Improving detection of diabetic retinopathy in low-resolution images via latent diffusion

Jasper Kyle Catapang[1], Darby E. Santiago[2], Iris Thiele Isip-Tan[2]
[1]Department of English Language and Linguistics, University of Birmingham, United Kingdom
[2]College of Medicine, University of the Philippines Manila, Philippines
Email: [1]jxc1354@alumni.bham.ac.uk,
[2]desantiago@up.edu.ph,[3]icisiptan@up.edu.ph

Abstract—**Diabetic retinopathy (DR) is the leading cause of blindness in working-age adults worldwide. Early detection and intervention are crucial in preventing blindness, but there is no national program for DR screening in the Philippines, and the current DR screening pathway is neither practical nor economical. The paper proposes a novel approach called Affinitive Diffusion-Augmented Vision for Interpretable Transformers (AffiDAVIT), which utilizes 5-fold cross-validation, diffusion-based data augmentation, and vision transformers to improve the detection of DR in low-resolution images. The approach leverages explainable ViT using Grad-CAM to highlight the regions of interest in fundus images for DR detection. Fuzzy logic is then applied whenever a no DR (negative) result is predicted by the model to lessen false negatives. AffiDAVIT outperforms state-of-the-art models on the same dataset. Overall, the paper presents a potential solution for improving the detection of DR in low-resolution images, and the findings contribute significantly to the field of DR detection. The approach presents a promising alternative for DR screening in areas with limited access to ophthalmologists and comprehensive diabetes care.**

Index Terms—**diabetic retinopathy, low-resolution images, vision transformer, diffusion model**

I. Introduction

Diabetic retinopathy (DR) is the leading cause of blindness in working-age adults worldwide [1], with early detection and intervention being crucial in preventing blindness. DR is a complication that results from poor glycemic control. Philippine data from the 2008 DiabCare-Asia project showed that only 15% achieved the American Diabetes Association HbA1c target of less than 7% with 20.1% exhibiting non-proliferative DR, 8.3% proliferative DR, and 7.4% requiring photocoagulation [2]. DR comprised 85% of patients in the retina clinic and 80% of retinal surgeries in a national tertiary hospital serving the underprivileged who have less access to comprehensive diabetes care [3].

There is no national program for diabetic retinopathy screening in the Philippines [4]. The traditional and gold standard method of assessing the presence or absence of DR requires funduscopic examination using an ophthalmoscope by a health professional. Though all physicians learned fundus examination in medical school, it is often relegated to ophthalmologists which has limited the assessment of DR, given the maldistribution of ophthalmologists in the Philippines. To address this, there is an imperative to establish telehealth programs using retinal imaging [5] especially as the prevalence of diabetes mellitus is expected to rise in Southeast Asia. A telemedicine screening program in community health centers in an urban city showed the feasibility of DR screening using fundus photographs sent to retina specialists, though the authors note that dilation was needed for 22% due to poor quality fundus photos [4]. This approach however still required retina specialists who are in short supply. Coupled with the increase in the prevalence of diabetes, this makes the current DR screening pathway neither practical nor economical, hence the interest in the use of portable smartphone-based fundus cameras [6] possibly integrated with artificial intelligence algorithms to detect DR [7].

Several issues and challenges exist in the use of artificial intelligence for DR screening. These include inconsistent annotations, less number of sample images, inappropriate performance evaluation metrics, and imbalanced data sets [8], [9]. Data augmentation addresses data imbalance by artificially increasing the size of the minority class in a dataset by creating new synthetic examples through various techniques. Traditional approaches to artificially increase image data include, but are not limited to, rotation, translation, flipping, but these may not be effective in the medical domain due to the limited number of medical images. Recent studies have explored the use of diffusion-based data augmentation on medical images, achieving promising results in skin disease classification [10] and medical image analysis [11]. However, no studies have explored the use of diffusion-based data augmentation in DR detection.

