tvst

Artificial Intelligence

Multimodal Artificial Intelligence Models Predicting Glaucoma Progression Using Electronic Health Records and Retinal Nerve Fiber Layer Scans

Abigail Koornwinder[1], Youchen Zhang[1], Rohith Ravindranath[1], Robert T. Chang[1], Isaac A. Bernstein[1], and Sophia Y. Wang[1]

[1] Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA, USA

Correspondence: Sophia Y. Wang, Department of Ophthalmology, Byers Eye Institute, Stanford University, 2452 Watson Court, Palo Alto, CA 94303, USA. e-mail: sophia.yw@gmail.com

Received: July 18, 2024
Accepted: February 16, 2025
Published: March 28, 2025

Keywords: glaucoma; multimodal; deep learning; optical coherence tomography (OCT); electronic health records (EHRs)

Citation: Koornwinder A, Zhang Y, Ravindranath R, Chang RT, Bernstein IA, Wang SY. Multimodal artificial intelligence models predicting glaucoma progression using electronic health records and retinal nerve fiber layer scans. Transl Vis Sci Technol. 2025;14(3):27, https://doi.org/10.1167/tvst.14.3.27

Purpose: The purpose of this study was to develop models that predict which patients with glaucoma will progress to require surgery, combining structured data from electronic health records (EHRs) and retinal fiber layer optical coherence tomography (RNFL OCT) scans.

Methods: EHR data (demographics and clinical eye examinations) and RNFL OCT scans were identified for patients with glaucoma from an academic center (2008–2023). Comparing the novel TabNet deep learning architecture to a baseline XGBoost model, we trained and evaluated single modality models using either EHR or RNFL features, as well as fusion models combining both EHR and RNFL features as inputs, to predict glaucoma surgery within 12 months (binary).

Results: We had 1472 patients with glaucoma who were included in this study, of which 29.9% ($N = 367$) progressed to glaucoma surgery. The TabNet fusion model achieved the highest performance on the test set with an area under the receiver operating characteristic curve (AUROC) of 0.832, compared to the XGBoost fusion model (AUROC = 0.747). EHR only models performed with an AUROC of 0.764 and 0.720 for the deep learning model and XGBoost models, respectively. RNFL only models performed with an AUROC of 0.624 and 0.633 for the deep learning and XGBoost models, respectively.

Conclusions: Fusion models which integrate both RNFL with EHR data outperform models only utilizing one datatype or the other to predict glaucoma progression. The deep learning TabNet architecture demonstrated superior performance to traditional XGBoost models.

Translational Relevance: Prediction models that utilize the wealth of structured clinical and imaging data to predict glaucoma progression could form the basis of future clinical decision support tools to personalize glaucoma care.

Introduction

Glaucoma is a chronic progressive disease of the optic nerve and is the leading global cause of irreversible blindness.[1] In its initial stages, glaucomatous damage to the optic nerve is often asymptomatic; however, if left undetected and untreated, glaucoma can progress to irreversible vision loss.[2] Thus, early detection and timely treatment of glaucoma are crucial for effective management and preservation of vision. Notably, some patients with glaucoma remain stable without progression for extended periods, whereas others experience progressive disease that eventually requires invasive interventions like incisional glaucoma surgery.[3] Prediction algorithms that could identify which patients are most likely to progress could allow clinicians to tailor treatments to individual disease profiles and more effectively prevent blindness.

Prior studies have demonstrated the potential of machine learning and deep learning techniques to predict glaucoma progression to surgery, highlight-

ing the importance of this question to glaucoma care.[4–7] Our own prior work has developed promising models that have leveraged the wealth of electronic health records (EHRs) data in structured and free-text formats to outperform glaucoma specialists in predicting which patients with glaucoma will progress to require surgery.[8–11] However, clinical glaucoma risk assessment usually also includes imaging and functional testing of the optic nerve, including retinal nerve fiber layer optical coherence tomography (RNFL OCT) scans, which provide detailed structural information about the optic nerve and helps clinicians assess how much glaucomatous damage may have occurred already. The optimal approach for integrating these imaging data with existing structured EHR data to improve the performance of prediction models is indeterminate. Prior work has applied deep learning models to data from RNFL OCT to predict glaucoma outcomes but work fusing data from both the RNFL OCT and the EHR clinical information is limited.[12]

Although deep learning methods have revolutionized many tasks involving text and images, its benefits for structured data have not been as remarkable. In many prediction models using structured data from EHRs, traditional machine-learning techniques, such as tree-based models, have been shown to be relatively effective.[11] The recently developed TabNet model is a cutting-edge attention-based deep learning model architecture especially designed for handling diverse tabular datasets, with minimal preprocessing, and it has been shown to perform exceptionally well in comparison to traditional machine learning models on a variety of benchmark tasks.[13] TabNet's sequential multistep architecture allows for informed decision making, where each step contributes to the final outcome based on a robust soft feature selection. TabNet thus holds great potential for healthcare applications using EHR data. However, the use of TabNet for building predictive algorithms using EHR data has been limited, and the integration of different modalities of data into TabNet models even more so. Hence, the objective of this study is to build on our previous work developing and assessing models that predict the progression of patients with glaucoma to surgery, comparing fusion models that incorporate EHR and RNFL OCT data to single-modality models. In doing so, we also compare the TabNet deep learning architecture to standard XGBoost modeling on a clinically relevant prediction task that uses real-world EHR and imaging data, which may provide insights into its suitability for modeling with these data types even beyond the scope of ophthalmology.

