RADBOUD UNIVERSITY NIJMEGEN

FACULTY OF SOCIAL SCIENCES

Real and synthetic multimodal retinal images for the diagnosis of Alzheimer's Disease with convolutional neural networks

THESIS MSC ARTIFICIAL INTELLIGENCE
NEUROTECHNOLOGY AND HEALTHCARE

Supervisor:
dr. ir. Patrick J. GONZÁLEZ

Author:
Ivan R. SLOOTWEG

Internal Supervisor:
dr. Leila BAGHERIYE

Second reader:
prof. dr. Johan H.P. KWISTHOUT

June 2024

I. Abstract

Abnormalities in the retinal nerve fiber layers and blood vessels are correlated with Alzhemer's Disease (AD) and can be identied with non-invasive retinal imaging. We show that a denoising diffusion probabilistic model (DDPM) can generate realistic and unique synthetic retinal images. Unimodal convolutional neural networks (CNNs) for predicting Positron Emission Tomography (AmyloidPET) biomarker status were pretrained on synthetic data and finetuned on real data or trained solely on real data. Multimodal classifiers combined unimodal CNN predictions with patient metadata. Our method for generating and leveraging synthetic data has the potential to improve AmyloidPET prediction. Our best unimodal and multimodal classifiers were not pretrained on synthetic data, however pretraining with synthetic data slightly improved classification performance for two out of the four modalities. and integration of metadata in multimodal CNNs achieved best results. Class activation maps show that CNNs for predicting AmyloidPET status can learn features at clinically relevant structures in the retina.

II. Introduction

A. Ocular pathologies in neurological diseases

Alzheimer's Disease (AD), a progressive neurodegenerative disease that begins many years before symptoms are present, poses an important public healthcare concern due to aging populations worldwide [1]. Current clinical diagnosis relies on detection of decreased amyloid-beta levels ($A\beta$), increased total tau and phosphorylated tau (pTau) levels in the cerebrospinal fluid (CSF) as well as Amyloid Positron Emission Tomography (AmyloidPET) and Tau-PET scanning. These techniques, however, are expensive and invasive and are hard to translate to community based screening for early onset of AD.

The retina, sharing an ontogenetic association and relationship with the brain, is hypothesized to provide a window to the brain. Several studies have associated retinal imaging parameters with AD. [2]–[6]. These studies revealed abnormalities in the retinal nerve fibre layer, blood vessels and the optic nerve that resemble changes in the brain of patients with AD. Modern retinal imaging techniques, such as Optical Coherence Tomography (OCT), OCT Angiography (OCT-A) and Fundus Scanning Laser Ophthalmoscopy photography (FSLO) allow for non-invase imaging of these features. These techniques are widely available which opens up the possibility of community based screening.

Advances in artificial intelligence (AI) have allowed for the development of parameter-based and image-based classification models for the prediction of AD using medical images. Most of these developments focus on brain MRI and only few studies focus on retinal imaging-based classification [7], [8]. As an example, Wang, Jiao, Liu, *et al.* compared machine learning algorithms such as linear regression and random forests to predict clinical AD diagnosis based on OCT parameters that were manually derived from automatic segmentations (Area Under the Receiver Operator Curve (AUROC) = 0.91) [9]. A drawback of this type of approach is that the inputs rely on the selection and extraction of handcrafted features which can introduce information loss. Moreover, these studies did not make use of multimodal imaging input.

Identification of AmyloidPET status from multimodal retinal images could allow for early identification of patients at risk of AD. To date several studies have attempted at training convolutional neural networks (CNNs) for identifying and combining relevant features of multimodal retinal images. Wisely, Wang, Henao, *et al.* showed that adding retinal imaging (AUROC = 0.836) to an existing set of inputs comprising ganglion cell-inner plexiform layer maps, quantitative data and patient data (AUROC = 0.841) does not improve performance of AD detection [10]. In a follow-up paper, distinction between mild cognitive impairment and normal cognition based on only OCT-A images (AUROC = 0.625) underperformed to classification on image-derived quantitative data (AUROC = 0.960) [11]. These results indicate that it can be difficult to extract meaningful features from retinal imaging, especially in a small dataset (284 eyes of 159 subjects). A multi-center study with 3888 subjects with optic nerve head (ONH)-centered and macula-centered fundus photographs predicted AD-dementia on both eyes (AUROC = 0.93 on internal test set, AUROC = $0,73$ - 0.91 on external test set) and unilaterally (AUROC = 0.93 on internal test set, AUROC = 0.62 - 0.84 on external test set) [12]. This study also predicted Amyloid PET status on both eyes (AUROC = 0.68 - 0.86 on external test set) and unilaterally (AUROC = 0.61 - 0.83 on external test set).

CNN's allow for multimodal retinal imaging input, however they rely on the availability of large image datasets, which is not always the case in the medical imaging domain. To overcome the limited availability of medical imaging data, generative artificial intelligence (AI) can provide a solution by synthesising image data and extending the dataset. Several approaches to generative AI exist, such as Generative Adverserial Networks (GANs) and Diffusion Probabilistic Models (DDPMs). GANs consist of two competing models, a generator to produce fake images and a discriminator that distinguishes between real and fake images. In this process the generator is secluded from the real images, so there is little risk of the generator producing copies of real images. Through iterative learning, the generator's output becomes more realistic. GANs have found numerous applications in medical image synthesis. For instance, one study employed six different GAN architectures for training several segmentation networks on MRI, CT and fundus image datasets [13]. The study revealed potential issues such as overfitting when training on synthesized data from a small original dataset and biased training data when synthesizing data from an

unbalanced dataset.

Another approach to data synthesis involves DDPMs (Figure 1) [14], [15]. DDPMs have gained significant popularity in the field of medical imaging and have been successfuly applied to image synthesis for MRI, CT hystopathology and dermathology [16]–[20]. They demonstrate superior output diversity compared to GANs and Variational Auto Encoders [16]. However, unlike GANS, which prioritize image fidelity over diversity and seclude the training images from the generator, DDPMs are more prone to memorisation [17], [21]. Peng, Adeli, Zhao, *et al.* developed a conditional DDPM that produces a subset of brain MRI slices at random locations conditioned on noise or on another subset of MRI of the same MRI [20]. This architecture, which incorporates a mask to encode the locations of the target and condition slices, outperformed five baseline GAN methods in terms of resemblance to the training data, diversity and capturing the distribution of the training data [20].

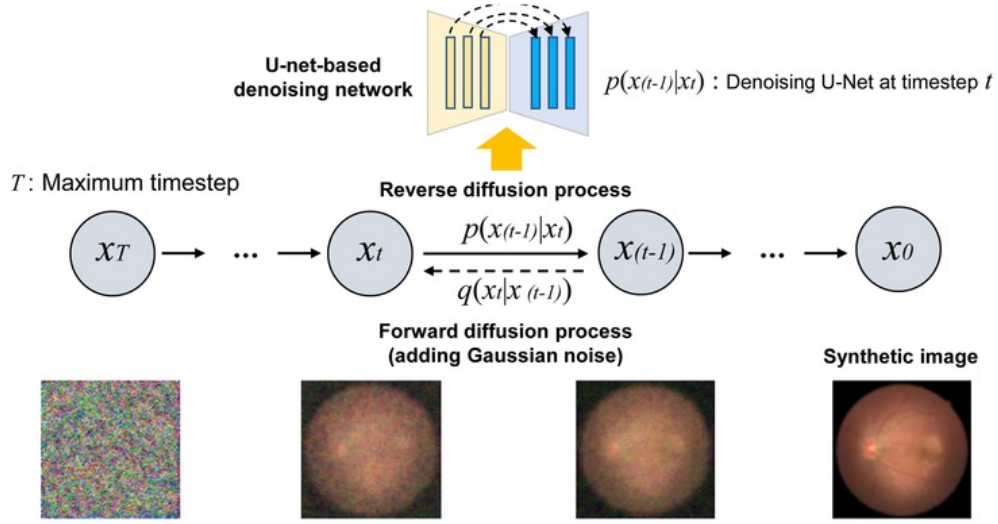


Figure 1: Forward and backward process in denoising diffusion models. These models initially map an input image to a noise image by gradually adding Gaussian noise in many small steps and subsequently learning to perform the reverse process in small steps as well [14]. A popular implementation of DDPMs involves using a neural network, typically with Unet-shaped architecture, ResNet layers, time embeddings and attention layers. The U-Net is trained to predict the noise to be removed from an input image given the noisy input and given the step in the dffusin process chain [14]. A trained DDPM can then generate synthetic images by iteratively removing noise from a pure Gaussian noisy input. Class-conditioning the DDPM allows one model to generate images with varying content, such as of a specific animal, or medical images belonging to a specific diagnosis. Image modified from Kim, Ryu, Choi, *et al.* [22].

In the present study we will examine the possibility to predict AmyloidPET status with multimodal retinal imaging and to improve the performance through pretraining with synthetic data. We develop a DDPM to generate synthetic images for four types of retinal scans and we develop a filter to recognize realistic synthetic images. Lastly, we train a multimodal classifier with heterogenous fusion of unimodal predictions and patient information. The design of this framework serves as proof of concept for leveraging generative AI in medical image classification tasks.

III. Materials and Methods

A. Participants

For this study we considered data from 183 patients from two retrospective cohorts with confirmed AmyloidPET status and opthalmic examinations including FSLO, conventional OCT and OCT-A. We included 328 eyes of 167 subjects: 116 eyes of 59 AmyloidPET+ subjects and 212 eyes of 108 AmyloidPET- subjects. We excluded patients without AmyloiPET status (N=6) and without good quality retinal scan recorded (N=11). 203 eyes are part of the PreclinAD cohort which is an extension of the Amsterdam sub-study of the European Medical Information Framework for AD (EMIF-AD) from the Amsterdam UMC, location VUmc [23], [24]. The PreclinAD cohort consist of cognitively healthy participants (monozygote twins) aged $\geq$ 60 from the Netherlands Twin Register who all underwent opthalmic evaluation [23], [25]–[27]. The remaining 125 eyes are part of an ongoing trial of the Amsterdam Alzheimer Center (METC 2019.623) which includes participants aged $\geq$ 48 years. Both studies followed the Tenets of the Declaration of Helsinki and the Medical Ethics Committee of the VU University Medical Center in Amsterdam approved the studies. All participants signed an informed consent and underwent a screenings protocol that included in short: ophthalmological examinations, (medical) history check-up, Mini Mental State Examination test and neuropsychological evaluations. Subjects with ischemic stroke, neurodegenerative disorders or systemic chronic conditions (i.e. Parkinson's disease (PD), Diabetes Mellitus (DM), multiple sclerosis) were excluded.

B. Datasets

The complete dataset derved from the examination of these patients contained in total 30 different modalities extracted from FSLO, OCT and OCT-A examinations. All scans were evaluated on image quality and availability of AmyloidPET status. Image quality for FSLO exams was reviewed based on movement or optical distortions such as lash or eyelid coverage and a straight eye gaze was required. OCT-A images were evaluated by focus and resolution and 3D OCT scans were evaluated on the same criteria. Four modalities from three scanner types covering the macula, ONH and fundus were selected for image synthesis and as input to the classification networks: 1) 2D OCT-A focusing on the superficial retinal layers of the macula (OCTA-SMAC); 2D OCT B-Scan (OCT-B) of 2) the ONH (OCT-BONH) and 3) macula (OCT-BMAC); 4) 2D FSLO auto fluorescence (FAF). These modalities were selected after the development of generative models for a larger set of eight modalities, which is described in the Appendix. For training the filter, images of 26 2D modalities were used, excluding 3D OCT of the macula and ONH for the volumetric nature of the data and excluding widefield FSLO recordings for the small dataset size of 74 auto-fluorescence and 76 red-green images. Furthermore, images without known AmyloidPET status were included for training the filter. Some eyes had duplicate FAF images with good image quality. In this case, one image per eye per modality was selected for classifier development. Duplicate images could be used for filter development and generative model development. This led to the following overlapping collections for distinctive tasks:

1) D_{synth} for development of generative models
2) D_{filter} for development of a modality recognition classifier
3) D_{uni} for development of unimodal classifiers
4) D_{multi} for development of multimodal classifiers

Of the 328 eyes in this dataset, 198 eyes were included in D_{multi}, 326 in D_{uni} and 328 in D_{synth} and D_{filter} (Table 1). The first split between testing, training and validation was created for D_{multi} which includes eyes with known AmyloidPET status and images for all selected modalities. These splits were supplemented to create the other collections. newline
We aimed to keep the proportion of AmyloidPET positives (AmyloidPET+) balanced while assigning 20% of the eyes to the test set and using 20% of the eyes in a collection for validation. Splits for these sets were performed at a family level to prevent information leakage as we dealt with monozygote twin pairs in our dataset, whose retinal images may hold strong resemblance.