Current literature on the use of vision transformers mostly focuses on skin diseases [10], COVID-19 detection [12], kidney cysts, stones, and tumors [13], and pneumonia identification [14]. However, there are a few studies that have also explored the use of vision transformers (ViT) for DR grading using fundus images, achieving promising results [15], [16]. Furthermore, explainable artificial intelligence (XAI) has become increasingly important in the medical domain, with studies showing the effectiveness of explainable ViT in COVID-19 detection [12], pneumonia identification [14], and DR grading [15]. One effective approach for explainability is using Grad-CAM [17], which has been

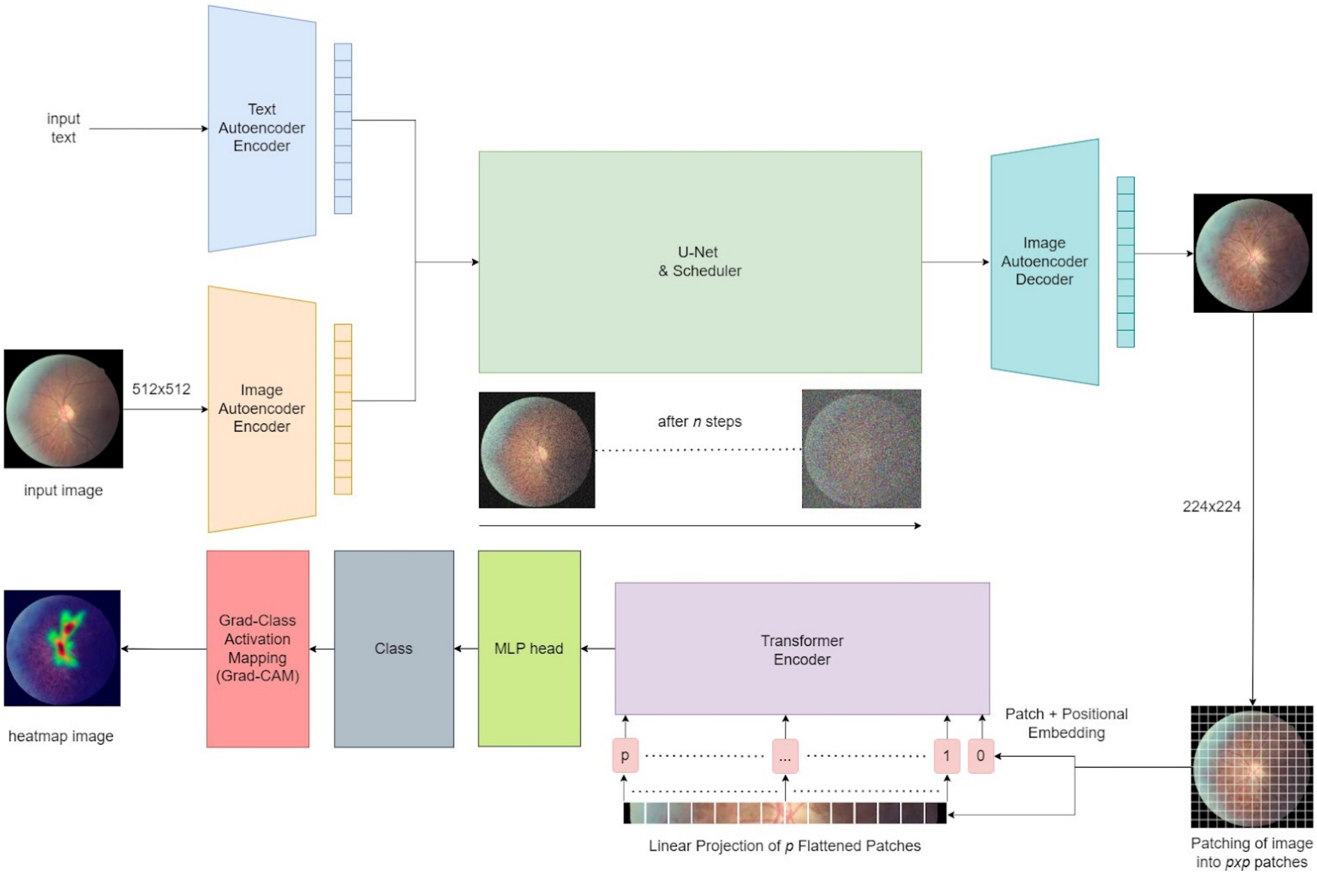

Fig. 1: The proposed architecture of Affinitive Diffusion-Augmented Vision for Interpretable Transformers (AffiDAVIT)

successfully applied in DR detection using EfficientNET [18].

This research paper addresses two main questions: can the performance of deep learning models in DR detection be improved using diffusion-based data augmentation, and are synthetically generated fundus images indistinguishable from authentic ones? To answer these questions, we propose a novel approach called Affinitive Diffusion-Augmented Vision for Interpretable Transformers (AffiDAVIT), which utilizes diffusion-based data augmentation and vision transformers to improve the detection of DR in low-resolution images. Our approach also leverages explainable ViT using Grad-CAM [17] to highlight the regions of interest in fundus images for DR detection. We test the impact of diffusion-based data augmentation via Stable Diffusion [19] on the generalization of our model, with the goal of minimizing the rate of false negatives. Reducing false negatives in the context of diabetic retinopathy detection ensures that we minimize the likelihood that we mislabel an eye image as DR-free. We consider the probabilities produced by ViT as our membership value. Overall, our approach presents a potential solution for improving the detection of DR in low-resolution images, and our findings contribute significantly to the field of DR detection.

II. Data