Methods

Data Source and Study Population

We identified patients from the Stanford Research Repository (STARR) who had encounters in the Department of Ophthalmology at Stanford University from 2008 to January 23, 2023. Our glaucoma cohort included individuals with a minimum of two encounters associated with a glaucoma-related International Classification of Diseases tenth revision diagnosis code (ICD-10 H40, H42, or Q15 and their descendants) or those who underwent glaucoma surgery based on Current Procedural Terminology codes (66150, 66155, 66160, 66165, 66170, 66172, 66174, 66175, 66179, 66180, 66183, 66184, 66185, 67250, 67255, 0191T, 0376T, 0474T, 0253T, 0449T, 0450T, 0192T, 65820, 65850, 66700, 66710, 66711, 66720, 66740, 66625, and 66540).[14] Patients with only glaucoma suspect codes (H40.0 and ICD 365.0 and their descendants) were excluded. This study adhered to the tenets of the Declaration of Helsinki.

Prediction Timeline

The goal was to develop models that could predict which patients with glaucoma are at the highest risk of progression to surgery and that could be flexible enough to be used at any time point during a patient's treatment trajectory by incorporating their latest clinical data. As patients have many encounters in their treatment trajectory, for training purposes, a single particular encounter was chosen as the "prediction date" and the models were trained to predict whether patients with glaucoma would undergo surgical intervention within the subsequent 12 months following that specific encounter, consistent with prior similar studies.[8,11] Thus, each patient was included in the cohort only once. We utilized clinical data from up to 1.5 years prior to and including the prediction encounter for model training. Thus, the prediction timeline for each patient was developed, defining a look-forward period over which the model would predict progression to surgery, and a lookback period of up to 1.5 years from which the models' input data were drawn. The prediction date was randomly chosen within specific criteria: for surgical patients, the prediction date was chosen to fall on an encounter date within 12 months prior to glaucoma surgery, ensuring the prediction timeline captured the period leading to surgery. For non-surgical patients, the prediction date was at least 12 months prior to the end of follow-up, ensuring that enough follow-up had occurred to

confirm that no surgery occurred during the look-forward period. This approach to prediction empowers the models to predict which patients are at the highest risk of progression to surgery over the next year. This flexibility maximizes the potential for future deployment of these models as predictions could continuously update with each visit. Example timelines for surgery and non-surgery patients are included in Supplementary Figure S1.

Feature Engineering

Our dataset encompasses two distinct data modalities: EHR data consisting of an eye examination and demographics information, as well as data extracted from RNFL OCT imaging scan reports, providing structural insights into the optic nerve. Our predictive models adopt a laterality-agnostic approach, forecasting future glaucoma surgery in either eye, at the patient level. This design choice accommodates cases where the decision to proceed with surgery in one eye is influenced by the status of the contralateral eye. All EHR and RNFL data within the lookback period of 1.5 years before the prediction date were acquired.

EHR Data

Demographic data included patient sex, age at prediction date, and race/ethnicity. Race/ethnicity and gender were categorical variables which were dummy encoded for model input. Structured eye examination data included visual acuity, intraocular pressure (IOP), central corneal thickness (CCT), and spherical equivalent from both eyes. Visual acuity was converted into logMar units and summarized into best recorded, worst recorded, last recorded, and mean. Spherical equivalents were calculated from refraction measurements for each eye. IOP was summarized into highest, lowest, median, and most recent values for this feature. The most recent CCT values for each patient were extracted. Where multiple spherical equivalent measurements were available, the highest absolute value across examinations for right and left eyes were used as input features for each patient. Continuous variables were appropriately standardized or scaled: IOP was standardized to mean 0 and standard deviation of 1; age was scaled by 100; and CCT was scaled by 1000; and spherical equivalent was scaled by 10. Mean imputation was performed for missing values of IOP, CCT, and spherical equivalent.

RNFL Data

Features extracted from the most recent eligible RNFL OCT scan in the lookback period for each patient included average RNFL thicknesses, cup-to-disc ratios, rim and disc areas, and quadrant thicknesses (superior, temporal, nasal, and inferior) from both eyes. Only data from scans with a signal strength ≥ 6 were included. For average RNFL thickness features, values below 40 or above 160 were considered extreme and filtered out. Average RNFL thickness and quadrant thickness values were scaled by dividing by 100. We used a last value carried forward method to address missing information in RNFL scans for patients who had multiple scans in the lookback period. All RNFL scans were captured using Zeiss Cirrus HD-5000 machines, and macular scans were not included.

Modeling

Dataset Formation

The cohort was split for model training, model validation, and evaluation, allocating 66% ($N = 972$) for training, 17% ($N = 250$) for validation, and 17% ($N = 250$) for the test set. The data were split using stratified sampling due to the class imbalance (29.9% were positive class patients who progressed to surgery). Three datasets were formed for modeling and evaluation purposes including an EHR-only dataset, an RNFL-only dataset, and a fusion dataset where the EHR and RNFL datasets were concatenated. The inclusion of the single modality datasets allowed for a comparison against the multimodal fusion approach. Importantly, the same set of patients were consistent across each modality and data split to ensure uniformity across dataset configurations. To address class imbalance, we applied Synthetic Minority Oversampling (SMOTE) on the training data.[15]

