International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-024-02485-w

ORIGINAL ARTICLE

Diffusion-based data augmentation methodology for improved performance in ocular disease diagnosis using retinography images

Burak Aktas[1] · **Doga Deniz Ates**[1] · **Okan Duzyel**[1] · **Abdurrahman Gumus**[1]

Received: 2 February 2024 / Accepted: 27 November 2024

Abstract

Deep learning models, integral components of contemporary technological landscapes, exhibit enhanced learning capabilities with larger datasets. Traditional data augmentation techniques, while effective in generating new data, have limitations, especially in fields like ocular disease diagnosis. In response, alternative augmentation approaches, including the utilization of generative AI, have emerged. In our study, we employed a diffusion-based model (Stable Diffusion) to synthesize data by faithfully recreating crucial vascular structures in the retina, vital for detecting eye diseases by using the Ocular Disease Intelligent Recognition dataset. Our goal was to augment retinography images for ocular disease diagnosis using diffusion-based models, optimizing the outputs of the fine-tuned Stable Diffusion model, and ensuring the generated data closely resembles real-world scenarios. This strategic approach resulted in improved performance in classification models and augmentation outperformed traditional methods, exhibiting high precision rates ranging from 85% to 76.2% and recall values of 86%, and 75% for 5 classes. Beyond performance enhancement, we demonstrated that the inclusion of synthetic data, coupled with data reduction using the t-SNE method, effectively addressed dataset imbalance. As a result of synthetic data addition, notable increases of 3.4% in the precision metric and 12.8% in the recall metric were observed in the 7-class case. Strategically synthesizing data addressed underrepresented classes, creating a balanced dataset for comprehensive model learning. Surpassing performance improvements, this approach underscores synthetic data's ability to overcome the limitations of traditional methods, particularly in sensitive medical domains like ocular disease diagnosis, ensuring accurate classification. The codes of the study will be shared on GitHub in a way that benefits everyone interested: https://github.com/miralab-ai/generative-data-augmentation.

Keywords Image classification · Data augmentation · Diffusion-based models · t-SNE · Medical image synthesis · Dataset imbalance

1 Introduction

Diffusion models [1–9], a subset of deep generative models, have surged in computer vision, displaying remarkable abilities in generating diverse and detailed high-quality images and visual content, especially exemplified by models like Latent Diffusion Models (LDMs) [1, 2], setting a new standard in generative modeling. Diffusion models are utilized in a range of generative tasks like image generation [1, 4, 10–14], inpainting [15–17], and image-to-image translation [18–21], demonstrating their adaptability across fields by fulfilling diverse data needs through synthetic data generation capabilities. Recently, they have started to be used to alleviate the data scarcity issues in the medical domain [22–27]. Their ability to produce realistic and diverse synthetic data has opened up new avenues for enhancing the

Burak Aktas and Doga Deniz Ates contributed equally to this work.

✉ Abdurrahman Gumus
abdurrahmangumus@iyte.edu.tr

Burak Aktas
imburakaktas@gmail.com

Doga Deniz Ates
dogadates@gmail.com

Okan Duzyel
okanduzyel@iyte.edu.tr

[1] Department of Electrical and Electronics Engineering, İzmir Institute of Technology, Izmir, Turkey

Published online: 11 December 2024

performance of medical image classification tasks, particularly in areas such as dermatology, pathology, and radiology.

Recent studies have demonstrated the effectiveness of LDMs in augmenting medical image datasets and mitigating bias in medical image classifiers. Akrout et al. [22] propose a diffusion-based data augmentation approach for skin disease classification in the Fitzpatrick 17K dataset. They use a diffusion model trained on real skin lesions to generate fully-synthetic skin lesions, achieving comparable classification accuracy even with a fully synthetic skin disease dataset. This suggests the potential value of diffusion-based data augmentation in enhancing medical image classifier performance. Ktena et al. [23] work on utilizing synthetic images from diffusion models in medical image classification highlights the potential of synthetic data in enhancing fairness under distribution shifts. Parallel to their findings, their observations demonstrate out-of-distribution improvements across different modalities: a 7.7% increase in histopathology, 5.2% in chest radiology, and a remarkable 63.5% enhancement in high-risk sensitivity for dermatology. These enhancements suggest the efficacy of synthetic data across various medical imaging domains. Sagers et al. [24] explore the use of latent diffusion models to extend medical image classifiers in datasets with synthetic data, including 3699 images from the Fitzpatrick 17K dataset and 656 images from the Stanford Diverse Dermatology Images (DDI) dataset. Their latent diffusion model generates synthetic images resembling real ones in the dataset, leading to enhanced performance in medical image classifiers. This suggests that latent diffusion models can effectively contribute to augmenting medical image classifiers with synthetic data. In a study by Sagers et al. [25], seven skin conditions were selected from the Fitzpatrick 17K dataset. Using OpenAI's DALL·E 2 model, synthetic variations were generated from randomly selected image samples of the lightest and darkest Fitzpatrick skin types for each condition. The experiments, involving training on lighter skin types and testing on darker skin types, and vice versa, evaluated the impact of synthetic images on classification performance using Fitzpatrick 17K images, baseline images, and DALL·E 2 produced synthetic images. These studies collectively demonstrate the promising role of LDMs in addressing data scarcity and bias challenges in medical image classification. By leveraging their ability to generate realistic and diverse synthetic data, LDMs can contribute to the development of more robust and equitable medical image classifiers, ultimately improving patient care and outcomes.

Fang et al. [28] explore the use of controllable diffusion models for data augmentation in object detection tasks. The authors utilize the Coco and Pascal VOC dataset to demonstrate how these models can generate synthetic images to enhance the performance of object detection models. The results indicate that this method significantly improves the accuracy and overall performance of detection models compared to traditional augmentation techniques. Feng et al. [29] introduces a method called DiffTPT, which leverages diffusion models to generate diverse synthetic data for enhancing test-time prompt tuning. By integrating both conventional data augmentation techniques and diffusion-based synthesis, this approach improves model adaptability to unseen test data. While our work focuses on improving ocular disease diagnosis through the augmentation of retinography images, their approach targets test-time prompt adaptation in vision-language models across various domains. Moreover, Fu et al. [30] explores how diffusion models can generate synthetic data to enhance model training, particularly in limited dataset scenarios. By creating high-quality, realistic data samples, DreamDA improves the performance of deep learning models across various tasks. In contrast, this research focuses on preserving critical features such as vascular structures in the retina, necessary for accurate medical diagnosis using the Ocular Disease Intelligent Recognition (ODIR) dataset. On a broder scale, Bennett et al. [31] explores the application of Stable Diffusion for synthetic data generation to create diverse image datasets. Using the Stable Diffusion 1.4 model, the authors generate synthetic images from text prompts, leveraging Wordnet's synsets to produce varied object representations. The dataset consists of images covering a wide range of concepts, with evaluations conducted using the ImageNet dataset to verify classification accuracy. Their findings reveal that while Stable Diffusion can generate realistic images, the quality and classification performance varies across different categories, with some images being less recognizable by models.

Generative Adversarial Networks (GANs) are among the various approaches used for generating synthetic data in medical image analysis. For example, Kebaili et al. [32] examines advanced generative models, such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models, to address challenges like limited data availability and the need for high-quality datasets in medical image analysis. While GANs excel at generating diverse image variations, the authors highlight their struggle with preserving subtle yet crucial details, underscoring the suitability of diffusion models for more sensitive tasks. Additionally, the challenge of the dataset imbalance, a key issue in medical imaging, is addressed through a combination of synthetic data and the t-SNE method, effectively enhancing the performance of the classification models. Smitha et al. [33] presented a semi-supervised GAN model for fundus image analysis, which includes a preprocessing module to enhance image quality. This module effectively preserves image details while improving input quality. Their model achieved a commendable accuracy of 90% when classifying Normal and AMD images, although the lowest accuracy was 72% for other classifications, indicating

that while the approach is effective, there is still room for further improvement in classification accuracy. Similarly, Gobinath et al. [34] proposed a semi-supervised Generative Adversarial Network (SGAN) for fundus image classification, focusing on the detection of retinal diseases such as diabetic retinopathy and glaucoma. Their model benefits from the use of both labeled and unlabeled datasets, achieving an accuracy of 0.88 and an F1-Score of 0.85. Despite these advancements, the performance of the model heavily relies on the quality and quantity of the training data, especially for pixel-wise annotated data, which remains a challenge for many deep learning methods. However, in our work, we investigated diffusion-based data augmentation methodology. Stable Diffusion presents a whole new ballgame compared to GAN-based approaches, offering distinct advantages in synthesizing high-quality images for augmentation while addressing some of the challenges faced by previous GAN models. This novel method holds the potential to enhance image diversity and boost model performance in classification tasks.

Our research demonstrates that the efficacy of deep learning-based classification models can be enhanced through the generation of realistic synthetic data, accomplished using diffusion-based models. The main goal of your study is to improve ocular disease diagnosis by addressing limitations in traditional data augmentation methods. Specifically, we aim to create high-quality synthetic data using a diffusion-based model (Stable Diffusion) to closely mimic real-world scenarios while preserving critical details like retinal vascular structures. This helps enhance classification model performance, particularly by mitigating class imbalance and improving representation for underrepresented classes. Our study utilizes the Ocular Disease Intelligent Recognition (ODIR) dataset [35], which encompasses a diverse range of eight ocular conditions. The research improves deep learning-based classification models using synthetic data generated by fine-tuning the diffusion-based model named Stable Diffusion [36]. By combining real and synthetic data, we enhanced model performance for a 5-class classification, addressing imbalance by creating 6015 synthetic samples for underrepresented classes. The VGG19 model, one of the Convolutional Neural Network (CNN) methods [37] was used to filter and create a clean synthetic dataset. Furthermore, we addressed the imbalance prevalent in the dominant classes (Normal and Diabetes) by employing the t-SNE method [38, 39]. Our focus revolved around assessing the performance enhancement resulting from the incorporation of synthetic data into classification models.

The novelties of the current study can be summarized as follows:

- Expanding the scope of diffusion-based data augmentation: This research challenges the conventional boundaries of diffusion-based data augmentation by utilizing a previously unexplored dataset within the field. This approach demonstrates the adaptability and expands the applicability of diffusion-based data augmentation techniques. For complex datasets like ours, where traditional augmentation methods often may not be suitable, it becomes essential to generate synthetic data that is both high-quality and diverse to provide the demanded generalization and performance. This approach ensures more robust and effective data augmentation, overcoming the limitations of conventional methods. This approach ensures more robust and effective data augmentation, overcoming the limitations of conventional methods.

- Uncovering the significance of generating prompts: Through our investigation, we have uncovered a series of critical factors that significantly influence the success of synthetic data generation using diffusion-based models. These insights hold substantial implications for future research endeavors, notably shedding light on the profound impact of prompts in shaping the quality and realism of synthetic data. It was discovered that when we trained Stable Diffusion with the specific prompt that identifies diseases with real data, the results we obtained were much better than those trained with general prompts.

- Examination of the effect of the model on the quantity of synthetic data: In our quest to fulfill our research goals, we conducted an exhaustive analysis of the connection between the performance fluctuations observed in classification models and the quantity of synthetic data utilized in our study. As a result of these analyses, we discovered that the idea of *"the more synthetic data we add, the better"* is a misconception. We discovered that the added data has a saturation point, and an ablation study was conducted to prove it.

- Comparative analysis between classic and diffusion-based data augmentation: We undertook a detailed comparison to assess the impact of data generated through classical data augmentation techniques versus that generated by the diffusion model on overall model performance. For complex datasets like ours, where traditional augmentation techniques may fall short, generating high-quality and diverse synthetic data is essential for achieving the desired generalization and performance. In this specific dataset, only a few methods can be applied, such as 180-degree rotation, are feasible, as other techniques can distort the delicate features of the eye images and hinder the extraction of key information. Therefore, using synthetic data offers a more reliable and efficient approach to data augmentation, overcoming the limitations of traditional methods and enhancing overall model effectiveness.

