, DenseNet, InceptionV3, InceptionResnetV2, Xception, SeNet, and so on. These networks demonstrate exceptional sensitivity and specificity in identifying and categorizing the nuances of DR[15–18]. Additionally, self-supervised CNNs employing techniques such as fuzzy clustering networks, category attention blocks, global attention blocks, or graph convolutional networks have been developed to improve the classification of DR lesions[9]. Transfer learning approaches, including Vision Transformer and lesion-aware transformer, have also contributed to training more robust and accurate models[9,19]. In order to produce stronger predictive results with reduced generalization error, ensemble learning based on multiple models has been utilized in the research[19,20].

With the emergence of generative adversarial networks (GANs), the landscape of AI in healthcare is experiencing a significant transformation. GANs are not only enhancing the quality of medical images[21,22], but also helping balance datasets by generating images of less common conditions, thus providing a more balanced view and reducing bias in machine learning models[23–25]. In addition, GANs' proficiency in translating images across various modalities and managing complex data distributions sets the stage for a promising application in healthcare: simulating future fundus images[26]. Such simulations can offer valuable insight that was previously unattainable by monitoring the progression of conditions. By accurately forecasting the progression of eye diseases, GANs could revolutionize the preventive strategies employed in ophthalmology, marking a significant leap toward proactive patient care. The impact of its capability is already evident in the generation of accurate visual representations for age-related macular degeneration, offering personalized predictions and detailed ocular graphics[27]. Yet, the exploration of GANs in creating deep learning models specifically designed for visual forecasting in DR is still in its infancy. Advancing this line of research could allow clinicians to anticipate the course of DR and intervene more effectively, thus preventing vision loss.

This study aims to propose a prediction framework DRForecastGAN (Diabetic Retinopathy Forecast Generative Adversarial Network) designed for synthesizing future fundus images in patients with DR. The framework can track the disease's progression while preserving patients' vessel identity based on an adversarial learning mechanism. Additionally, the study will investigate the clinical value of this framework in predicting the development of DR.

Our DRForecastGAN model incorporates two discriminators, indicating its ability to learn and generate realistic fundus images, with potentially greater accuracy than traditional GANs. The detection of lesion changes is also fundamental in the diagnosis of DR, so we guide the model to focus on these changes by utilizing lesion segmentation masks.

Methods
DRForecastGAN model
In preparation for the training, all fundus images and lesion segmentation masks were standardized to a consistent size. The image and mask are structured with dimensions $V \in R^{H \times W \times 4}$, signifying height ($H$), width ($W$), and four accounts for the three RGB channels used in fundus images along with one channel dedicated to the mask.

The DRForecastGAN model (Fig. 1) consists of the generator and the discriminator of fundus images based on conditional GANs. The generator
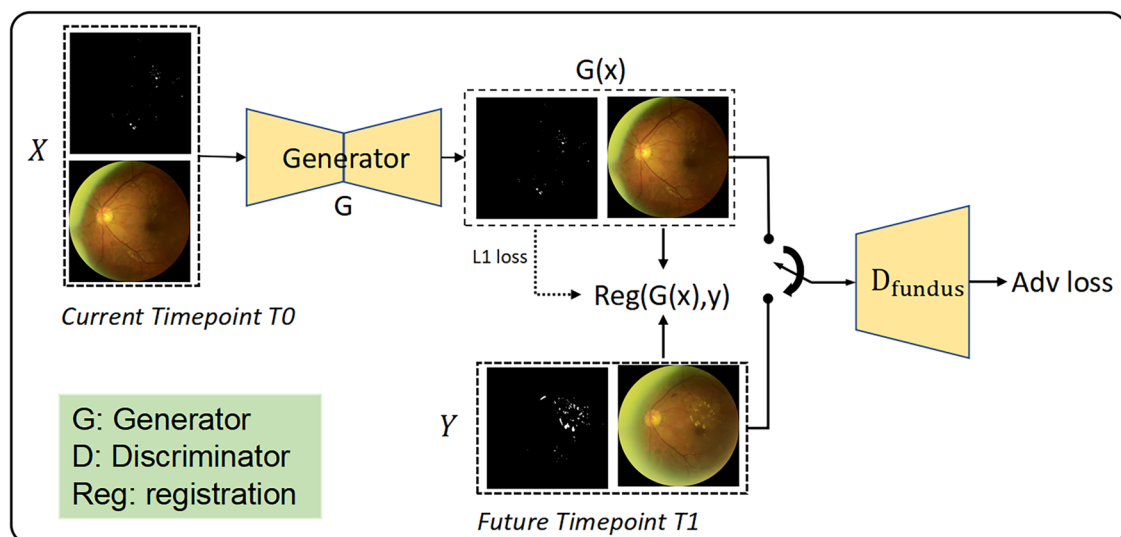
Fig. 1 | An overview of the proposed DRForecastGAN model. The model consists of the generator and the discriminator of fundus images with a registration network, to learn the mapping between current (T0) and future (T1) timepoint images.

