

Learning Spatio-Temporal Model of Disease Progression With NeuralODEs From Longitudinal Volumetric Data

Dmitrii Lachinov¹, Arunava Chakravarty², Christoph Grechenig, Ursula Schmidt-Erfurth³, and Hrvoje Bogunović⁴, for the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Abstract—Robust forecasting of the future anatomical changes inflicted by an ongoing disease is an extremely challenging task that is out of grasp even for experienced healthcare professionals. Such a capability, however, is of great importance since it can improve patient management by providing information on the speed of disease progression already at the admission stage, or it can enrich the clinical trials with fast progressors and avoid the need for control arms by the means of digital twins. In this work, we develop a deep learning method that models the evolution of age-related disease by processing a single medical scan and providing a segmentation of the target anatomy at a requested future point in time. Our method represents a time-invariant physical process and solves a large-scale problem of modeling temporal pixel-level changes utilizing NeuralODEs. In addition, we demonstrate the approaches to incorporate the prior domain-specific constraints into our method and define temporal Dice loss for learning temporal objectives. To evaluate the applicability of our approach across different age-related diseases and imaging modalities, we developed and tested the proposed method on the datasets with 967 retinal OCT volumes of 100 patients with Geographic Atrophy and 2823 brain MRI volumes of 633 patients with Alzheimer's Disease. For Geographic Atrophy, the proposed method outperformed the related baseline models in the atrophy growth prediction. For Alzheimer's Disease, the proposed method demonstrated remarkable performance in predicting the brain ventricle changes induced by the disease, achieving the state-of-the-art result on TADPOLE cross-sectional prediction challenge dataset.

Index Terms—Disease progression, deep learning, longitudinal imaging, retina, geographic atrophy, Alzheimer's disease.

I. INTRODUCTION

THE progress in medical imaging techniques brought vast possibilities for the diagnosis and monitoring of patients with age-related diseases. Tools such as Optical Coherence Tomography (OCT) or Magnetic Resonance Imaging (MRI) are the current clinical gold standard and contribute to an overwhelming number of diagnostic findings. The amount of information within these scans, however, is so large, that it requires many years of training and practice to interpret them. Yet, we believe, a lot stays hidden from expert eyes. For instance, we hypothesize that the track of the natural, age-related disease progression is determined by the subtle physiological changes which are captured and encoded by medical scans. The class of computational methods trying to extract such information is called Disease Progression Models (DPMs), which can, e.g. calculate the expected future change in the target quantitative biomarkers. A more sophisticated class of methods can forecast morphological change with respect to the shape and size of an anatomical structure or lesion that is expected to occur in the future, from an image acquired at the time of screening. The task of modeling future changes is extremely challenging as the early biomarkers that indicate either a fast or slow speed of progression are often very subtle and subclinical, i.e., not well-understood in clinical practice.

DPMs have already been employed as decision support and as tools complementing phase II and III clinical trials [1], [2], [3]. They typically model the evolution of a set of predefined biomarkers over the course of the disease and map it to the global timeline of disease evolution, where the biomarkers are typically extracted from clinical reports and imaging data. However, such biomarkers relevant to model the disease progression have to be predetermined by the human experts, which is often a non-trivial task, and likely does not cover all the relevant ones. Thus, the search for predictive biomarkers is currently a complicated and tedious procedure.

The state-of-the-art methods that automatically extract features from the input volume also have some weaknesses, as some are designed for forecasting fixed intervals only;

Manuscript received 22 August 2023; revised 17 October 2023; accepted 29 October 2023. Date of publication 7 November 2023; date of current version 4 March 2024. This work was supported in part by the Christian Doppler Research Association; in part by the Austrian Federal Ministry for Digital and Economic Affairs; and in part by the National Foundation for Research, Technology and Development. (Corresponding author: Dmitrii Lachinov.)

Dmitrii Lachinov and Hrvoje Bogunović are with the Christian Doppler Lab for Artificial Intelligence in Retina, the Department of Ophthalmology and Optometry, Medical University of Vienna, 1090 Vienna, Austria (e-mail: dmitrii.lachinov@meduniwien.ac.at; hrvoje.bogunovic@meduniwien.ac.at).

Arunava Chakravarty, Christoph Grechenig, and Ursula Schmidt-Erfurth are with the Department of Ophthalmology and Optometry, Medical University of Vienna, 1090 Vienna, Austria (e-mail: arunava.chakravarty@meduniwien.ac.at; christoph.grechenig@meduniwien.ac.at; ursula.schmidt-erfurth@meduniwien.ac.at).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TMI.2023.3330576>, provided by the authors.

Digital Object Identifier 10.1109/TMI.2023.3330576

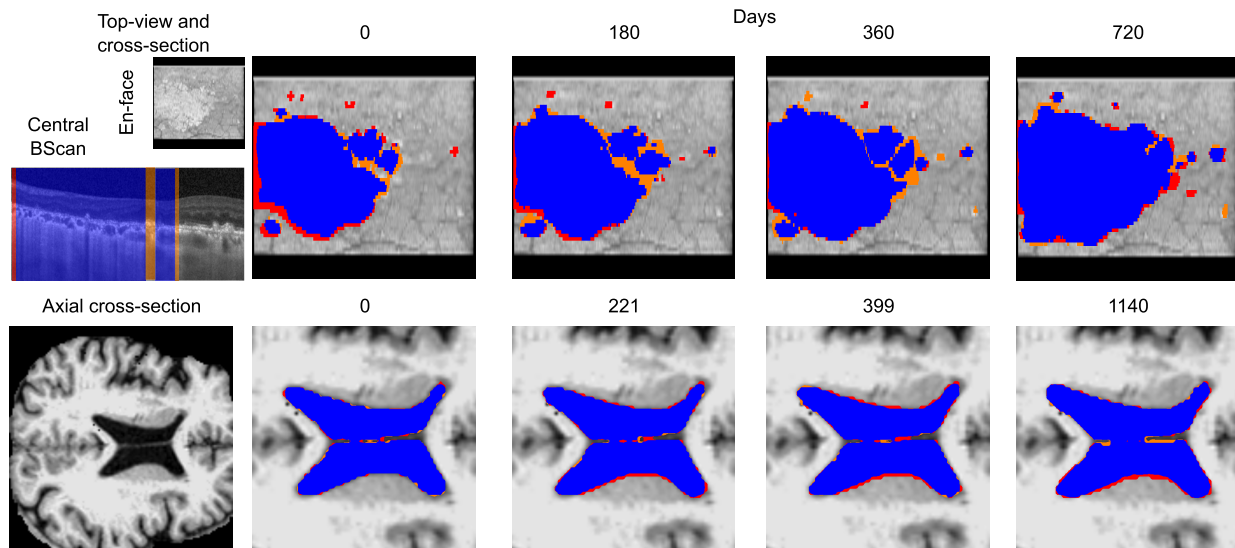


Fig. 1. The proposed method predicts the future target anatomy segmentation from a single volumetric scan acquired at baseline (day 0). The top row represents the en-face projections of the volumetric 3D OCT scan with the atrophy segmentation predicted at different future visits, denoted by the number of days since the baseline visit. The raw baseline OCT volume is shown as an en-face projection and with a central BScan (cross-section). Note that new islands of atrophy form in the central right and top left locations of the scan. The bottom row represents an axial slice of the brain T1 MRI cropped around the ventricles, with their predicted segmentations at future time points. The growth in a 2D slice is marginal, but it is substantial in 3D. In both rows true positives are marked in blue, false positives in orange, and false negatives in red. The ground truth is represented as blue + red and prediction as blue + orange.

others are time-variant systems, as the same input volume can produce different outcomes at different times. In this work, we propose a novel disease progression model capable of learning directly from longitudinal imaging data to predict the target area or volume change from a single scan, acquired at baseline visit. The proposed model not only predicts continuous-in-time changes in the target size, but it is time-invariant and inherently interpretable, as it simultaneously produces a detailed segmentation map of the current and the future target anatomy.

A. Clinical Background

We focus on developing a DPM method for two clinically-relevant scenarios: (1) modeling the growth of Geographic Atrophy (GA) from longitudinal 3D OCT volumes of the retina to monitor Age-Related Macular Degeneration (AMD), and (2) modeling Alzheimer's Disease (AD) progression by capturing the morphological changes in the brain ventricles from longitudinal 3D MRI volumes of the brain.

1) Geographic Atrophy: AMD is a leading cause of blindness, affecting 8.7% of the world population [4]. In the 80+ age group, this number grows to 25%, with the prevalence of advanced AMD of 3.3%. AMD in its late stages can either be *dry* characterized by the presence of Geographic Atrophy (GA) or *wet* characterized by the presence of Macular Neovascularization. OCT has become a gold standard modality for imaging the retina in the clinical setting due to its non-invasive 3D acquisition and widespread availability in clinics. It acquires a detailed volumetric scan of the retina made of a series of cross-sectional slices or BScans (Fig. 1). On a BScan, GA is represented as atrophy of the outer retinal tissues [5]. Thus, each column (AScan) in a BScan can be denoted as atrophic or not. This makes the GA lesion to be 2D in

nature (Fig. 1), with its size typically measured as an en-face area in mm^2 . In contrast to the late-wet stage of AMD, the late-dry stage is more prevalent, affecting around 5 Million people worldwide, and only recently a very first treatment for GA received FDA approval [6]. Thus, modeling the progression of GA can play a critical role in clinical research to understand the pathophysiology and the simulation of the future atrophy progression. It will also be fundamental to patient-specific disease management in order to identify the fast progressors suitable for more frequent treatment.

2) Alzheimer's Disease: AD is a gradually progressing, irreversible neurodegenerative disorder that is the primary of Dementia. Around 50 million patients suffer from this impairment worldwide, with the projected number of patients effectively doubling every 5 years and reaching 152 million by 2050 [7]. Currently, only symptomatic treatment is available. Volumetric MRI of the brain is commonly employed in the clinical examination of AD to detect the early pathological changes in the cortex, hippocampus, and ventricles. Recent studies by [8] and [9] have correlated the ventricle volume extracted from MRI images to the progression of AD, thereby making it a potential biomarker for future interventions and disease progression modeling. The recent challenge, TAD-POLE [10] aimed at predicting the future evolution of AD patients, where the ventricles' volume was used as one of the main outcome measures. Thus, we focus on the development of ventricles captured by brain MRI, and predicting their 3D segmentation evolution over a three-year period (Fig. 1).

B. Related Work

Addressing DPM with Deep Learning is an active area of research. Existing methods can be broadly categorized into two approaches, those that operate on attributes of tabular data,

and those that operate on pixels of the imaging data. In the first category, the methods rely on Electronic Health Records (EHR), sex, age, anthropometric parameters, and quantitative imaging biomarkers, which have been derived from the images through manual or automated segmentation. Such *attribute-level* methods typically employ Recurrent Neural Networks (RNNs) [11], [12] to fit the temporal variation of the measured attributes. For the prediction of diabetic kidney disease progression, Chan et al. [13] employed a random forest model using a combination of EHR and imaging biomarkers. Zheng and Hu [14] focused on the disease progression modeling of Alzheimer's Disease, and proposed a longitudinal data fusion approach to merge EHR data from multiple time points. The RNN was used by Bai et al. [15] to process EHR codes of the visits, where the authors introduced a mechanism to learn time decay factors for each EHR code. For monitoring COVID-19 patients, Dayan et al. [16] trained a CNN using a Federated Learning (FL) strategy. The model used EHR data in addition to X-Ray images and was designed to predict outcomes 24 and 72 hours from admission to the emergency room.