Subset	[N] eyes (All)	[N] Images (All)	[N] eyes (Dev)	[N] images (Dev)	[%] AmyloidPET+ (Dev)	[N] eyes (Test)	[N] images (Test)	[%] AmyloidPET+ (Test)
Multimodal classification	198	198	160	160	0.394	38	38	0.368
Unimodal classification	326	1080	285	869		83	211	
- OCTA-SMAC	276	276	223	223	0.4	53	53	0.396
- OCT-BONH	278	278	220	220	0.429	58	58	0.379
- OCT-BMAC	276	276	220	220	0.426	56	56	0.411
- FAF	250	250	206	206	0.354	44	44	0.364
Synthesis	328	1102						
- OCTA-SMAC	276	276	223	223	0.386	53	53	0.396
- OCT-BONH	278	278	220	220	0.377	58	58	0.379
- OCT-BMAC	276	276	220	220	0.382	56	56	0.411
- FAF	254	272	210	225	0.351	44	47	0.34

Table 1: Description of datasets used for training + validation (development, Dev) and testing (Test). Making these splits on the family level put a restriction on the distribution of AmyloidPET labels, nonetheless the proportions between the development and test sets differ at most 5%. Images without known AmyloidPET status were excluded from the classification set. The first split between testing, training and validation was created for D_{multi} with 198 eyes with known AmyloidPET status and available images for all selected modalities. The remaining 128 eyes with images for at least one of the selected modalities and known AmyloidPET status were distributed over these splits to create the D_{uni} collection. This totals to 1080 images used for unimocal classification. 22 images with known AmyloidPET status from duplicate FAF recordings were added to these splits to create D_{synth}. The train, validation and test splits of the collections were created in this way to prevent imformation leakage. For example, the eyes of the training split in the unimodal classification collection are part of the same split in all the other collections. The collection used for synthesis is a superset of the unimodal classification collection which in turn is a superset of the multimodal classification collection.

C. Models

The pipeline for data synthesis and for training the classification networks are depicted in Figure 2. Our approach involves the generation of a synthetic image dataset, development of a filter to ensure high-quality synthetic images, and training unimodal and multimodal classifiers to predict AmyloidPET status. Details of the network architectures and hyperparameters governing the training trajectories are provided in the Appendix. Hyperparameters were optimized by hyperparameter optimisation with 250 trials per model. Each trial performed training with one set of hyperparameters.

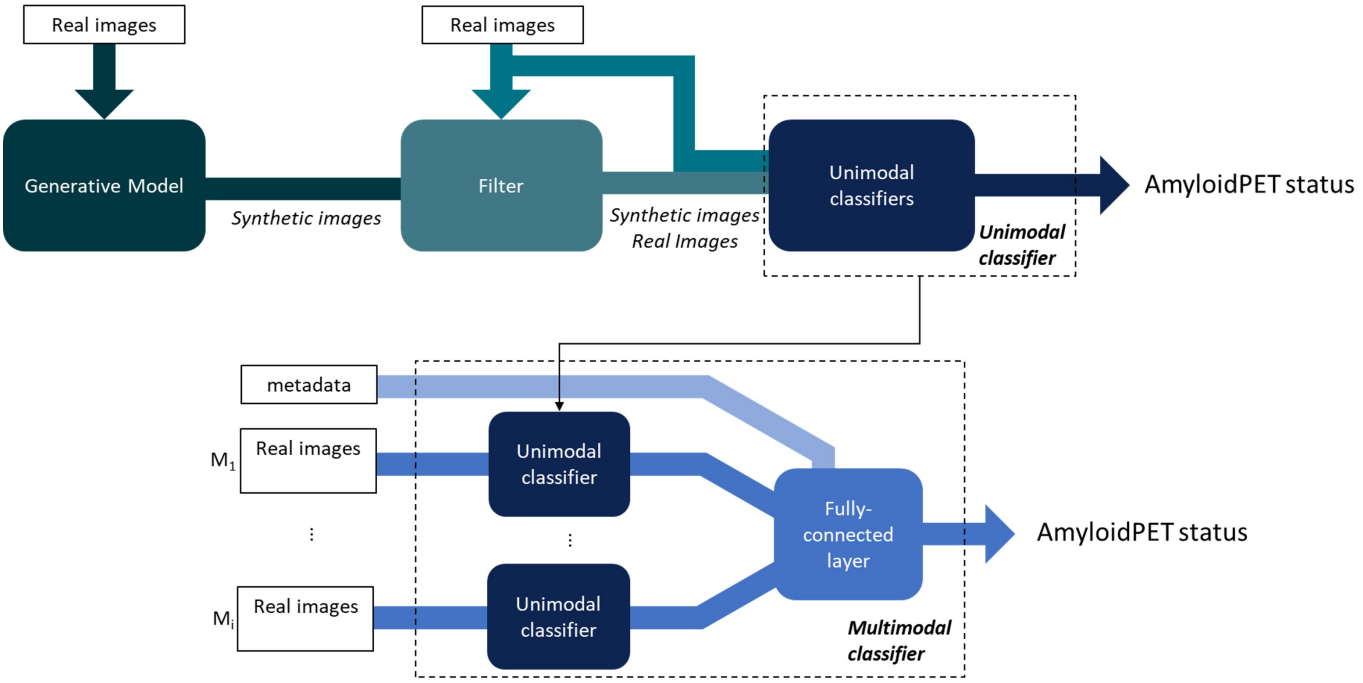


Figure 2: Illustration of the pipeline. (Top): synthetic images are created by a DDPM. The synthetic images for which the filter can recognize the modality were added to the training budget of 1000 synthetic images per class. Both synthetic and real images are used to train unimodal classifiers for each modality separately. We create 'baseline' unimodal classifiers trained on real images, and unimodal classifiers pretrained on synthetic images and finetuned on real images. Unimodal classifiers are not trained with metadata inputs because synthetic data, which has no associated metadata, will be used for pretraining, (Bottom): We perform separate experiments with the baseline and pretrained unimodal classifiers. Unimodal classifiers generate predictions for the four different recordings of one eye. If metadata was used as inputs, the unimodal predictions would be incorporated with age (binary) and gender (scaled by 0.01) metadata in a multi-layer fully connected network. Output of the model is a score between 0 and 1 for the probability of AmyloidPET negative status.

1) Synthetic image generation: We used a DDPM for the generation of the synthetic image dataset. Different configurations for attention levels, filter channels, residual connections and architecture variations such as latent diffusion networks and ControlNets were tested. The final diffusion model was selected based on the quality of the synthetics images and the required training time. The selected model to produce images corresponding tot specific AmyloidPET status is a conditional U-Net DDPM which performs conditioning on the timestep embedding that is used for predicting the noise that is to be removed. This conditioning is achieved through addition of the class label embedding vector to the time step embedding vector.

2) Filter: We assumed that for images with large artifacts or deformations it is more difficult to recognize from what modality the image is generated. For this reason a CNN was trained to recognize the modality of images and then applied during sampling of the DDPM to detect synthetic images with large artifacts or deformations. those images for which the modality could not be correctly recognized were discarded. The performance of the filter is with Matthew's Correlation Coefficient (MCC) for a balanced evaluation of the filter's overall performance by comparing the outputs, between 0 and 1 for each class, and the ground truth labels [28].

3) Classification: Unimodal classifiers were developed for all modalities to predict AmyloidPET status. Resnet and the Efficientnet family architectures were explored in combination with and without pre-trained ImageNet weights as starting points. The Efficientnet-B0 backbone without pretrained ImageNet weights showed best loss reduction in small (25 trials) hyperparameter-optimisation experiments. To leverage the synthetic data, unimodal classifiers were pre-trained with a training budget of 2000 synthetic images (1000 per class) per modality and were subsequently fine-tuned on real data. This totals to three hyperparameter-optimisation experiments per modality: 1) training on real images only, 2) pretraining on synthetic images, and 3) pretraining on synthetic images and subsequent finetuning on real images. In addition to modality-specific classifiers we also experimented with training one shared modality-aware classifier with feature-wise linear modulation (FiLM)[29]. This approach is further discussed in the Appendix.

The multimodal classifiers were trained to take as input the unimodal networks' outputs for each of the four different image modalities of an eye, in some cases together with age (binary) and gender (scaled by 0.01) metadata, to predict AmyloidPET status. Initial experiments were performed to explore methods for integrating feature vectors produced by the unimodal classifiers and to incorporate metadata through either convolutions or fully connected (FC) layers. We found that fusion with convolutions

and FC layers with too many layers were prone to overfitting. We selected a multi-layer FC network with three layers to perform late heterogenous fusion of the unimodal predictions with metadata into one probability prediction. We also compared multimodal classifiers with and without metadata. The multimodal classifier was trained to predict for individual eyes, and only the fully connected layer of the multimodal classifier was trained, the weights of the unimodal classifiers were frozen. All classification models were trained with early stopping: the training process was terminated when evaluation on the validation st showed no improvement. All AmyloidPET classifier were trained with Focal Loss to emphasise on hard, misclassified examples [30]. The models were trained to output the probability for AmyloidPET- status. The performance of the classifiers were evaluated on the validation and test sets, reported by AUROC, area under the precision recall curve (AUPR), sensitivity, specificity and F1-score as harmonic mean between sensitivity and precision. As the model was trained for outputting the probability for AmyloidPET-, these metrics are calculated with inverse ground truth labels and predictions. We calculate the predictions for AmyloidPET+ as $1 - outputs$ and the groundtruth labels for AmyloidPET as $1 - groundtruth$, this way the sensitivity and specificity reflect the model's performance with respect to detecting AmyloidPET+ cases.

D. Synthetic data evaluation

There exists a potential for memorisation of the training images in a dataset with DDPMs. This is for the reason that the generator learns to replicate training data during the learning process and is not secluded from the real images [17]. Therefore, we evaluated the diversity and uniqueness of generated images with intra- and inter-set correlations using a random sample of 200 synthetic images and all real images [17]. The maximum correlation score (Pearson's correlation coefficient) of one image from a set A with all images in a different set B expresses how similar this image is at most to set B. This can be used to evaluate memorisation of the training images in the synthetic dataset, in the case of memorisation the distribution of maximum correlation values of all synthetic images with the set of real images would be high. The distribution of maximum correlation scores that results from comparing all images within the same set informs the diversity of this set. Therefore, maximum correlation values among real images and among synthetic images were compared to evaluate the diversity of the synthetic dataset. Wasserstein Distance (WD) was used to express the difference between two distributions of correlation values and the Kolmogorov–Smirnov (KS) test informed the significance of such differences.

E. Class activation maps

With class activation maps (CAMs) we attempt at providing explanation to the black-box CNN predictions. CAMs indicate which discriminative image regions contribute to a model's output value [31]. We implemented Gradient-weighted CAM (Grad CAM) with the library `pytorch-grad-cam` to produce heatmaps that give insight into the model's activations after the last convolutional layer [32]. Since the activity of convolutional layers often maps spatially to the input, we can upsample the GradCAM attributions to mask the input. GradCAM applied to our binary classification network computes the gradient of the binary output layer with respect to each of the network's activations at the selected convolutional layer. To produce the heatmap, the gradient at the output layer is computed and subsequently multiplied with the layer's activations. As our model outputs the probability for AmyloidPET-, the resulting heatmap displays the image regions that, according to the model, contribute to an output of a higher probability of AmyloidPET-. This is because our model has only 1 output layer for binary predictions, so the gradCAMs can only be calculated with respect to this one output node. If the output layer had been a 2-node layer we could have displayed heatmaps with respect to AmyloidPET+ predictions by calculating the gradient with respect to the corresponding outpu node.

IV. Results

A. Synthetic data

Several examples of synthetic images are shown in Figure S1 in the Appendix. Figure 3 displays, for each modality, the three synthetic images with the highest correlation with any real image and the corresponding real image. Visual inspection of these examples illustrates that Pearson's correlation coefficient of the pixel values is in good agreement with the visual similarity of retinal images in our dataset. Figure 4 displays distributions of the maximum correlation for synthetic images with all real images (SvR), among the real images (RvR), and among synthetic images (SvS). We asses SvR distributions to evaluate memorisation of the real images in the generated samples. Direct interpretation of the SvR distributions implies little memorisation in the generated data as the correlations do not reach close to 1. However, we observe a positive trend between SvR and RvR values: pearson's correlation coefficient $= 0.997$ for the SvR and RvR distributions of correlation values. This indicates that if for a modality the similarity among real images is high, the similarity between real images and synthetic images is also high. This could be a result of the characteristics of the modality. FAF, for example, contains a predominantly grey background which contributes to a high correlation between any two images of this modality. To account for this trend, we compared SvR with RvR for each modality to provide extra context on the extend of memorisation. For no modality was the distribution of SvR higher than RvR, indicating that the similarity of synthetic images to real images is not higher than the similarity among real images. Therefore there is little concern for memorisation of real training images in the

generated images. Similarity among synthetic images is stronger than the similarity among real images for OCTA-SMAC (WD = 0.025, p=$1.22e^{-6}$) and FAF (WD = 0.027, p=$8.079e^{-11}$). This could imply reduced diversity of the synthetic images of these modalities. However, the distribution plots in Figure 4 display that the differences between the RvR and SvS distributions for OCTA-SMAC and FAF are minimal. This eliminates concerns for severe memorisation.

