NEURAL ODE-BASED DISEASE FORECASTING FROM RETINAL IMAGING WITH TEMPORAL CONSISTENCY

Arunava Chakravarty, Taha Emre, Dmitrii Lachinov, Antoine Rivail,

Ursula Schmidt-Erfurth & Hrvoje Bogunović *

ABSTRACT

Efficient clinical trial recruitment and personalized treatment depend on the ability to predict future disease progression from medical images. However, often there is a lack of well-defined biomarkers that can predict future disease development and a wide inter-subject variation in disease progression speed. We address these issues in the context of predicting the onset of late dry Age-related Macular Degeneration (dAMD) from retinal OCT scans. To model the CDF of future dAMD onset, we propose jointly training an AMD stage classifier with a Neural-ODE that predicts the future disease trajectory. A temporal ordering is imposed that inversely relates the distance from the decision hyperplane of the classifier to the time-to-conversion. Furthermore, we ensure intra-subject temporal consistency by incorporating pairs of longitudinal scans from the same eye during training. Our method is evaluated on a longitudinal dataset comprising 235 eyes (3,534 OCT scans), including 40 converters. The results demonstrate the efficacy of our approach, achieving an average eye-level AUROC of 0.83 in predicting conversion within the next 6,12,18 and 24 months, outperforming several popular survival analysis methods.

1 INTRODUCTION

Forecasting the risk of disease progression is crucial to prioritizing high risk patients for personalized treatment, and recruitment in clinical trials. However, it is a challenging task due to: (i) lack of well-established *clinical biomarkers* indicative of future disease progression; (ii) *Data Censoring* leading to unknown time-to-conversion labels due to missing followups or non-conversion within a limited study duration; (iii) *Class imbalance* as only a small proportion of the patients being monitored actually convert, with all scans from the non-converter cases and scans from converter patients before the conversion visit constituting the negative samples; (iv) *Discretizing time* into bins to pose conversion prediction as a binary or multi-label classification results in imprecise labels during training and inability to predict conversions at arbitrary continuous time during inference.

In this work, we aim to address these issues by considering Age-related Macular Degeneration (AMD), a progressive retinal disease which is the leading cause of blindness among the elderly population. As AMD progresses through early and intermediate stages (iAMD) characterized by drusen, it gradually progresses to a late-stage that is either dry (dAMD) or neovascular (nAMD), leading to irreversible vision loss. dAMD is indicated by the onset of Geographic atrophy due to the loss of the retinal pigment epithelium (RPE) layer while nAMD is marked by abnormal vessel growth that leaks fluid into the retina. Longitudinal OCT imaging is routinely employed to montior AMD progression. Although dAMD is more prevalent, most existing work has focussed on predicting nAMD onset, with few exceptions Rivail et al. (2019), Rivail et al. (2023). In contrast to handcrafted quantitative biomarker based methods Sleiman et al. (2017); Schmidt-Erfurth et al. (2018); Banerjee et al. (2020); de Sisternes et al. (2014); Lad et al. (2022), the recent deep learning (DL) models bypass the need for automated segmentation of retinal layers/pathologies by directly utilizing the OCT scans for end-to-end training. Temporal self-supervision has also been used to learn features from unlabeled longitudinal datasets in Emre et al. (2022); Rivail et al. (2019). A hybrid approach using both

*OPTIMA Lab, Department of Ophthalmology, Medical University of Vienna, Austria
{arunava.chakravarty,hrvoje.bogunovic}@meduniwien.ac.at.

biomarker and image features for predicting nAMD conversion is presented in Yim et al. (2020). Few methods have explored survival analysis such as the discrete logistic hazard model in Rivail et al. (2023) which utilizes OCT scans directly, while the linear Cox Proportional Hazards (CoxPH) model was used with pre-extracted biomarkers in Schmidt-Erfurth et al. (2018).