Methods in the second category utilize the raw imaging data directly for spatio-temporal modeling of the disease, typically in the form of respective future lesion segmentation maps. We will refer to these methods as *pixel-level* progression methods, and focus on them as our proposed method follows this approach.

For modeling the GA progression, Zhang et al. [17] proposed a pixel-level multi-scale deep Convolutional Neural Network (MS-DCNN) for joint segmentation and prediction of GA. The proposed method takes preprocessed baseline en-face OCT projection and produces segmentation at the specified future visits. However, the method lacks the ability to interpolate the progression between the visits. In addition, the dataset used in the experiments was small, consisting of only 29 patients. In follow-up work, Zhang et al. [18] proposed a Bi-LSTM-based prediction model and a UNet-based refinement module. However, it requires two visits to make a prediction and the dataset used was still small, consisting of only 25 patients. Recently, Gigon et al. [19] proposed a pixel-level model that uses a Taylor series with learnable time derivatives to approximate patient-specific continuous GA growth. The method takes a set of retinal layer thickness maps as input, which makes the method application-specific and limits the amount of information the model can learn from. This method was also evaluated on only 20 test samples.

For the AD progression, Wang et al. [11] proposed to utilize an attribute-level method with LSTM [20] to forecast the AD stage at the upcoming visits using 78 features derived from demographics, patient history, and other sources. In this method, the prediction interval is provided as part of the input. Nguyen et al. [12] employed a similar approach utilizing MinimalRNN [21], where the authors linked each recursive update with a one-month interval. The authors used the set of features provided as a part of the TADPOLE [10] challenge. Among teams participating in TADPOLE [10], Marinescu et al. [22] as a part of team GlassFrog employed high-dimensional regression and disease state-specific slope models and ensembling for ventricle volume forecasting.

Marinescu et al. [22] as a part of BravoLab team proposed to use LSTM for making forecasts. Marinescu et al. [22] fitted a linear mixed-effects model to estimate ventricle volume expansion rate. Gruffaz et al. [23] applied Riemannian metric learning to mixed effects models. The method embeds the data into the Riemannian manifold and learns patient-specific trajectories. The authors employed the proposed method for AD progression on ADNI and TADPOLE data, surpassing the performance of the methods participating in the TADPOLE challenge. The method, however, is not suitable for imaging data and works only with EHR or biomarkers. In general, the reviewed methods do not exploit the information from the entire volume, since they rely on a set of input features, which were pre-defined by human experts.

A notable pixel-level method, was proposed by Ezhov et al. [24] which fit the parameters of the partial differential equation (PDE), that defines the brain tumor growth, to the real imaging data. Unfortunately, the solution is not generalizable outside the tumor growth tasks, as it would require knowing a form of PDE beforehand. In addition, the method was only tested on 2 real rat MRI scans. Another pixel-level method was proposed by Petersen et al. [25]. The authors trained a deep learning model to learn tumor growth dynamics directly from the imaging data. The work focuses on modeling the distribution of the possible tumor appearances. However, as the authors mentioned, the method is neither suitable for predicting a single correct growth trajectory nor designed for it. Expanding the work on probabilistic disease modeling, Petersen et al. [26] introduced a continuous progression model, based on UNet [27] and attention mechanism [28]. As before, the model wasn't explicitly designed for predicting a single trajectory, but rather a distribution of possible outcomes. Additionally, the model was trained to utilize two or more images as prior information, which reduces the method's applicability in certain scenarios, such as screening for fast progressors from a single scan for enriching the clinical trials.

In this paper, we base our proposed pixel-level DPM method on Neural Ordinary Differential Equation (NeuralODE) [29]. The Ordinary Differential Equation (ODE) consists of an unknown function of one variable, typically time, and its derivatives. Together with the observed value of the unknown function at the initial time (Initial Conditions), ODE forms Initial Value Problem (IVP). We are interested in finding such functions that satisfy initial conditions and set ODE into equality. The framework introduced within NeuralODEs models the right-hand side of the ODE with a neural network and approximates the solution using a numerical solver. As a part of training, the neural network is updated to match the solution with the observed training data.

The distinctive features of the NeuralODEs that make them especially suitable for DPM are that they are inherently continuous, which allows forecasting at any future time-point; they represent some underlying physical process, and, at the simplest level, the neural network learned by NeuralODEs corresponds to the instantaneous change in the target variable. In addition, NeuralODEs allow for controlling the solution's properties and imposing domain-specific constraints on the learned function. This builds an ideal foundation for

NeuralODEs as a disease progression model. With the mentioned advantages, NeuralODEs are steadily finding an application in learning disease progression. Hao et al. [30] proposed a NeuralODE-based diffusion model, conditioned on an individual's connectivity, for Amyloid pathology progression, which is a primary pathological event in Alzheimer's disease. Rubanova et al. [31] investigated NeuralODEs in application to irregularity-sampled time series and introduced ODE-RNNs for handling multiple prior observations. The method was tested on multiple tasks including survival prediction of ICU patients using a set of quantitative biomarkers. De Brouwer et al. [32] used their previously proposed GRU-ODE-Bayes model [33] for the prediction of disability progression in Multiple Sclerosis patients in a two-year interval using a set of biomarkers.

C. Contribution

The existing DPM methods discussed above have at least one of the following drawbacks: they rely on hand-crafted features [13], [14], [15], [16]; they are inherently non-interpretable [22], as many provide only numerical value as a prediction; they are designed for fixed forecasting intervals [17]; they are time-variant systems [26], as they can produce different results for the same forecasting interval. In contrast, we introduce a framework for pixel-level DPM based on NeuralODEs [29] that addresses the above-mentioned drawbacks. To the best of our knowledge, this is the first work that incorporated NeuralODEs [29] for target anatomy segmentation and shape progression modeling effectively addressing disease progression tasks. The proposed method: (1) extracts the relevant features directly from the image and provides progression estimate with corresponding segmentation masks at a future time-point for reliability and feasibility assurance; (2) learns from data with missing visits and variable visit intervals; (3) interprets disease progression as autonomous ODE and introduces methods to incorporate prior domain knowledge. (4) introduces temporal Dice loss and highlights a successful application of NeuralODE as shown by our large-scale evaluation

We trained the proposed method using two large-scale datasets of 100 GA patients with 967 scans and 633 AD patients with 2537 scans, with the total amount of data far larger than in related works. We used one of the public datasets of 143 patients with 286 scans, TADPOLE D3-D4, as a hold-out test set and compared our results with those of the participants in the TADPOLE challenge [22], achieving state-of-the-art prediction performance on the cross-sectional prediction task.

II. METHOD

We formulate the disease progression as an IVP comprising of an ODE with an initial condition that specifies the value of the unknown function at baseline. The initial conditions take the form of the embedding vectors which are generated for each pixel of the baseline image using any segmentation architecture such as UNet [27]. The goal of the initial conditions is to map the baseline scan to the embedding space, encoding the current disease state. The mapping to the chronological time

frame of the disease happens implicitly through embedding space. As demonstrated by prior work [34], the deep learning models are able to effectively predict the pathological state from OCT scans, thus being a good candidate for encoding chronological disease state. The ODE system defined on the embedding and logit space is numerically solved for the given time points in the future. The obtained logits corresponding to the future time point are then converted to the segmentation maps (Fig. 2) of the target anatomy. We utilize the method introduced in [29] for training which requires a constant amount of memory independent of the solver's step size and integration time.

A. Basic ODE Model

We aim to learn a continuous spatiotemporal model of disease progression from a longitudinal dataset comprising a series of n -dimensional images (3D in our case) for each subject, which are acquired across multiple visits at irregular time intervals. The relative time $t \in \mathbb{R}^+$ represents the time since the baseline and is normalized to a scale of 1 year. Thus, $t = 0.5$ and $t = 1.5$ corresponds to the 6 and 18-month time points respectively from the baseline visit at $t = 0$. The entire stack of the longitudinal volumetric data for each patient i is represented by a continuous function $x_i : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{R}$ such that $x_i(\omega, t)$ is the intensity value at the spatial coordinate $\omega \in \Omega$ in the image acquired at time t . The entire image of the patient acquired at time t is denoted by $x(\cdot, t)$. The ground-truth segmentation masks for each image in the training set can be similarly represented by a continuous ground-truth function $g : \Omega \times \mathbb{R}^+ \rightarrow [0, 1]$, that corresponds to the state of the target biomarker. The continuous functions x and g are not directly observable but sampled at specific time-points t corresponding to the visits during which the medical images were actually acquired. Our objective is to learn a continuous function $f : \Omega \times \mathbb{R}^+ \rightarrow [0, 1]$ to approximate g using a Deep Learning (DL) framework inspired by the NeuralODE [29]. The proposed model attempts to approximate f using only a single image $x(\cdot, 0)$ of the baseline visit as input. The estimated $f(\cdot, t)$ can then be sampled at arbitrary future time points to obtain the target anatomy segmentation in the future.

First, the image voxels of $x_i(\cdot, 0)$ are projected onto the latent embedding in the form of convolutional feature maps ξ_0 using a CNN with learnable parameters Θ_{CNN} . In this work, the 3D-UNet [35] was used for MRI and the 3D \rightarrow 2D PSC-UNet [36] was employed for OCT. Subsequently, the evolution of ξ_t over time is modeled as an IVP with the differential equation $\frac{\partial \xi}{\partial t} = p(\xi)$ and the initial condition $\xi = \xi_0$. The $p(\xi)$ is modeled using another CNN with three convolutional layers (see Fig 2 for details) with parameters Θ_{ODE} . This IVP can be numerically solved to obtain the latent feature embedding for any arbitrary future target time-point T and the corresponding evolving segmentation map is modeled as a linear transformation at each pixel (or voxel) of the 2D (or 3D) feature embedding. Thus, $f(\cdot, t) = \sigma(\phi_t)$, where $\phi_t = W^T \xi_t + b$ represents the unnormalized logits before applying the sigmoid activation to the linear transformation with the learnable parameters W and b . During training, the