TabNet and XGBoost Models

We trained deep learning models based on the TabNet architecture, utilizing the PyTorch version 1.7.1 package.[13,16] TabNet's architecture leverages attention mechanisms to focus on important features, enhancing interpretability. TabNet allows modelers to specify groupings for related features, enabling the models to share attention among related features.[16] Our model grouped demographic features together, leaving the rest of the RNFL and eye examination features ungrouped so as to allow the model to learn attention across the modalities. Hyperparameters were fine-tuned on the validation set to optimize the area under the receiver operating characteristic curve (AUROC). For the TabNet fusion model, training parameters included using the Adam optimizer with a learning rate of 0.02, and a step size of 25 with a gamma value of 0.9 for the learning rate scheduler, using the StepLR scheduler function. Addition-

ally, entmax was used as the masking function, and the maximum number of training epochs was set to 100, with early stopping implemented through a patience parameter of 20.[17] Batch size was also selected through hyperparameter tuning on the validation set and each training iteration involved a batch size of 128 samples, with a virtual batch size of 32. Baseline tree-based models (XGBoost) were fit using the Python sklearn version 1.3.0 package. XGBoost was chosen as a baseline comparison model due to its high performance in previous studies using similar EHR data to perform prediction tasks for glaucoma.[8,11] Hyperparameter tuning was conducted using a grid search approach and using a 5-fold cross-validation. Tuned hyperparameters for the XGBoost fusion model included a learning rate of 0.2, maximum depth of 4, and 150 as the number of estimators. The classification thresholds for XGBoost and TabNet were tuned to optimize F1 score on the validation set. To ensure robust evaluation given the relatively small dataset, in a sensitivity analysis, we performed additional model training and evaluation using nested cross-validation, with five-fold cross-validation for hyperparameter tuning and performance metrics averaged across the five folds.

Evaluation

Metrics

We used standard classification metrics to evaluate model performances on the test set including sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, F1-score (the harmonic mean of recall and precision), AUROC, and area under the precision-recall curve (AUPRC).

Explainability

We performed model explainability studies utilizing model-agnostic and model-specific techniques. As a model-agnostic method, we used SHapley Additive exPlanation (SHAP) values to directly compare TabNet and XGBoost models.[18,19] This method uses a game theory approach to determine feature importance by analyzing the magnitude of feature attributions, thus enhancing interpretability for artificial intelligence (AI) models. SHAP values represent the marginal contribution to the model predictions for each feature, calculated across all possible combinations or feature subsets. We estimated SHAP values on the test set for both fusion models.

In addition to our Shapley studies, we specially leveraged TabNet's model-specific explainability functions to uncover a more direct understanding of the decision-making process underlying the deep learning architecture. TabNet's model-specific explainability method is derived from the attention mechanisms of the architecture itself, resulting in more accurate feature analysis.[13] Using these built-in functions, we calculated overall feature importances as well as instance-wise feature importance for the test set. TabNet's unique instance-wise feature importance identifies the most relevant features for each individual in the dataset. By conducting TabNet-specific explainability studies alongside the commonly used SHAP approach, we can assess differences and similarities across explainability methods.

Code Availability

Code for this project is available at https://github.com/akoornwinder4/multimodal-glaucoma-surgery-prediction.[20]

Results

Study Population

Population characteristics for the entire study cohort of 1472 patients with glaucoma are summarized in Table 1. Of these patients, 29.9% ($N = 367$) progressed to glaucoma surgery. The majority of the cohort was Asian (38.5%, $N = 566$) and White (28.9%, $N = 425$), and the overall mean age was 67.71 years (standard deviation = 15.43). Mean vCDR was 0.68 for both the right and left eye, whereas mean RNFL thickness was 75.86 (OD) and 75.44 (OS). The full distribution of vCDR and RNFL thickness for this cohort is shown in Supplementary Figure S2.

Model Performance

Receiver operating characteristic and precision recall curves for XGBoost and TabNet models are shown in Figure 1. Fusion models using both EHR and RNFL modalities of data generally outperformed models trained with only EHR data, and both of these approaches were superior to models trained with only RNFL data. Overall, the TabNet fusion model achieved the highest performance with AUROC of 0.832, and the XGBoost fusion model achieved AUROC of 0.747. The TabNet fusion model achieved an AUPRC of 0.541, whereas the XGBoost fusion model achieved AUPRC of 0.510. The concavity of the TabNet fusion model precision-recall curve, in contrast to the decreasing curve of the XGBoost model, indicates a well-balanced trade-off between precision and recall across a wider range of decision thresholds. Additional classification metrics

Table 1. Population Characteristics

	No Surgery $N = 1105$		Surgery $N = 367$		Total $N = 1472$	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Age, y	67.24	15.57	69.12	14.94	67.71	15.43
Best logMAR VA, OD	0.15	0.47	0.2	0.52	0.16	0.48
Best logMAR VA, OS	0.18	0.45	0.18	0.45	0.19	0.53
IOP maximum, OD, mm Hg	20.21	6.42	25.19	8.96	21.45	7.45
IOP maximum, OS, mm Hg	20.59	7.65	26.82	10.28	22.14	8.81
Spherical equivalent, OD	−1.41	6.69	−1.34	4.87	−1.4	6.28
Spherical equivalent, OS	−1.32	5.49	−1.13	4.97	−1.27	5.37
CCT, OD, um	548.33	45.27	545.56	54.51	547.62	47.81
CCT, OS, um	551.11	54.74	546.54	47.73	549.94	53.05
RNFL average thickness, OD	75.86	14.26	75.14	16.78	75.68	14.93
RNFL average thickness, OS	75.44	14.16	74.06	16.18	75.1	14.7
Vertical cup-to-disc ratio, OD	0.68	0.14	0.69	0.15	0.69	0.15
Vertical cup-to-disc ratio, OS	0.68	0.15	0.7	0.16	0.69	0.15
	N	%	*N*	%	*N*	%
Female	574	51.95	185	50.41	759	51.56
Race/ethnicity						
White	308	27.87	116	31.61	425	28.87
Black	43	3.89	17	4.63	60	4.08
Asian	453	41.00	113	30.79	566	38.45
Hispanic	118	10.68	71	19.35	189	12.84
Other	162	14.66	46	12.53	208	14.13
Declines to state	21	1.90	3	0.82	24	1.63

CCT, central corneal thickness; IOP, intraocular pressure; OD, right eye; OS, left eye; RNFL, retinal nerve fiber layer; VA, visual acuity.

for TabNet and XGBoost models are summarized in Table 2. Results from the supplementary sensitivity analysis using nested cross-validation were closely aligned with the primary findings (Supplementary Table S1).