processes a fundus image along with its accompanying lesion segmentation mask. During the training of DRForecastGAN, the model takes the current timepoint fundus image and its corresponding lesion segmentation mask as input, while using the real future fundus image and its lesion mask as the target for supervision. The lesion mask used in this study is a single-channel image with four representative lesions, including microaneurysm, hemorrhage, soft exudate, and hard exudate. All the lesions were uniformly regarded as abnormal regions and labeled as one.

The generator would synthesize a fundus image with the prediction of future lesions by combining information from the current image and lesion mask. Simultaneously, the discriminator would attempt to learn the distribution of synthetic and real future fundus images to distinguish the synthetic fundus images.

However, slightly different lesion locations may occur between longitudinal fundus images of the same patient due to the difference in angles between the device and the patient, and the changes in the device during twice image capture. Therefore, a registration network was employed to correct for these spatial inconsistencies. The registration network takes the generated image as input and produces a refined version of it. It essentially acts as a guide to adjust or align the generated images to a more consistent and accurate representation, considering the potential differences in positioning.

Assuming x_n(fundus image), $\widetilde{y}_n$(segmentation mask), $V \in R^{H \times W \times 4}$ are two modalities and $\widetilde{y}_n$ is the label for x_n, the optimization direction of the generator (G) and the discriminator (D) is given by formulas 1 and 2, respectively. $\mathcal{L}_{\text{Adv}}$ is the adversarial loss.

$$\widehat{G} = \arg \min_G \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(G(x_n), \widetilde{y}_n) \qquad (1)$$

$$\mathcal{L}_{\text{Adv}} = \min_G \max_D \mathcal{L}_{\text{Adv}}(G, D) = \mathbb{E}_y \left[\log \left(D(\widetilde{y}_n) \right) \right] + \mathbb{E}_x \left[\log \left(1 - D(G(x_n)) \right) \right] \qquad (2)$$

The registration network can reduce pixel-level differences between the current fundus images and real future fundus images (within 5 years) by adjusting them into a consistent scale and orientation. A regularization loss ($\mathcal{L}_1$) is used to optimize the registration network (formula 3). The loss function of DRForecastGAN model ($\mathcal{L}_{\text{Total}}$) is a combination of the regularization loss and the adversarial loss, with weights λ_1 and λ_2 (formula 4).

$$\mathcal{L}_1 = \min_{G,R} \mathcal{L}_1(G, R) = \mathbb{E}_{x,\widetilde{y}} \left[|| \widetilde{y} - G(x) \circ R(G(x), \widetilde{y}) ||_1 \right] \qquad (3)$$

$$\min_{G,R} \max_D \mathcal{L}_{\text{Total}}(G, R, D) = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{Adv}} \qquad (4)$$

Datasets

The data were retrospectively collected from ten iKANG Health Care centers and the Chinese PLA General Hospital between June 2018 and June 2023. The inclusion criteria are as follows: (1) patients diagnosed with DR; (2) patients underwent at least twice fundus image examination during a follow-up period of up to 5 years. All fundus images were anonymized before the analysis, removing all labels that contain patient personal information.

In total, 2969 patients from iKANG and 1689 patients from Chinese PLA General Hospital were included. Out of 2969 patients from iKANG, 2375 patients (80.0%) with 12,852 fundus images set as training datasets and 594 patients (20.0%) with 2734 fundus images set as internal validation datasets. The 1689 patients with 8523 fundus images from the Chinese PLA General Hospital were set as external test datasets. All datasets include fundus images of both left and right eyes. Patients in these datasets underwent a varying number of fundus image examinations over the follow-up period ($0 < N \leq 5$ years), with detailed information provided in Supplementary Table 1. The changes in DR severity during the follow-up period are also illustrated in Supplementary Table 2.