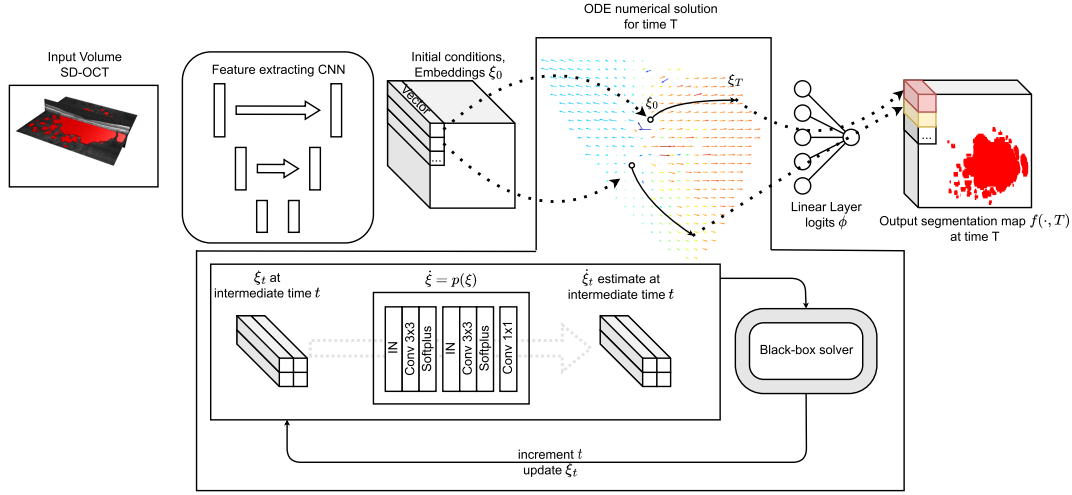


Fig. 2. Illustration of the proposed method. The input image at the baseline visit is first processed with a CNN to extract the feature embeddings ξ_0 for each pixel. The feature embedding ξ_0 acts as the initial condition for an ODE which evolves it over time to form a trajectory for each pixel in the phase space. The solution trajectory is shown on a phase plot of Eq. 2 projected on 2D plane using PCA, where blue represents completely healthy pixels, and red represents pixels with GA. The computed solutions ξ_T are projected with a linear layer and sigmoid activation to obtain the output segmentation map $f(\cdot, T)$. The abstract procedure of numerical integration is presented at the bottom of the figure. The learnable function $p(\xi)$ representing time derivative of ξ takes the form of a small CNN with 3 convolutional blocks. The computed embeddings ξ_t at intermediate time t and function $p(\xi)$ form an input to the black-box solver, which iteratively updates ξ_t .

backpropagation is performed as standard except for $\frac{\partial L}{\partial \Theta_{ODE}}$ and $\frac{\partial L}{\partial \xi_0}$, whose computations are based on [29] and described in Appendix. The training and inference are summarized in the Algorithms 1 and 2, where ODE is one of the three proposed variants discussed below, ODESolve is a black box ODE solver, and Θ_{ODE} are the learnable weights of the ODE.

Algorithm 1 Forward Pass, $f(\cdot, T, x(\cdot, 0), \Theta_{CNN}, \Theta_{ODE}, W)$

Require: Baseline image $x(\cdot, 0)$, target time T , parameters Θ_{CNN} , Θ_{ODE} and W

- Initial conditions $\xi_0 = \text{CNN}(x(\cdot, 0), \Theta_{CNN})$,
- Solve ODE, $\xi_T = \text{ODESolve}(\xi_0, p(\xi), 0, T, \Theta_{ODE})$,
 $\phi_t = W^T \xi_t + b$,
- Compute $f(\cdot, t)$ as $\sigma(\phi_t)$.

Algorithm 2 Backward Pass

Require: Baseline image $x(\cdot, 0)$, annotations $g(\cdot, 0)$ and $g(\cdot, T)$, pre-dimension $f(\cdot, 0)$, $f(\cdot, T)$, target time T , parameters Θ_{CNN} , Θ_{ODE} and W

- compute loss $L(f(\cdot, \cdot), g(\cdot, \cdot)) = \bar{D}(f(\cdot, \cdot), g(\cdot, \cdot))$ (Eq. 6),
- compute $\frac{\partial L}{\partial \xi_T}$ and $\frac{\partial L}{\partial W}$ as regular,
- define $a(t) = \frac{\partial L}{\partial \xi_t}$, set initial augmented state $s_0 = [\xi_T, \frac{\partial L}{\partial \xi_T}, 0]$,
- define $\text{aug_dynamics}([\xi_t, a(t), \cdot], t, \Theta_{ODE})$:
return $[p(\xi_t), -a(t)^T \frac{\partial p}{\partial \xi}, -a(t)^T \frac{\partial p}{\partial \Theta_{ODE}}]$,
- $[\xi_0, \frac{\partial L}{\partial \xi_0}, \frac{\partial L}{\partial \Theta_{ODE}}] = \text{ODESolve}(s_0, \text{aug_dynamics}, T, 0, \Theta_{ODE})$,
- compute $\frac{\partial L}{\partial \Theta_{CNN}}$ as regular,
- perform an update of Θ_{ODE} , Θ_{CNN} and W .

B. Variations of the Proposed ODE

In certain irreversible medical conditions, once a pathology occurs at a spatial location, the damaged tissue can never be

recovered. For example in GA [37], the affected tissue can no longer recover, and tumors [38], [39] are expected to enlarge in size without treatment. In such cases, the target biomarker can only increase in size over time. The proposed IVP-based formulation for DPM gives us a higher degree of control over the evolution of the disease. It allows the incorporation of additional domain-specific anatomical constraints by employing different systems of ODE in the IVP. In this section, we first summarize the basic IVP formulation without constraints as described previously. Next, we propose two modifications to the ODE system to enforce the *non-decreasing* constraints on the size of the target anatomy progression. The key idea of the described constraints is to force logits' derivative $\dot{\phi}$ to be non-negative. By following the reasoning described below, the proposed constraints can also be turned into non-positive ones to accommodate shrinking structures.

- *No constraints:* This simple approach does not impose any constraints on the way the biomarker segmentation evolves over time. In this case, ϕ_t is obtained as a linear transformation of the embedding feature space ξ that evolves over time as defined by the ODE

$$\dot{\xi} = p(\xi), \quad (1)$$

where the function $p(\xi)$ is modeled using a learnable CNN to predict $\dot{\xi}$, which is a shorthand notation for the time derivative $\frac{\partial \xi}{\partial t}$. The ODE is assumed to be autonomous or time-invariant implying that the function $p(\xi)$ is independent of time, ie., the CNN neither takes t as input nor does its network parameters change with time.

- *Constraints on the embeddings:* The non-decreasing target anatomy constraint can be incorporated into our ODE system by enforcing a non-negativity constraint on the derivative of ϕ with respect to time, ie., $\dot{\phi} \geq 0$. Given

the linear dependency of the logits $\phi = W^T \xi + b$, we need $\dot{\phi}$ to be non-negative. Taking the derivatives of both sides $\dot{\phi} = W^T \dot{\xi} = W^T p(\xi)$ we want to enforce constraints on $W^T p(\xi)$. Considering the unit vector $\hat{W} = \frac{W}{\|W\|}$, we can decompose $p(\xi)$ into orthogonal and collinear components $p(\xi) = W^T p(\xi) \hat{W} + (p(\xi) - \hat{W}^T p(\xi) \hat{W})$. We explicitly constrain the projection of $p(\xi)$ on \hat{W} in the collinear term to be non-negative by clipping its negative values to 0 using a rectified unit function denoted by $|\cdot|^+$. Thus, the proposed ODE becomes

$$\dot{\xi} = |\hat{W}^T p(\xi)|^+ \hat{W} + p(\xi) - \hat{W}^T p(\xi) \hat{W}. \quad (2)$$

- *Constraints on the logits:* In contrast to the above, where $\dot{\phi}$ and $\dot{\xi}$ had a linear dependency, we want to decouple these two variables and introduce a non-linear dependency between them. Thus, we introduce the time derivative of ϕ as a learnable function $q(\xi, \phi)$, and apply explicit non-negative constraint by means of rectified unit function. In this case, we introduce the following system:

$$\begin{cases} \dot{\xi} = p(\xi, \phi), \\ \dot{\phi} = |q(\xi, \phi)|^+, \end{cases} \quad (3)$$

where we used *softplus* as rectified unit function $|\cdot|^+$.

By introducing a non-negativity constraint on $\dot{\phi}$ we are effectively adding a non-decreasing constraint on the volume of the target biomarker over time. The target biomarker segmentation is obtained by thresholding ϕ , denoted using the indicator function as $I_{[\phi > 0.5]}(\phi)$. The corresponding biomarker volume is given by $V = \sum_{\omega \in \Omega} I_{[\phi > 0.5]}(\phi(\omega))$. The time derivative $\dot{V} = \sum_{\omega \in \Omega} \delta(\phi(\omega) - 0.5) \dot{\phi}$ is non-negative because the delta function δ is non-negative by definition and $\dot{\phi}$ is also constrained to be non-negative. As a result, the introduced pixel-value constraint guarantees the non-decreasing volume of the target anatomy over time.

During implementation, $p(\xi)$ is modeled with two consecutive residual blocks (Fig. 2) consisting of instance normalization [40], convolution with 3×3 filters and softplus activation followed by 1×1 convolution (for 2D, with natural extension to the 3D case). For implementing Eq. 3, we concatenate ξ and ϕ and the architecture of q is kept identical to p .

C. Loss Function

The proposed method is trained by obtaining $f(\cdot, t)$ from the ODE at specific discrete time points $\{0 \leq t \leq T\}$, corresponding to the actual patient visits for which the GT $g(\cdot, t)$ are available. The total loss at a subject level is then defined as the mean of the segmentation losses across all visits. The binary cross-entropy could be used at each time step:

$$\begin{aligned} \bar{D}(f, g) = & - \sum_{t=0}^T \sum_{\omega \in \Omega} g(\omega, t) \log(f(\omega, t)) \\ & + (1 - g(\omega, t)) \log(1 - f(\omega, t)) \end{aligned} \quad (4)$$

Another alternative is the soft Dice loss [41]:

$$\bar{D}(f, g) = 1 - \frac{1}{T} \sum_{t=0}^T \frac{2 \sum_{\omega \in \Omega} f(\omega, t) g(\omega, t)}{\sum_{\omega \in \Omega} f^2(\omega, t) + \sum_{\omega \in \Omega} g^2(\omega, t)} \quad (5)$$

Below, we approach the problem of temporal loss using functional analysis and derive the temporal Dice loss function. We define $f, g : \Omega \times \mathbb{R}^+ \rightarrow [0; 1]$, depending on time, and assume that f and g are continuous almost everywhere on $\Omega \times \mathbb{R}^+$, and that both integrals of f, g and their squares are finite. In turn, the space of such square integrable functions $F = \{f : \Omega \times \mathbb{R}^+ \rightarrow [0; 1]\}$ with inner product defined as $(f, g) = \int_{\Omega \times \mathbb{R}^+} f(\omega, t) g(\omega, t) d\omega dt$ is a Hilbert space.

We want our method to learn the function f by minimizing the respective squared distance $\|f - g\|^2$.

$$\|f - g\|^2 = \|f\|^2 + \|g\|^2 - 2(f, g) \leq \|f\|^2 + \|g\|^2,$$

due to non-negativity of (f, g) .

The minimum of $\|f - g\|^2$ is achieved at $f \sim g$: f and g being from the same equivalence class. Introducing a surrogate function $D(f, g)$, we limit it's the upper bound to 1: $D(f, g) = \frac{\|f - g\|^2}{\|f\|^2 + \|g\|^2}$. We can show that $D(f, g)$ has a minimum at $f \sim g$. As a result, by finding the minimum of $D(f, g)$ we also minimize the distance $\|f - g\|^2$. Next, we discretize the space Ω and time T , thus getting the definition of Soft Dice loss [41] from $D(f, g)$ with temporal objective:

$$\begin{aligned} \bar{D}(f, g) = & 1 - \frac{2 \sum_{t=0}^T \sum_{\omega \in \Omega} f(\omega, t) g(\omega, t)}{\sum_{t=0}^T \sum_{\omega \in \Omega} f^2(\omega, t) + \sum_{t=0}^T \sum_{\omega \in \Omega} g^2(\omega, t)}. \end{aligned} \quad (6)$$

Effectively, this implies that we should compute Dice loss not for individual visits independently, but for the entire stack of visits. In the experiments, as part of ablation, we compare the temporal Dice formulation with the other two loss functions and study their influence on performance.

