Multi-task learning for activity detection in neovascular age-related macular degeneration

Murat Seçkin Ayhan[1,§], Hanna Faber[1,2,§], Laura Kühlewein[1,2], Werner Inhoffen[2], Gulnar Aliyeva[2], Focke Ziemssen[2,3], and Philipp Berens[1,4,‡]

[1]*Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany*
[2]*University Eye Clinic, University of Tübingen, Tübingen, Germany*
[3]*University Eye Clinic, University of Leipzig, Leipzig, Germany*
[4]*Tübingen AI Center, Tübingen, Germany*
[‡]*Corresponding author: philipp.berens@uni-tuebingen.de*
[§]*Equal contribution*

August 2, 2022

Meeting Presentation: The manuscript is accepted for presentation (ID PDo04-01) at the 120. Congress of the DOG (Deutsche Ophthalmologische Gesellschaft), 29.09.–02.10.2022, Berlin, Germany

Financial support: Financial support was provided by German Ministry of Science and Education (BMBF) through the Tübingen AI Center (FKZ 01IS18039A) and the German Science Foundation for funding through a Heisenberg Professorship (BE5601/4-2) and the Excellence Cluster "Machine Learning — New Perspectives for Science" (EXC 2064, project number 390727645), Junior Clinician Scientist Program of the Faculty of Medicine, Eberhard Karls University of Tübingen, Germany (application number 463-0-0) (HF) and the Novartis AG. The sponsor or funding organization had no role in the design or conduct of this research.

Conflict of Interest: PB holds shares of eye2you GmbH. FZ has received consulting fees from Allergan, Bayer HealthCare, Boehringer-Ingelheim, Novo Nordisk, MSD and Novartis and speaker fees from Alimera, Allergan, Bayer HealthCare and Novartis. FZ was involved in research funded by grants from Bayer Healthcare, Biogen, Clearside, Ionis, Kodiak, Novartis, Ophtea, Regeneron and Roche/Genentech. LK receives, via third-party accounts of the University Eye Hospital, research funding and honoraria from Novartis and research funding from the Tistou and Charlotte Kerstan Foundation. HF received medical training event costs from Novartis. MSA, WI and GA declare no competing interest.

Short title: Multi-task learning for activity detection in nAMD

Word Count: 4810

Section Code: RE

Keywords: machine-learning, age-related macular degeneration, anti-VEGF treatment, multi-task learning, saliency maps

Address for preprints: Prof. Dr. rer. nat. P. Berens, Werner Reichardt Centre for Integrative Neuroscience (CIN) Institute for Ophthalmic Research, University of Tübingen, Otfried-Müller-Str. 25, D-72076 Tübingen, Germany. Phone: +49 (0)7071 29-88833, philipp.berens@uni-tuebingen.de

Abstract

Purpose:

Comparison of performance and explainability of a multi-task convolutional deep neuronal network to single-task networks for activity detection in neovascular age-dependent macular degeneration.

Methods:

From n = 70 patients (46 female, 24 male) who attended the University Eye Hospital Tübingen 3762 optical coherence tomography B-scans (right eye: 2011, left eye: 1751) were acquired with Heidelberg Spectralis, Heidelberg, Germany. B-scans were graded by a retina specialist and an ophthalmology resident, and then used to develop a multi-task deep learning model to predict disease activity in neovascular age-related macular degeneration along with the presence of sub- and intraretinal fluid. We used performance metrics for comparison to single-task networks and visualized the DNN-based decision with t-distributed stochastic neighbor embedding and clinically validated saliency mapping techniques.

Results:

The multi-task model surpassed single-task networks in accuracy for activity detection and visualizations via t-distributed stochastic neighbor embedding and saliency maps highlighted that multi-task networks' decisions for activity detection in neovascular age-related macular degeneration are indeed based on the presence of sub- and intraretinal fluids, the optical coherence tomography characteristics used for treatment decision in clinical routine.

Conclusions:

Multi-task learning increases the performance of neuronal networks for predicting disease activity, while providing clinicians with an easily accessible decision control, which resembles human reasoning.

1 Introduction

Neovascular age-related macular degeneration (nAMD) is a sight-threatening disease affecting the elderly and among the most common causes of vision loss worldwide [38, 62, 17]. Among the basic features of nAMD are subretinal (SRF) and intraretinal fluid (IRF), which serve as surrogate markers of nAMD activity and can be monitored using optical coherence tomography (OCT) [52, 42] (Fig. 1).

In nAMD, increased levels of vascular endothelial growth factor (VEGF) lead to formation of new vessels from the choroidal and/or retinal vasculature. If leakage from these vessels exceeds local clearance rates, liquid builds up, leading to IRF and SRF [52]. IRF is assumed to originate from vascular leakage from intraretinal neovasculaturisation and/or retinal vasculature or from diffusion through the outer retina due to changes within the external limiting membrane [52]. In contrast, SRF formation likely results from malfunction of the retinal pigment epithelium with reduced removal rates [52]. Due to the partially different pathophysiology, IRF and SRF can occur simultaneously as well as independently from each other [52, 41]. Furthermore, different lesion characteristics based on presence of IRF and SRF have been observed to indicate the potential functional outcome [59]. For both fluid compartments, automatic detection and quantification have been shown to be possible [47].

Treatment with intravitreal anti-VEGF agents can efficiently restore the balance between liquid formation and retinal removal and are standard of care for nAMD, when IRF or SRF is detected via OCT [42]. Since delay of treatment is associated with vision loss [25, 56, 3], treatment has to be initiated promptly. Also, therapy monitoring using OCT has to take place on up to four-weekly basis in some cases until the end of life. Due to this high frequency of visits, the therapy has put a considerable burden on patients, their families and ophthalmological care since its initial approval in 2006 [14, 2, 43, 51]. Additionally, a future increased need for AMD care has to be expected, since the number of patients suffering from AMD are thought to rise from 196 million in 2020 to 288 million in 2040 [62]. Hence, automated solutions making the diagnostic processes more efficient have considerable appeal. For example, deep neural networks (DNNs) have been used for automatic referral decisions [15] and predicting disease conversion to nAMD [64]. Automated algorithms have been shown to detect both SRF and IRF more reliably than retinal specialists especially in less conspicuous cases [26]. DNNs have been shown to be able to accurately detect retinal fluids caused by various diseases with OCT scans acquired from different devices [47, 26]. Ideally, such automated tools serve to support retinal specialists in their decision making. To this end, computational tools need to explain their decisions and communicate their uncertainty to the treating ophthalmologist [20, 21]. In collaboration, a retina specialist assisted by an artificial intelligence (AI) tool can outperform the model alone, e.g. for the task of diabetic retinopathy grading [46].

Here, we develop a convolutional deep learning model based on the concept of multi-task learning [10, 60], that simultaneously detects IRF, SRF and disease activity in nAMD. The localization of the fluid plays a decisive role in the treatment outcome [48, 31, 45] with the simultaneous presence of IRF and SRF being associated with the worst prognosis [56]. To this end, we visualize the representation driving the DNN-based decisions using t-distributed stochastic neighbor embedding (t-SNE) [58, 28] and investigate the model's decisions using clinically validated saliency mapping techniques [6]. Thus, our work provides an interpretable tool for the ophthalmologist to rapidly access the neural network's decision process both on a population-based as well as an individual-patient level as a prerequisite for clinical application.

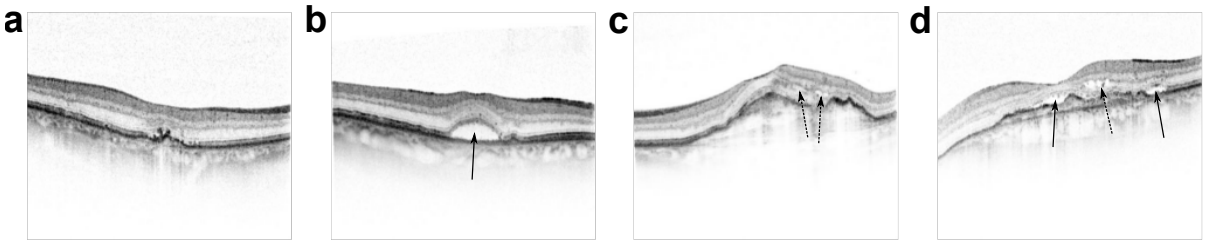

Figure 1: Exemplary retinal images (B-scans) with neovascular age-related macular degeneration (nAMD). Solid and dotted arrows indicate subretinal and intraretinal fluid, respectively. **(a)**: no nAMD activity. **(b)**: nAMD activity due to subretinal fluid (SRF) (arrow). **(c)**: nAMD activity due to intraretinal fluid (IRF) (dotted arrow). **(d)**: nAMD activity due to both SRF (arrow) and IRF (dotted arrow).

