Predicting Glaucoma Visual Field Progression: Performance Evaluation of Machine Learning Algorithms

by

Runjie (Bill) Shi
Supervisor: Prof. Moshe Eizenman

April 2019

Predicting Glaucoma Visual Field Progression: Performance Evaluation of Machine Learning Algorithms

by

Runjie (Bill) Shi
Supervisor: Prof. Moshe Eizenman
April 2019

A thesis submitted in conformity with the requirements
for the degree of Bachelor of Applied Science in Engineering Science
Undergraduate Division of Engineering Science
University of Toronto

Abstract

Predicting Glaucoma Visual Field Progression: Performance Evaluation of Machine Learning Algorithms

Runjie (Bill) Shi
Bachelor of Applied Science in Engineering Science
Undergraduate Division of Engineering Science
University of Toronto
2019

Glaucoma progression prediction is crucial for making clinical decisions. Visual field testing is one of the most important tools to functionally detect vision loss due to glaucoma. Current standard prediction method (ordinary least-squares linear regression (OLSLR) on Mean Deviation (MD)) requires at least 5–6 tested visual fields, or approximately 2 years assuming a semi-annual follow-up interval, to obtain a good estimate of the progression rate. This thesis investigates if machine learning methods can produce better predictions of (1) future MD and (2) point-wise field thresholds using the Rotterdam dataset ($N = 249$ eyes). All models, including linear models, convolutional neural network (CNN), and recurrent neural network (RNN), performed similarly on this dataset and performed better than OLSLR-based extrapolation in terms of mean absolute error (MAE). However, the performance of these algorithms is similar to simply assuming no change in the field because the dataset consists primarily of stable patients. Future studies should be based on a more comprehensive dataset by including older patients and/or different glaucoma types that are more likely to progress.

Acknowledgements

I would like to acknowledge first and foremost my supervisor Professor Eizenman who has now supervised my work for three years. As I embark on my post-graduate studies and future career, I will never forget the impact of his mentoring on my training to become a researcher. For this specific project, I would like to thank my colleagues, especially Yan (Leo) Li, for their support, teaching me various aspects of machine learning, and the thoughtful discussions that we had together. I want to thank all of my friends in the Engineering Science program, without whose support during the difficult times I would not be here today. Last but not least, I want to thank my family for their support in the last five years of my undergraduate studies that have allowed me to pursuit my dream of becoming a successful researcher. I will always be grateful for their unreserved support in the past, present, and future of my life.

Contents

List of Tables

List of Figures

List of Abbreviations

CGS	Canadian Glaucoma Study
CNN	convolutional neural network
DLS	differential light sensitivity
EMGT	Early Manifest Glaucoma Trial
GHT	Glaucoma Hemifield Test
HFA	Humphrey Field Analyzer
IOP	intraocular pressure
LSTM	long short-term memory
MAE	mean absolute error
MD	Mean Deviation
MLP	multi-layer perceptron
ML	machine learning
MSE	mean squared error
OCT	optical coherence tomography
OLSLR	ordinary least-squares linear regression
PSD	Pattern Standard Deviation
RL	reinforcement learning
RNFL	retinal nerve fibre layer

RNN	recurrent neural network
VF	visual field
VFI	Visual Field Index

Chapter 1

Introduction and Literature Review

Glaucoma is a group of progressive optic neuropathies where vision loss results from slow progressive degeneration of retinal ganglion cells and their axons. [1] Patients are typically elderly. The loss of vision is irreversible. As a result, it is a leading cause of blindness worldwide.

If glaucoma is detected and monitored in its early stages, it is treatable and can be reasonably well managed. This detection and monitoring rely upon imaging, and psychophysical tests. A typical procedure includes examination of the optic disk, angle closure (gonioscopy), retinal nerve fibre layer (RNFL) with optical coherence tomography (OCT), measurement of intraocular pressure (IOP), and visual field test.

Currently, the pathophysiology of glaucoma is not well understood and there are no models that can robustly characterize glaucoma progression [2]. Meanwhile, glaucoma treatment relies heavily on accurate and timely prediction of the progress of the disease. Patients with slowly progressing glaucoma might only require active surveillance, while fast progressors would require immediate intervention. Since glaucoma can eventually cause blindness, an accurate prediction of the progress of the disease will support optimal treatment decisions that are critical for the patient's quality of life.

1.1 Background: Visual Field Test

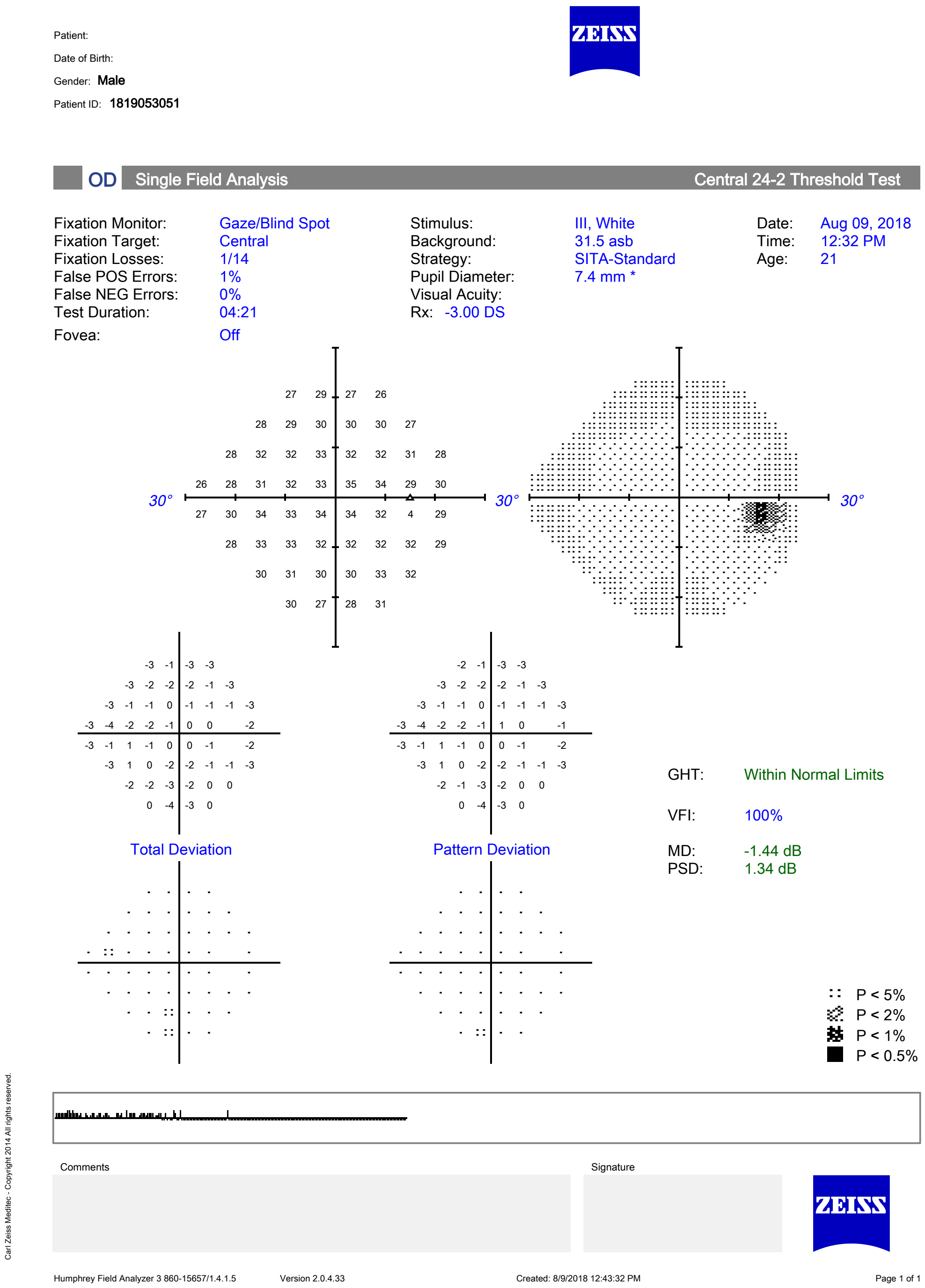
The primary focus of this thesis is on the visual field test. Visual field testing is the current gold standard in clinical functional evaluation of glaucoma patients. It is used to both diagnose glaucoma, and in patients with confirmed or suspected glaucoma, to monitor the progression of the disease. Global visual field indices such as Mean Deviation (MD) (also known as Mean Deviation Index), Pattern Standard Deviation (PSD), and Visual Field Index (VFI) (also known as Glaucoma Progression Index) are used to monitor

Figure 1.1: The Humphrey Field Analyzer (HFA) is one of the most popular visual field examiner used clinically ("Humphrey VF" by Sej licensed under CC BY-SA 4.0)

the integrity of the visual field. These indices characterize each visual field with a few statistical measures that attempt to capture the relevant clinical information to diagnose glaucoma and determine the progression of the disease.

1.1.1 Single Field Analysis Metrics

Historically Heijl et al. developed the two classic metrics for summarizing a visual field: MD and PSD. [3] First and foremost, to statistically analyzer one's visual field, it is necessary to know the appropriate reference threshold (N) with the associated variance (σ^2) at each field location. It is widely accepted that the human differential light sensitivity (DLS) thresholds, in units of dB, can be modeled as a linear function of age. [4] In general, the slope of loss of sensitivity is larger in mid-periphery than para-centrally, ranging from 0.5 to 1 dB/decade. The variance is also higher in mid-periphery than para-centrally.