III. EVALUATION

A. Data

Table I provides an overview of the three datasets used to develop and evaluate the proposed method. The private MUV-GA dataset is used to evaluate the performance on the task of predicting the segmentation maps for GA at future time points from a retinal OCT or Fundus Autofluorescence (FAF) of the first baseline visit. The two public datasets TADPOLE [10] and ADNI1 are employed for the task of predicting the changes in the ventricle volume over time, which is an important biomarker for tracking AD progression. We selected ventricle volume over the hippocampus as it is also a relevant biomarker, which has been demonstrated by a number of studies [9], [42], [43]. In addition, the ventricle volume is one of the endpoints of the TADPOLE [10] challenge, which allowed us to compare our performance with independent algorithms developed by the teams participating in the challenge.

1) *MUV-GA*: is a longitudinal in-house dataset from the Department of Ophthalmology at Medical University of Vienna [44] which originated from an observational study of natural GA progression. It consists of 3D OCT scans of the eyes of patients older than 50 years, diagnosed with GA, and imaged with Spectralis (Heidelberg Engineering) device. Study visits occurred every 3 months. The patients were diagnosed

TABLE I
AN OVERVIEW OF THE LONGITUDINAL DATASETS. OF NOTE,
IN MUV-GA, BOTH EYES WERE IMAGED
AT A VISIT WHEN POSSIBLE

Dataset	Scan	#Patients (#Eyes)	Total #scans	#Visits per patient	Avg. time span (days)
MUV-GA	OCT, FAF	100 (184)	967	5.95	970
ADNI	MRI	633	2537	4.01	704
TADPOLE D3-D4	MRI	143	286	2	1082

with GA secondary to AMD and were excluded from the study when other confounding pathologies were found. It consists of 3D retinal OCT and 2D FAF scans of subjects older than 50 years who have been diagnosed with GA. Longitudinal scans were acquired for both eyes of each subject using the SPECTRALIS (Heidelberg Engineering, DE) device. The ground truth pixel-level annotations of GA lesions were first performed manually, using an in-house built reading center-certified annotation tool (OCTAVO, Vienna Reading Center). Following the definition of Arikian et al. [46], the area of atrophy was defined as well-demarcated area with a very low or extinguished degree of autofluorescence. The annotations on the corresponding 2D FAF scans were done by a retinal expert at all available visits [44]. As the final step, the annotated FAF scans were registered to the en-face infrared images of OCT to obtain the corresponding enface OCT annotations [45]. The dataset features both missing visits and irregular time intervals between visits. While the OCT scans are 3D, the reference standard for the GA segmentation is available as a 2D en-face binary map, as the pathology corresponds to the absence of the layers at the given location. The example of the en-face lesion is depicted in Fig. 1. Thus, the CNN architectures that can be employed for this task are limited to 3D→2D networks, such as [36], which reduce the dimensionality of the output predictions.

Despite the number of challenges, such as 3D→2D reduction, we focus on the building progression models from 3D OCT data rather than 2D FAF. Even though FAF primarily reflects the autofluorescence of the RPE layer making it ideal to segment GA which is primarily characterized by RPE loss, FAF is not widely available outside large clinical centers, and its blue light is uncomfortable to the patients during image acquisition. In contrast, OCT is more widely available in clinics and more comfortable to patients. Moreover, it is established as a standard of care for monitoring AMD and provides much richer 3D information on the morphology of all intra-retinal tissue layers, and not just the state of RPE. Especially, its cross-sectional slices offer a wealth of information on the status of retinal layers and photoreceptors in the junctional zone around the atrophy (loss of RPE cells). Photoreceptor degeneration in the junctional zone has already been shown to be an early indicator of disease progression [46]. We assume that the additional morphological information, not present on FAF, can improve predictive capabilities, and, in the future,

contribute towards biomarker discovery. In terms of precision of the measurements, Mai et al. [47] have demonstrated that OCT measurements are consistent with FAF measurements of GA size and growth rate.

a) Preprocessing: First, we resample the scans to have $5.671\mu\text{m}$ spacing across the BScans. Then, we register the follow-up visits. To control and fix the alignment of the scans and corresponding annotations between visits, we first perform retinal blood vessel segmentation [36], and then use phase correlation algorithm [48] on the blood vessels to estimate the shift between the given image and the baseline scan. After the registration of the follow-up visits is done, we run layer segmentation algorithm [49] and flatten the retinal appearance according to the outer RPE boundary. Flattening helps the methods to focus on retinal layer appearance and structure, rather than spatial position and inclination.

2) ADNI: ADNI¹ is an open-access dataset - part of the initiative launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The dataset was created for measuring the progression of mild cognitive impairment and early Alzheimer's Disease. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease.

TADPOLE [10] is a challenge guided towards the accurate prediction of AD progression. It aims to predict AD progression by estimating the number of biomarker values. In this work, we were specifically interested in the imaging biomarkers, thus we focused on the prediction of the ventricle volume. The TADPOLE challenge has a series of datasets, derived from the ADNI database. In this study, we use only **D3 and D4** datasets, designed for making predictions from just a single baseline scan, where D3 contains a single visit for a patient, and D4 plays a role of the test set with the follow-up measurements. We use TADPOLE data exclusively as a test set and train on **ADNI** data.

As the proposed method requires imaging data to train, we query T1 MRI scans of the patients participating in ADNI1 and having at least 2 visits. Due to the absence of the manual ground truth ventricle annotation, we obtain pseudo-label segmentations automatically using FastSurfer [50] at all available time points following the protocols used in the TADPOLE challenge. We then trained the baseline methods and the proposed method using such pseudo-label segmentations and tested using TADPOLE D3-D4 data. Both the input MRI images and the corresponding segmentation labels are in 3D. Similar to MUV-GA, TADPOLE has missing visits and irregular time intervals between visits.

a) Preprocessing: The MRI scans were first imported to the FreeSurfer, normalized [51] and skull-stripped [52]. Then, FastSurfer [50] was used to process the scans and segment the brain structures. For training, follow-up visits of the same patients were aligned using the longitudinal stream of FreeSurfer [53]. For alignment, several processing

¹adni.loni.usc.edu

steps, such as skull stripping, Talairach transforms, atlas registration, spherical surface maps, and parcellations are then initialized with common information from the within-subject template [53], [54]. Since we explicitly focus on the evolution of the ventricle volume, we discard all other labels. The preprocessed scans have the size of $256 \times 256 \times 256$. To reduce the memory burden for the storage and data processing, after FreeSurfer processing we additionally crop the cube of size $128 \times 128 \times 128$. We center the cropped region on the ventricles of the baseline visit. The center is calculated as a mean of the ventricles segmentation as if it were a probability density function. The follow-up scans are cropped with the same region from the baseline. For the new test subjects, only processing of the baseline scan is required, thus longitudinal alignment step is omitted.

B. Experiments

1) *Data Split*: To assess the performance of the proposed method, we run a series of experiments for each dataset. In our experiments, we utilize 5-Fold Cross Validation (CV) and split patient-wise **MUV-GA** and **ADNI** datasets into five groups, where four folds are used as a train-val dataset, and one fold is used as a test dataset with ratios of 72%+8%, 20% for train+val and test. We use **TADPOLE D3-D4** exclusively as a test set, and those scans don't overlap with **ADNI**, however, they contain follow-ups of the same patients.

2) *Baselines*: Since the datasets represent different tasks, we compare the proposed method with different baselines depending on the application.

a) *Oracle*: Each of the datasets has a corresponding *Oracle* model as an estimate of the upper bound that a predictive model can achieve. To obtain an *Oracle* model, we directly segmented both the baseline and the endpoint scan. For this, we trained PSC-UNet [36] for automated GA segmentation, and UNet3D [35] to perform an automated segmentation of the brain ventricles. For GA segmentation, we employed a model that performs dimensionality reduction taking 3D volume as an input and producing 2D en-face mask. As highlighted by Lachinov et al. [36], dimensionality reduction methods deliver better performance in the task of GA segmentation.

b) *Geographic Atrophy*: We compare the proposed method against the approach introduced in Zhang et al. referred as MS-DCNN [17], Nguyen et al. [12], Gigon et al. [19] and the *Oracle* model. The nature of the task of predicting growth on 2D en-face masks from 3D volumetric data ($3D \rightarrow 2D$) poses an additional challenge and limits a set of baseline methods that can be utilized. In the paper, we validated the methods that use 2D input in the form of OCT projections or layer thickness maps [17], [19], the mixed approaches [12], [36], and purely 3D input/output [17]. As a standalone experiment, to show that the approach works on both 3D OCT and 2D FAF scans, we applied the proposed method to the FAF scans of the same patients. The performance of progression prediction is evaluated at the follow-up visit, one year from the baseline scan.

c) *ADNI & TADPOLE*: For ADNI data, we compare the proposed method with Nguyen et al. [12], Petersen et al. Deep Glioma Growth (DGG) [26] and *Oracle* model. The

prediction performance is evaluated at the 3-year follow-up visit. For TADPOLE, we evaluated the method using only cross-sectional tasks, namely D3-D4 datasets. For comparison, we additionally include the methods from TADPOLE challenge such as *GlassFrog*, *BravoLab*, *EMCI* [22]. To put the proposed method into perspective, we also included methods from the longitudinal task, where methods could use more than a single scan to form a prediction. These methods include *EMCI* [22] and Gruffaz et al. [23].

3) *Ablation Experiments*: We ran a series of ablations experiments, to investigate:

a) *Constraints in the ODE system*: we train the proposed method with three different ODE systems, one introducing no constraints (Eq. 1), one constraining the embedding space (Eq. 2), and one constraining the logits space (Eq. 3). For this ablation experiment, we employed only **MUV-GA** OCT dataset, since GA can only grow in size as the atrophic tissue cannot recover, which is not necessarily the case with brain ventricles. As a loss, we employed the function described in Eq. 6. To demonstrate the applicability of the method for different scenarios, we applied non-increasing constraints on embedding space (Eq. 2) for the Hippocampus progression model using **ADNI** data (Appendix B.2).

b) *Loss functions*: for learning temporal segmentation objective we investigate the performance of the loss function derived from a functional distance (Eq. 6), compared to the alternative standard loss functions (Eq. 4, and 5). We perform 5-fold cross-validation experiments using the proposed method on both **MUV-GA** and **ADNI** datasets and further apply the models trained on **ADNI** to **TADPOLE** hold-out test set. For **MUV-GA** dataset, we employed the ODE system with constraints defined by Eq. 2. For **ADNI** data, we used the ODE system described by Eq. 1.

