MEDICRAFT

MEDICAL SYNTHETIC DATASET GENERATOR

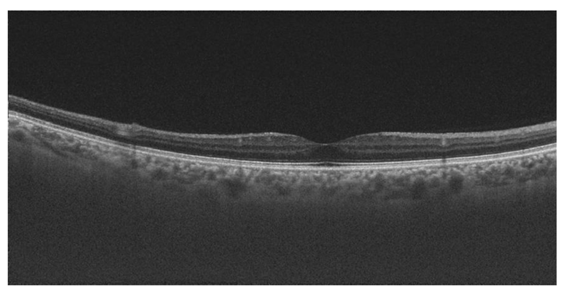
UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Matematyki i Informatyki

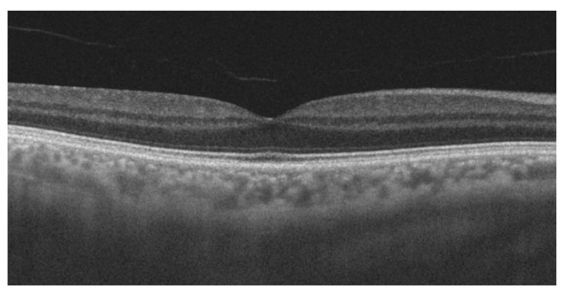
INTRODUCTION

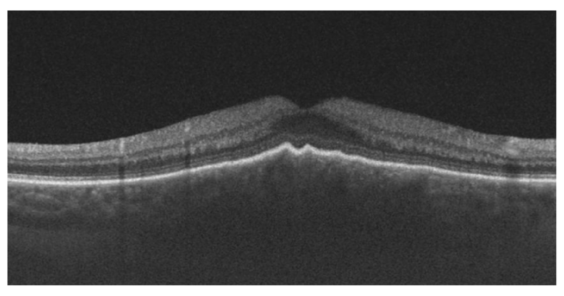
The medical imaging analysis using machine learning encounters challenges such as limited and imbalanced datasets, often constrained by privacy concerns related to patient information. This innovative project addresses these limitations by developing software capable of generating synthetic and diverse medical datasets from imaging information. Notably, this tool utilizes OCT eye scans, featuring images with abnormalities like tumors and melanomas. By mitigating the scarcity of real-world data, the solution facilitates improved research and education in machine learning for medical image analysis and classification.

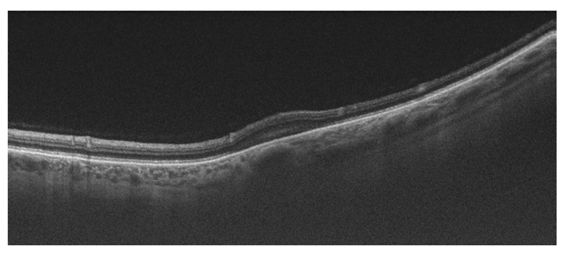
LEARNING DATA

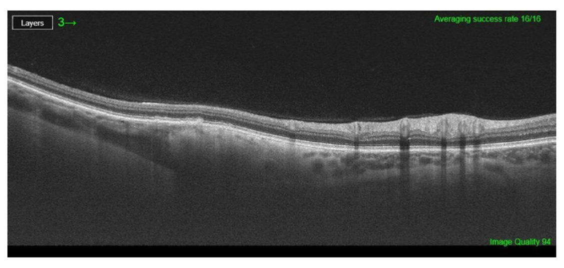
The dataset was collected in cooperation with University Clinical Hospital in Poznan. It consists of OCT scans of the eye fundus taken in 12 planes every 30° for one examination. The collection has been annotated and divided according to the medical condition diagnosed, including i.a.: healthy reference eyes, with benign precancerous lesions, with serious lesions in the form of melanomas.

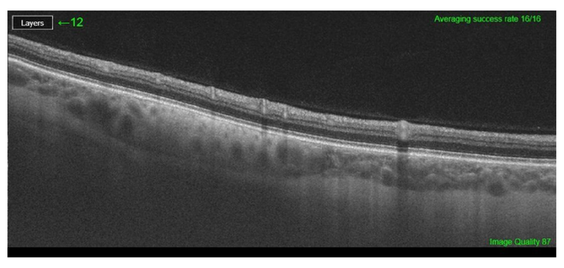

healthy reference eye fundus


benign precancerous lesions


foveolar scan at the margin of choroidal melanoma


precancerous lesions with fluid


cancerous lesions transformation


cancerous lesions transformation

OBJECTIVE

The goal is to support research and education in machine learning for medical image analysis and classification for the diagnosis of diseases e.g tumours.

