A Multitask Learning Framework for Predicting Diabetic Retinopathy Progression via State Transition Matrices

Rachid Zeghlache[1,2,3], Pierre-Henri Conze[1,2], Tinashe Ernest Mutsvangwa[2], Pascale Massin[4], Béatrice Cochener[1,5], Ikram Brahim[6], and Gwenolé Quellec[1]

[1] LaTIM UMR 1101, Inserm, Brest, France
[2] IMT Atlantique, Brest, France
[3] Heriot-Watt University Dubai, Dubai, United Arab Emirates
[4] Service d'Ophtalmologie, Hôpital Lariboisière, AP-HP, Paris, France
[5] Service d'Ophtalmologie, CHU de Brest, Brest, France
[6] Inserm UMR 1227, LBAI, Univ Brest, CHU Brest, Brest, France
rz4025@hw.ac.uk

Abstract. Diabetic retinopathy (DR) is a chronic disease characterized by progressive vision loss, with its progression posing an increasing risk of impairment among individuals with diabetes. Predicting DR progression over extended periods is essential for guiding patient management and therapeutic strategies. We propose a matrix-based multi-task learning (MTL) framework that redefines progression prediction through three contributions. First, we present an MTL framework that simultaneously predicts disease progression across multiple time points. Second, instead of predicting isolated future states, we predict a matrix of state transition probabilities between successive time points, offering a structured and richer representation of disease evolution. Third, we introduce a progression consistency loss to align predicted transitions with empirical progression dynamics observed in training data. We evaluate the proposed method for predicting DR across one to five years on a large-scale longitudinal dataset. Our results highlight the potential of the matrix-based approach, particularly for long-term DR outcome prediction and early intervention planning.

Keywords: Diabetic Retinopathy · Disease Progression · Multi-task Learning · State Transition Matrix · Long-term Prediction.

1 Introduction

Traditional models represent disease progression as a probabilistic vector at a given time but struggle to capture the structured nature of stage transitions. We address this limitation through a matrix-based multi-task learning (MTL) framework with two key components. First, we introduce an MTL setup that predicts disease progression across multiple time points simultaneously. Conventional methods either train separate models per time horizon [1,13] or treat

each time point as an independent task. While MTL improves generalization and efficiency, training remains challenging due to task balancing and potential negative transfer [16,4,18]. Fragmented approaches often result in inconsistencies in modeling disease dynamics.

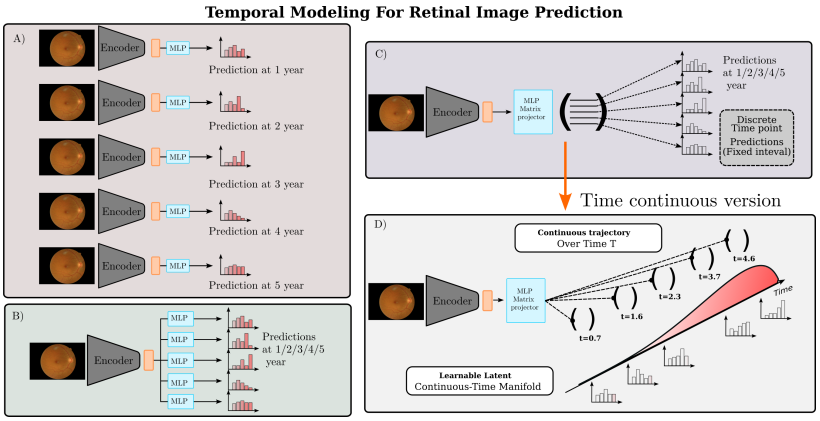

Fig. 1. Illustration of the proposed approach: instead of predicting an isolated future state vector (left), our method predicts a full state transition matrix (right), capturing structured progression dynamics across DR stages over time.