4) *Evaluation Metrics*: For cross-validation experiments on **MUV-GA** and **ADNI** datasets, we report: Dice index of baseline and predicted visit, Hausdorff distance of the baseline and the last visit, coefficient of determination R^2 and coefficient of correlation Pearson's r of the change in area or volume. For **MUV-GA**, we additionally report the mean squared error (MSE) of the change in square root transformed area in mm. For **ADNI** and **TADPOLE**, we report the metrics utilized in TADPOLE [10] challenge. Namely, MAE, WES, calculated as a percentage of intracranial volume, and CPA:

$$MAE = \frac{1}{N} \sum_{i=1}^N \frac{1}{ICV_i} |V_i - \bar{V}_i|,$$

$$WES = \frac{\sum_{i=1}^N \frac{1}{\bar{V}_i^{0.75} - \bar{V}_i^{0.25}} |V_i - \bar{V}_i|}{\sum_{i=1}^N \frac{1}{ICV_i} \frac{1}{\bar{V}_i^{0.75} - \bar{V}_i^{0.25}}},$$

$$CPA = |0.5 - \frac{1}{N} \sum_{i=1}^N 1_{[\bar{V}_i^{0.25}; \bar{V}_i^{0.75}]}(V_i)|,$$

where V and \bar{V} are predicted ventricle volumes, $\bar{V}_i^{0.25}$ and $\bar{V}_i^{0.75}$ are 25% and 75% percentile of the predicted volume of an individual patient, together forming 50% CI, ICV is the intracranial volume, and 1 is an indicator function.

TABLE II

CROSS-VALIDATION EXPERIMENTS ON MUV-GA DATASET WITH 3D OCT VOLUMES AND 2D FAF IMAGES. THE METRICS⁴ WERE COMPUTED WITH RESPECT TO THE BASELINE AND THE 1-YEAR VISITS. IN CONTRAST TO THE OTHER METHODS, THE ORACLE MODEL DOES NOT PERFORM A FUTURE PREDICTION BUT EMPLOYS THE SCAN AT THE 1-YEAR VISIT TO HIGHLIGHT AN UPPER LIMIT OF THE PERFORMANCE THAT CAN BE ACHIEVED

Method	Dice ₀	Dice _t	HD ₀ ⁹⁵	HD _t ⁹⁵	R ²	r	MAE (in mm)
Optical Coherence Tomography (OCT)							
Oracle: PSC-UNet [36]	0.82 ± 0.15	0.87 ± 0.13	0.16 ± 0.21	0.10 ± 0.28	0.73	0.86	0.46 ± 0.59
Zhang et al. [17]	0.62 ± 0.25***	0.70 ± 0.21***	1.50 ± 1.00***	1.05 ± 0.74***	-0.18***	0.52***	1.19 ± 0.96***
Zhang et al. 3D [17]	0.76 ± 0.21***	0.81 ± 0.14**	0.78 ± 1.12***	0.47 ± 0.69***	-0.29***	0.38***	1.15 ± 1.10***
Nguyen et al. [12]	0.80 ± 0.17	0.84 ± 0.11	0.30 ± 0.54	0.19 ± 0.29	0.28*	0.66*	0.89 ± 0.80***
Gigon et al. [19]	0.76 ± 0.22**	0.79 ± 0.18****	0.57 ± 0.89**	0.78 ± 1.12***	0.36	0.61*	0.85 ± 0.75***
Proposed	0.80 ± 0.19	0.84 ± 0.13	0.30 ± 0.54	0.21 ± 0.30	0.59	0.78	0.64 ± 0.63
Fundus Autofluorescence (FAF)							
Proposed	0.85 ± 0.19	0.83 ± 0.18	0.53 ± 0.7**	0.52 ± 0.41***	0.53	0.74	0.68 ± 0.68

In addition, we performed a statistical analysis of the results. For Dice, Hausdorff distance, and MAE we utilized the one-sided Mann-Whitney U test. For R^2 and r we performed a random permutation test. The significance level is denoted as * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$.

C. Training Details

Each dataset was split into five folds. For each method, we performed a hyperparameter optimization, described in the Appendix, based on the validation subsplit of the training folds. We chose between ADAM and SGD optimizers, a number of learning rates, weight decay, and batch size. For the development We used PyTorch framework, and we trained all the models using NVidia 2080Ti GPU. Training of each model took approximately 24h.

a) *Proposed method:* We used the function defined by Eq. 6 as a loss function. For solving the ODE system, we used RK4 scheme with step size $h = \frac{1}{12}$ equivalent to a one-month interval. For optimization, we used ADAM optimizer with a learning rate of 10^{-3} and batch size of 2 for 3D data, and 8 for 2D FAF images, found with hyperparameter optimization for all 3 variants of the proposed method. For MUV-GA, the patch size was set to $48 \times 128 \times 512$ for OCT, 256×256 for FAF, and $96 \times 96 \times 96$ for ADNI. For NeuralODE part, we used the code provided by Chen et al.² During training the proposed method takes a constant amount of GPU memory, independent of the time interval over which the ODE is evolved by using the adjoint method [29] for backpropagation. This allowed our method to be trained over large time intervals with limited GPU memory.

b) *Baseline methods:* For MS-DCNN [17], we removed dropout and added InstanceNorm [40], which greatly improved training stability and performance. We also extended the method of Zhang et al. to handle 3D OCT volumes instead of en-face projections, it is referred to as MS-DCNN 3D. We extended the approach of Nguyen et al. [12] to handle the imaging data instead of tabular features, introducing convolutional filters into the recurrent blocks. The work of Gigon et al. [19] was implemented as in the paper, however, we used [49] for layer segmentation. For the implementation of the Deep Glioma Growth [26], we used code provided by

Petersen et al.,³ with the exception that only the baseline image and its segmentation was used for generating the prior.

IV. RESULTS

A. Progression of Geographic Atrophy

A qualitative result of our method is shown in Fig. 1, and the quantitative cross-validation results are provided in Table II. The top row represents the segmentation performed with *Oracle* PSC-UNet [36] using baseline and 1-year images. The proposed method clearly outperformed the baseline methods on all the metrics with the exception of Nguyen et al. [12], where they both achieved a mean Dice score of 0.8 and 0.84 for predicting the atrophy segmentation at baseline and 1-year visits, respectively. However, the proposed method performed significantly better in estimating the atrophy growth, achieving $R^2 = 0.59$, $r = 0.78$, and $MAE = 0.64$ mm using OCT data and $R^2 = 0.53$, $r = 0.74$, $MAE = 0.68$ using FAF data, which is very close to the performance of the *Oracle* model. The model trained on FAF images showed slightly worse results in growth estimation and forecasting of the segmentation at year 1, which can be attributed to a lesser amount of morphological information compared to OCT. The similar performance of the proposed method and the RNN-based solution of [12] may be attributed to the fact that both belong to the same family with the RNN model being a discrete approximation of the NeuralODE. The RNN can be represented as NeuralODE with a less sophisticated explicit Euler method for numerical integration at discrete time steps.

The recently introduced method of Gigon et al. [19], exploring the idea of Taylor series decomposition, performs a bit worse both in segmentation and prediction. The method of Zhang et al. [17] failed in the quantitative prediction of the lesion change, at the same time showing moderate performance in terms of segmentation.

B. Progression of Alzheimer's Disease

A qualitative result of our method for predicting the ventricle segmentation in brain MRI at future time points is depicted in Fig. 1. All the models achieved a high segmentation performance of the baseline visits, with Dice score ranging from

²<https://github.com/rtqichen/torchdiffeq>

³<https://github.com/MIC-DKFZ/image-time-series>

TABLE III

CROSS-VALIDATION EXPERIMENTS ON ADNI DATASET. THE METRICS⁴ WERE COMPUTED WITH RESPECT TO THE BASELINE AND THE 3-YEAR VISITS. THE ORACLE MODEL HIGHLIGHTS THE PREDICTIVE CEILING THAT CAN BE ACHIEVED

Method	Dice ₀	Dice _t	HD ₀ ⁹⁵	HD _t ⁹⁵	R ²	r	MAE, % ICV	WES	CPA
Oracle: 3D-UNet [35]	0.97 ± 0.02	0.97 ± 0.2	1.08 ± 4.20	1.28 ± 4.32	0.98	0.99	0.06 ± 0.10	0.05	0.46
Nguyen et al. [12]	0.96 ± 0.02***	0.93 ± 0.04***	1.46 ± 4.31**	1.90 ± 5.07**	0.32	0.58	0.24 ± 0.20*	0.23	0.04
Petersen et al. [26], mean sample	0.95 ± 0.02***	0.91 ± 0.04***	1.10 ± 4.20***	1.38 ± 4.40***	-0.11**	0.34**	0.24 ± 0.26***	0.19	0.25
Petersen et al. [26], best sample	0.95 ± 0.02***	0.91 ± 0.04***	1.10 ± 4.20***	1.38 ± 4.30***	-0.09**	0.35**	0.24 ± 0.26***	0.18	0.25
Proposed	0.97 ± 0.02	0.94 ± 0.03	1.39 ± 4.21	1.63 ± 4.63	0.36	0.61	0.21 ± 0.19	0.20	0.08

TABLE IV

RESULTS FOR TADPOLE D4 TEST SET, EVALUATED USING A SINGLE BASELINE SCAN. THE SEGMENTATION PERFORMANCE IS EVALUATED⁴ ONLY AT THE BASELINE SCAN AS THE FOLLOW-UP SCANS ARE NOT AVAILABLE

Method	Dice ₀	HD ₀ ⁹⁵	MAE, % ICV	WES	CPA
Nguyen [12]	0.96 ± 0.01	1.07 ± 0.46	0.46 ± 0.85**	0.39	0.02
Petersen [26], mean sample	0.95 ± 0.01*	1.0 ± 0.4	0.56 ± 0.85***	0.4	0.2
Proposed	0.96 ± 0.01	1.05 ± 0.44	0.44 ± 0.83	0.35	0.08
GlassFrog-LCMEM-HDR [22]	-	-	0.48	0.38	0.24
BravoLab [22]	-	-	0.64	0.64	0.42
GlassFrog-Average [22]	-	-	0.68	0.55	0.24
EMC1-Std [22]	-	-	0.80	0.62	0.48
Methods utilizing multiple prior visits					
Gruffaz [23]	-	-	0.42	-	-
EMC1-Std [22]	-	-	0.41	0.29	0.43

0.95 of Petersen et al. [26] to 0.97 of the proposed method (Table III). It's important to note that we used automatically generated annotations by FastSurfer, which potentially might not fully reflect the complexity of the ventricle anatomy and might be easy to learn by automated methods. The proposed method appeared to be significantly better than all other methods in terms of Dice and HD95. The method also achieved the best performance in predicting segmentation at a 3-year visit in terms of R², r, and MAE. The other methods demonstrated similarly good results in WES and CPA. As Petersen et al. claim, the Continuous-Time Deep Glioma Growth model wasn't designed for precise prediction of the changes, but rather for their probabilistic description. However, the Deep Glioma Growth model achieved good results in predicting the changes, with mean prediction being close to the best sample from the predicted segmentation mask distribution w.r.t 3-year Dice score.