All the fundus images were categorized into three grades according to published criteria[28] by experienced ophthalmologists: (1) mild non-proliferative DR (NPDR): presence of only microaneurysms on the retina; (2) moderate NPDR: presence of microaneurysms, hemorrhages, cotton-wool spots, venous beading, etc; (3) severe NPDR: presence of one of the following signs without PDR: more than 20 retinal hemorrhages in each of the 4 quadrants; venous beading in 2 quadrants; abnormal retinal microvascular ties in 1 quadrant. The detailed information of the datasets is shown in Table 1.

Model training and evaluation

In the training of the DRForecastGAN model, an Adam optimizer with a learning rate of 0.0002, where $\beta 1$ (mean of the gradients) = 0.0 and $\beta 2$ (uncentered variance of the gradients) = 0.9 was utilized. The model was trained for 200 epochs, with each batch containing 4 images on the NVIDIA RTX 3090 GPU. The weights for the two discriminators in the model were set to the default value of 1. Data augmentation techniques were applied, including horizontal flipping, translation, scaling, and rotation. For the patients with more than one examination during the follow-up period, the two nearest examinations were regarded as paired source and future fundus images.

The performance of the DRForecastGAN model was compared with the performance of the CycleGAN model[29] and Pix2pix model[30] in the internal validation and external test datasets. All three models were trained on the same data. Compared with Pix2pix and CycleGAN models, DRForecastGAN integrated a registration network, a lesion-guided generator, and a dual discriminator architecture. The detailed architecture of Pix2Pix, CycleGAN and DRForecastGAN models is shown in the Supplementary Fig. 1. To ensure a fair comparison of the three models, the same Adam optimizer (learning rate = 0.0002, $\beta 1 = 0.0$, $\beta 2 = 0.9$), training epochs (200), and data augmentation techniques (Flip, Translation, Scaling, Rotation). Pix2Pix model retained its original loss function (adversarial loss + L1 loss), while CycleGAN used adversarial loss and cycle-consistency loss ($\lambda = 10$). The evaluation indexes including Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) were calculated based on synthetic and real fundus images. A lower FID value indicates that the synthetic images are closer to the distribution of real images. A higher PSNR value indicates better image quality of synthetic images. SSIM values range from −1 to 1, with a value closer to 1 indicating greater similarity between synthetic and real images. A higher PSNR value indicates better image quality of synthetic images. SSIM values range from −1 to 1, with a value closer to 1 indicating greater similarity between synthetic and real images. The comparison of three models was conducted with the same training settings.

Table. 1 | The detailed information of training, internal validation and external test datasets

	Training	Internal validation	External test
Number of individuals	2375	594	1689
Number of images	12,852	2734	8523
Age (range)	56 (18–96)	56 (20–96)	51 (19–87)
Gender (male, female)	(63.83%, 36.17%)	(63.56%, 36.44%)	(66.69%, 33.31%)
Eyes (OD/OS)	6392/6460	1470/1264	4830/3693
Follow-up years (n)	2.5 ± 0.47	2.3 ± 0.4	2.5 ± 1.2
DR level (individuals/images)[a]			
Mild NPDR	267/1413	62/308	174/895
Moderate NPDR	2052/11,182	519/2377	1423/7170
Severe NPDR	56/257	13/49	92/458

[a]The DR level represents the DR grading at the initial time point.

Clinical value assessment

The future fundus images generated by the DRForecastGAN, CycleGAN, and Pix2Pix models were compared with real future fundus images from the datasets to assess DR severity. The DR severity of the real images was determined directly from ophthalmologists' assessments in the datasets, whereas the severity of the generated images was classified by a pre-trained ResNet50 model[31]. Initially, the ResNet50 model categorized DR severity into three classes: mild NPDR, moderate NPDR, and severe NPDR.

Statistics and reproducibility

The area under the receiver operating characteristic curve (AUC) was calculated to evaluate the performance of all three models in predicting DR severity. This assessment measured the agreement between the DR severity classifications derived from the generated images and those from the real images. The accuracy, sensitivity, and specificity were also calculated for each model. Moreover, at the pixel level, the Dice score was computed for the three models by comparing the similarity between these assessments were performed on both the internal validation and external test datasets.