Our second component extends MTL by forecasting a matrix of state transition probabilities, offering a more granular view of disease evolution and detailed patient trajectories. Fig. 1 contrasts conventional vector-based prediction with our matrix-based approach. The formulation relates to Markov state modeling, effective in speech recognition and NLP [8] and in financial forecasting [6] for capturing temporal dependencies [14]. An expectation-maximization approach for continuous-time hidden Markov models was introduced in [11] to model disease progression, though optimization remains complex. Our formulation mitigates this by sharing a single transition operator across horizons and leveraging MTL's shared representations.

Arcadu et al. [1] predicted individual DR progression from fundus images and Nderitu et al. [13] 1–3 year emergent referable DR; Zeghlache et al. [19] proposed longitudinal self-supervised learning for DR detection, and Liu et al. [9] modeled image-level trajectories from irregularly-sampled data. Unlike these approaches, which predict isolated future states or severity scores, our method predicts a full transition matrix, jointly modeling all stage transitions and enforcing consistency with observed progression patterns.

Our contributions are: (1) a transition matrix prediction framework that extends MTL by forecasting state transition probabilities rather than isolated future states; (2) a progression consistency loss computed exclusively on training data to regularize predictions; and (3) a comprehensive evaluation on a large-

scale DR dataset demonstrating improved long-term prediction, particularly for severe DR stages [2].

2 Proposed Method

Our method models DR progression by encoding input images into latent representations, capturing temporal dynamics, and predicting transition probabilities between DR stages. We employ two training setups: *fixed-time* and *time-dependent*, as illustrated in Fig. 2. A progression consistency loss aligns predicted transitions with empirical dynamics.

Let $\mathbf{x}_t \in \mathbb{R}^d$ denote an input fundus image at time t with dimensionality d, and $\mathbf{s}_t \in \{0, 1, 2, 3, 4\}$ represent the DR severity on the ICDR scale (0: no DR, 1: mild, 2: moderate, 3: severe, 4: proliferative DR). Our goal is to model the transition $\mathbf{s}_t \to \mathbf{s}_{t+\Delta}$ for arbitrary time horizons Δ.

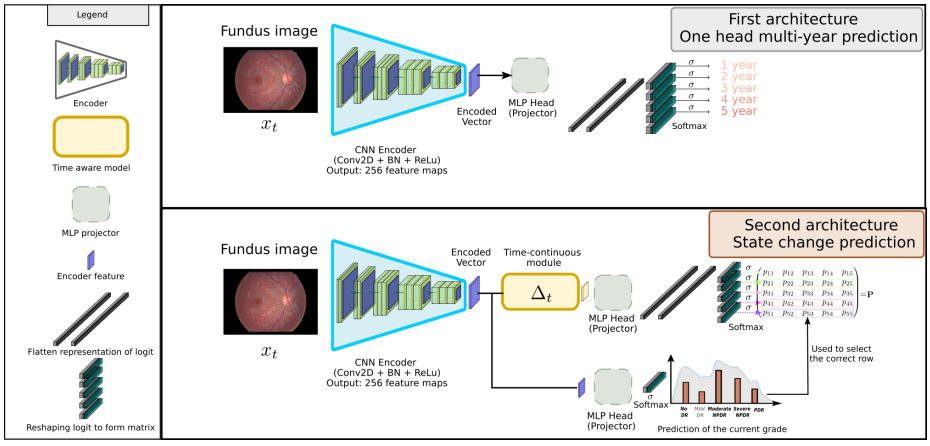

Fig. 2. Overview of the proposed framework. An encoder f maps the input image $\mathbf{x}_t$ to a latent representation $\mathbf{h}_t$. In the time-dependent setup, a Neural ODE-RNN module Φ_θ incorporates the time interval Δ_t to produce a time-aware representation $\mathbf{z}_t$. An MLP then outputs a 5×5 logit matrix $\mathbf{L}_t$, which is converted via row-wise softmax into the transition probability matrix $\mathbf{P}_t$ (corresponding to $\mathbf{G}_t$ in the fixed-time setup). Each row i of $\mathbf{P}_t$ provides the probability distribution over future DR stages given current stage i.

The input $\mathbf{x}_t$ is encoded into a latent representation $\mathbf{h}_t \in \mathbb{R}^m$ through an encoder f (ResNet50 [5]):

$$\mathbf{h}_t = f(\mathbf{x}_t) \tag{1}$$

Temporal dynamics modeling.

