Deep Learning Models to Predict Primary Open-Angle Glaucoma Using Longitudinal Visual Field Measurements

Anonymous Authors[1]

Abstract

Glaucoma is a major cause of blindness and vision impairment worldwide and visual field (VF) tests are essential for monitoring the conversion of glaucoma. Existing research often uses VF data at a single time point to predict glaucoma; few explored the longitudinal trajectories. Additionally, many deep learning techniques treat the time-to-glaucoma prediction as a binary classification problem (glaucoma Yes/No), resulting in the misclassification of some censored subjects into the non-glaucoma category and decreased power. To tackle these challenges, we propose and apply several deep-learning approaches that naturally incorporate temporal and spatial information in longitudinal visual field data and predict time-to-glaucoma. The proposed methods' prediction performance is validated on the large Ocular Hypertension Treatment Study (OHTS) dataset. Extensive experiments show that the proposed LSTM and Bi-LSTM have better prediction performance than the traditional Cox proportional hazards model, ResNet50-LSTM, and CNN-LSTM methods.

1. Introduction

Primary open-angle glaucoma (POAG) is an irreversible optic neuropathy associated with glaucomatous damage by the progressive loss of retinal ganglion cells (RGCs), which leads to ultimate structural changes that cause visual field (VF) defects. This chronic disease is one of the leading causes of blindness worldwide. Therefore, monitoring visual field examination and making predictions of time-to-glaucoma are vital to prevent disease progression and irreversible vision loss. The visual field data are collected from the 24-2 Humphrey Visual Field map shown in Figure 1 (Feldon et al. 2006). Among 54 test points on the Humphrey Visual Field map, two blind points are excluded from the analysis, resulting in 52 points for measurement. The total deviation (TD) at each location is the difference of the test result (in dB) from that of a "normal" patient's field of the same age (Feldon et al. 2006), while a large TD indicates a poor eye condition. During a clinical visit, the TD of these 52 points are recorded for both eyes, which could be used for the prediction of glaucoma.

However, early diagnosis with longitudinal glaucomatous VF can be difficult since it contains random errors and fluctuates over time which vary between patients and locations. Therefore, using VF observation only at baseline or a fixed time point may fail to capture disease progression trends, leading to poor prediction performance on time-to-conversion. To overcome this barrier, traditional statistical methods (e.g., linear and exponential regression models proposed by McNaught et al. (1995) and Caprioli et al. (2011) respectively), machine learning methods (e.g., variational Bayes proposed by Murata et al. (2014)), and deep learning methods (e.g., recurrent neural network (RNN) (Park, Kim, and Lee, 2019)) were used to capture such variation in the longitudinal point-wise VF data. However, those methods cannot be directly used for disease diagnosis or time-to-glaucoma prediction. They may be further extended for diagnosis purposes with the help of classification methods. For example, Asaoka et al. (2016) and Kucur et al. (2018) proposed deep-learning methods to classify patients into early glaucoma or normal group. It is worth noting that converting the time-to-event prediction into classification can be less powerful and biased for disease diagnosis, because classification techniques do not take censoring into account and misclassify some right-censored subjects to a none-event group.

To predict the time-to-glaucoma using longitudinal VF data, we propose to combine landmark analysis with artificial intelligence methods, including long-short-term-memory (LSTM, Graves 2012), bidirectional long-short-term-memory (Bi-LSTM, Schuster 1997), convolutional neural network (CNN)-LSTM (Kucur et al. 2018), and

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ResNet50-LSTM (Gheisari et al. 2021) to capture the temporal trends and small localized VF defects. Of note, these methods were originally introduced in the classification framework, while we are among the first to adapt them to the landmark analysis for a survival outcome.