Patient:
Date of Birth:
Gender: Male
Patient ID: 1819053051

OD Single Field Analysis — Central 24-2 Threshold Test

Fixation Monitor: Gaze/Blind Spot
Fixation Target: Central
Fixation Losses: 1/14
False POS Errors: 1%
False NEG Errors: 0%
Test Duration: 04:21
Fovea: Off

Stimulus: III, White
Background: 31.5 asb
Strategy: SITA-Standard
Pupil Diameter: 7.4 mm *
Visual Acuity:
Rx: -3.00 DS

Date: Aug 09, 2018
Time: 12:32 PM
Age: 21

```
                27  29 | 27  26
            28  29  30 | 30  30  27
        28  32  32  33 | 32  32  31  28
    26  28  31  32  33 | 35  34  29  30
30° ---------------------------------------- 30°
    27  30  34  33  34 | 34  32   4  29
        28  33  33  32 | 32  32  32  29
            30  31  30 | 30  33  32
                30  27 | 28  31
```

```
            -3  -1 | -3  -3
        -3  -2  -2 | -2  -1  -3
    -3  -1  -1   0 | -1  -1  -1  -3
-3  -4  -2  -2  -1 |  0   0      -2
-------------------------------------
-3  -1   1  -1   0 |  0  -1      -2
    -3   1   0  -2 | -2  -1  -1  -3
        -2  -2  -3 | -2   0   0
             0  -4 | -3   0
```

Total Deviation

```
            -2  -1 | -3  -3
        -3  -2  -2 | -2  -1  -3
    -3  -1  -1   0 | -1  -1  -1  -3
-3  -4  -2  -2  -1 |  1   0      -1
-------------------------------------
-3  -1   1  -1   0 |  0  -1      -2
    -3   1   0  -2 | -2  -1  -1  -3
        -2  -1  -3 | -2   0   0
             0  -4 | -3   0
```

Pattern Deviation

GHT: Within Normal Limits

VFI: 100%

MD: -1.44 dB
PSD: 1.34 dB

∷ P < 5%
▨ P < 2%
▩ P < 1%
■ P < 0.5%

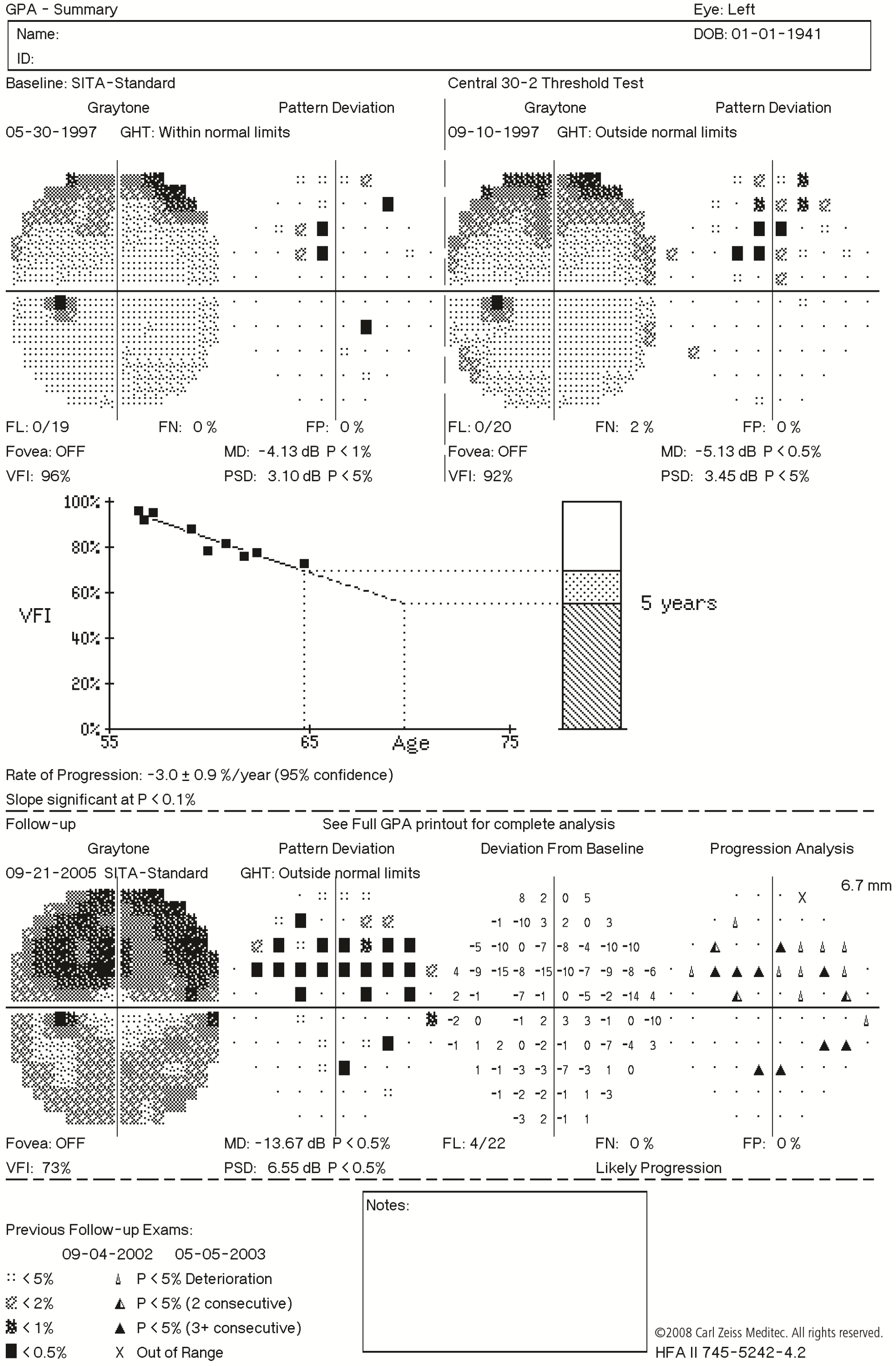
Comments

Signature

Humphrey Field Analyzer 3 860-15657/1.4.1.5 Version 2.0.4.33 Created: 8/9/2018 12:43:32 PM Page 1 of 1

Carl Zeiss Meditec - Copyright 2014 All rights reserved.

Figure 1.2: Example of an HFA visual field test report



GPA - Summary Eye: Left

Name: DOB: 01-01-1941

ID:

Baseline: SITA-Standard Central 30-2 Threshold Test

Graytone Pattern Deviation | Graytone Pattern Deviation

05-30-1997 GHT: Within normal limits | 09-10-1997 GHT: Outside normal limits

FL: 0/19 FN: 0 % FP: 0 % | FL: 0/20 FN: 2 % FP: 0 %

Fovea: OFF MD: -4.13 dB P < 1% | Fovea: OFF MD: -5.13 dB P < 0.5%

VFI: 96% PSD: 3.10 dB P < 5% | VFI: 92% PSD: 3.45 dB P < 5%

Rate of Progression: -3.0 ± 0.9 %/year (95% confidence)

Slope significant at P < 0.1%

Follow-up See Full GPA printout for complete analysis

Graytone Pattern Deviation Deviation From Baseline Progression Analysis

09-21-2005 SITA-Standard GHT: Outside normal limits 6.7 mm

```
                  8   2 | 0   5
             -1 -10   3 | 2   0   3
         -5 -10   0  -7 |-8  -4 -10 -10
     4   -9 -15  -8 -15 |-10 -7  -9  -8  -6
     2   -1      -7  -1 | 0  -5  -2 -14   4
    -2    0      -1   2 | 3   3  -1   0 -10
    -1    1   2   0  -2 |-1   0  -7  -4   3
          1  -1  -3  -3 |-7  -3   1   0
             -1  -2  -2 |-1   1  -3
                 -3   2 |-1   1
```

Fovea: OFF MD: -13.67 dB P < 0.5% FL: 4/22 FN: 0 % FP: 0 %

VFI: 73% PSD: 6.55 dB P < 0.5% Likely Progression

Notes:

Previous Follow-up Exams:

09-04-2002 05-05-2003

∷ < 5%	△ P < 5% Deterioration
▨ < 2%	◭ P < 5% (2 consecutive)
▩ < 1%	▲ P < 5% (3+ consecutive)
■ < 0.5%	X Out of Range

HFA II 745-5242-4.2

Figure 1.3: Example of an HFA progression analysis test report. The ordinary least-squares linear regression (OLSLR) trend-based analysis can be found in the middle section of the report. (Image from Carl Zeiss Meditec, Inc.)

MD is a measure of a visual field's general mean value as compared to the reference. Mathematically, it is the mean of the difference between the measured threshold (x) and the reference N, weighted by the inverse of the variance. Points with higher variance are considered less reliable and given less weight in the MD metric; blind spots (i.e. (15°, ±3°)) are not included in the analysis. MD values typically range from negative to slightly positive. Negative MD values indicate less than normal sensitivity in the entire visual field and the patient may be suspected of glaucoma and/or cataract.

$$\mathrm{MD} = \frac{\sum\limits_{i=1}^{n} \left\{ \frac{1}{\sigma_i^2}(x_i - N_i) \right\}}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}} \tag{1.1}$$

While the MD is an intuitive summary metric for a visual field, it captures the general reduction in sensitivity, which is typical of cataract, but does not capture local asymmetries (variations) in the field, which is typical in a glaucomatous field. PSD is another metric that tries to capture this asymmetry. Mathematically, it is defined as the weighted variance of the field, as defined in eq. (1.2). n refers to the numbers of locations, less the two blind spots, in the test pattern used for calculation (e.g. $n = 54 - 2$ for the 24-2 pattern). PSD is always positive, with higher value indicating a more varying field with a higher likelihood of glaucoma.

$$\mathrm{PSD}^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \times \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - N_i - \mathrm{MD})^2}{\sigma_i^2} \tag{1.2}$$

Recently the new VFI metric has become available on HFA field analysis reports. It attempts to reduce the influence of cataract on MD and is argued to be a better indicator for glaucoma. It is reported as a percentage with 100% being a perfectly healthy field and 0% being a blind field. [5]

Another metric developed is Glaucoma Hemifield Test (GHT). It combines measurement of the overall field sensitivity and differences between the top and half hemi-fields into a few hand-crafted "if" branches to classify a field as one of the following categories for easy interpretation: [6]

- Abnormally high sensitivity
- Outside normal limits
- Borderline
- General reduction of sensitivity

- Within normal limits

These four metrics can be seen in fig. 1.2.

1.1.2 Trend-Based Progression Analysis

Currently, the main approach to determine glaucoma progression is a trend-based analysis by performing OLSLR on MD. The HFA Guided Progression Analysis reports OLSLR on its VFI that offers similar information. The fitted trend is typically extended five years into the future, and the slope of the line is used to classify patients into three categories (mild: 0 to -0.4 dB/year, moderate: -0.5 to -2 dB/year, and severe: < -2 dB/year). [7] Glaucoma specialists combine these data with additional information such as the thickness of the RNFL in the macula, IOP and the cup-to-disk ratio to determine the optimal course of treatment.

1.1.3 Limitations

The current method of evaluating visual field information is limited in its robustness. For example, VFI may remain at 100% in 22% of patients who have MD of -5 dB or better. Hence early progression (up to -5 dB) may be missed if VFI alone is used. Moreover, based on a semi-annual follow-up schedule, at least five fields are required to produce an accurate prediction for a disease that is progressing at moderate rates (the slope of the OLSLR on MD is -1.0 dB/year). [7] This means that meaningful decision about the progression of the disease cannot be made until at least two years after the initial visit.

Last but not least, OLSLR is a simple model chosen based on its mathematical simplicity but not upon physiology. Since glaucoma can be caused by multiple neurological pathways, it is argued that the progression trend is likely nonlinear in nature. [8]

In short, the use of linear regression to predict visual field loss and the need for a long observation period can impede timely and accurate decision making by the clinician. A better method that can provide more accurate information to the clinician in fewer visual field tests can allow more effective intervention for glaucoma patients in the disease's early stages. This is the primary motivation for the alternative models reviewed below and for the current work.

1.2 Literature Review: Alternative Progression Models

This section will review alternatives to the OLSLR on MD model for visual field progression prediction.