- Resolving the dataset imbalance by jointly using synthetic data generation and the t-SNE-based image selection: We worked on addressing the dataset's imbalance issue by applying the t-SNE method to the dominant classes and adding synthetic data to the remaining 5 classes.

- Implementation of advanced metrics: Usage of advanced evaluation metrics beyond accuracy (e.g., F1-Score, precision, recall) provides a comprehensive assessment of the classification model's performance, especially in the context of class imbalance.

The rest of the paper is organized as follows: Section 2 provides the methodologies for the proposed approach, referring to the algorithm flow and outlining the key methods used. Section 3 presents the results, focusing on how the synthetic data influenced classification and offering discussions about the findings. Lastly, Section 4 summarizes the study's contributions and highlights future research opportunities.

2 Methods

In this section, we provide a detailed overview of the methodologies employed in our study, focusing on the algorithmic flow, dataset selection, and augmentation techniques. We begin by outlining the systematic steps involved in our algorithm flow, which serves as the foundation for our approach to enhancing classification performance. The dataset section highlights the specific ocular disease images utilized, ensuring relevance to our objectives of improving diagnostic accuracy in ocular disease detection through advanced synthetic data generation. We then discuss classic data augmentation methods, addressing challenges related to data scarcity and their application in ocular eye disease diagnosis. Following this, we introduce diffusion-based data augmentation, emphasizing the significance of fine-tuning and prompt engineering in generating high-quality synthetic data. The impact of t-SNE-based image set dilution on data interpretation is also examined. Finally, we conclude with an overview of model training and evaluation processes, demonstrating how these methodologies collectively contribute to our findings.

2.1 Algorithm flow

In our comprehensive methodology as shown in Fig. 1, we embark on a six-step process to leverage the power of synthetic data for enhancing classification performance. This systematic approach ensures a robust assessment of the effectiveness of our synthetic data augmentation strategy.

Step 1: Fine-tune the Diffusion-Based Data Augmentation Model: In the first step of our process, we fine-tuned the diffusion-based image generation model (Stable Diffusion) based on specific classes and instances that we wish to generate. In order to generate synthetic data that satisfies our desired criteria, this customized fine-tuning is essential.

Step 2: Generating a Synthetic Dataset: Following the model's fine-tuning, we employed a Stable Diffusion model to generate a substantial synthetic dataset. This dataset forms the basis for the creation of our synthetic data (Fig. 2).

Step 3: Labeling the Synthetic Data: Upon the creation of the synthetic dataset, which consists of unlabeled data, it becomes necessary to assign labels to this data. To accomplish this, we constructed a disease classification model. In this model, we utilized a pre-trained VGG19. This model was fine-tuned using real data, and its classification accuracy became a benchmark for our subsequent steps (Fig. 2).

Step 4: Data Reduction for Dominant Classes (Diabetes and Normal) using t-SNE: Image selection of dominant classes, namely diabetes and normal, was made using the t-SNE method referred to as "Selected Images" (Fig. 3).

Step 5: Integrating Synthetic Data, Real Data, and Selected Images: With the completion of the labeling process through the classification model, the synthetic data and selected images were integrated with our existing real data. This union resulted in the formation of new datasets.

Step 6: Evaluation and Testing for Different Cases: The final step involves subjecting the newly formed Hybrid datasets to a series of tests. We carefully assessed the impact of the synthetic data addition on accuracy. Once again, a pre-trained VGG19 model was utilized. This step

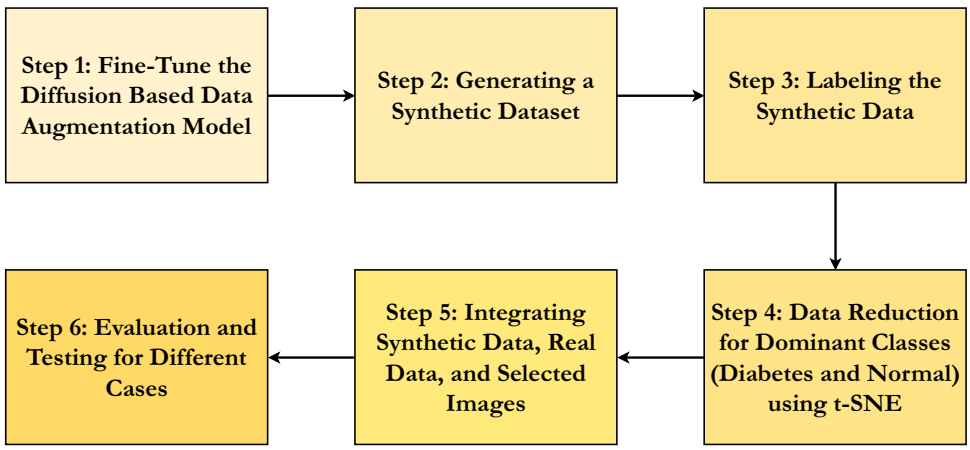


Fig. 1 Flow diagram of the used methods in the study

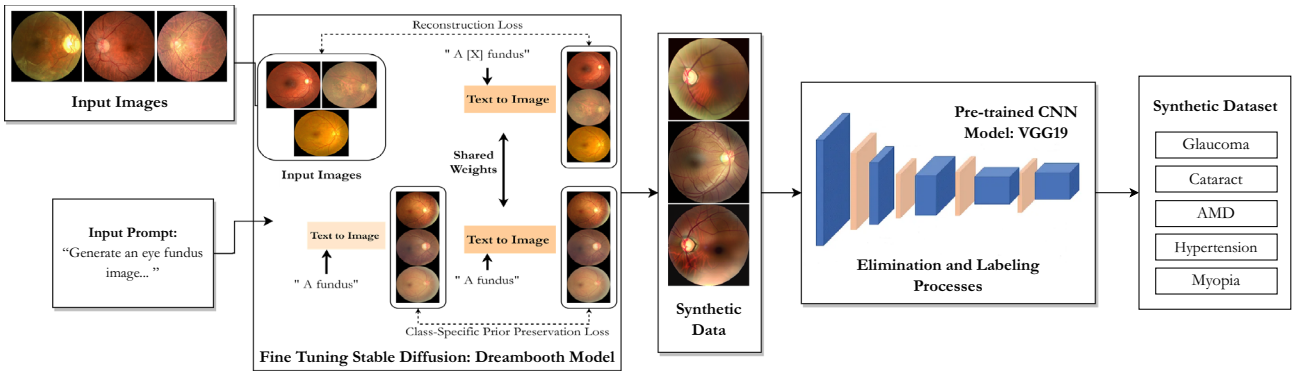


Fig. 2 Illustration of the diffusion-based synthetic data generation methodology for ocular disease diagnosis. The process involves using a Stable Diffusion model that has been fine-tuned on a large dataset of real ocular disease images, allowing it to capture the statistical relationships and patterns within the data. These synthetic images are then labeled by a custom disease classification model, using a pre-trained VGG19 network, that has been trained to identify and classify eye diseases

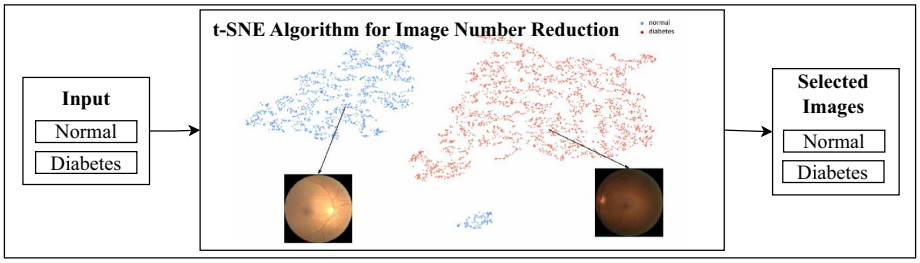


Fig. 3 t-SNE dimensionality reduction technique was employed to down-sample the ODIR dataset, focusing on the prevalent Normal (N) and Diabetes (D) classes. This reduction facilitates a more targeted analysis and investigation of features in the context of the Normal and Diabetes classes, contributing to a detailed understanding of these common ocular disorders

allowed us to draw conclusions and provided insights into the effectiveness of our synthetic data augmentation in enhancing classification performance.

The study comprises six distinct scenarios for 5-class and 7-class classifications, as illustrated in Fig. 4.

Case 1: Classification using only real data from 5 classes. These classes are Glaucoma, Cataract, AMD (Age-related Macular Degeneration), Hypertension, and Myopia. The classification process is conducted using only real data fed into the VGG19 model.

Case 2: Synthetic data is added to the dataset of 5 classes. A hybrid dataset is created by combining synthetic data with real data. The classification is performed using the VGG19 model with both synthetic and real data, aiming to improve performance through the inclusion of synthetic data.

Case 3: Classification using only real data from 7 classes. The additional classes in this case are Normal and Diabetes. The VGG19 model is trained on the real dataset consisting of Normal, Diabetes, Glaucoma, Cataract, AMD, Hypertension, and Myopia classes.

Case 4: Synthetic data is added to the dataset of 7 classes, which initially consisted of only real data.

Case 5: Classification of selected images combined with real images from the remaining 5 classes. The model is trained on a dataset consisting of selected real images and additional data from the 5-class set.

Case 6: Classification is conducted after adding real and synthetic data from 5 classes to the selected images. The hybrid dataset, consisting of both real and synthetic data, is used to train the VGG19 model for classification.

2.2 Dataset

Ocular Disease Intelligent Recognition (ODIR) [35] is a database of 5000 patients with information about their age, color fundus photographs from both eyes and doctors' diagnostic keywords. The dataset was collected by Shanggong Medical Technology Co., Ltd. from different hospitals and medical centers in China. Fundus images were captured in these institutions. In a total of 6392 images, the distribution can be seen in Fig. 5 under the title **"Original Dataset"** .

While creating experiment datasets, to prove that the novel methodology is working, the ***"Other"*** category canceled out. It contains a diverse range of diseases; however, the number of images for each disease is insufficient to extract meaningful

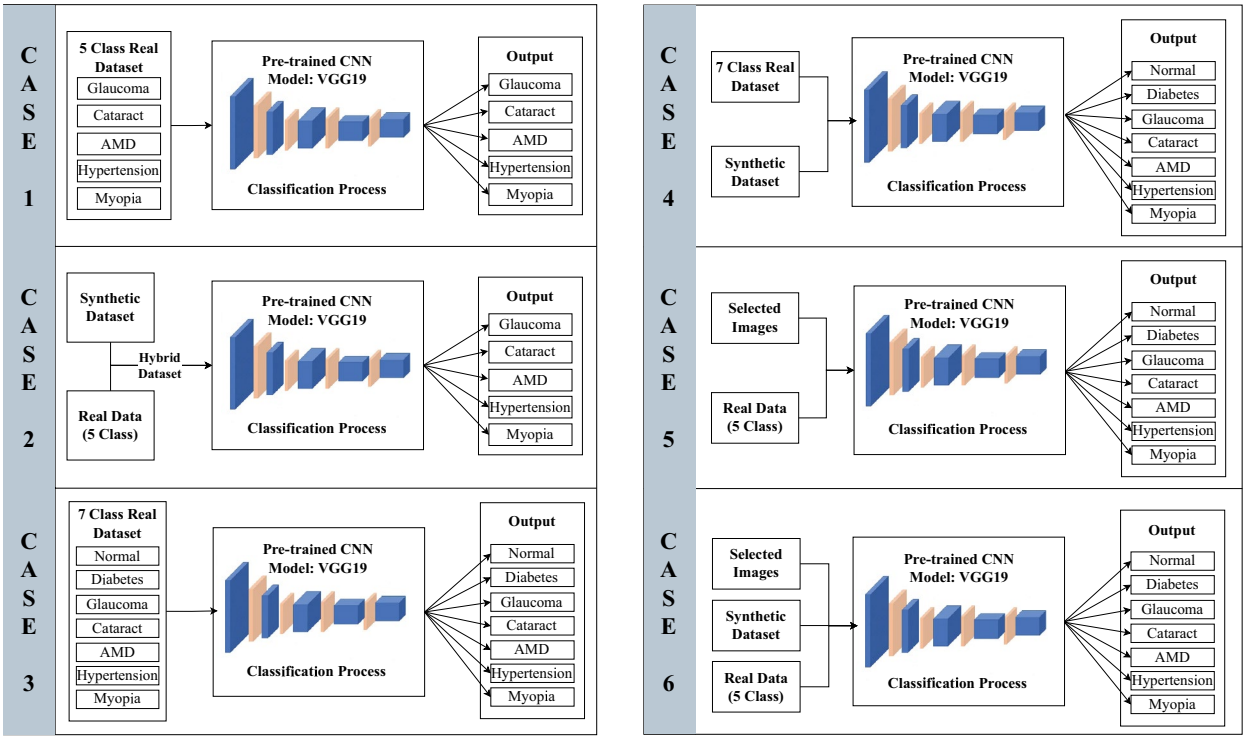


Fig. 4 Different cases of classification processes in the study using the VGG19 model. In Case 1, classification is performed using only real data from 5 classes. Case 2 involves the addition of synthetic data to the 5-class dataset to enhance the model's performance. Case 3 extends the classification to 7 classes, using only real data. Case 4 builds on this by adding synthetic data to the 7-class dataset. Case 5 focuses on classifying selected images combined with real data from the remaining 5 classes, while Case 6 incorporates both real and synthetic data from the 5 classes to classify the selected images

features. In order to create new examples with the Stable Diffusion model, we need to have some specific features so that the model can learn from these features and generate new data accordingly.