Ethics statement

Ethics approval for the study was obtained from the Institutional Ethics Committee of iKANG Guobin Healthcare Group Co., Ltd (LLPJ-002) and the Chinese PLA General Hospital (S2023-215). Written informed consent was obtained in the datasets. The individual informed consent of each participant was exempted.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Evaluation of DRForecastGAN

The FID of the DRForecastGAN model was lowest compared with the Pix2Pix and CycleGAN models on both internal validation (27.3 vs. 32.8 vs. 34.4) and external test datasets (37.6 vs. 45.1 vs. 48.4). The PSNR and SSIM of the DRForecastGAN model were highest compared with the Pix2Pix and CycleGAN models on both internal validation (25.3 vs. 17.0 vs. 16.9; 0.93 vs. 0.79 vs. 0.65) and external test datasets (20.7 vs. 15.2 vs. 14.7; 0.86 vs. 0.69 vs. 0.63). The input source fundus image, real future fundus image, and synthetic future fundus images from the CycleGAN model, Pix2Pix model, and the DRForecastGAN model on internal validation and external test datasets are shown in Fig. 2a, b, respectively. More visualized results of the DRForecastGAN model are shown in Supplementary Fig. 2.

Clinical value of the DRForecastGAN model

The AUC in predicting the severity of DR using the DRForecastGAN model, CycleGAN model, and Pix2Pix model on internal validation datasets was 0.87 (95% confidence interval (CI): 0.85–0.89), 0.76 (95% CI: 0.74–0.79), and 0.75 (95% CI: 0.72–0.78), respectively (Fig. 3a), and 0.85 (95% CI: 0.83–0.87), 0.70 (95% CI: 0.67–0.73), and 0.69 (95% CI: 0.66–0.71) on external test datasets, respectively (Fig. 3b). The accuracy, sensitivity, specificity, and Dice score of proposed DRForecastGAN model, CycleGAN model, and Pix2Pix model in predicting the severity of DR on internal validation datasets and external test datasets are shown in Table 2.