1.2.1 Statistical Models

A complement to trend-based analysis is event-based analysis. A popular implementation is the Glaucoma Progression Analysis (GPA) since it is readily provided by the HFA. Instead of fitting a trend on a global index, the progression of each point in the field with respect to a baseline measurement is considered. On average, this method is found to have low false-positive rate. However, this is dependent upon achieving a good baseline and does not work for patients with severe visual field loss. [9]

Other non-linear models suggest fitting an exponential model to the visual field indices. For example, Pathak et al. argued that an exponential model is better supported by recent knowledge of structure-function relationship than a linear model. [8] They demonstrated that a linear mixed-effect (LME) approach with an exponential model provided significantly better prediction of glaucoma progression than linear models. However, it is also pointed out that even though the LME approach has better results than the OLSLR, it still cannot capture the full extent of glaucomatous visual field change.

Other innovative approaches to the glaucoma progression problem in the literature include:

- Point-wise linear regression (PLR): Developed for improving early detection, PLR combines event-based and trend-based analysis [10].

- Analysis with Non-Stationary Weibull Error Regression and Spatial Enhancement (ANSWERS): Using spatial correlation between points and incorporating non-stationary variability; progression detection was found to be better especially in short time series [11].

- Spatially filtering visual field data before PLR analysis: By applying a spatial filter that incorporates physiological relationships between measured contrast sensitivities at test points within the visual field, the specificity of the PLR method is not affected but the sensitivity of detecting the rate of progression is improved [12].

1.2.2 Machine Learning Models

The above methods use different approaches to improve the prediction of glaucoma progression and demonstrate the complexity of the prediction task. This is not surprising due to the complex nature of the disease. A natural step forward involves leveraging the power of modern machine learning techniques to integrate features such as non-linear trends, correlation between visual field test locations, field patterns, etc. into a single model.

Existing studies have demonstrated the usefulness of machine learning in glaucoma care. For example, Asaoka et al. compared the performance of traditional machine learning classifiers with that of a deep feed-forward neural network (FNN) in diagnosing preperimetric glaucoma with visual field data [13]. Their FNN model performed significantly better than other methods. In another study, Yousefi et al. applied clustering algorithms to extract visual field patterns as features, then adapted the traditional linear regression algorithm to model the features to generate the predictions [14]. The machine learning-based index for the detection of glaucoma progression outperformed current methods.

However, these studies either did not directly address the problem of predicting visual field progression or only used traditional machine learning methods for initial feature extraction. In a recent study that fully utilized the power of machine learning for the prediction task, Wen et al. [15] used deep learning network to predict future visual field given the measurements of a single current visual field. Their deep learning network demonstrated amazing capability to generate prediction for future visual fields for up to 5.5 years with a correlation of 0.92 between the predicted MD and actual future MD (average difference of 0.41 dB). Their results suggest the tremendous potential of this research area.

Another benefit of using a machine learning model is the ability to use visual field data (i.e. functional indication of visual field integrity) and OCT data (i.e. structural indication for the integrity of the retina) at the same time. It is known that by utilizing both visual field and OCT data the sensitivity of glaucoma detection can be improved [16, 17]. Multivariate models including both visual field and OCT have also been shown to be successful [18]. However, there is limited research on combining visual field and OCT features through a machine learning model. A limited attempt to explore this idea by Silva et al. did not produce better results than those obtained by visual field only parameters [19].

1.3 Outline of Current Work

This thesis will focus on describing the evaluation of performance of common machine learning architectures on a publicly available longitudinal visual field dataset. Chapter 2 will introduce the Rotterdam dataset used and performance of existing non-learning-based prediction methods. Chapter 3 will describe the learning-based algorithms evaluated in this work. Chapter 4 will compare the performance of the algorithms evaluated and address issues arising from the current investigation.

Chapter 2

Rotterdam Dataset and Non-learning Methods

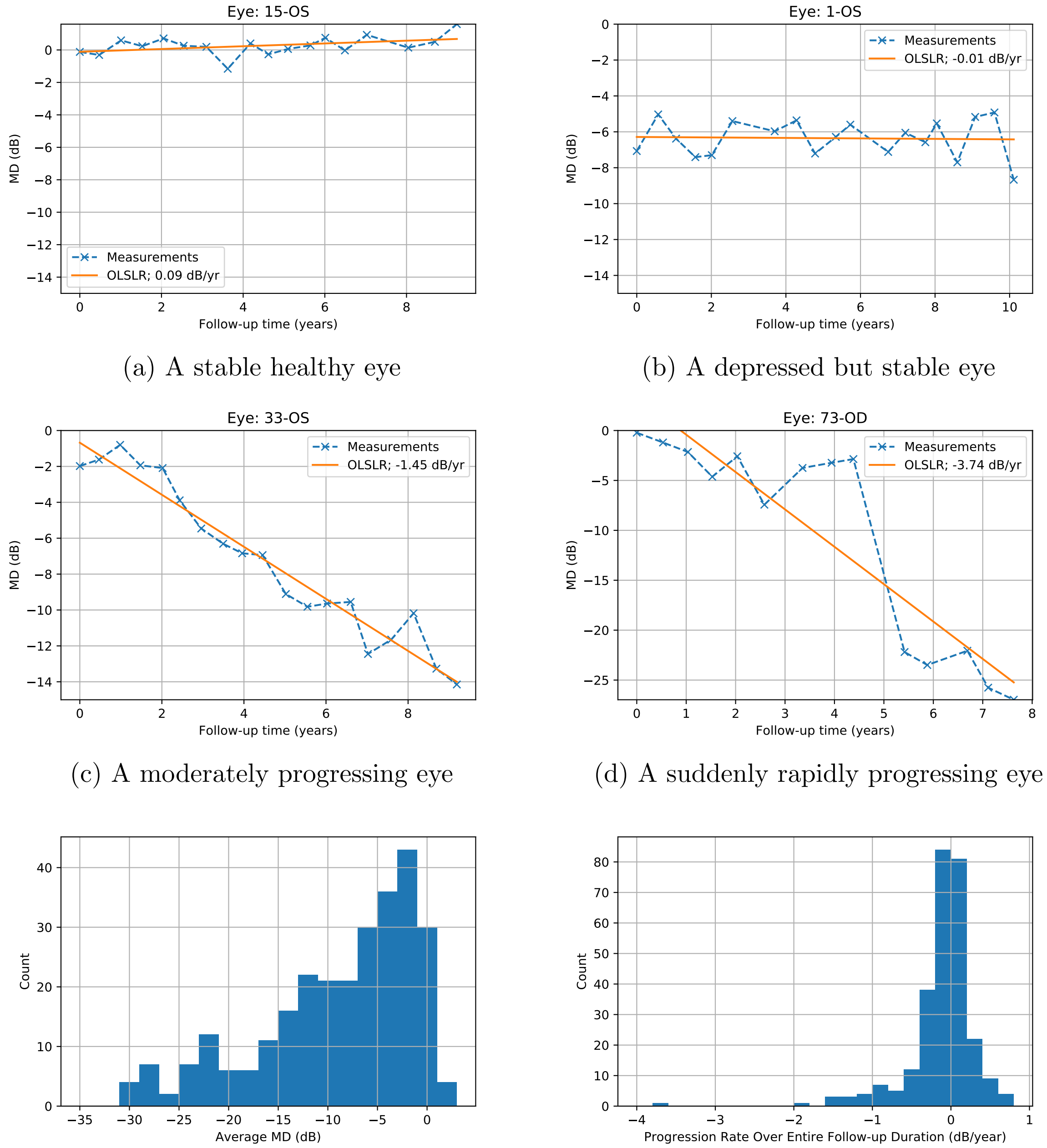
The Longitudinal Glaucomatous Visual Field data from Rotterdam Ophthalmic Data Repository [25] consists of data from 139 patients' (80 male versus 59 female) 278 eyes. A total of 4863 24-2 visual field test results with thresholds and Mean Deviation (MD) are available. On average each eye has 17.5 fields available with mean total follow-up duration of 9.2 years. 270 (97.1%) eyes have at least 14 fields with a minimum of 7.6 years of follow-up. The follow-ups are originally scheduled for every 6 months. The mean and median actual follow-up interval between tests is 203 and 189 days respectively; the standard deviation of follow-up time is 72.3 days. 346 (7.5%) of follow-ups had an interval of more than 270 days.

2.1 Dataset Characteristics

To investigate the composition of healthy versus glaucomatous patients the dataset, the characteristic of MD values in the dataset is investigated. In figure 2.1e the distribution of average MD value calculated from all tests administers on each of the 278 eyes is shown. The mean and median of the distribution are -8.9 and -6.8 dB respectively. The dataset contains mostly eyes with mild to moderate reduced MD values (75% of eyes have average MD > -13.2 dB).

To investigate the progression rate in the sample, an ordinary least-squares linear regression (OLSLR) line is fit to each eye's MD history. The distribution of the slope (db/year) is shown in fig. 2.1f.

It is important to note that glaucoma is a very slowly progressing disease. In addition, the data is collected from patients who are undergoing standard treatment. Moreover,



(a) A stable healthy eye

(b) A depressed but stable eye

(c) A moderately progressing eye

(d) A suddenly rapidly progressing eye

(e) Mean MD

(f) Progression rate ΔMD/yr

Figure 2.1: Overview of the Rotterdam Longitudinal Glaucomatous Visual Field Dataset. (a-d) Select examples of the longitudinal visual field test results. (e) Distribution of mean MD value over each eye's follow-up duration. A lower mean MD value represents a more depressed eye and likely indicates more severe disease. Not that since perimeters typically have a dynamic range of 0–34 dB, an MD of -30 essentially indicates a blind/almost blind eye. (f) Distribution of progression rate of each eye over the follow-up duration as measured by the rate of change of MD. (e-f) shows that a large number of eyes are relatively healthy and most eyes did not progress. This is likely due to either the slowly progressing nature of the glaucoma disease or due to appropriate intervention by clinicians.

both eyes of a glauocma patient are tested, and a patient can often have one glaucomatous eye and another healthy eye. As a result, we see many patients who are healthy (close to 0 MD) and not progressing (close to 0 MD/yr).

The inclusion/exclusion criteria as well as further information about the dataset can be found in appendix B.

2.2 Performance of Simple Extrapolators

In this section, using the Rotterdam dataset, we establish a baseline prediction performance using simple non-learning based extrapolation methods.

2.2.1 Tasks

Two prediction tasks are evaluated:

- MD prediction:

  This is the traditional, current clinical routine task of predicting future field index value(s) from current known fields and their summaries.