The dataset created with an additional synthetic quantity, twice the size of the original dataset, incorporating five classes, is observed in Fig. 5 under the title **"Hybrid Dataset"**.

The imbalance issue identified in the original dataset has been alleviated through the implementation of our methodologies, which involve synthetic data integration and the use of t-SNE. This improvement has been visually alleviated, as shown in the dataset distribution named **"Balanced Hybrid Dataset"** in the same figure.

The ODIR dataset is meant to represent real-world patient information and can be used to develop and evaluate machine-learning models for ocular disease detection and diagnosis. The samples of the dataset are represented below in Fig. 6.

2.3 Classical data augmentation

2.3.1 Data scarcity and classical data augmentation

The data used in deep learning models may not always be sufficient. Data scarcity often leads to a negative result in model performance, at which point classical data augmentation methods step in such as rotation, flipping, cropping, brightness, and contrast adjustment [40]. More importantly, these strategies can also be effective in addressing dataset imbalance and further improve the overall robustness of the model. By adding diversity, data augmentation techniques can help prevent overfitting and empower models to be trained with more variation, which improves the model's performance and adaptability under various conditions. Additionally, these techniques can address the problem of data insufficiency by making the most out of existing data. However, ensuring data augmentation methods are appropriate to the data type and application is essential. A model becomes more robust and exhibits a higher generalization ability only when the appropriate data augmentation strategies are implemented.

2.3.2 Data augmentation in ocular eye disease diagnosis

In this study, a dataset containing eye diseases was examined. Retinal scans are performed when diagnosing eye diseases. The retina is the light-sensitive tissue layer located on the inner side of the eye. Retinal scans are performed to identify any abnormalities in the retina, such as

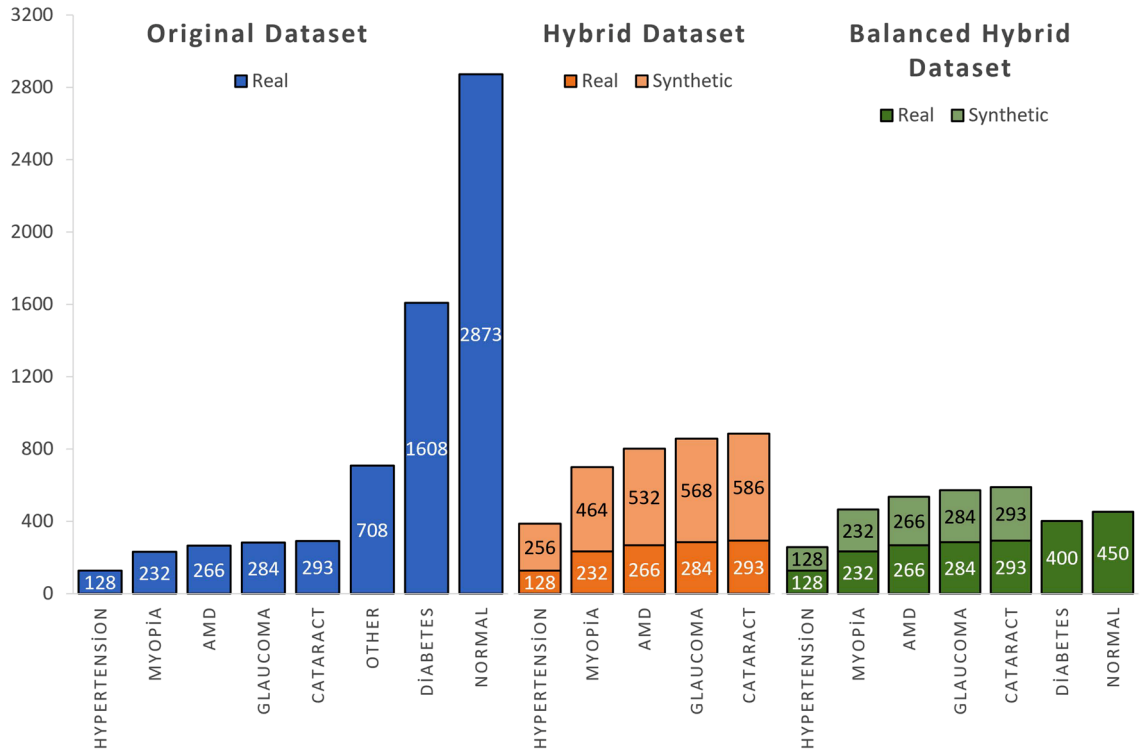


Fig. 5 Trained human readers with quality control management annotated the images, classifying patients into eight categories: Normal (N) with 2873, Diabetes (D) with 1608, Glaucoma (G) with 284, Cataract (C) with 293, Age-related Macular Degeneration (A) with 266, Hypertension (H) with 128, Pathological Myopia (M) with 232, Other diseases/abnormalities (O) with 708 images. Three different datasets are shown in the figure as the original dataset (blue color), a hybrid dataset with twice the amount of synthetic data combined with the real data (orange color), and the dataset generated after applying the t-SNE technique to the clusters of normal and diabetes data (green color). The "other" class, which includes various abnormalities, was excluded from the augmentation and classification processes due to insufficient features representing multiple diseases. The Hybrid Dataset consists of five classes, with dominant classes removed and all the remaining 5 classes augmented to twice their original size. The Balanced Hybrid Dataset, containing seven classes, is formed by combining the normal and diabetes classes with the Hybrid Dataset after the t-SNE process

hemorrhages, spots, or vascular structure [41]. Therefore, when data augmentation methods are used in medical diagnoses such as eye diseases, a delicate approach is essential. Making excessive changes to such data can potentially disrupt important features such as eye vascular structure. For instance, operations like blurring, contrast changes, and similar operations can cause data loss in images and reduce the accuracy of diagnosis. It is crucial to strike a careful balance between augmenting data for better model performance and maintaining the integrity of medically relevant features.

2.4 Diffusion-based data augmentation

In the ever-evolving field of deep learning, the insatiable hunger for data to train more robust models has led to the development of various data augmentation techniques. In response to this challenge, the field of Generative AI [4, 42], particularly diffusion-based models, was leveraged to generate synthetic data replicating complex vascular structures within the retina. Diffusion-based data augmentation is a technique used to enrich datasets and improve the performance of machine learning and deep learning models by creating additional data points through a diffusion process. It involves generating new data points, starting from existing ones, and allowing them to propagate or *"diffuse"* in a way that preserves their similarity to the original data. These augmented data points are then added to the training dataset, making it larger and more diverse. By leveraging diffusion-based data augmentation, machine learning models can better generalize and learn complex patterns, especially when facing issues of limited or imbalanced data, ultimately improving their performance. This technique is widely applied in fields such as image processing and natural language processing.

2.4.1 Stable diffusion fine-tuning

The cornerstone of our approach lies in the implementation of Stable Diffusion, a state-of-the-art generative model that exhibits exceptional stability and accuracy in

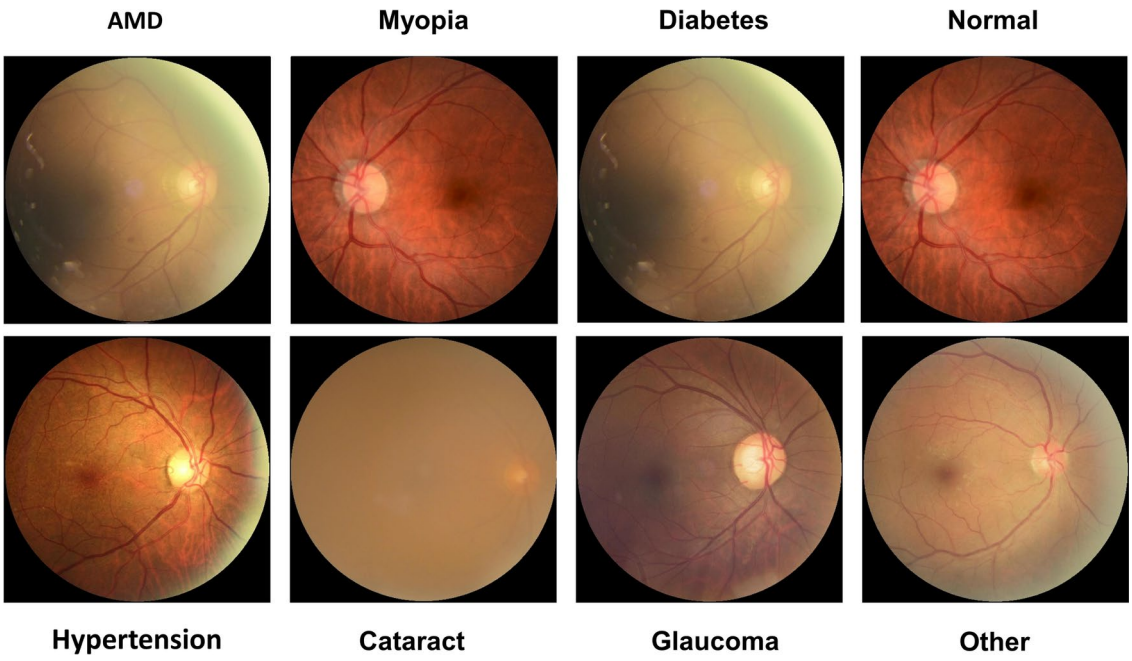


Fig. 6 Sample images from the dataset. Each subfigure displays a representative example from different classes

generating high-quality synthetic data. A journey of fine-tuning was embarked upon to leverage the full potential of Stable Diffusion for our mission. Stable Diffusion v2.1 [1] was used. Fine-tuning involved training the Stable Diffusion model on a selected subset of the ODIR dataset; This allowed the model to learn and copy complex patterns and structures within retinal images. Fine-tuned Stable Diffusion model serves as the cornerstone of our data augmentation pipeline and facilitates the creation of synthetic data that closely mimics the complexities found in real-world retinal images.

To fine-tune the Stable Diffusion model, DreamBooth [43] was utilized. It offers a powerful and user-friendly environment, enabling precise fine-tuning of Stable Diffusion models. DreamBooth's intuitive interface streamlines prompt creation, data selection, and fine-tuning for effortless model customization. Empower users with granular control over text prompts, data access, and domain-specific adjustments to tailor model behavior precisely to their needs. Having specified the desired parameters, prompts, and data, while also ensuring proper configuration, the training process utilized a batch size of 1, a learning rate set at 1e-6, and 1000 training steps for every 10 photos. For each class, 100 photographs were fed to the model as data. For this reason, 500 photos were uploaded as data from the 5 classes we trained. Therefore, the model was trained in 50,000 steps with a Tesla T4 GPU for 5 h.

