Response-Aware Multimodal Learning for Post-Treatment Visual Acuity Forecasting

Phuoc-Nguyen Bui[1], Van-Vi Vo[1], Duc-Tai Le[2], Junghyun Bum[3], Van-Nguyen Pham[1], Kiyoung Kim[4], Seung-Young Yu[4], and Hyunseung Choo[5*]

[1] Convergence Research Institute, Sungkyunkwan University, Korea
[2] Dept. of AI Systems Engineering, Sungkyunkwan University, Korea
[3] College of Computing and Informatics, Sungkyunkwan University, Korea
[4] Dept. of Ophthalmology, Kyung Hee University Hospital, Kyung Hee University, Korea
[5] Dept. of Electrical and Computer Engineering, Sungkyunkwan University, Korea
{phuocnguyen,vovanvi,ldtai,bumjh,nguyenpv195,choo}@skku.edu, {pourma,syyu}@khu.ac.kr

Abstract. Long-term visual acuity (VA) forecasting after anti-VEGF therapy is important for counseling and follow-up planning in diabetic macular edema (DME), yet remains challenging when only early post-treatment findings are available. While prior OCT-based methods mainly focus on short-term response or single-endpoint prediction, multi-horizon VA forecasting from early longitudinal data remains insufficiently under-explored. In this study, we assembled a real-world cohort of 188 anti-VEGF–treated DME patients with paired baseline and month-1 OCT scans, along with tabular OCT-derived biomarkers and non-imaging clinical variables. Using only these early data, we formulate a multi-horizon VA forecasting problem aimed at predicting visual outcomes at 3, 6, 12, 18, and 24 months, reflecting clinically meaningful follow-up intervals. We propose **ReVA**, a response-aware multimodal framework that combines baseline and month-1 OCT features with tabular variables to capture disease status and early treatment response. ReVA integrates spatial OCT attention, dependency-aware tabular encoding, and cross-modal fusion to predict patient-specific long-term VA trajectories. The proposed framework achieves MAE = 0.1246, RMSE = 0.1621, and R^2 = 0.6064 for 24-month VA prediction, with consistent performance across all forecast horizons. Our findings show that incorporating early treatment-response signals enables clinically meaningful long-term visual acuity forecasting, supporting data-driven decision support for routine anti-VEGF management. Code and pretrained models will be released on `https://github.com/nguyenpbui/ReVA`.

Keywords: Anti-VEGF treatment · Diabetic macular edema · Long-term prognosis · Multimodal learning · Visual acuity prediction.

1 Introduction

Visual acuity (VA) is the principal functional endpoint in macular disease and directly informs anti-VEGF treatment decisions, follow-up scheduling, and pa-

tient counseling. Optical coherence tomography (OCT) is routinely acquired at high resolution, enabling learning-based models that predict VA or treatment outcomes from image data, often augmented with tabular variables. Yet robust multi-horizon VA forecasting remains difficult in real-world cohorts due to heterogeneous baseline status, variable treatment response trajectories, structure–function discordance, and pervasive missing follow-up and label noise.

Existing VA forecasting approaches broadly span (i) classical machine learning on structured tabular variables for short- to mid-term prognosis [17], and (ii) deep learning that extracts prognostic representations from OCT, via either engineered imaging biomarkers or end-to-end modeling for cross-sectional and future outcomes [4, 18]. Multimodal variants further combine OCT with tabular variables to improve post-treatment VA and anatomical prediction [12], and attention-based or generative methods have been explored for response characterization [16]. Despite steady progress, two limitations constrain clinical deployment: most models target a single endpoint or short-term response rather than forecasting multiple future horizons, and many OCT encoders compress scans into global thickness/fluid-driven descriptors, underutilizing localized microstructural cues (e.g., outer retinal and photoreceptor integrity) that more directly bound achievable vision.