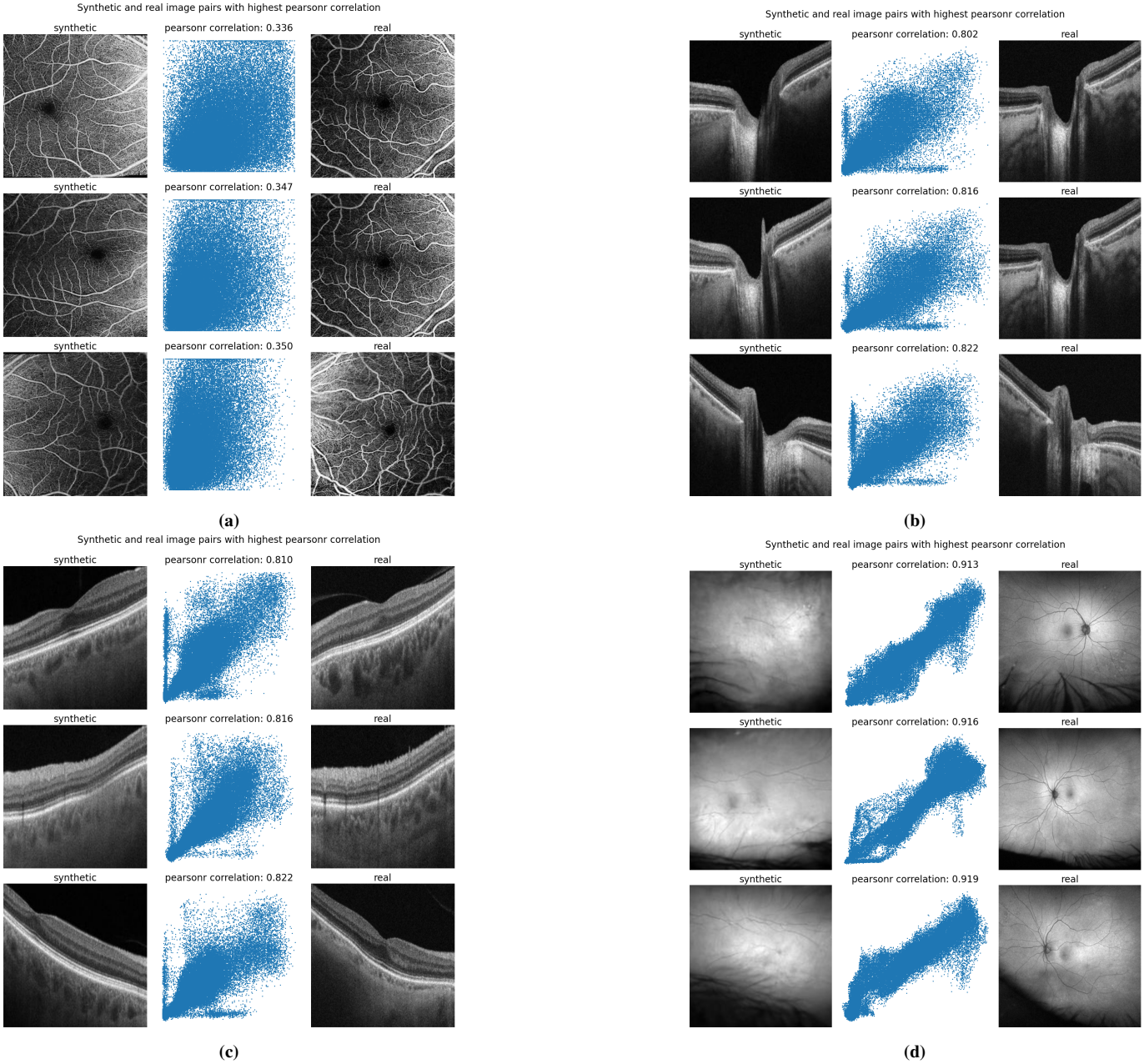


Figure 3: Examples of the synthetic images with the highest correlation to any real image. Displayed are pairs of synthetic image and the corresponding real image that it most closely resembles together with scatter plot of the pixel values accompanied wth the correlation value. (a) OCTA-SMAC; (b) OCT-BONH; (c) OCT-BMAC; (d) FAF. For OCT-BMAC and OCT-BONH the synthetic images have strong resemblance to the real images but are not exact copies. For OCTA-SMAC and FAF the images with the highest correlations show only slight resemblance. For OCTA-SMAC his is also reflected in the lower correlation values. FAF has the highest max correlation values, which is likely due to the large amount of homogeneous grey background of the fundus.

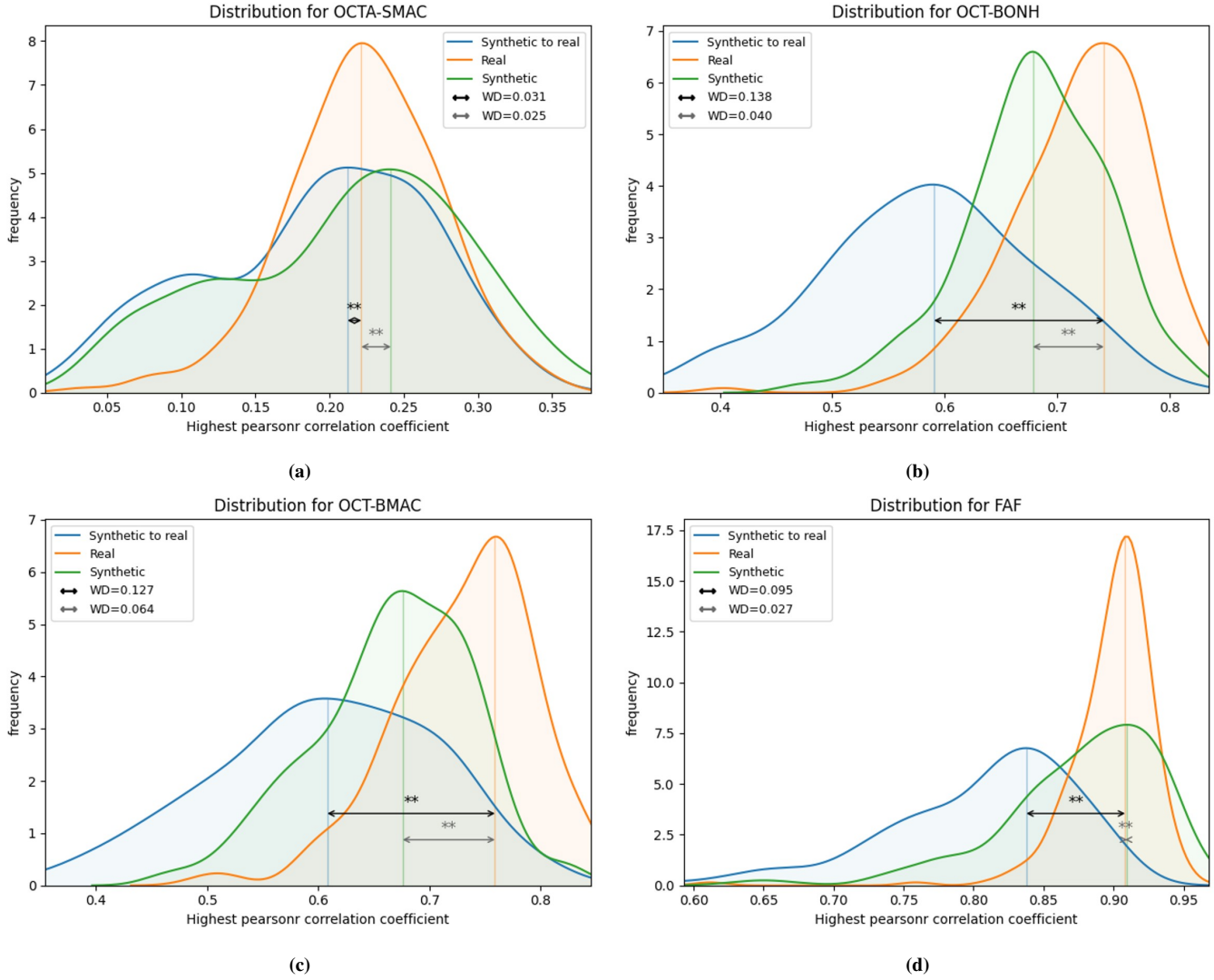


Figure 4: (A-D): Distributions for maximum pearson correlation values computed for 200 synthetic images and all real images. Distributions display the highest correlation for image pairs among real images (RvR, orange), among synthetic images (SvS, green) and for all synthetic images with any real image (SvR, blue). Arrows indicate the differences between SvR and RvR distributions (black) and the differences between RvR and SvS distributions (grey). WD values express the distance between two distributions with KS test p-values for the significance of such differences. pearsonr = Pearson's correlation coefficient. ** = $p < 0.005$

.

B. Filter

The trained filter model achieves 99% accuracy for recognising the modalities on the test set with MCC of 0.990 for the predictions on the validation set and MCC of 0.9968 on the test set (Table S1). This indicates a good agreement between the model's outputs and the true classes. The model performed so well that it even achieved to correctly recognize the modality of unrealistic synthetic images of type OCTA-SMAC, OCT-SONH and OCT-SMAC, diminishing its role in detecting unrealistic images. We tried applying a filter with lower validation accuracy, by training for fewer epochs, in the hope that an image would have to represent its modality very strongly and therefore unrealistic images would be discarded more often. However, this did not make any difference. Therefore, we applied a manually set threshold on the model outputs for these three modalities; 0.9, 0.99 and 0.96, respectively. Any sampled synthetic image will be included in our synthetic dataset if the predicted modality was correct and if the confidence for this modality satisfied this threshold.

C. Classification

Table 2 displays the results of classification experiments. Given the unbalanced nature of AmyloidPET labels in the dataset, AUPR values were computed in addition to AUROC values. Pretraining on synthetic data shows slight improvement in terms of AUPR for OCT-BONH, OCT-BMAC AND FAF in the validation set and for the test set this is the case for OCTA-SMAC and OCT-BMAC. For mulitimodal classification, pretraining on synthetic data and finetuning on real data improved AUPR for

	AUPR			AUROC			F1-score			Sensitivity			Specificity		
	Real	Synth.	Pretr.	Real	Synth,	Pretr.	Real	Synth,	Pretr.	Real	Synth,	Pretr.	Real	Synth,	Pretr.
Valdation															
Unimodal															
- OCTA-SMAC	0.583	**0.632**	0.436	0.646	**0.690**	0.583	0.635	**0.695**	0.565	**0.909**	0.740	0.591	0.364	0.610	**0.667**
- OCT-BONH	0.573	0.593	**0.617**	0.602	0.617	**0.707**	0.553	0.560	**0.690**	0.542	0.510	**0.833**	0.688	**0.690**	0.563
- OCT-BMAC	0.611	0.569	**0.665**	0.668	0.577	**0.728**	0.655	0.605	**0.667**	**0.826**	0.650	0.739	0.484	0.500	**0.645**
- FAF	0.545	**0.864**	0.576	0.744	**0.858**	0.664	0.629	**0.811**	0.519	0.647	**0.835**	0.412	0.774	0.775	**0.903**
Multimodal															
- No metadata	0.406	-	**0.632**	0.567	-	**0.708**	0.600	-	0.571	**0.800**	-	0.533	0.458	-	**0.792**
- With metadata	0.456	-	**0.467**	**0.631**	-	0.592	0.588	-	0.636	0.667	-	**0.933**	**0.625**	-	0.375
Test															
Unimodal															
- OCTA-SMAC	0.338	**0.613**	0.455	0.381	**0.647**	0.586	0.381	**0.541**	0.522	0.381	0.476	**0.571**	0.594	**0.813**	0.594
- OCT-BONH	**0.488**	0.368	0.481	0.530	0.412	**0.583**	0.533	**0.545**	0.489	0.727	**0.955**	0.500	0.389	0.056	**0.667**
- OCT-BMAC	0.350	0.451	**0.579**	0.391	**0.569**	0.615	0.426	0.571	**0.596**	0.435	0.696	**0.739**	**0.576**	0.485	0.485
- FAF	**0.390**	0.310	0.347	**0.549**	0.326	0.478	**0.571**	0.100	0.489	**0.875**	0.063	0.688	0.321	**0.893**	0.357
Multimodal															
- No metadata	0.486	-	**0.441**	**0.622**	-	0.491	**0.500**	-	0.333	**0.500**	-	0.286	0.708	-	**0.750**
- With metadata	**0.634**	-	0.306	**0.729**	-	0.369	**0.625**	-	0.531	0.714	-	**0.929**	**0.667**	-	0.083

Table 2: Classification results on the validation set and test set. Real = trained on real data. Synth. = trained on synthetic data. Pretr. = pretrained on synthetic data and finetuned on real data.

models with and without metadata in the validation set but not in the test set. Including metadata improved AUPR only for the model trained on real data and improved F1 and sensitivity for both the pretrained model, with the bist improvement for the pretrained model (F1-score improved from 0.333 to 0.531). The best performing overall model on the test set in terms of AUPR is the multimodal classifier trained on real data (AUPR 0.634). The best-perfoming unimodal classifier in terms of AUPR is OCTA-SMAC trained on synthetic data and finetuned on (AUPR 0.613). Intresting is the high sensitivity on the test set for OCT-BONH classifier (0.955 trained on synthetic data and sensitivity of the pretrained multimodal classifier with metadata (0.929). The specificity for these models, on the other hand, is low (0.056 and 0.083 respectively). This indicates that these models can be good at detecting AmyloidPET+ cases but at the cost of many false positives. With respect to the F1-score, the best-performing model is the multimodal classifier with metadata (0.625). The best-performing unimodal classifier is trained on synthetic data and finetuned on real data for OCT-BMAC (0.596). Finetuning on real data led to reduction of performance compared to the models trained on synthetic data in several models, the biggest detorioration is seen for OCTA-SMAC.

