Predicting Progressive Vision Loss in Glaucoma Patients Using Functional Principal Component Analysis and Electronic Health Records

Stanford University

DATA INSTITUTE — UNIVERSITY OF SAN FRANCISCO

Rithvik Krishna Donnipadu, MS[1], Maxim Sivolella, MS[1],
Cody Carroll, PhD[1], Sophia Y. Wang, MD, MS[2]

[1]Data Institute, University of San Francisco, California, USA , [2]Stanford University, Palo Alto, California, USA

Introduction

Glaucoma, the Silent Thief of Sight

Figure 1: Typical Vision Progression Resulting from Glaucoma

Background

Glaucoma is a chronic progressive disease of the optic nerve and is one of the leading causes of irreversible blindness, affecting an estimated 76 million people in 2020, a number that is expected to grow to 119 million by 2040[1]. It is estimated that around half of the individuals suffering from glaucoma do not realize they have the disease, which attacks the peripheral vision first. This makes it difficult to notice vision deterioration and seek a diagnosis. Due to this, glaucoma has been characterized as the "silent thief of sight," as many are unaware of the disease until irreversible damage has been done[2].

Things to know

- Visual field (VF) testing quantifies vision loss and is summarized by the mean deviation (MD).
- Fast progressors can lose more than 1 dB/year of MD[3,4]. Slow progressors may be stable for years.
- Predicting a patient's future trajectory would help physicians decide on the ideal treatment strategy.

Objective

We propose a method of estimating a patient's future MD values by combining clinical data from electronic health records (EHR) and Functional Principal Component Analysis (FPCA). FPCA is a longitudinal technique which estimates a population-level mean trajectory and its major modes of variation, then leverages this shared information across patients to tailor predictions for individuals' future curves.