In this work, we propose a Neural Ordinary Differential Equation (N-ODE) based solution to predict dAMD onset. Our key contributions are: (i) Time-to-conversion from iAMD to dAMD is modeled in continuous time, rather than discrete time-intervals as used in most existing methods. Our model can therefore use actual continuous conversion times as ground-truths (GT) during training and also predict conversion probabilities within arbitrary continuous times. (ii) We combine an AMD stage classifier (iAMD vs. dAMD) with a N-ODE to predict the future trajectory of disease progression in a shared embedding space. We directly model the Cumulative Distribution Function (CDF) of the future conversion time in contrast to the existing methods Tang et al. (2022) that models the cumulative hazard function. We extend the ODE-GRU architecture for the N-ODE by stacking multiple layers, with multiple parallel heads in each layer. (iii) We employ a rank loss on the logits of the linear AMD stage classifier to derive a risk score to stratify patients into different risk groups relevant for personalized treatment. (iv) We incorporate intra-subject consistency by requiring the N-ODE estimates of the feature and risk at future time-points to be consistent with the values obtained using the actual OCT scan of the future visit by considering pair of visits per eye during training.

2 METHOD

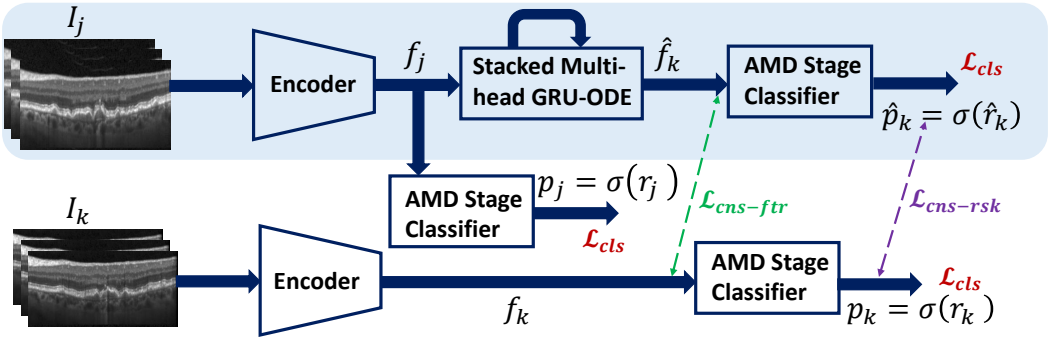

Figure 1: The same encoder and linear AMD stage classifier are used in both branches with shared weights during training. Only the top branch (highlighted in blue) is used during inference.

An overview of our method is depicted in Fig 1. Each training batch comprises a set of image-pairs from different eyes, where each random image-pair $(\boldsymbol{I}_j, \boldsymbol{I}_k)$ are two OCT scans of the same eye, acquired at different patient visits at time-points j and k with $j < k$. Following survival analysis, the GT for each scan $\boldsymbol{I}_j$ is denoted by (T_j, E_j) where the binary event indicator $E_j = 1$ if the eye to which $\boldsymbol{I}_j$ belongs, converts to dAMD during the duration of the study, else $E_j = 0$. T_j represents time duration from the current visit to either the first visit of conversion (if $E_j = 1$) or the last visit of the eye in the study (if $E_j = 0$). Both $\boldsymbol{I}_j, \boldsymbol{I}_k$ are passed through an Encoder (ConvNeXt-Tiny initialized with image-net pretrained weights Liu et al. (2022) was used), to obtain the features $\boldsymbol{f}_j, \boldsymbol{f}_k$ respectively. They are fed to a linear, binary iAMD vs. dAMD stage classifier to obtain the logits r_j, r_k and the corresponding prediction probabilities for conversion within time j (or k) denoted by $p_j = \sigma(r_j)$ (or p_k) with the sigmoid activation (σ). Additionally, $\boldsymbol{f}_j$ is evolved with a N-ODE for the time-interval $(j - k)$ as: $\frac{df(t)}{dt} = \boldsymbol{v}(\boldsymbol{f}(t))$, with the initial condition $f(t = 0) = \boldsymbol{f}_j$ to obtain an estimate of the future feature $\hat{\boldsymbol{f}}_k$ and the corresponding prediction $\hat{\boldsymbol{p}}_k$ directly from the previous scan $\boldsymbol{I}_j$. The instantaneous velocity $\boldsymbol{v}(f(t))$ is modeled with a DL network as follows.