While the use of longitudinal biomarkers for survival prediction has been extensively studied by statisticians, deep learning methods have also shown great success in identifying temporal dependencies in repeated measurements. However, there is a scarcity of literature that merges the traditional landmark model with deep learning methods for time-to-event prediction using longitudinal biomarkers. Our contribution is to introduce a framework that seamlessly combines deep learning techniques (e.g., LSTM, Bi-LSTM, CNN-LSTM) with traditional landmark analysis to improve the performance of time-to-event prediction. Additionally, we provide implementation code in which these deep learning methods are optimized using the negative partial likelihood function loss function specific to landmark analysis. Our deep learning approaches are presented in section 2. We then compare the prediction performance of our proposed deep learning structures on four prespecified landmark time points using the Ocular Hypertension Treatment Study (OHTS) dataset. Estimation results are given in section 3. Some discussion and concluding remarks are given in section 4.

2. Model Specification

2.1. Data Sources

In this study, we analyze the data from the Ocular Hypertension Treatment Study (OHTS) - a randomized trial testing whether topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma (POAG; Gordon et al. 1999, 2020, Kass et al. 2002, 2021). OHTS randomized 1,636 participants in 22 clinics nationwide, 25% of whom self-identified as African origin, to either treatment with topical ocular hypotensive mediation or close observation. In OHTS, measures of variables were recorded every 6 months for a median follow-up of 11 years and reassessed after 20 years, using the same schedule of tests and measures in both randomization arms.

For all the methods below, we use the visual field (VF) data collected from the 24-2 Humphrey VF map. In this map, 52 total deviation (TD) values in a plane are assessed quantitatively on their threshold sensitivity to spots of light, which are stretched into a 52×1 vector at each observation time. Figure 1 shows one example of the 24-2 Humphrey VF map (Feldon et al. 2006).

2.2. Model Development

Landmark Analysis

Landmark analysis is widely applied for survival prediction with longitudinal covariates or features, where a survival model is fitted from a series of time origins (landmarks) only on those patients still at risk at the landmark time (Anderson et al. 1983). The risk function is estimated based on features measured up to the landmark time. All methods considered in this paper are within the landmark analysis framework.

Cox's proportional hazards model

Under the Cox's proportional hazards model, let s denote the prespecified landmark time, and v denote a clinically relevant prediction time period, the hazard for subject i conditional on predictors from time s up to time $s+v$ can be obtained by

$$h_i\left(t \mid X_i, \mathcal{Y}_i(s), s, v\right) = h_{0,s}(t \mid s) \exp\left\{\alpha Y_i(s)\right\},$$

where $t \in (s, s+v)$ and $Y_i(s)$ is longitudinal VF observations of length 52 (for the 52 VF locations) of subject i at time s, $\mathcal{Y}_i(s)$ is the longitudinal VF observation history up to time s, and $h_{0,s}(t \mid s)$ denotes the baseline hazard function at time t for landmark time s.

LSTM Architecture (Hochreiter and Schmidhuber (1996))

Suppose that the number of observations before a given landmark time s is l, then the VF inputs feeding into the LSTM for each subject is l vectors of length 52. The structure of LSTM is shown in Figure $2(A)$. Stretched inputs are fed into two layers of LSTM, followed by a dropout layer and another two layers of LSTM. The outputs from previous layers of LSTM are then used as inputs for a feed-forward neural network layer with a Sigmoid activation function to calculate the final risk score prediction. The output for subject i at landmark time s is denoted as $r_\theta(Y_i(s))$, where θ denotes the unknown parameters in the proposed LSTM structure. With the predicted risk scores, for $t \in (s, s+v)$, the hazard of developing glaucoma at time t for subject i under the Cox proportional hazards model is defined as

$$h_i\left(t \mid X_i, \mathcal{Y}_i(s), s, v\right) = h_{0,s}(t \mid s) \exp\left\{r_\theta(Y_i(s))\right\}.$$

Bidirectional LSTM Architecture (Schuster 1997)

LSTM only preserves the information in one direction, from past to future or from future to past. However, bidirectional LSTM (Bi-LSTM) has the inputs flowing in two ways to preserve future and past information, which is usually more suitable for complex structures. The proposed bidirectional LSTM structure for landmark analysis is similar to the proposed LSTM structure introduced above. For each subject at every observational time point, VF data are stretched into a 52×1 vector. If the number of VF observations before a given landmark time s is l, then the inputs of dimension

$52 \times l$ are fed into the bidirectional LSTM. The structure of Bi-LSTM is shown in Figure $2(B)$, with two layers of bidirectional LSTM followed by a dropout layer, and then the hidden nodes are fed into two bidirectional LSTM layers followed by a feed-forward neural network layer with a Sigmoid activation to make the final risk score predictions.