2.4.2 Prompt engineering

Stable Diffusion uses the CLIP (Contrastive Language-Image Pretraining) text [44] encoder to understand and respond to text prompts. During our development phase for this work, this interaction between the model and the text prompts was limited by the 77-token limit, which means that the prompts provided to the model must be no more than 77 tokens long to maintain efficient computation and memory usage [44]. When working with Stable Diffusion, researchers and practitioners must be thoughtful about the prompts they provide to the model. Despite its limitation, the 77 token limit has not hindered innovation in Stable Diffusion and in our development, we were able to successfully work around the limitation. We note that the limit is no longer an issue [45], so future development may not require consideration of this issue. However, we believe our strategy was successful so we do not believe revisiting the development would have significant advantages. Researchers and practitioners have developed a variety of techniques to optimize their prompts and extract high-quality and contextually relevant responses from the model. This has led to the development of new and creative applications for Stable Diffusion, such as generating images from text descriptions, creating photo-realistic images, and editing existing images. Related to this, we have invested in the art of prompt engineering [46–48] to achieve the highest efficiency and detail in the outputs of our fine-tuned Stable Diffusion model. Leveraging the

power of language models, we crafted prompts that provide precise instructions to the model, directing it to produce synthetic images that not only reflect the richness of vascular structures but also exhibit a striking resemblance to reality. Prompt engineering was crucial in bridging the gap between synthetic and original, enabling our synthetic data to integrate seamlessly with the original dataset. While creating the model, a concept list needed to be arranged. This concept list includes `instance_prompt, class_prompt, instance_data_dir, and class_data_dir`. These prompt and data directory variables are set according to the above narrative. Below is an explanation of the functions of these concepts.

`instance_prompt`: An instance prompt is a specific instruction or explanation provided to Stable Diffusion during fine-tuning. It helps the model understand and generate specific data instances or instances. For example, in fine-tuning Stable Diffusion to create images of dogs, the ***"a photo of a Golden Retriever"*** sample prompt can be used. In our case, we are trying to fine-tune a model that cannot fully encompass the demanded visual forms. Therefore, it is important to find a prompt that will not overlap or intersect with any existing prompts. In our example, ***"a photo of ocudisAMD"*** serves as the instance prompt. This prompt indicates that the model should focus on generating images related to ocudisAMD(Age-related macular degeneration), which is a specific condition class of eye disease.

`class_prompt`: A class prompt is a broader instruction that defines the properties and characteristics that the generated data should exhibit. For example, in fine-tuning Stable Diffusion to create images of dogs, the class prompt ***"photo of a dog"*** can be used. In our case, the class prompt is quite detailed, specifying that the generated image should represent an eye fundus image with particular attributes related to age-related macular disease. This prompt provides overarching guidance for the generated data. This part is making a real difference. Since these text-to-image models require highly detailed prompts to generate optimal images, providing precisely what is asked for becomes essential.

`instance_data_dir`: The sample data directory is the location from which Stable Diffusion accesses specific sample-level data during fine-tuning. This data may include real-world images, text descriptions, or other data types. In our case, ***"/content/data/ocudisAMD"*** contains the real-world images related to ocudisAMD. Real data is uploaded to this directory. During fine-tuning, the model can learn from these actual examples to generate synthetic data that aligns with the characteristics of ocudisAMD.

`class_data_dir`: The class data directory is the location from which Stable Diffusion accesses class-level or category-level data during fine-tuning. This data may include real-world images, text descriptions, or other types of data representing different classes or categories. In our case, ***"/content/data/eye_fundus"*** contains data related to different eye fundus images. The model creates this data by itself and uses it to understand the broader context and characteristics of eye fundus images, especially those associated with different diseases or conditions.

2.5 t-SNE-based image set dilution

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a data mining and visualization technique that converts multidimensional data into a lower-dimensional space. It projects data onto a lower-dimensional space while preserving crucial structure, placing similar points close and dissimilar ones far apart The main purpose of t-SNE is to visualize or analyze multidimensional data in a more understandable and interpretable form.

Suppose we have a dataset with n data points and a similarity matrix S; Here S*ij* represents the similarity between data points *i* and *j*. The first thing we do is compute a similarity matrix P in the multidimensional space. Given the Gaussian kernel centered at j, each element *pi*|*j* reflects the conditional probability that x data point *i* is a neighbor of data point *j*. The *pi*|*j* equation is:

$$p_{j|i|} = \frac{e^{\frac{-\|x_i - x_j\|^2}{2\sigma_i 2}}}{\sum_{k \neq i}^{N} e^{\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}}} \tag{1}$$

In this expression, x_i, and x_j stand for individual data points within a high-dimensional space while σi represents the variance of a Gaussian distribution centered on data point *i*. The denominator is a normalization constant, and it is designed in a way to make sure that the overall value of $\sum j \cdot p_{ij}$ equals 1.

Next, a similarity matrix Q is calculated within the low-dimensional space. In this matrix, each q*ij* element signifies the probability of data point *i* and data point *j* being neighbors via t-distribution. The equation for q_{ij} is used to calculate this probability.

$$q_{j|i} = \frac{\left(1 + \| y_i - y_j \|^2\right)^{-1}}{\sum_{k \neq i}^{N} \left(1 + \| y_i - y_k \|^2\right)^{-1}} \tag{2}$$

In this expression, y_i and y_j are the respective points located in the low-dimensional space. Thirdly, our next step involves the computation of the Kullback–Leibler (KL) divergence between P and Q, which measures the dissimilarity between these two probability distributions. The equation used for calculating the KL divergence is as follows:

$$D_{KL}(P \parallel Q) = \sum_{j=1}^{N} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{3}$$

In order to refine the low-dimensional embedding and minimize the KL divergence, an iterative process is employed. The gradient descent approach is essential to this quest. It involves moving data points around in the low-dimensional space. This update is carried out using the equation denoted as 4.

$$y_i(t+1) = y_i(t) + \eta \sum_{j=1}^{N} \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right)\left(1 + \parallel y_i - y_j \parallel^2\right)^{-1} \tag{4}$$

In this expression, the learning rate is represented by η, and 't' stands for the current iteration number. The positions of data point 'i - th' in the lower-dimensional space before and after the update is denoted as '$y_i(t)$' and '$y_i(t+1)$' respectively.

2.6 Model training and evaluation

In this work, the VGG19 architecture was utilized to create a robust model for image classification. A base model was constructed using a pre-trained VGG19 network from the extensive ImageNet [49] dataset. Subsequent adjustments were made to fit this base model to the dimensions of our visual data. During training, VGG19's weights were not frozen, allowing the network properties to better adapt to our data. The inclusion of Global Average Pooling and Batch Normalization layers further customized our model. A dropout layer was introduced as a precautionary measure against overfitting. The final classification layer was attained by incorporating fully connected layers equipped with a softmax activation function capable of classifying into five different classes. This meticulously crafted model is now poised for training, geared towards the creation of a high-performance image classification model. When compiling our model, the Adam optimization algorithm [50] is employed, along with the specified learning rate, to optimize it. The multi-class classification task is tackled using ***"categorical_crossentropy"*** as the chosen loss function. Subsequently, the metrics parameter is utilized to specify the metrics we want the model to monitor during training. These metrics enable the detailed monitoring of the model's performance, aiding in the evaluation of training results.

The dataset was divided into three distinct partitions for training %70, validation %15, and testing %15. These partitions were named training data, validation data, and test data. The test data was consistently maintained to achieve stable results and enhance the model's generalization capabilities. Model training was carried out on this fixed test set. Additionally, **no synthetic data was added to the test data**. In this way, the model's performance and success criteria were meticulously evaluated to better suit it for real-world applications.

In our model evaluation, various metrics were considered to assess its performance, given the presence of a class imbalance problem. While accuracy and loss were deemed important, a comprehensive understanding of our method's effectiveness was sought. Consequently, additional metrics, including F1-Score, recall, and precision, were closely examined. These metrics facilitated a more thorough comprehension of our model's performance, especially in situations where class imbalances could influence the accuracy of the results.

Accuracy measures the overall correctness of the model by calculating the proportion of true predictions (both positives and negatives) out of the total number of predictions. While it's a popular metric, it can be misleading in cases of imbalanced datasets where one class significantly outweighs the others, its formulation can be seen in Eq. 5 [51].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where:

TP = True Positives (correctly predicted positive cases)
TN = True Negatives (correctly predicted negative cases)
FP = False Positives (incorrectly predicted as positive)
FN = False Negatives (incorrectly predicted as negative)

Precision evaluates how well the model identifies positive predictions by measuring the proportion of true positives among all instances predicted as positive. High precision indicates a low rate of false positives, making it an important metric in scenarios where false positives are costly, its formulation is presented in Eq. 6 [52].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall measures the ability of the model to correctly identify all actual positive instances, by calculating the proportion of true positives out of all actual positives. This metric is particularly important when missing true positives (i.e., false negatives) could have serious consequences, such as in medical diagnoses, the formulation is provided in Eq. 7 [53].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

F1-Score is the harmonic mean of precision and recall, offering a balanced metric when there is a trade-off between precision and recall. It's especially useful for imbalanced datasets where accuracy may not provide an accurate representation of model performance, the formulation is given in Eq. 8 [52].

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

3 Results and discussions

In this section, we examine the impact of custom prompts on generating synthetic data, with a particular focus on the role of fine-tuning and prompt engineering. The main objective is to demonstrate that the methodology employed for synthetic data generation, particularly for the complex dataset used in this study, offers superior generalization compared to traditional data augmentation techniques.

The "other" class was excluded from the entire process. This decision was made because the"other" class encompasses a wide range of abnormalities, and there were insufficient distinguishing features to represent multiple diseases simultaneously. As a result, this class was not considered in the augmentation or classification stages, leaving a 7-class dataset. Within this dataset, diabetes and normal classes dominate the remaining five classes after excluding the "other" category.

To evaluate the effectiveness of synthetic data, we first removed the dominating classes and formed a 5-class dataset. The results demonstrated a positive impact of synthetic data on model performance, particularly within this 5-class configuration. We further compared the outcomes of Diffusion-Based Data Augmentation with traditional augmentation techniques. Additionally, this section explores the impact of incorporating synthetic data into the imbalanced 7-class dataset, with performance analyzed through t-SNE operations on the diabetes and normal classes. Lastly, we discuss the limitations of synthetic data augmentation, including performance saturation, and the implications for model accuracy.

3.1 Creating synthetic data using customed prompts

In the ever-evolving field of deep learning, the pursuit of abundant data to train robust models has led to the development of diverse data augmentation techniques. While classical augmentation methods serve their purposes effectively in many domains, certain datasets present unique challenges. Notably, the Ocular Disease Intelligent Recognition (ODIR) dataset, containing essential retinal images for diagnosing eye diseases, posed such challenges. Traditional data augmentation techniques [54, 55] were found insufficient for this intricate dataset. In response, Generative AI, specifically Stable Diffusion, was delved into to generate synthetic data replicating complex vascular structures within the retina.

One of the significant challenges encountered in working with the ODIR dataset was the presence of class imbalance. Specifically, five classes were notably underrepresented compared to others. This class imbalance issue, commonly found in real-world datasets, including medical imaging, can introduce bias in models. In order to mitigate this challenge, data augmentation played a pivotal role. Traditional data augmentation techniques turned out to be inadequate for this complex dataset. With this, we aimed to balance class representation within the dataset by using the Stable Diffusion Model. This strategic expansion provided more equitable exposure to less frequent classes, enabling effective learning and generalization across all classes. As a result, our model's diagnostic capabilities became more balanced and robust, alleviating the skew toward the majority classes.