We use the training images in the dataset provided through a Kaggle competition entitled “Diabetic Retinopathy Detection” [1]. The California Healthcare Foundation hosted the competition with the images provided by EyePACS. There are 35,126 training images from the dataset. Of these images, only 20% were initially considered for this study [2], to simulate data scarcity, since this study effectively serves as a first experiment to a series of AI-related research on diabetic retinopathy in developing communities. And from this subset, 30% of the images were treated as the validation data and 70% were considered the training data, as suggested by the data split survey in DR detection by Rahman et al. [9]. The resulting dataset can be categorized as follows: 4,604 training images with no diabetic retinopathy, 1,305 training images with diabetic retinopathy, 558 validation images with no diabetic retinopathy, and 558 validation images with diabetic retinopathy. Two additional sets of diabetic retinopathic and non-diabetic retinopathic images (560 images each, respectively) are considered for further evaluation, acting as the test set.

[1] https://www.kaggle.com/competitions/diabetic-retinopathy-detection

[2] This selected subset is expanded to test the validity of the methodology in larger subsets in a smaller set of experiments, later on.

III. Methodology

Our proposed architecture consists of three main modules, latent diffusion augmentation (DA), the ViT module, and the XAI module, as seen in Figure 1. In Figure 1, the input is the input text and the 512x512 input image—passed through the text and image encoders. These are processed through Stable Diffusion (U-Net & Scheduler) and a 224x224 variation of the image is produced by the decoder. This variation is then passed to the ViT model [21] where the image is split into multiple patches and used to train the model. After classifying the images, the classes and images are processed by Grad-CAM [17], wherein the heatmap images are generated. This procedure simulates an explainable classifier for low-resolution images of diabetic retinopathy. For our experiment, we employ two setups. The first setup does not use data augmentation while the second setup employs the DA module. This produces two distinct sets of results to quantify the performance of the ViT module with and without augmentation. In this series of experiments, we use a local machine with an Intel i9 CPU, RTX 3070 GPU with 8 GB VRAM, and 32 GB RAM. Lastly, we use Python 3.10.6 for programming and PyTorch 1.12.1 with CUDA GPU acceleration. The code, data, and best model can be found in our GitHub repository [3]. However, for the additional experiment, we use Google Colab's GPU T4 runtime type for more processing power.