CNN-LSTM Architecture (Kucur et al. 2018)

CNNs require input images with fixed height, width, and depth. However, visual field maps cannot be used directly as inputs for CNNs because they lack a standard representation for perimetric data points. To maintain the spatial information in the visual field map and convert it into a suitable input format for a CNN, in the beginning, the 52 visual field data points are converted into a 61×61 image. Following Kucur, Hollo, and Sznitman (2018), we convert the visual field map into a new image using a novel Voronoi representation and then train the transformed Voronoi images in the designed CNN-LSTM algorithm. For a subject with l number of visual field examinations before landmark time s, the converted images of dimension $61 \times 61 \times l$ first pass through the CNN. The structure of the CNN-LSTM is shown in Figure $2(C)$. After two CNN layers with two max-pooling in the middle of the two layers, the output is flattened into a vector followed by a fully connected layer. The visual field map features are condensed into vectors of length 52, representing a summary of information at 52 locations on the map, and then used as inputs for the subsequent LSTM. The LSTM architecture is the same as in Figure $2(A)$. The final outputs from the CNN-LSTM architecture are risk predictions for each subject at a specified landmark time.

ResNet50-LSTM Architecture (Gheisari et al. 2021)

ResNet-50 is a 50-layer CNN with 48 convolutional layers, one max-pooling layer, and one average pooling layer. Similar to CNN-LSTM, in applying the ResNet50-LSTM architecture proposed by Gheisari et al. (2021), we also use transformed Voronoi images (Kucur, Hollo, and Sznitman, 2018) as input to ResNet50 to extract spatial features. The outputs of ResNet50 are vectors of length 52. The output length is set to be 52 to be consistent with other methods. Then the outputs from ResNet50 are fed into LSTM, which shares the same structure as the LSTM method shown in Figure $2(A)$. The detailed structure of ResNet50-LSTM is shown in Figure $2(D)$. In the beginning, ResNet50 is applied to extract spatial information from each visual field Voronoi image. Then the longitudinal extracted features are fed into the LSTM accounting for time-dependency to make a prediction on the risk scores.

Of note, the CNN-LSTM method trains the convolutional neural network and LSTM simultaneously, while ResNet50-LSTM trains the convolutional neural network and LSTM in completely separate phases.

Loss Function

The loss function is defined as the negative partial log-likelihood:

$$l(\theta) := -\sum_{i:\delta_i=1}\left(\hat{r}_\theta\left(Y_i(t)\right) - \log \sum_{j\in\Re(T_i)} e^{\hat{r}_\theta(Y_j(t))}\right)$$

where T_i is the true failure time, C_i is the censoring time, and $V_i = min(T_i, C_i)$ denotes the observed event time for the i^{th} subject. The censoring indicator δ_i is defined as $\delta_i = I_{(T_i<C_i)}$. The risk set $\Re(t) = \{i : T_i \geq t\}$ is the set of patients still at risk of failure at time t.

3. Experimental settings

We compare the five methods on their time-to-glaucoma prediction with the Ocular Hypertension Treatment Study (OHTS) dataset, including conventional landmark analysis under Cox proportional hazards model (CPH), the LSTM method, the Bi-LSTM method, the CNN-LSTM method, and the ResNet50-LSTM method. The conventional landmark analysis under Cox proportional hazards model is regarded as the baseline method. The proposed LSTM, Bi-LSTM, CNN-LSTM, and ResNet50-LSTM are implemented by Keras and Tensorflow.

