VISUAL FIELD TESTING USING FORECASTING AND VIRTUAL REALITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EMRE BÜLBÜL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2022

Approval of the thesis:

VISUAL FIELD TESTING USING FORECASTING AND VIRTUAL REALITY

submitted by **EMRE BÜLBÜL** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** ________________

Prof. Dr. Gözde Bozdağı Akar
Supervisor, **Electrical and Electronics Engineering, METU** ________________

Examining Committee Members:

Prof. Dr. İlkay Ulusoy
Electrical and Electronics Engineering, METU ________________

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering, METU ________________

Prof. Dr. Abdullah Aydın Alatan
Electrical and Electronics Engineering, METU ________________

Assoc. Prof. Dr. Elif Sürer
Graduate School of Informatics, METU ________________

Assoc. Prof. Dr. Sertan Serte
Electrical and Electronics Engineering, NEU ________________

Date: 08.02.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Emre Bülbül

Signature :

ABSTRACT

VISUAL FIELD TESTING USING FORECASTING AND VIRTUAL REALITY

Bülbül, Emre
M.S., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. Gözde Bozdağı Akar

February 2022, 60 pages

The standard method for assessing a patient's visual field is visual field testing. Many diseases, including glaucoma, which affects more than 80 million people, require visual field testing for monitoring and diagnosis. The testing is done by sending light to fixed locations with different luminosities while the patient is fixating at a certain point, then the sensitivities to light at each location are determined by recording the responses of the patient to seen stimuli. Because of their form and digital screens, virtual reality headsets have lately begun to be utilized to conduct visual field examinations. However, since the testing duration is long, it causes fatigue in patients, which decreases cooperation and test accuracy. Also, it limits how many tests a clinic can conduct in a day. In this thesis, using a digital screen, the number of testable point locations is increased, and the effect of using an optimal subset of locations, which is found by employing a reinforcement learning method to decrease test duration, is investigated. Also, the effect of using forecasted future visual field test results in testing to the test duration is compared with standard testing strategies.

Keywords: Visual Field Testing, Virtual Reality, Perimetry, Forecasting, Reinforcement Learning, Deep Learning

ÖZ

ÖNGÖRÜ VE SANAL GERÇEKLİK KULLANILARAK GÖRME ALANI TESTİ

Bülbül, Emre
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Gözde Bozdağı Akar

Şubat 2022 , 60 sayfa

Bir hastanın görme alanını değerlendirmenin standart yöntemi görme alanı testidir. 80 milyondan fazla insanı etkileyen glokom dahil birçok hastalık, takip ve teşhis için görme alanı testi gerektirir. Hasta belirli bir noktaya sabitlenirken farklı parlaklıklara sahip sabit yerlere ışık gönderilerek test yapılır, daha sonra hastanın görülen uyaranlara verdiği tepkiler kaydedilerek her bir lokasyondaki ışığa duyarlılıkları belirlenir. Sanal gerçeklik gözlükleri, formları ve dijital ekranları nedeniyle son zamanlarda görme alanı testlerini gerçekleştirmek için kullanılmaya başlandı. Ancak test süresinin uzun olması hastalarda yorgunluğa neden olmakta, bu da kooperasyonu ve test doğruluğunu azaltmaktadır. Ayrıca, bir kliniğin bir günde kaç test yapabileceğini sınırlar. Bu tezde, bir dijital ekran kullanılarak, test edilebilir nokta konumlarının sayısı artırılmış ve takviyeli öğrenme yöntemi kullanılarak bulunan konumların optimal bir alt kümesinin kullanılmasının test süresini kısaltmaya etkisi araştırılmıştır. Ayrıca, testte öngörlen gelecek görme alanı test sonuçlarının kullanılmasının test süresine etkisi standart test stratejileri ile karşılaştırılır.

Anahtar Kelimeler: Görme Alanı Testi, Sanal Gerçeklik, Perimetre, Öngörü, Takviyeli Öğrenme, Derin Öğrenme

To my family.

ACKNOWLEDGMENTS

I'd would like to express my sincere gratitude to my thesis supervisor Prof. Dr. Gözde Bozdağı Akar for her invaluable support and guidance. Also, to great people at Oculera Healthcare Technologies for their technical consultancy.