Table 1: OCT Data distribution of subretinal fluid (SRF), intraretinal fluid (IRF) and active nAMD in B-Scans in training, validation and test sets, respectively. Absolute and relative numbers are shown.

	Training			Validation			Test		
	Subretinal fluid	*Intraretinal fluid*	*Active nAMD*	*Subretinal fluid*	*Intraretinal fluid*	*Active nAMD*	*Subretinal fluid*	*Intraretinal fluid*	*Active nAMD*
Yes	639 (0.232)	286 (0.104)	848 (0.308)	69 (0.170)	58 (0.143)	101 (0.248)	161 (0.267)	153 (0.253)	269 (0.445)
No	2112 (0.768)	2465 (0.896)	1903 (0.692)	338 (0.830)	349 (0.857)	306 (0.752)	443 (0.733)	451 (0.747)	335 (0.555)

2 Methods

2.1 Data Collection

70 consecutive patients (46 females, 24 males) with nAMD at least in one eye, seen by GA in the Macula clinic at the University Eye Hospital Tübingen were included in this study. Exclusion criteria were any other cause of neovascularisation, any coexisting retinal pathology (e.g. epiretinal membrane, macular hole, diabetic retinopathy), glaucoma and media opacity that did not allow to take images of sufficient quality.

3762 B-scans (2011 right eye, 1751 left eye) of 440 x 512 pixels with Heidelberg Spectralis OCT (Heidelberg Engineering, Heidelberg,Germany) were included in the study. A retina specialist (WI) assessed the presence of IRF and SRF as well as disease activity on each individual image (Fig. 1). Disease activity was also graded by an ophthalmologist resident (GA). The degree of inter-annotator agreement according to Cohen's kappa statistic was 0.86. B-scans were assigned to a training, validation or test set (Table 1), where care was taken to assign all images from one patient to one of the sets to avoid information leakage. The relationship between the nAMD activity and SRF or IRF were captured by Cohen's kappa statistic (Table 2), which indicated the independence of the two retinal fluid types. Ethical approval was granted by the local institutional ethics committee of the University of Tübingen. Due to the retrospective character of the study, the requirement for patient consent was waived by the ethics committee. The study was conducted in accordance with the tenets of the Declaration of Helsinki.

Table 2: Agreement of task-specific labels across training, validation and test sets, measured via Cohen's kappa statistic, which is essentially a number between -1 and 1. While 1 indicates a full agreement, lower scores mean less agreement. Negative scores indicate disagreement.

	Training			Validation			Test		
	Subretinal fluid	*Intraretinal fluid*	*Active nAMD*	*Subretinal fluid*	*Intraretinal fluid*	*Active nAMD*	*Subretinal fluid*	*Intraretinal fluid*	*Active nAMD*
Subretinal fluid	n.a.	-0.02	0.79	n.a.	0.26	0.75	n.a.	-0.02	0.59
Intraretinal fluid	-0.02	n.a.	0.37	0.26	n.a.	0.65	-0.02	n.a.	0.57

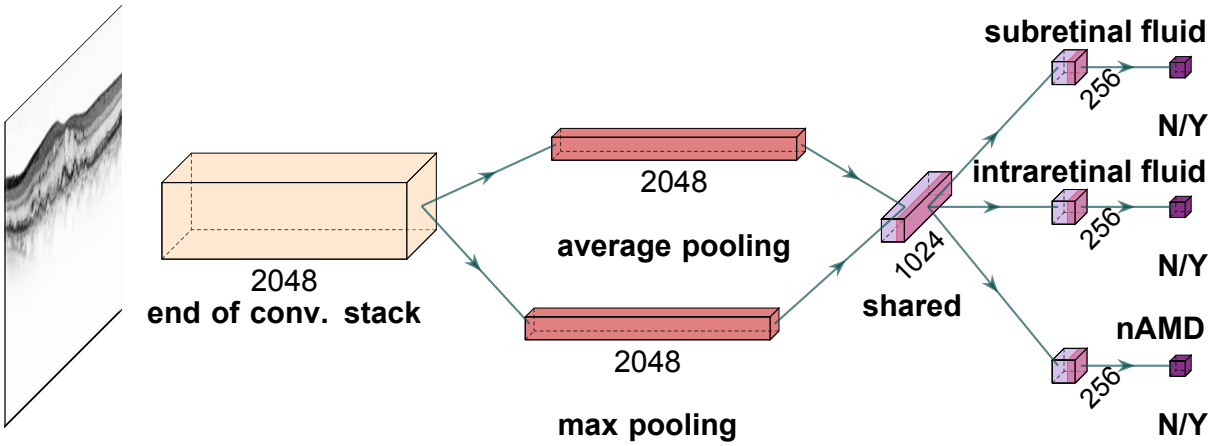

Figure 2: A deep neural network for simultaneous detection of subretinal and intraretinal fluid as well as the nAMD activity from OCT B-scans. Given a B-scan, convolutional stack of the InceptionV3 architecture extracts 2048 feature maps. These are average and max pooled, and fed into a fully connected (dense) layer with 1024 units for shared representation. Then, task-specific heads specialize into individual tasks and single units with sigmoid function achieve binary classification based on 256 task-specific features.

2.2 Diagnostic Tasks, Network Architecture and Model Development

We developed a multi-task DNN to detect the presence of SRF and IRF as well as the nAMD activity from OCT B-scans. While these tasks could have been performed by different networks trained for each particular task, we adopted a multi-task learning approach and trained a single network to perform these tasks simultaneously (Fig. 2). As backbone, we used the InceptionV3 architecture [54] via Keras [13]. The backbone was pretrained on ImageNet [44] for 1000-way classification via a softmax function. We used the InceptionV3 DNN's convolutional stack as is but adapted the deeper layers to our multi-task scenario as follows. First, we linked max pooling and average pooling to the end of convolutional stack. They were followed by a dense layer, which yielded a shared representation with 1024 features. Following the shared representation we added task-specific heads with 256 units. These specialized into their respective tasks and extracted their own 256-dimensional feature representations. Then, task-specific binary decisions were achieved by single units equipped with sigmoid functions.

We trained our networks with equally weighted cross-entropy losses for all tasks on the training images: $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$, where $\mathbf{y}_n$ was a vector of binary labels indicating nAMD activity and the pres-

ence of IRF or SRF in an image $\mathbf{x}_n$. Parameterized by θ, a DNN $f_\theta(\cdot)$ was optimized with respect to the total cross-entropy on the training data: $\mathcal{L}(\mathcal{D}, f_\theta(\cdot)) = \frac{1}{N}\sum_{n=1}^{N} l(\mathbf{y}_n, f_\theta(\mathbf{x}_n))$, where $l(\mathbf{y}_n, f_\theta(\mathbf{x}_n)) = -\sum_{t=1}^{T} \mathbf{y}_{n,t} \log p_{n,t} + (1-\mathbf{y}_{n,t}) \log (1-p_{n,t})$, p_n was a list of probabilities estimated via the sigmoid functions for different tasks and t was an index into T tasks. For $T = 1$, multi-task learning reduced to single-task learning. To address the class imbalance in data (Table 1), we used random oversampling (see Section 2.2.2 for details). We also used Stochastic Gradient Descent (SGD) with Nesterov's Accelerated Gradients (NAG) [35, 53], minibatch size of eight, a momentum coefficient of 0.9, an initial learning rate of $5 \cdot 10^{-4}$, a decay rate of 10^{-6} and a regularization constant of 10^{-5} for 120 or 150 epochs (see Section 2.2.1 for longer training). During the first five epochs, the convolutional stack was frozen and only dense layers were trained. Then, all layers were fine-tuned to all tasks. The best models were selected based on total validation loss after each epoch and used for inference on the test set.