In the ODIR dataset, five classes were significantly dominated by other, diabetes and normal classes regarding available data. In order to rectify this imbalance and prevent the model from neglecting these critical categories, we implemented a selective boosting strategy. Extensive augmentation efforts were directed towards these non-dominated classes. This approach aimed to level the playing field and ensure that the model received ample exposure to even the most underrepresented conditions. Practical implementation required feeding the model with more extensive data and specific prompts during fine-tuning. Consequently, our model emerged as a more capable diagnostic tool, demonstrating enhanced performance across all classes, regardless of their initial representation in the dataset.

During the model's creation, we organized a concept list including instance_prompt, class_prompt, instance_data_dir, and class_data_dir. Figure 7 shows the prompts we used, the resulting synthetic data as a result of fine-tuning with these prompts, and their comparison with the real data of the specified classes and the synthetic data produced. As shown in Fig. 7, generated synthetic data looks like a variation of real data as expected. It broadens up the data pool without adding any additional features. It used only the features provided to it and did not deviate from these. These features were provided both with the prompt and with real data.

3.2 Fine-tuning and prompt engineering significantly influenced the quality of synthetic data

In order to understand the importance of prompt engineering and fine-tuning, we have investigated their impact. Figure 8 below contains both basic Stable Diffusion model outputs and Stable Diffusion outputs fine-tuned according to written prompts. The results of our study underscore the pivotal role of prompt engineering and fine-tuning in enhancing data augmentation for Glaucoma, Hypertension, Cataract, Age-related Macular Degeneration, and Pathological

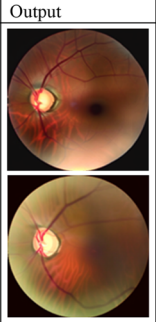
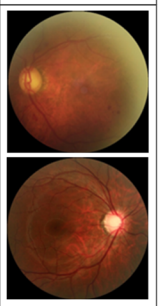
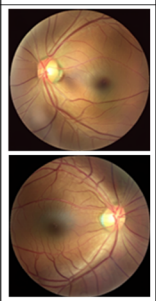
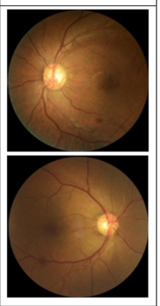
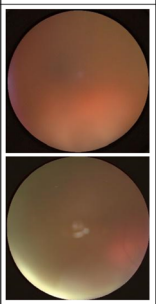
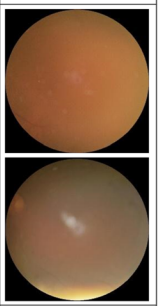
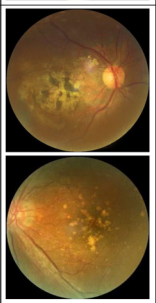
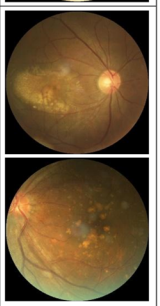
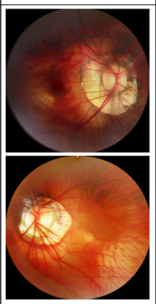
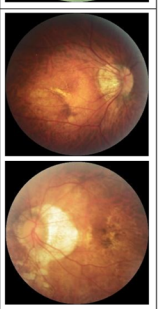
Disease	Prompt	Fine-Tuned Stable Diffusion Output	Real Image
Glaucoma	*"Generate an eye fundus image. The entire eye fundus should be considered as the foreground, and the background (areas outside the eye) should be **completely black**. Glaucoma can cause the retina to thin. This makes the retina appear lighter in color than a normal retina and blood vessels appear redder than normal. Disease: glaucoma"*		
Hypertension	*"Generate an eye fundus image. The entire eye fundus should be considered as the foreground, and the background (areas outside the eye) should be **completely black**. In this disease, vascular wall changes, flame-shaped hemorrhages, and cotton-wool spots are seen in retina. Disease: hypertension."*		
Cataract	*"Generate a left eye fundus image. The **entire eye fundus** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black**. This **disease clouding the retina. Retina is more opaque and blurry. Blood vessels cannot be seen properly.** Disease:cataract"*		
Age-related Macular Degeneration	*"Generate a left eye fundus image. The **entire eye fundus** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black**. In this disease, **retina is less pigmented and more flecked than normal retina.** Disease: age-related macular disease"*		
Pathological Myopia	*"Generate a left eye fundus image. The **entire eye fundus** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black**. In this disease **retina is stretched and thinned, with thin blood vessels, a distorted optic nerve head, some folds in the thin retina, and pigmentation at the fovea.** Disease: Pathological myopia"*		

Fig. 7 Comparison between real image and synthetic image output generated by fine-tuned Stable Diffusion. The synthetic images were produced to augment the five classes in the dataset. These classes include various ocular diseases, where synthetic data generation helped preserve critical features, such as vascular structures, crucial for accurate diagnosis. This comparison illustrates how well the fine-tuned model captures key visual characteristics from the real images, aiding in improving classification accuracy and addressing class imbalance

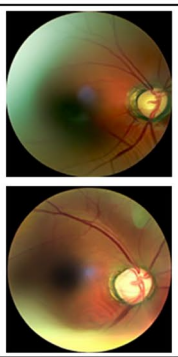
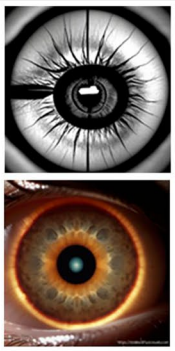
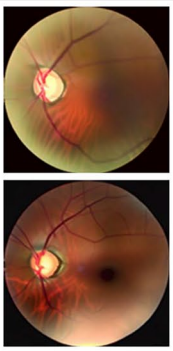
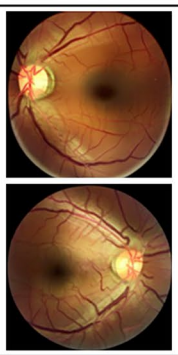
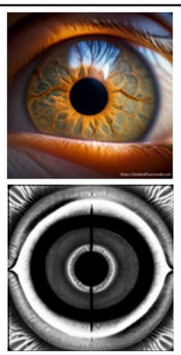
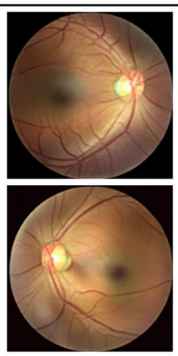
Disease	Prompt	Stable Diffusion Output	Fine-Tuned Stable Diffusion Output
Glaucoma with a generic prompt	*"Generate an eye fundus image where **the blood vessels are prominently displayed as a foreground** within the eye. The **entire eye** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black.** The image should capture **fine details of blood vessel structure.** Ensure that image adheres to the circular shape of eye fundus. Disease: glaucoma"*		
Glaucoma with illness specific prompt	*"Generate an eye fundus image. The **entire eye fundus** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black**. Glaucoma can cause the **retina to thin.** This makes the **retina appear lighter** in color than a normal retina and **blood vessels appear redder** than a normal. Disease: glaucoma"*		
Hypertension with a generic prompt	*"Generate an eye fundus image where the **blood vessels are prominently displayed as a foreground** within the eye. The **entire eye** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black. The image should capture fine details of blood vessel structure.** Ensure that image adheres to the circular shape of eye fundus. Disease: hypertension"*		
Hypertension with illness specific prompt	*"Generate an eye fundus image. The **entire eye fundus** should be considered as the **foreground**, and the **background** (areas outside the eye) should be **completely black.**In this disease, **vascular wall changes, flame-shaped hemorrhages, and cotton-wool spots are seen** in retina. Disease: hypertension"*		

Fig. 8 Demonstrating the effectiveness of prompt engineering and fine-tuning in data augmentation for Glaucoma and Hypertension disease. Generic prompts in data augmentation yielded Glaucoma images lacking vessel details and Hypertension images with exaggerated structures. Specific prompts significantly improved realism, accurately capturing disease characteristics in synthetic data

Myopia diseases. Focusing on Glaucoma, the use of generic prompts during synthesis led to synthetic images that lacked detailed blood vessel structures, resulting in a suboptimal representation of the actual data. Conversely, employing specific prompts significantly improved the synthesis process, capturing intricate details accurately and providing a more faithful representation of Glaucoma characteristics. In the context of Hypertension, generic prompts tended to exaggerate blood vessel structures, deviating from realistic data representation. For Cataract, generic prompts produced synthetic images with overly smooth retinas, failing to capture the characteristic cloudiness and opacity. In age-related macular disease, generic prompts led to synthetic images that missed the distinct flecked and depigmented appearance of the retina, resulting in less realistic outputs. Lastly, for pathological myopia, generic prompts produced synthetic images that lacked the characteristic retinal thinning and distortion of the optic nerve head, resulting in an oversimplified depiction. This overemphasis was effectively corrected through the application of specific prompts, ensuring a more authentic depiction of the disease's characteristics in the generated images. These findings emphasize the critical importance of tailored prompt selection in achieving meaningful and accurate data augmentation. Fine-tuning, as demonstrated in Fig. 8, is crucial, but the choice of prompt during this process is equally significant. Figure 8 illustrates that when examining the model's output images based on the prompts, categorized as generic and specific, the significance of prompt selection becomes evident.

3.3 Demonstrating the positive impact of synthetic data on model performance

Addressing the issue of class imbalance, the application of synthetic data augmentation was considered. However, in order to assess the quality of these generated synthetic data and to understand how they influence the patterns recognized by the classification model, a specific case was studied. This involved the removal of dominant classes, focusing solely on the underrepresented classes, and augmenting these underrepresented classes with synthetic data generated by diffusion-based models. As a result, a noticeable increase in accuracy, F1-Score, precision, and recall was observed as shown in Table 1, while improvements in class-specific performance were observed as shown in Fig. 9.

The accuracy and loss graphs are illustrated in Fig. 10. As can be seen from the figures, it takes much less time to reach a plateau and shows a higher performance. This provided clear evidence that synthetic data augmentation effectively enhances the performance of the model, prompting the progression to the next phase of the study. The decision to move to the next phase is motivated by the imbalance in the dataset, specifically the need to experiment with all seven classes.

Table 1 5 class classification results with synthetic data augmentation. Scenario 1 (sc1): real data, Scenario 2 (sc2): real data + x2 synthetic data

Metrics	sc1	sc2
accuracy	80.8%	**86%**
F1-Score	79.2%	**85.2%**
precision	79.4%	**85%**
recall	79.4%	**86%**

3.4 Impact of diffusion-based data augmentation on model performance versus traditional methods

The utilization of Stable Diffusion-based data augmentation proved to be significantly more effective than traditional data augmentation methods. Diffusion-based data augmentation is a technique that enhances the integrity and relevance of data by using a range of advanced techniques. Unlike conventional methods that often distorted or over-processed the data, diffusion-based techniques consistently produced augmented data closely aligned with the underlying data distribution. This resulted in models better equipped to generalize from the augmented data to unseen examples, ultimately improving accuracy and performance. Additionally, our decision to focus on 5-class scenarios was influenced by the need to demonstrate the saturation point of data augmentation benefits more clearly, as traditional methods typically struggle to replicate the complexities present in ocular disease images. Diffusion-based data augmentation is a beneficial option for strengthening model training and prediction quality because it maintains the key features of the original data, which helps reduce the risk of overfitting and improves the robustness of machine learning models.

As shown in Table 2 above, in the best-case scenarios, diffusion-based data augmentation has a much better impact on model performance compared to traditional data augmentation.

3.5 Impact of adding synthetic data for a 7-class dataset

After establishing the effectiveness of synthetic data, the next phase of the study involved the inclusion of dominant classes into the training process. An attempt was made to bridge the gap between the underrepresented classes and the dominant ones by augmenting the underrepresented classes with synthetic data. Subsequently, certain increases were observed as shown in Table 3.