In the time-dependent setup, we use a Neural ODE combined with an RNN (ODE-RNN) [3], denoted Φ_θ, to produce a time-aware representation. Neural

ODEs model continuous-time dynamics by parameterizing the derivative of the hidden state with a neural network, enabling the model to handle *irregularly-sampled* longitudinal observations—a common scenario in clinical practice where patient visit intervals vary. The RNN component maintains a summary of past observations, while the ODE solver evolves the hidden state between visits. This yields:

$$\mathbf{z}_t = \Phi_\theta(\mathbf{h}_t, \Delta_t), \quad \Phi_\theta : \mathbb{R}^m \times \mathbb{R}_{\geq 0} \to \mathbb{R}^m \tag{2}$$

where $\Delta_t \in \mathbb{R}_{\geq 0}$ denotes the time interval between observations. In the fixed-time setup, Δ_t is implicit (fixed at one-year intervals), and $\mathbf{z}_t = \mathbf{h}_t$.

An MLP produces a 5×5 logit matrix $\mathbf{L}_t$:

$$\mathbf{L}_t = \mathrm{MLP}(\mathbf{z}_t), \quad \mathrm{MLP} : \mathbb{R}^m \to \mathbb{R}^{5\times 5} \tag{3}$$

Row-wise softmax yields the transition probability matrix $\mathbf{P}_t$, where each entry represents the probability of transitioning from stage i to stage j:

$$\mathbf{P}_t(i,j) = \frac{\exp(\mathbf{L}_t(i,j))}{\sum_{k=0}^{4} \exp(\mathbf{L}_t(i,k))} \tag{4}$$

Prediction. Given the patient's current DR stage $\mathbf{s}_t$, the predicted probability distribution over future stages is obtained by selecting row $\mathbf{s}_t$ from $\mathbf{P}_t$. The predicted future stage is $\hat{\mathbf{s}}_{t+\Delta} = \arg\max_j \mathbf{P}_t(\mathbf{s}_t, j)$.

Why predict a full matrix? Only row $\mathbf{s}_t$ is directly supervised by the classification loss; the other rows are shaped by the consistency term (Sec. 2.2) and by weight sharing across the training population. They are not patient-specific predictions but a *counterfactual* progression operator: row i answers "given this retina's appearance, how would a patient at grade i evolve?". Three properties follow that a plain five-class vector lacks: supervision from patients at every grade is pooled into one structured output, which regularises the low-data regime of long follow-ups; the row-stochastic form lets the empirical matrix $\mathbf{Q}$ inject clinical prior knowledge; and the operator composes, so $\mathbf{G}_t^h$ yields multi-year forecasts at no extra parameter cost.

2.1 Training Setups

We denote the transition matrix as $\mathbf{G}_t$ in the fixed-time setup and $\mathbf{P}_t$ in the time-dependent setup.

Fixed-Time Setup. This setup predicts DR progression at discrete one-year intervals. Given $\mathbf{x}_t$, the model outputs a *single* one-year transition matrix $\mathbf{G}_t \in \mathbb{R}^{5\times 5}$, where each row i represents the predicted transition probabilities from stage i. Note that $\mathbf{G}_t$ corresponds to $\mathbf{P}_t$ in Fig. 2 but without the ODE-RNN module. A single head is shared across horizons rather than one head or one model per horizon: the prediction at horizon $h \in \{1, \ldots, 5\}$ years is obtained by applying the

same matrix h times, $\mathbf{G}_t^h$, as follows from the Markov assumption. All horizons are thus trained jointly, with losses summed:

$$\mathcal{L}_{\text{fixed}} = -\sum_{h=1}^{5}\sum_{j=0}^{4}\mathbf{y}_{t+h}(j)\log\left[\mathbf{G}_t^h\right](\mathbf{s}_t, j) \tag{5}$$

where $\mathbf{y}_{t+h}$ one-hot encodes $\mathbf{s}_{t+h}$ and horizons without an available label are masked. This sharing is what distinguishes *Matrix-head* from the *One-head* and *Multi-task* baselines, which allocate independent parameters per horizon.