2.2.1 Data augmentation and preprocessing

First, we used *mixup* [66] for data augmentation during training. Mixup generates artificial examples through the convex combinations of randomly sampled data points. We adapted *mixup* to our multi-task learning scenario as follows:

$$\hat{\mathbf{x}} = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \quad \hat{\mathbf{y}} = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j, \quad \lambda \in [0,1]. \tag{1}$$

Mixing was controlled by $\lambda \sim Beta(\alpha, \alpha)$, where $\alpha \in (0, \infty)$. For $\alpha = 0$, λ is either 0 or 1, and there is no mixing. Typical values to enable mixing are in $[0.1, 0.4]$. While large values may lead to underfitting, longer training aids in mixing for large α [66]. We used 0, 0.05 0.1, and 0.2 for α and trained networks for 120 epochs when not mixing and 150 epochs when mixing. Also, to allow for a warm-up period when mixing [66], we set $\alpha = 0$ for the first five epochs.

As a second step in data augmentation, we applied common data augmentation operations such as adjustment of brightness within $\pm 10\%$, horizontal and vertical flipping, up and down scaling within $\pm 10\%$, translation of pixels horizontally and vertically within ± 30 positions and random rotation within ± 45 degrees. After all data augmentation operations, we used an appropriate preprocessing function[1] from the Keras API [13].

2.2.2 Quantification of uncertainty via *mixup* and Deep Ensembles

Quantification of diagnostic uncertainty is crucial for treatment decisions. With a proper management of uncertainty, diagnostic errors, delays or excess healthcare utilization can be minimized [8]. However, DNNs are typically overconfident about their predictions and they do not generate well-calibrated and reliable uncertainty estimates for their decisions [22, 27, 29, 32, 16]. *mixup* [66] improves the calibration of DNN outputs by smoothing labels through their convex combinations (Eq. 1) [55]. On top of *mixup*, we used Deep Ensembles [29] consisting of multiple DNNs. These DNNs are randomly initialized and then allowed to follow different optimization trajectories to explore different modes in function space [29, 18]. The ensemble, then, exploits the diversity of multiple predictors in decision-making and improves upon the single network performance both in accuracy and calibration, even with small numbers of DNNs trained on standard datasets [29, 18, 39]. Also in a DR detection scenario [5], an ensemble of three DNNs already performed well in both aspects.

[1] `keras.applications.inception_v3.preprocess_input`

Table 3: Accuracy of ensembles for various degrees of mixing (indicated by α). Gray row indicates the ensemble of choice for further analysis based on the validation performance for the activity detection task.

	Single task								
	Training			Validation			Test		
	Subretinal fluid	*Intraretinal fluid*	*Active nAMD*	*Subretinal fluid*	*Intraretinal fluid*	*Active nAMD*	*Subretinal fluid*	*Intraretinal fluid*	*Active nAMD*
$\alpha = 0$	1.000	1.000	1.000	0.988	0.971	0.958	0.924	0.950	0.914
$\alpha = 0.05$	0.983	0.994	0.975	0.971	0.963	0.951	0.906	0.919	0.909
$\alpha = 0.1$	0.978	0.994	0.948	0.948	0.919	0.929	0.868	0.891	0.856
$\alpha = 0.2$	0.983	0.991	0.851	0.975	0.946	0.853	0.881	0.909	0.702
	Multiple tasks								
$\alpha = 0$	1.000	0.995	0.998	0.973	0.973	0.961	0.914	0.935	0.940
$\alpha = 0.05$	0.999	0.998	1.000	0.971	0.971	0.966	0.917	0.937	0.942
$\alpha = 0.1$	1.000	0.997	0.998	0.983	0.968	0.966	0.916	0.957	0.939
$\alpha = 0.2$	1.000	0.998	1.000	0.971	0.966	0.966	0.894	0.937	0.906

Using the network architecture, hyperparameters and training procedures described above, we constructed our ensembles with three DNNs. During their training, we also used oversampling with a twist. For each DNN, we oversampled training images with respect to a particular task's labels. This enabled DNNs to train on a balanced dataset for their respective tasks while also learning about other tasks, even though the data was not balanced for the other tasks. Overall, this contributed to the diversity of DNNs, which is essential for ensemble models. DNNs were further diversified by the randomness in the initialization of dense layers, shuffling of training examples as well as mixing and data augmentation. In the end, we used the ensemble's mean output for predictions and quantified uncertainty in terms of entropy, given the average predictive probabilities.

2.3 Low-dimensional embedding of images

We used t-SNE [58] to obtain further insights into the decision-making process of our ensemble model. t-SNE is a non-linear dimensionality reduction method, that embeds high-dimensional data points into a low-dimensional space. To evaluate ensemble-based representations, we concatenated features from ensemble members' predetermined read-out layers and performed t-SNE based on them, embedding each B-scan into the two-dimensional plane. We used *openTSNE* [40] with PCA initialization to better preserve the global structure of the data and improve the reproducibility [28]. We used a perplexity of 200 for 1500 iterations with an early exaggeration coefficient of 12 for the first 500 iterations, according to best-practice strategies [28]. Similarities between data points were measured by Euclidean distance in the feature space.

2.4 Saliency Maps

We used Layer-wise Relevance Propagation (LRP) [7] to compute saliency maps, to highlight the regions in the OCT images which contributed to the DNN decisions. We have recently shown that a propagation rule known as *LRP-PresetBFlat* performs best in obtaining clinically relevant saliency maps from InceptionV3 networks trained to detect active nAMD from OCT B-scans [6]. Using this rule, we created three saliency

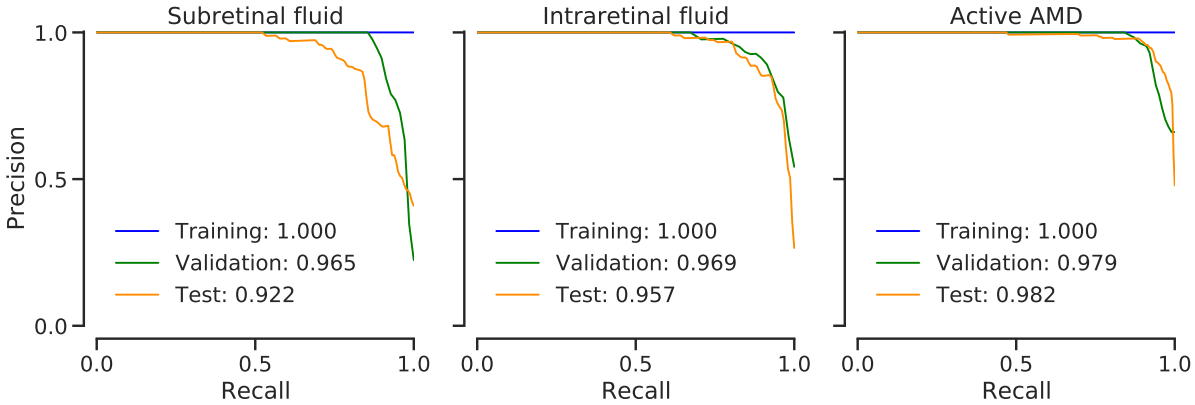

Figure 3: Precision-recall curves for the selected ensemble model. Area under the curve (AUC) values given for partitions also summarize the overall performance into one number (higher is better).

maps for each OCT slice, namely, one for each task: subretinal (cyan), intraretinal (magenta) and diesease activity in nAMD (yellow) (Fig. 6). To improve the visualization of the salient regions, saliency maps were postprocessed and the maps of each task were combined into one [6]. Saliency maps were only shown for predictions with an estimated probability greater than 0.5 since previous work has shown, that especially in absence of disease, saliency maps can lead physicans to overdiagnosis [46].

3 Results

We developed an ensemble of three multi-task DNNs to simultaneously detect SRF, IRF and activity of nAMD on OCT B-scans (Fig. 1). Each DNN consisted of a shared convolutional core combined with pooling operations which yielded a shared representation (Fig. 2). This representation served as the basis for the decision of the three task heads. The idea behind this approach is that the DNN can benefit from the shared *general purpose* representation induced by combining information from different tasks. We first investigated the performance of the multi-task model in the three tasks. To this end, we compared the multi-task model with more specialized single task models, where we constructed three DNNs, one for each task, which did not share any representation but were trained independently. All DNNs were trained on a training set acquired during clinical routine at the University Eye Hospital in Tübingen (see Table 1 and Methods). We selected the multi-task model with the best accuracy for the activity detection task on the validation set and report accuracy values computed on an independent test set (Table 3). Overall, we found the multi-task model to be well calibrated on the test set (Adaptive expected calibration error [16] of 0.0147 for SRF, 0.0104 for IRF and 0.0263 for active nAMD), suggesting that uncertainty reported for the decisions of the DNN-based ensemble reflect the true model uncertainty.