For the DA module, we make use of Stable Diffusion 1.5. It is open-source via GitHub[4]. We utilize DDIM as the diffusion sampling method, as described in the vanilla Stable Diffusion paper [19]. Qualitative analysis is performed to determine the best diffusion model hyperparameters. We adjust the values of sampling steps, classifier free guidance (CFG) scale, and denoising strength to achieve the most convincing synthetic images. The range of values considered for the denoising strength should be near zero—making the diffusion "affinitive" or closely related to the base image. The images considered in the DA module are all resized to 512px x 512px. For the ViT module, we use Google's ViT model [21], pre-trained on ImageNet-21k at 224px x 224px resolution, and fine-tuned on ImageNet 2012 at 224px x 224px resolution, similar to Mohan et al.'s (2022) [15] architecture. We chose to replicate this paper's [15] choice of the model since it is state-of-the-art, at the time of initial experimentation. This ViT model segments the input image into 16px x 16px resolution patches. We finetune [21]'s model via Huggingface on the images described in an earlier section. The images considered in the ViT module are all resized to 224px x 224px. For the XAI module, we use pytorch-grad-cam[5], a PyTorch implementation of [17]'s paper, similar to the framework proposed by [18]. The finetuned ViT module serves as the classifier utilized in pytorch-grad-cam.

We evaluate our proposal in two ways. First, we assess the (in)distinguishability of the synthetically-generated images by having an ophthalmologist/retina specialist distinguish authentic eye images from synthetic ones. We obtain the confusion matrix and measure the accuracy, precision, recall, and f1 scores for this comparison. Another evaluation focuses on the finetuning of the ViT model, measuring validation loss and accuracy and testing accuracy. To improve generalizability, we also use 5-fold cross-validation.

IV. Results and Discussion

In Stable Diffusion, we make use of the base text prompt "eye fundus image" and the negative text prompt "eyelashes, iris" since the diffusion model tends to hallucinate eyelashes and irises within the eye images themselves. We also use the built-in BLIP [20] function of the Stable Diffusion software to reverse-engineer the possible text prompt for the eye images. The values tested for the various hyperparameters are listed in Table 1. After a series of quality testing, we found that the output images produced with hyperparameter values of 20 for sampling steps, 7.0 for the CFG scale, and 0.2 for denoising strength, proved to look the most convincing. We generate using the same values 2,998 synthetic images of eyes with diabetic retinopathy, the underrepresented class in the dataset.

TABLE I: Stable Diffusion hyperparameter optimization through qualitative analysis

Hyperparameter	Values
Sampling steps	20, 25, 30
Classifier free guidance scale	7.0, 7.5, 8.0, 8.5, 9
Denoising strength	0.1, 0.2, 0.3

We asked a retina specialist with 15 years of professional experience to sort a mixed set of images containing 354 real/synthetic samples. This sample size yields results with a 95% confidence interval and a 5% error. The results of the manual labeling compared to the ground truth are shown in Figure 2.

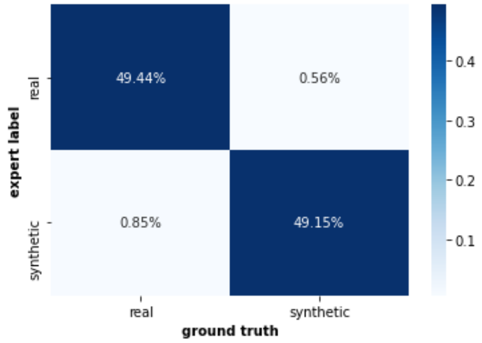

Fig. 2: Manual labeling of retina specialist versus ground truth

Figure 2 suggests that the synthetic images produced by Stable Diffusion are easily distinguishable from authentic fundus images when examined by a retina specialist. The distinctions between real and synthetic images are as follows: (1) loss of the normal dichotomous branching of the retinal