Cohort Creation & Data Sources

Visual Field (VF) Data

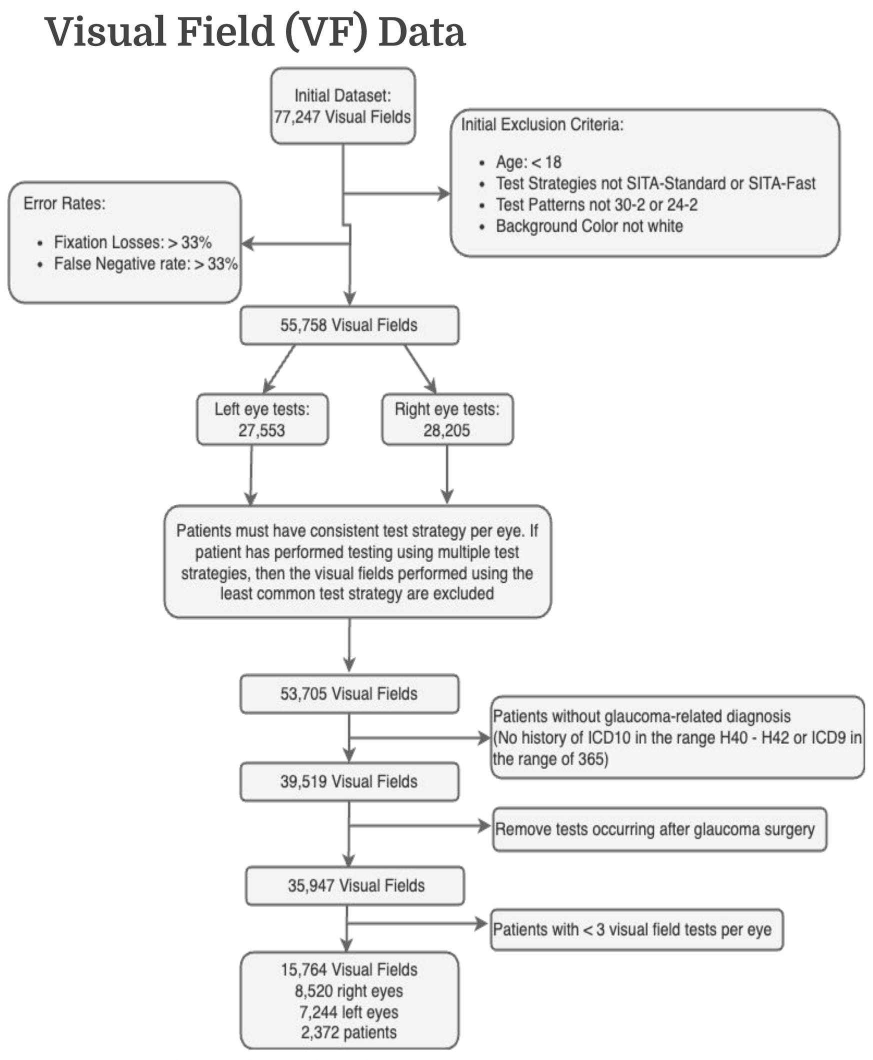

Figure 2: Flowchart outlining the cohort creation process

Electronic Health Record (EHR) Data

1. EHR data taken only from data recorded on or before a patient's first visual field test.
2. Processed EHR data, creating features such as intraocular pressure (IOP), corneal thickness, cup to disc ratio, visual acuity, spherical equivalent, and more.
3. One hot encoded (OHE) systemic and ocular medications and diagnoses (as determined by International Classification of Disease [ICD] codes).
4. Applied a variance threshold set at 1% to our OHE columns, eliminating rare diagnoses and medications.
5. These features were joined with the visual field data, resulting in 473 features altogether.

Methods

Overview: We have proposed a 3 stage model to predict the Mean Deviation of our patients (Fig 5):

Stage 1: Use Logistic Regression to classify patients into fast/slow progressors, with baseline EHR and first VF as inputs.

For each of the categories of progression:

Stage 2: Use Functional Principal Component Analysis on the longitudinal VF data to predict future Mean Deviation (MD) values for a given patient

Stage 3: Use a Random Forest regressor to incorporate baseline EHR data to fine-tune the FPCA predictions and reduce errors between stage 2 predictions and actual MDs.

Stage 1 : Logistic Regression

For classification of eyes into fast or slow progressors, we used Logistic Regression. LR had similar performance as TabNet but due to ease of implementation and explainability we opted to use LR. Detailed steps are as follows.

- Identify date of first VF for each eye and filter the recorded EHR clinical information on or prior to that date resulting in a cohort of (N=4,081 eyes).
- Data was split 80:10:10 into a train, validation and test set, ensuring that eyes from the same patient remained in the same split.
- Outliers, consisting of eyes whose slopes were greater than 5 or less than -20 dB/year were removed.
- Hyperparameters were tuned on a validation set and and classification threshold was set to 0.185, such that predicted probabilities >0.185 were considered fast progressors.
- Logistic Regression achieved a best AUC score of 0.77 on the test set for distinguishing between fast and slow progressors.

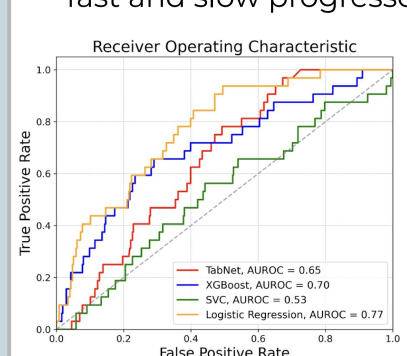

Figure 3: ROC curve of different ML models

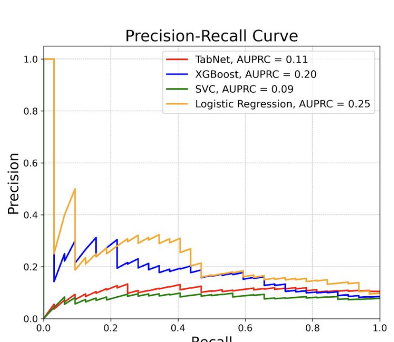

Figure 4: Precision Recall curve of different ML models

Stage 2 : Functional PCA Model

Based on the classification of patients into fast/slow progressors from the first stage, we split the following process into two separate steps, one for each of the progressor category.