Discussion

In this study, the DRForecastGAN model demonstrated superior performance in synthesizing the future fundus image for DR patients by achieving the lowest FID, highest PSNR, and highest SSIM compared with the CycleGAN and Pix2Pix models. The future fundus images from the proposed DRForecastGAN model also presented the greatest AUC, sensitivity, and specificity in predicting the severity of DR among the three models. By providing ophthalmologists with a predictive visualization of the disease's trajectory, the proposed DRForecastGAN model offers a valuable tool for personalized patient care. This may aid clinicians in making more informed decisions regarding the timing and intensity of treatments, thereby enhancing the strategic planning of clinical interventions and potentially improving long-term patient outcomes.
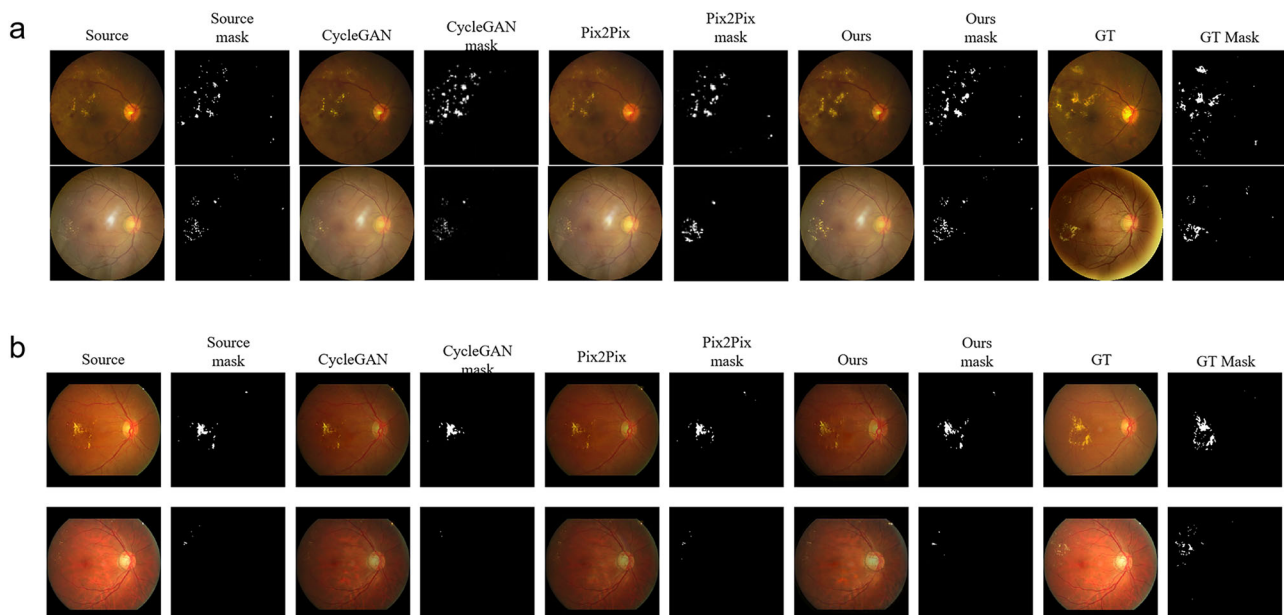
Fig. 2 | The input and output images of three models. Visual comparison of three models on the internal validation (**a**) and external test (**b**) datasets. From top to bottom are fundus images of a 51-year-old male patient's right eye, a 49-year-old female patient's right eye, a 57-year-old female patient's right eye, and a 46-year-old male patient's right eye. For the first and third cases, where the input and ground truth (GT) are moderate DR, all three models, including DRForecastGAN, Cycle-GAN, and Pix2Pix, predicted moderate DR. In the second case, where the input and GT are mild DR, the DRForecastGAN and Pix2Pix models predicted mild DR, while the CycleGAN model predicted moderate DR. In the fourth case, where the input and GT are moderate DR, the DRForecastGAN model predicted moderate DR, while the CycleGAN and Pix2Pix models predicted mild DR. Overall, the DRForecastGAN model achieved the lowest FID and the highest PSNR, SSIM, and Dice values across all four cases, demonstrating superior performance compared to the CycleGAN and Pix2Pix models.
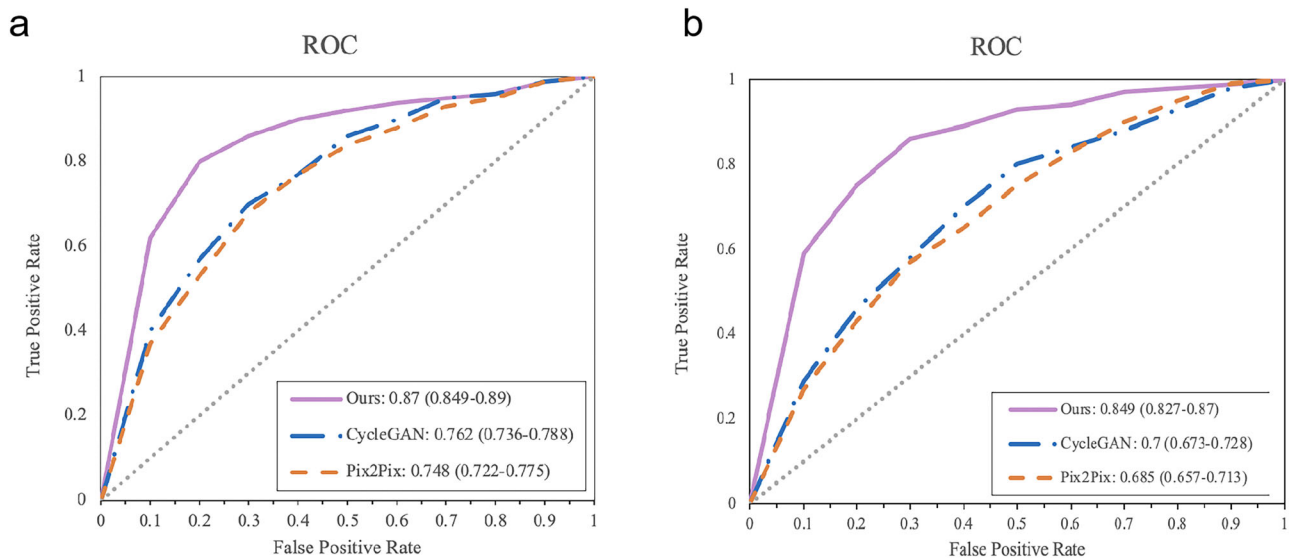
Fig. 3 | ROC curves of three models in predicting the progression of DR. Our DRForecastGAN model outperforms both Pix2Pix and CycleGAN models, achieving the highest AUC values on both internal validation (0.87 vs.0.76 vs.0.75) (**a**) and external test (0.85 vs.0.70 vs.0.69) (**b**) datasets.