We investigate the prediction performance at four landmark time points: 1.5 years, 2 years, 2.5 years, and 3 years. Subjects in the analysis were expected to examine every 6 month from baseline, which results in 4, 5, 6 and 7 number of observations (including baseline) for the four landmark times, respectively. The number of glaucoma events for each landmark time point 1.5 years, 2 years, 2.5 years, and 3 years are 301, 286, 262, and 248, respectively. It is desirable to find an early landmark time point with adequate prediction power for clinical practice. In the application, we exclude subjects who have missing observations at the first three observation times for reliable estimation. The sample used for training and testing is 1436 subjects with 2872 eyes. For subjects with intermittent missing VF maps before the landmark time, we use the last observation carried forward (LOCF) method to impute those intermediate missing VF data.

For the LSTM method and bidirectional LSTM method, each visual field map is transformed to a length 52 vector. For ResNet50-LSTM, we use the transformed Voroni images with size 61×61 as inputs for ResNet50, the extracted features for each patient are of length 52×1 and are then fed into LSTM for training and prediction with a batch size

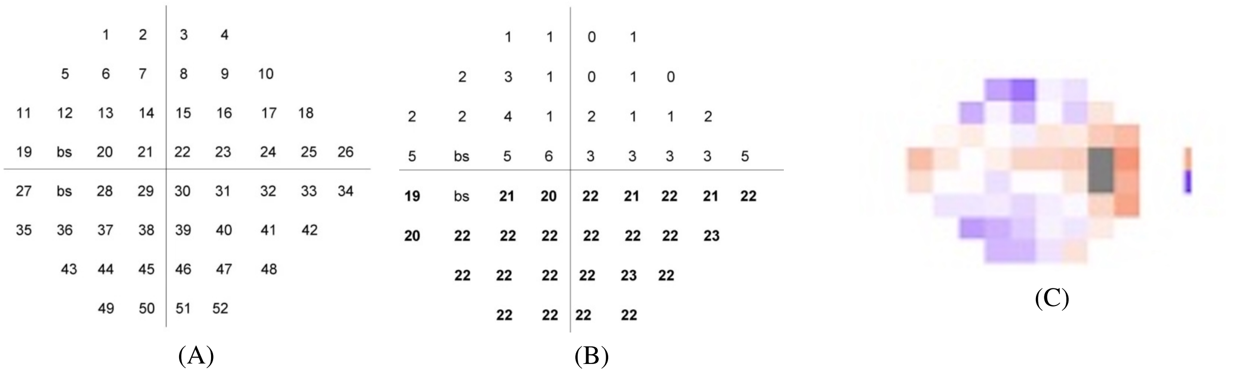

Figure 1. (A). Schematic of a 52-point (program 24-2) Humphrey Visual Field for a left eye (bs: blind spot). The numbers on the plot indicate location indexes in a visual field map. (B). Example of a Humphrey Visual Field with an inferior altitudinal field defect. The numbers are the difference between the values in decibels of each point in the linear array between a single visual field and those of age-matched controls. Bold points indicate locations with inferior altitudinal defects. (C).Example of heatplot of a visual field map with TD values at 52 locations. Purple means low values, and red means high values.