D. Class activation maps

Figures 5-7 display input images with GradCAMs that identify the regions of the image that contribute to a higher model output. We depict the GradCAM as a heatmap in which the red-orange colour regions are the most discriminative areas for the model to predict AmyloidPET- and the green-blue areas are the least salient. Figures 5-7 contain the same retinal images, to allow for direct comparison of the heatmaps of different classifiers. We compared the GradCAMS in Figure 5 and 6 to interpret and compare the outputs of the baseline and pretrained classifiers. We also review the GradCAMS of the classifier trained on synthetic images (Figure 7) to discover whether a model trained on synthetic data can identify relevant areas in the images.

The GradCAM heatmaps for the three different classifiers trained on OCTA-SMAC show different patterns. The classifier trained on real images shows highest response to blood vessels in the periphery and there is a larger area of high response to real images than synthetic images. In contrast, the pretrained classifier shows a larger response area in the center of the image. The response to the synthetic images is different, with high responses to a large central area. The model trained on synthetic images displays attention to small areas in the periphery. We do not observe large differences in the size of the heatmap areas with high response between true negative an true positive images, except for the heatmaps on real images produced by the classifier trained on synthetic images. This is also reflected by the small differences in output values.

The OCT-BONH classifier trained on real images shows high response to small areas in the layers of the retina. The pretrained classifier shows responses to similar areas, with slightly larger areas of high activation in the pretrained images. The shape and location of the high response areas imply that these models learned to identify meaningful features, however the output values for the different classes are very close to each other (0.494 and 0.485 compared to 0.471 and 0.489) which implies that the classes are not well distinguished by these features. The classifier trained on synthetic data shows larger areas of high response, even more so in the real images compared to synthetic images.

The pretrained OCT-BMAC classifier shows very little response in any of the images, except for the heatmap on the real AmyloidPET+ images which depicts a localised response in the photoreceptor layer and retinal pigment epithelium. Interestingly, the classifier trained on real images shows localised response to a synthesis artifact in the synthetic AmyloidPET- images. The

heatmaps produced by this classifier seem different for the images, nonetheless the confidence for AmyloiPET- is similar. In contrast, the classifier trained on synthetic images shows a low response to the synthesis artifcact. The responses are more localised than in the classifier trained on real images.

The GradCAM heatmaps for the three different classifiers trained on FAF show different patterns. The pretrained model identifies areas around the fovea and ONH. For both synthetic and real images there are slightly larger areas of high response which is also reflected in the output scores. The model trained on synthetic images shows responses of very different shapes when comparing the synthetic and real images. The synthetic images show a large squared area of high response whereas the response to real images is more restricted to specific areas, mostly in the periphery of the fundus. The model trained on real images shows strongest responses in small areas of the far periphery of the fundus. This model produces larger areas of high response to the synthetic images. These responses are more localised around the ONH and fovea.

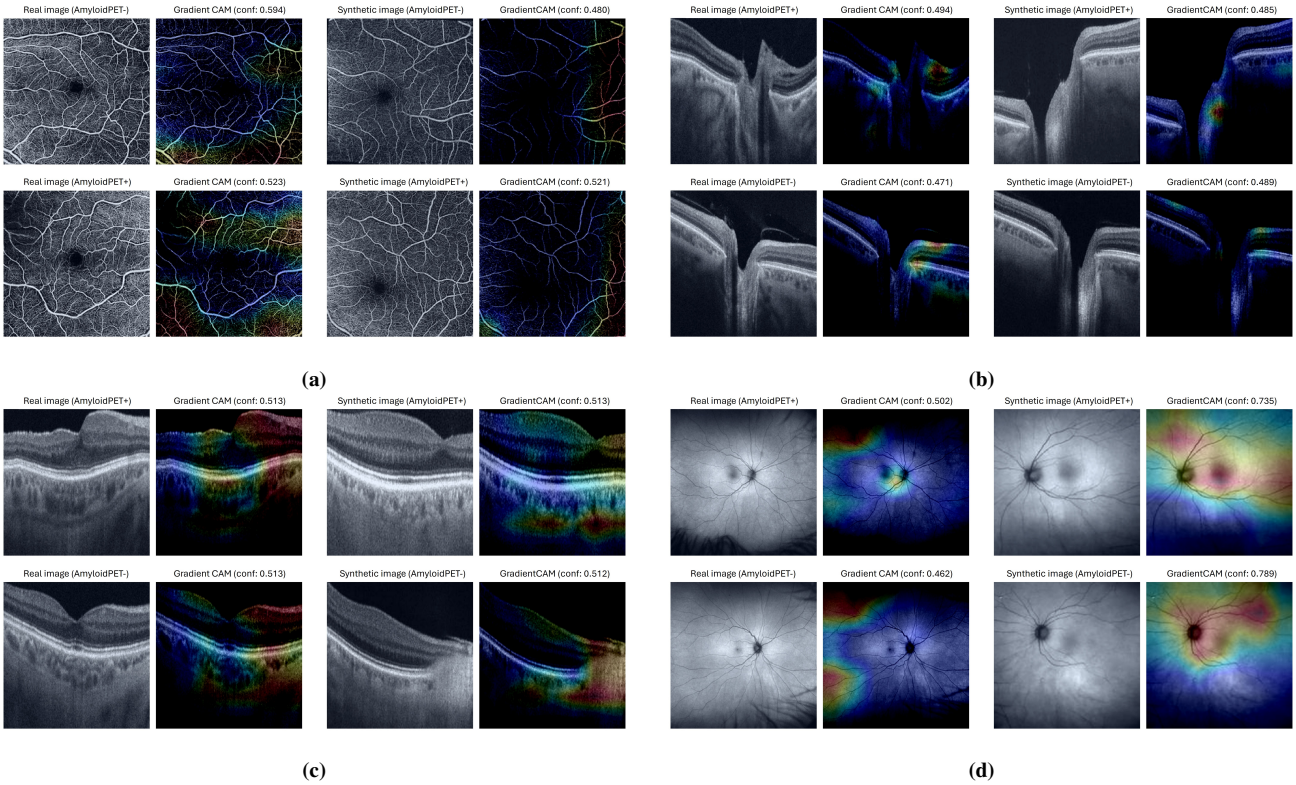


Figure 5: GradCAMs generated for classifiers trained on real images only. The subfigures show pairs of input images and GradCAM heatmaps for OCTA-SMAC (a), OCT-BMONH (b), OCT-BMAC (c), FAF (d). Each subfigure has four images. Left top to bottom: real images for AmyloidPET+ and AmyloiPET-. Right top to bottom: synthetic Images idem for for AmyloidPET+ and AmyloiPET-.

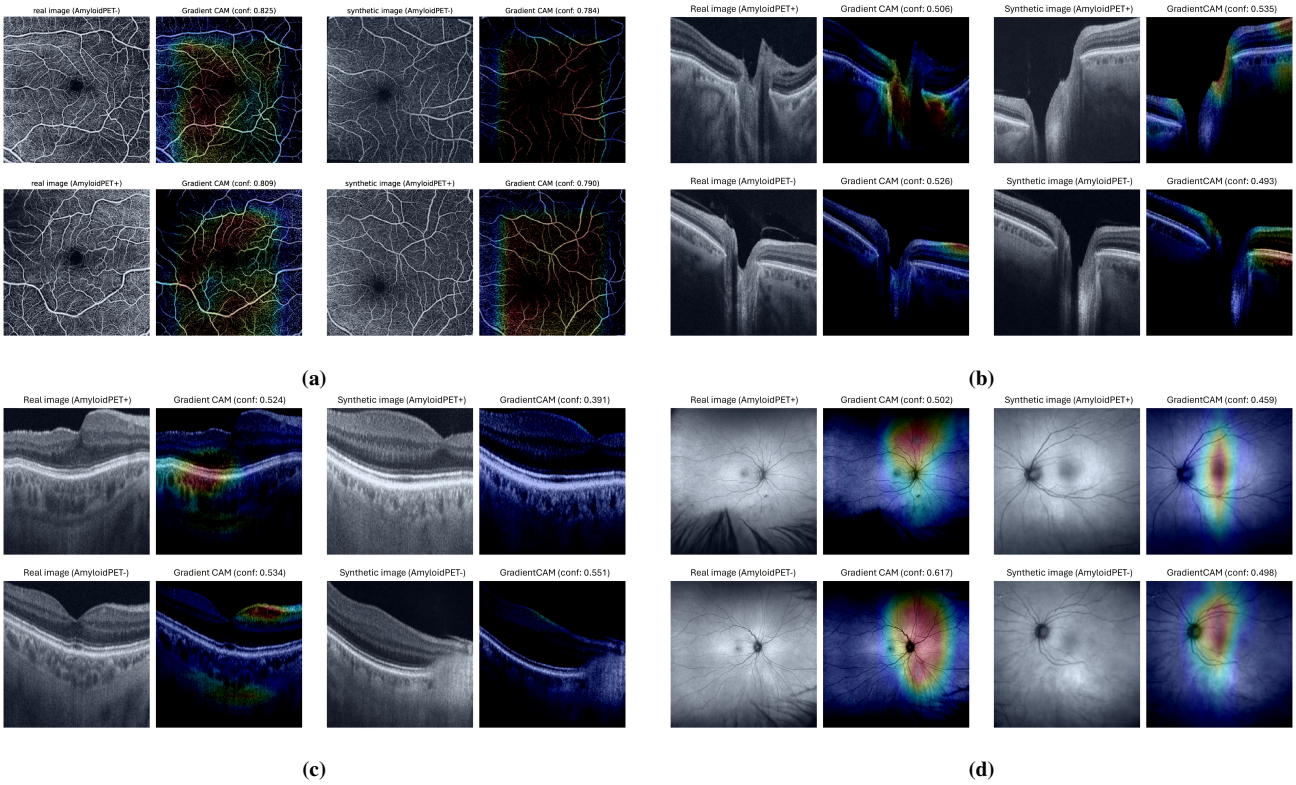


Figure 6: GradCAMs generated for classifiers pretrained on synthetic images and finetuned on real images. The subfigures show pairs of input images and GradCAM heatmaps for OCTA-SMAC (a), OCT-BMONH (b), OCT-BMAC (c), FAF (d). Each subfigure has four images. Left top to bottom: real images for AmyloidPET+ and AmyloiPET-. Right top to bottom: synthetic Images idem for for AmyloidPET+ and AmyloiPET-.