Creation of an end-to-end Python library multiplying the dataset .

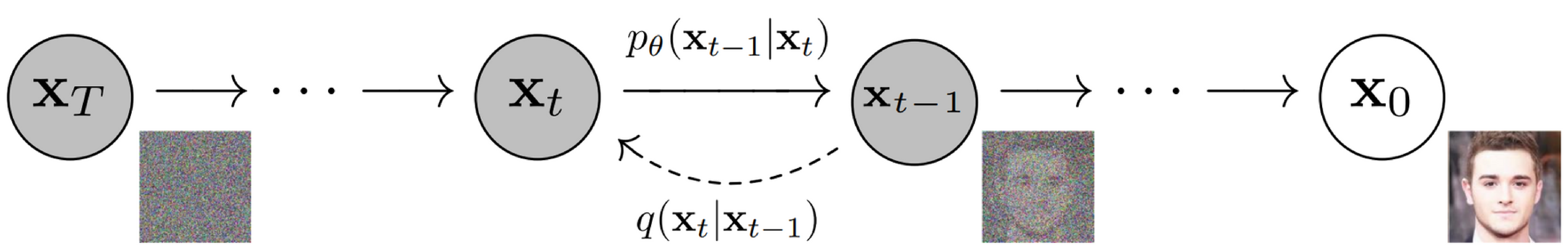
METHODOLOGY

Manually collected OCT eye scans from University Clinical Hospital in Poznan are used in data-driven learning using convolutional neural networks.

Instead of GANs, a state-of-the-art diffusion model solution is used to generate synthetic datasets.

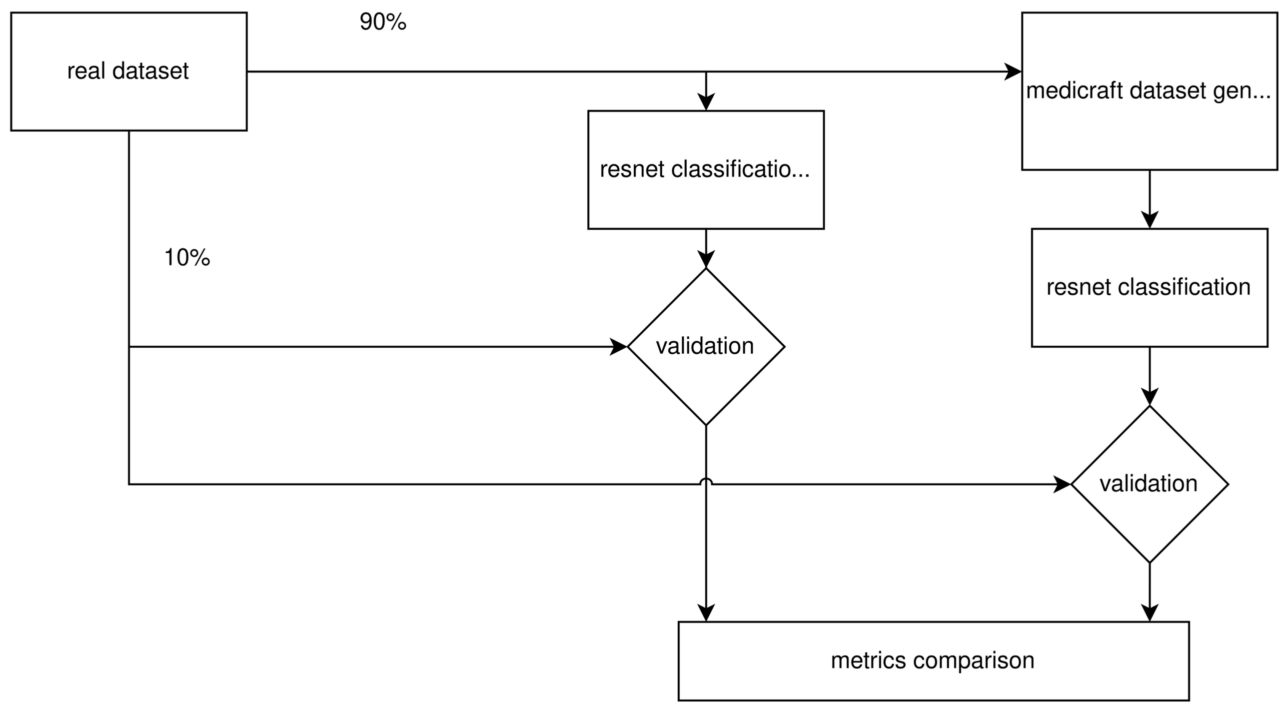
DIFFUSION MODEL

The Denoising Diffusion Probabilistic Model (DDPM) is a powerful probabilistic generative model inspired by nonequilibrium thermodynamics, achieving high-quality image synthesis through a diffusion process. With a novel weighted variational bound and progressive lossy decompression, DDPM demonstrates state-of-the-art performance and effectively models complex data distributions, providing a diverse solution for tasks demanding noise reduction and high-quality data synthesis across diverse domains.