- An FPCA was fit to predict future MD trajectories up through ~10 years after their first VF exam using the R package *fdapace*[4,5] (Fig 6).
- Input features included patients' first year of VF exam data.
- The predicted trajectories were fitted using the first 3 principal components, which together explained roughly 99% of the variation in the original training data (Fig. 7(left))

Stage 3 : Random Forest

- Finally, residuals from the FPCA model were calculated for each training observation.
- These were taken as the targets for a random regressor model. EHR data, VF data, and FPCA fitted values across the timeline were taken as input features.
- On adding the residual to the prediction, we were able to improve the R^2 value from ~67% to ~74% for slow progressors and ~55% to ~71% for fast progressors (Fig. 7(right)).

Mean Function and Eigenfunctions from Functional PCA

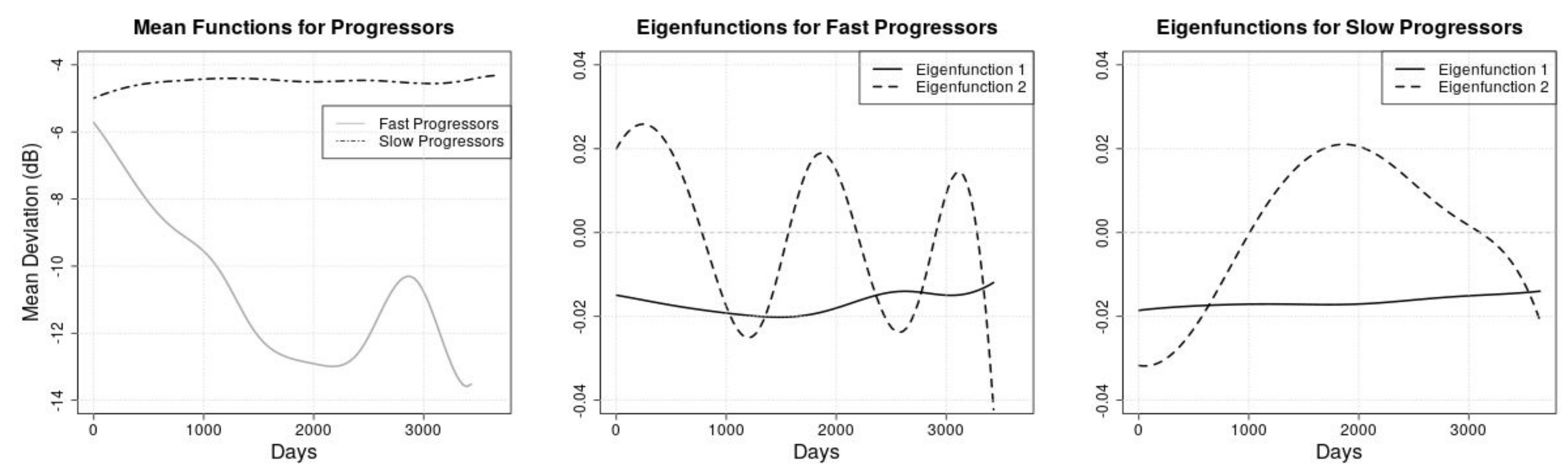

Figure 6: Mean function of both the progression categories (left), first 2 eigenfunctions (middle) which explain 99.5% of the variance for the fast progressors, and first 2 eigenfunctions (right) which explain 99.9% of the variance for the slow progressors.