While Table 3 and Fig. 11 demonstrate improved performance with synthetic data augmentation, a saturation point emerges beyond which further augmentations lead

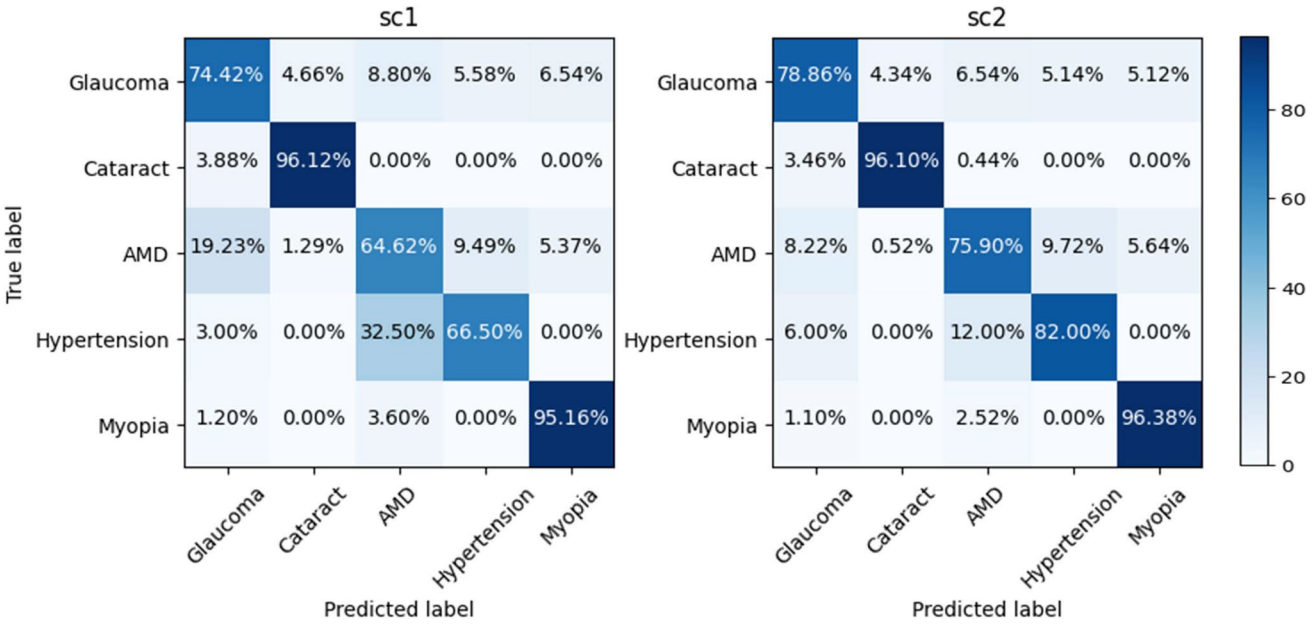


Fig. 9 Confusion matrices of sc1 and sc2 for 5 classes

to diminishing returns. As a result, it becomes difficult to perfectly align underrepresented classes with the dominant classes, leading to some distortions.

Moreover, performance declines when real data density falls below a critical threshold. Adding smaller quantities of data beyond the saturation point yielded limited improvements, suggesting the presence of a more fundamental issue: class imbalance. To address this challenge, we implemented a dedicated selection method that reduces the number of samples from dominant classes, thereby promoting a more balanced data distribution. Further research explored advanced techniques to refine this approach and optimize performance.

3.6 Performing t-SNE operation on two dominant classes, diabetes, and normal classes

After demonstrating the utility of synthetic data in five classes, the issue of dataset imbalance was addressed in Sect. 2.2. To accomplish this, the t-distributed Stochastic Neighbor Embedding (t-SNE) method was applied to the *"normal"* and *"diabetes"* classes, with the exclusion of the *"other"* class at this specific step. Subsequent to the addition of synthetic data to the remaining five classes, as depicted in Fig. 4, testing was completed for all seven classes. The t-SNE process was executed within the Matlab platform. t-SNE effectively visualized distinct underlying data structures in diabetic and normal classes. This implies divergent cluster formations, due to variations in feature distributions between the classes. Within the *"diabetes"* and *"normal"* classes, we employed an image selection process within a two-dimensional plane to effectively reduce the dataset size while ensuring the preservation of class-specific information. This targeted approach enabled the retention of crucial features relevant to these classes, simultaneously achieving the desired data reduction. While making this reduction, adjustments were made to ensure that the dataset was as close to a uniform distribution as possible and to solve the imbalance problem. In Fig. 12, it is evident that images chosen from distant points in the one-dimensional space represent different features of the same disease. Simultaneously, the images outlined in blue in Fig. 12 demonstrate the similarity among images of closely situated feature points within the diabetes class.

3.7 Advancing 7-class classification accuracy through t-SNE and synthetic data integration

After the t-SNE operation, the amount of data for the normal class was reduced from 2873 to 450, and for the diabetes class, it was reduced from 1608 to 400. To balance the resulting dataset, synthetic data as much as themselves were added to the remaining 5 underrepresented classes, and the results in Table 4 were obtained. Accuracy alone can be misleading and may not provide a comprehensive understanding of the model's classification performance. The table results show that the values (F1-Score, recall, precision) obtained with real data after t-SNE are higher than the values obtained with real data before t-SNE. This indicates that t-SNE transforms real data into a feature space that better represents the underlying patterns. Key metrics for assessing data balance-the F1-Score, precision, and recall-also showed

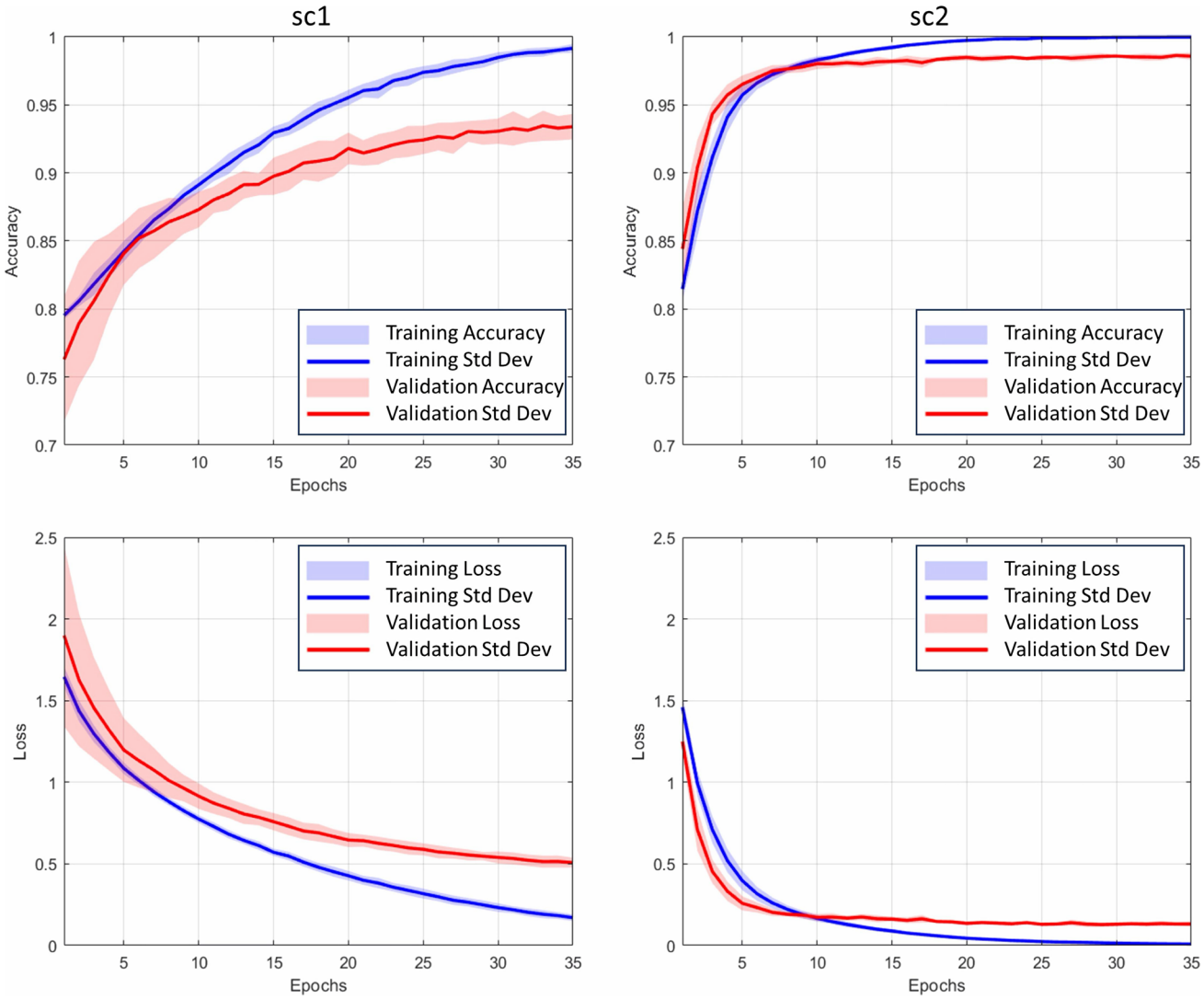


Fig. 10 The accuracy and loss plots for scenario 1 (sc1) and scenario 2 (sc2), exclusively showcase the model's performance on real data and real data + 2x synthetic data, illustrating the impact of synthetic data augmentation on overall performance

Table 2 5 class classification comparison between traditional and synthetic data augmentation results for different scenarios. sc1: real data, sc2: real data + x1 synthetic data augmentation, sc3: real data + x1 traditional data augmentation, sc4: real data + x2 synthetic data augmentation, sc5: real data + x2 traditional data augmentation

Metrics	sc1	sc2	sc3	sc4	sc5
Accuracy	80.8	83.6%	71.8%	**86%**	75.2%
F1-Score	79.2%	82%	70.4%	**85.2%**	73.4%
Precision	79.4%	82.6%	71.2%	**85%**	76.2%
Recall	79.4%	82.2%	71.2%	**86%**	75%
Glaucoma	74.42%	77.2%	67.9%	**78.86%**	67.83%
Cataract	96.12%	**98.7%**	90.86%	96.1%	89.1%
AMD	64.62%	69.22%	49.22%	**75.9%**	58.46%
hypertension	66%	72%	67%	**82%**	74%
Myopia	95.16%	94.54%	80.6%	**96.38%**	84.8%

Table 3 7 class classification results with synthetic data augmentation for different scenarios. sc1: real data, sc2: real data + x1 synthetic data (before t-SNE), sc3: real data + x4 synthetic data (before t-SNE)

Metrics	sc1	sc2	sc3
accuracy	65%	**67.6%**	67.2%
F1-Score	52.4%	**56%**	52.8%
precision	60.2%	61.8%	**62.2%**
recall	50.8%	**54.4%**	51%

significant improvement following the t-SNE operation. The addition of synthetic data further improved these metrics, as detailed in Table 4. The combination of these approaches proved to be an effective strategy for addressing the issue of class imbalance in the dataset. Scenario 1 exhibited higher

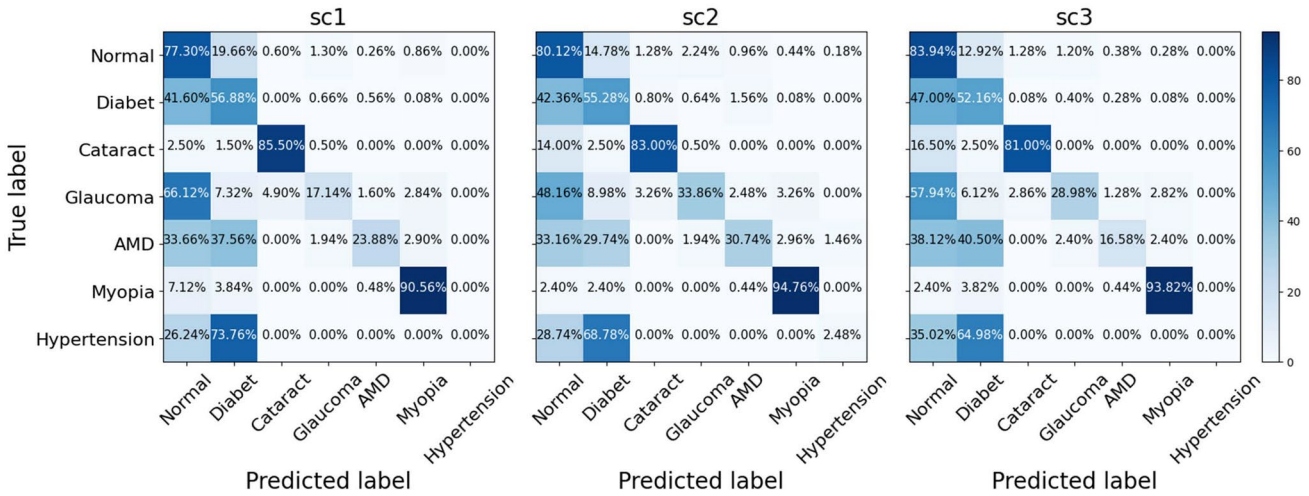


Fig. 11 Confusion matrices of 7 classes before t-SNE with synthetic data augmentation classification. sc1: real data, sc2: before t-SNE real data + x1 synthetic data, sc3: before t-SNE real data + x4 synthetic data