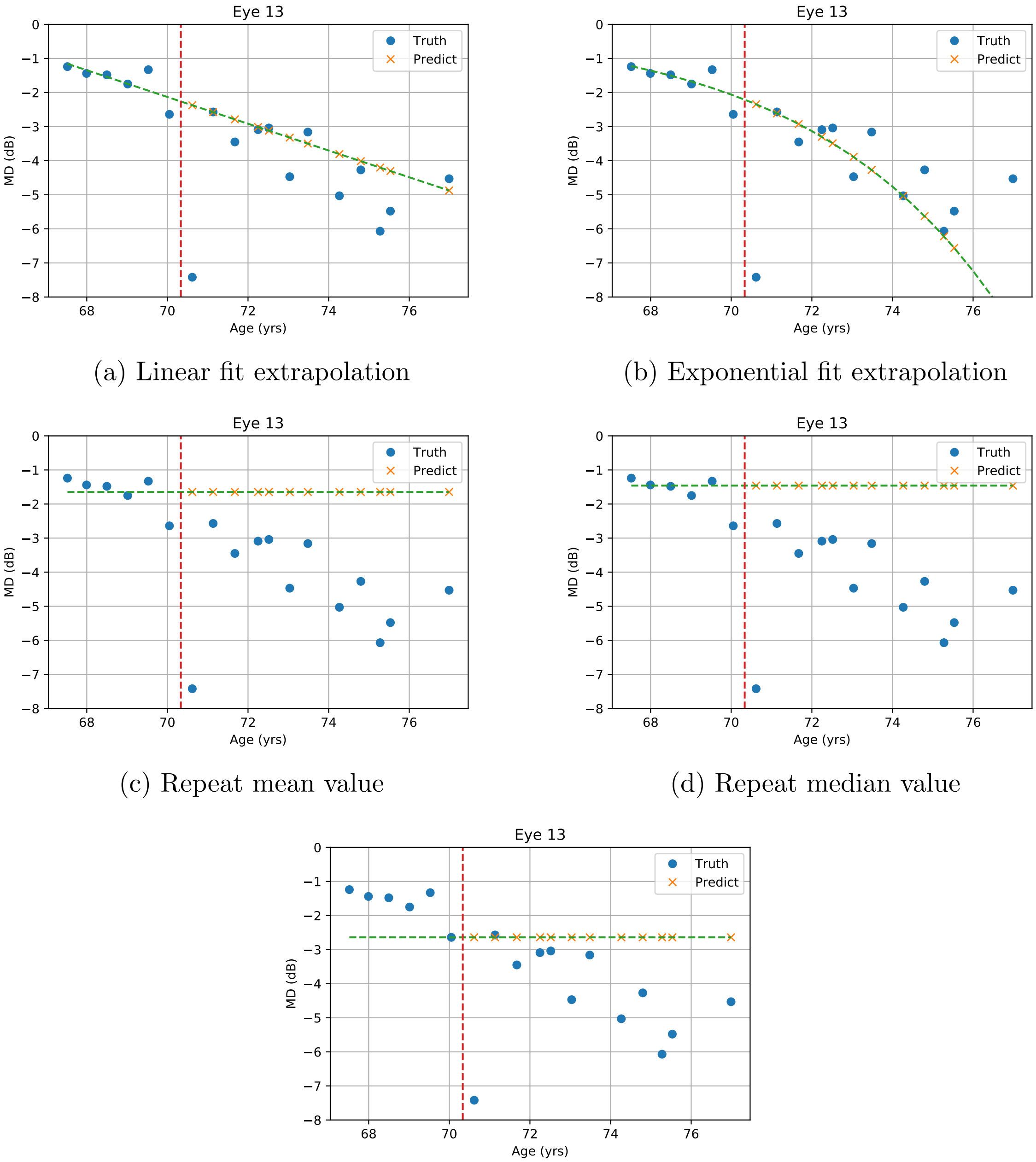
- Point-wise prediction:

  In this task, instead of predicting one field index that summarizes each future field, the algorithm will output full future field(s) at all n locations. (For 24-2, $n = 52+2$ points). This task is usually not attempted traditionally, but in recent years is a common goal of modern machine learning algorithms in this field.

The input to the algorithms consists of time (age) at each field examination, and the respective MD or point-wise field thresholds. Both tasks are evaluated with 3 and 6 input ("known") fields to predict the next future first, second, third, . . . visual field results. Having 6 inputs is similar to the current clinical guidelines [7]. Using only 3 inputs would be more ideal for earlier detection and used for training learning algorithms.

To fully utilize the data available, all combinations ("prediction series") of 3 or 6 consecutive fields approximately 6 months apart are used for evaluation. For example, for a patient with fields 1 to 5, fields 1 to 3 are used to predict 4 and 5 and fields 2 to 4 are used to predict 5. Details on the generation of testing dataset is described in section 3.1.2.

.



(a) Linear fit extrapolation

(b) Exponential fit extrapolation

(c) Repeat mean value

(d) Repeat median value

(e) Repeat last value

Figure 2.2: Illustration of the five fitting methods described in section 2.2.2. Shown is an example of the MD prediction problem with 6 inputs.

2.2.2 Methods

Inspired by the current widely accepted methods and work by Chen et al. [2], the following extrapolation methods are assessed on the dataset:[1]

- Linear fit extrapolation, i.e. ordinary least-squares linear regression (OLSLR):

  Based on known MD and threshold values, predict future corresponding values by linearly fitting and extrapolating on a line that minimizes the sum of squared errors. Mathematically,

  $$\hat{y}(x) = b + ax = \mathbf{w} \cdot [1, x] \tag{2.1}$$

  where $\mathbf{w} = [b, a] \in \mathbb{R}^2$ are the linear fitting parameters. The closed form solution implemented is:

  $$\mathbf{w} = (X^T X)^{-1} X^T y \tag{2.2}$$

  where $X \in \mathbb{R}^{m \times 2}$ is the vector of time of m known measurements padded with a column of ones for the bias term. $y \in \mathbb{R}^{m \times n}$ is a matrix of dependent values consisting of rows of measurements (e.g. $n = 1$ for MD or $n = 54$ for 24-2 field thresholds).

- Exponential extrapolation:

  The exponential model is an alternative to the traditional linear model that has been proposed in literature [2, 8], where

  $$\ln \hat{y}(x) = \mathbf{w} \cdot [1, x] \tag{2.3}$$

  Since to take the logarithm of y, one needs to ensure $y > 0$, so the following modifications are made: (1) when used for MD prediction, $y = -MD$; (2) when not all values of y are positive (e.g. some tests yield MD> 0), an appropriate bias is added to make all y positive. The equation above can also be fitting using eq. (2.2).

- Repeating mean value:

[1] These methods are intentionally called extrapolators to avoid confusion with the learning-based regression models introduced later in chapter 3.

Predict with a constant value that is the mean of known measurements.

$$\hat{y}(x) = \text{mean}(y) \tag{2.4}$$

- Repeating median value:

  Predict with a constant value that is the median of known measurements.

$$\hat{y}(x) = \text{median}(y) \tag{2.5}$$

- Repeating last value, i.e. nearest neighbor extrapolator:

  Extrapolate with the closest value. In the predict task, it means simply taking the last observed value.

$$\hat{y}(x) = y_m \tag{2.6}$$

Lastly, the following physiological assumptions are added to limit the range of predictions from the linear and exponential fit methods to increase prediction stability: [2]

1. The minimum MD value is restricted to -30 dB;
2. The differential light sensitivity (DLS) threshold value is restricted to between -1 and 40 dB;[3]
3. Visual field sensitivity does not improve, i.e. the prediction value is always less sensitive than the last data point available.

The average MD prediction mean absolute error (MAE) for prediction data series n is calculated as: (figs. 2.3a and 2.3b)

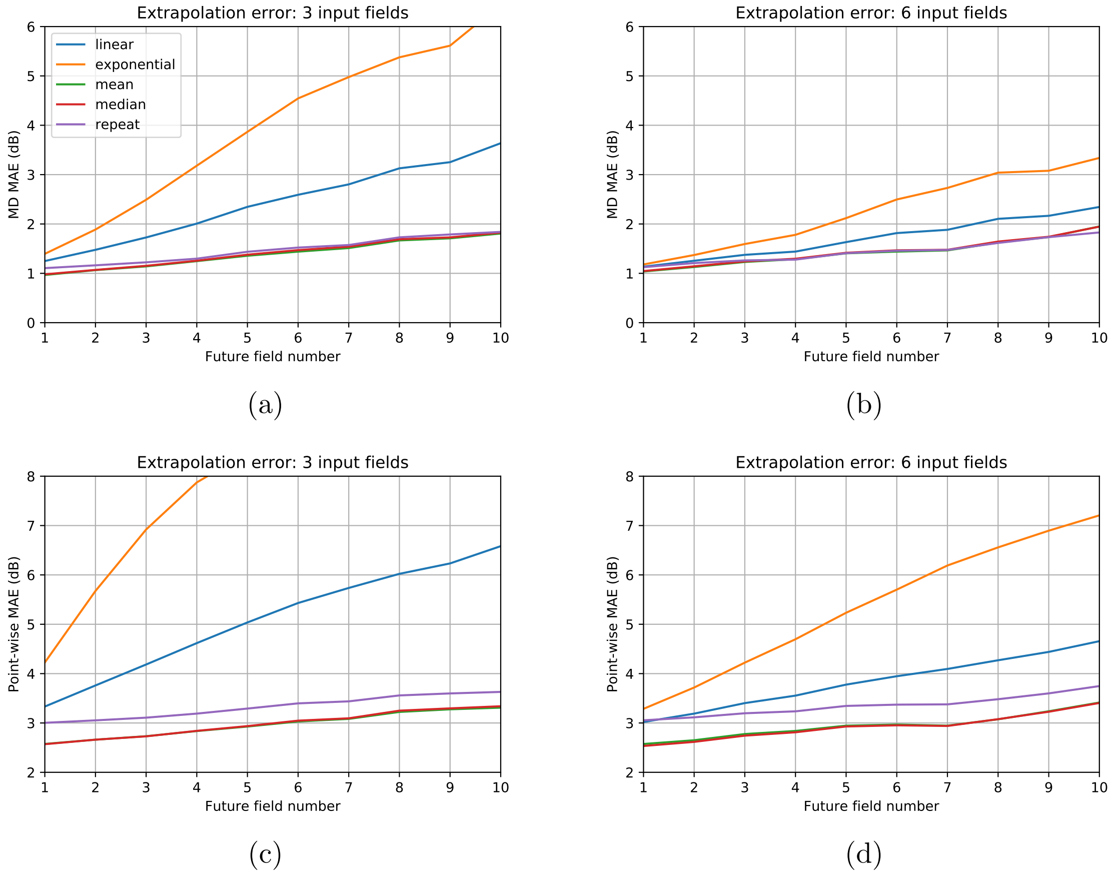
$$\text{MAE}_{\text{MD}} = \frac{1}{N} \left| \widehat{\text{MD}}^{(n)} - \text{MD}^{(n)} \right| \tag{2.7}$$

The average whole field point-wise prediction MAE for prediction data series n is calculated as: (figs. 2.3c and 2.3d)

$$\text{MAE}_{\text{VF}} = \frac{1}{54N} \sum_{i=1}^{54} \left| \widehat{\text{VF}}_i^{(n)} - \text{VF}_i^{(n)} \right| \tag{2.8}$$

[2] These restrictions are also added because without them, the performance of extrapolators deteriorates very quickly in a few fields.