We study long-term VA prognosis in a clinically practical two-visit setting that explicitly leverages early treatment response. We curate a real-world cohort of 188 anti-VEGF–treated DME patients with paired baseline and month-1 OCT scans, together with tabular OCT-derived biomarkers and non-imaging clinical variables, and predict VA at 3, 6, 12, 18, and 24 months using only these early data. Our premise is that the change from baseline to month-1 encodes patient-specific treatment sensitivity and latent disease dynamics, providing a strong signal for long-range forecasting when represented with appropriate inductive bias. To this end, we propose a response-aware multimodal architecture that adapts a foundation-pretrained OCT transformer via LoRA, aggregates patch tokens with multi-head spatial attention to retain localized prognostic patterns, encodes tabular variables with a dependency-aware tabular module, and performs cross-attention fusion to obtain visit-specific multimodal representations. A lightweight temporal aggregation module then composes baseline state and early response into a compact embedding for multi-output VA prediction using temporal convolutional networks (TCN). Our main contributions are three-fold:

- We curate a real-world two-visit DME cohort and formulate an early-response, multi-horizon forecasting task to predict VA at 3–24 months from baseline and month-1 data.
- We develop **ReVA**, a response-aware multimodal framework that fuses paired OCT and tabular variables and explicitly models early post-treatment change for long-range VA forecasting.
- We achieve strong multi-horizon forecasting performance, with best month-24 results of MAE $= 0.1246$, RMSE $= 0.1621$, and $R^2 = 0.6064$, and validate key components via ablation studies.

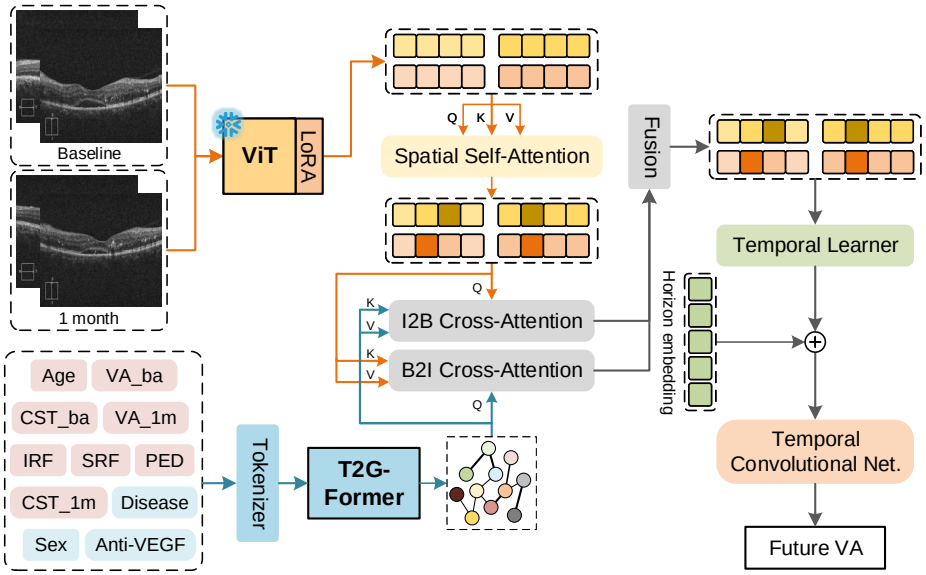

Fig. 1: Overview of the proposed response-aware multimodal framework for multi-horizon post-treatment VA forecasting from baseline and month-1 scans and tabular variables. The model fuses LoRA-adapted OCT features and dependency-aware tabular embeddings via cross-attention, then aggregates the two visits for multi-horizon VA prediction. *VA: Visual acuity; CST: Central subfield thickness; IRF: Intra-retinal fluid; SRF: Sub-retinal fluid; PED: Pigment epithelial detachment.*

2 Method

2.1 Problem Formulation

For each patient, we observe OCT scans from two visits (baseline and month-1) with two orthogonal scan views (horizontal/vertical), together with tabular OCT-derived biomarkers and non-imaging clinical variables. The OCT input is

$$\mathbf{I} = \{\mathbf{I}_H^{(0)}, \mathbf{I}_V^{(0)}, \mathbf{I}_H^{(1)}, \mathbf{I}_V^{(1)}\}, \tag{1}$$

where superscripts (0) and (1) denote baseline and month-1. Let $\mathbf{c}$ denote the tabular variables, consisting of numerical features (Age, $\mathrm{CST}_{\mathrm{base}}$, $\mathrm{CST}_{1\mathrm{m}}$, $\mathrm{VA}_{\mathrm{base}}$, $\mathrm{VA}_{1\mathrm{m}}$) and categorical features (Sex, IRF, SRF, PED, Disease, Anti-VEGF type). Our goal is multi-horizon regression of future VA:

$$f_\theta : (\mathbf{I}, \mathbf{c}) \mapsto \hat{\mathbf{y}} \in \mathbb{R}^T, \tag{2}$$

where $\hat{y}_t$ is the predicted VA at horizon $t \in \{1, \ldots, T\}$ (we use T=5 in our implementation). When follow-up labels are missing, we use an observation mask $\mathbf{m} \in \{0, 1\}^T$ to compute loss only over available horizons.