The TADPOLE hold-out test set consists of on average longer forecasting intervals, testing the extrapolation capabilities of the methods. Quantitatively (Table IV), the proposed and baseline models demonstrated good generalization in baseline visit segmentation, achieving the Dice score of 0.95 or above and showing a mean HD⁹⁵ smaller than 1.07 mm. Importantly, as in the cross-validation exper-

TABLE V

ABLATION EXPERIMENTS OF EXPLICIT DOMAIN-SPECIFIC CONSTRAINTS. EQ. 1 CORRESPONDS TO THE SYSTEM WITHOUT CONSTRAINTS. TO ENCODE THAT THE LESION ONLY GROWS OVER TIME, EQ. 2 USES EXPLICIT CONSTRAINTS IN THE EMBEDDING SPACE, WHILE EQ. 3 APPLIES THE CONSTRAINTS ON THE LOGITS⁴

ODE	Dice ₀	Dice _t	R ²	r	MAE, mm
Eq. (1)	0.79 ± 0.19	0.83 ± 0.15	0.44	0.72	0.78 ± 0.69**
Eq. (2)	0.8 ± 0.19	0.84 ± 0.13	0.59	0.78	0.64 ± 0.63
Eq. (3)	0.8 ± 0.19	0.84 ± 0.14	0.56	0.76	0.68 ± 0.63

iments, the proposed method excelled in accurate disease evolution prediction, surpassing the challenge contenders in all metrics. It significantly outperformed the TADPOLE participants, in the prediction of the future ventricle size achieving 0.44 MAE. The proposed method also reached 0.35 WES, and demonstrated a competitive 0.02 CPA. The method of Petersen et al. [26] showed competitive predictive performance, where we sampled the mean prediction since we were interested in a single prediction output. In addition, we added a comparison illustrating the distance between the methods utilizing multiple prior scans and the methods utilizing a single scan, including the proposed method. As described in the TADPOLE results [22], the gap between the two mentioned classes is modest, and it's further reduced by the proposed method reaching MAE of 0.44 and Gruffaz et al. [23] achieving 0.42 MAE. In general, all of the imaging-based methods demonstrated competitive results, with the proposed method surpassing the challenge contenders in the cross-sectional task on all the metrics.

C. Ablation Experiments

To investigate the effect of the explicit constraint enforcement as a part of the ODE system, we compared three different ODE systems defined by Eq. 1-3 (Table V) on the MUV-GA dataset. The effect of the constrained ODE system was slightly pronounced on the segmentation performance of the baseline and 1-year scans as both ODE systems with constraints (Eq. 2 and 3) performed slightly better than the unconstrained system. However, for the GA growth estimation, the effect of the constraints was more pronounced, where the system with the constraints in the embedding space (Eq. 2) performed clearly the best with the significantly lower MAE compared to the unconstrained counterpart. The Eq. 2 assumes linearity of

⁴* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$

TABLE VI

ABLATION EXPERIMENTS OF THE PROPOSED METHOD TRAINED WITH DIFFERENT LOSS FUNCTIONS. EQ. 4 REPRESENTS BINARY CROSS-ENTROPY LOSS, EQ. 5 REFLECTS A SOFT DICE LOSS APPLIED ON EACH INDIVIDUAL SCAN AND THEN AVERAGED, EQ. 6 REPRESENTS THE DICE LOSS DESCRIBED IN THE PAPER, CONSIDERING ALL SCANS AS A SINGLE 4D VOLUME⁴

Dataset	MUV-GA, 5-Fold CV					ADNI, 5-Fold CV					TADPOLE D4		
	Dice ₀	Dice _t	R ²	r	MAE, mm	Dice ₀	Dice _t	R ²	r	MAE, % ICV	MAE, % ICV	WES	CPA
Eq. (4)	0.73 ± 0.22***	0.77 ± 0.18***	0.46	0.68	0.71 ± 0.74*	0.97 ± 0.02	0.94 ± 0.03	0.36	0.6	0.22 ± 0.19	0.46 ± 0.8*	0.36	0.12
Eq. (5)	0.78 ± 0.18	0.83 ± 0.13	0.58	0.76	0.67 ± 0.61	0.97 ± 0.02	0.94 ± 0.03	0.36	0.6	0.22 ± 0.19	0.45 ± 0.82	0.36	0.05
Eq. (6)	0.8 ± 0.19	0.84 ± 0.13	0.59	0.78	0.64 ± 0.63	0.97 ± 0.02	0.94 ± 0.03	0.36	0.61	0.21 ± 0.19	0.44 ± 0.83	0.35	0.08

the embedding space w.r.t the progression, while Eq. 3 does not. In the provided experiment both constraint variants performed well, however, we conjecture that Eq. 3 will perform better for tasks with complex progression patterns.

The effect of different loss functions on the predictive performance is reported in Table VI. In all experiments, the binary cross-entropy function performed the worst, except in ADNI cross-validation, where all loss functions performed equally well. This can be attributed to the nature of the dataset's reference standard, which was produced by an automated segmentation method. Both Dice-based losses performed well, however, the loss we derived from the functional distance (Eq. 6), performed better in GA. The difference between the proposed loss and the binary cross-entropy was statistically significant across most of the metrics. In the comparison with the Dice loss, we have not observed a significant difference, however, the p-values on MUV-GA and TADPOLE data were in the range (0.05; 0.1) for Dice and MAE, and in (0.1; 0.4) for r and R^2 . We believe that with the increased number of samples in the datasets, or on the more challenging tasks, the proposed loss function will clearly outperform the baselines. Thus we recommend using Eq. 6 for working with temporal data.

D. Limitations

The paper comes with a number of limitations, which will be addressed in future work. The main drawback of the proposed method is that it cannot model the progression of Electronic Health Records data. The method can potentially be extended to model the progression of such tabular data by introducing an auxiliary task through exploring the latent space produced by the method. However, it poses a non-trivial problem and requires extensive investigation. Addressing the progression modelling of segmentation inferred biomarkers, the method often produces an embedding space with a high number of steady states, to which the solution converges. In the experiments, we observed that they are located at the borders of the embedding space and reachable with a large value of time. This puts a certain limit on the forecasting potential of the method. Analyzing and regularizing the embedding space is a promising direction for future work. Additionally, the proposed method can only process a single scan, instead of a series of scans representing the patient's history. At the same time, the ability to provide a robust disease progression model from a single scan is a strength, as it can be performed at the time of the screening visit. Furthermore, the datasets employed in the study are potentially biased with respect to age, gender, and scanning protocols. The bias of the scanning protocol is partially addressed by data processing techniques

with image registration and intensity normalization. As we use fairly large longitudinal datasets, we did not explore the performance of the method trained with fewer samples. Lastly, one of the evaluation tasks contains a limitation inherited from the TADPOLE [22] challenge, namely, we utilized automated segmentation of the ventricles on MRI as a reference. In contrast, GA was annotated manually by a retinal expert, which is considered a gold standard.

V. CONCLUSION

In this work, we introduced a DPM to predict the future progression of the morphological changes in pathologies (such as GA in retinal OCT) and clinically relevant anatomical structures (such as the ventricles in brain MRI) from a single scan of the baseline visit. Our method produces the segmentation maps for the desired object of interest for varying future time-points which is clinically useful in the early detection of fast progressors for personalized treatment and the recruitment of patients in clinical trials.

In contrast to the existing methods, the proposed method employs a continuous time-invariant Neural-ODE formulation to constrain the solution space and can be trained on datasets with missing data and variable time intervals between the visits. We demonstrated two approaches to constrain the solution to the physiologically plausible ones and introduced temporal Dice loss for training with longitudinal data. The state-of-the-art performance of our method on the two distinct tasks of predicting the progression of GA in retinal OCT and the ventricles in brain MRI demonstrates that the proposed methodology can be applied to disease progression modeling in a variety of settings involving different anatomical structures and imaging modalities.

In this study, we focused on the task of providing predictions from a single scan, targeting the scenario of patient recruitment for clinical trials, where typically only a screening scan is available. In addition to enriching the trials with fast progressors, our method could serve as a control digital twin of a patient receiving a treatment to predict the expected outcomes without the treatment, thereby reducing the need for costly sham arms in clinical trials. An extension of this work to utilize multiple prior scans and update the predictions as the new data becomes available offers a promising direction for future research

APPENDIX

A. Method Implementation Details

1) *Backpropagation Through the NeuralODE*: For calculating the derivatives of the network at a given time point, we follow

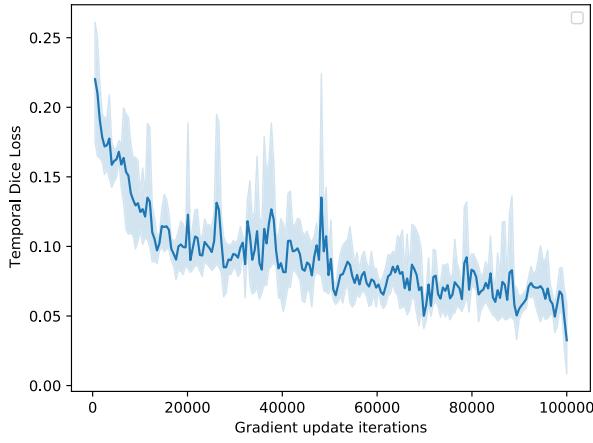


Fig. 3. Training loss curve of the proposed method. The blue region stands for the standard deviation derived from the five-fold cross-validation.

the method described by Chen et al. [29]. We are interested in finding $\frac{\partial L}{\partial \Theta_{ODE}}$ and $\frac{\partial L}{\partial \xi_0}$. Following adjoint method of NeuralODEs, we define $a(t) = \frac{\partial L}{\partial \xi_t}$ and, as shown in [29], $\frac{da(t)}{dt} = -a(t)^T \frac{\partial p(\xi_t, \Theta_{ODE})}{\partial \xi}$. After introducing IVP with $a(T) = \frac{\partial L}{\partial \xi_T}$ and $\frac{da(t)}{dt}$ and solving it in reverse time sequence from T to 0 using ODEsolve, we obtain $\frac{\partial L}{\partial \xi_0}$. A similar augmented state for Θ_{ODE} is introduced as $a_{\Theta_{ODE}}(t) = \frac{\partial L}{\partial \Theta_{ODE}(t)}$ and $\frac{da_{\Theta_{ODE}}(t)}{dt} = -a(t)^T \frac{\partial p(\xi_t, \Theta_{ODE})}{\partial \Theta_{ODE}}$. Solving these IVPs requires knowing ξ_t along the entire trajectory from 0 to T. Instead of storing the intermediate results from the forward pass, we recompute the values by solving the IVP for ξ backward, alongside the augmented IVPs. Next, we form an ODE system:

$$\begin{bmatrix} \frac{d\xi}{dt} \\ \frac{da}{dt} \\ \frac{da_{\Theta_{ODE}}}{dt} \end{bmatrix} = \begin{bmatrix} p(\xi_t, \Theta_{ODE}) \\ -a(t)^T \frac{\partial p(\xi_t, \Theta_{ODE})}{\partial \xi} \\ -a_{\Theta_{ODE}}(t)^T \frac{\partial p(\xi_t, \Theta_{ODE})}{\partial \Theta_{ODE}} \end{bmatrix}$$

Following this, we define the initial state as $[\xi_T, \frac{\partial L}{\partial \xi_T}, \frac{\partial L}{\partial \Theta_{ODE}}]$, and solve the IVP backwards in time. Since we are not using the intermediate time points in the training, we can obtain the solutions with a single call to the solver.