We found that the performance of the multi-task model surpassed the single-task model performance in disease activity detection, reaching an accuracy of 94.2 % for the multi-task model vs. 91.4% for the single task model (Table 3, Fig. 3). Interestingly, this multi-task model optimized for AMD activity detection performed slightly worse than the single-task models for the two tasks of detecting SRF and IRF (SRF: accuracy of 0.917 vs. 0.924 for multi-task vs. single-task; IRF: 0.937 vs. 0.950). This suggests that the representations learned by the multi-task DNNs are indeed a trade-off between achieving high performance on all three tasks, and as a result on activity detection, but somewhat sacrifice single-task detection performance.

We thus further studied the representations learned by the multi-task model to gain insight into its

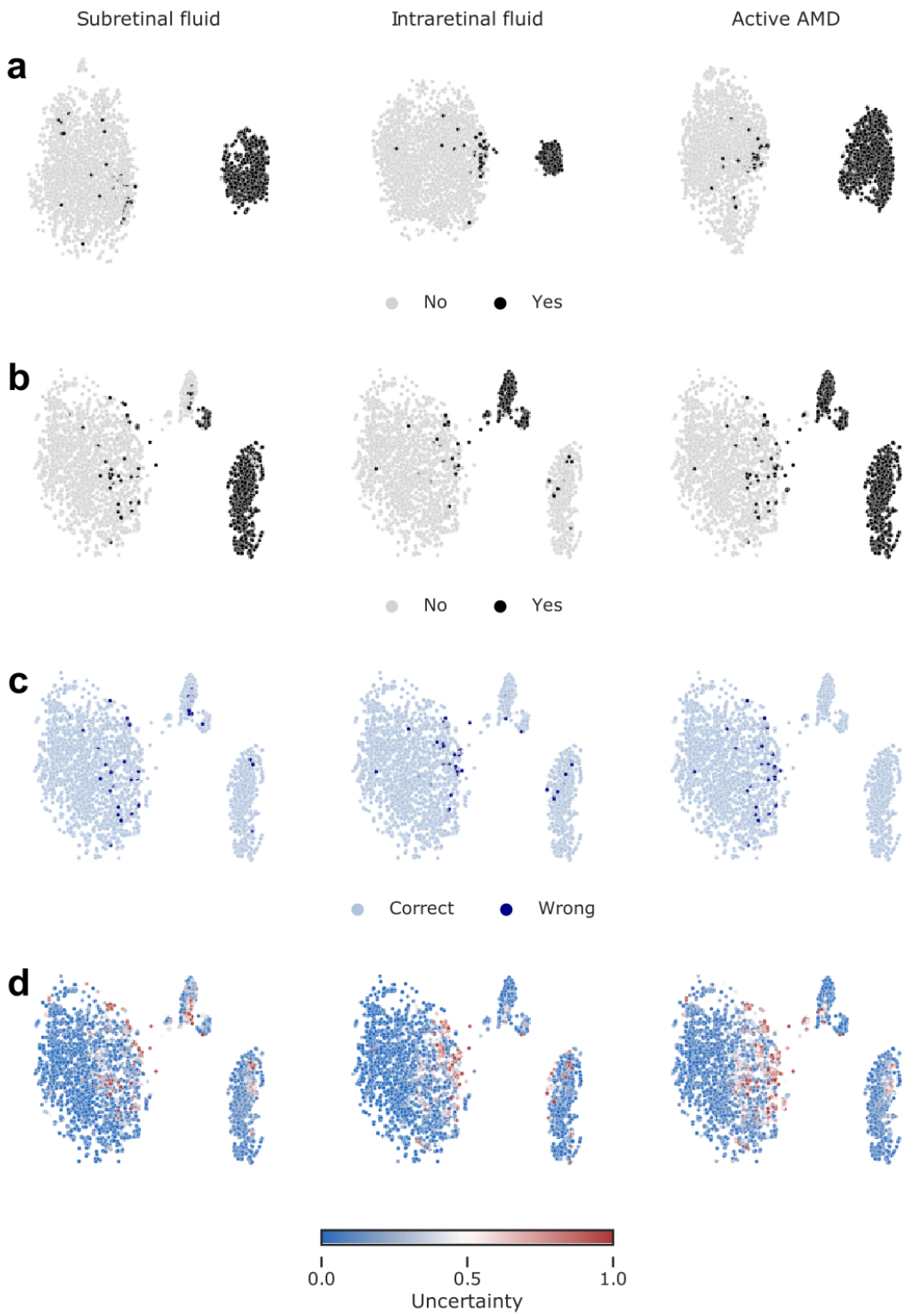

Figure 4: Visualization of data via t-SNE of ensemble-based representations. **(a)** Low dimensional embedding of images based on the penultimate layer features from single-task networks. Training, validation and test data aligned together and colored with respect to the task-specific labels. **(b)** Same as in (a) but w.r.t. features from the shared representation layer of multi-task networks. **(c)** Same map as in (b) but colored w.r.t. correct and wrong predictions. **(d)** Same map as (b) but colored w.r.t. uncertainty min-max normalized to $[0, 1]$.

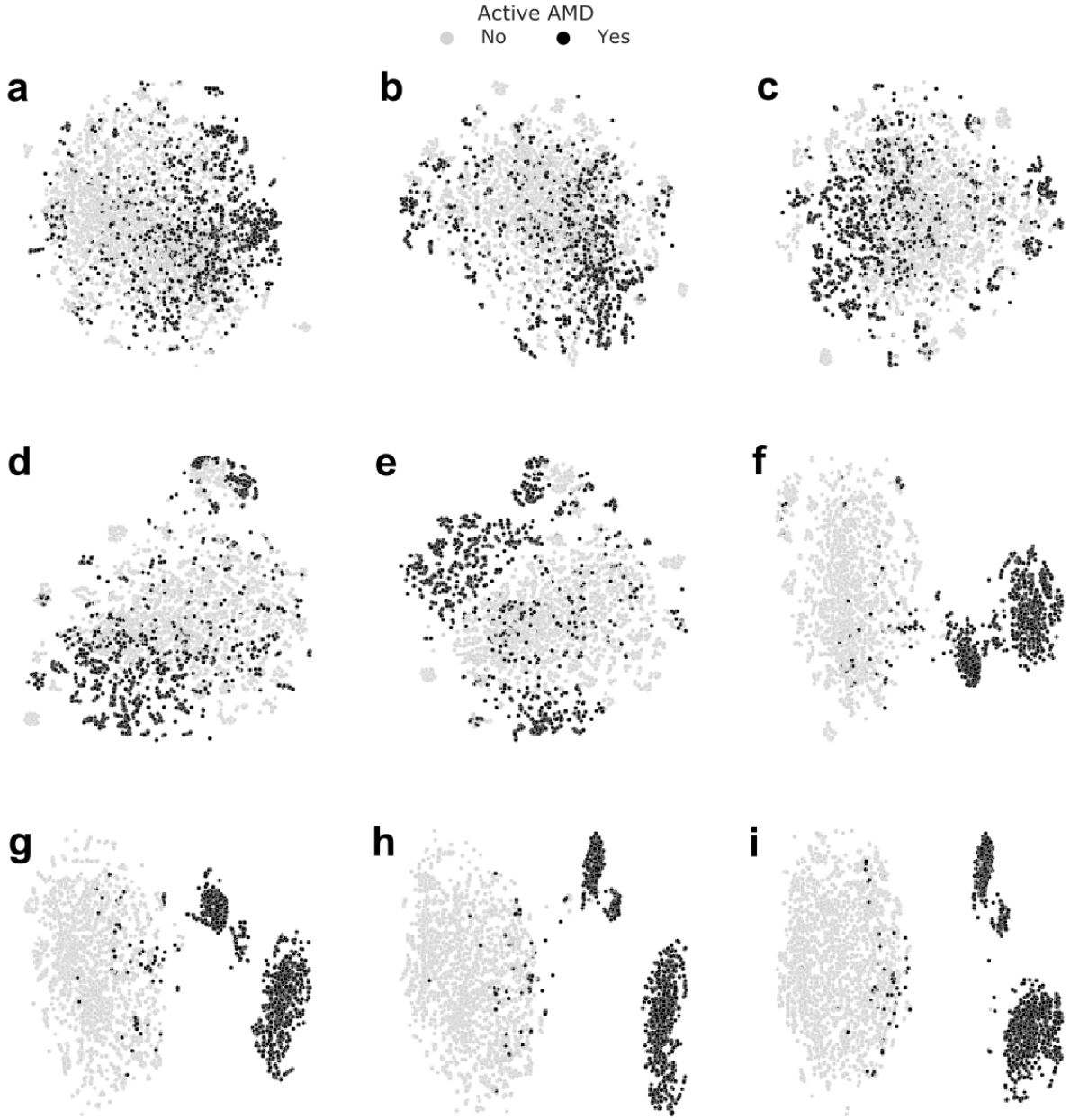

Figure 5: Layer-wise visualization of data via t-SNE. Starting just before the first Inception module **(a)** and reading out feature representations yielded by every other module **(b-f)** along with the last Inception module **(g)**, the shared representation layer **(h)** and the nAMD activity detection head's penultimate layer **(i)**, we performed t-SNE with the aforementioned settings. Useful representations emerged towards the end of convolutional stack and the task-specific representation allowed the best separation of nAMD active cases from those inactive. Exact read-out locations can be found in Appendix (Fig. 7).