2.2 Foundation OCT Encoding and Tabular Tokenization

Each patient provides two visits (baseline t=0 and month-1 t=1) and two orthogonal B-scan views (horizontal/vertical). Each scan is encoded by a RETFound-initialized ViT-Large, producing a CLS token and N patch tokens:

$$\mathbf{Z} = [\mathbf{z}_{\mathrm{CLS}}, \mathbf{z}_1, \ldots, \mathbf{z}_N] \in \mathbb{R}^{(N+1)\times D}. \tag{3}$$

To enable parameter-efficient fine-tuning, we inject LoRA [7] adapters into attention projections (e.g., `qkv`, `proj`) by reparameterizing

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \tag{4}$$

with rank r and scaling α. When used, the backbone can be frozen so optimization is restricted to adapter parameters. Given patch tokens $X \in \mathbb{R}^{B\times N\times D}$, we apply multi-head self-attention pooling to obtain a single view descriptor while retaining spatial attributions:

$$A^{(h)} = \mathrm{softmax}\left(\frac{Q^{(h)}K^{(h)\top}}{\sqrt{d}}\right), \qquad \tilde{X} = \mathrm{LN}\big(X + \mathrm{Proj}(\mathrm{Concat}_h(A^{(h)}V^{(h)}))\big), \tag{5}$$

where $d = D/H$. The view embedding is computed by token pooling,

$$\mathbf{v} = \frac{1}{N}\sum_{n=1}^{N}\tilde{\mathbf{x}}_n \in \mathbb{R}^D. \tag{6}$$

We represent tabular variables (numerical and categorical features) as a token sequence with a learnable readout token,

$$\mathbf{T} = [\mathbf{t}_{\mathrm{CLS}}, \mathbf{t}_1, \ldots, \mathbf{t}_L] \in \mathbb{R}^{(L+1)\times d}. \tag{7}$$

Numerical features are embedded by feature-wise scaling of learnable token vectors, whereas categorical ones are embedded via lookup tables. We then encode the tabular tokens using T2G-FORMER [19] to capture higher-order feature interactions, and use the final readout token as the clinical embedding $\mathbf{g} \in \mathbb{R}^d$.

2.3 Cross-Modal Fusion and Delta-Gated Temporal Learner

Given OCT view descriptors $\mathbf{v}_k$ and a clinical embedding $\mathbf{g}$, we obtain view-wise multimodal features via cross-attention fusion:

$$\mathbf{f}_k = \mathrm{Fuse}(\mathbf{v}_k, \mathbf{g}), \qquad k \in \{1, \ldots, 4\}. \tag{8}$$

We then concatenate horizontal/vertical features within each visit,

$$\mathbf{b} = [\mathbf{f}_H^{(0)}; \mathbf{f}_V^{(0)}], \qquad \mathbf{m} = [\mathbf{f}_H^{(1)}; \mathbf{f}_V^{(1)}], \tag{9}$$

and summarize early response with a gated delta update:

$$\mathbf{s} = f(\mathbf{b}), \quad \Delta = g(\mathbf{m}-\mathbf{b}), \quad \boldsymbol{\gamma} = \sigma\big(h([\mathbf{s}; \Delta; |\Delta|; \|\Delta\|_2])\big), \quad \mathbf{z} = \mathbf{s} + \boldsymbol{\gamma}\odot\Delta. \tag{10}$$

Here f, g, h are MLPs and $\boldsymbol{\gamma}$ adaptively suppresses noisy month-1 changes. We further apply *delta dropout* by randomly zeroing $(\mathbf{m} - \mathbf{b})$ with probability p during training.