2) Simultaneous Training of NeuralODE and CNN: During the model development, we faced a series of computational challenges, revolving around the training stability of the proposed model. When training from scratch, the network sometimes converged to local minimums representing population average progression and formed attractor points where the embedded space-time trajectories merged leading to a high rate of misclassification or they diverged, depending on the randomly initialized weights. Analyzing the results, the trained network produces highly clustered embedding space with many attractor points. Even though NeuralODE captures the dynamics of the disease evolution, it fails to keep the embedding space linearly separable, making it impossible for linear classification layers to classify individual pixels. In the implementation, we addressed this issue by pretraining of the segmentation backbone on the training dataset, thus giving good initializations for the prediction training. The resulting training curve of the method is plotted in Fig. 3. To make a fair comparison, we equally pretrained all the baseline backbones.

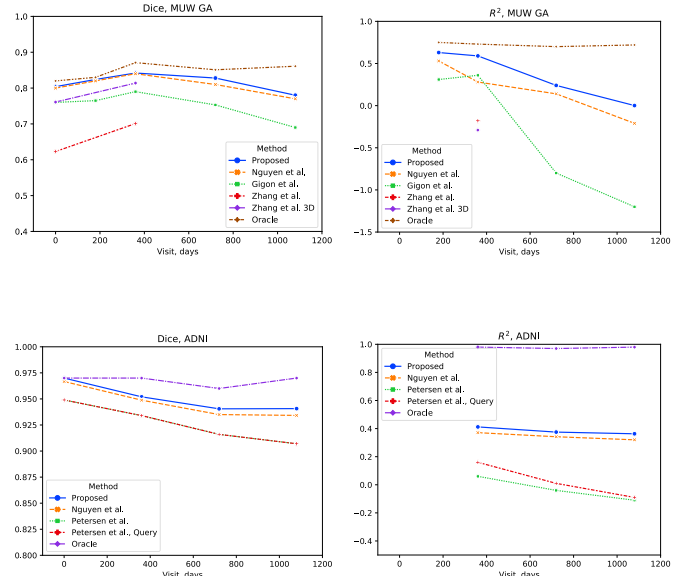


Fig. 4. The result for intermediate visits. The top row corresponds to the models trained on MUV-GA data with the 1-year target interval. The evaluation, however, was made even for longer time intervals of up to 3 years. The second row corresponds to the models trained on ADNI data with 3 years target interval.

During the development, we also tried methods described by Massaroli et al. [55], however, none of them produced superior results to the vanilla NeuralODEs.

To facilitate batched training, we sort all unique time points within the batch and construct a list of the observations (patients) for each time point. If the observation for the requested patient and time is missing, we skip it for the loss calculation. This technique was also employed in [33], [56]. It allows solving underlying ODEs for each batch element jointly and scales linearly with the batch size.

3) Hyperparameter Optimization: To ensure a fair comparison, in all the experiments, including baseline methods, we ran a hyperparameter optimization using grid search with the following values:

- Optimizer: Adam, SGD
- Learning rate: Adam $[10^{-3}, 10^{-5}]$, SGD $[10^{-1}, 10^{-2}]$
- Weight decay: 0, 10^{-5}
- Batch size: 1, 2, 4
- NODE only - Solver: RK4 step 1/4 or 1/12, dopri5

Due to the extensive volume of the experiments, we only performed the hyperparameter optimization on the first fold and transferred the hyperparameters to the training of the other folds. For the proposed method, the set of hyperparameters was: Adam optimizer with a learning rate of 10^{-3} , no weight decay, batch size of 2, RK4 scheme with step size $h = \frac{1}{12}$.

B. Extended Results

To evaluate the interpolation and extrapolation capabilities of the methods, we calculated the Dice score and R^2 of the predicted growth changes for the follow-up visits, with the results presented in Fig. 4.

1) MUV-GA: On this dataset, the models were trained to forecast a 1-year time interval. The results beyond the 1-year time interval represent how well the model can extrapolate and handle large forecasting intervals. Zhang et al. [17] methods

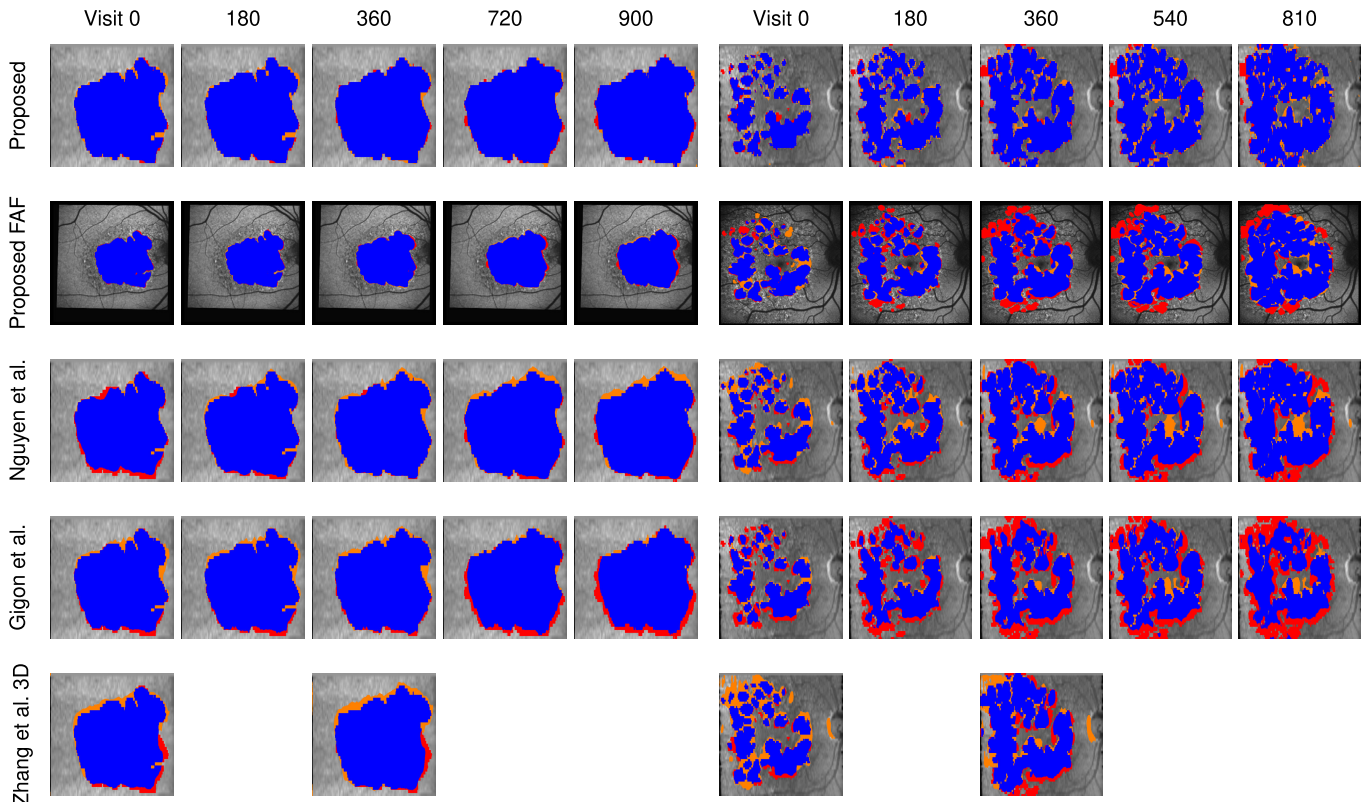


Fig. 5. Visualization of the GA progression prediction from a single baseline scan for two patients. The predictions were performed at baseline (day 0) and available follow-up visits (in days) using baseline OCT. The proposed FAF method employed Fundus Autofluorescence images. True positives are highlighted in blue, false positives in orange and false negatives in red.

have fewer evaluation points as they are designed to produce segmentations only at two fixed visits, namely, baseline and year 1. Overall, the proposed method outperformed the others in Dice score and R^2 , showing slightly better performance for longer forecasting time. In terms of the Dice score, all of the methods decay at a similar rate with increasing forecasting time intervals. In terms of R^2 of the growth rate, the method of Gigon et al. [19] deteriorates exceptionally fast. This can be linked to the design of the method, which relies on function approximation by Taylor series.

The qualitative example of longitudinal prediction with a comparison of multiple methods is shown in Fig. 5. Of note, the method of Zhang et al. can only predict the two time points it was trained to. The method of Gigon et al. when extrapolating the progression beyond 1-year time point starts to shrink the GA segmentation, which is not physiologically plausible. The method of Nguyen et al. produces more false positives and false negatives. The model trained using FAF data performed worse on the cases with complex lesions, likely due to the absence of the rich morphological information present in OCT.

To ensure the correct behaviour of the proposed method, it's crucial to evaluate whether the t_0 embeddings reflect the pathological state. In the absence of reliable tools, we propose to utilize the GA segmentation performance as a fuzzy marker of the quality of the latent representations. In the paper we specifically address the baseline segmentation quality using Dice metric at t_0 as a marker for correct judgment. In our other study, dedicated to the clinical validation of the baseline GA segmentation performance (Mai et al. [57]), we extensively demonstrated the ability of the backbone to correctly estimate

the size and location of GA and to generalize to an external independent dataset. This gives us confidence that t_0 segmentation is robustly estimated across a multitude of patients.

2) *ADNI*: For ADNI data, the models were trained to forecast a 3-year time interval. Thus, the results in Fig. 4 represent intermediate visits inside the forecasting interval. Compared to the GA experiment, the models experience a slower decline in performance in terms of R^2 . In terms of both Dice and R^2 the proposed method showed consistently higher scores compared to other methods. For the Petersen et al. baseline method, we observed an extremely close Dice score of the mean prediction and the best-sampled prediction. In terms of R^2 of the growth, however, the best-sampled prediction was consistently better. The performance of the Nguyen et al. method was slightly lower and followed the same decline pattern. This is attributed to the fact, that RNN can be represented as NeuralODE with the explicit Euler solver.

In the following experiment we provide the evidence of the applicability of the method to different scenarios, namely Hippocampus development under AD influence, thus we didn't perform ablation study or comparison with state of the art methods. In the experiment, we trained the proposed model with non-increasing constrain discussed in the section II-B with Hippocampus labels provided through FastSurfer segmentation on ADNI, and data using the same hyperparameters, as for other ADNI experiments. For comparison, we trained a linear mixed model with random intercept using visit date to predict 3-year Hippocampus size from the baseline, and a linear mixed model with random intercept and slope to predict 1-year Hippocampus size using visit date. We compare the

TABLE VII

DISEASE PROGRESSION MODELS OF HIPPOCAMPUS AFFECTED BY ALZHEIMER'S DISEASE. THE PROPOSED METHOD INCLUDES NON-INCREASING CONSTRAINTS FOR HIPPOCAMPUS VOLUME (SECTION II-B)

Method	Interval	Dice _t	HIPP, %	R ²	r
Proposed	1 year	0.89 ± 0.03	3.4 ± 2.8	0.15	0.48
Proposed	3 year	0.87 ± 0.04	7.5 ± 5.2	0.11	0.42
Linear	3 year	-	11.2 ± 15.3***	0.02*	0.2**
Methods utilizing multiple prior visits					
Gruffaz [23]	Unknown	-	2.8	-	-
Linear	1 year	-	10.1 ± 16.6***	-0.17***	0.38*

performance against the one reported in Gruffaz et al. [23], which includes experiments with predicting Hippocampus volume change on ADNI data. The subsets of ADNI used in these experiments were likely different, however, due to the large size of the datasets, and the fact that the data comes from the same population, we believe that the performance indicated by Gruffaz et al. can still serve for comparison.