[3] https://github.com/jcatapang/AffiDAVIT

[4] https://github.com/CompVis/stable-diffusion

[5] https://github.com/jacobgil/pytorch-grad-cam

TABLE II: The validation set performance of finetuning ViT base 224px, patch 16 on training data

Finetune data	Steps	Learning rate	Validation loss	Validation accuracy
Base	700	1.00E-05	0.7167	0.6548
Base	400	3.00E-05	0.6835	0.6594
Base + diffusion augmentation (AffiDAVIT)	700	1.00E-05	0.7096	0.6521
AffiDAVIT	700	3.00E-05	0.6871	0.663

TABLE III: The test set performance of finetuning ViT base 224px, patch 16 on training data

Finetune data	Learning rate	Test loss	Test accuracy	True no DR	False with DR	False no DR	True with DR
Base	1.00E-05	7.35E-01	0.6554	529	31	355	205
Base	3.00E-05	6.96E-01	0.6509	530	30	361	199
AffiDAVIT	1.00E-05	6.85E-01	0.6545	521	39	348	212
AffiDAVIT	3.00E-05	6.21E-01	0.7	515	45	291	269

TABLE IV: Comparing the validation set performance of AffiDAVIT against annual state-of-the-art models

Model name	Validation loss	Validation accuracy
AffiDAVIT	**0.6871**	**0.663**
ViT-DR (2022) [15]	0.7178	0.6478
ViT-DR (2022) [15] + diffusion augmentation	0.7220	0.6548
Wu et al. (2021) [16]	0.7192	0.6424
Wu et al. (2021) [16] + diffusion augmentation	0.7238	0.6452
EfficientNET (2020) [18]	0.7253	0.6272
EfficientNET (2020) [18] + diffusion augmentation	0.7312	0.6389

TABLE V: Comparing the test set performance of AffiDAVIT against annual state-of-the-art models

Model name	Test loss	Test accuracy	True no DR	False with DR	False no DR	True with DR
AffiDAVIT	**0.621**	**0.7**	**515**	**45**	**291**	**269**
ViT-DR (2022) [15]	0.678	0.67	480	80	246	334
ViT-DR (2022) [15] + diffusion augmentation	0.682	0.67	495	85	259	307
Wu et al. (2021) [16]	0.684	0.65	470	110	238	328
Wu et al. (2021) [16] + diffusion augmentation	0.686	0.65	485	115	251	315
EfficientNET (2020) [18]	0.698	0.62	445	155	223	343
EfficientNET (2020) [18] + diffusion augmentation	0.702	0.64	475	125	264	302

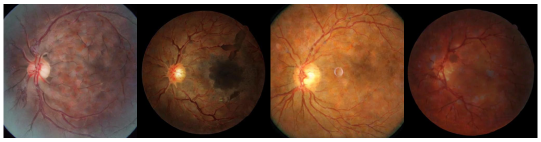

Fig. 3: Sample synthetic fundus images produced by Stable Diffusion

vessels, (2) distortion of the choroidal background, (3) loss of the normal color and vascular architecture of the foveal avascular zone, (4) abnormal appearance of the optic nerve head and the vessels arising from it, (5) “disappearance” of some blood vessels, (6) abnormal distribution of “retinal colors,” (7) repeating or similar pockets of “white” spots, and (8) presence of large dark “laser marks” in the posterior pole and foveal center. Several examples of synthetic images produced by Stable Diffusion are illustrated in Figure 3.

We finetuned Google’s ViT model [21] with 32 batch sizes for training and evaluation, 1000 max steps, 0.01 weight decay, on two learning rates: 1e-5 and 3e-5, using a cosine learning rate scheduler with restarts. Table 2 shows the best result for each variant of the model on the validation set and Table 3 shows the performance evaluation of the best checkpoint for each variant on the test set.