2.4 Time-Conditioned TCN for Multi-Horizon VA Regression

Given the temporally aggregated patient representation $\mathbf{z}$, we predict VA at T future horizons by casting multi-horizon forecasting as sequence regression over *ordered horizons*. This enables the predictor to share information across horizons and to model smooth, structured progression patterns without enforcing a rigid parametric trajectory. We first compute a shared latent descriptor

$$\mathbf{h} = \psi(\mathbf{z}) \in \mathbb{R}^H, \tag{11}$$

where $\psi(\cdot)$ is an MLP. To specialize the prediction for each horizon $t \in \{0, \dots, T-1\}$, we add a learnable horizon embedding $\mathbf{e}_t \in \mathbb{R}^H$ and form a horizon-conditioned token as follows:

$$\mathbf{s}_t = \mathbf{h} + \mathbf{e}_t, \qquad \mathbf{S} = [\mathbf{s}_0, \dots, \mathbf{s}_{T-1}] \in \mathbb{R}^{T \times H}. \tag{12}$$

This construction yields a compact horizon sequence $\mathbf{S}$ whose tokens share patient context ($\mathbf{h}$) while remaining horizon-aware via $\mathbf{e}_t$.

We then model inter-horizon dependencies with a dilated causal TCN, which captures both short-range coupling (nearby horizons) and long-range effects (distant horizons) using exponentially increasing receptive fields:

$$\tilde{\mathbf{S}} = \mathrm{TCN}(\mathbf{S}) \in \mathbb{R}^{T \times H}. \tag{13}$$

The TCN is implemented as a stack of residual temporal blocks with causal, dilated 1D convolutions and dropout, preserving the horizon order while preventing information leakage from later to earlier horizons. A linear readout produces per-horizon predictions

$$\hat{\mathbf{y}} = \rho(\tilde{\mathbf{S}}) \in \mathbb{R}^T, \tag{14}$$

where $\rho(\cdot)$ is a point-wise linear layer (equivalently a 1×1 temporal convolution). With ground-truth $\mathbf{y} \in \mathbb{R}^T$ and observation mask $\mathbf{m} \in \{0,1\}^T$, we optimize a masked regression objective

$$\mathcal{L}_{\mathrm{reg}} = \frac{\sum_{t=0}^{T-1} m_t \, \ell(\hat{y}_t, y_t)}{\sum_{t=0}^{T-1} m_t + \epsilon}, \tag{15}$$

where ℓ is ℓ_1 (MAE loss) and ϵ ensures numerical stability.

3 Experiments

3.1 Dataset and Evaluation Metrics

We curated a retrospective real-world cohort of N=188 anti-VEGF–treated DME patients with paired OCT scans acquired at baseline and month-1 after treatment initiation. Each visit includes two orthogonal B-scan views, yielding four OCT inputs per patient, together with OCT-derived biomarkers [2] and non-imaging clinical variables. The cohort had an age of 61.7±11.6

years, with 98 female/90 male patients and 97 left/91 right eyes. Baseline VA/CST were 0.499±0.248 logMAR and 425.5±97.8 μm, while month-1 VA/CST were 0.583±0.263 logMAR and 339.3±73.3 μm. IRF/SRF/PED were present in 176/73/3 of 188 cases. Mean VA at 3/6/12/18/24 months was 0.572±0.262/0.577±0.263/0.566±0.266/0.545±0.259/0.562±0.262 logMAR.

All splits were performed at the **patient level** to prevent data leakage [3]. OCT slices were resized to 224×224 and center-cropped around the fovea. Numerical variables were z-score normalized using training-fold statistics, and categorical variables were index-encoded. The primary endpoint was 24-month best-corrected VA in logMAR, with 3/6/12/18-month labels used as auxiliary supervision. Performance was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2), reported as mean±std over 5-fold cross-validation. Statistical significance was assessed by Wilcoxon signed test using out-of-fold patient-level absolute errors (MAE).

3.2 Implementation Details

Unless otherwise specified, we use RETFound ViT-Large as the OCT backbone (output dimension D=768) with frozen weights during training. The clinical encoder is a T2G-Former with token dimension d=128, three graph-attention layers, and four attention heads. Spatial attention uses four heads, and multimodal fusion uses a hidden dimension of 512. A GRU-based temporal aggregator models two-visit dynamics, followed by a TCN head with a time-embedding dimension of 16. Dropout is set to p=0.1. Models were optimized using Adam with early stopping based on validation MAE. All experiments were implemented in PyTorch and trained on NVIDIA A6000 GPU.