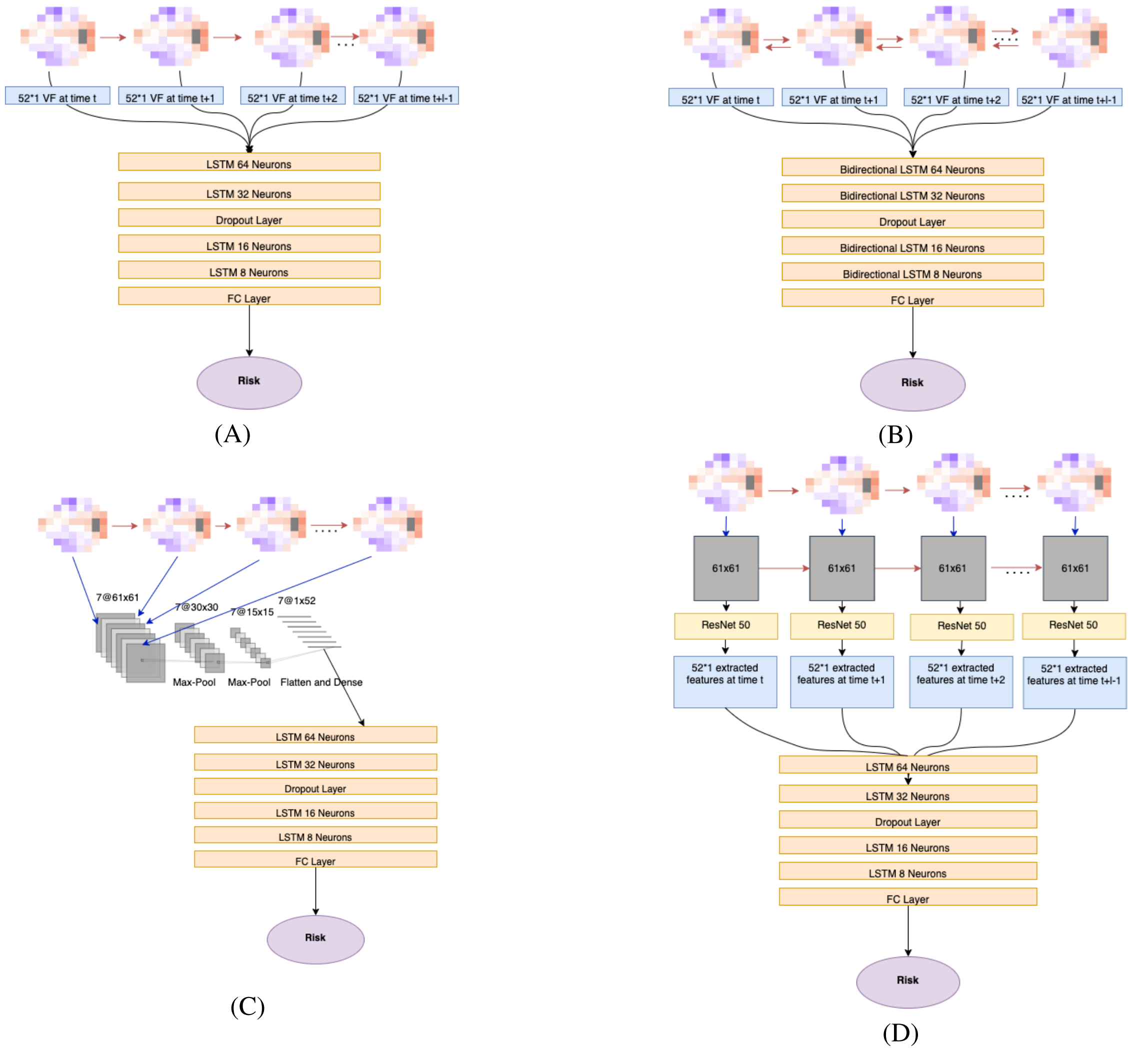

Figure 2. (A). LSTM with l number of longitudinal VF observations. (B). Bi-LSTM with l number of longitudinal VF observations. (C). CNN-LSTM with $l = 7$ number of longitudinal VF observations. (D). ResNet50-LSTM with l number of longitudinal VF observations. Red arrows represent time-dependence among VF maps. Blue arrows represent Voronoi transformation.

Table 1. The averaged C-index on the 10 cross-validation sets as the predictive performance by the five methods.