Table 2 | The accuracy, sensitivity, specificity, and Dice score of three models in predicting the severity of DR on internal validation and external test datasets

Model		DRForecastGAN model	CycleGAN model	Pix2Pix model
Internal validation datasets	Accuracy	0.82 (0.80, 0.85)	0.76 (0.73, 0.79)	0.68 (0.65, 0.71)
	Sensitivity	0.77 (0.75, 0.80)	0.56 (0.53, 0.59)	0.72 (0.69, 0.75)
	Specificity	0.85 (0.82, 0.87)	0.84 (0.81, 0.86)	0.66 (0.64, 0.69)
	Dice	0.52	0.35	0.40
External test datasets	Accuracy	0.75 (0.72, 0.77)	0.61 (0.58, 0.64)	0.66 (0.63, 0.69)
	Sensitivity	0.85 (0.83, 0.87)	0.77 (0.74, 0.79)	0.58 (0.56, 0.61)
	Specificity	0.71 (0.68, 0.73)	0.55 (0.52, 0.58)	0.69 (0.66, 0.71)
	Dice	0.46	0.31	0.38

The proposed DRForecastGAN model has the best performance in synthesizing future fundus images, which might benefit from the registration network and the input of segmented lesion mask. The ablation experiment results of the segmented lesion mask and registration network are shown in Supplementary Table 3. GANs' models can synthesize realistic images, which have a wide range of applications, such as image-to-image translation, image quality enhancement, and data augmentation[32]. However, GANs may produce system bias for the small abnormality regions[33]. As a result, the CycleGAN and Pix2Pix models struggle to accurately reflect the progression of DR on the synthetic future fundus images in this work. Our DRForecastGAN model incorporates two discriminators, indicating its ability to learn and generate realistic fundus images, with potentially greater accuracy than traditional methods. Within the proposed DRForecastGAN model, the registration network realized the alignment of longitudinal fundus images to mitigate the influence of the lesion's slight variation. Besides, the segmented lesion mask provided additional information to guide the proposed model to focus on changes of DR. To further understand the behavior of segmented masks, we tested the DRForecastGAN in healthy subjects with lesion-free masks. The results demonstrated its ability to generate realistic, lesion-free images without hallucinating false lesions (Supplementary Table 4). The registration network and segmented lesion mask improved the performance of the DRForecastGAN model and made the model more generalized and applicable in clinical settings.

The progression of DR is influenced by follow-up duration. Longer follow-up periods may lead to more pronounced progression. However,

finally, the time factor was not accounted for in the proposed DRForecastGAN, CycleGAN and Pix2Pix models. The models' effectiveness could be associated with the 2-year interval between retinal imaging sessions in this study. Incorporating the time factor is necessary for the practical application of these models[27]. For accurate prediction of DR progression over varying follow-up periods, datasets should encompass images with broader time spans and diverse changes. The models can also use a recurrent neural network or an attention network to better capture the image changes over different follow-up years[34].

In this study, the individualized visualization of the development of DR is helpful for the improvement of the utilization efficiency of medical resources. The visualization of DR progression can help doctors explain possible trends of DR disease and the necessary treatment and monitoring. In addition, the recommendation of retina imaging for DR patients was based on the general population, which might not be quite valuable economically and socially for the individual sometimes[35]. The precise prediction can be used to optimize the frequency of follow-up retinal examination for the individual, which can make the medical resources more efficiently utilized. Moreover, the communication based on the future fundus images could make patients better understand their health condition and increase their understanding and cooperation with the treatment.

While the DRForecastGAN model in this study can generate future retinal images that closely resemble real images and achieve promising diagnostic results, its clinical application still faces challenges regarding generalization, stability, and interpretability. As a data-driven approach, the

DRForecastGAN model might be sensitive to the training datasets; for instance, it may produce biased results when certain lesion types are underrepresented[36]. To improve its generalization capability, it is crucial to diversify the data samples by including a broader spectrum of DR progression stages and collecting retinal images from multiple centers, thereby enhancing the model's applicability across varied patient populations and clinical settings. Additionally, during training, the generator and discriminator may encounter "mode collapse,"[37] which restricts the variety of generated images and further impedes the model's diversity. Regarding stability and interpretability, the "black-box" nature of GANs makes it difficult to understand the relationship between the generated images and actual disease progression[38], limiting the model's utility for clinical decision-making. Future efforts should therefore focus on integrating explainable AI techniques, physical information, or knowledge distillation to improve the model's stability and interpretability[39–41].