[3] -1 dB because the Rotterdam dataset includes some field thresholds at -1 dB, likely indicating a < 0 dB result on the Humphrey Field Analyzer (HFA).



(a)

(b)

(c)

(d)

Figure 2.3: Accuracy of extrapolation methods over different prediction lengths (a,b) MD extrapolation result (c,d) Whole field point-wise extrapolation result (a,c) Extrapolation from 3 fields (b,d) Extrapolation from 6 fields. As expected, the longer the prediction length on the x-axis, the less the predictive power of any method. The MD prediction results are as expected better in dBs than point-wise prediction. Having 6 input fields yield better result for linear and exponential models as it yields a more stable result, while increasing the number of fields has little effect on the mean/median/repeat extrapolator. This illustrates the limitation of the dataset where many subjects are not progressing.

2.2.3 Results and Discussion

Figure 2.3b illustrates the typical clinical progression prediction method. In a typical situation where 6 fields are used to estimate future MD value after 5 years (i.e. approximately "next 10th"), the mean and median MAE is 2.3 and 1.6 dB, respectively. Simply repeating the mean (1.8 dB), median (1.8 dB), or last (1.8 dB) observation achieves lower prediction error after 10 fields. (See table 2.1) This somewhat surprising result in fact agrees with the previous observation that much of the dataset was not progressing. This illustrates that glaucoma, in many cases and definitely in this dataset, is a very slowly progressing disease. Fitting a predetermined function may severely overfit a non-existent

Table 2.1: Extrapolation method performance for predicting the 10th field (25%/50%/75% are the 25-th percentile/median/75-th percentile respectively)

		Extrapolation Method				
Prediction MAE (dB)		Linear	Exp.	Mean	Median	Repeat
MD (3 inputs)	mean	3.6	6.4	1.8	1.8	1.8
	std	4.3	8.4	2.2	2.3	2.2
	25%	0.8	0.9	0.5	0.5	0.5
	50%	1.9	2.3	1.1	1.1	1.2
	75%	4.9	8.5	2.2	2.1	2.3
MD (6 inputs)	mean	2.3	3.3	2.0	1.9	1.8
	std	2.5	4.5	2.3	2.4	2.1
	25%	0.7	0.7	0.6	0.6	0.5
	50%	1.6	1.7	1.2	1.2	1.2
	75%	3.0	3.9	2.4	2.3	2.4
Point-wise (3 inputs)	mean	6.6	9.8	3.3	3.3	3.6
	std	7.4	10.9	4.2	4.4	4.6
	25%	1.0	1.0	0.7	1.0	1.0
	50%	4.0	4.0	2.0	2.0	2.0
	75%	10.0	20.0	4.3	4.0	5.0
Point-wise (6 inputs)	mean	4.7	7.2	3.4	3.4	3.7
	std	5.4	8.5	4.3	4.5	4.7
	25%	1.0	1.0	0.7	0.5	1.0
	50%	3.0	3.3	2.0	2.0	2.0
	75%	6.5	11.0	4.3	4.0	5.0

trend, a potential limitation of the current OLSLR approach.

In fig. 2.3, it is observed that within each table the error increases as fields further in the future are to be predicted, as expected. Fewer input fields (3 versus 6) also resulted in higher error, as expected. The point-wise prediction task is harder than the MD prediction task due to the nature of MD being a summary statistics that averages errors across the field.

These values establish a baseline against which the algorithm designs in chapter 3 will be compared.

Chapter 3

Learning Methods

This chapter provides framing of the visual field progression prediction tasks introduced previously as learning problems, and introduces the learning algorithm architectures to be evaluated against the Rotterdam dataset.

3.1 The Learning Task

3.1.1 Formulation and Normalization

Given the information in the Rotterdam dataset, we can write each patient visit as one feature vector as follows:

$$v^{(i)} = \begin{bmatrix} \mathrm{VF}_1^{(i)} & \mathrm{VF}_2^{(i)} & \cdots & \mathrm{VF}_{54}^{(i)} & \mathrm{MD}^{(i)} & \mathrm{IOP}^{(i)} & t^{(i)} \end{bmatrix} \in \mathbb{R}^{57} \tag{3.1}$$

where $t^{(i)}$ is the time of the visit, and $\mathrm{VF}_j^{(i)}$ is the differential light sensitivity (DLS) for the j-th point in the visual field. By convention of the visual field analysis tools developed by Marin-Franch et al. [26], all fields of the left (OS) eye are flipped horizontally (unmodified vertically) to match with the coordinates of the right eye, and then each visual field location is indexed from left to right, then top to bottom (same as the reading direction in English). Each of the vector components is normalized to approximately $[0, 1]$ in the range shown in table 3.1.

For three input visits (i.e. the input to our learning task), we can concatenate the three visits and write the entire input vector as:

$$x^{(n)} = \begin{bmatrix} v^{(i)} & v^{(i+1)} & v^{(i+1)} \end{bmatrix} \in \mathbb{R}^{171} \tag{3.2}$$

Table 3.1: Feature range normalization with appropriate physiological ranges

	Normalization Range			
Feature	Lower (0)	Upper (1)	Unit	Comments
VF	0	40	dB	
MD	0	40	dB	Negative sign preserved
IOP	0	20	mmHg	
Age (t)	50	80	years	

3.1.2 Learning Dataset Generation

The entire dataset can be put into matrix form, where the input variable is:

$$\mathbf{X} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times 171} \tag{3.3}$$

and the prediction variable is simply the DLS at each location:

$$\mathbf{Y} = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(N)} \end{bmatrix} = \begin{bmatrix} \mathrm{VF}_1^{(1)} & \mathrm{VF}_2^{(1)} & \cdots & \mathrm{VF}_{54}^{(1)} \\ \mathrm{VF}_1^{(2)} & \mathrm{VF}_2^{(2)} & \cdots & \mathrm{VF}_{54}^{(2)} \\ & \vdots & & \\ \mathrm{VF}_1^{(N)} & \mathrm{VF}_2^{(N)} & \cdots & \mathrm{VF}_{54}^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times 54} \tag{3.4}$$

Instead of training a separate regressor, the Mean Deviation (MD) values will be calculated from the predicted point-wise values.

The glaucoma patients are followed up on an interval of 6 months. Therefore, the prediction series are generated such that:

$$\Delta t = 183 \text{ days} \approx (t_{v^{(i+1)}} - t_{v^{(i)}}) \approx (t_{v^{(i+2)}} - t_{v^{(i+1)}}) \tag{3.5}$$

and the output field is K fields (approximately $0.5K$ years) after the last input field:

$$t_{v^{(Y)}} - t_{v^{(i+1)}} \approx 183K \text{ days} \tag{3.6}$$

Learning sets in the Rotterdam dataset that satisfy eqs. (3.5) and (3.6) for all $K \in [1, 10]$ are generated for learning.

3.1.3 Error Definition

The average point-wise mean absolute error (MAE) error is defined as:

$$\mathrm{MAE}_{\mathrm{VF}} \triangleq \frac{1}{54N} \sum_{n=1}^{N} \sum_{j=1}^{54} \left| \widehat{\mathrm{VF}}_j^{(n)} - \mathrm{VF}_j^{(n)} \right| \tag{3.7}$$

The average MAE for MD is defined as:

$$\mathrm{MAE}_{\mathrm{MD}} \triangleq \frac{1}{N} \sum_{n=1}^{N} \left| \widehat{\mathrm{MD}}^{(n)} - \mathrm{MD}^{(n)} \right| \tag{3.8}$$

The calculation of MD was introduced as eq. (1.1), which requires (1) age-adjusted baseline DLS values and (2) variance at each location of the field (i.e. weights when averaging the deviation values). These values, specifically the ones used to generate MD values in the Rotterdam dataset, is not available. However, one can modify eq. (3.8) by substituting in eq. (1.1):

$$\begin{aligned} \mathrm{MAE}_{\mathrm{MD}} &= \frac{1}{N} \sum_{n=1}^{N} \left| \sum_{j=1}^{54} w_j \left(\widehat{\mathrm{VF}}_j^{(n)} - N_j \right) - \sum_{j=1}^{54} w_j \left(\mathrm{VF}_j^{(n)} - N_j \right) \right| && (3.9) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left| \sum_{j=1}^{54} w_j \left(\widehat{\mathrm{VF}}_j^{(n)} - \mathrm{VF}_j^{(n)} \right) \right| && (3.10) \end{aligned}$$

where N_j is the age-adjusted population normal value at the j-th visual field point, and

$$w_j = \frac{\frac{1}{\sigma_j^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \tag{3.11}$$

This expression is only dependent upon the weights used in calculating MD, since the age-adjusted baseline cancels out. Sample weights are available from Heijl et al. [3] or Marin-Franch et al. [26]

3.1.4 Evaluation Procedure

70% of the data is used for training, 15% is used for validation to fine tune the hyper-parameters of the model. Then, the hyper-parameters are set and 5-fold cross validation is performed to report model performance.

3.2 Models

This section describes a range of simple to complex learning models that are proposed for evaluation for the learning task. Simpler models might be more robust to noise in the visual field dataset and generalize better for the limited data available, while a complex model might extract more powerful visual field and patient features from the data at the peril of over-fitting.

3.2.1 Ridge Regression

Ridge regression is an L_2-regularized linear regression[1]that aims to minimize the combination of the sum squared cost and the L_2 norm of the weight parameters. By weighting on the model parameters, it is expected to yield a simpler model that is fit less toward specific noisy trends in the training set and generalizes better. The model prediction is:

$$\hat{\mathbf{y}} = [1, \mathbf{x}] \cdot \mathbf{W} \tag{3.12}$$

The objective is to minimize the regularized sums squared cost:

$$\mathbf{W} \leftarrow \arg\min_{\mathbf{W}} ||\hat{\mathbf{Y}} - \mathbf{Y}||_2^2 + \lambda ||\mathbf{W}||_2^2 \tag{3.13}$$

This has a closed form solution, which is directly implemented:

$$\mathbf{W} = \left(\mathbf{X}_1^T \mathbf{X}_1 + \lambda I\right)^{-1} \mathbf{X}_1^T \mathbf{Y} \tag{3.14}$$

where $\mathbf{X}_1 \triangleq \begin{bmatrix} \mathbf{1} & \mathbf{X} \end{bmatrix}$ includes the bias feature.