Time-Dependent Setup. This setup incorporates variable time intervals via the ODE-RNN module. Following [15], additional clinical features (e.g., current DR grade) are concatenated with $\mathbf{z}_t$ before the MLP. **Setup 1** predicts $\mathbf{P}_t$ accounting for Δ_t:

$$\mathcal{L}_1^{\text{time}} = -\sum_{j=0}^{4}\mathbf{y}_{t+1}(j)\log\mathbf{P}_t(\mathbf{s}_t, j) \tag{6}$$

Setup 2 jointly predicts both the current state (via an auxiliary classifier $\hat{\mathbf{y}}_t$) and the future state. The current-state prediction $\hat{\mathbf{s}}_t = \arg\max_i \hat{\mathbf{y}}_t(i)$ is used to index the transition matrix:

$$\mathcal{L}_2^{\text{time}} = \underbrace{-\sum_{i=0}^{4}\mathbf{y}_t(i)\log\hat{\mathbf{y}}_t(i)}_{\mathcal{L}_{\text{current}}} + \underbrace{-\sum_{j=0}^{4}\mathbf{y}_{t+1}(j)\log\mathbf{P}_t(\hat{\mathbf{s}}_t, j)}_{\mathcal{L}_{\text{future}}} \tag{7}$$

2.2 Progression Consistency Loss

To align $\mathbf{P}_t$ with empirical DR progression, we introduce a consistency loss. The empirical transition matrix $\mathbf{Q}$ is computed *exclusively from training data* to prevent information leakage. We define both a global version computed over the full training set and a per-batch version:

$$\mathbf{Q}(i,j) = \frac{N_{\text{train}}(i,j)}{N_{\text{train}}(i)}, \quad \mathbf{Q}_{\text{batch}}(i,j) = \frac{N_{\text{batch}}(i,j)}{N_{\text{batch}}(i)} \tag{8}$$

where $N_{\text{train}}(i,j)$ and $N_{\text{batch}}(i,j)$ count transitions from stage i to j in the training set and mini-batch, respectively. The consistency loss is:

$$\mathcal{L}_{\text{consistency}} = \|\mathbf{P}_t - \mathbf{Q}\|_2^2 \tag{9}$$

This regularization prevents unrealistic transitions and stabilizes training. While regularizing toward training-set statistics may reduce flexibility for out-of-distribution patterns, DR progression follows well-established clinical trajectories, making this a reasonable inductive bias and placing the loss in line with the constrained-learning literature [7,10].

Implementation details. The consistency term is added with *unit weight*; no weighting factor was tuned, so Eq. 10 is a plain sum. If grade i is absent from a mini-batch then $N_{\text{batch}}(i) = 0$ and row i of $\mathbf{Q}_{\text{batch}}$ is undefined; such rows are masked out for that batch rather than imputed. Finally, $\mathbf{Q}$ pools all consecutive-visit transitions *without* conditioning on Δ_t, so the network sees the interval but the regulariser does not. As short- and long-interval transition rates differ, this interval-marginal prior is an approximation (see Limitations). The final losses combine the main task loss with consistency regularization:

$$\mathcal{L}_{\text{fixed}}^{\text{combined}} = \mathcal{L}_{\text{fixed}} + \mathcal{L}_{\text{consistency}}, \quad \mathcal{L}_{\text{time}}^{\text{combined}} = \mathcal{L}_{2}^{\text{time}} + \mathcal{L}_{\text{consistency}} \tag{10}$$

Relationship between the two setups. The setups are trained and evaluated separately and serve different ends. The fixed-time setup discretises follow-up onto an annual grid, matching the protocol of the DR-progression baselines we compare against [1,13], and isolates the contribution of the matrix head from that of the temporal module. The time-dependent setup is the general model and the one intended for deployment, since screening visits are irregular and a yearly grid discards visits that do not align with it. It subsumes the fixed-time model in principle by querying $\Delta_t \in \{1, \ldots, 5\}$ years, but the two are trained on different cohorts (annual-grid subsets vs. all consecutive visit pairs) and are therefore not interchangeable in practice.