There are some limitations in our study. First, more generalized lesion masks should be segmented and put into the model. The proposed DRForecastGAN model focused on exudates in the lesion mask and might overlook the movement or disappearance of hemorrhages as the disease progresses. Second, the model's generality is limited by the datasets. The current model, focusing solely on DR patients, may have limited applicability to a broader context. For the generalization of models, more images with comprehensive lesion types should be included in datasets in the future. Third, although the registration network, data augmentation, and lesion segmentation mask in the DRForecastGAN model help mitigate lesion deviations, they cannot fully address the variations in fundus images caused by changes in illumination, optical aberrations, and image warping resulting from the scan and optics configuration. In the future, more specific data augmentation, image normalization, and more diverse datasets might be included in the model's training. Finally, enhancing the model's interpretability is crucial, making it more comprehensible and trustworthy for physicians and clinicians.

In conclusion, the proposed DRForecastGAN model can effectively visualize the development of DR by synthesizing future fundus images, offering potential utility for both treatment and ongoing monitoring of DR.

Data availability

All data are private data and not available to the public. Collaborations are welcome, and applications for controlled access to private data are assessed based on each individual case and subject to approval from the institutional review board (ce.rong@ikang.com and 301irb@sina.com). The numerical results of Fig. 3 are provided in the Supplementary Data 1.

Code availability

The code for DRForecastGAN is archived on Zenodo with https://doi.org/10.5281/zenodo.15747872[42].

References

1. Zhou, C. et al. Visual impairment and blindness caused by retinal diseases: a nationwide register-based study. *J. Glob. Health* **13**, 04126 (2023).
2. Yau, J. W. et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* **35**, 556–564 (2012).
3. Teo, Z. L. et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology* **128**, 1580–1591 (2021).
4. Ting, D. S., Cheung, G. C. & Wong, T. Y. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin. Exp. Ophthalmol.* **44**, 260–277 (2016).
5. Solomon, S. D. et al. Diabetic retinopathy: a position statement by the American Diabetes Association. *Diabetes Care* **40**, 412–418 (2017).
6. Goh, J. K. et al. Retinal imaging techniques for diabetic retinopathy screening. *J. Diabetes Sci. Technol.* **10**, 282–294 (2016).
7. Rajesh, A. E., Davidson, O. Q., Lee, C. S. & Lee, A. Y. Artificial intelligence and diabetic retinopathy: AI framework, prospective studies, head-to-head validation, and cost-effectiveness. *Diabetes Care* **46**, 1728–1739 (2023).
8. Vujosevic, S. et al. Screening for diabetic retinopathy: new perspectives and challenges. *Lancet Diabetes Endocrinol.* **8**, 337–347 (2020).
9. Atwany, M. Z., Sahyoun, A. H. & Yaqub, M. Deep learning techniques for diabetic retinopathy classification: a survey. *IEEE Access* **10**, 28642–28655 (2022).
10. Wong, T. Y. et al. Guidelines on diabetic eye care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. *Ophthalmology* **125**, 1608–1622 (2018).
11. Jones, C. D., Greenwood, R. H., Misra, A. & Bachmann, M. O. Incidence and progression of diabetic retinopathy during 17 years of a population-based screening program in England. *Diabetes Care* **35**, 592–596 (2012).
12. Agardh, E. & Tababat-Khani, P. Adopting 3-year screening intervals for sight-threatening retinal vascular lesions in type 2 diabetic subjects without retinopathy. *Diabetes Care* **34**, 1318–1319 (2011).
13. Looker, H. C. et al. Predicted impact of extending the screening interval for diabetic retinopathy: the Scottish Diabetic Retinopathy Screening programme. *Diabetologia* **56**, 1716–1725 (2013).
14. Vujosevic, S. et al. A decade-long telemedicine screening program for diabetic retinopathy in the north-east of Italy. *J. Diabetes Complications* **31**, 1348–1353 (2017).
15. Selvachandran, G., Quek, S. G., Paramesran, R., Ding, W. & Son, L. H. Developments in the detection of diabetic retinopathy: a state-of-the-art review of computer-aided diagnosis and machine learning methods. *Artif. Intell. Rev.* **56**, 915–964 (2023).
16. Asiri, N., Hussain, M., Al Adel, F. & Alzaidi, N. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: a survey. *Artif. Intell. Med.* **99**, 101701 (2019).
17. Vives-Boix, V. & Ruiz-Fernández, D. Diabetic retinopathy detection through convolutional neural networks with synaptic metaplasticity. *Comput. Methods Programs Biomed.* **206**, 106094 (2021).
18. Liu, Y. P., Li, Z., Xu, C., Li, J. & Liang, R. Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network. *Artif. Intell. Med.* **99**, 101694 (2019).
19. Tsiknakis, N. et al. Deep learning for diabetic retinopathy detection and classification based on fundus images: a review. *Comput. Biol. Med.* **135**, 104599 (2021).
20. Bhulakshmi, D. & Rajput, D. S. A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images. *PeerJ Comput. Sci.* **10**, e1947 (2024).
21. Alwakid, G., Gouda, W. & Humayun, M. Enhancement of diabetic retinopathy prognostication using deep learning, CLAHE, and ESRGAN. *Diagnostics* **13**, 2375 (2023).
22. Li, W. et al. Quality improvement of adaptive optics retinal images using conditional adversarial networks. *Biomed. Opt. Express* **11**, 831–849 (2020).
23. Xie, Y. et al. Fundus image-label pairs synthesis and retinopathy screening via GANs with class-imbalanced semi-supervised learning. *IEEE Trans. Med. Imaging* **42**, 2714–2725 (2023).
24. Zhou, Y., Wang, B., He, X., Cui, S. & Shao, L. DR-GAN: conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *IEEE J. Biomed. Health Informatics* **26**, 56–66 (2022).
25. Zheng, R. et al. Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. *Biomed. Opt. Express* **9**, 4863–4878 (2018).