Last but not least, I would like to thank my wife, family and friends for their constant support.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

VF	Visual Field
VFA	Visual Field Analyzer
VR	Virtual Reality
SAP	Static Automated Perimetry
TD	Total Deviation
PD	Pattern Deviation
PMAE	Point-wise Mean Absolute Error
RMSE	Root Mean Square Error
RL	Reinforcement Learning
NN	Neural Network

CHAPTER 1

INTRODUCTION

Perimetry or Visual Field Testing is a standard method for assessing a patient's visual field in ophthalmology and optometry. It measures the patient's visual function over their whole field of vision. Visual Field Analyzers (or Perimeters) are the equipment that is utilized to do this examination. Perimetry is used for a variety of purposes, including the diagnosis of pathologies, the assessment of disease state, the monitoring of pathologies over time to identify progression or disease stability, the evaluation of treatment efficacy, and the testing of visual capacity. The most common use case of Visual Field Analyzers is aiding the diagnosis of glaucoma, which is estimated that by the year 2040, it will affect more than 118 million people worldwide [17]. However, it can be used in the diagnosis of many other diseases like brain tumors and a variety of retinal diseases. Throughout this thesis, perimetry testing and visual field testing are used interchangeably.

1.1 Motivation and Problem Definition

Perimetry tests are vital tools for aiding the diagnosis and monitoring of countless diseases. However, each test can take from 3 minutes to 15 minutes per eye, depending on the type. This can be tiring for patients. Since perimetry is a subjective test that heavily depends on patient attention and cooperation, increasing test duration impacts the reliability of the results. This effect is more prominent in elderly patients, who are the majority of the patients that take perimetry tests.

Conventionally, perimetry test is conducted with automated machines that have mounted light bulbs inside a hemisphere that emit light in a given luminosity and duration determined by an algorithm. The lowest intensity of light that can be seen by the patient at a given point is saved as the threshold sensitivity of that location, and it is represented in decibels away from the maximum intensity.

With the developing technology, a new alternative to those conventional devices has emerged. The alternative consists of a proprietary or commercial Virtual Reality Headset and a software specifically tailored to the technical specifications of the device to conduct perimetry tests. Unlike the conventional devices where light bulbs are mounted in fixed positions, VR headsets have continuous, high-resolution screens.

These high-resolution screens give the ability to test any point on the human visual field, contrary to the fixed-positioned test points on conventional machines. Stimuli on conventional machines are presented using light bulbs. In this thesis, this ability is used to find new test points that have non-conventional positions to estimate the same visual field representation with less number of points is investigated. Therefore, by testing fewer number of locations, conducting shorter visual field tests is possible.

Another condition that affects test duration is the starting luminosity at the tested locations. The closer the starting point to the patient's threshold sensitivity, the faster the test duration. Recently, with the integration of convolutional filters, novel methods employing neural network models are very successful in forecasting feature VF results by incorporating one or more past VF test results. This thesis compares the improvement in duration by simulating a benchmark perimetry test (e.g., Full Threshold) with a forecasted VF test result.

1.2 Contribution of the Thesis

In this thesis, two different ways of decreasing VF tests are proposed. The first one is by decreasing the number of test locations. The second one is by adjusting starting luminosities of test locations.

To decrease test duration by reducing the number of test locations, a reinforcement learning method that extracts new and fewer points by using the continuousness of a VR headset screen is proposed. Only a single VF test is needed from a patient for this

(a) Humphrey Visual Field Analyzer [1].

(b) Oculera Visual Field Analyzer [14].

Figure 1.1: Two different technologies of Visual Field Analyzers. a. Conventional. b. VR Headset based.

method. RL agent explores new VF test positions between the conventional points to check if multiple points can be represented by a single one in the area of replaced points.

To decrease test duration by adjusting starting luminosities of test locations, the output of a forecasting neural network is used. The CascadeNet neural network model is trained using the Rotterdam VF dataset [5]. The output of the forecasting network is used as the starting luminosities. The test duration is simulated and compared with Full Threshold testing strategy with starting luminosities as age-normal values [3], previous VF test, and forecasted values.

The structure of the thesis is as follows; in Chapter 2 the background information about the common concepts of the thesis is given and in Chapter 3, literature review is detailed. In Chapter 4, the methods that are proposed and used in simulations and testing is documented and in Chapter 5, the results are presented and discussed. Finally, Chapter 6 concludes the thesis.