3.2.2 Huber Regression

The Huber linear regressor uses a cost metric that is more robust to outliers, with the following definition:

$$L(\hat{y}_j, y_j) = \begin{cases} \frac{1}{2}(\hat{y}_j - y_j)^2, & |\hat{y}_j - y_j| \leq \delta \\ \delta|\hat{y}_j - y_j| - \frac{1}{2}\delta^2, & |\hat{y}_j - y_j| > \delta \end{cases} \tag{3.15}$$

Intuitively, for points close to the target, a squared cost is used, while for points further away than a threshold (e.g. outliers), a linear cost is used. Therefore it is also

[1] In this document the term "linear regression" is intentionally avoided to avoid confusion with the simple linear extrapolation method by extrapolating each variable with a linear fit over time, which is often called the ordinary least-squares linear regression (OLSLR) method.

known as a robust regressor. One may hope that it may be better at rejecting noise from visual field recordings.

One Huber regressor is needed for each output dimension, so essentially 54 regressors are fit with the same input. Implementation wise, this is done by chaining `HuberRegressor` into `MultiOutputRegressor` using the Scikit-learn library. [27] There is no closed-form solution for Huber regression so an iterative solution is found.

3.2.3 Multi-Layer Perceptron

The multi-layer perceptron (MLP) (also known as fully-connected neural network) model takes the 171 features as input and outputs 54 DLS values. Considering the high dimensionality of the output, a standard three-layer (two hidden layers) network is investigated with number of neural units being the hyperparameter. L_2 regularization is applied to all weights. ReLU is used as the activation function at all layers except the output layer. The model is implemented in Tensorflow. [28] The average point-wise MAE loss is used as the objective function. The weights are optimized iteratively with the Adam Optimizer using batch size of 32.

3.2.4 Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of neural network architecture that, thanks to the tremendous improvement in computational power recently, has shown tremendous performance in automatically learning to extract image features instead of requiring hand-crated features detection algorithms. Each visual field can be seen as an 8×9 image. This relatively low dimensionality, compared to the inputs to typical image recognition tasks in which convolutional neural networks (CNNs) are applied, may leave limited room of improvement in the performance of CNN over a simpler flat architecture such as an MLP. However, spatial patterns in the visual field are have been known to important to glaucoma diagnosis and progression (see fig. 3.1), and CNN is an architecture that is known to be able to extract such spatial information.

A typical CNN architecture is implemented. Specifically, this CNN consists of:

1. Input layer: Input is three visual field “images” stacked ($8 \times 9 \times 3$). Locations outside the 24-2 examination area are replaced with zeros.

2. First convolutional layer with 3×3 kernel, stride length 1. Output is passed through ReLU activation and batch normalization.

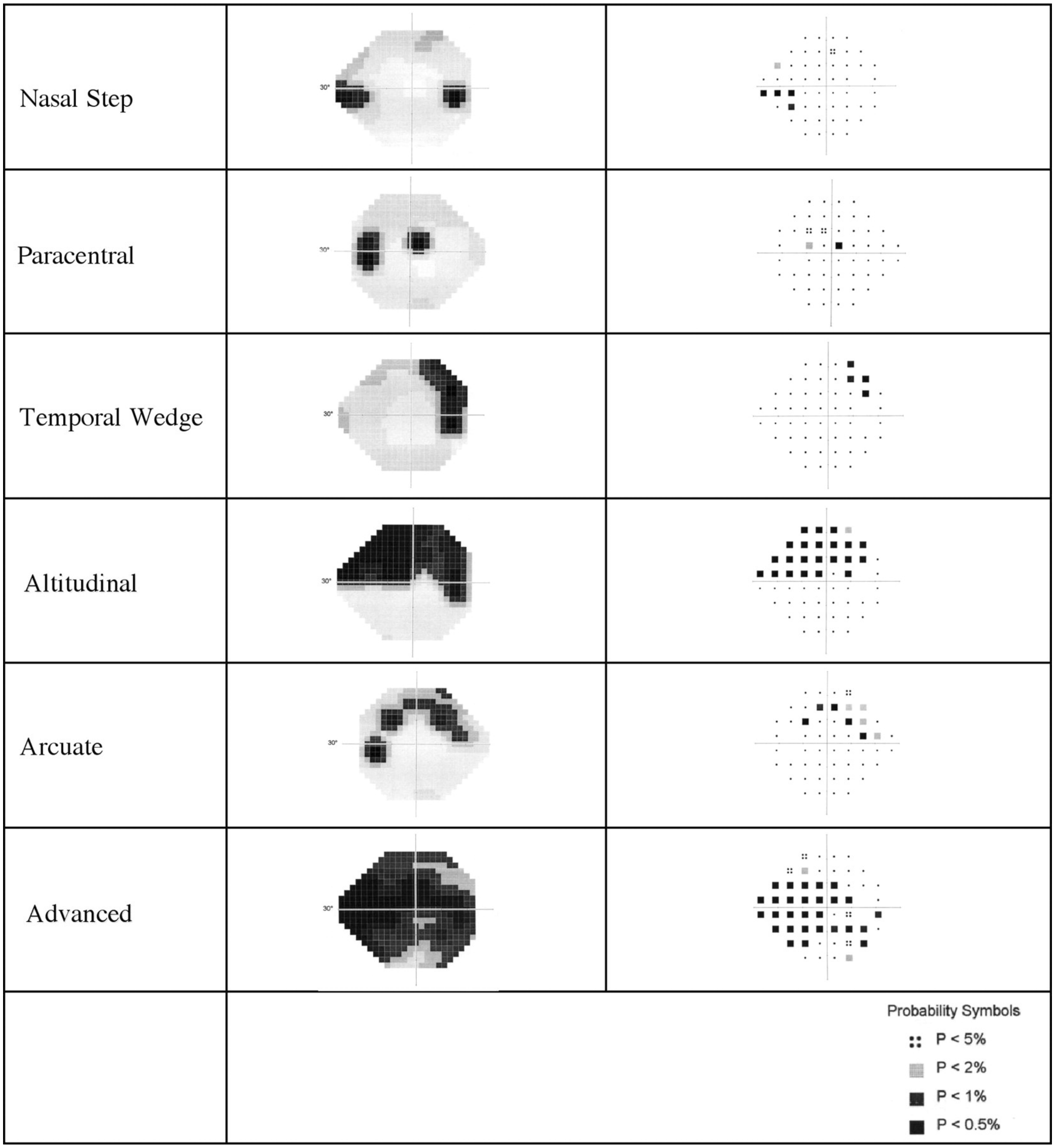


Figure 3.1: Examples of typical patterns of glaucomatous visual field loss that are known to be physiologically and clinically important for diagnosis and prediction. Such evidence is the motivation behind employing a CNN architecture that is likely to preserve and extract useful spatial information in the visual field. (Figure from Sample et al. [29])

3. Max pool layer with pool size and stride length of 2.
4. Second convolutional layer with 3×3 kernel, stride length 1. Output is passed through ReLU activation and batch normalization, then flattened.
5. Additional features (i.e. MD, IOP, and Age) are concatenated to the vector.
6. First fully connected hidden layer. Output is passed through ReLU activation and a dropout layer.
7. Second fully connected hidden layer.
8. Output layer. ($d = 54$)

The modeled is trained using Adam Optimizer to minimize the L_2 regularized MAE loss, and implemented with Tensorflow. [28] The number of channels in the CNN, neural units in the fully-connected layer, and regularization are fine-tuned as hyperparameters.

3.2.5 Deep CNN+LSTM Network

The final proposed framework involves adding the long short-term memory (LSTM) network to the CNN network. The LSTM module is a type of recurrent neural network (RNN). Its recurrent nature allows modeling of sequential data; in the case of the current prediction task, it is used to model the sequential visual field inputs at different time points.

The LSTM network also has the benefit of only needing one network to predict multiple outputs in the future due to its recurrent nature. Because of this, the training dataset used to train this network is slightly different from networks above. When training this network, it is necessary to generate a sequence of all fields $i = 0, 1, 2, (2+1), (2+2), \ldots, (2+K)$ where K is the maximum number of fields in the future that is desired to be predicted. At test time, the model takes fields $i = 0, 1, 2$ as input, and outputs all field predictions at $i = (2+1), (2+2), \ldots, (2+K)$. $K = 5$ and $K = 10$ are tested, where $K = 5$ is expected to include more training examples.

The detailed procedure of this network is shown as algorithm 1 in appendix A. The network is implemented using the PyTorch framework. [30]

Chapter 4

Results and Discussion

4.1 Model Parameters

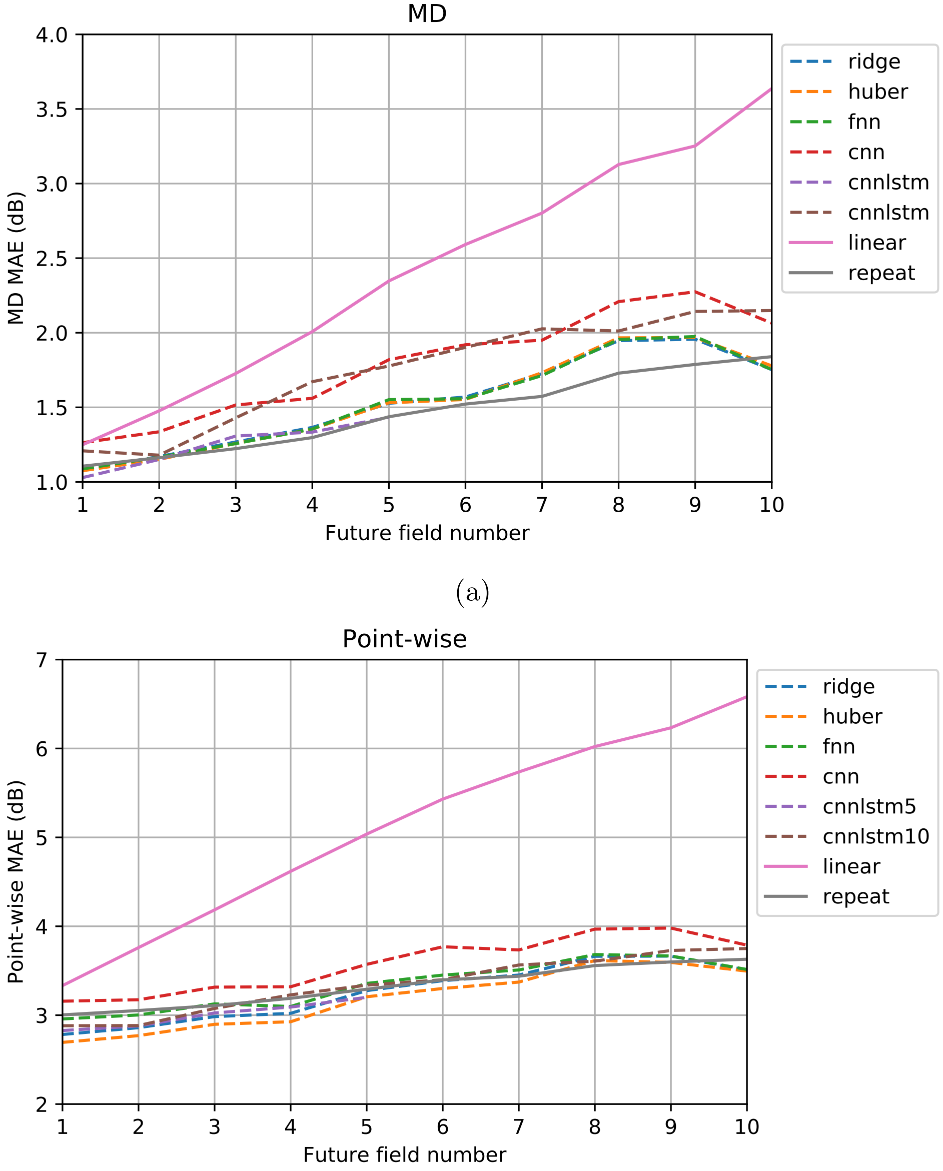
Using the 70% training set, 15% validation set, and 15% testing set, the hyper-parameters chosen for each learning model is shown in table 4.1.