decision making-process. To this end, we extracted the representation of individual OCT scans at various levels of processing throughout DNNs (Fig. 7) and created two-dimensional embeddings of these via t-SNE (Fig. 4 and 5). In these visualizations, each point in two-dimensions corresponds to an individual OCT scan. OCT scans, which are similar to each other according to the learned representation, are mapped to nearby points. While t-SNE representations are generally useful for exploratory analysis if some guidelines are followed, one should be careful interpreting distances in the embedded space — e.g. the size of the white space between clusters is rather an effect of the algorithm not the data [28, 9].

We first investigated the final representation based on which the single task DNNs and the individual task heads of the multi-task DNNs make their decision (Fig. 4). We colored the individual points according to whether the OCT scan was labeled as containing evidence for SRF or IRF, as well as overall AMD activity. Reflecting the high task accuracy, most inconspicuous OCT scans were placed in a clearly separated island, clearly distinct from the OCT scans with any of the disease labels (Fig. 4a, b). For the single-task DNNs, additional well-separated clusters were found, indicating the learned task-label (Fig. 4a). For example, OCT scans with SRF present formed a single cluster, clearly distinct from the OCT scans without this label. Interestingly, this was also the case for the active AMD task, for which no clearly distinct subclusters could be seen.

In contrast, for the embedding extracted from the shared representation of the multi-task model, OCT scans labeled with SRF formed a well-separated cluster at the bottom right, as did scans with IRF labels at the top right (Fig. 4b). Interestingly, there was a small cluster in between these two which contained scans labeled with both. Consequently, OCT scans labeled with active AMD encompassed all three of these major disease related clusters, suggesting the multi-task DNNs indeed learned separate representations of the two fluid types which were then used by the individual task heads. The few incorrectly classified OCT scans could be found within their clusters to be placed close towards other clusters (Fig. 4c) in areas where we also found examples with high classifier uncertainty. Thus, decisions were more uncertain e.g. for inactive OCT scans that were more similar to OCT scans with signs of sub- or intraretinal fluid, sometimes leading the DNN to incorrect decisions (Fig. 4d). In clinical application of such an algorithm, high uncertainty could thus be used to select individual B-scans warranting further scrutiny through experienced clinicians.

We next studied how the multi-task representation emerged through processing in the network (Fig. 5). While in the initial layers data points representing active nAMD were still uniformly distributed (Fig. 5, a-c), a clear separation of active nAMD cases developed gradually in later layers of the DNN (Fig. 5, d-g), leading to best separation in the shared representation (Fig. 5, h). The decision head for active AMD refined this representation only very little (Fig. 5, i). This analysis is in agreement with previous work showing that lower layers in DNNs typically extract very general task-independent image features that are gradually refined to disentangle the representation of the task-relevant image classes [65, 19].

We next analyzed if well-known saliency maps can also be used in case of the multi-task DNNs to identify which image regions in individual OCT scans were relevant for the decision. Specifically, we were interested in whether the saliency maps for the subtasks of SRF and IRF detection obtained from the multi-task model allowed reasoning about evidence specific to these tasks. To this end, we generated saliency maps on four exemplary OCT scans using LRP [7] (Figure 6). For each OCT scan, we generated three maps, one for each of the three tasks, propagating the task-information back from the task head.

We first analyzed an OCT scan with clearly active AMD and both SRF and IRF present (Figure 6a). The active AMD saliency map focused on intraretinal fluids, which were also clearly visible in the task-specific saliency map, and faintly highlighted regions with SRF. The SRF saliency map, however, clearly highlighted SRF. In two further example scans with either IRF or SRF, respectively, active AMD saliency maps clearly corresponded to the individual task maps (Figure 6b,c), indicating that the saliency maps obtained from

the multi-task DNN can support clinical decision making about active AMD, but also allow clinicians to identify evidence in the relevant sub-tasks of finding SRF and IRF. We also identified rare failure cases of the obtained saliency maps (Fig. 6d): In one example, an OCT scan was falsely classified positive for SRF with a confidence of 0.614, because IRF was falsely classified as SRF. We hypothesize that the DNN misclassified the superior border of the intraretinal fluid as photoreceptor layer detached from the retinal pigment epithelium. The assumption, that the DNN primarily recognizes contrast-rich interfaces such as SRF and IRF is further supported by the false labeling of cystoid spaces within choroid in Fig. 6b and d, while in a smoother, lower-contrast choroid saliency maps do not highlight any structures (Fig. 6. This suggests that beyond such proof of principle studies, larger and more variable datasets will be needed to train multi-task DNNs to more completely rule out such artefacts.

We additionally generated saliency maps from the single task DNNs (Fig.8). Compared to the saliency maps generated from the multi-task models, those saliency maps appear slightly more defined, but highlighted similar regions, indicating that single task relevant information could be extracted from the multi-task DNN. Interestingly, Fig.8)d provides additional support for the multi-task DNNs, showing that independently trained single task DNNs can make serious mistakes in the lack of information shared between diagnostic tasks. Multi-task networks are more informed about their tasks (Fig. 6d).

4 Discussion

In this study, we developed a machine learning model based on the concept of multi-task learning to simultaneously detect SRF, IRF as well as disease activity in OCT B-scans of nAMD patients. We showed that a multi-task model, which takes the presence of IRF and SRF into account to detect disease activity in nAMD, surpassed a single task model regarding accuracy in this task. Furthermore, our visualization of the multi-task model's decision-making process via t-SNE demonstrated that in later layers of the multi-task model, inactive and active nAMD B-scans increasingly formed different clusters. Additionally, within the active AMD B-scans, we observed a growing separation in three distinct clusters, each containing OCT B-scans with either subretinal, intraretinal or both fluid types. In contrast, this separation could not be seen in the single task models of the respective tasks. Saliency maps of exemplary individual B-scans of the three tasks further corroborate that task-relevant information can be extracted from the multi-task networks, suggesting that a multi-task DNN can serve as a basis for an explainable clinical decision support system for nAMD activity.

Overall treatment burden of nAMD measured in disability-adjusted life years as well as the economic burden have decreased since the approval of anti-VEGF [63, 33]. However, there is still a high number of patients who discontinue treatment [61]. Patients named the need for assistance, either in the form of a travel companion or a family member, as the main reason for dropping out of therapy [51], and quoted traveling illness as a major reason for therapy discontinuation [24]. Additionally, recurrence of quiescent disease requiring prompt treatment is common, making life-long monitoring necessary [4]. For these reasons, automated solutions allowing monitoring close home or even at home are promising technologies to increase treatment rates, even more in a patient population, that has difficulties seeing an ophthalmologist [50, 12]: They provide easier access and reduce the disease burden on the individual [34]. Automated solutions for fluid detection have further gained popularity during the Covid-19 pandemic, which showed the devastating effects of delay or interruption in VEGF treatment of nAMD on visual function [4, 56]. However, despite promising results in laboratory settings, real-world data revealed significantly lower performance rates of home-based OCT with in particular SRF being overlooked by the system [30]. This shows the necessity of further developments on the machine learning side to guarantee reliable use, with multi-task