3 Experiments and Results

Dataset and Setup. The OPHDIAT dataset [12], a colour fundus photography database collected by the OPHDIAT telemedical DR screening network in the Île-de-France region, was used. It comprises 763,848 images from 101,383 patients (2004–2017), with 673,000 assigned DR severity grades on the ICDR scale [17]. Patients aged 9 to 91 had multiple examinations with at least two images per eye; one image per eye was randomly selected per examination. Data was split 60/20/20 into training, validation, and test *at the patient level*: all examinations and both eyes of a patient fall in the same fold, so no patient, eye, or repeated visit crosses folds, ruling out the leakage an image-wise split would cause. The follow-up sub-cohorts of Tab. 1 are nested subsets of these folds. Models were trained for 400 epochs with AdamW optimizer, OneCycleLR scheduler, learning rate 1×10^{-3}, weight decay 1×10^{-4}, and batch size 128 on a single Nvidia A6000 GPU. Neural ODEs were implemented using Torchdiffeq [3] with dense layers, tanh activation, and the dopri5 solver. The encoder is ResNet50 [5] with a two-layer MLP projector.

Evaluation Metrics. We report AUC at three clinically relevant DR severity thresholds: at least Mild (AUC1: stages $\geq$1), at least Moderate (AUC2: stages $\geq$2), and at least Severe (AUC3: stages $\geq$3). These thresholds correspond to key clinical decision points for referral and treatment. Cohen's κ measures agreement between predicted and observed 5-class DR stages.

Fixed-Time Setup. We evaluate three approaches: (1) *One-head*: a separate single-head classifier for each year; (2) *Multi-task*: a multi-head classifier predicting all years concurrently; and (3) *Matrix-head*: our proposed method outputting a 5×5 transition matrix. We stratify patients by follow-up length (2–5 years) to assess how data availability affects performance.

Time-Dependent Setup. We evaluate our time-dependent matrix-based model with varying consistency constraints (none, per-batch, global) and with/without using the current DR grade as input (Tab. 2). "Next visit" refers to the chronologically subsequent examination in the patient record, with the time interval Δ_t provided as input to the ODE-RNN.

Method	Prediction 1 year				Prediction 2 year				Prediction 3 year				Prediction 4 year				Prediction 5 year			
	AUC1	AUC2	AUC3	κ	AUC1	AUC2	AUC3	κ	AUC1	AUC2	AUC3	κ	AUC1	AUC2	AUC3	κ	AUC1	AUC2	AUC3	κ
Sub-Dataset 1: 5-year follow-up																				
One-head	0.626	0.659	0.655	0.012	0.688	0.686	0.713	0.059	0.663	0.687	0.698	0.093	**0.702**	0.715	0.736	0.078	0.679	0.672	0.705	0.062
Multi-task	0.686	0.636	0.655	0	0.685	0.641	0.661	0	0.680	0.648	0.655	0	0.675	0.650	0.671	0	0.682	0.645	0.667	0
Matrix-head (ours)	**0.716**	**0.759**	**0.767**	**0.182**	**0.705**	**0.731**	**0.776**	**0.152**	**0.693**	**0.723**	**0.764**	**0.143**	0.684	**0.718**	**0.769**	**0.126**	**0.697**	**0.710**	**0.767**	**0.110**
Sub-Dataset 2: 4-year follow-up																				
One-head	0.722	0.770	0.718	**0.249**	0.707	0.750	**0.747**	0.194	0.707	0.730	0.699	0.191	0.704	0.723	0.706	0.143	-	-	-	-
Multi-task	**0.725**	**0.780**	**0.742**	0.230	**0.711**	**0.762**	0.736	0.195	**0.710**	**0.750**	**0.742**	0.177	**0.705**	**0.736**	**0.746**	0.141	-	-	-	-
Matrix-head (ours)	0.724	0.770	0.716	0.247	0.705	0.747	0.718	**0.210**	0.705	0.742	0.712	**0.213**	0.699	0.725	0.710	**0.199**	-	-	-	-
Sub-Dataset 3: 3-year follow-up																				
One-head	**0.837**	0.833	0.799	**0.569**	**0.795**	0.821	0.785	**0.471**	**0.777**	0.809	0.781	**0.429**	-	-	-	-	-	-	-	-
Multi-task	0.826	**0.834**	**0.802**	0.540	0.791	**0.824**	**0.804**	0.447	0.771	**0.812**	**0.809**	0.373	-	-	-	-	-	-	-	-
Matrix-head (ours)	0.835	0.833	0.788	0.551	**0.795**	0.820	0.790	0.469	**0.777**	0.809	0.790	0.397	-	-	-	-	-	-	-	-
Sub-Dataset 4: 2-year follow-up																				
One-head	0.834	0.847	0.856	0.572	0.812	0.837	0.846	0.529	-	-	-	-	-	-	-	-	-	-	-	-
Multi-task	0.837	0.849	0.866	**0.589**	0.810	0.839	0.857	0.529	-	-	-	-	-	-	-	-	-	-	-	-
Matrix-head (ours)	**0.840**	**0.855**	**0.868**	**0.589**	**0.818**	**0.847**	**0.866**	**0.538**	-	-	-	-	-	-	-	-	-	-	-	-