ARCHITECTURE

The solution utilizes a sophisticated architecture for interpreting input and generate data collaboratively with a medical team, refining the learning process iteratively, making the model to generate different disease changes on images. The evaluation of the quality of the generator will include the use of the classifier model, the comparison of its indicators on both real and synthetic data to assess performance and the quality of the dataset.



CLASSIFIER RESULTS

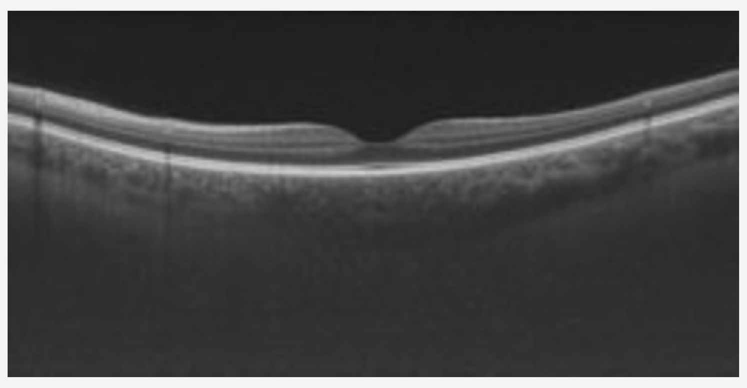
To prove a concept an effectiveness of overall architecture, the binary classifiers were trained on balanced real and synthetic datasets and then validated. The results are similar, as follows:

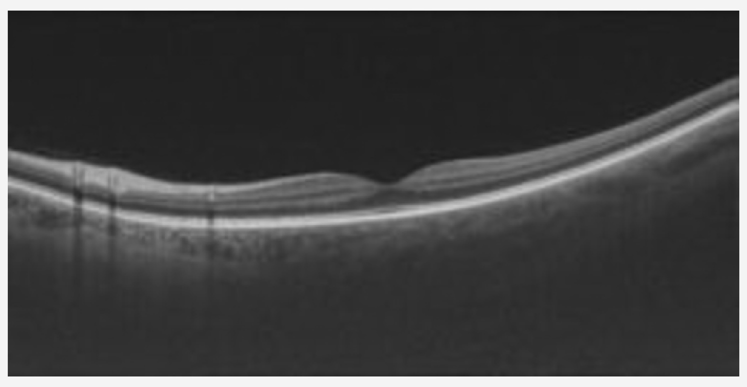
	Real data	**Synthetic data**
Accuracy	0.93	0.88

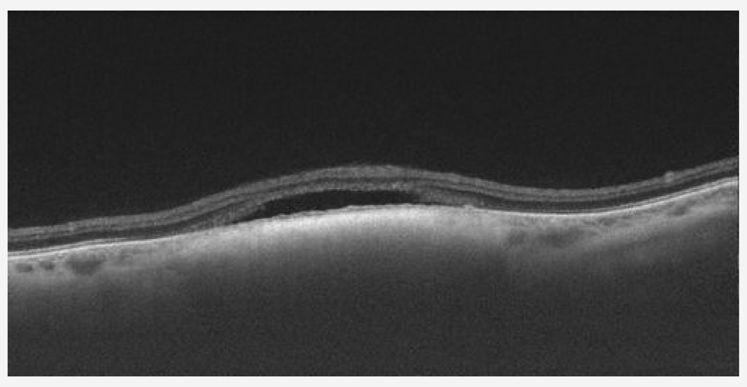
GENERATED IMAGES AND ANOMALIES

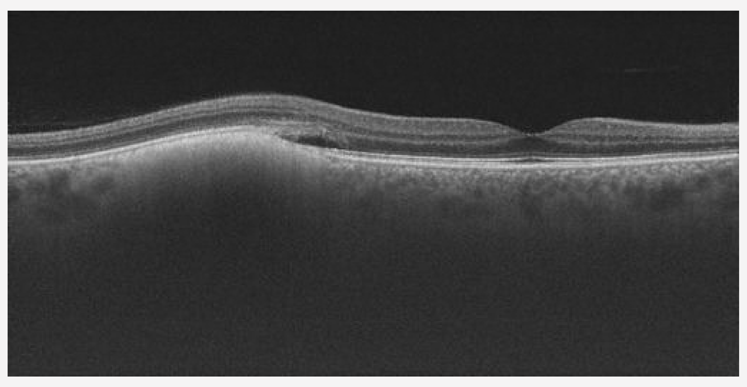
Below are correctly generated images and errors which were observed in the evaluation process.
After the first iteration of the model, the medical team found 66% anatomical correctness of the generated images for one of generated classes.