Feature Importance and Explainability

We conducted explainability analyses to investigate which features from the EHR and RNFL modalities most impacted model predictions toward surgery or no surgery. Shapley values were calculated for the test set for both the TabNet and XGBoost fusion models (Fig. 2). This is a model-agnostic approach which calculates the marginal contribution of each feature to the prediction based on analysis of model outputs with different combinations of input features. A negative Shapley value indicates that the feature influences the model toward predicting no surgery, whereas a positive Shapley value indicates that the feature influences the model toward predicting surgery. Vision and IOP-related features were among the top most important features for both TabNet and XGBoost models. Race and ethnicity were also among the important features for the TabNet model. RNFL features represented 8 of the top 20 most important features for the XGBoost fusion model, and 7 for the TabNet model.

We also utilized TabNet's unique model-specific explainability methods to further investigate feature importance for the TabNet fusion model (Fig. 3). TabNet's model architecture incorporates feature selection masks, which can be used to identify which features have the greatest attention at sequential layers of the model architecture. Feature selection occurs at an instance-wise level, which enables identification of features most relevant to a specific patient's prediction. Instance-wise importance is then aggregated to reveal which features were most important globally. RNFL features included 9 of the top 20 most important features for the TabNet model, whereas race/ethnicity features do not emerge at all among the most important features. Average RNFL thickness, an important global structural measure of the optic nerve, is among the top features according to TabNet's innate feature importance, but not in the Shapley analysis.

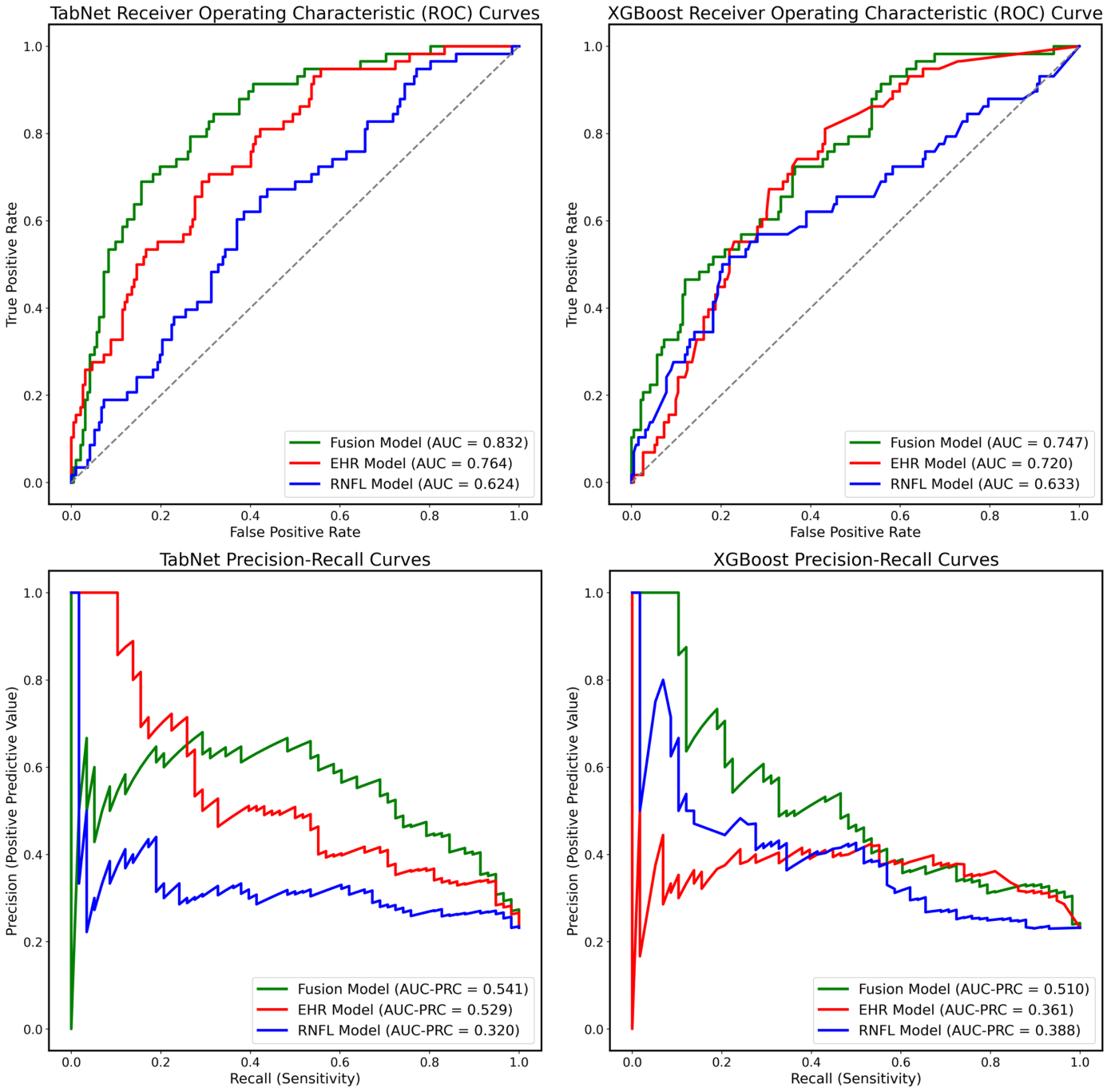

Figure 1. TabNet and XGBoost area under the receiver operating characteristic curve and precision recall curves.