N-ODE architecture: GRU-ODE in De Brouwer et al. (2019) was extended by stacking 3 layers and incorporating 12 parallel heads within each layer[1]. $\boldsymbol{f}(t)$ acts as the initial hidden state for each

[1]Pytorch code of our deep Multi-head GRU-ODE architecture is available at `https://github.com/arunava555/Multihead_GRU_ODE_based_Survival_Analysis`

layer, while the output from the previous layer is provided as an *external input* (except for the first layer). Both the external input and the hidden state in each layer were projected to independent 64-dimensional sub-spaces for each head, to which a GRU-ODE cell with the *reset* and *update* gates was employed. The outputs from all heads were conatenated ($768 = 12$ heads $\times 64$ dimensions) and passed through a fully connected layer. The output from the final layer $\boldsymbol{v}(\boldsymbol{f}(t))$ is unnormalized (without activation) to ensure that it could take values in any range (even negative).

Training Losses The primary loss for training, $\mathcal{L}_{cls} = \mathcal{L}_{bce}(y_i^c, p_i) + \mathcal{L}_{bce}(y_j^c, p_j) + \mathcal{L}_{bce}(y_j^c, \hat{\boldsymbol{p}}_k)$ is the classification loss, where $\mathcal{L}_{bce}$ denotes binary cross-entropy. The GT y_j^c indicates if $\boldsymbol{I}_j$ is currently in the iAMD or dAMD stage ($y_j^c = 1$ if $E_j = 1$ & $T_j \leq 0$, or 0 otherwise).

The learned feature embedding is regularized to ensure that the N-ODE learns meaningful future trajectories of disease progression with intra-eye consistency between the features ($\mathcal{L}_{cns-ftr}$) and predictions ($\mathcal{L}_{cns-rsk}$) with the loss terms $\mathcal{L}_{cns-ftr} = ||\boldsymbol{f}_k - \hat{\boldsymbol{f}}_k||_2^2$ and $\mathcal{L}_{cns-rsk} = \mathcal{L}_{bce}(p_k, \hat{p}_k)$.

Finally, we impose a ranking loss $\mathcal{L}_{rnk}$ on the logits r_j to use it as a risk score. Given, two arbitrary scans $\boldsymbol{I}_m, \boldsymbol{I}_n$ (they can be scans from different eyes, unlike $\boldsymbol{I}_j, \boldsymbol{I}_k$) s.t. $\boldsymbol{I}_m$ converts before $\boldsymbol{I}_n$, then their risk scores $r_m > r_n$. Since, r_m is proportional to the distance of $\boldsymbol{f}_m$ from the decision hyperplane separating iAMD and dAMD, the $\mathcal{L}_{rnk}$ loss ensures: (a) regularization of the feature manifold to encourage temporal ordering, i.e., the iAMD samples closer to the decision hyperplane separating iAMD and dAMD will have a smaller time-to-conversion and viceversa, and (b) the risk score can be used to stratify eyes in different risk groups, enabling personalized treatment. $\mathcal{L}_{rnk}$ solves an additional logistic regression task with scalar input $(r_m - r_n)$ to predict the probability $p_{m>n}$ of $\boldsymbol{I}_m$ converting before $\boldsymbol{I}_n$ as $\mathcal{L}_{rnk} = -\frac{1}{|S_{m<n}|+|S_{m>n}|} \cdot \Big[\sum_{S_{m<n}} log\,(p_{m<n}) + \sum_{S_{m>n}} log\,(1 - p_{m<n})\Big]$, where $S_{m<n}$ represents a subset of all possible image-pairs in a training batch where I_m converts before I_n for which ideally, $p_{m<n} \approx 1$. Similarly, $S_{m>n}$ contains image-pairs where $\boldsymbol{I}_m$ converts after $\boldsymbol{I}_n$ and ideally, $p_{m<n} \approx 0$. The set $S_{m<n}$ is defined by taking censoring into account as the image-pairs for which $(T_m < T_n)$ & either $E_m = 1$ or $\boldsymbol{I_m}$, $\boldsymbol{I_n}$ are scans from different time-points of the same eye. For image-pairs coming from different visits of the same eye, the risk for the scan from a later visit $\boldsymbol{I}_m$ (and hence a smaller time T_m from the last monitored visit) will always be higher than the former visit $\boldsymbol{I}_n$, even for censored cases, since AMD progression is irreversible and can only increase with time. Similarly, $S_{m>n}$ is defined as pairs where $(T_m > T_n)$ & either $E_n = 1$ or $\boldsymbol{I_m}$, $\boldsymbol{I_n}$ come from the same eye. The relative weights of $\mathcal{L}_{cls}$, $\mathcal{L}_{cns-ftr}$, $\mathcal{L}_{cns-rsk}$ and $\mathcal{L}_{rnk}$ are not handcrafted but dynamically adjusted during training using the Multi-Task ADAM optimizer Malkiel & Wolf (2021) that normalizes the gradients from each loss term to a similar range.