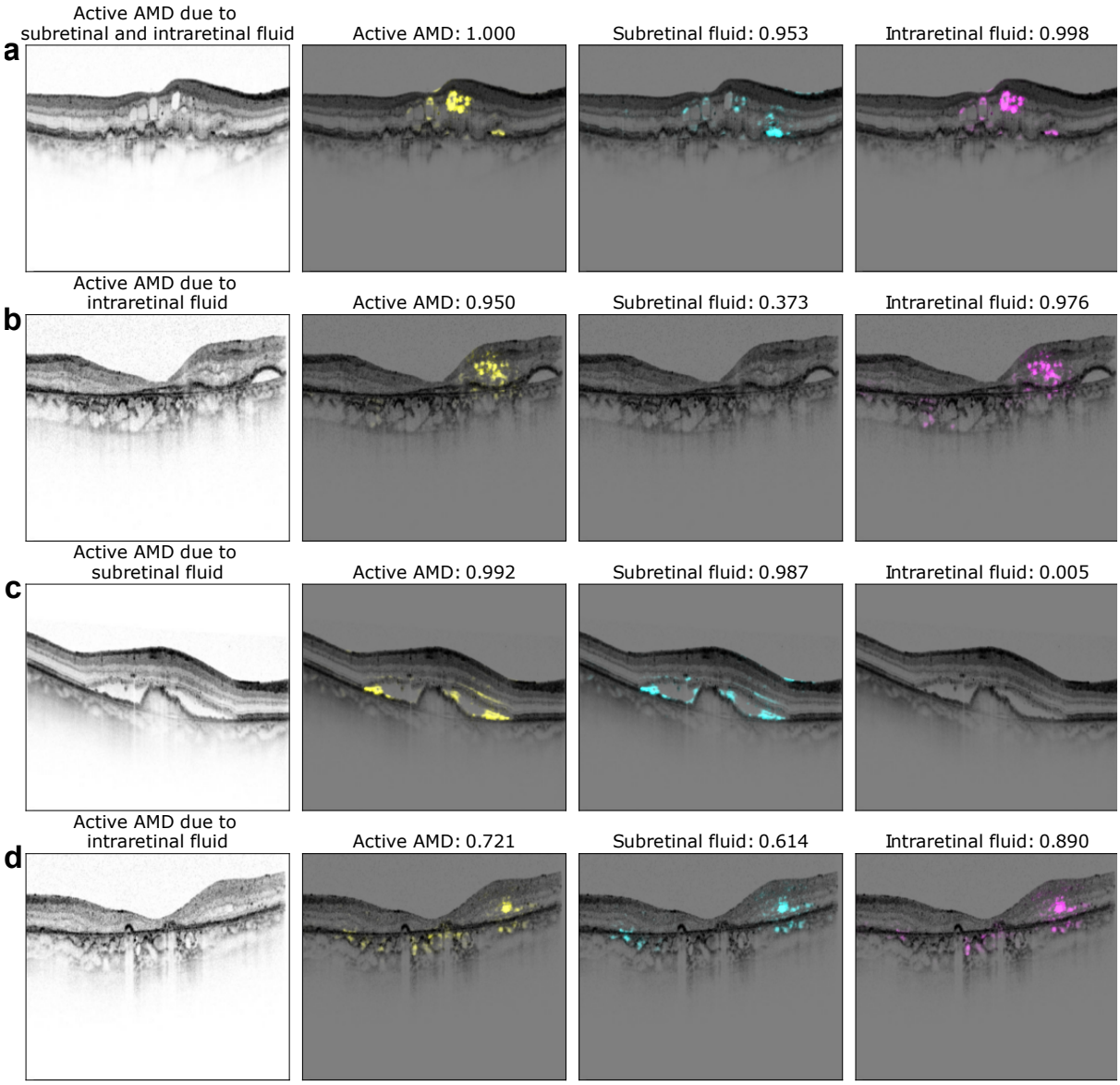

Figure 6: Exemplary saliency maps for four optical coherence tomography (OCT) images. The first column displays the OCT B-scan with the corresponding labeling of a retinal specialist. Second to fourth column show saliency maps and the network's confidence for active nAMD (yellow), subretinal fluid (SRF) (cyan) and intraretinal fluid (IRF) (magenta). Note, that saliency maps are only shown in case of confidence > 0.5. Supplementary saliency maps obtained from single-task models can also be found in Fig.8.

learning as suggested in this study being a viable option.

Beyond that, a recent meta-analysis provided evidence of varying influences of SRF and IRF on the visual outcome in nAMD patients [11]. Stable SRF might not affect visual outcome, while fluctuations in IRF during treatment seem to negatively influence visual acuity [11]. For this reason, treatment decisions in nAMD solely on a yes/no basis may not meet future treatment guidelines, which might rather require a sophisticated decision depending on the present fluid type for or against an anti-VEGF injection. Our analysis shows that this insight is not provided by single task DNNs for nAMD activity detection and thus argues for multi-task DNNs as backbone in clinician support system.

Ophthalmology has recently seen a development of various artificial intelligence systems, yet their use in clinical routine remains rare, despite a few systems now being available on the market [1, 37]. One big barrier is the concern of potential harm of the patient-physician relationship going hand in hand with the lack of trust in those systems [23]. Here, we combined multi-task DNNs with different visualization methods to give an insight into the DNNs' reasoning and increase transparency. First, we used t-SNE as visualization method for high-dimensional data [58, 28] (Fig. 4) to present the decision-making process of the model. We showed that the two-dimensional embedding of the shared representation of the multi-task model nicely separated OCT B-scans in distinct clusters according to the presence of SRF or IRF or both fluid types (Fig. 5). In comparison, single task DNNs for active nAMD detection only separated two clusters of OCT-scans, indicating absence or presence of disease (Fig. 4). The visualization of the multi-task learning via t-SNE provides thus a rationale for why certain OCT B-scans were graded as active for nAMD, which cannot be seen in a visualization of the single task algorithm (Fig. 4). It suggests that in concurrently learning basic features of nAMD activity, namely IRF and SRF [52], multi-task learning increases prediction accuracy for the main task of active AMD. Multi-task learning therefore potentially increases ophthalmologist's confidence in an algorithm since visualization via t-SNE shows, that reasoning resembles their own (Fig. 5). In the future, the multi-task system could also be extended for other signs indicative of active nAMD such as hard exsudates, pigment epithelial detachment or hyperreflective foci, which we did not study here due their comparably rare occurrence [52].

Overall high accuracy and reliability of a DNN might not be sufficient for trust and use in clinical routine, since best medical advice has to be given to an individual patient. In a second step, we therefore analyzed the multi-task model's decision on saliency maps of individual OCT-scans. Saliency maps highlight critical regions for the model's decision and thus allow a quick visual control of its reasoning. However, it needs to be kept in mind, that first various methods of saliency map generation exist with different degrees of agreement with clinical validation [6, 49, 57] and secondly, saliency maps can lead to overdiagnosis [46], while some methods have also been shown to generate maps independent of the final decision of the algorithm [36]. Therefore we only displayed saliency maps in case of a confidence of the algorithm > 0,5. Compared to saliency maps of single task DNNs, multi-task saliency maps seem to draw slightly less sharp contours, however, there is good overlap between regions used for active AMD detection and those for SRF and IRF.

Future studies will need to assess how well these multi-task learning results transfer from the data sample acquired at a tertiary center in Germany. The generalization to other populations, different recording qualities as well as OCT devices from other devices and in particular recently developed mobile devices, needs to be assessed. However, we show as a proof-of-principle study that multi- task learning enhances performance in a complex main task, namely activity recognition in nAMD, while increasing the overall explainability of neural network decision-making and the interpretability of the DNN decision for individual patient outcomes. It thus helps overcome the barriers to clinical application by mimicking the human process of diagnosis considering multiple disease features.

5 Acknowledgments

We thank the German Ministry of Science and Education (BMBF) for funding through the Tübingen AI Center (FKZ 01IS18039A) and the German Science Foundation for funding through a Heisenberg Professorship (BE5601/4-2) and the Excellence Cluster "Machine Learning — New Perspectives for Science" (EXC 2064, project number 390727645). H. Faber thanks the Faculty of Medicine, Eberhard Karls University of Tuebingen, Germany (application number 463–0–0) for additionally funding her research through the Junior Clinician Scientist Program (application number 463–0–0). We further thank Novartis AG for funding part of the research. The funding bodies did not have any influence in the study planning and design.

Author contribution statement

MSA, HF and PB designed the research. MSA performed the experiments. GA, WI, FZ, LK were involved in data acuqisition. FZ, HF, GA, LK and WI provided medical advice. MSA, HF and PB wrote the manuscript with input from all authors. All authors approved the final version of the manuscript and agreed on being accountable for the work.

References

[1] Michael D Abràmoff et al. "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices". In: *NPJ digital medicine* 1.1 (2018), pp. 1–8.