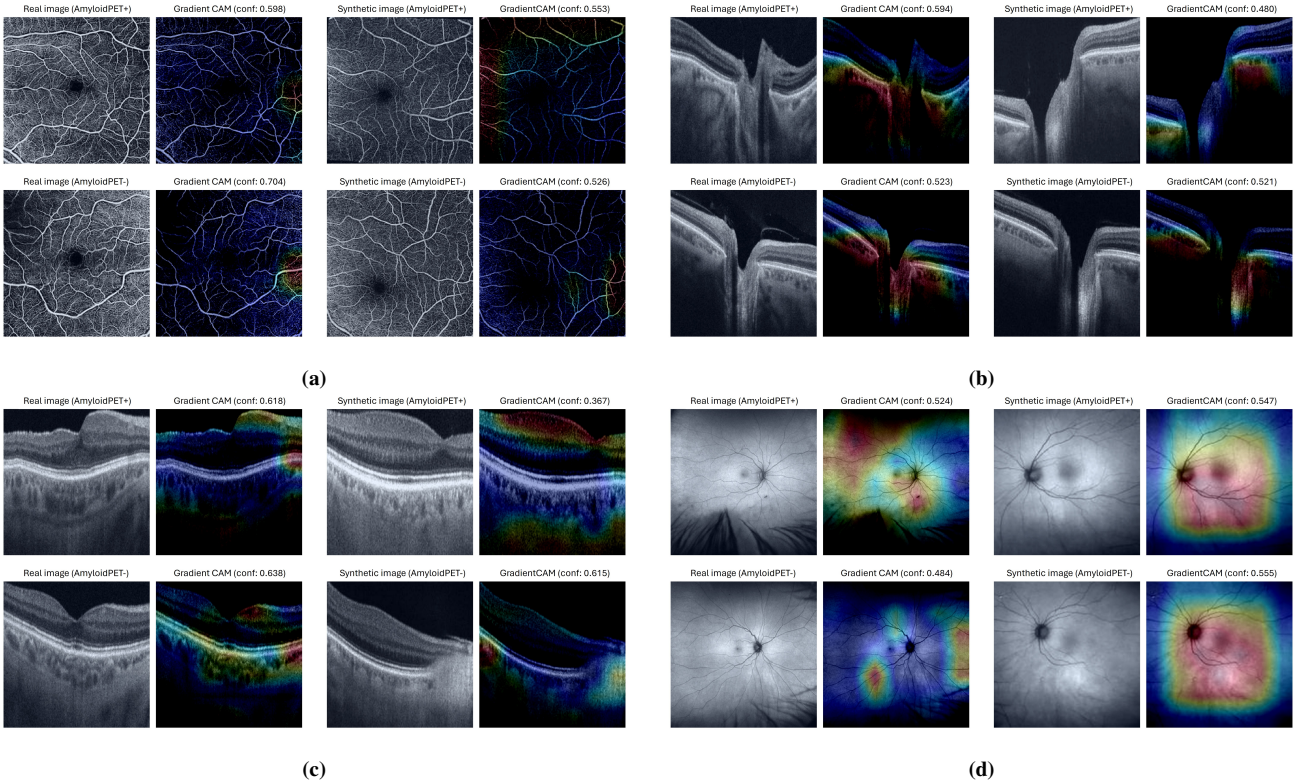


Figure 7: GradCAMs generated for classifiers trained on synthetic images only. The subfigures show pairs of input images and GradCAM heatmaps for OCTA-SMAC (a), OCT-BMONH (b), OCT-BMAC (c), FAF (d). Each subfigure has four images. Left top to bottom: real images for AmyloidPET+ and AmyloiPET-. Right top to bottom: synthetic Images idem for for AmyloidPET+ and AmyloiPET-.

V. Discussion

To our knowledge, this is the first study to generate multimodal synthetic image data to detect AmyloidPET status with retinal imaging. The first aim of the present study is to create synthetic multimodal retinal images. Akbar, Wang, and Eklund observed that diffusion models are more likely to memorize the training images, compared to StyleGAN, especially for small datasets [17]. Nonetheless our experiments demonstrated how to construct a U-Net DDPM that is capable of generating diverse images that are unique and not copies of the real images. By incorporating a CNN in the sampling procedure that is trained to recognized image modalities it is ensured that realistic synthetic images are used for classification.

In the present study, the dataset with real images is limited in size, posing the risk that the model could not learn to recognize salient features in an image without overfitting on the training set. This would mean that the model fits well to the training data but performs poorly on unobserved testing data. To deal with this issue, the development set was complemented with 1000 images per class, generated by a conditional DDPM. We hypothesized that synthetic images contain information that is relevant for learning to classify AmyloidPET status. Therefore, we assumed that pretraining CNNs with synthetic data could learn the model to recognize important features, leading to better performance when subsequent training on real data. Pretraining with synthetic data slightly improved classification performance for two out of the four modalities in terms of AUPR and improved F1-score for one modality. With this we have shown that our method for exploiting synthetic data has the potential to improve CNN performance in small datasets in medical image classification. Further research into the effect of the training budget of the synthetic data, different synthesis methods and different ways for exploiting synthetic data should give insights into the potential of generative AI for future applications in deep learning in medical imaging.

Our best performing unimodal and multimodal classifiers were not pretrained on synthetic data. For some models, finetuning on real data led to reduction of performance compared to the models trained on synthetic data in several models. One possible explanation for this could be the low similarity between real and synthetic images. Our overall best performing model takes multimodal FSLO, OCT and OCT-A inputs and metadata, achieving AUPR of 0.634 and AUROC of 0.729 on the test set. This outperforms our best-performing unimodal classifiers. Our motivation for designing a multimodal classifier and to incorporate metadata comes from prior studies on multimodal CNN for AD diagnosis [10], [12], [33]. Incorporation of age and gender metadata as inputs to the multimodal classifier improved performance from AUPR 0.486 (AUROC 0.622) to 0.634 (AUROC 0.729). Such improvement was also found by Wisely, Wang, Henao, *et al.* for AD diagnosis. However, it is difficult to make a direct comparison between our studies as their proportion of cognitively healthy individuals is only 23% of the 222 eyes [10]. Furthermore, AmyloidPET detection is different from AD diagnosis as cognitively healthy individuals can be positive for AmyloidPET biomarker. The experiments in by Cheung, Ran, Wang, *et al.* on AmyloidPET status prediction are conducted on unilateral and bilateral inputs [12]. As the model with bilateral inputs improved (AUROC = 0.68 - 0.86 on external validation) compared to unilateral predictions (AUROC = 0.61 - 0.83 on external validation), it would be worthwile to investigate this effect for our datasets and models. However, the size of our dataset would results in very few training examples for the bilateral predictions.

FAF-based models perform bad compared to other classifiers, which is in line with [10], which concluded that FSLO images provide low utility for predicting AD diagnosis. It would be interesting to explore in future experiments the incorporation of metadata into the unimodal classifiers. The unimodal classifier would then have to be adapted to facilitate pretraining without metadata and subsequent finetuning with metadata. Incorporating metadata can have favourable and infavourable effects. Cheung, Ran, Wang, *et al.* argue that the fact that their model does not require patient data is an advantage to their approach [12]. Although not incorporating patient information like age and gender could prevent CNNs from bias towards patient groups of certain demographics, it can be hypothesised that learning from patient age and gender is valuable. As these patient characteristics can affect the retinal structure, it can be stated that a model should be able to exploit this patient information in order to extract better features and consequently make better predictions [34], [35]. One hypothesized reason why Wisely, Wang, Henao, *et al.* and our study are able to improve performance with patient information can lie in the type of retinal images used; Cheung, Ran, Wang, *et al.* used exclusively retinal photographs whereas Wisely, Wang, Henao, *et al.* used additional OCT-derived modalities which look at retinal structures at much higher axial resolution [10], [12]. This diversity of inputs can possibly grant more opportunity for utilisation of metadata in the feature extraction process.

As this is the first study to our knowledge to use synthetic retinal images for AmyloidPET or AD prediction, this study evaluated the models' internal states with heatmaps to visualise what knowledge is learned from synthetic images. The GradCAM heatmaps in Figures 7 indicate that training on synthetic images can teach a model recognize salient regions in real images. This study demonstrated that that attention of a CNN based on OCTA-SMAC pretrained on synthetic images and finetuned on real images can bring attention towards the center of the images, which is where the foveal avascular zone (FAZ) is located. Several meta analyses show that FAZ size has been associated with AD [5], [36]. Cheung, Ran, Wang, *et al.* provided heatmaps with their fundus-based classifier and one other existing study provides saliency maps for fundus-based

AD prediction [12], [37]. This study showed that small blood vessels are most salient for AD prediction. The heatmaps of our FAF classifier trained on real images projected on synthetic images shows similarities with that of [12] as the high attention areas follow along the main vascular branches. The FAF classifier trained on synthetic images projects similar attention on the small vascular branches in real images.

Limitations to this study are related to the nature of the training and validation sets. With our small dataset it is difficult to fit models that generalise well on unseen data. Furthermore, performance on the evaluation dataset may not be good indicators of the performance of our methods, as the composition of the evaluation dataset influence the performance and may greatly vary depending on the split. We attempted to tackle this problem with stratified splits, but additional cross-validation experiments would give more reliable insights. Evidently, a larger dataset could also counteract these problems.

REFERENCES

[1] X. Li, X. Feng, X. Sun, N. Hou, F. Han, and Y. Liu, "Global, regional, and national burden of alzheimer's disease and other dementias, 1990–2019," *Frontiers in Aging Neuroscience*, vol. 14, 2022, ISSN: 1663-4365. DOI: 10.3389/fnagi.2022.937486.

[2] J. Zhang, L. Shi, and Y. Shen, "The retina: A window in which to view the pathogenesis of alzheimer's disease," *Ageing Research Reviews*, vol. 77, p. 101 590, May 2022. DOI: 10.1016/j.arr.2022.101590.

[3] C. Y. Cheung, V. Mok, P. J. Foster, E. Trucco, C. Chen, and T. Y. Wong, "Retinal imaging in alzheimer's disease," *Journal of Neurology, Neurosurgery &amp Psychiatry*, vol. 92, no. 9, pp. 983–994, Jun. 2021. DOI: 10.1136/jnnp-2020-325347.

[4] P. Zabel, J. J. Kaluzny, K. Zabel, *et al.*, "Quantitative assessment of retinal thickness and vessel density using optical coherence tomography angiography in patients with alzheimer's disease and glaucoma," *PLOS ONE*, vol. 16, no. 3, A. S. Lewin, Ed., e0248284, Mar. 2021. DOI: 10.1371/journal.pone.0248284.

[5] Q. Jin, Y. Lei, R. Wang, H. Wu, K. Ji, and L. Ling, "A systematic review and meta-analysis of retinal microvascular features in alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 13, Jun. 2021. DOI: 10.3389/fnagi.2021.683824.

[6] J. Salazar, A. Ramirez, R. de Hoz, and P. Rojas, "Amyotrophic lateral sclerosis, a neurodegenerative motor neuron disease with retinal involvement," *Neural Regeneration Research*, vol. 17, no. 5, p. 1011, 2022. DOI: 10.4103/1673-5374.324841.

[7] S. Basaia, F. Agosta, L. Wagner, *et al.*, "Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks," *NeuroImage: Clinical*, vol. 21, p. 101 645, 2019, ISSN: 2213-1582. DOI: 10.1016/j.nicl.2018.101645.

[8] N. Amoroso, D. Diacono, A. Fanizzi, *et al.*, "Deep learning reveals alzheimer's disease onset in mci subjects: Results from an international challenge," *Journal of Neuroscience Methods*, vol. 302, pp. 3–9, May 2018, ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2017.12.011.

[9] X. Wang, B. Jiao, H. Liu, *et al.*, "Machine learning based on optical coherence tomography images as a diagnostic tool for alzheimer's disease," *CNS Neuroscience &amp Therapeutics*, vol. 28, no. 12, pp. 2206–2217, Sep. 2022. DOI: 10.1111/cns.13963.

[10] C. E. Wisely, D. Wang, R. Henao, *et al.*, "Convolutional neural network to identify symptomatic alzheimer's disease using multimodal retinal imaging," *British Journal of Ophthalmology*, vol. 106, no. 3, pp. 388–395, Nov. 2020. DOI: 10.1136/bjophthalmol-2020-317659.

[11] C. E. Wisely, A. Richardson, R. Henao, *et al.*, "A convolutional neural network using multimodal retinal imaging for differentiation of mild cognitive impairment from normal cognition," *Ophthalmology Science*, vol. 4, no. 1, p. 100 355, Jan. 2024, ISSN: 2666-9145. DOI: 10.1016/j.xops.2023.100355.

[12] C. Y. Cheung, A. R. Ran, S. Wang, *et al.*, "A deep learning model for detection of alzheimer's disease based on retinal photographs: A retrospective, multicentre case-control study," *The Lancet Digital Health*, vol. 4, no. 11, e806–e815, Nov. 2022, ISSN: 2589-7500. DOI: 10.1016/s2589-7500(22)00169-8.

[13] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for medical image synthesis: An empirical study," *Journal of Imaging*, vol. 9, no. 3, p. 69, Mar. 2023. DOI: 10.3390/jimaging9030069.

[14] L. Yang, Z. Zhang, Y. Song, *et al.*, "Diffusion models: A comprehensive survey of methods and applications," 2022. DOI: 10.48550/ARXIV.2209.00796.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. DOI: 10.48550/ARXIV.2006.11239.

[16] A. Kazerouni, E. K. Aghdam, M. Heidari, *et al.*, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, p. 102 846, May 2023. DOI: 10.1016/j.media.2023.102846.

[17] M. U. Akbar, W. Wang, and A. Eklund, "Beware of diffusion models for synthesizing medical images – a comparison with gans in terms of memorizing brain mri and chest x-ray images," 2023. DOI: 10.48550/ARXIV.2305.07644.

[18] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," 2019. DOI: 10.48550/ARXIV.1903.07291.

[19] M. Akrout, B. Gyepesi, P. Holló, *et al.*, "Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images," 2023. DOI: 10.48550/ARXIV.2301.04802.

[20] W. Peng, E. Adeli, Q. Zhao, and K. M. Pohl, "Generating realistic 3d brain mris using a conditional diffusion probabilistic model," 2022. DOI: 10.48550/ARXIV.2212.08034.

[21] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021. DOI: 10.48550/ARXIV.2105.05233.