3.3 Comparison with State-of-the-Art

We compared our approach with representative CNN- and Transformer-based OCT backbones, as well as previously reported VA prognostic models. Quantitative results are summarized in Table 1. Compared with the strongest prior method, Rohm et al. [17], our approach reduced MAE by 10.17% and improved R^2 by 9.80% relative, with statistically significant differences ($p < 0.05$). From a clinical perspective, an MAE of approximately 0.12 logMAR corresponds to roughly one line on a standard VA chart. While a one-line difference may not independently alter treatment decisions, many clinical trials define meaningful visual change as a two- to three-line difference; thus, the primary value of our approach lies in improving long-term trajectory estimation and setting more informed expectations for patients. Beyond accuracy, spatial attention maps, shown in Fig. 2, consistently highlight pathologically relevant regions, including subfoveal fluid and photoreceptor disruption. This agreement between model attribution and clinically recognized features improves interpretability and supports the biological plausibility of the predictions.

Table 1: Comparison of 24-month VA regression (logMAR). Lower MAE/RMSE and higher R^2 indicate better performance (underline: second best, **bold**: best).

Method	MAE ↓	RMSE ↓	R^2 ↑	*p value*
R.-Bucheli et al. [18]	0.2358±0.0199	0.2806±0.0252	0.1552±0.0802	<0.001
Ko et al. [11]	0.1427±0.0147	0.1895±0.0143	0.4591±0.1254	<0.001
Leng et al. [12]	0.1536±0.0201	0.1911±0.0227	0.4633±0.0785	<0.001
Han et al. [5]	0.1657±0.0240	0.2091±0.0353	0.3541±0.1545	<0.001
Holste et al. [6]	0.1705±0.0167	0.2117±0.0236	0.3388±0.0986	<0.001
Kim et al. [10]	0.1588±0.0226	0.1972±0.0249	0.4275±0.0965	<0.001
Anderson et al. [1]	0.1486±0.0166	0.1874±0.0154	0.4809±0.0666	<0.001
Rohm et al. [17]	0.1387±0.0056	0.1741±0.0083	0.5470±0.0762	0.0027
Mondal et al. [15]	0.1396±0.0064	0.1729±0.0114	0.5496±0.1002	<0.001
ReVA	**0.1246**±0.0066	**0.1621**±0.0056	**0.6064**±0.0727	–

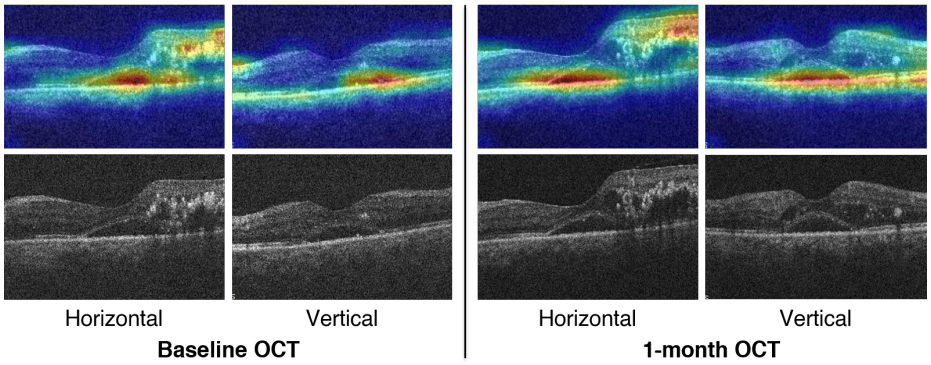

Fig. 2: Spatial attention heatmaps on representative OCT scans. Warmer colors indicate higher attribution to the predicted VA.