3 RESULTS

Dataset: The proposed method was evaluated on a dataset comprising 3,534 OCT scans from 235 eyes (40 converters and 195 censored), collected at the Department of Ophthalmology, Medical University of Vienna Schlanitz et al. (2017). The images were acquired with 49 B-scans (slices), each of size $512 - 1024 \times 496$ pixels. The eyes were imaged every 3-6 months, with a follow-up period of 2-7 years. The GT time-to-conversion was computed for each scan of a converter eye by measuring the time interval between its acquisition and the first conversion visit.

Experiments: A stratified five-fold cross-validation was employed to reduce any bias due to a train-test data split. Each fold was divided at an eye-level with 8 converter and 39 non-converter cases with 667-707 OCT scans. The training set in each fold was further randomly divided where 80% was used to train and the remaining 20% for validation to track the best performing model weights and perform early stopping. The average Area under the ROC curve (AUROC) for predicting the conversion to dAMD within $6, 12, 18$ and 24 months was computed and Concordance Index (CCI) was used to evaluate the risk score at an eye-level across the five-folds. An eye-level *bootstraping* was employed which involved constructing multiple random re-samplings of the test OCT scans. In each re-sampling, only one OCT scan was selected from each eye (by randomly selecting one of the patient visits) in the test set. The average performance across the 1000 (bootstrap re-samplings)$\times$5(folds)=5000 sample estimates is reported in Table 3. Three consecutive central B-scans around the macula were employed as RGB channels for input to the Encoder.

Results: We compared our performance against many popular Survival Analysis based methods. The censored cross-entropy loss Wulczyn et al. (2020) and the logistic hazard model Rivail et al.

(2023) are discrete-time survival models and employed 6-month time-interval bins. DeepSurv Katzman et al. (2018) extends CoxPH with DL, while SODEN Tang et al. (2022) is a Neural-ODE based method, proposed for tabular data. These methods were also trained with ConvNeXt-Tiny as the encoder (similar to our method) but with modified classification layers and losses. Notably, all of these methods do not employ intra-subject regularization, hence require training a single branch network. The results in Table 3 demonstrate the superiority of our proposed method, which outperforms the existing methods at all time-points. SODEN, another N-ODE-based method showed signs of overfitting with good performance on the validation set (used for selecting the best-performing models in each fold) but led to a drastic drop in performance on the test sets across all folds.

Identifying risk groups: We calibrated the risk scores in each fold to lie in the $[0, 1]$ by mapping each value to its corresponding percentile with a bicubic interpolation. The test set predictions of the calibrated risk scores were combined from the five folds to obtain a risk score for each OCT scan. The scans were then stratified into 3 groups with low risk ($0 \leq r \leq 0.33$), moderate risk ($0.33 < r \leq 0.67$) and high risk ($0.67 < r \leq 1$). A population-level survival function for these groups is plotted in Fig. 2(a) using the Kaplan–Meier estimator on the GT conversion time. It depicts the mean and standard deviation of the survival probability for each population group, computed across 1000 re-samplings using eye-level bootstrapping. The survival curves for the three risk groups show a clear separation, thereby demonstrating the effectiveness of the proposed risk score.