Discussion and Conclusions

In this study, we developed and evaluated models that predict whether patients with glaucoma will progress to require surgery, fusing multiple EHR and RNFL OCT imaging features and comparing XGBoost and TabNet model architectures. Models designed using a single modality of data, either EHR or RNFL, were compared against those trained using both data modalities as inputs. We found that performance improved when both the RNFL and EHR data were integrated into the TabNet and XGBoost models, compared with models using single modalities of data, which highlights the value of integrating multimodal data into prediction models for glaucoma. Moreover, the TabNet fusion model outperformed the conventional tree-based XGBoost fusion model, highlighting

Table 2. Model Performance Metrics

	TabNet	XGBoost
Fusion model		
AUROC	0.832	0.747
F1	0.587	0.469
Accuracy	0.764	0.728
Sensitivity, recall	0.724	0.517
Specificity	0.776	0.792
Positive predictive value, precision	0.494	0.429
Negative predictive value	0.903	0.844
EHR only model		
AUROC	0.764	0.720
F1	0.480	0.323
Accuracy	0.584	0.732
Sensitivity, recall	0.828	0.276
Specificity	0.510	0.870
Positive predictive value, precision	0.338	0.390
Negative predictive value	0.907	0.799
RNFL only model		
AUROC	0.624	0.633
F1	0.406	0.343
Accuracy	0.344	0.740
Sensitivity, recall	0.966	0.293
Specificity	0.156	0.838
Positive predictive value, precision	0.257	0.415
Negative predictive value	0.938	0.804

AUROC, area under the receiver operating characteristic curve; EHR, electronic health record; RNFL, retinal nerve fiber layer.

the promise of TabNet as a flexible deep learning architecture suitable for multiple modalities of healthcare data.

This study expands upon prior efforts in predicting progression to surgery in patients with glaucoma. Our previous models which leverage structured and free-text EHR data, achieved AUROC values ranging from approximately 0.70 to 0.90.[8–10] However, these models lacked integration of baseline optic nerve imaging data, which can provide crucial information on glaucoma severity and influence surgical decisions. Wang et al. attempted to bridge this gap by incorporating RNFL data alongside visual field and EHR data to forecast future surgeries among patients with glaucoma across different timelines in the future, achieving AUROCs ranging from 0.77 for long-term prediction to 0.85 for predictions within the 0.5 to 1 year timeframe.[21] Their tri-modality fusion approach involved a custom deep learning architecture, combining a vision transformer and a fully connected neural network. This method required complex and idiosyncratic preprocessing steps to convert the numerical results from imaging and visual field tests into resized and color-coded pixel arrays for input into the vision transformer. In contrast, TabNet offers a distinct advantage in its simplicity, as it can be applied directly and intuitively to diverse tabular datasets without requiring extensive customization or preprocessing. This enabled robust performance in predicting glaucoma progression to surgery, comparable to more complex fusion architectures, making TabNet a compelling choice for modeling with various healthcare data types structured in tabular formats.

In ophthalmology and across the broader healthcare domain, there have been relatively few prior studies using TabNet, although these have been promising. TabNet was among the architectures used to predict stroke mortality using EHR data in Hong Kong, achieving AUROC of 0.840 for predicting death by ischemic stroke.[22] Additionally, fusion models with EHR and extracted features from computed tomography (CT) data have been found to outperform single modality models in predicting pulmonary embolism mortality, demonstrating the potential of using multiple modalities of data with TabNet.[23] Our study is one of the first to explore the applicability of TabNet