3.4 Ablation Studies

We conducted ablation analyses under identical protocols, focusing on 24-month VA prediction in Table 2. First, multimodal fusion outperformed both OCT-only and tabular variables-only models. Clinically, this suggests that structural OCT captures localized microstructural correlates of vision, whereas non-imaging tabular variables provide contextual information that may explain structure–function mismatch when OCT findings are subtle. Second, modeling dependencies among clinical variables using T2G-FORMER [19] consistently outperformed simpler tabular encoders, supporting the notion that clinical features do not act independently but interact in meaningful ways. Among OCT encoders, RETFound [20] yielded the strongest performance, consistent with the idea that foundation pretraining on retinal data enhances robustness in relatively small

Table 2: Ablation studies for 24-month VA regression (logMAR).

(a) Impact of different input modalities on prediction performance.

Model	MAE ↓	RMSE ↓	R^2 ↑
OCT-only	0.1571±0.0163	0.1957±0.0179	0.4354±0.0675
VA-only	0.1396±0.0122	0.1763±0.0103	0.5368±0.0742
Tabular-only	0.1298±0.0073	0.1651±0.0589	0.5916±0.0734
Multi-modal	**0.1246**±0.0066	**0.1621**±0.0056	**0.6064**±0.0727

(b) Comparison of clinical tabular variables and OCT image encoders.

Model	MAE ↓	RMSE ↓	R^2 ↑
MLP	0.1532±0.0196	0.1900±0.0208	0.4685±0.0725
TabTransformer [9]	0.1329±0.00684	0.1693±0.0113	0.5745±0.0567
T2G-FORMER [19]	**0.1246**±0.0066	**0.1621**±0.0056	**0.6064**±0.0727
Swin-T [13]	0.1453±0.0161	0.1828±0.0123	0.5038±0.0691
DenseNet [8]	0.1448±0.0124	0.1822±0.0128	0.5124±0.0141
ConvNeXt [14]	0.1432±0.0149	0.1785±0.0128	0.5268±0.0703
RETFound [20]	**0.1246**±0.0066	**0.1621**±0.0056	**0.6064**±0.0727

(c) Effect and results of intermediate-horizon auxiliary supervision (aux.).

Model	MAE ↓	RMSE ↓	R^2 ↑
ReVA w/o aux.	0.1341±0.0057	0.1667±0.0099	0.5829±0.0827
ReVA-3m	0.1135±0.0019	0.1495±0.0114	0.6687±0.0429
ReVA-6m	0.1261±0.0118	0.1640±0.0228	0.5846±0.1042
ReVA-12m	0.1218±0.0087	0.1556±0.0175	0.6420±0.1005
ReVA-18m	0.1283±0.0105	0.1631±0.0111	0.5914±0.0851
ReVA-24m	0.1246±0.0066	0.1621±0.0056	0.6064±0.0727

(d) Effectiveness of two-visit temporal aggregation strategies.

Model	MAE ↓	RMSE ↓	R^2 ↑
Naive concat.	0.1296±0.0081	0.1668±0.0079	0.5951±0.0758
Transformer-based	0.1287±0.0083	0.1646±0.0076	0.5924±0.0933
GRU-based	**0.1246**±0.0066	**0.1621**±0.0056	**0.6064**±0.0727

real-world cohorts with long-term outcome noise. Next, trajectory-level supervision improves long-term prognostic learning as shown in Table 1 (c). Finally, gated early-change modeling outperformed simple concatenation and generic sequence models. In clinical terms, explicitly modeling early treatment response appears more informative than treating visits as independent observations, reinforcing the importance of the early response signal in long-term prognosis.

4 Conclusion

We present a clinically practical two-visit framework for long-term visual prognosis after anti-VEGF therapy in DME patients, forecasting VA from 3 to 24 months using only baseline and 1-month OCT scans combined with tabular variables from a real-world cohort of 188 patients. By explicitly incorporating early treatment response through multimodal modeling, our approach achieved strong predictive performance. Beyond numerical improvements, the model provides trajectory-level estimates that may help clinicians anticipate disease course, individualize follow-up intensity, and support patient counseling. While further validation in larger and external cohorts is required, these findings suggest that response-aware multimodal learning offers a clinically meaningful step toward scalable, data-driven prognostic support in routine anti-VEGF care.

Acknowledgments. This work was supported in part by the Korea government (MSIT), IITP, under IITP-2026-RS-2020-II201821 (60%) and RS-2019-II190421 (10%); and by the Ministry of SMEs and Startups (MSS) under RS-2024-00514724 (30%). Professors D.T. Le and H. Choo are also with SKAI X Inc.