[22] H. K. Kim, I. H. Ryu, J. Y. Choi, and T. K. Yoo, "A feasibility study on the adoption of a generative denoising diffusion model for the synthesis of fundus photographs using a small dataset," *Discover Applied Sciences*, vol. 6, no. 4, Apr. 2024, ISSN: 3004-9261. DOI: 10.1007/s42452-024-05871-9.

[23] E. Konijnenberg, S. F. Carter, M. ten Kate, *et al.*, "The emif-ad preclinad study: Study design and baseline cohort overview," *Alzheimer's Research amp; Therapy*, vol. 10, no. 1, Aug. 2018, ISSN: 1758-9193. DOI: 10.1186/s13195-018-0406-7.

[24] J. A. van de Kreeke, H.-T. Nguyen, J. den Haan, *et al.*, "Retinal layer thickness in preclinical alzheimer's disease," *Acta Ophthalmologica*, vol. 97, no. 8, pp. 798–804, May 2019, ISSN: 1755-3768. DOI: 10.1111/aos.14121.

[25] D. I. Boomsma, E. J. C. d. Geus, J. M. Vink, *et al.*, "Netherlands twin register: From twins to twin families," *Twin Research and Human Genetics*, vol. 9, no. 6, pp. 849–857, Dec. 2006, ISSN: 1839-2628. DOI: 10.1375/twin.9.6.849.

[26] M. ten Kate, C. H. Sudre, A. den Braber, *et al.*, "White matter hyperintensities and vascular risk factors in monozygotic twins," *Neurobiology of Aging*, vol. 66, pp. 40–48, Jun. 2018, ISSN: 0197-4580. DOI: 10.1016/j.neurobiolaging.2018.02.002.

[27] J. A. van de Kreeke, H. T. Nguyen, E. Konijnenberg, *et al.*, "Retinal and cerebral microvasculopathy: Relationships and their genetic contributions," *Investigative Opthalmology amp; Visual Science*, vol. 59, no. 12, p. 5025, Oct. 2018, ISSN: 1552-5783. DOI: 10.1167/iovs.18-25341.

[28] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7.

[29] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," 2017. DOI: 10.48550/ARXIV.1709.07871.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017. DOI: 10.48550/ARXIV.1708.02002.

[31] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2015. DOI: 10.48550/ARXIV.1512.04150.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7.

[33] T.-C. Yeh, C.-T. Kuo, and Y.-B. Chou, "Retinal microvascular changes in mild cognitive impairment and alzheimer's disease: A systematic review, meta-analysis, and meta-regression," en, *Front. Aging Neurosci.*, vol. 14, p. 860 759, Apr. 2022.

[34] B. W. Polascik, A. C. Thompson, S. P. Yoon, *et al.*, "Association of oct angiography parameters with age in cognitively healthy older adults," *Ophthalmic Surgery, Lasers and Imaging Retina*, vol. 51, no. 12, pp. 706–714, Dec. 2020, ISSN: 2325-8179. DOI: 10.3928/23258160-20201202-05.

[35] M. R. Munk, T. Kurmann, P. Márquez-Neila, M. S. Zinkernagel, S. Wolf, and R. Sznitman, "Assessment of patient specific information in the wild on fundus photography and optical coherence tomography," *Scientific Reports*, vol. 11, no. 1, Apr. 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-86577-5.

[36] G. Ashraf, M. McGuinness, M. A. Khan, C. Obtinalla, X. Hadoux, and P. van Wijngaarden, "Retinal imaging biomarkers of alzheimer's disease: A systematic review and meta-analysis of studies using brain amyloid beta status for case definition," *Alzheimer's amp; Dementia: Diagnosis, Assessment amp; Disease Monitoring*, vol. 15, no. 2, Apr. 2023, ISSN: 2352-8729. DOI: 10.1002/dad2.12421.

[37] J. Tian, G. Smith, H. Guo, *et al.*, "Modular machine learning for alzheimer's disease classification from retinal vasculature," en, *Sci. Rep.*, vol. 11, no. 1, p. 238, Jan. 2021.

APPENDIX

A. Dataset

The full image dataset covers retinal scans from three types of retinal scans: Fundus SLO (Optos), OCT-A (Zeiss Angioplex) and volumetric OCT (Heidelberg OCT). The initial dataset included a set of 26 modalities. Optos: FAF and red-green fundus (FRG); Zeiss Angioplex: depth-encoded, and five layer-specific OCT-A and structural B-scan images for both the ONH and macula; Heidelberg: 3D volumetric cubes of the optic nerve and macula. Initial experiments were performed to create synthetic images for FAF, FRG, depth-encoded macula angiography (OCTA-EMAC), depth-encoded ONH angiography (OCTA-EONH), layer-specific OCT-A of the deep and superficial layers of the ONH (OCTA-DONH, OCTA-SONH) and macula (OCTA-DMAC, OCTA-SMAC). We also experimented with 3D DDPM for volumetric scans of the ONH (OCT-VONH) and macula (OCT-VMAC). The experiments failed to create realistic volumetric, depth-encoded and FRG images. Therefore we made a selection of four modalities extracted from FSLO, OCT-A and structural OCT scans to maintain the variety of the inputs to the classification models. For FSLO we selected the FAF modality as we could not create realistic synthetic images for FRG. For OCT-A we used the superficial layer of the macula instead of the depth-encoded OCTA-EONH and OCTA-EMAC as most research to date suggest changes in FAZ and vessel density in superficial macula may resemble AD progression [1]–[3]. Instead of volumetric structural OCT we used 2D structural OCT-BONH and OCT-BMAC that were generated during the angiography recordings on the Zeiss Angioplex. A comparison of the 2D real and synthetic images for the initial and final DDPM's is depicted in Supplementary Figure S1.

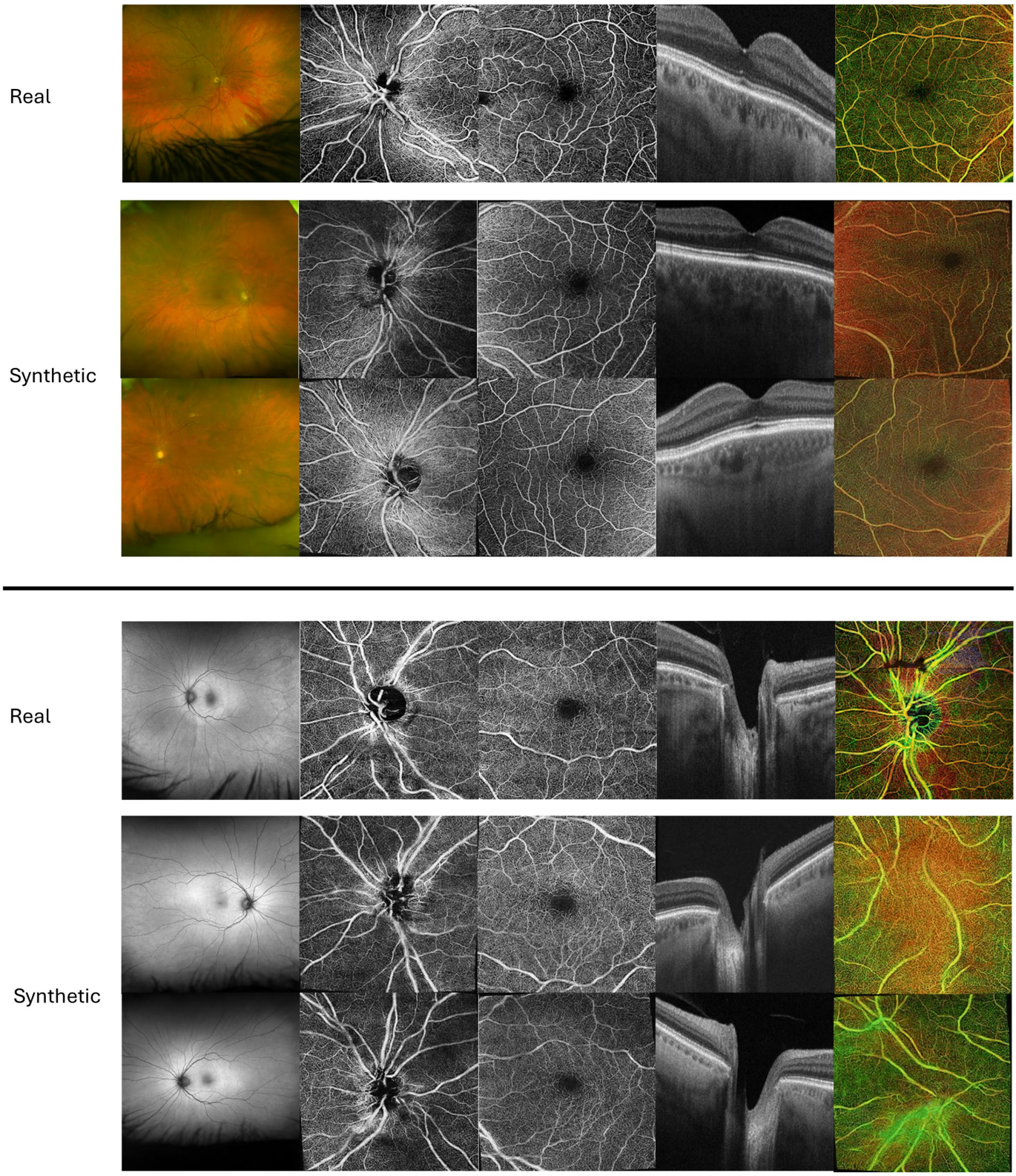


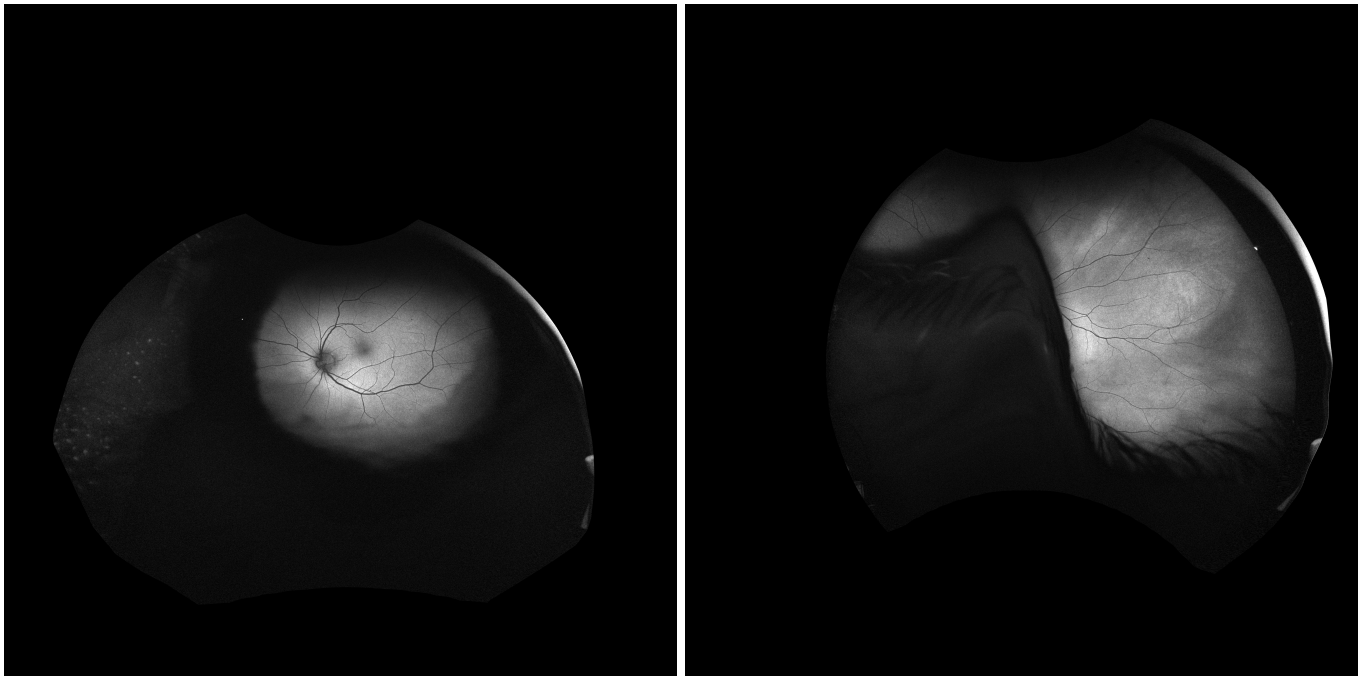
Figure S1: Examples of generated synthetic images and real images for eight modalities. Top row from left to right: FRG, OCTA-SONH, OCTA-SMAC; OCT-BMAC; OCTA-EMAC. Bottom row left to right: FAF; OCTA-DONH; OCTA-DMAC; OCT-BONH; OCTA-EONH. The synthetic FAF images best resemble their real counterpart, with often accurate branching of the bloodvessels and even replication of the eye lashes at the periphery of the image. Synthetic FRG images fail to replicate the vasculature. Furthermore, the images were not sharp and most of them lacked accurate colours and would be green or yellow similar to the the depth-encoded images. Most of the OCT-A images failed to replicate accurate branching of the blood vessels. OCT B-Scan images are overall quite realistic, although sometimes replication of one or two layers in the retina would occur. Depth-encoded OCT-A synthetic images are the least realistic, with malformations in the vasculature as well as in the colouring.