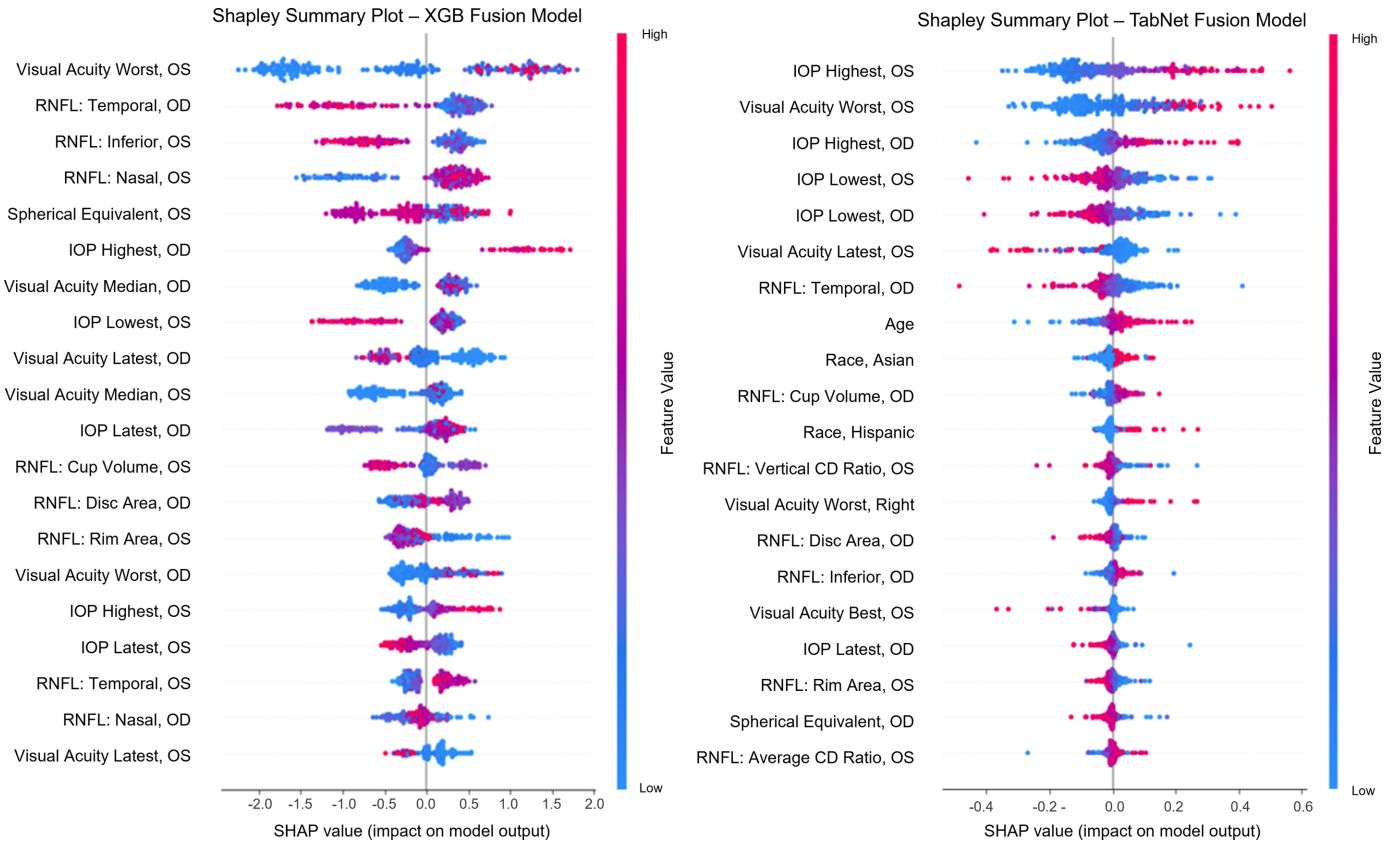

Figure 2. **Explainability evaluation using Shapley values for XGBoost and TabNet fusion models.** The figure depicts the Shapley values for predicting whether a patient with glaucoma will progress to require surgery. The top 20 most important features are displayed for the XGBoost and TabNet fusion models as calculated across the test set. Each point represents an individual observation (patient) in the test set. The color of each point signifies whether the feature's value was high or low for that particular patient. If a feature has a positive Shapley value for a patient, this indicates it influenced the model toward predicting surgery. Conversely, a negative Shapley value indicates influence toward a model prediction of no surgery. CD, cup-to-disc; IOP, intraocular pressure; OD, right eye; OS, left eye; RNFL, retinal nerve fiber layer denoting features that come from the retinal nerve fiber layer optical coherence scan results.

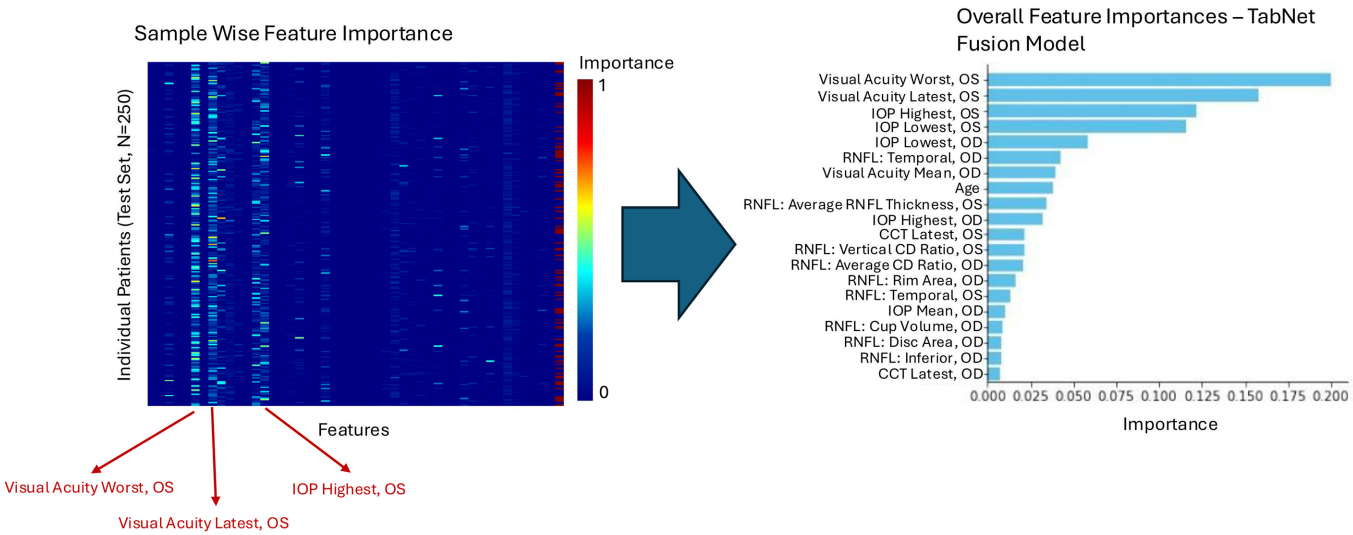

Figure 3. **Instance-wise feature importance and aggregate feature importance.** On the *left* is the instance-wise feature importance for the test set, with patients represented on the y-axis and features on the x-axis, and the color corresponding to the magnitude of feature importance for that particular patient. *Columns* with brighter colors correspond to features that are more globally important; the top three features important across all patients are labeled in *red*. On the *right*, is the resultant global importance for each feature. CCT, central corneal thickness; CD, cup-to-disc; IOP, intraocular pressure; OD, right eye; OS, left eye; RNFL, retinal nerve fiber layer.