Table 4.1: Final hyperparameters for learning models

Model	Parameters
Ridge Regression	$\lambda = 10$
Huber Regression	$\epsilon = 1.0$ $\alpha = 0.001$ (ϵ and α are specific to the Scikit-learn implementation [27])
MLP	Size of first hidden layer: 200 Size of second hidden layer: 150
CNN	L_2 Regularization: $\lambda = 0.8$ Convolution layers depth: $(128, 256)$ Fully connected layers size: $(512, 256)$
CNN+LSTM	L_2 Regularization: $\lambda = 0.0001$ LSTM hidden state size: 4096 Convolution block size for three convolution blocks: $(64, 128, 256)$

4.2 Cross Validation Results

5-fold cross validation is performed on the learning models with hyperparameters listed in table 4.1. The same cross-validation set is used for ridge regression, Huber regression,



(a)

(b)

Figure 4.1: MD and point-wise field prediction results for learning algorithms with 3 inputs, compared to the linear extrapolator and repeating the last observed value "repeat". There is no significant difference between the learning methods, and due to the dataset containing mostly stable patients, there is no significant difference between the learning methods and "repeat". However, all methods performed better than the linear extrapolation methods. [a]

[a] In an earlier version of this figure, it was shown that the CNN+LSTM model performed slightly better than other learning methods. This has been corrected in this latest batch of experiments.

Table 4.2: Learning algorithm performance for the 5-th and 10-th prediction (25%/50%/75% are the 25-th percentile/median/75-th percentile respectively)

Prediction		Learning Algorithm						Extrapolation	
MAE (dB)		Ridge	Huber	MLP	CNN	CNN+LSTM5	CNN+LSTM10	Linear	Repeat
MD, 5-th	mean	1.53	1.53	1.55	1.82	1.43	1.78	2.35	1.44
	std	1.7	1.73	1.76	1.83	1.44	1.63	2.75	1.59
	25%	0.45	0.45	0.44	0.59	0.49	0.59	0.64	0.43
	50%	1.05	1.02	1.04	1.3	1.03	1.44	1.41	0.99
	75%	1.93	1.93	1.97	2.5	1.93	2.3	2.95	1.85
MD, 10-th	mean	1.75	1.78	1.75	2.06		2.15	3.64	1.84
	std	1.75	1.76	1.78	2.01		2.16	4.28	2.17
	25%	0.57	0.6	0.57	0.64		0.65	0.81	0.51
	50%	1.27	1.27	1.23	1.45		1.47	1.91	1.18
	75%	2.32	2.29	2.24	2.8		2.79	4.94	2.3
Point-wise, 5-th	mean	3.28	3.21	3.36	3.57	3.2	3.34	5.04	3.29
	std	3.55	3.63	3.85	4.08	3.69	3.87	5.85	4.18
	25%	0.95	0.89	0.88	0.94	1	1	1	1
	50%	2.14	2.02	2.04	2.21	1.91	1.94	3	2
	75%	4.25	4.1	4.27	4.53	4	4.1	7	4
Point-wise, 10-th	mean	3.51	3.5	3.51	3.79		3.75	6.58	3.63
	std	3.58	3.74	3.86	4.16		4.33	7.39	4.6
	25%	1.11	1.05	1	1.03		1	1	1
	50%	2.42	2.34	2.24	2.41		2.21	4	2
	75%	4.61	4.54	4.54	4.94		4.64	10	5

multi-layer perceptron (MLP), and convolutional neural network (CNN). A different sequence is used for the CNN+LSTM model due to the recurrent structure of the long short-term memory (LSTM) module; sequences extending to 5 and 10 fields into the future are investigated for CNN+LSTM. These results are compared to the linear extrapolation method and the repeat-last-value ("repeat") method, which were described in section 2.2.3. A detailed comparison of the performance of different methods at predicting 5 and 10 fields in the future is tabulated in table 4.2.

Using the Rotterdam dataset, there is no significant difference between the learning algorithms evaluated for both the Mean Deviation (MD) prediction task mean absolute error (MAE) and the point-wise whole field task MAE. Furthermore, there is no significant difference between the learning and repeat method for both tasks. The linear extrapolation method performed much worse than all other methods in terms of mean, median, and 75-th percentile error, but the 25-th percentile error is similar.

Within each method, the error increases as one extends the prediction further into the future. In addition, the MD prediction MAE is in general approximately one half that of the point-wise prediction MAE. These observations are in agreement with expectation.

4.3 Discussion

The distribution of glaucoma progression rates, expressed in dB/year, is known to be asymmetric with a long tail for highly negative progression rates—some but only limited patients show very rapid progression, but most are reasonably stable as mild progressors. [31] In the Early Manifest Glaucoma Trial (EMGT), Heijl et al. found that the rate of visual field loss in their study sample is: mean -1.08 dB/year, median -0.40 dB/year, interquartile range 1.05 dB/year. [32] Chauhan et al. reported in the Canadian Glaucoma Study (CGS): mean -0.54 dB/year, median -0.35 dB/year, interquartile range 0.64 dB/year. [33] Both studies reported that elder patients are more likely to progress faster. Heijl et al. reported that exfoliation glaucoma progresses much faster than non-exfoliation glaucoma patients.

For the Rotterdam dataset used in this work, the rate of change values are: mean

Table 4.3: Glaucoma progression rates and study dataset statistics of EMGT, CGS, and Rotterdam studies

	Study		
	EMGT [32]	CGS [33]	Rotterdam
Progression (db/year)			
mean	-1.08	-0.54	-0.10
std	2.07	1.10	0.46
25%		-0.76	-0.21
50%	-0.40	-0.35	-0.04
75%		-0.12	0.12
interquartile range	1.05	0.64	0.33
Age (years)			
N (< 68)	53 (45%)		108 (78%)
N (≥ 68)	65 (55%)		31 (22%)
mean	68.0		59.7
std	5.1		10.3
25%		63.5	52.7
50%	68	68.2	61.3
75%		75.1	67.6

−0.10 dB/year, median −0.04, interquartile range 0.33. Relevant statistics comparison is shown in table 4.3. Overall, the progression rate is slower (less negative) than the two major studies. The patient population is younger. Furthermore, pseudo-exfoliation glaucoma was not included in the Rotterdam dataset. This causes the major limitation of current work in terms of algorithm training and testing, because most patients in the dataset are not demonstrating progression. As a result, the “repeat” prediction has the same effectiveness as learning methods. On the other hand, it also shows that the learning methods successfully picked up this trend.

It is also found that the linear extrapolation methods, which is the current clinically adopted method, performed much worse than the other methods. Whether this suggests limitations in the linear extrapolation method for accurate prediction and slope estimation requires further investigation. Previous in fig. 2.3, it was also demonstrated that the exponential model, despite its potential physiological interpretation, performs even worse than linear extrapolation. The fact that all learning methods managed to perform better suggests there is a potential application for such algorithms in the glaucoma prediction task.

4.4 Future Work

The next step in the project schedule is to collect a glaucoma patient dataset from our own sources, upon which these and other methods will be evaluated again. Such a dataset should include more examples of glaucoma progression by including older patients, patients with exfoliation and other types of glaucoma, and a larger patient sample size.

Most current study datasets also do not include any patients with history of surgical intervention. However, patients with surgical intervention are also likely ones with the most severe progression rate and require the most attention for a carefully-determined treatment decision. It may be effective to also collect some patients with significant interventions, along with other potentially important details such as their medication history, family history, ethnicity, specific diagnosis, angle closure, etc., all of which are known to affect the trajectory of glaucoma progression. One may also collect optical coherence tomography (OCT) data from the patients who will be included in our study, though its usefulness is still an open question. The challenge then lies in designing an algorithm that can appropriately represent such a multi-model and multi-variate dataset.

Finally, another area of application of machine learning algorithm is glaucoma screening, specifically the classification of patients in a screening setting. The input to the algorithm will be one visual field produced from a screening setting, potentially with intraocular pressure (IOP) information which is also routinely examined, and the algorithm will label if the patient is healthy or a glaucoma suspect. The performance of such an algorithm will be primarily compared against using field indices, especially Glaucoma Hemifield Test (GHT) which is designed for a similar classification task.

Chapter 5

Conclusion

Glaucoma is often referred to as the "silent thief of vision" because its irreversible progression is unnoticeable until the very late stages without visual field testing. Therefore, medical doctors need to have the best information available as early as possible to determine the risks and benefits of different treatment approaches. It is widely accepted that 5–6 visual fields, or approximately 2 years, are required before a clinically useful progression rate statistic can be determined.

The motivation behind the current work is to investigate if a machine learning based approach may yield earlier and better clinical information for treatment decisions. Specifically, the Rotterdam public dataset is used to evaluate 5 different machine learning algorithms of different complexities, including linear models (ridge and Huber regression), neural network models (MLP and CNN), and a recurrent neural network (RNN) approach (CNN+LSTM). The motivation behind the CNN architecture is its ability to extract spatial features, while the motivation for the LSTM architecture is its ability to model sequential temporal inputs.

The different approaches are evaluated on two tasks: (1) predicting future Mean Deviation (MD) and (2) predicting future point-wise field thresholds using 3 input fields. It is found that all machine learning methods produce similar results on the Rotterdam dataset, with lower mean absolute error (MAE) in both tasks than using than linear extrapolation based on OLSLR, but do not outperform simple repetition of the last observed value. This is due to the fact that the Rotterdam dataset consists primarily of patients who are not progressing, and therefore a non-progressing prediction performed quite well. Therefore, future studies should be based on a more comprehensive dataset, including a significant number of progressing patients. This also illustrates the challenges associated with achieving both high sensitivity and specificity in the glaucoma prediction tasks in general.

Bibliography

[1] R. N. Weinreb and P. Tee Khaw, "Primary open-angle glaucoma," in *Lancet*, vol. 363, no. 9422, 2004, pp. 1711–1720.

[2] A. Chen, K. Nouri-Mahdavi, F. J. Otarola, F. Yu, A. A. Afifi, and J. Caprioli, "Models of glaucomatous visual field loss," *Investigative ophthalmology & visual science*, 2014.