Landmark time	CPH	LSTM	Bi-LSTM	ResNet50-LSTM	CNN-LSTM
1.5 years	0.597	0.657	0.657	0.523	0.579
2 years	0.559	0.642	0.644	0.516	0.520
2.5 years	0.559	0.657	0.666	0.558	0.656
3 years	0.568	0.697	0.704	0.613	0.642

64. As for the CNN-LSTM method, similarly, visual field maps with 52 points are resized to $61 \times 61 \times l$ Voronoi images, where l is the number of longitudinal observation times at each landmark time. For example, $l = 7$ when the landmark time point is 3 years. The transformed Voronoi images are fed into the convolutional layer (kernel size 3×3, hidden size 8) and a max-polling layer (pool size 2) twice to extract image features. Then the extracted features are flattened and fed into a fully connected layer with the Relu activation function (size 52). During training, we use the Adam optimizer with a learning rate $= 10^{-5}$ and a batch size of 64.

4. Results

We report the concordance index (C-index) to compare the prediction performance of different methods. C-index is the most commonly used evaluation metric in survival analysis (Harrell et al. 1984) as a measure of the rank correlation between predicted risk scores and observed time points. If C-index$= 1$, the prediction model has the best performance in the sense that the ranking of predicted risk scores perfectly matches that of the observed event times.

We use ten-fold cross-validation to train the model. In each fold, to mimic the patients' event distribution in the original dataset, we use stratified sampling on the OHTS data with the same ratio of censored and uncensored patients in each sample and divide the data by ten folds. C-indexes shown in Table 1 are those averaged on each of 10 cross-validation folds.

Overall, the LSTM and bidirectional LSTM methods have the highest C-indexes among the five methods on all landmark time points. Lower C-indexes from the ResNet50-LSTM and CNN-LSTM methods suggest that it might be unnecessary to convert the VF data to the Voronoi image for the prediction purpose. Furthermore, it appears that as the duration of landmark time increases (more VF data available), the accuracy of predictions tend to improve. According to our analysis using the LSTM and Bi-LSTM methods, a three-year VF data set can achieve an AUC of 0.70.

5. Conclusion

We employed a variety of deep-learning techniques, including LSTM, Bi-LSTM, CNN-LSTM, and ResNet50-LSTM, to analyze longitudinal visual field data with the goal of predicting the empirical distribution of the future event time for subjects who had not yet experienced the event or censoring prior to the landmark time. The proposed structures are intended to capture both spatial and temporal information and predict the survival probability in the presence of right-censored data. For landmark analysis of time-to-POAG prediction with longitudinal VF observations, the proposed LSTM and Bi-LSTM demonstrate better prediction performance than the traditional Cox proportional hazards model, ResNet50-LSTM, and CNN-LSTM methods. Our results illustrate the potential benefits of using deep learning methods in time-to-POAG prediction with longitudinal features.

All methods used in the experiment are within the landmark analysis framework. An alternative approach is the joint modeling framework, which models both the longitudinal measures and the survival outcomes jointly, utilizing shared random effects to account for correlation (Rizopoulos et al. 2017, Suresh et al. 2017, Tanner et al. 2021, Lin and Luo 2022). It would be interesting to extend the proposed deep learning structures into the joint modeling framework and perform time-to-event predictions.

Besides visual field data, the OHTS dataset also contains other covariates such as age, intraocular pressure, central corneal thickness, pattern standard deviation, and vertical cup disc ratio, which are informative on time-to-glaucoma prediction and might be worthy of investigation in the future study to improve prediction performance.

In this paper, we treat each eye independently. To account for dependency between two eyes on the same patient, we recommend to use frailty method for cluster effect (Paik, Tsai, and Ottman (1994), Ripatti and Palmgren (2000)).

References

Asaoka, R., Murata, H., Iwase, A., and Araie, M. (2016). Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. *Ophthalmology*, 123(9), 1974-1980.

Anderson, J. R., Cain, K. C., and Gelber, R. D. (1983). Analysis of survival by tumor response. *J Clin Oncol*, 1(11), 710-719.

Caprioli, J., Mock, D., Bitrian, E., Afifi, A. A., Yu, F., Nouri-Mahdavi, K., and Coleman, A. L. (2011). A method to measure and predict rates of regional visual field decay in glaucoma. *Investigative ophthalmology & visual science*, 52(7), 4765-4773.