in developing prediction models within the field of ophthalmology, and the first in glaucoma. A previous ophthalmic prediction model used TabNet to predict which patients may benefit from a corneal topographical scan based on ophthalmic examination information, demonstrating superior performance compared to XGBoost and a fully connected neural network in a Korean population.[24] Another study, which predicted the presence of sarcopenia based on eye examination information, showed no substantial differences among TabNet, XGBoost, and logistic regression models.[25] Taken together, these studies suggest a promising role for TabNet in ophthalmology, while also highlighting the need for ongoing investigation for how best to incorporate the diversity of medical data types into prediction models using TabNet. Our study particularly focuses on this question by using TabNet for the development of fusion models that integrate data from EHR alongside results from RNFL imaging studies, which are important for assessing the health of the optic nerve.

A strength of this study was our investigation into model explainability, using both model-agnostic approaches to compare between TabNet and XGBoost, as well as TabNet-specific approaches that give further insight into TabNet's attention-based feature importance. In general, many features which were important for our models, such as IOP, age, and visual acuity, were clinically reasonable features that would influence the clinicians' patient care decisions for glaucoma. In addition, many features from RNFL scans were also among the top most important features for model prediction, including global structural metrics of the nerve such as cup-to-disc ratio, cup volume, rim area, and disc area, as well as individual quadrant thicknesses. These features are fairly consistent with results of explainability studies on previous work, where IOP, visual acuity, rim area, and cup volume were highly important.[9,12,24] Shapley values offer a convenient model-agnostic way to ascertain feature importance, and can be used across different model architectures. The relative Shapley importance of RNFL features differed between XGBoost and TabNet models, but this may be expected as two independent models would not necessarily emphasize all of the same feature inputs to produce their predictions. Some prior studies have also suggested that Shapley explainability can sometimes be inaccurate and misleading, as it does not directly rely upon information encoded in the model structure itself, but merely computes explainability based on observed patterns of model inputs and outputs.[26,27] In our study, results from the Shapley feature importance analysis for TabNet did not exactly mirror the TabNet model-specific feature importance results; race/ethnicity as a feature was comparatively de-emphasized, whereas visual acuity and certain RNFL features were more important in the model-specific feature importance analysis. This ability for direct interpretability analyses sets TabNet apart from many other deep learning models. Moreover, TabNet's instance-wise feature selection aids efficient learning by fully utilizing model capacity for the most salient features, leading to an easily explainable decision-making process.

We acknowledge that this study also has several limitations. The models developed and validated in this investigation are based on a dataset from patients receiving care at a single academic center, which may reduce generalizability. Furthermore, the cohort was limited to those patients who did undergo RNFL OCT scans during their care, limiting the sample size. A limited cohort size also precludes model performance analyses in subgroups, such as by glaucoma subtype, which would be valuable information, as well as prediction over longer time horizons, requiring larger numbers of patients with longer periods of follow-up. Future studies can consider modeling using multi-institutional registries, such as the newly established Sight OUtcomes Research Collaborative (sourcecollaborative.org), after imaging results become integrated into this registry. Additionally, we recognize that the criteria for performing glaucoma surgery can vary among physicians due to differences in practice patterns, with some opting for earlier intervention whereas others may delay until later stages. This variability reflects the lack of universal standards and the personalized nature of glaucoma care. Incorporating larger and more diverse datasets in future work could aid in addressing this limitation by capturing wider variation in surgical practice patterns. Additionally, future work could also incorporate direct prediction of glaucoma-related findings, such as future RNFL or visual field progression, which are less dependent on surgical practice patterns. Another potential limitation is that the present models included only demographic and eye examination features from the EHR, and did not include medication or diagnosis data. In doing so, this study more heavily emphasizes the clinical measurements obtained from ophthalmic examinations and the structural features of the eye. Future work could explore the incorporation of other elements from the EHR, although medication and diagnosis features carry considerably more noise than documented eye examination measurements. In addition, future work could also incorporate results from visual field testing in TabNet fusion models. Although we acknowledge that our approach does not include raw image data derived from the OCT scans, such data are often proprietary and difficult

to obtain, store, and analyze, and their incorporation into models limits the ability to deploy such models because data ingestion requirements into them becomes more complex. We have demonstrated that a simpler approach using OCT imaging results stored in tabular form is still highly effective. Future research could explore different methods of image representation to better encapsulate the spatial information inherent in imaging scans to potentially augment performance.

In conclusion, we developed models that predict the patients with glaucoma progression to surgery using data from EHR and RNFL OCT scans, comparing TabNet and XGBoost modeling techniques. We found that models incorporating both EHR and RNFL data outperformed single-modality models. In addition, TabNet outperformed XGBoost, achieving the highest AUROC at 0.832. Our research highlights the simplicity and versatility of TabNet for data fusion models in healthcare, which may have broad applicability for researchers in the healthcare domain. Future research can investigate incorporating additional modalities, such as visual field test results. Such endeavors hold promise for enhancing predictive modeling and augmenting decision-making processes for patients with glaucoma.