Functional PCA predicted trajectories

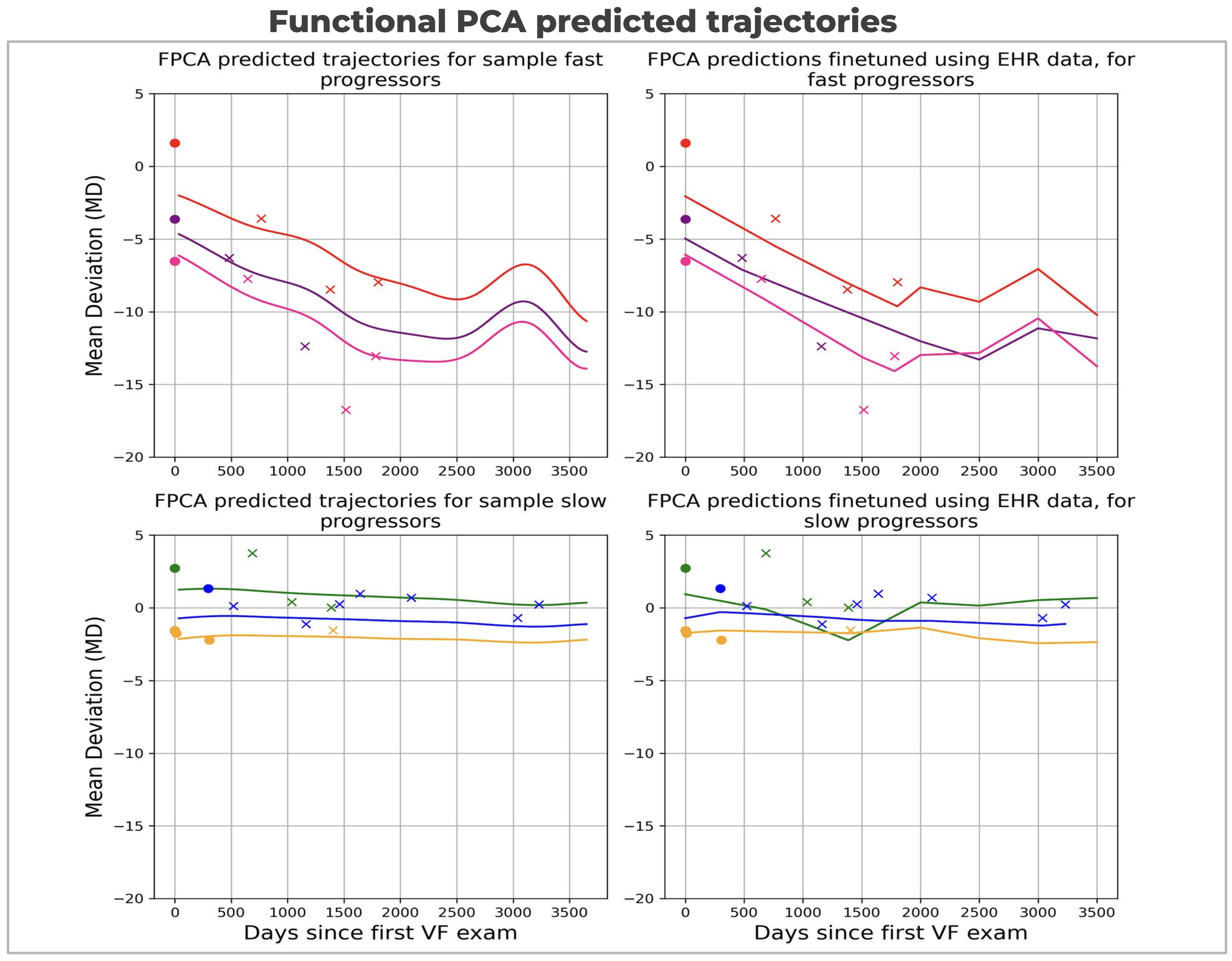

Figure 7:Predicted VF trajectories compared to actual observed VF MD. The figure illustrates example Stage 2 predicted trajectories for Fast progressors (top) and Slow progressors (bottom) over a 10-year period. Using one year of test data (indicated by dotted circles), trajectories (indicated by the continuous line) were fitted for each patient employing Functional Principal Component Analysis (FPCA). The final predicted trajectories for three Fast progressors (top right) and three Slow progressors (bottom right) over a 10-year period are also shown.