[3] A. Heijl, G. Lindgren, and J. Olsson, "A package for the statistical analysis of visual fields," in *Documenta Ophthalmologica*. Springer, Dordrecht, 1987, vol. 49, pp. 153–168.

[4] ——, "Normal Variability of Static Perimetric Threshold Values Across the Central Visual Field," *Archives of Ophthalmology*, vol. 105, no. 11, pp. 1544–1549, nov 1987.

[5] B. Bengtsson and A. Heijl, "A Visual Field Index for Calculation of Glaucoma Rate of Progression," *American Journal of Ophthalmology*, 2008.

[6] P. Åsman and A. Heijl, "Glaucoma Hemifield Test: Automated Visual Field Evaluation," *Archives of Ophthalmology*, 1992.

[7] B. C. Chauhan, D. F. Garway-Heath, F. J. Goñi, L. Rossetti, B. Bengtsson, A. C. Viswanathan, and A. Heijl, "Practical recommendations for measuring rates of visual field change in glaucoma," 2008.

[8] M. Pathak, S. Demirel, and S. K. Gardiner, "Nonlinear, multilevel mixed-effects approach for modeling longitudinal standard automated perimetry data in glaucoma," *Investigative Ophthalmology and Visual Science*, 2013.

[9] A. A. Aref and D. L. Budenz, "Detecting Visual Field Progression," 2017.

[10] K. Nouri-Mahdavi, J. Caprioli, A. L. Coleman, D. Hoffman, and D. Gaasterland, "Pointwise Linear Regression for Evaluation of Visual Field Outcomes and Com-

parison With the Advanced Glaucoma Intervention Study Methods," Tech. Rep., 2005.

[11] H. Zhu, D. P. Crabb, T. Ho, and D. F. Garway-Heath, "More accurate modeling of visual field progression in glaucoma: ANSWERS," *Investigative Ophthalmology and Visual Science*, 2015.

[12] N. G. Strouthidis, A. Scott, A. C. Viswanathan, D. P. Crabb, and D. F. Garway-Heath, "Monitoring Glaucomatous Visual Field Progression: The Effect of a Novel Spatial Filter." [Online]. Available: https://iovs.arvojournals.org/

[13] R. Asaoka, H. Murata, A. Iwase, and M. Araie, "Detecting Preperimetric Glaucoma with Standard Automated Perimetry Using a Deep Learning Classifier," *Ophthalmology*, 2016.

[14] S. Yousefi, T. Kiwaki, Y. Zheng, H. Sugiura, R. Asaoka, H. Murata, H. Lemij, and K. Yamanishi, "Detection of Longitudinal Visual Field Progression in Glaucoma Using Machine Learning," *American Journal of Ophthalmology*, 2018.

[15] J. C. Wen, C. S. Lee, P. A. Keane, S. Xiao, Y. Wu, A. Rokem, P. P. Chen, and A. Y. Lee, "Forecasting future humphrey visual fields using deep learning," Tech. Rep. 206, 2018.

[16] N. N. Shah, C. Bowd, F. A. Medeiros, R. N. Weinreb, P. A. Sample, E. M. Hoffmann, and L. M. Zangwill, "Combining Structural and Functional Testing for Detection of Glaucoma," *Ophthalmology*, 2006.

[17] A. T. H. Lu, M. Wang, R. Varma, J. S. Schuman, D. S. Greenfield, S. D. Smith, and D. Huang, "Combining Nerve Fiber Layer Parameters to Optimize Glaucoma Diagnosis with Optical Coherence Tomography," *Ophthalmology*, 2008.

[18] J.-C. Mwanza, J. L. Warren, and D. L. Budenz, "Combining Spectral Domain Optical Coherence Tomography Structural Parameters for the Diagnosis of Glaucoma With Early Visual Field Loss," *Investigative Opthalmology & Visual Science*, 2013.

[19] F. R. Silva, V. G. Vidotti, F. Cremasco, M. Dias, E. S. Gomi, and V. P. Costa, "Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using spectral domain oct and standard automated perimetry," *Arquivos Brasileiros de Oftalmologia*, 2013.

[20] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, 2016.

[21] Y. Zhao, M. R. Kosorok, and D. Zeng, "Reinforcement learning design for cancer clinical trials," *Statistics in Medicine*, 2009.

[22] Y. Zhao, D. Zeng, M. A. Socinski, and M. R. Kosorok, "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer," *Biometrics*, 2011.

[23] D. Ernst, G.-B. Stan, J. Goncalves, and L. Wehenkel, "Clinical data based optimal STI strategies for HIV: a reinforcement learning approach," in *Proceedings of the 45th IEEE Conference on Decision and Control*, 2006.

[24] S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy, "Informing sequential clinical decision-making through reinforcement learning: An empirical study," *Machine Learning*, 2011.

[25] S. R. Bryan, K. A. Vermeer, P. H. Eilers, H. G. Lemij, and E. M. Lesaffre, "Robust and censored modeling and prediction of progression in glaucomatous visual fields," *Investigative Ophthalmology and Visual Science*, 2013.

[26] I. Marin-Franch and W. H. Swanson, "The visualFields package: A tool for analysis and visualization of visual fields," *Journal of Vision*, vol. 13, no. 4, pp. 10–10, mar 2013.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[29] P. A. Sample, K. Chan, C. Boden, T. W. Lee, E. Z. Blumenthal, R. N. Weinreb, A. Bernd, J. Pascual, J. Hao, T. Sejnowski, and M. H. Goldbaum, "Using unsupervised learning with variational bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects," *Investigative Ophthalmology and Visual Science*, vol. 45, no. 8, pp. 2596–2605, aug 2004.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[31] A. J. Anderson, "Estimating the true distribution of visual field progression rates in glaucoma," *Investigative Ophthalmology and Visual Science*, vol. 56, no. 3, pp. 1603–1608, 2015.

[32] A. Heijl, B. Bengtsson, L. Hyman, and M. C. Leske, "Natural History of Open-Angle Glaucoma," *Ophthalmology*, vol. 116, no. 12, pp. 2271–2276, dec 2009.

[33] C. G. S. Group, "Canadian Glaucoma Study: 3. Impact of Risk Factors and Intraocular Pressure Reduction on the Rates of Visual Field ChangeRisk Factors, IOP Reduction, Visual Field Change," *Archives of Ophthalmology*, vol. 128, no. 10, pp. 1249–1255, oct 2010. [Online]. Available: https://doi.org/10.1001/archophthalmol.2010.196

Appendices

Appendix A

Algorithms

Algorithm 1 CNN LSTM Forward Propogation

```
procedure CNNLSTMForward(input)
    $out \leftarrow 0$                                        ▷ Initialize LSTM
    $cell \leftarrow 0$
    $list \leftarrow [out]$
    for $input_i$ in all inputs do                             ▷ LSTM network calculation
        $(out, cell) \leftarrow \text{LSTM}(input_i, out, cell)$
        Append $out$ to $list$
    end for
    $\overline{out} \leftarrow \text{mean}(list)$                ▷ LSTM average pooling
    $out_{LSTM} \leftarrow \text{linear}(list)$                  ▷ LSTM fully connected layer

    $x_{down} \leftarrow \text{Conv2d}(input, \text{kernel} = (3,2), \text{padding} = (1,0))$    ▷ Down-convolution
    $x_0 \leftarrow \text{ReLU}(\text{BatchNormalization}(x_{down}))$
    $x_1 \leftarrow \text{ConvBlock}(x_0) + x_0$                 ▷ Residual Structure
    $x_2 \leftarrow \text{ConvBlock}(x_1) + x_0 + x_1$
    $x_3 \leftarrow \text{ConvBlock}(x_2) + x_0 + x_1 + x_2$
    $out_{CNN} \leftarrow \text{Conv2d}(x_3, \text{kernel} = (3,2), \text{padding} = (1,1))$    ▷ Up-convolution

    return $out_{LSTM} + out_{CNN}$
end procedure

procedure ConvBlock(x)
    $x \leftarrow \text{Conv2d}(x, \text{kernel} = (3,3), \text{padding} = (1,1))$
    $x \leftarrow \text{BatchNormalization}(\text{ReLU}(x))$
    $x \leftarrow \text{Conv2d}(x, \text{kernel} = (3, 3), \text{padding} = (1, 1))$
    $x \leftarrow \text{BatchNormalization}(\text{ReLU}(x))$
    $x \leftarrow \text{Conv2d}(x, \text{kernel} = (3, 3), \text{padding} = (1, 1))$
    $x \leftarrow \text{BatchNormalization}(\text{ReLU}(x))$
    $x \leftarrow \text{Dropout}(x)$
    return $x$
end procedure
```

Appendix B

Longitudinal Glaucomatous Visual Field Dataset

The Longitudinal Glaucomatous Visual Field dataset is publicly available online from Rotterdam Ophthalmic Institute. [25]

The following is an excerpt from the Rotterdam dataset information:

> The data set contains longitudinal visual field data of 139 glaucoma patients. These patients were recruited from the Rotterdam Eye Hospital (Rotterdam, The Netherlands). Informed consent was obtained from all subjects.
>
> Inclusion criteria:
>
> - Between 18 and 85 years at time of inclusion.
> - Glaucoma diagnosis: Two of the following conditions are met: pattern standard deviation significant at p=0.05, abnormal hemifield test result, cluster of ≤ 3 points depressed at p=0.05 level or 1 point at p=0.01. VF defects must be reproducible on at least one occasion.
> - Primary open-angle glaucoma, normal tension glaucoma, pigmentary glaucoma, and treated angle closure glaucoma. (Pseudo-exfoliation glaucoma was not included.)
>
> Exclusion criteria:
>
> - Secondary glaucomas except pigmentary.
> - Evidence of SAP VF abnormality consistent with other disease.
> - BCVA > 0.3 (LogMAR).
> - Refractive error outside -10.0 to +5.0 D range.

- Cataract surgery in previous 12 months.
- Previous refractive or vitreoretinal surgery.
- Evidence of diabetic retinopathy, diabetic macular edema, or other vitreoretinal disease.
- Previous keratoplastic surgery.
- Diabetes, leukemia, AIDS, uncontrolled systemic hypertension, multiple sclerosis or (other) life threatening disease.

Both eyes of each participant were included. Visits were scheduled every 6 months. At each visit, standard clinical ophthalmic examinations were performed, including visual acuity, intra-ocular pressure, gonioscopy and ophthalmoscopy. At each visit, standard automated perimetry was also performed. Visual fields were acquired on a Humphrey Visual Field Analyzer (Carl Zeiss Meditec) with a standard white-on-white 24-2 field with the full threshold program. The provided data set contains information on the visual field and on the individual visual field test locations.

Included data:

- Patient's gender
- Visit data (per eye): Patient's age at visit, IOP, MD.
- Local visual field data (54 points per field): Threshold, total deviation