According to the results shown in Table 2, the use of diffusion augmentation slightly decreased the validation loss and accuracy for the base ViT model with a learning rate of 1e-05, while it slightly increased the validation loss but improved the accuracy for the model with a learning rate of 3e-05. These results suggest that diffusion augmentation did not have a significant impact on the performance of the model during validation. However, the results shown in Table 3 indicate that diffusion augmentation significantly improved the performance of the model on the test set, particularly for the model with a learning rate of 3e-05. In this case, the diffusion-augmented model achieved a lower test loss and a higher test accuracy compared to the base model, suggesting that the additional data generated by diffusion augmentation helped the model to generalize better to new data. Note that this version of the model also predicted the least number of false negatives and the most number of true positives.

Table 4 presents a comparison of the validation set performance of AffiDAVIT against several state-of-the-art models in the field. The models have been evaluated based on their validation loss and accuracy, with a lower validation loss and a higher validation accuracy being indicative of better model performance. Based on the results presented in the table, it can be observed that AffiDAVIT outperforms all other models in terms of validation loss and accuracy. AffiDAVIT achieved a validation loss of 0.6871 and a validation accuracy of 0.663, which are the best values among all the models. This indicates that AffiDAVIT has a better ability to generalize to new data, making it more robust and reliable. The other models included in the comparison are ViT-DR (2022) with and without diffusion augmentation, Wu et al. (2021), and EfficientNET (2020) with and without diffusion augmentation. While these models are all state-of-the-art, they have achieved lower validation accuracy and higher validation loss than AffiDAVIT, indicating that they may not perform as well when tested on new data.

Table 5 shows the comparison of the test set performance of AffiDAVIT against the other state-of-the-art models, evaluated based on test loss, test accuracy, and the number of true/false predictions for diabetic retinopathy (DR) cases. The results indicate that AffiDAVIT outperforms all other models in terms of test accuracy, achieving a score of 0.7. The number of true positive (TP) and true negative (TN) predictions for DR cases are also important to consider, as they reflect the model's ability to correctly identify individuals with and without the disease. Among the other models, ViT-DR (2022) with and without diffusion augmentation achieved the highest test accuracy after AffiDAVIT, followed by Wu et al. (2021) and EfficientNET (2020) models. However, ViT-DR (2022) without augmentation had the lowest TN and highest FP predictions among all models, indicating that it may have difficulty in correctly identifying individuals without DR. In contrast, Wu et al. (2021) + diffusion augmentation had the highest TN and lowest FP predictions, indicating that it performed well in correctly identifying individuals without DR. However, this model had a lower overall test accuracy compared to AffiDAVIT and ViT-DR (2022). Overall, the results suggest that AffiDAVIT is the best model among all the models compared in terms of test accuracy, and it also achieves a good balance between TP and TN predictions.

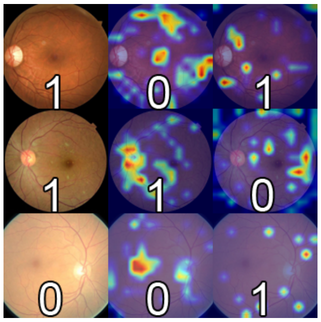

Fig. 4: Grad-CAM on test data using the best finetuned ViT model

Figure 4 illustrates several heatmaps illustrating the regions of interest where the best finetuned ViT model focused, using its attention mechanism. The first column is the ground truth, the second column is the most confident prediction made by the ViT model, and the third column is the other class. In this diagram, 1 indicates diabetic retinopathy and 0 indicates its absence. The first row is the false negative, the second row is the true positive, and the third row is the true negative. For the false negative example, the no DR probability is 80.21%, and the DR probability is 19.79%. These probabilities act as membership values.