Qualitative Results: The UMAP visualization of the features is depicted in Fig. 2(b). The censored OCT scans with unknown conversion time are shown in gray. The remaining scans are plotted with a colormap transitioning from red (small time-to-conversion) to blue (long time-to-convert). A smooth transition in conversion time is observed in the feature manifold. Grad-CAM maps for two OCT scans are also presented in Fig. 2(c) corresponding to the risk score. The saliency maps indicate the sensitivity of the proposed method towards irregularities around the RPE layer(top row) and pathologies such as Hyper-reflective Foci (HRF) that have been linked to fast AMD progression.

Table 1: Eye-level performance (mean $\pm$ std. dev.) with best values highlighted column-wise.

	AUROC				
	6	12	18	24	CCI
Proposed	$\mathbf{0.863 \pm 0.10}$	$\mathbf{0.827 \pm 0.10}$	$\mathbf{0.808 \pm 0.07}$	$\mathbf{0.816 \pm 0.07}$	$\mathbf{0.769 \pm 0.06}$
Cens. Cross-Entropy	0.775 ± 0.14	0.772 ± 0.103	0.773 ± 0.10	0.790 ± 0.08	0.762 ± 0.06
Logistic Hazard	0.769 ± 0.19	0.768 ± 0.12	0.763 ± 0.09	0.786 ± 0.08	0.749 ± 0.08
DeepSurv	0.769 ± 0.18	0.710 ± 0.16	0.712 ± 0.14	0.723 ± 0.14	0.752 ± 0.07
SODEN	0.675 ± 0.24	0.674 ± 0.17	0.673 ± 0.13	0.698 ± 0.11	0.673 ± 0.09

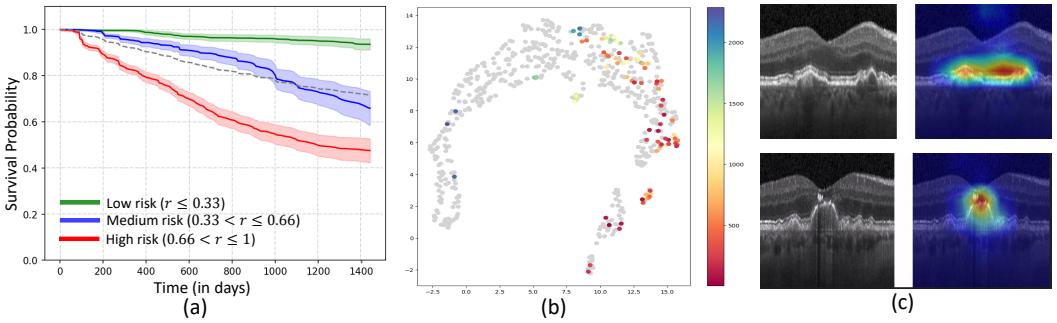

Figure 2: (a) Kaplan-Meier curves for different risk groups; (b) UMAP plot of feature embedding for one of the five folds. The censored scans are depicted with gray dots and the converters colored by their time to conversion (red indicates fast conversion); (c) Grad-CAM maps for the risk score.

4 CONCLUSION

In this work, we propose a novel framework that combines an AMD stage classifier with a Neural ODE to forecast dAMD onset at continuous future times. To learn meaningful feature embedding from limited data, it was regularized by enforcing intra-subject consistency between the predicted features and risk scores over pairs of longitudinal visits. Furthermore, temporal ordering was also imposed where a scan's proximity to the AMD classifier's decision hyperplane is inversely related to its time-to-conversion. These constraints enabled our model to outperform several existing deep survival analysis methods. Additionally, temporal ranking provides a scalar risk score to stratify eyes into low and high risk groups. Our method has the potential to facilitate patient-specific disease management and enrich clinical trial populations with high-risk patients.