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

References

1. Anderson, M., Corona, V., Stankiewicz, A., Habib, M., Steel, D.H., Obara, B.: Enhancing post-treatment visual acuity prediction with multimodal deep learning on small-scale clinical and oct datasets. In: Medical Imaging with Deep Learning (2025)
2. Bui, P.N., Le, D.T., Bum, J., Han, J.C., Pham, V.N., Choo, H.: Multi-scale feature enhancement in multi-task learning for medical image analysis. Artificial Intelligence in Medicine p. 103338 (2025)
3. Bui, P.N., Le, D.T., Bum, J., Kim, S., Song, S.J., Choo, H.: Multi-scale learning with sparse residual network for explainable multi-disease diagnosis in oct images. Bioengineering **10**(11), 1249 (2023)
4. Fu, D.J., Faes, L., Wagner, S.K., Moraes, G., Chopra, R., Patel, P.J., Balaskas, K., Keenan, T.D., Bachmann, L.M., Keane, P.A.: Predicting incremental and future visual change in neovascular age-related macular degeneration using deep learning. Ophthalmology Retina **5**(11), 1074–1084 (2021)
5. Han, J.M., Han, J., Ko, J., Jung, J., Park, J.I., Hwang, J.S., Yoon, J., Jung, J.H., Hwang, D.D.J.: Anti-vegf treatment outcome prediction based on optical coherence tomography images in neovascular age-related macular degeneration using a deep neural network. Scientific reports **14**(1), 28253 (2024)
6. Holste, G., Lin, M., Zhou, R., Wang, F., Liu, L., Yan, Q., Van Tassel, S.H., Kovacs, K., Chew, E.Y., Lu, Z., et al.: Harnessing the power of longitudinal medical imaging for eye disease prognosis using transformer-based sequence modeling. NPJ Digital Medicine **7**(1), 216 (2024)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)

8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
9. Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 (2020)
10. Kim, N., Lee, M., Chung, H., Kim, H.C., Lee, H.: Prediction of post-treatment visual acuity in age-related macular degeneration patients with an interpretable machine learning method. Translational Vision Science & Technology **13**(9), 3–3 (2024)
11. Ko, Y.C., Peng, C.W., Ho, H.C., Chiu, S.H., Chen, S.J., Lee, C.Y.: Deep learning assisted prediction of long-term visual outcome after 3 monthly anti-vascular endothelial growth factor injections in patients with central-involved diabetic macular edema. Investigative Ophthalmology & Visual Science **63**(7), 3778–F0199 (2022)
12. Leng, X., Shi, R., Xu, Z., Zhang, H., Xu, W., Zhu, K., Lu, X.: Development and validation of cnn-mlp models for predicting anti-vegf therapy outcomes in diabetic macular edema. Scientific Reports **14**(1), 30270 (2024)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
15. Mondal, A., Nandi, A., Pramanik, S., Mondal, L.K.: Application of deep learning algorithm for judicious use of anti-vegf in diabetic macular edema. Scientific Reports **15**(1), 4569 (2025)
16. Moon, S., Lee, Y., Hwang, J., Kim, C.G., Kim, J.W., Yoon, W.T., Kim, J.H.: Prediction of anti-vascular endothelial growth factor agent-specific treatment outcomes in neovascular age-related macular degeneration using a generative adversarial network. Scientific reports **13**(1), 5639 (2023)
17. Rohm, M., Tresp, V., Müller, M., Kern, C., Manakov, I., Weiss, M., Sim, D.A., Priglinger, S., Keane, P.A., Kortuem, K.: Predicting visual acuity by using machine learning in patients treated for neovascular age-related macular degeneration. Ophthalmology **125**(7), 1028–1036 (2018)
18. Romo-Bucheli, D., Erfurth, U.S., Bogunović, H.: End-to-end deep learning model for predicting treatment requirements in neovascular amd from longitudinal retinal oct imaging. IEEE Journal of Biomedical and Health Informatics **24**(12), 3456–3465 (2020)
19. Yan, J., Chen, J., Wu, Y., Chen, D.Z., Wu, J.: T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10720–10728 (2023)
20. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)