CHAPTER 2

BACKGROUND INFORMATION

2.1 Human Visual Field

A person's visual field is defined as the area in which he or she can see at any given time relative to the direction of fixation, without moving the head or eyes (i.e., it specifies the bounds of the area beyond which nothing can be seen). The extent of one's visual field is an important aspect of one's visual function since a restricted visual field has a major detrimental influence on everyday activities, and as a result, quality of life.

In the visual field of each eye there is an area that is called the blind spot where there is no vision. The position of the blind spot corresponds to the optic nerve head's location where there are no photoreceptors. From the center of the eye, the blind spot is 1.5° below the horizontal line and 12°-15° to the right for the right eye and vice versa in the left eye.

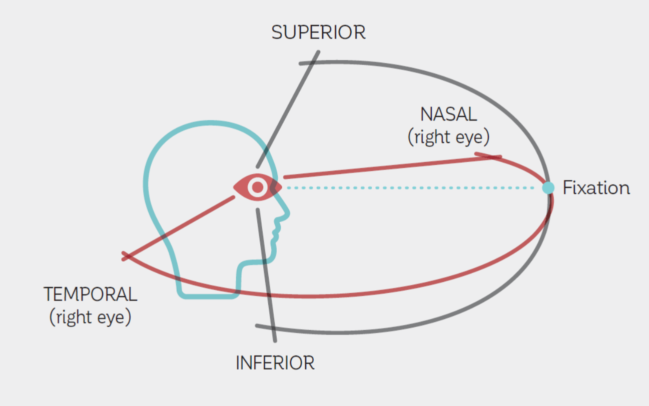


Figure 2.1: Monocular visual field. [18]

2.2 Sensitivity To Light In the Visual Field

To represent the human visual accurately, it is not enough to just test where a patient sees but to test how a patient sees is also necessary. To describe how a patient sees at a given moment sensitivity to light is measured in decibels with the formula below,

$$Threshold\,Value = 10 * log(L_{max}/L) \tag{2.1}$$

L_{max} is the maximum luminance that can be displayed by the VFA and L is the luminance of the stimulus, and the resulting threshold value is the sensitivity threshold at the given point.

Most VFAs have a representation range of 0 to 50.

2.3 Test Strategies and Test Patterns

There are many different perimetry testing strategies and test patterns. The most common test patterns are 24-2 and 30-2. The first number represents the degree of visual field testing from the center to a single corner, meaning 60° of the testing area in an axis. The second point represents two extra points placed in the blind spot. 30-2 pattern tests 76 locations and 24-2 tests 54 locations, proportionally changing test duration with the number of tested points.

Each perimetry testing strategy, in its essence, is a ladder algorithm, changing direction with the response of the patient. The Full Threshold algorithm, which is a staircase strategy, is explained in Algorithm 1.

2.4 Representation of Test Results

Perimetry test thresholds for SAP are presented in a single page layout per eye. The threshold values are shown in a 2D mapping where points are placed relative to the positions they are shown during the test. All other metrics are derived from threshold

Algorithm 1 Full Threshold

```
while There are non-tested locations do
    Pick a location
    if Never tested before then
        Test the location and record if seen or not
        Record last tested luminosity value in dB
    else if First reversal then
        if Last answer is seen then
            Test last luminosity + 4bB
        else
            Test last luminosity - 4bB
        if Answer is not same as the last answer then
            Mark as Second Reversal
    else if Second Reversal then
        if Last answer is seen then
            Test last luminosity - 2bB
        else
            Test last luminosity + 2bB
        Mean of the last positive and first negative patient's reply is recorded as threshold value.
        Mark as completed
```