Table 1. Fixed-time prediction results. Comparison of One-head, Multi-task, and Matrix-head (ours) for predicting DR over 1–5 years. Results are stratified by patient follow-up length. Best results per sub-dataset and year are in **bold**.

Model	Constraint	Current grade	DR severity at next visit			
			AUC1	AUC2	AUC3	κ
NODE classifier	-	no	0.564	0.617	0.657	0.083
Matrix-head (ours)	-	no	0.562	0.618	0.617	0.057
	per-batch	no	0.598	0.683	0.773	0.197
	global	no	0.582	0.659	0.681	0.143
	-	yes	0.603	0.675	0.745	0.179
	per-batch	yes	**0.653**	**0.741**	**0.863**	**0.320**
	global	yes	0.613	0.699	0.758	0.196

Table 2. Time-dependent prediction results. The NODE baseline uses a standard classifier head; the Matrix-head is evaluated under different consistency constraints, with and without the current DR grade as auxiliary input.

3.1 Results Analysis

Fixed-Time Predictions (Tab. 1). The benefit of the Matrix-head is cohort-dependent rather than uniform. On the 5-year cohort it leads on 19 of 20 entries, with the best AUC3 (0.767 at 1 year vs. 0.655 for One-head) and markedly better κ (0.182 vs. 0.012); the exception is AUC1 at 4 years (0.684 vs. 0.702). Multi-task collapses to a single class here ($\kappa = 0$ throughout). On the 2-year cohort the Matrix-head is best on every entry (AUC3 = 0.868, $\kappa = 0.589$ at 1 year). On the 3- and 4-year cohorts it does *not* consistently outperform the baselines: Multi-task takes most AUCs on the 4-year cohort and One-head the best AUC1 and κ on the 3-year cohort, the Matrix-head's edge being confined to κ at the longer 4-year-cohort horizons (0.199 at 4 years vs. 0.143 and 0.141). All models show declining κ with horizon, consistent with prior work [1,13]. Sample size plausibly drives part of the cross-cohort spread, since longer follow-ups form smaller subsets, but the sub-cohorts also differ in grade distribution and transition counts, so we do not attribute it to a single factor.

Time-Dependent Predictions (Tab. 2). The per-batch consistency constraint yields the largest improvements, particularly combined with the current grade as input (AUC3: 0.863, κ: 0.320); the latter is expected, as the grade provides a strong prior for the transition matrix. The per-batch constraint outperforms the global one, likely because batch-level statistics better capture local progression patterns and add useful stochastic regularization during training.