B. Data extraction

Extraction of the image data from the scanners involved decoding of binary files, anonymization of patient identifiers and association of files to the respective examinations. Anonymous patient identifiers were already created in the context of the two studies where the image data originate from. The eye examinations were stored in the scanners under these IDs. We replaced these identifiers with new pseudo-anonymous keys for the family, patient and eye such that our dataset cannot be directly related to information stored in the scanners. Fundus SLO was extracted as DICOM files and directly anonymized. Image quality was reviewed based on movement or optical distortions such as lash or eyelid coverage and a straight eye gaze was required. Two examples of excluded FAF scans are shown in Figure S2. Initially we intended to use widefield fotos which allowed for a more complete depiction of the fundus, but this dataset contained only a small set of images with 74 auto-fluorescence and 76 red-green images. OCT-A images from the Zeiss Angioplex were evaluated by focus and resolution. Images were exported as .bmp and converted to jpg while sensitive metadata such as study date were removed from the filename. Exported files from the Heidelberg OCT lacked any reference to the examination it contained. Furthermore, these files were encoded as .E2E binary files, for which the scanner manufacturer provides no software to read the data. Therefore we had to decode the files ourselves. By adapting code from https://github.com/marksgraham/OCT-Converter to the structure of our data files we managed to extract binary image data, convert it into .npy files and to extract patient identifiers with image metadata that allowed for identification of the recording type and follow-up order of examinations as multiple visits per patients were recorded.

C. Preprocessing

After building the database, images were preprocessed to be used in the neural networks. Preprocessing was the same for the generative models, the filter and the classification models. Pixel values were normalised according to the untrained or pretrained neural networks (Resnet, EfficientNet). In case grayscale images were used on networks designed for RGB, grayscale channels were duplicated to create three-channel tensors. Age inputs were rescaled by 0.01 and sex was encoded as binary 0 (Male) 1 (Female). Fundus SLO images (originally 4000×4000 pixels) were cropped to remove the black background as visible in Figure S2 and then zoomed in to bring the fundus in the field of view and remove the eye lashes. Background and irrelevant structures of the sclera in B-scan OCT images (originally 1536×1536 pixels) were removed by cropping the top and bottom regions with pixel intensity lower then a manually set threshold. All 2D Images were resized to 256×256 pixels. 3D OCT images were resized to $32 \times 128 \times 128$ (raw images: $73 \times 384 \times 496$ (OCT-VONH); $49 \times 512 \times 496$ (OCT- VMAC))



(a) **(b)**

Figure S2: Example of excluded fundus. The two most often occurring reasons for exclusion were insufficient focus (a) and coverage of the fundus by the eyelid or eye lashes (b).

D. Augmentation

For augmenting the 2D images, random affine (shear, rotate) transformations, color (brightness, contrast), crop and zoom were applied. The zoom factor for B-Scan OCT was left unchanged as interpolation methods resulted in poor image quality. Volumetric 3D OCT was augmented by randomly cropping $75 - 10\%$ of the original volume in the X-Y plane.

E. Hyperparameters of the neural networks

The best learning rate, learning rate scheduler and optimizer hyperparameters were obtained through hyperparameter optimisation. Model designs and approaches were also selected by comparison of small (25-trial) hyperparameter optimisation experiments. Hyperparameter optimisation used the Tree-structured Parzen Estimator (TPE) algorithm through the Optuna Framework. Furthermore, training batches were formed with weighted random sampling to address class imbalance of AmyloidPET labels by oversampling the AmyloidPET+ samples.

1) DDPM: A DDPM with U-Net backbone was used from the MONAI Generative open source project [4]. The model is trained to take inputs of 256×256 pixels with three blocks in the encoders and decoders $(64, 128$ and 128 channels$)$, two residual blocks per encoder/decoder block and spatial transformer attention mechanisms with 32 channels per head in the last encoder/decoder block. The spatial transformer learns to model complex spatial transformations to align the feature maps more accurately with the target distribution. The diffusion model is trained to produce images corresponding to the desired AmyloidPET status by conditioning the UNet and therefore conditioning the denoising process during training and sampling. This conditioning is achieved through incorporation of the class label embedding into the timestep embedding. The timestep embedding is incorporated into the output by passing it through a linear layer and subsequent summation with the intermediate representation of the input at each block of the UNet. We used a scaled linear beta noise schedule function to add noise to the training images, initial learning rate of $1e^{-3}$, cosine annealing learning rate scheduler and Adam optimizer without weight decay.

2) Filter: The network for the filter is an EfficientnetB0 model pretrained on Imagenet. All layers in the model were finetuned during training. The network was trained for 100 epochs with learning rate $1e^{-5}$, Adam optimizer with weight decay of 0.1 and no learning rate scheduler. The training objective was to correctly recognize the modality of an image out of 26 different classes. The model performed well, with Matthew's Correlation Coefficient (MCC) of 0.9900 for the predictions on the validation set and MCC of 0.9968 on the test set. Results for all modalities are shown in Table S1 and Figure S3. Abbrevations:

Modality (abbreviation)	Modality (explanation)	precision	recall	f1-score
COL	Fundus SLO red-green	1.00	1.00	1.00
FAF	Fundus SLO autofluorescence	1.00	1.00	1.00
OCTA-EMAC	OCT-A Macula Angiography Depth Encoded	1.00	1.00	1.00
OCTA-EONH	OCT-A ONH Angiography Depth Encoded	1.00	1.00	1.00
OCTA-WONH	OCT-A ONH Angiography WholeEye	0.99	1.00	0.99
OCTA-WMAC	OCT-A Macula Angiography WholeEye	1.00	1.00	1.00
OCTA-ORCCMAC	OCT-A Macula Angiography ORCC	1.00	1.00	1.00
OCTA-ORCCONH	OCT-A ONH Angiography ORCC	1.00	1.00	1.00
OCTA-RMAC	OCT-A Macula Angiography Retina	0.95	1.00	0.98
OCTA-RONH	OCT-A ONH Angiography Retina	1.00	1.00	1.00
OCTA-DMAC	OCT-A Macula Angiography Deep	1.00	0.99	0.99
OCTA-DONH	OCT-A ONH Angiography Deep	1.00	1.00	1.00
OCTA-SMAC	OCT-A MaculaAngiography Superficial	1.00	0.96	0.98
OCTA-SONH	OCT-A ONH Angiography Superficial	1.00	1.00	1.00
OCT-WONH	OCT ONH Structure WholeEye	1.00	1.00	1.00
OCT-WMAC	OCT Macula Structure WholeEye	1.00	1.00	1.00
OCT-ORCCMAC	OCT Macula Structure ORCC	1.00	1.00	1.00
OCT-ORCCONH	OCT ONH Structure ORCC	1.00	1.00	1.00
OCT-RMAC	OCT Macula Structure Retina	0.99	1.00	0.99
OCT-RONH	OCT ONH Structure Retina	1.00	1.00	1.00
OCT-DMAC	OCT Macula Structure Deep	1.00	0.97	0.99
OCT-DONH	OCT ONH Structure Deep	1.00	1.00	1.00
OCT-SMAC	OCT Macula Structure Superficial	1.00	1.00	1.00
OCT-SONH	OCT ONH Structure Superficial	1.00	1.00	1.00
OCT-BMAC	OCT Macula B-Scan	1.00	1.00	1.00
OCT-BONH	OCT ONH B-Scan	1.00	1.00	1.00
weighted avg		1.00	1.00	1.00

Table S1: Accuracy of filter

.

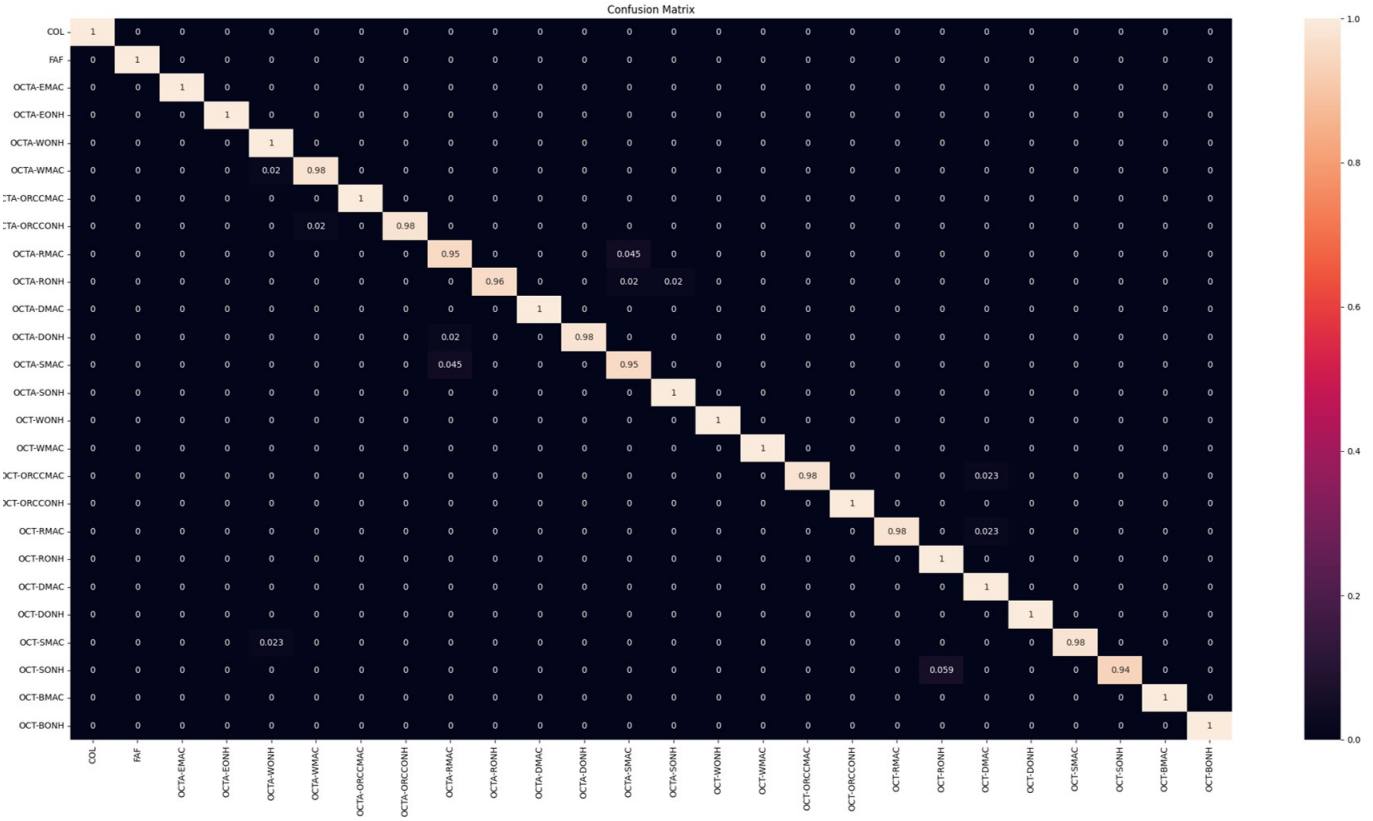
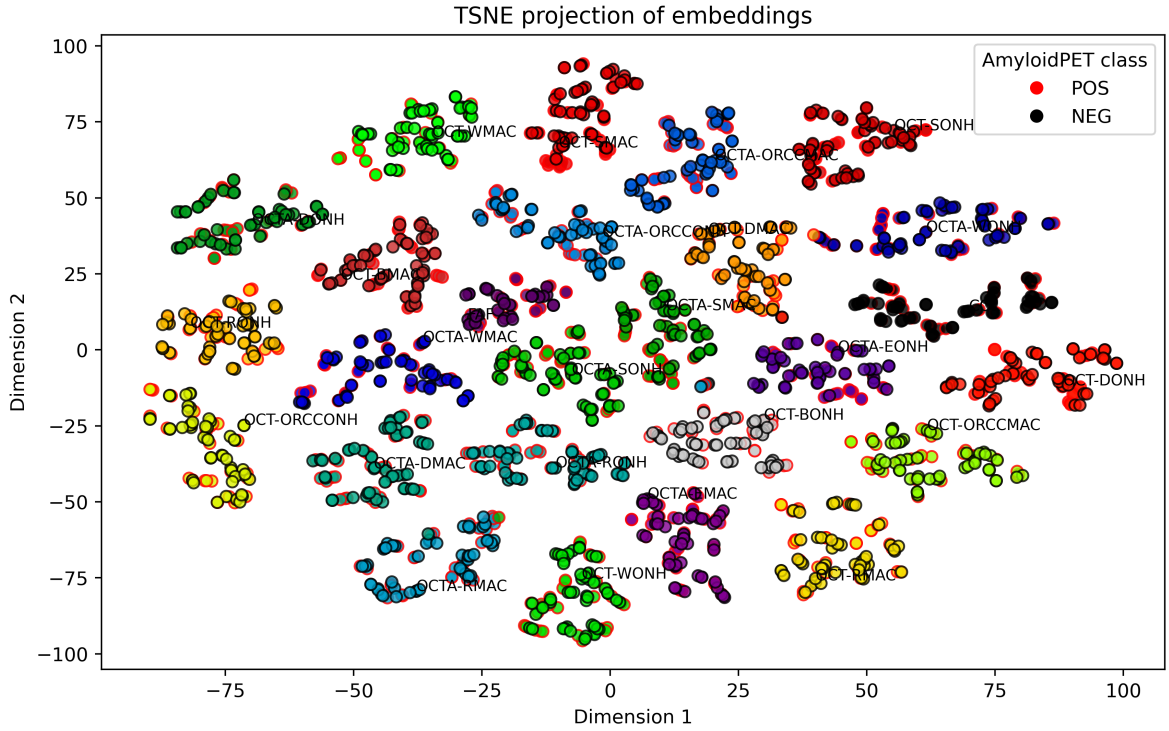
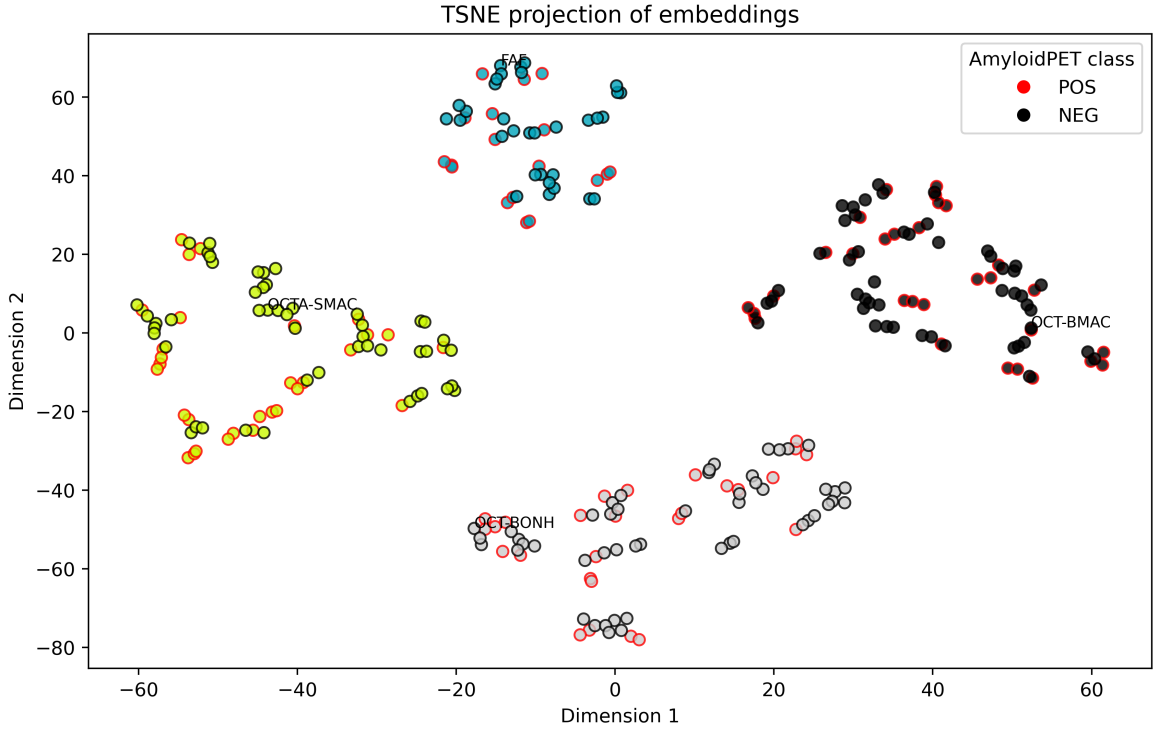