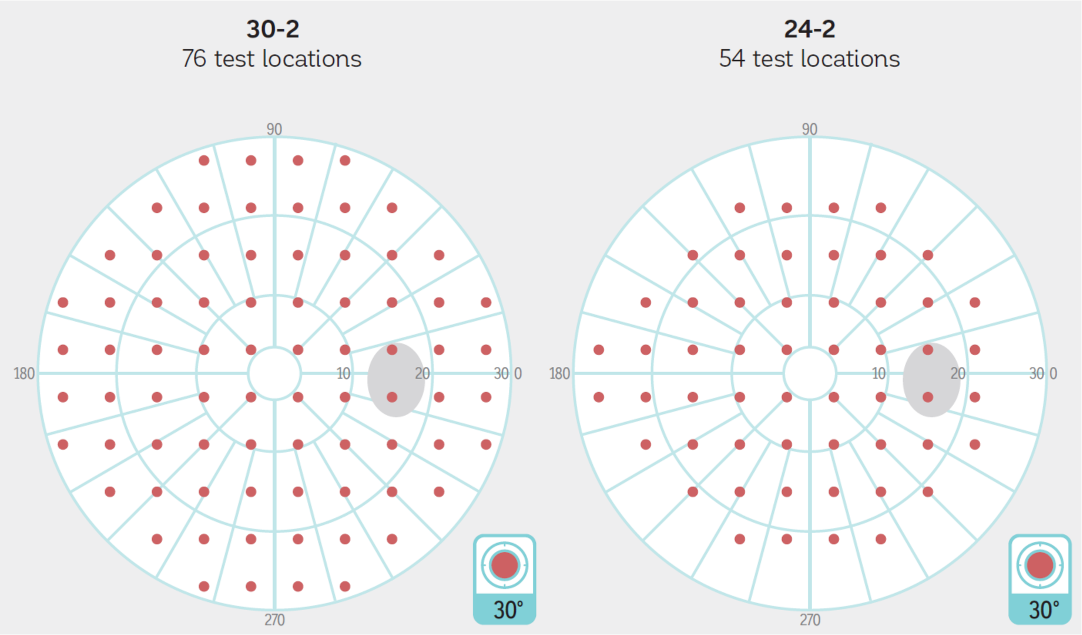
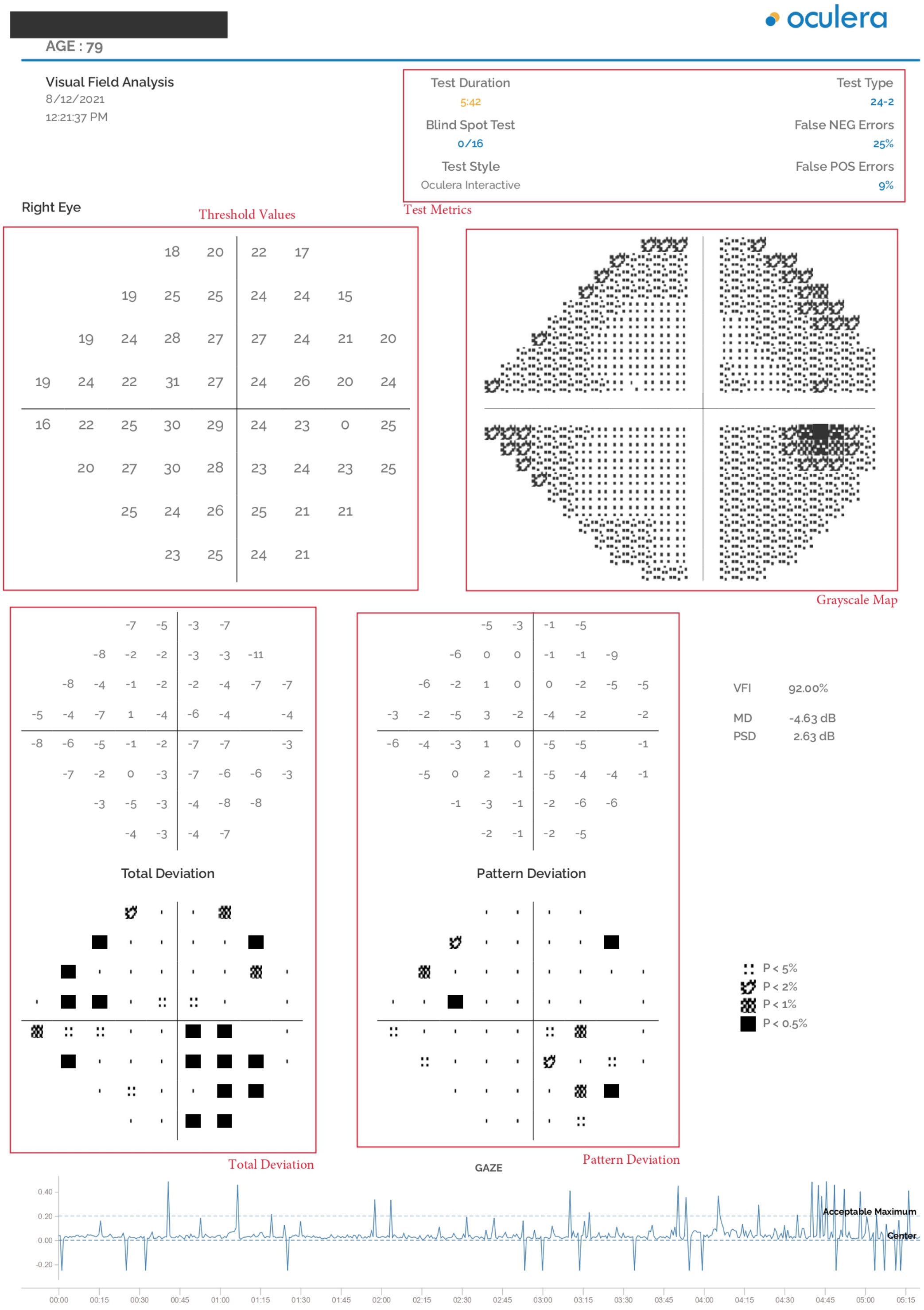