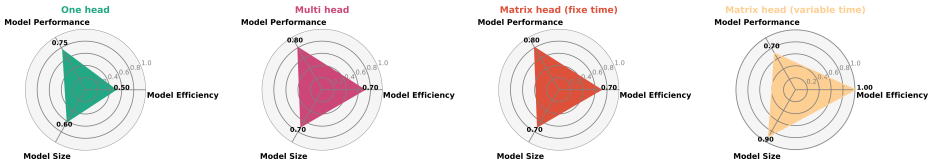

Fig. 3. Comparison of approaches in terms of AUC performance, number of parameters, and task efficiency (number of prediction horizons handled by a single model).

Efficiency Analysis (Fig. 3). The efficiency of the Matrix-head lies in horizon coverage rather than in the size of its output layer, which is modest but not negligible: a linear map from $m = 2048$ features to a 5×5 matrix holds $25m+25$ weights, and our two-layer MLP more. Because one matrix is composed as $\mathbf{G}_t^h$ (Eq. 5), a single encoder and head span all five horizons, whereas One-head trains five independent networks (five ResNet50 encoders, ≈ 23.5M parameters each) and Multi-task shares the encoder but adds a head per horizon. Against One-head this approaches a fivefold reduction in total parameters, driven by encoder sharing.

4 Discussion and Conclusion

We presented a matrix-based MTL framework predicting DR progression via state transition probability matrices rather than isolated severity labels: a 5×5 matrix jointly models all stage transitions from one fundus image, and the progression consistency loss anchors predictions to empirical dynamics. On OPHDIAT across 1–5 year horizons, it is strongest on the long-horizon, data-limited 5-year cohort and the large 2-year cohort, particularly for severe DR detection (AUC3 = 0.767 vs. 0.655 at 1 year), while matching—not beating—the baselines on the 3- and 4-year cohorts. It covers every horizon with one shared encoder and one composable matrix.

Limitations. The framework requires labeled longitudinal data with DR grades at successive visits; semi-supervised approaches [20] could exploit unlabeled follow-ups. The empirical matrix $\mathbf{Q}$ encodes population-specific rates and may need recalibration across demographics or treatment practices; estimated without conditioning on the follow-up interval, it mixes short- and long-interval transitions into a single prior, so stratifying $\mathbf{Q}$ by Δ_t is a natural next step. Cohort size and grade distribution are confounded across the follow-up sub-cohorts, leaving the source of their performance differences open. Patient-level sequences and covariates (e.g., HbA1c, diabetes duration) could personalize estimates. Finally, the declining κ from 0.182 at 1 year to 0.110 at 5 years reflects inherent forecasting uncertainty and motivates calibration-aware objectives.

Acknowledgments. The work takes place in the framework of Evired, an ANR RHU project. This work benefits from State aid managed by the French National Research Agency under the "Investissement d'Avenir" program bearing the reference ANR-18-RHUS-0008.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., Prunotto, M.: Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Digital Medicine **2**(1), 92 (2019). https://doi.org/10.1038/s41746-019-0172-3
2. Bora, A., Balasubramanian, S., Babenko, B., Virmani, S., Venugopalan, S., Mitani, A., de Oliveira Marinho, G., Cuadros, J., Ruamviboonsuk, P., Corrado, G.S., Peng, L., Webster, D.R., Varadarajan, A.V., Hammel, N., Liu, Y., Bavishi, P.: Predicting the risk of developing diabetic retinopathy using deep learning. The Lancet Digital Health **3**, e10–e19 (1 2021). https://doi.org/10.1016/S2589-7500(20)30250-8
3. Chen, R.T.Q.: torchdiffeq (2018), https://github.com/rtqichen/torchdiffeq
4. Crawshaw, M.: Multi-task learning with deep neural networks: A survey (2020), https://arxiv.org/abs/2009.09796