TABLE VI: Percent subset simulated accuracy

Percentage	Test accuracy
20%	0.7
40%	0.74
60%	0.77
80%	0.82
100%	0.86

To further validate the effectiveness of diffusion as an augmentation technique, we also tried expanding the initial subset considered earlier[6], from 20% and incremented in steps of 20. As expected, the accuracy is shown to increase as we increase the subset coverage. The effect on accuracy is shown in Table 6.

While it demonstrates promising results for this particular application, the transferability and adaptability of this method to other medical domains or early detection of other medical issues would require careful consideration and evaluation. The success of applying Stable Diffusion or similar generative models to other medical domains would depend on the nature of the medical issue being addressed. Some medical conditions might exhibit visual characteristics that can be effectively modeled and generated using image-based techniques. However, other conditions might involve more complex interactions and require additional types of data, such as time-series data, textual data, or other modalities, to accurately capture and represent the condition. This methodology can be extended to diverse industries such as agriculture, logistics, and environmental science. In agriculture, diffusion augmentation can assist with crop yield prediction and disease detection. Logistics can utilize diffusion augmentation to enhance solutions for visual traffic and routing problems. Environmental science benefits from diffusion augmentation for climate modeling and pollution analysis. However, the effectiveness of pipelines implementing the use of generative models, in each industry, depends on domain-specific data, appropriate feature representation, and careful evaluation. Ethical considerations and legal implications must also be addressed when deploying generative models in real-world scenarios. Lastly, ethical considerations and legal implications specific to each industry must be taken into account when deploying such methodologies, as AI-generated data is considered ethically

[6] Please contact the corresponding author for access to the resources of this additional series of experiments.

questionable in several disciplines, such as the arts.

V. Conclusion

In this study, we evaluated the performance of the ViT model with and without diffusion augmentation and the (in)distinguishability of synthetically-generated eye images from authentic eye images. Our findings indicate that diffusion augmentation had a minor impact on the performance of the model during validation, but significantly improved its performance on the test set. In particular, a lower test loss and a higher test accuracy compared to the base model were demonstrated on the test set. This occurred due to the models having different environments, in terms of evaluation during validation. We further gauged AffiDAVIT's performance by comparing the losses and accuracies against state-of-the-art models—of which AffiDAVIT's is the best performing for both validation and test sets. The results suggest that the additional data generated via diffusion augmentation helped the model to generalize better to new data, despite the photo quality. Our results also showed that the low-resolution synthetic images produced by Stable Diffusion are easily distinguishable from real low-resolution fundus images when examined by a retina specialist. Overall, our study provides evidence that synthetic data generated by Stable Diffusion can be a useful tool for training deep learning models in medical image analysis, and that diffusion augmentation can enhance the performance of the ViT model on new data. The low-resolution synthetic images may be unconvincing to a retina specialist, but it is sufficient to improve the computational performance of deep learning models. These findings have important implications for improving the accuracy and efficiency of medical image analysis, which can ultimately lead to better patient outcomes.

VI. Recommendations

Our suggestions include investigating the finetuning of various ViT models concerning the number of patches and pixels. Additionally, it is advisable to evaluate the models' performance on genuine low-resolution images, given that the low-resolution images used in this study were artificially created. Furthermore, a quantitative approach to hyperparameter optimization of the diffusion model would reduce toil.

Acknowledgements

We would like to thank Kaggle, the California Healthcare Foundation, and EyePACS for the dataset. We are also grateful for the inspiration and support provided by several Y Combinator founders (who wish to not be named) that made the authors consider the use of diffusion models in this study. Last but not the least, we thank Moses Visperas for his pointers for minor revisions.

References

[1] Bourne, R. R., Jonas, J. B., Bron, A. M., Cicinelli, M. V., Das, A., Flaxman, S. R., ... & Resnikoff, S. Prevalence and causes of vision loss in high-income countries and in Eastern and Central Europe in 2015: magnitude, temporal trends and projections. British Journal of Ophthalmology, 102(5), 575-585. 2018.