Feldon, S. E., Levin, L., Scherer, R. W., Arnold, A., Chung, S. M., Johnson, L. N., ... and Dickersin, K. (2006). Development and validation of a computerized expert system for evaluation of automated visual fields from the Ischemic Optic Neuropathy Decompression Trial. BMC ophthalmology, 6(1), 1-21.

Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.

Gordon, M. O., Gao, F., Huecker, J. B., et al. (2020). Evaluation of a primary open-angle glaucoma prediction model using long-term intraocular pressure variability data: a secondary analysis of 2 randomized clinical trials. *JAMA ophthalmology*, 138(7), 780-788.

Gordon MO, Kass MA, for the Ocular Hypertension Treatment Study Group. (1999). The ocular hypertension treatment study: design and baseline description of the participants. *Archives of Ophthalmology*. 117, 573-583.

Gleiss, A., Oberbauer, R., and Heinze, G. (2018). An unjustified benefit: immortal time bias in the analysis of time-dependent events. *Transplant international*, 31(2), 125-130.

Gheisari, S., Shariflou, S., Phu, J., Kennedy, P. J., Agar, A., Kalloniatis, M., and Golzan, S. M. (2021). A combined convolutional and recurrent neural network for enhanced glaucoma detection. *Scientific reports*, 11(1), 1-11.

Harrell Jr, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2), 143-152.

Hochreiter, S., and Schmidhuber, J. (1996). LSTM can solve hard long time lag problems. *Advances in neural information processing systems*, 9.

Kass, M. A., Heuer, D. K., Higginbotham, E. J., Johnson, C. A., Keltner, J. L., Miller, J. P., ... and Ocular Hypertension Treatment Study Group. (2002). The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Archives of ophthalmology*, 120(6), 701-713.

Kass, M. A., Heuer, D. K., Higginbotham, E. J., Parrish, R. K., Khanna, C. L., Brandt, J. D., ... and Ocular Hypertension Study Group. (2021). Assessment of cumulative incidence and severity of primary open-angle glaucoma among participants in the ocular hypertension treatment study after 20 years of follow-up. *JAMA ophthalmology*, 139(5), 558-566.

Kucur, ş. S., Hollo, G., and Sznitman, R. (2018). A deep learning approach to automatic detection of early glaucoma from visual fields. *PloS one*, 13(11), e0206081.

Kass, M. A., Gordon, M. O., Gao, F., Heuer, D. K., Higginbotham, E. J., Johnson, C. A., ... and Wilson, M. R. (2010). Delaying treatment of ocular hypertension: the ocular hypertension treatment study. *Archives of ophthalmology*, 128(3), 276.

Lin, J., and Luo, S. (2022). Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in Medicine.* 41(15), 2894-2907.

Murtagh, F., and Heck, A. (2012). Multivariate data analysis (Vol. 131). *Springer Science & Business Media*.

McNaught, A. I., Hitchings, R. A., Crabb, D. P., and Fitzke, F. W. (1995). Modelling series of visual fields to detect progression in normal-tension glaucoma. *Graefe's archive for clinical and experimental ophthalmology*, 233(12), 750-755.

Murata, H., Araie, M., and Asaoka, R. (2014). A new approach to measure visual field progression in glaucoma patients using variational Bayes linear regression. *Investigative ophthalmology and visual science*, 55(12), 8386-8392.

Park, K., Kim, J., and Lee, J. (2019). Visual field prediction using recurrent neural network. *Scientific reports*, 9(1), 1-12.

Paik, M. C., Tsai, W. Y., and Ottman, R. (1994). Multivariate survival analysis using piecewise gamma frailty. *Biometrics*, 975-988.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6), 1261-1276.

Ripatti, S., and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4), 1016-1022.

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673-2681.

Suresh, K., Taylor, J. M., Spratt, D. E., Daignault, S., and Tsodikov, A. (2017). Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59(6), 1277-1300.

Tanner, K. T., Sharples, L. D., Daniel, R. M., and Keogh, R. H. (2021). Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 184(1), 3-30.