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
6. Hu, Z., Zhao, Y., Khushi, M.: A survey of forex and stock price prediction using deep learning. Applied System Innovation **4**(1), 9 (Feb 2021). https://doi.org/10.3390/asi4010009, http://dx.doi.org/10.3390/asi4010009
7. Kałuża, A., Morkisz, P.M., Mulewicz, B., Przybyłowicz, P., Wiącek, M.: Deep learning-based estimation of time-dependent parameters in markov models with application to nonlinear regression and sdes (2023). https://doi.org/10.48550/ARXIV.2312.08493, https://arxiv.org/abs/2312.08493
8. Kuo, C.Y., Chien, J.T.: Markov recurrent neural networks. In: 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE (Sep 2018). https://doi.org/10.1109/mlsp.2018.8517074, http://dx.doi.org/10.1109/MLSP.2018.8517074
9. Liu, C., Xu, K., Shen, L.L., Huber, G., et al.: ImageFlowNet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images. In: ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE (2025)
10. Liu, W., Lai, Z., Bacsa, K., Chatzi, E.: Physics-guided deep markov models for learning nonlinear dynamical systems with uncertainty. Mechanical Systems and Signal Processing **178**, 109276 (Oct 2022). https://doi.org/10.1016/j.ymssp.2022.109276, http://dx.doi.org/10.1016/j.ymssp.2022.109276
11. Liu, Y.Y., Li, S., Li, F., Song, L., Rehg, J.M.: Efficient learning of continuous-time hidden markov models for disease progression. Advances in neural information processing systems **28**, 3599–3607 (2015), https://api.semanticscholar.org/CorpusID:8316852
12. Massin, P., Chabouis, A., Erginay, A., Viens-Bitker, C., Lecleire-Collet, A., Meas, T., Guillausseau, P.J., Choupot, G., André, B., Denormandie, P.: OPHDIAT: a telemedical network screening system for diabetic retinopathy in the Île-de-France. Diabetes & Metabolism **34**(3), 227–234 (2008). https://doi.org/10.1016/j.diabet.2007.12.006
13. Nderitu, P., Bergeles, C., Sivaprasad, S.: Predicting 1, 2 and 3 year emergent referable diabetic retinopathy and maculopathy using deep learning. Communications Medicine **4**(1), 167 (2024). https://doi.org/10.1038/s43856-024-00590-x
14. Petropoulos, A., Chatzis, S.P., Xanthopoulos, S.: A hidden markov model with dependence jumps for predictive modeling of multidimensional time-series. Information Sciences **412–413**, 50–66 (Oct 2017). https://doi.org/10.1016/j.ins.2017.05.038, http://dx.doi.org/10.1016/j.ins.2017.05.038
15. Rabhi, S., Blanchard, F., Diallo, A.M., Zeghlache, D., Lukas, C., Berot, A., Delemer, B., Barraud, S.: Temporal deep learning framework for retinopathy prediction in patients with type 1 diabetes. Artificial Intelligence in Medicine **133**, 102408 (Nov 2022). https://doi.org/10.1016/j.artmed.2022.102408, http://dx.doi.org/10.1016/j.artmed.2022.102408
16. Vandenhende, S., Georgoulis, S., Proesmans, M., Dai, D., Gool, L.V.: Revisiting multi-task learning in the deep learning era. ArXiv **abs/2004.13379** (2020), https://api.semanticscholar.org/CorpusID:216562469
17. Wilkinson, C., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003). https://doi.org/https://doi.org/10.1016/S0161-6420(03)00475-5, https://www.sciencedirect.com/science/article/pii/S0161642003004755

18. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. ArXiv **abs/2001.06782** (2020), https://api.semanticscholar.org/CorpusID:210839011
19. Zeghlache, R., Conze, P.H., Daho, M.E.H., Tadayoni, R., Massin, P., Cochener, B., Quellec, G., Lamard, M.: Detection of diabetic retinopathy using longitudinal self-supervised learning. In: International Workshop on Ophthalmic Medical Image Analysis. pp. 43–52. Springer (2022)
20. Zeghlache, R., Conze, P.H., El Habib Daho, M., Li, Y., Le Boité, H., Tadayoni, R., Massin, P., Cochener, B., Rezaei, A., Brahim, I., et al.: Latim: Longitudinal representation learning in continuous-time models to predict disease progression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 404–414. Springer Nature Switzerland Cham (2024)