[2] Jimeno, C., Sobrepeña, L., Mirasol, R. DiabCare 2008: Survey on Glycaemic Control and the Status of Diabetes Care and Complications Among Patients with Type 2 Diabetes Mellitus in the Philippines. Phillippine Journal of Internal Medicine. 50(1):15-22. 2012.

[3] Fajardo-Gomez, M.F., Uy, H.S. Prevalence of diabetic retinopathy among diabetic patients in a tertiary hospital. Philippine Journal of Ophthalmology 30(4): 178-180. 2005.

[4] Daza, J., Sy, J., Rondaris, M. V., & Uy, J. P. Telemedicine Screening of the Prevalence of Diabetic Retinopathy Among Type 2 Diabetic Filipinos in the Community. Journal of Medicine, University of Santo Tomas, 6(2), 999–1008. 2022.

[5] Silva, P., Cavallerano, J., Paz-Pacheco, E., & Aiello, L. P. Diabetic Retinopathy in Southeast Asia:A Call for Ocular Telehealth Programs. Journal of the ASEAN Federation of Endocrine Societies, 27(2), 176–179. https://doi.org/10.15605/jafes.027.02.07. 2012.

[6] Rajalakshmi, R., Prathiba, V., Arulmalar, S., & Usha, M. Review of retinal cameras for global coverage of diabetic retinopathy screening. Eye, 35(1), 162-172. 2021.

[7] Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C., & Rajalakshmi, R. Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. Eye (London, England), 33(1), 97–109. https://doi.org/10.1038/s41433-018-0269-y. 2019.

[8] Saini, M., & Susan, S. Diabetic retinopathy screening using deep learning for multi-class imbalanced datasets. Computers in Biology and Medicine, 149, 105989. 2022.

[9] Mujeeb Rahman, K. K., Mohamed Nasor, and Ahmed Imran. "Automatic Screening of Diabetic Retinopathy Using Fundus Images and Machine Learning Algorithms." Diagnostics 12.9: 2262. 2022.

[10] Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincső, B., Solis, S., ... & Fazekas, I. Diffusion-based Data Augmentation for Skin Disease Classification: Impact Across Original Medical Datasets to Fully Synthetic Images. arXiv preprint arXiv:2301.04802. 2023.

[11] Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., & Merhof, D. Diffusion models for medical image analysis: A comprehensive survey. arXiv preprint arXiv:2211.07804.

[12] Chetoui, M., & Akhloufi, M. A. (2022). Explainable vision transformers and radiomics for covid-19 detection in chest x-rays. Journal of Clinical Medicine, 11(11), 3013. 2022.

[13] Islam, M. N., Hasan, M., Hossain, M. K., Alam, M. G. R., Uddin, M. Z., & Soylu, A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Scientific Reports, 12(1), 11440. 2022.

[14] Ukwuoma, C. C., Qin, Z., Heyat, M. B. B., Akhtar, F., Bamisile, O., Muaad, A. Y., ... & Al-Antari, M. A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. Journal of Advanced Research. 2022.

[15] Mohan, N. J., Murugan, R., Goel, T., & Roy, P. ViT-DR: Vision Transformers in Diabetic Retinopathy Grading Using Fundus Images. In 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 167-172). IEEE. 2022.

[16] Wu, J., Hu, R., Xiao, Z., Chen, J., & Liu, J. Vision Transformer-based recognition of diabetic retinopathy grade. Medical Physics, 48(12), 7850-7863. 2021.

[17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626). 2017.

[18] Chetoui, M., & Akhloufi, M. A. Explainable diabetic retinopathy using EfficientNET. In 2020 42nd annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC) (pp. 1966-1969). IEEE. 2020.

[19] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695). 2022.

[20] Li, J., Li, D., Xiong, C., & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900). PMLR. 2022.

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.