Scatter plot for actual vs predicted Mean Deviation values

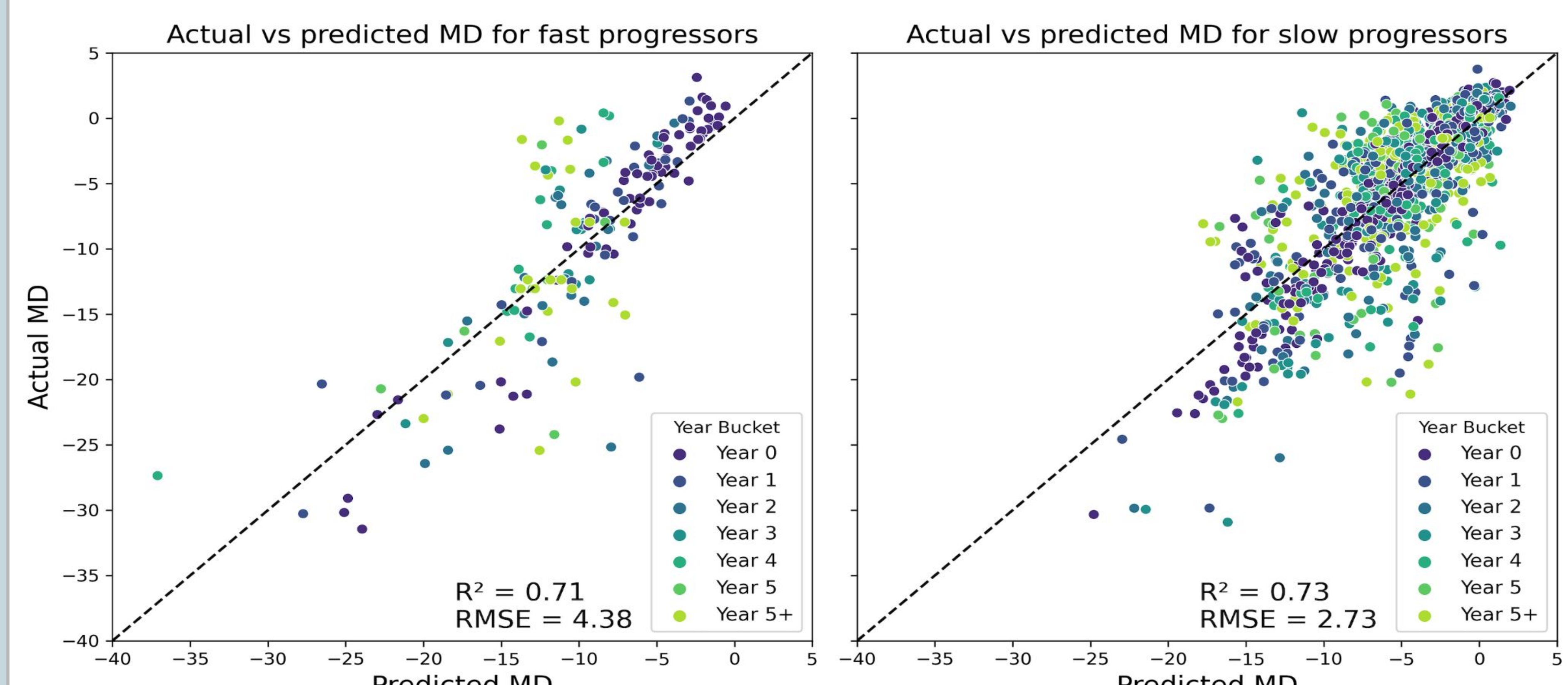

Figure 8:Scatterplots for the actual vs predicted MD values for fast and slow progressors. The points on each plot are distinguished by color, representing different prediction time horizons.

Results

- For the fast progressors, on a test set of **364** visual fields, prediction of MD trajectories ~10 years into the future using 1 year of historical data yielded an R^2 of 0.708 and an RMSE of 4.38.
- For the slow progressors, on a test set of **3717** visual fields, prediction of MD trajectories ~10 years into the future using 1 year of historical data yielded an R^2 of 0.735 and an RMSE of 2.73.