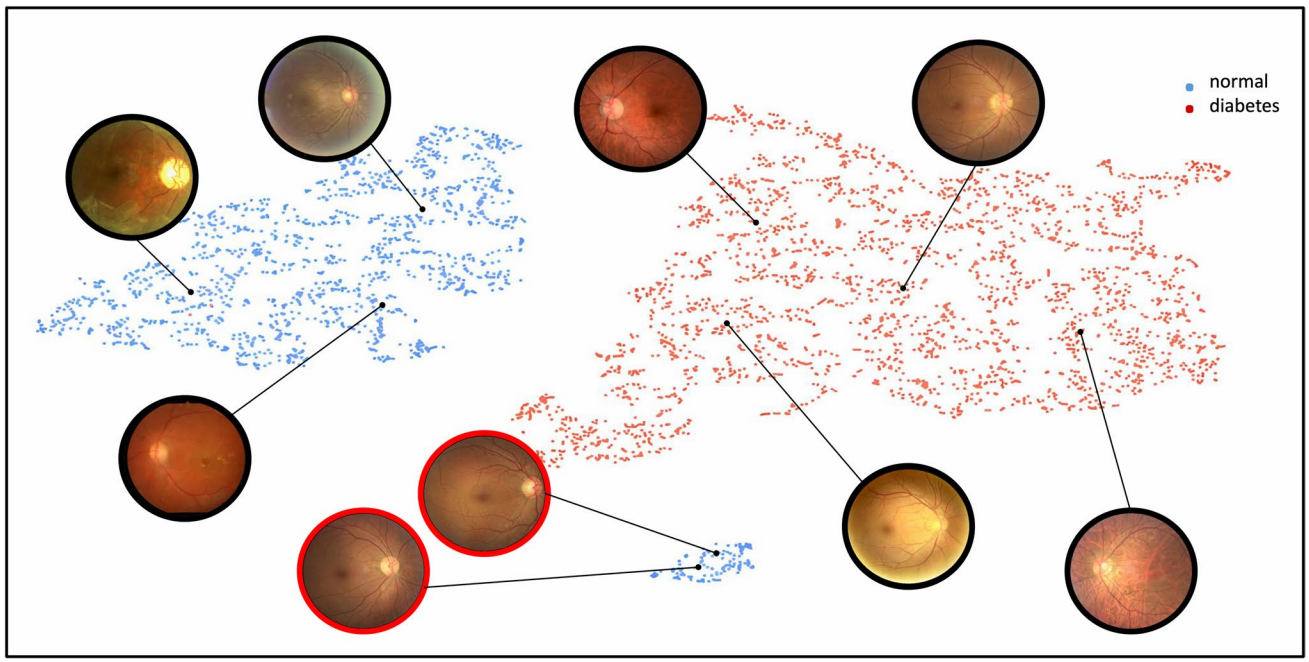


Fig. 12 Selecting images using t-SNE-based dimensionality reduction technique to down-sample the ODIR dataset. Different clustering patterns are observed in t-SNE embedding. Images are selected from different regions to have a uniform distribution across classes. During this reduction process, modifications were implemented to achieve a dataset that closely resembles a uniform distribution, addressing the issue of imbalance. Blue dots indicate the data of the diabetes class, and red dots indicate the data of the normal class. Image differences and similarities can be seen in the randomly selected images in both classes

accuracy compared to Scenario 2 due to an imbalance favoring the normal and diabetes classes. In Scenario 1, accuracy did not represent the overall model behavior accurately. This understanding becomes evident when examining other metrics. Despite the overall model accuracy decreasing after the t-SNE operation, other metrics increased, enabling better recognition of previously dominated classes as the data for these classes reduced. A comparison of Scenario 2 and Scenario 3 reveals that the addition of synthetic data to the data obtained after the t-SNE process significantly improves the model performance. As a result, the findings in the table demonstrate that the combined use of t-SNE and synthetic data augmentation can significantly enhance the classification performance for a 7-class classification problem.

Table 4 7 class classification results with synthetic data augmentation and t-SNE-based images selection for different scenarios. sc1: before t-SNE real data, sc2: after t-SNE real data, sc3: after t-SNE real data + x1 synthetic data

Metrics	sc1	sc2	sc3
accuracy	65%	63.4%	**65.2%**
F1-Score	52.4%	60.6%	**63%**
precision	60.2%	62%	**63.6%**
recall	50.8%	61.4%	**63.6%**

In the enhanced dataset, improvements were noticeable across all classes as can be seen in Fig. 13. Notably, detection rates in the previously undetected hypertension class rose to 16%. This trend of increased detection accuracy was consistent in every class.

3.8 Analyzing the limitations of synthetic data augmentation: the saturation point and its impact on model accuracy

Adding synthetic data to a dataset often improves model performance by introducing more diversity and information, aiding the model in better generalization. However, the benefits of adding synthetic data diminish beyond a certain point, leading to a saturation effect. This occurs as the model, having already captured most of the relevant patterns and information from the real data, gains little from additional synthetic data. Moreover, excessive synthetic data, being model-generated and based on limited initial data, may start deviating from real-world representations. In some cases, excessive synthetic data can lead to overfitting. The model becomes too tailored to the synthetic data and loses its ability to generalize to new, unseen real data. This can result in decreased accuracy in real-world scenarios.

The production of synthetic data can be likened to photocopying. Each copy introduces slight distortions, which can accumulate over time. Similarly, adding synthetic data incrementally enhances performance up to a point. Beyond this threshold, the distortions become detrimental, overshadowing the characteristics of the real data and leading the model to prioritize less representative features. This occurs because, as previously mentioned, the data density undergoes distortion, causing the inherent characteristics of the real data to lose prominence. Consequently, the model begins to emphasize features that resemble more of a *"photocopy"*. While the synthetic data, akin to these photocopies, possesses similar traits and lends support to the model, an excessive presence of synthetic data leads to a reduction in the significance of the real data's features. To determine the optimal balance between real and synthetic data, and identify the saturation point, we conducted an ablation study. Table 5 presents this study's findings, including the best ratio of real to synthetic data and the procedure for incrementally adding synthetic data until saturation is reached.

The study found that the 'c5' case, involving doubling the synthetic data, yielded the best performance. Beyond this point, the model showed signs of saturation and performance began to decrease.

3.9 Limitations of the study and future perspectives

Our approach offers several key advantages that distinguish it from existing methodologies in the field. First, we focus on targeted data generation specifically aimed at ocular diseases, utilizing advanced models like Stable Diffusion to create high-quality synthetic data that closely resembles

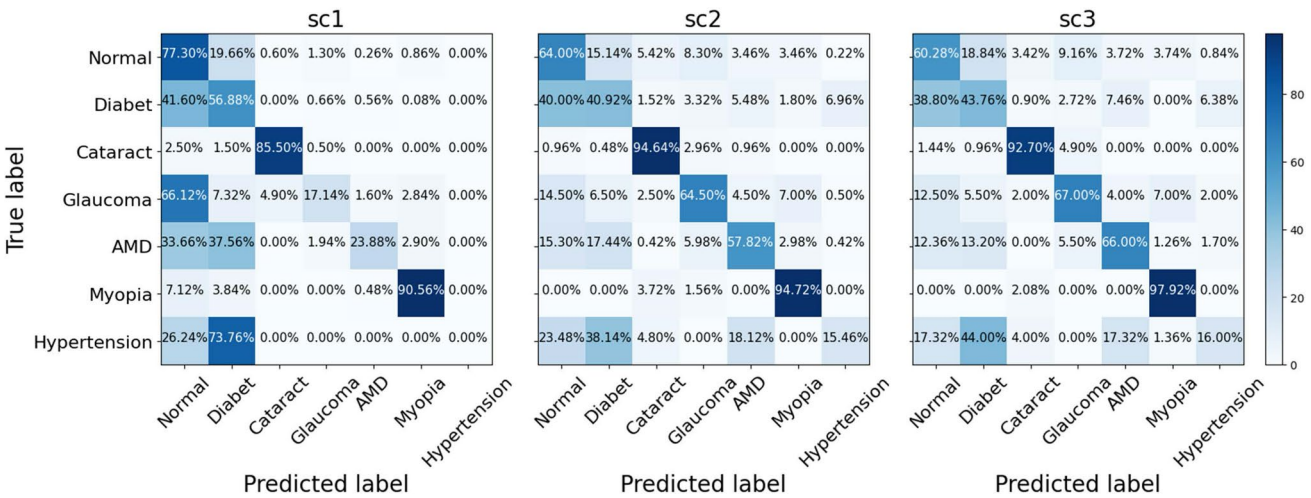


Fig. 13 Confusion matrices of 7 class with t-SNE and synthetic data augmentation classification. sc1: before t-SNE real data, sc2: after t-SNE real data, sc3: after t-SNE real data + x1 synthetic data

Table 5 5 class classification results with synthetic data augmentation for different scenarios. sc1: real data, sc2: real data + x0.5 synthetic data, sc3: real data + x1 synthetic data, sc4: real data + x1.5 synthetic data, sc5: real data + x2 synthetic data, sc6: real data + x4 synthetic data

metrics	sc1	sc2	sc3	sc4	sc5	sc6
accuracy	80.8%	82%	83.6%	84.4%	**86%**	84.2%
F1-Score	79.2%	79.8%	82%	83%	**85.2%**	82.8%
precision	79.4%	79.8%	82.6%	83%	**85%**	83.2%
recall	79.4%	80.8%	82.2%	83.6%	**86%**	83.4%
glaucoma	74.42%	74.42%	77.2%	79.98%	78.86%	**80%**
cataract	96.12%	97.4%	**98.7%**	97.82%	96.1%	95.26%
AMD	64.62%	66.66%	69.22%	69.76%	**75.9%**	71.82%
hypertension	66%	68%	72%	76%	**82%**	75%
myopia	95.16%	**96.98%**	94.54%	93.92%	96.38%	94.52%

real-world scenarios. This capability sets our method apart, as it effectively mitigates class imbalance through synthetic data production, enhancing the model's ability to learn from underrepresented classes. Furthermore, our technique excels in complex datasets, such as the Ocular Disease Intelligent Recognition (ODIR) dataset, where traditional data augmentation methods often fall short. Unlike many existing approaches, we prioritize the preservation of critical details, such as vascular structures, which are essential for accurate medical diagnoses. The resulting improvements in precision and recall demonstrate the efficacy of our method, further complemented by innovative techniques like prompt engineering and fine-tuning that optimize model performance. In addition to the advantages provided by our methodology, future work could focus on investigating the effects of the "other" class and developing more effective data augmentation strategies for this category. For instance, defining more detailed subclasses or using advanced generative models could provide better representations for this class. Additionally, exploring alternative CNN architectures and multi-labeling strategies could enhance the labeling process, further improving the overall performance of the model. Also, obtaining clinical approval for the generated synthetic data may be a crucial step toward validating our methodology's real-world applicability.