Figure 2.2: 24-2 and 30-2 test patterns. [18]

values and the patient's age. Each VFA has its own result sheet and way of displaying results. Figure 2.3 shows an example VFA result sheet.

Threshold values are the resulting light sensitivities of the eye in decibel scale. The grayscale map is only a graphical representation to easily identify defects and depressed areas. Dark portions represent the lower sensitivities, while lighter portions represent higher sensitivities. Total deviation gives the difference between normal values of the population categorized by age and the measured values. Pattern deviation, on the other hand, removes the measurement losses that can be present because of cataracts, uncorrected refractive error, and other unrelated eye conditions. Therefore, local problems are highlighted in this map, and because of this, it is the one that used in diagnosis mostly. The smaller maps right below the TD and PD maps are statistical maps showing the probability of seeing such a result in the population of the same age.



[redacted]

AGE : 79

oculera

Visual Field Analysis
8/12/2021
12:21:37 PM

Test Duration 5:42	Test Type 24-2
Blind Spot Test 0/16	False NEG Errors 25%
Test Style Oculera Interactive	False POS Errors 9%

Test Metrics

Right Eye

Threshold Values

			18	20	22	17		
		19	25	25	24	24	15	
	19	24	28	27	27	24	21	20
19	24	22	31	27	24	26	20	24
16	22	25	30	29	24	23	0	25
	20	27	30	28	23	24	23	25
		25	24	26	25	21	21	
			23	25	24	21		

Grayscale Map

			-7	-5	-3	-7		
		-8	-2	-2	-3	-3	-11	
	-8	-4	-1	-2	-2	-4	-7	-7
-5	-4	-7	1	-4	-6	-4		-4
-8	-6	-5	-1	-2	-7	-7		-3
	-7	-2	0	-3	-7	-6	-6	-3
		-3	-5	-3	-4	-8	-8	
			-4	-3	-4	-7		

Total Deviation

			-5	-3	-1	-5		
		-6	0	0	-1	-1	-9	
	-6	-2	1	0	0	-2	-5	-5
-3	-2	-5	3	-2	-4	-2		-2
-6	-4	-3	1	0	-5	-5		-1
	-5	0	2	-1	-5	-4	-4	-1
		-1	-3	-1	-2	-6	-6	
			-2	-1	-2	-5		

Pattern Deviation

VFI	92.00%
MD	-4.63 dB
PSD	2.63 dB

∷ P < 5%
✿ P < 2%
▦ P < 1%
■ P < 0.5%

Total Deviation

Pattern Deviation

GAZE

Figure 2.3: Example VFA result sheet.

2.5 Reinforcement Learning Methods

2.5.1 Maskable Proximal Policy Optimization

As the reinforcement learning method the Maskable PPO [21] is used. Maskable PPO is a version of the proximal policy optimization [19]. It implements an action mask that indicates if an action is valid or not. In bigger action spaces like those used here, eliminating invalid actions speeds up the process and increases accuracy.