Model Architecture

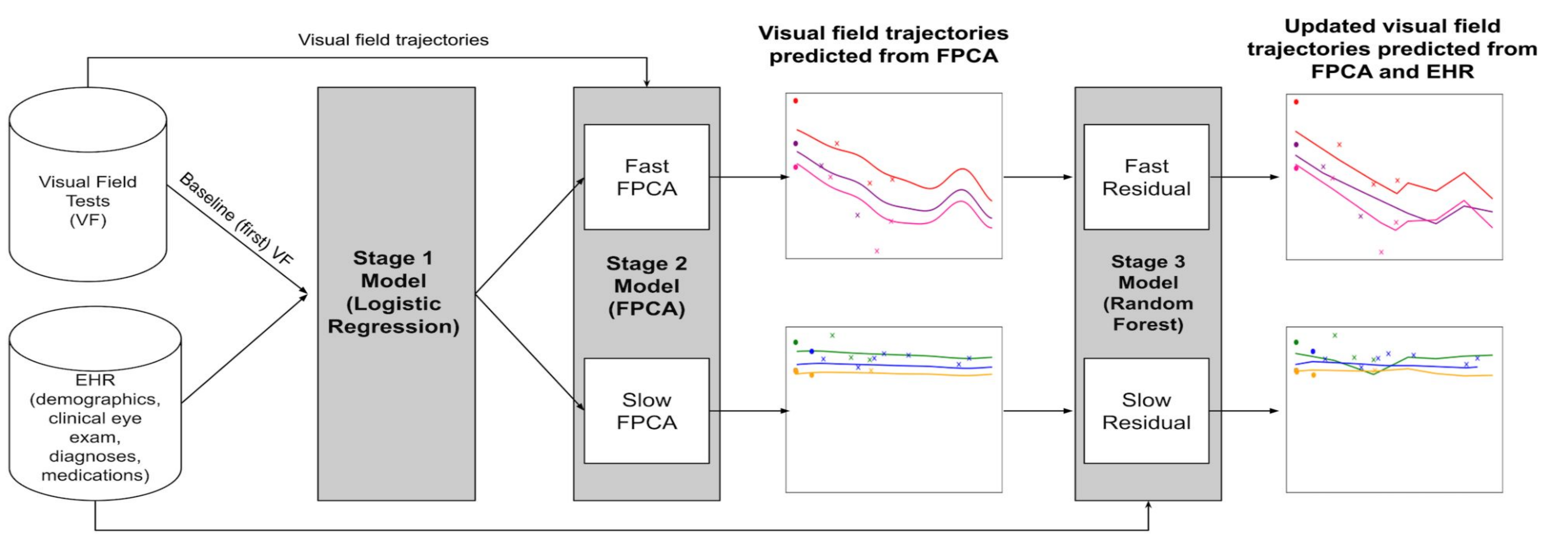

Figure 5: Architecture of the 3 stage model for predicting future vision loss

Conclusion

In this study, we present a novel application of a three-stage machine learning model to predict glaucoma patients' future visual field results, achieving an test set RMSE of ~4.3(fast) and ~2.7(slow), a performance surpassing prior benchmarks[6,7]. An advantage of this model is that it can be used in sparse and irregular observation schedules and can also incorporate baseline EHR data.

Using one year of visual field test data, our model can predict the next 10 years of a patient's disease progression at state-of-the-art accuracy. This predictive power enables doctors to time their surgical interventions optimally and better prevent irreparable losses in vision.

References

1. Allison K, et al. Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future. Cureus. 2020.
2. Weinreb RN, et al. The pathophysiology and treatment of glaucoma: a review. JAMA. 2014.
3. Jackson AB, et al. Fast Progressors in Glaucoma: Prevalence Based on Global and Central Visual Field Loss. Ophthalmology. 2023.
4. Carroll C, et al. fdapace: Functional data analysis and empirical dynamics. R package version 0.5.7, 2023.
5. Yao F, et al. Functional Data Analysis for Sparse Longitudinal Data. J Am Stat Assoc. 2005.
6. Wen J, et al. Forecasting future Humphrey Visual Fields using deep learning. Plos One. 2019.
7. Keunheung P, et al. Visual Field Prediction using Recurrent Neural Network. Scientific Reports. 2019.