[2] Arghavan Almony et al. "Clinical and economic burden of neovascular age-related macular degeneration by disease status: a US claims-based analysis". In: *Journal of managed care & specialty pharmacy* 27.9 (2021), pp. 1260–1272.

[3] Carolina Arruabarrena et al. "Impact on visual acuity in neovascular age related macular degeneration (nAMD) in Europe due to COVID-19 pandemic lockdown". In: *Journal of clinical medicine* 10.15 (2021), p. 3281.

[4] Stamatios Aslanis et al. "Recurrent Neovascular Age-Related Macular Degeneration after Discontinuation of Vascular Endothelial Growth Factor Inhibitors Managed in a Treat-and-Extend Regimen". In: *Ophthalmology Retina* 6.1 (2022), pp. 15–20.

[5] Murat Seçkin Ayhan et al. "Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection". In: *Medical Image Analysis* (2020), p. 101724.

[6] Murat Seçkin Ayhan et al. "Clinical validation of saliency maps for understanding deep neural networks in ophthalmology". In: *Medical Image Analysis* (2022), p. 102364.

[7] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[8] Viraj Bhise et al. "Defining and measuring diagnostic uncertainty in medicine: a systematic review". In: *Journal of general internal medicine* 33.1 (2018), pp. 103–115.

[9] Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. "Attraction-repulsion spectrum in neighbor embeddings". In: *Journal of Machine Learning Research* 23.95 (2022), pp. 1–32.

[10] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[11] Usha Chakravarthy et al. "Impact of macular fluid volume fluctuations on visual acuity during anti-VEGF therapy in eyes with nAMD". In: *Eye* 35.11 (2021), pp. 2983–2990.

[12] Emily Y Chew et al. "Randomized trial of the ForeseeHome monitoring device for early detection of neovascular age-related macular degeneration. The HOme Monitoring of the Eye (HOME) study design—HOME Study report number 1". In: *Contemporary clinical trials* 37.2 (2014), pp. 294–300.

[13] Francois Chollet et al. *Keras*. 2015. url: `https://github.com/fchollet/keras`.

[14] Shelley Day et al. "Medicare costs for neovascular age-related macular degeneration, 1994–2007". In: *American journal of ophthalmology* 152.6 (2011), pp. 1014–1020.

[15] Jeffrey De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". In: *Nature medicine* 24.9 (2018), pp. 1342–1350.

[16] Yukun Ding et al. "Evaluation of Neural Network Uncertainty Estimation with Application to Resource-Constrained Platforms". In: *arXiv preprint arXiv:1903.02050* (2019).

[17] Frederick L Ferris, Stuart L Fine, and Leslie Hyman. "Age-related macular degeneration and blindness due to neovascular maculopathy". In: *Archives of ophthalmology* 102.11 (1984), pp. 1640–1642.

[18] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. "Deep ensembles: A loss landscape perspective". In: *arXiv preprint arXiv:1912.02757* (2019).

[19] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[20] Thomas Grote and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare". In: *Journal of medical ethics* 46.3 (2020), pp. 205–211.

[21] Thomas Grote and Philipp Berens. "How competitors become collaborators—Bridging the gap (s) between machine learning algorithms and clinicians". In: *Bioethics* 36.2 (2022), pp. 134–142.

[22] Chuan Guo et al. "On calibration of modern neural networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1321–1330.

[23] Joshua James Hatherley. "Limits of trust in medical AI". In: *Journal of Medical Ethics* 46.7 (2020), pp. 478–481.

[24] B Heimes et al. "Compliance von Patienten mit altersabhängiger Makuladegeneration unter Anti-VEGF-Therapie". In: *Der Ophthalmologe* 113.11 (2016), pp. 925–932.

[25] Frank G Holz et al. "Key drivers of visual acuity gains in neovascular age-related macular degeneration in real life: findings from the AURA study". In: *British Journal of Ophthalmology* 100.12 (2016), pp. 1623–1628.

[26] Tiarnan DL Keenan et al. "Retinal specialist versus artificial intelligence detection of retinal fluid from OCT: age-related eye disease study 2: 10-year follow-on study". In: *Ophthalmology* 128.1 (2021), pp. 100–109.

[27] Alex Kendall and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in Neural Information Processing Systems*. 2017, pp. 5580–5590.

[28] Dmitry Kobak and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics". In: *Nature communications* 10.1 (2019), pp. 1–14.

[29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6405–6416.

[30] Yingna Liu, Nancy LM Holekamp, and Jeffrey S Heier. "Prospective, longitudinal study: daily self-imaging with home OCT in neovascular age-related macular degeneration". In: *Ophthalmology Retina* (2022).

[31] Sara Llorente-González et al. "The role of retinal fluid location in atrophy and fibrosis evolution of patients with neovascular age-related macular degeneration long-term treated in real world". In: *Acta Ophthalmologica* (2021).

[32] Andrey Malinin and Mark Gales. "Predictive uncertainty estimation via prior networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 7047–7058.

[33] Karen Mulligan et al. "Economic value of anti–vascular endothelial growth factor treatment for patients with wet age-related macular degeneration in the United States". In: *JAMA ophthalmology* 138.1 (2020), pp. 40–47.

[34] Kester Nahen, Gideon Benyamini, and Anat Loewenstein. "Evaluation of a self-imaging SD-OCT system for remote monitoring of patients with neovascular age related macular degeneration". In: *Klinische Monatsblätter für Augenheilkunde* 237.12 (2020), pp. 1410–1418.

[35] Yurii E Nesterov. "A method for solving the convex programming problem with convergence rate O (1/k^ 2)". In: *Dokl. akad. nauk Sssr*. Vol. 269. 1983, pp. 543–547.

[36] Weili Nie, Yang Zhang, and Ankit Patel. "A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, July 2018, pp. 3809–3818.

[37] Abraham Olvera-Barrios et al. "Diagnostic accuracy of diabetic retinopathy grading by an artificial intelligence-enabled algorithm compared with a human standard for wide-field true-colour confocal scanning and standard digital retinal images". In: *British Journal of Ophthalmology* 105.2 (2021), pp. 265–270.

[38] World Health Organization et al. "World report on vision". In: (2019).

[39] Yaniv Ovadia et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13991–14002.

[40] Pavlin G. Poliar, Martin Stražar, and Blaž Zupan. "openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding". In: *bioRxiv* (2019). doi: 10.1101/731877.

[41] Sophie Riedl et al. "Impact of Intra-and Subretinal Fluid on Vision Based on Volume Quantification in the HARBOR Trial". In: *Ophthalmology Retina* 6.4 (2022), pp. 291–297.

[42] Philip J Rosenfeld. "Optical coherence tomography and the development of antiangiogenic therapies in neovascular age-related macular degeneration". In: *Investigative ophthalmology & visual science* 57.9 (2016), OCT14–OCT26.

[43] José M Ruiz-Moreno et al. "Economic burden of age-related macular degeneration in routine clinical practice: the RAMDEBURS study". In: *International ophthalmology* 41.10 (2021), pp. 3427–3436.

[44] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[45] Manuel Saenz-de-Viteri et al. "Role of intraretinal and subretinal fluid on clinical and anatomical outcomes in patients with neovascular age-related macular degeneration treated with bimonthly, treat-and-extend and as-needed ranibizumab in the In-Eye study". In: *Acta Ophthalmologica* 99.8 (2021), pp. 861–870.

[46] Rory Sayres et al. "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy". In: *Ophthalmology* 126.4 (2019), pp. 552–564.

[47] Thomas Schlegl et al. "Fully automated detection and quantification of macular fluid in OCT using deep learning". In: *Ophthalmology* 125.4 (2018), pp. 549–558.

[48] Sumit Sharma et al. "Macular morphology and visual acuity in the second year of the comparison of age-related macular degeneration treatments trials". In: *Ophthalmology* 123.4 (2016), pp. 865–875.

[49] Amitojdeep Singh et al. "What is the Optimal Attribution Method for Explainable Ophthalmic Disease Classification?" In: *Ophthalmic Medical Image Analysis*. Ed. by Huazhu Fu et al. Cham: Springer International Publishing, 2020, pp. 21–31. isbn: 978-3-030-63419-3.

[50] Frank A Sloan et al. "Longitudinal analysis of the relationship between regular eye examinations and changes in visual and functional status". In: *Journal of the American Geriatrics Society* 53.11 (2005), pp. 1867–1874.