The main contributions of the PPO algorithm are; the Clipped Surrogate Objective and for each policy update, performing multiple epochs of gradient ascent. The vanilla policy gradient approach tracks the impact of actions using the log probability of the action. The Clipped Surrogate Objective is a simple drop-in substitute for the vanilla policy gradient. The clipping restricts the effective change that can be made at each step to increase stability, and the minimization allows for error correction if the error becomes too significant.

Unlike traditional policy gradient approaches, PPO allows for numerous epochs of gradient ascent on samples without causing destructively large policy changes due to the Clipped Surrogate Objective function. This helps us to extract more information from the data while also reducing sample inefficiency. PPO performs the policy with N parallel actors, each gathering data, and then using the Clipped Surrogate Objective function, it samples mini-batches of this data to train for K epochs. Furthermore, in practically all continuous control environments, PPO outperforms prior approaches such as A2C, A3C, and Vanilla PG.

2.6 Forecasting Models

2.6.1 CascadeNet

CascadeNet [13] is the main neural network model that is used in this study. Moreover, it is the most performant compared to the other networks that have been trained. CascadeNet, in its essence, is a modified ResNet [7] model. It introduces the cascade block concept. Cascade blocks are internal convolutional paths to forward knowledge from the middle and previous ones. The number of convolutional layers is denoted by adding a number at the end of the model name. For example, CascadeNet with five convolutional layers is called CascadeNet-5. Figure 2.4 shows a part of the cascade architecture.

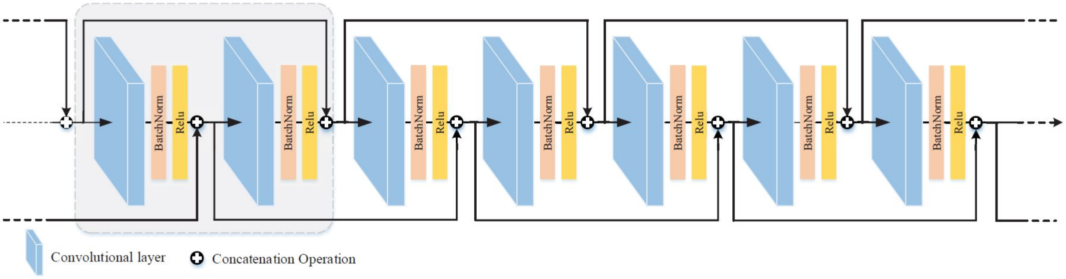


Figure 2.4: A sample partial representation of CascadeNet model architecture. [13]

CHAPTER 3

LITERATURE REVIEW

3.1 Decreasing Visual Field Test Duration

The duration of a Visual Field Test has always been a concern for the patients and the doctors since the duration of the test affects the reliability of the test due to patient fatigue. Also, the duration affects patient motivation and keeps them from taking their periodic examinations. Because of the aforementioned reasons, researchers have always been looking for different methods attacking different aspects of the VF test to decrease test duration. For example, decreasing the number of stimuli presented, decreasing the number of locations tested, decreasing inter-presentation times of stimuli, and many others.

One of the main focus points for decreasing test duration in this research is using sparse observation of points. There are similar studies that try using sparse observations to decrease test time. For example, authors in [12] used a training set to determine which locations are more prominent to reconstruct the entire VF from sparse observations and used the resulting weights in new examinations for the reconstruction task. This method is called Sequentially Optimized Reconstruction Strategy (SORS). It is shown that healthy eyes need fewer points for accurate testing. On the other hand, glaucomatous eyes needed more points for similar accuracy, and while the number of tested locations increased, the accuracy also increased in both cases. This relation can be seen in Figure 3.1.