Figure S3: Confusion matrix for distinguishing between 26 modalities by the filter. The wrong predictions happen among modalities that are relatively similar.

3) Modality-aware classifier with FiLM: As the filter model is successful at distinguishing modalities, we experimented with training one modality-aware classification network that is capable of predicting AmyloidPET status for any of the four modalities through an additional input that informs the model of the modality type. This additional input is an embedding vector extracted from the first fully connected layer of the filter model. Figure S4 displays these embeddings on reduced 2D-dimensional axes for all images in the test set and demonstrates that this embedding is suitable for distinguishing modality types. The embeddings were used in the model by conditioning intermediate activations of the classification networks in FiLM layers. Through FiLM, the intermediate outputs of selected layers of a neural network are transformed with scale and bias parameters. These parameters are mapped from a given conditioning embedding, in our case the filter embeddding. The mapping from embedding vector to a scale and bias parameter is learned during model training, as part of the same back-propagation chain as the rest of the model training. As a result, the scale and bias parameters learn to modulate the model operations according to the embedding vector that describes the image modality. We added two FiLM layers in each network: one between the stem and Block 1 and one layer between Block 2 and Block 3. The adjusted pipeline that implements this modality-aware classifier is depicted in Figure S5. The mixed-modality classifier did not perform well on the validation set and therefor we did not experiment with FiLM for the multimodal classifiers.



(a)



(b)

Figure S4: TSNE projection of filter embedding of (a) images of all modalities and (b) of the modalities used in the classification experiments. Modalities can be distinguished by filling colour. AmyloidPET status is distinguished by edge colour, with which we want to display that the embedding space is not correlated to AmyloidPET status. TSNE embeds the points from a higher dimension to a lower dimension trying to preserve the neighborhood of that point by minimizing the Kullback–Leibler divergence between distributions with respect to the locations of the points in the map. Projections (a) and (b) come from TSNE calculations with different sets of modalities which results in different projections of the OCT-SMAC, OCT-BONH, OCT-BMAC and FAF points on the 2D spaces.

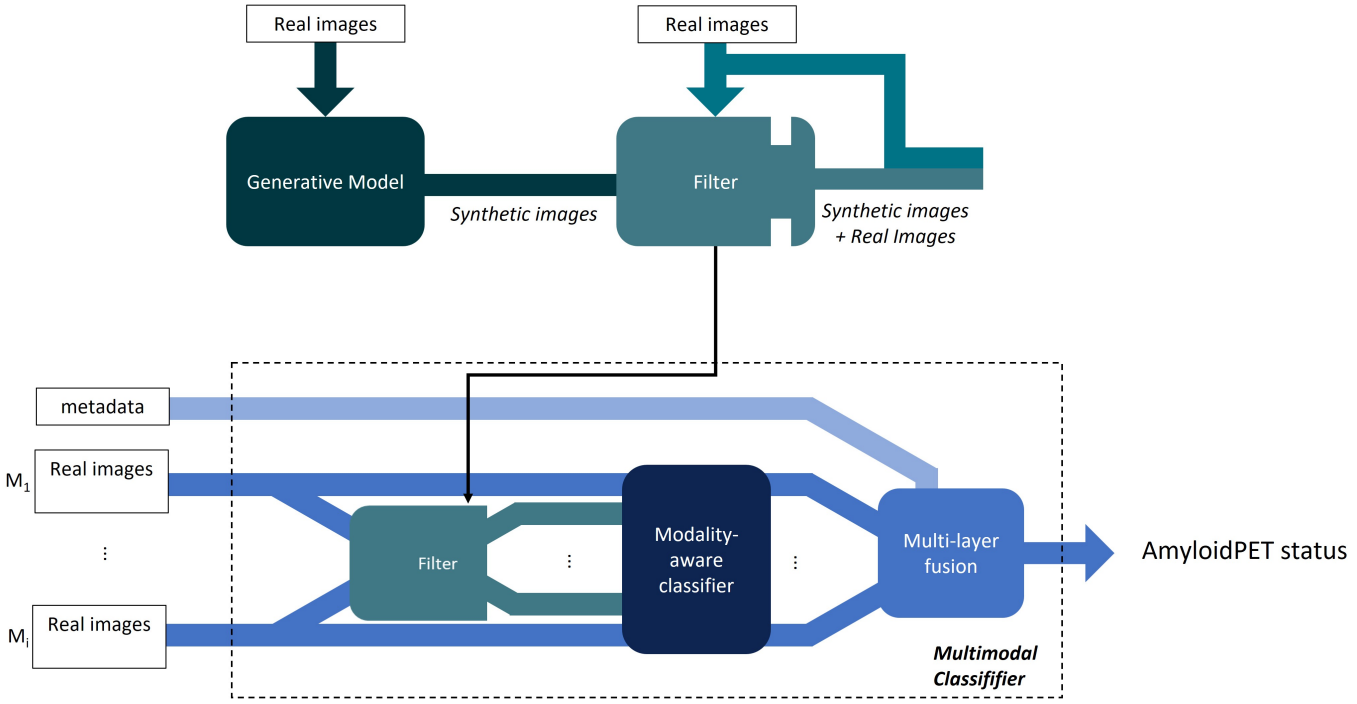


Figure S5: Proposed pipeline for using a modality-aware classifier in the multimodal classifier. (Top): A DDPM is trained to create synthetic images. The synthetic images for which the filter can recognize the modality were included. (Bottom): The embedding vectors generated by the filter model are given as input to the modality-aware model one by one to generate a prediction for each modality seperately. These binary prediction probabilities (between 0-1 for AmyloidPET positive) together with age (binary) and gender (scaled by 0.01) metadata is inputted to a three-layer fully connected network. Output of the model was a score between 0 and 1 for the probability of AmyloidPET positive status.

	AUROC on validation		AUPR on validation		AUROC on test		AUPR on test	
Classifier	Baseline	Pretrained	Baseline	Pretrained	Baseline	Pretrained	Baseline	Pretrained
Mixed-modality								
- without FiLM	0.5853		0.6716		0.4492		**0.5952**	
- with FiLM	0.5899		0.6858		0.4157		0.5544	

Table S2: AUROC and AUPR for classification by one mixed-modality classifier that predicts AmyloidPET status for all types of images. For comparison we implemented the mixed-modality classifier without FiLM layer and thus this model has no context on the modality of an input image. This network slightly underperforms on the test set compared to the classifier that is modality aware.

4) Unimodal classifiers: Starting point for the unimodal classifiers is a EfficientnetB0 network without pretrained weights. The final layer was changed to a binary prediction layer. Classifiers were trained for 200 epochs with various initial learning rates, a step LR scheduler and Adam optimizer without weight decay. Binary cross entropy was compared with Focal Loss in 20-trial hyperparameter experiments. As the loss stopped converging throughout the 200 epochs we would select model checkpoints taken from the best epoch.

5) Multimodal classifiers: The multimodal classifier is a fusion of the unimodal classifiers through fully connected layers to fuse the unimodal predictions and metadata. We experimented with various fusion methods, none of these fusion methods outperformed our selected approach (i.e. heterogenuous fusion). The methods we experimented with are:

- Fusion of embedding vectors produced by the unimodal classifiers by transforming the vectors into 2d arrays and applying convolution operations on the resulting multi-channel vectors and several fully connected layers at the end.
- Fusion of embedding vectors produced by the unimodal classifiers through a multiple fully connected layers.
- Early or late fusion of metadata with classification features through either incorporation of metadata early in the fully connected network or late in the fully connected network.

As the loss on the validation set stopped converging but would not deteriorate throughout the 200 epochs, we would select the model checkpoint from the last epoch.

References

[1] A. Katsimpris, A. Karamaounas, A. M. Sideri, J. Katsimpris, I. Georgalas, and P. Petrou, "Optical coherence tomography angiography in alzheimer's disease: A systematic review and meta-analysis," *Eye*, vol. 36, no. 7, pp. 1419–1426, Jun. 2021, ISSN: 1476-5454. DOI: 10.1038/s41433-021-01648-1.

[2] Q. Jin, Y. Lei, R. Wang, H. Wu, K. Ji, and L. Ling, "A systematic review and meta-analysis of retinal microvascular features in alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 13, Jun. 2021. DOI: 10.3389/fnagi.2021.683824.

[3] G. Ashraf, M. McGuinness, M. A. Khan, C. Obtinalla, X. Hadoux, and P. van Wijngaarden, "Retinal imaging biomarkers of alzheimer's disease: A systematic review and meta-analysis of studies using brain amyloid beta status for case definition," *Alzheimer's amp; Dementia: Diagnosis, Assessment amp; Disease Monitoring*, vol. 15, no. 2, Apr. 2023, ISSN: 2352-8729. DOI: 10.1002/dad2.12421.

[4] W. H. L. Pinaya, M. S. Graham, E. Kerfoot, *et al.*, "Generative ai for medical imaging: Extending the monai framework," 2023. DOI: 10.48550/ARXIV.2307.15208.