For evaluation, alongside Dice coefficient for endpoint segmentation, we use R^2 and Pearson's r of the Hippocampus volume reduction, and HIPP metric employed by Gruffaz, et al. [23] representing mean absolute error of the Hippocampus volume divided by the Hippocampus volume at the baseline visit. In statistical analysis, for HIPP we applied one-sided Mann-Whitney U test, while for R^2 and r we performed a random permutation test. The results of the experiments are presented in table VII. In the baseline segmentation of Hippocampus, the proposed model achieved 0.91 ± 0.02 Dice score. As highlighted by HIPP metric, the longer the forecasting interval, the less accurate the predictions become. In all cases, the proposed method outperformed simple linear models. The performance of the proposed method over 1-year forecasting interval is close to the one of Gruffaz et al.

3) Performance: We analyzed the runtime characteristics and the number of the multiply-accumulate (MAC) operations. The experiment was performed on MUV-GA dataset with a 1-year forecasting interval for training and a 3-year interval for inference. The training batch size was set to 2 with the patch size of $24 \times 128 \times 512$, while the batch size for inference was set to 1 with the volume size $64 \times 128 \times 1024$. As a result, the proposed model had 5.19 million parameters, required 772 GMACs, 0.4 sec and 3.05 GB VRAM for inference; and 4.85 GB VRAM for training. As expected, the proposed method required a large amount of computation, as it requires solving ODE. For inference, the memory requirements were modest and can fit most of the modern GPUs. During the training, however, the proposed method can utilize larger batch sizes to stabilize the training and speed up the convergence as it provides better gradient estimates.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: <http://adni.loni.usc.edu/>

wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd., and its affiliated company Genentech Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Company Inc.; Meso Scale Diagnostics LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education and the study is coordinated by the Alzheimer's Therapeutic Research Institute, University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging, University of Southern California.

REFERENCES

- [1] J. S. Barrett, T. Nicholas, K. Azer, and B. W. Corrigan, "Role of disease progression models in drug development," *Pharmaceutical Res.*, vol. 39, no. 8, pp. 1803–1815, Aug. 2022.
- [2] N. Carrillo et al., "Safety and efficacy of N-acetylmannosamine (ManNAc) in patients with GNE myopathy: An open-label phase 2 study," *Genet. Med.*, vol. 23, no. 11, pp. 2067–2075, Nov. 2021.
- [3] C. Hu, "Exposure-response modeling of clinical end points using latent variable indirect response models," *CPT, Pharmacometrics Syst. Pharmacol.*, vol. 3, no. 6, p. e117, Jun. 2014.
- [4] W. L. Wong et al., "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic review and meta-analysis," *Lancet Global Health*, vol. 2, no. 2, pp. 106–116, Feb. 2014.
- [5] R. H. Guymer et al., "Incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration: Classification of atrophy meeting report 4," *Ophthalmology*, vol. 127, no. 3, pp. 394–409, 2020.
- [6] D. S. Liao et al., "Complement C3 inhibitor pegcetacoplan for geographic atrophy secondary to age-related macular degeneration: A randomized phase 2 trial," *Ophthalmology*, vol. 127, no. 2, pp. 186–195, Feb. 2020.
- [7] Z. Breijyeh and R. Karaman, "Comprehensive review on Alzheimer's disease: Causes and treatment," *Molecules*, vol. 25, no. 24, p. 5789, Dec. 2020.
- [8] B. R. Ott et al., "Brain ventricular volume and cerebrospinal fluid biomarkers of Alzheimer's disease," *J. Alzheimer's Disease*, vol. 20, no. 2, pp. 647–657, Apr. 2010.
- [9] S. M. Nestor et al., "Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database," *Brain*, vol. 131, no. 9, pp. 2443–2454, 2008.
- [10] R. V. Marinescu et al., "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," in *Proc. Int. Workshop PRedictive Intell. MEDicine*, vol. 11843, Oct. 2019, pp. 1–10.

- [11] T. Wang, R. G. Qiu, and M. Yu, "Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks," *Sci. Rep.*, vol. 8, no. 1, p. 9161, Jun. 2018.
- [12] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, and B. T. T. Yeo, "Predicting Alzheimer's disease progression using deep recurrent neural networks," *NeuroImage*, vol. 222, Nov. 2020, Art. no. 117203.
- [13] L. Chan et al., "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease," *Diabetologia*, vol. 64, no. 7, pp. 1504–1515, Jul. 2021.
- [14] Y. Zheng and X. Hu, "Healthcare predictive analytics for disease progression: A longitudinal data fusion approach," *J. Intell. Inf. Syst.*, vol. 55, no. 2, pp. 351–369, Oct. 2020.
- [15] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 43–51.
- [16] I. Dayan et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Med.*, vol. 27, no. 10, pp. 1735–1743, 2021.
- [17] Y. Zhang, Z. Ji, S. Niu, T. Leng, D. L. Rubin, and Q. Chen, "A multi-scale deep convolutional neural network for joint segmentation and prediction of geographic atrophy in SD-OCT images," in *Proc. 16th Int. Symp. Biomed. Imag.*, 2019, pp. 565–568.
- [18] Y. Zhang et al., "An integrated time adaptive geographic atrophy prediction model for SD-OCT images," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101893.
- [19] A. Gigon et al., "Personalized atrophy risk mapping in age-related macular degeneration," *Translational Vis. Sci. Technol.*, vol. 10, no. 13, p. 18, Nov. 2021.
- [20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [21] M. Chen, "MinimalRNN: Toward more interpretable and trainable recurrent neural networks," 2017, *arXiv:1711.06788*.
- [22] R. V. Marinescu et al., "The Alzheimer's disease prediction of longitudinal evolution (TADPOLE) challenge: Results after 1 year follow-up," *Mach. Learn. Biomed. Imag.*, vol. 1, pp. 1–60, Dec. 2021. [Online]. Available: <https://melba-journal.org/2021:019>, doi: 10.59275/j.melba.2021-2dccc.
- [23] S. Gruffaz, P.-E. Poulet, E. Maheux, B. M. Jedynak, and S. Durrleman, "Learning Riemannian metric for disease progression modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 23780–23792.
- [24] I. Ezhov et al., "Neural parameters estimation for brain tumor growth modeling," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*, D. Shen et al., Eds. Cham, Switzerland: Springer, 2019, pp. 787–795.
- [25] J. Petersen et al., "Deep probabilistic modeling of glioma growth," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*, D. Shen et al., Eds. Cham, Switzerland: Springer, 2019, pp. 806–814.
- [26] J. Petersen et al., "Continuous-time deep glioma growth models," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, M. de Bruijne et al., Eds. Cham, Switzerland: Springer, 2021, pp. 83–92.
- [27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.
- [28] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [29] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *Proc. NIPS*, 2018, pp. 1–13.
- [30] W. Hao et al., "Learning amyloid pathology progression from longitudinal PIB-PET images in preclinical Alzheimer's disease," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 572–576.
- [31] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
- [32] E. De Brouwer et al., "Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106180.
- [33] E. D. Brouwer, J. Simm, A. Arany, and Y. Moreau, "GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series," in *Proc. NIPS*, 2019, pp. 1–12.
- [34] L. S. Shigueoka, E. B. Mariottoni, A. C. Thompson, A. A. Jammal, V. P. Costa, and F. A. Medeiros, "Predicting age from optical coherence tomography scans with deep learning," *Transl. Vis. Sci. Technol.*, vol. 10, no. 1, p. 12, Jan. 2021.
- [35] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016, pp. 424–432.
- [36] D. Lachinov, P. Seeböck, J. Mai, F. Goldbach, U. Schmidt-Erfurth, and H. Bogunovic, "Projective skip-connections for segmentation along a subset of dimensions in retinal OCT," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, M. de Bruijne et al., Eds. Cham, Switzerland: Springer, 2021, pp. 431–441.
- [37] F. G. Holz, E. C. Strauss, S. Schmitz-Valckenberg, and M. V. L. Campagne, "Geographic atrophy: Clinical features and potential therapeutic approaches," *Ophthalmology*, vol. 121, no. 5, pp. 1079–1091, 2014.
- [38] B. Ribba et al., "A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis," *CPT, Pharmacometrics Syst. Pharmacol.*, vol. 3, no. 5, pp. 1–10, May 2014.
- [39] A. K. Laird, "Dynamics of tumor growth," *Brit. J. Cancer*, vol. 18, no. 3, pp. 490–502, Sep. 1964.
- [40] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.
- [41] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
- [42] L. Ferrarini, W. M. Palm, H. Olofsen, M. A. van Buchem, J. H. C. Reiber, and F. Admiraal-Behloul, "Shape differences of the brain ventricles in Alzheimer's disease," *NeuroImage*, vol. 32, no. 3, pp. 1060–1069, Sep. 2006.
- [43] F. L. Giesel et al., "Temporal horn index and volume of medial temporal lobe atrophy using a new semiautomated method for rapid and precise assessment," *Amer. J. Neuroradiol.*, vol. 27, no. 7, pp. 1454–1458, 2006.
- [44] P. T. A. Bui et al., "Fundus autofluorescence and optical coherence tomography biomarkers associated with the progression of geographic atrophy secondary to age-related macular degeneration," *Eye*, vol. 36, no. 10, pp. 2013–2019, 2021.
- [45] M. Arikian, A. Sadeghipour, B. Gerendas, R. Told, and U. Schmidt-Erfurt, "Deep learning based multi-modal registration for retinal imaging," in *Proc. IMIMIC*, 2019, pp. 75–82.
- [46] M. V. Pasricha et al., "Local anatomic precursors to new-onset geographic atrophy in age-related macular degeneration as defined on OCT," *Ophthalmol. Retina*, vol. 5, no. 5, pp. 396–408, May 2021.
- [47] J. Mai et al., "Comparison of fundus autofluorescence versus optical coherence tomography-based evaluation of the therapeutic response to pegcetacoplan in geographic atrophy," *Amer. J. Ophthalmol.*, vol. 244, pp. 175–182, Dec. 2022.
- [48] E. De Castro and C. Morandi, "Registration of translated and rotated images using finite Fourier transforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 700–703, Sep. 1987.
- [49] M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1436–1447, Sep. 2009.
- [50] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "FastSurfer—A fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, Oct. 2020, Art. no. 117012.
- [51] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, 1998.
- [52] F. Ségonne et al., "A hybrid approach to the skull stripping problem in MRI," *NeuroImage*, vol. 22, no. 3, pp. 1060–1075, Jul. 2004.
- [53] M. Reuter, H. D. Rosas, and B. Fischl, "Highly accurate inverse consistent registration: A robust approach," *NeuroImage*, vol. 53, no. 4, pp. 1181–1196, Dec. 2010.
- [54] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, Jul. 2012.
- [55] S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama, "Dissecting neural ODEs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3952–3963.
- [56] M. Habiba and B. A. Pearlmuter, "Neural ordinary differential equation based recurrent neural network model," in *Proc. 31st Irish Signals Syst. Conf. (ISSC)*, Jun. 2020, pp. 1–6.
- [57] J. Mai et al., "Clinical validation for automated geographic atrophy monitoring on OCT under complement inhibitory treatment," *Sci. Rep.*, vol. 13, no. 1, p. 7028, Apr. 2023.