3.9.1 Generalizability across classification models

In this study, we primarily focused on demonstrating the effectiveness of synthetic data generation using a diffusion-based model, Stable Diffusion, in enhancing the performance of a selected image classification model for ocular disease diagnosis. Our goal was to show how high-quality synthetic data could mitigate data imbalance and improve model accuracy, particularly for underrepresented classes.

However, as noted, validating the generalizability of our approach across a broader range of classification models is an important next step. We acknowledge that relying on a single model does not fully capture the potential of our data augmentation technique. In the future, various convolutional neural network (CNN) architectures could be applied, including more advanced models, to further validate the generalizability of synthetic data and the t-SNE technique. This would help assess whether the improvements seen with Stable Diffusion-augmented data in the current study can extend to other classification tasks.

The introduction of multiple classification models in future studies will allow for a more comprehensive assessment of synthetic data generation techniques, establishing broader relevance and utility in various medical image classification tasks. We believe that the foundation laid in this study highlights the potential for generalization, setting the stage for more detailed comparisons in future work.

3.9.2 Inclusion of the "other" class and 8-class experiments

The "other" class was excluded from the entire process because it encompasses a wide range of abnormalities and lacks sufficient distinguishing features to represent multiple diseases simultaneously. This situation made it challenging to create accurate and reliable representations during synthetic data generation. Therefore, the exclusion of the "other" class was implemented to enhance the model's overall performance and achieve clearer classification results.

In the remaining 7-class dataset, diabetes and normal classes dominate compared to the other five classes. This class imbalance could impact the model's learning during the training process; thus, it is essential to utilize synthetic data to improve the performance of the other classes. In the future, investigating the effects of the "other" class and developing more effective data augmentation strategies for this class could be a significant step toward enhancing the model's overall accuracy. For example, defining more detailed subclasses for this class or using advanced generative models could provide better representations.

3.9.3 Enhancing labeling through alternative CNNs and multiple labeling processes

In future studies, enhancing the labeling process could be achieved by exploring the application of alternative CNNs and implementing a multiple labeling strategy. Different CNN architectures, such as ResNet and DenseNet each offer unique strengths in feature extraction and representation. By leveraging these diverse architectures, it may be possible to improve the model's capability to capture a broader range of features and nuances within the dataset. This diversity in modeling could help address the complexity of the ocular diseases present, allowing for more nuanced classifications.

Furthermore, implementing a multiple-labeling strategy could significantly increase the robustness of the labeling process. This approach involves using several models to independently predict the labels for each instance in the dataset. By aggregating these predictions, it could reduce the impact of individual model biases or errors, leading to more reliable and consistent labeling.

Adopting these strategies could enhance the accuracy and reliability of the labeling process, potentially improving overall model performance. Future research may focus on evaluating the effectiveness of these approaches in generating more accurate labels for the dataset.

3.9.4 Clinical approval and real-world applicability

In future work, obtaining clinical approval for the generated synthetic data could be an essential step toward validating the real-world applicability of our methodology. While our current focus was to demonstrate the effectiveness of diffusion-based data augmentation in enhancing model performance, future studies could explore the integration of clinical validation processes to ensure that the synthetic data meets the required medical standards. This would not only strengthen the reliability of our approach but also pave the way for its use in clinical settings, particularly for training diagnostic models. Additionally, establishing collaborations with healthcare professionals and institutions could provide further insights into the clinical relevance of the generated data, ensuring that it aligns with real-world diagnostic requirements.

4 Conclusions

This article has showcased the innovative impact of generative AI, particularly diffusion-based models, in the realm of synthetic data generation. In our research, the Stable Diffusion model was applied to the Ocular Disease Intelligent Recognition (ODIR) dataset, a rich source of ocular health data but highly imbalanced. By fine-tuning and conducting rigorous experiments, we successfully combined synthetic data with real data to mitigate class imbalance and improve data representation for dominant classes. In essence, the objective is to improve diagnostic accuracy in sensitive medical fields, such as ocular disease detection, through advanced synthetic data generation techniques. Our analysis involved a thorough comparison between traditional data augmentation techniques and the output from the Stable Diffusion model. In this comparison, diffusion-based data augmentation showed promising results. It achieved precision rates ranging from 76.2% to 85% and recall values between 75% and 86%, indicating a noteworthy improvement over traditional augmentation methods. The results were notable: the diffusion model consistently showed improved the classification performance over traditional methods, highlighting its capacity to enhance machine-learning model accuracy through advanced data generation. This observation suggests the potential usefulness of diffusion-based models in data augmentation and model performance optimization. The study's results indicate that integrating synthetic data, particularly that generated by diffusion models, can significantly improve classification model performance. After synthetic data addition, notable increases of 3.4% in the precision metric and 12.8% in the recall metric were observed in the 7-class case. This strategy is particularly promising for boosting accuracy and robustness in machine-learning models across various medical fields, including dermatology, pathology, and radiology. Given the rapid advancements in Latent Diffusion-Based Models, further exploration into their applications in data augmentation and classification tasks is both necessary and promising. Our study not only confirms the effectiveness of these models in tackling data scarcity but also paves the way for future research in using synthetic data to enhance healthcare diagnostics and more.

Author Contributions Conceptualization, methodology, implementation, experiments, results analysis, and manuscript writing were performed by BA, DDA, and AG. OD contributed to the methodology of the study and manuscript review. BA and DDA are equally contributed first authors.

Data Availability Ocular Disease Intelligent Recognition (ODIR), used in this study, is publicly available at https://odir2019.grand-challenge.org/.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695
2. Galatolo FA, Cimino MG, Vaglini G (2021) Generating images from caption and vice versa via clip-guided generative latent space search. arXiv preprint arXiv:2102.01645
3. Azizi S, Kornblith S, Saharia C, Norouzi M, Fleet DJ (2023) Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466
4. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T (2022) Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inform Process Syst 35:36479–36494
5. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inform Process Syst 33:6840–6851
6. Watson D, Chan W, Ho J, Norouzi M(2022) Learning fast samplers for diffusion models by differentiating through sample quality. arXiv preprint arXiv:2202.05830
7. Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. Adv Neural Inform Process Syst 34:8780–8794
8. Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR
9. Song J, Meng C, Ermon S (2020) Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502
10. Liu N, Li S, Du Y, Torralba A, Tenenbaum JB (2022) Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision, pp. 423–439. Springer
11. Sinha A, Song J, Meng C, Ermon S (2021) D2c: Diffusion-decoding models for few-shot conditional generation. Adv Neural Inform Process Syst 34:12533–12548
12. Song Y, Durkan C, Murray I, Ermon S (2021) Maximum likelihood training of score-based diffusion models. Adv Neural Inform Process Syst 34:1415–1428
13. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M (2021) Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741
14. Song Y, Ermon S (2020) Improved techniques for training score-based generative models. Adv Neural Inform Processing Syst 33:12438–12448
15. Chung H, Sim B, Ye JC (2022) Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12413–12422
16. Lugmayr A, Danelljan M, Romero A, Yu F, Timofte R, Van Gool L (2022) Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11461–11471
17. Batzolis G, Stanczuk J, Schönlieb C-B, Etmann C. (2021) Conditional image generation with score-based diffusion models. arXiv preprint arXiv:2111.13606
18. Saharia C, Chan W, Chang H, Lee C, Ho J, Salimans T, Fleet D, Norouzi M (2022) Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10
19. Wang T, Zhang T, Zhang B, Ouyang H, Chen D, Chen Q, Wen F (2022) Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952
20. Choi J, Kim S, Jeong Y, Gwon Y, Yoon S (2021) Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938
21. Li B, Xue K, Liu B, Lai Y-K (2022) Vqbb: Image-to-image translation with vector quantized brownian bridge. arXiv preprint arXiv:2205.07680
22. Akrout M, Gyepesi B, Holló P, Poór A, Kincső B, Solis S, Cirone K, Kawahara J, Slade D, Abid L (2023) et al.: Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. arXiv preprint arXiv:2301.04802
23. Ktena I, Wiles O, Albuquerque I, Rebuffi S-A, Tanno R, Roy AG, Azizi S, Belgrave D, Kohli P, Karthikesalingam A (2023) et al.: Generative models improve fairness of medical classifiers under distribution shifts. arXiv preprint arXiv:2304.09218
24. Sagers LW, Diao JA, Melas-Kyriazi L, Groh M, Rajpurkar P, Adamson AS, Rotemberg V, Daneshjou R, Manrai AK (2023) Augmenting medical image classifiers with synthetic data from latent diffusion models. arXiv preprint arXiv:2308.12453
25. Sagers LW, Diao JA, Groh M, Rajpurkar P, Adamson AS, Manrai AK (2022)Improving dermatology classifiers across populations using images generated by large diffusion models. arXiv preprint arXiv:2211.13352
26. Rajotte J-F, Bergen R, Buckeridge DL, El Emam K, Ng R, Strome E (2022) Synthetic data as an enabler for machine learning applications in medicine. Iscience **25**(11)
27. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F (2021) Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 5(6):493–497
28. Fang H, Han B, Zhang S, Zhou S, Hu C, Ye W-M (2024) Data augmentation for object detection via controllable diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1257–1266
29. Feng C-M, Yu K, Liu Y, Khan S, Zuo W (2023) Diverse data augmentation with diffusions for effective test-time prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2704–2714
30. Fu Y, Chen C, Qiao Y, Yu Y (2024) Dreamda: Generative data augmentation with diffusion models. arXiv preprint arXiv:2403.12803
31. Bennett B (2022) Synthetic Image Datasets with Stable Diffusion and Data Augmentation
32. Kebaili A, Lapuyade-Lahorgue J, Ruan S (2023) Deep learning approaches for data augmentation in medical imaging: a review. J Image 9(4):81
33. Smitha A, Jidesh P (2022) Classification of multiple retinal disorders from enhanced fundus images using semi-supervised gan. SN Comput Sci 3(1):59
34. Gobinath C, Gopinath M (2022) Deep classification of fundus images using semi supervised gan. In: 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 1–4 . IEEE
35. Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019). https://odir2019.grand-challenge.org/. Accessed: 2022-02-10 (2019)
36. Mostaque E (2022) Stable diffusion public release. Stability AI
37. Mascarenhas S, Agarwal M (2021) A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In: 2021 International Conference on Disruptive Technologies for Multi-disciplinary Research and Applications (CENTCON), vol. 1, pp. 96–99. IEEE
38. Maaten L, Hinton G (2008) Visualizing data using t-sne. Journal of machine learning research **9**(11)
39. Düzyel O (2023) A comparative study of gan-generated handwriting images and mnist images using t-sne visualization. arXiv preprint arXiv:2305.09786
40. Goceri E (2023) Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intell Rev 56(11):12561–12605

41. Abràmoff MD, Garvin MK, Sonka M (2010) Retinal imaging and image analysis. IEEE Rev Biomed Eng 3:169–208
42. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR
43. Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K (2023) Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510
44. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 . PMLR
45. AI S (2024) Stable Diffusion. https://github.com/Stability-AI/stablediffusion. Accessed: 2024-11-05
46. Wang J, Liu Z, Zhao L, Wu Z, Ma C, Yu S, Dai H, Yang Q, Liu Y, Zhang S (2023) et al.: Review of large vision models and visual prompt engineering. arXiv preprint arXiv:2307.00855
47. Oppenlaender J (2022) Prompt engineering for text-based generative art. arXiv preprint arXiv:2204.13988
48. Witteveen S, Andrews M (2022) Investigating prompt engineering in diffusion models. arXiv preprint arXiv:2211.15462
49. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
50. Bera S, Shrivastava VK (2020) Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification. Int J Remote Sens 41(7):2664–2683
51. Goodfellow I (2016) Deep learning. MIT press
52. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inform Process Manag 45(4):427–437. https://doi.org/10.1016/j.ipm.2009.03.002
53. Powers DM (2020) Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061
54. Mumuni A, Mumuni F (2022) Data augmentation: A comprehensive survey of modern approaches. Array, 100258
55. Mikołajczyk A, Grochowski M (2018) Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW), pp. 117–122 . IEEE