26. Huang, K. et al. Lesion-aware generative adversarial networks for color fundus image to fundus fluorescein angiography translation. *Comput. Methods Programs Biomed.* **229**, 107306 (2023).

27. Pham, Q. T. M., Ahn, S., Shin, J. & Song, S. J. Generating future fundus images for early age-related macular degeneration based on generative adversarial networks. *Comput. Methods Programs Biomed.* **216**, 106648 (2022).

28. Sahlsten, J. et al. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Sci. Rep.* **9**, 10750 (2019).

29. Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)* 2242–2251 (IEEE, 2017).

30. Huang, X., Liu, M. Y., Belongie, S. & Kautz, J. Multimodal unsupervised image-to-image translation. In *Proc. European Conference on Computer Vision (ECCV) (Munich)* 172–189 (Springer, 2018).

31. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).

32. Yi, X., Walia, E. & Babyn, P. Generative adversarial network in medical imaging: a review. *Med. Image Anal.* **58**, 101552 (2019).

33. Cohen, J. P., Luck, M. & Honari, S. Distribution matching losses can hallucinate features in medical image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 529–536 (2018).

34. Yoon, J. S., Zhang, C., Suk, H. I., Guo, J. & Li, X. SADM: sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging* 388–400 (SpringerNature, 2023).

35. Graham-Rowe, E. et al. Barriers to and enablers of diabetic retinopathy screening attendance: a systematic review of published and grey literature. *Diabet. Med.* **35**, 1308–1319 (2018).

36. Paproki, A., Salvado, O. & Fookes, C. Synthetic data for deep learning in computer vision & medical imaging: a means to reduce data bias. *ACM Comput. Surv.* https://doi.org/10.1145/3663759 (2024).

37. Li, W., Fan, L., Wang, Z., Ma, C. & Cui, X. Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recogn.* **110**, 107646 (2021).

38. Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).

39. Kim, J. & Park, H. Limited discriminator GAN using explainable AI model for overfitting problem. *ICC Express* **9**, 241–246 (2023).

40. Brown, E. E. et al. Physics-informed deep generative learning for quantitative assessment of the retina. *Nat. Commun.* **15**, 6859 (2024).

41. Mi, J., Wang, L., Liu, Y. & Zhang, J. KDE-GAN: a multimodal medical image-fusion model based on knowledge distillation and explainable AI modules. *Comput. Biol. Med.* **151**, 106273 (2022).

42. Qiao, H. DRForecastGAN (v1.0.0). Zenodo https://doi.org/10.5281/zenodo.15747872 (2025).

Author contributions

H.Q.: study design, data curation, writing and revising manuscript; F.T.: data curation, model building and evaluation, revising manuscript; H.Z.: study design and evaluating images; Y.C.: investigation and methodology; K.G.: data cleaning; J.W.: project administration; T.M.: revising manuscript and methology; L.J.: data curation; W.F.: model building; Z.M.: reviewing images; J.C.: image analysis; Y.L.: data cleaning and image analysis; B.W.: study design and investigation; Z.G.: supervision, writing—review and editing; Q.Y.: study design, investigation, supervison and reviewing manuscript.

Competing interests

The authors declare no competing interests.

Additional information