Acknowledgments

Supported by the National Eye Institute (K23EY03263501; S.Y.W.); American Glaucoma Society Young Clinician Scientist Award (S.Y.W.), an unrestricted departmental grant from Research to Prevent Blindness (S.Y.W., A.K.); and a departmental grant from the National Eye Institute (P30-EY026877 [S.Y.W., A.K.]

Disclosure: **A. Koornwinder**, None; **Y. Zhang**, None; **R. Ravindranath**, None; **R.T. Chang**, None; **I.A. Bernstein**, None; **S.Y. Wang**, None

References

1. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006;90(3):262–267.
2. Susanna R, Jr, De Moraes CG, Cioffi GA, Ritch R. Why do people (still) go blind from glaucoma? *Transl Vis Sci Technol*. 2015;4(2):1.
3. Chauhan BC, Malik R, Shuba LM, Rafuse PE, Nicolela MT, Artes PH. Rates of glaucomatous visual field change in a large clinical population. *Invest Ophthalmol Vis Sci*. 2014;55(7):4135–4143.
4. Wang R, Bradley C, Herbert P, et al. Deep learning-based identification of eyes at risk for glaucoma surgery. *Sci Rep*. 2024;14(1):599.
5. Hussain S, Chua J, Wong D, et al. Predicting glaucoma progression using deep learning framework guided by generative algorithm. *Sci Rep*. 2023;13(1):19960.
6. Baxter SL, Marks C, Kuo T-T, Ohno-Machado L, Weinreb RN. Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records. *Am J Ophthalmol*. 2019;208:30–40.
7. Gonzalez R, Huynh J, Walker E, et al. Predicting surgical interventions for glaucoma with clinically available data. *Invest Ophthalmol Vis Sci*. 2023;64(8):387.
8. Jalamangala Shivananjaiah SK, Kumari S, Majid I, Wang SY. Predicting near-term glaucoma progression: an artificial intelligence approach using clinical free-text notes and data from electronic health records. *Front Med*. 2023;10:1157016.
9. Wang SY, Tseng B, Hernandez-Boussard T. Deep learning approaches for predicting glaucoma progression using electronic health records and natural language processing. *Ophthalmol Sci*. 2022;2(2):100127.
10. Hu W, Wang SY. Predicting glaucoma progression requiring surgery using clinical free-text notes and transfer learning with transformers. *Transl Vis Sci Technol*. 2022;11(3):37.
11. Wang SY, Ravindranath R, Stein JD, SOURCE Consortium. Prediction models for glaucoma in a multicenter electronic health records consortium: the sight outcomes research collaborative. *Ophthalmol Sci*. 2024;4(3):100445.
12. Wang P, Shen J, Chang R, et al. Machine Learning models for diagnosing glaucoma from retinal nerve fiber layer thickness maps. *Ophthalmol Glaucoma*. 2019;2(6):422–428.
13. Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. *Proc AAAI Conf Artificial Intelligence*. 2021;35(8):6679–6687.
14. American Medical Association. *CPT 2003: Professional Edition*. Chicago, IL: American Medical Association Press; 2002.
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
16. *Tabnet: PyTorch Implementation of TabNet Paper: https://arxiv.org/pdf/1908.07442.pdf*. n.d. Github. Accessed March 13, 2024, https://github.com/dreamquark-ai/tabnet.

17. entmax: The entmax mapping and its loss, a family of sparse softmax alternatives [Internet]. Github; [cited 2024 Mar 14]. Available from: https://github.com/deep-spin/entmax.
18. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *NIPS'17; Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017;4768–4777.
19. SHAP: A game theoretic approach to explain the output of any machine learning model [Internet]. Github; [cited 2024 Mar 13]. Available from: https://github.com/shap/shap.
20. Koornwinder A. multimodal-glaucoma-surgery-prediction [Internet]. Github; [cited 2024 Mar 15]. Available from: https://github.com/akoornwinder4/multimodal-glaucoma-surgery-prediction.
21. Wang R, Bradley C, Herbert P, et al. Deep learning-based identification of eyes at risk for glaucoma surgery. *Sci Rep*. 2024;14(1):599.
22. Huang R, Liu J, Wan TK, et al. Stroke mortality prediction based on ensemble learning and the combination of structured and textual data. *Comput Biol Med*. 2023;155:106176.
23. Cahan N, Klang E, Marom EM, et al. Multimodal fusion models for pulmonary embolism mortality prediction. *Sci Rep*. 2023;13(1):7544.
24. Ahn H, Kim NE, Chung JL, et al. Patient selection for corneal topographic evaluation of keratoconus: a screening approach using artificial intelligence. *Front Med*. 2022;9:934865.
25. Kim BR, Yoo TK, Kim HK, et al. Oculomics for sarcopenia prediction: a machine learning approach toward predictive, preventive, and personalized medicine. *EPMA J*. 2022;13(3):367–382.
26. Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. In: Iii HD, Singh A, eds. *Proceedings of Machine Learning Research*. Cambridge, MA: PMLR: 2020;119:5491–5500. (Proceedings of the 37th International Conference on Machine Learning). [Google Scholar].
27. Huang X, Marques-Silva J. The inadequacy of Shapley values for explainability. *arXiv [csLG]* 2023, http://arxiv.org/abs/2302.08160. Accessed March 13, 2024.