[51] Bianka Sobolewska, Muhammed Sabsabi, and Focke Ziemssen. "Importance of Treatment Duration: Unmasking Barriers and Discovering the Reasons for Undertreatment of Anti-VEGF Agents in Neovascular Age-Related Macular Degeneration". In: *Clinical Ophthalmology (Auckland, NZ)* 15 (2021), p. 4317.

[52] Richard F Spaide et al. "Consensus nomenclature for reporting neovascular age-related macular degeneration data: consensus on neovascular age-related macular degeneration nomenclature study group". In: *Ophthalmology* 127.5 (2020), pp. 616–636.

[53] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning." In: *ICML (3)* 28.1139-1147 (2013), p. 5.

[54] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[55] Sunil Thulasidasan et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks". In: *Advances in Neural Information Processing Systems* 32 (2019).

[56] Alicia Valverde-Megıas et al. "Effect of COVID-19 Lockdown in Spain on Structural and Functional Outcomes of Neovascular AMD Patients". In: *Journal of Clinical Medicine* 10.16 (2021), p. 3551.

[57] Toon Van Craenendonck et al. "Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection". In: *Translational vision science & technology* 9.2 (2020), pp. 64–64.

[58] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[59] Sebastian M Waldstein et al. "Morphology and visual acuity in aflibercept and ranibizumab therapy for neovascular age-related macular degeneration in the VIEW trials". In: *Ophthalmology* 123.7 (2016), pp. 1521–1529.

[60] Yuyan Wang et al. "Small Towers Make Big Differences". In: *CoRR* abs/2008.05808 (2020). arXiv: 2008.05808. url: https://arxiv.org/abs/2008.05808.

[61] Inger Westborg et al. "Treatment for neovascular age-related macular degeneration in Sweden: outcomes at seven years in the Swedish Macula Register". In: *Acta Ophthalmologica* 95.8 (2017), pp. 787–795.

[62] Wan Ling Wong et al. "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis". In: *The Lancet Global Health* 2.2 (2014), e106–e116.

[63] Xiayan Xu et al. "Regional differences in the global burden of age-related macular degeneration". In: *BMC Public Health* 20.1 (2020), pp. 1–9.

[64] Jason Yim et al. "Predicting conversion to wet age-related macular degeneration using deep learning". In: *Nature Medicine* 26.6 (2020), pp. 892–899.

[65] Jason Yosinski et al. "How Transferable Are Features in Deep Neural Networks?" In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 3320–3328.

[66] Hongyi Zhang et al. "mixup: Beyond Empirical Risk Minimization". In: *International Conference on Learning Representations*. 2018.

Appendix

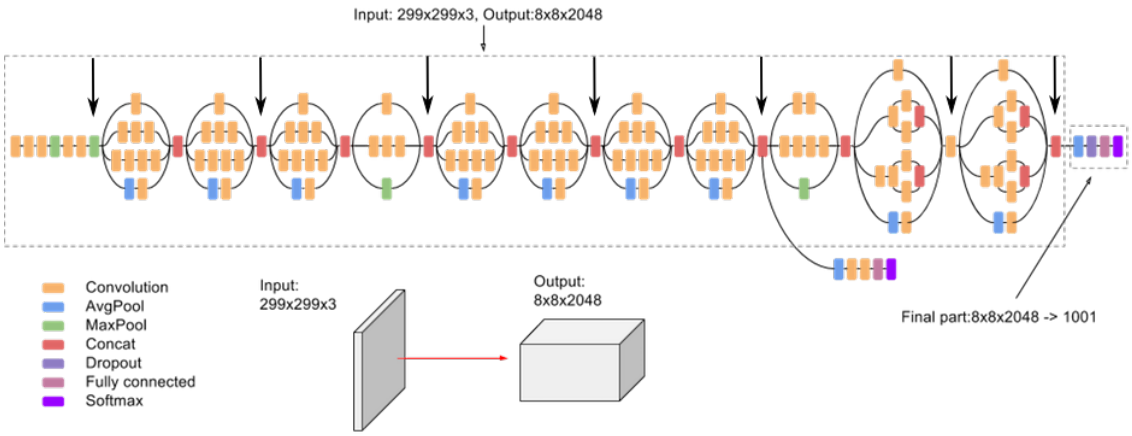

Figure 7: Read-out locations within the convolutional stack of the InceptionV3 architecture (indicated by big black arrows). In addition to these, we used the shared representation layer and task-specific layers of our multi-task networks (see Fig. 2). Base figure was obtained from `https://cloud.google.com/tpu/docs/inception-v3-advanced` .

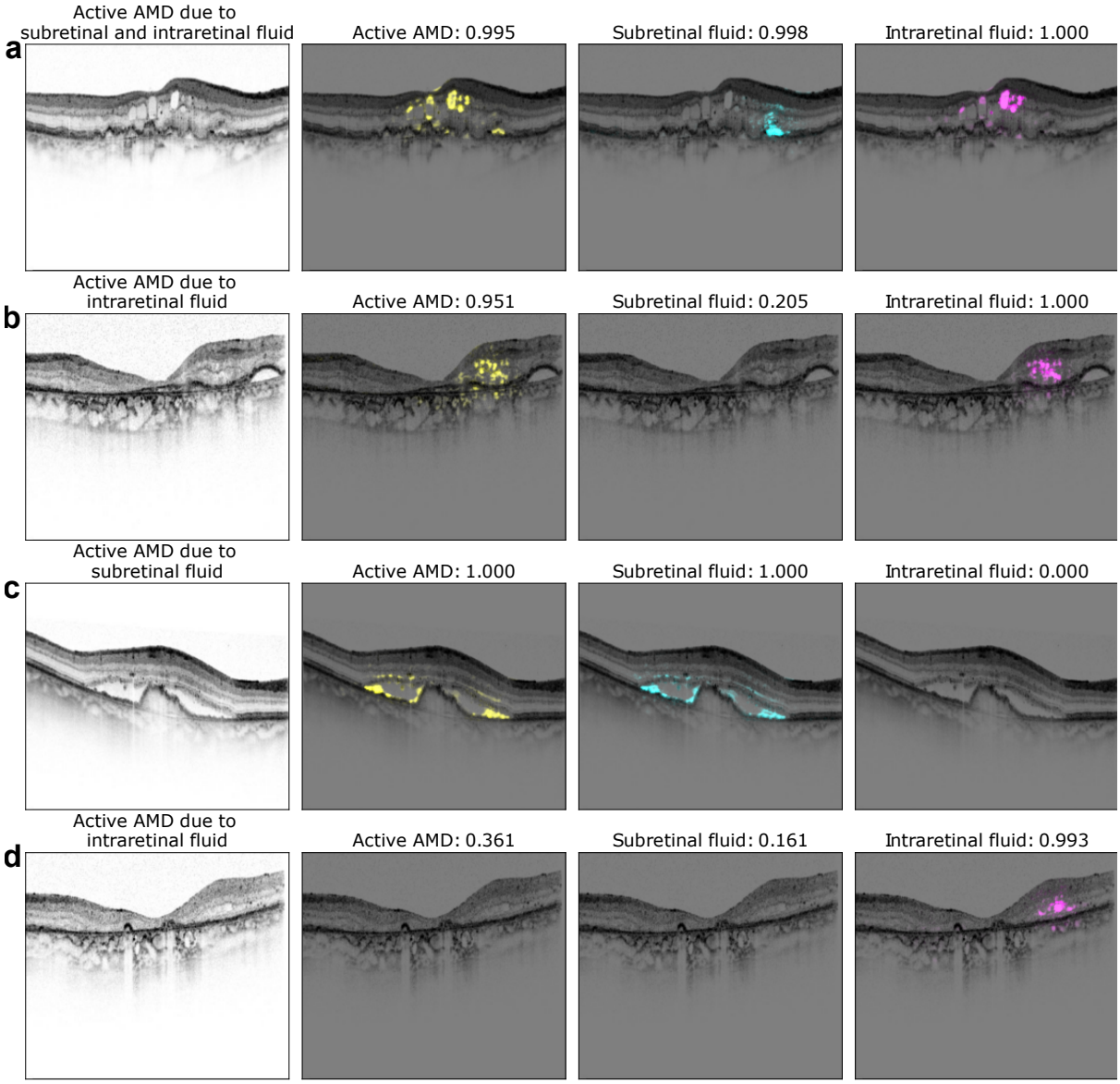

Figure 8: Supplementary saliency maps for the OCT images shown in Fig.6. These were obtained from single-task models.