REFERENCES

Imon Banerjee, Luis de Sisternes, Joelle A Hallak, Theodore Leng, Aaron Osborne, Philip J Rosenfeld, Giovanni Gregori, Mary Durbin, and Daniel Rubin. Prediction of age-related macular degeneration disease using a sequential deep learning approach on longitudinal sd-oct imaging biomarkers. *Scientific reports*, 10(1):15434, 2020.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

Luis de Sisternes, Noah Simon, Robert Tibshirani, Theodore Leng, and Daniel L Rubin. Quantitative sd-oct imaging biomarkers as indicators of age-related macular degeneration progression. *Investigative ophthalmology & visual science*, 55(11):7093–7103, 2014.

Taha Emre, Arunava Chakravarty, Antoine Rivail, Sophie Riedl, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Tinc: Temporally informed non-contrastive learning for disease progression modeling in retinal oct volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 625–634. Springer, 2022.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

Eleonora M Lad, Karim Sleiman, David L Banks, Sanjay Hariharan, Traci Clemons, Rolf Herrmann, Daniyar Dauletbekov, Andrea Giani, Victor Chong, Emily Y Chew, et al. Machine learning oct predictors of progression from intermediate age-related macular degeneration to geographic atrophy and vision loss. *Ophthalmology Science*, 2(2):100160, 2022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Itzik Malkiel and Lior Wolf. Mtadam: Automatic balancing of multiple training loss terms. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10713–10729, 2021.

Antoine Rivail, Ursula Schmidt-Erfurth, Wolf-Dieter Vogl, Sebastian M Waldstein, Sophie Riedl, Christoph Grechenig, Zhichao Wu, and Hrvoje Bogunovic. Modeling disease progression in retinal octs with longitudinal self-supervised learning. In *Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2*, pp. 44–52. Springer, 2019.

Antoine Rivail, Wolf-Dieter Vogl, Sophie Riedl, Christoph Grechenig, Leonard M Coulibaly, Gregor S Reiter, Robyn H Guymer, Zhichao Wu, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Deep survival modeling of longitudinal retinal oct volumes for predicting the onset of atrophy in patients with intermediate amd. *Biomedical Optics Express*, 14(6):2449–2464, 2023.

Ferdinand G Schlanitz, Bernhard Baumann, Michael Kundi, Stefan Sacu, Magdalena Baratsits, Ulrike Scheschy, Abtin Shahlaee, Tamara J Mittermüller, Alessio Montuoro, Philipp Roberts, et al. Drusen volume development over time and its relevance to the course of age-related macular degeneration. *British Journal of Ophthalmology*, 101(2):198–203, 2017.

Ursula Schmidt-Erfurth, Sebastian M Waldstein, Sophie Klimscha, Amir Sadeghipour, Xiaofeng Hu, Bianca S Gerendas, Aaron Osborne, and Hrvoje Bogunović. Prediction of individual disease conversion in early amd using artificial intelligence. *Investigative ophthalmology & visual science*, 59(8):3199–3208, 2018.

Karim Sleiman, Malini Veerappan, Katrina P Winter, Michelle N McCall, Glenn Yiu, Sina Farsiu, Emily Y Chew, Traci Clemons, Cynthia A Toth, Wai Wong, et al. Optical coherence tomography predictors of risk for progression to non-neovascular atrophic age-related macular degeneration. *Ophthalmology*, 124(12):1764–1777, 2017.

Weijing Tang, Jiaqi Ma, Qiaozhu Mei, and Ji Zhu. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *The Journal of Machine Learning Research*, 23(1):1516–1544, 2022.

Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PloS one*, 15(6):e0233678, 2020.

Jason Yim, Reena Chopra, Terry Spitz, Jim Winkens, Annette Obika, Christopher Kelly, Harry Askham, Marko Lukic, Josef Huemer, Katrin Fasler, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine*, 26(6):892–899, 2020.