CORRECTLY GENERATED

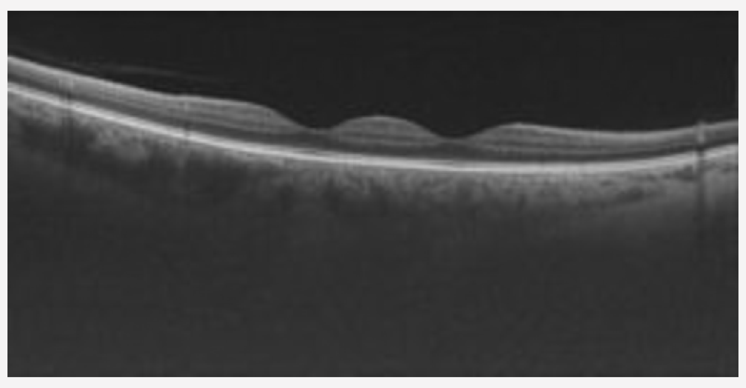

healthy eye fundus

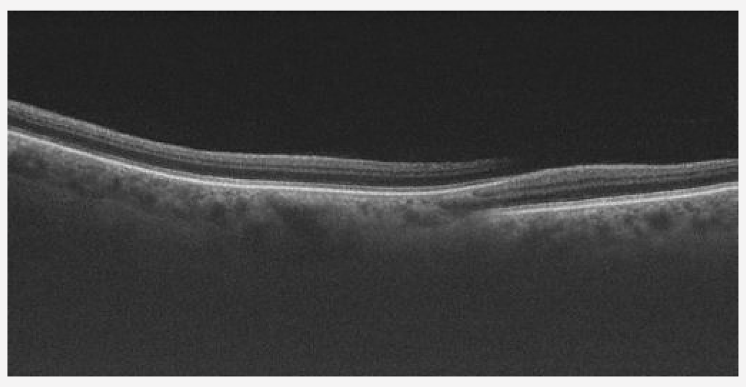

healthy eye fundus

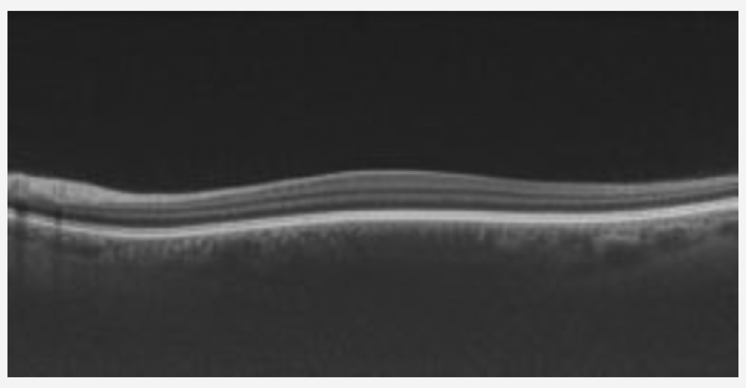

precancerous lesions with fluid

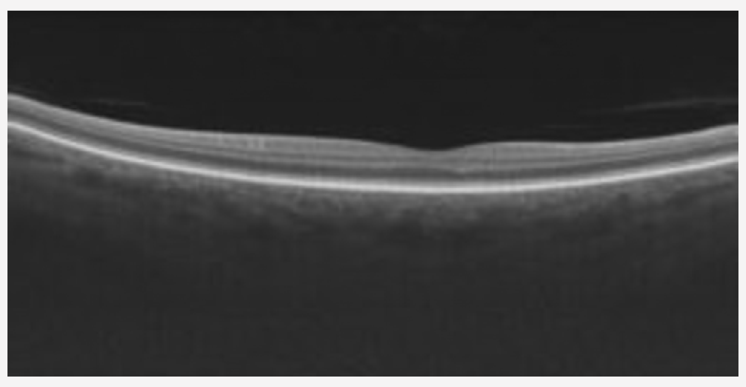

benign precancerous lesions

INCORRECTLY GENERATED


artificially doubled fovea


totally corrupted image


absence of fovea


hypoplastic fovea

fp.patyk@gmail.com
https://github.com/drfifonz/medicraft
https://www.linkedin.com/in/filip-patyk/