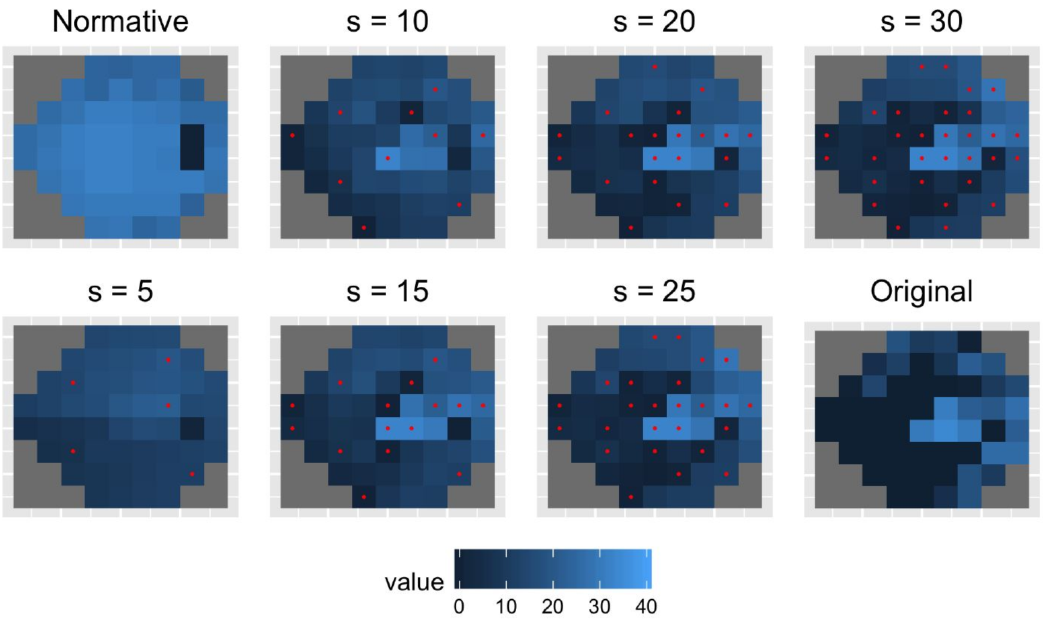


Figure 3.1: A comparison of the number of tested locations and resulting VF representations for SORS. [12]

In [11], while taking SORS as a basis, researchers developed a patient-attentive sequential strategy (PASS) for VF testing. PASS is a RL based method. PASS uses two independent networks for VF testing, a policy network and a reconstruction network. In its simplest terms, reconstruction network reconstructs VFs from sparse observations and the policy network decides which location to test after observing the reconstructed VF.

3.2 Forecasting Future Visual Field Results

Even though there are a lot of studies that examine visual field progression [26] [27] [10] [15], there are a handful of studies that forecast the progression. Authors in [20] also predict visual field progression they use a 3 previous tests to forecast 3 visual tests for different periods in the future and use those results for prediction.

In [25] researchers use nearly 25,000 VF tests to train a CascadeNet-5 neural network model to predict future visual field tests for 1 to 5.5 years from a single test. It differs from other methods because it only uses a single VF test and predicts with a PMAE

Table 3.1: Different NN models that used in [25]

	Layers (n)	Batch Size	Optimizer with learning rate	Space Complexity (MB)	Parameters
FullyConnected	2	32	Adam1x10-3	4	336,968
FullBN-3	10	32	Adam1x10-3	23	1,921,795
FullBN-5	16	32	Adam1x10-3	40	3,472,771
FullBN-7	22	32	Adam1x10-3	58	5,023,747
Residual-3	12	32	Adam1x10-3	27	2,332,163
Residual-5	18	32	Adam1x10-3	45	3,883,139
Residual-7	24	32	Adam1x10-3	63	5,434,115
Cascade-3	10	32	Adam1x10-3	81	6,992,086
Cascade-5	16	32	Adam1x10-3	238	20,694,754

of 2.47dB. The equation 3.1 shows the PMAE calculation, x_i representing a point that belongs to the forecasted VF for location i and y_i representing a point that belongs to the actual result for location i. n represents the number of point locations for the VF. Figure 3.2 shows example results of the aforementioned method for different time periods.

Also, authors compared 9 different NN models for the forecasting task. Cascade-5 was the best performing one between them and it has the highest number of trainable parameters. Different NN models compared can be seen in Table 3.1. Results are presented in Chapter 5.

$$PMAE = \frac{1}{n}\Sigma_{i=1}^{n}|x_i - y_i| \tag{3.1}$$

Long Short-Term Memory (LSTM)[9] is a type of recurrent neural network (RNN) that has feedback connections on previous neurons on the network. Because of this ability, LSTM can store memory type information from previous states. Moreover, LSTM can be used with time-series data. A different method is employed for forecasting in [16] with LSTM. Previous 5 VF tests are given as input to a 6-cell LSTM network last cell only containing the duration to the forecasted test. The duration between tests are not fixed, and given to the model in days as input. Other cells contain the TD values, reliability metrics and time displacement values. Figure 3.3 shows the LSTM architecture being used. The method outperformed Point-wise Linear Regres-

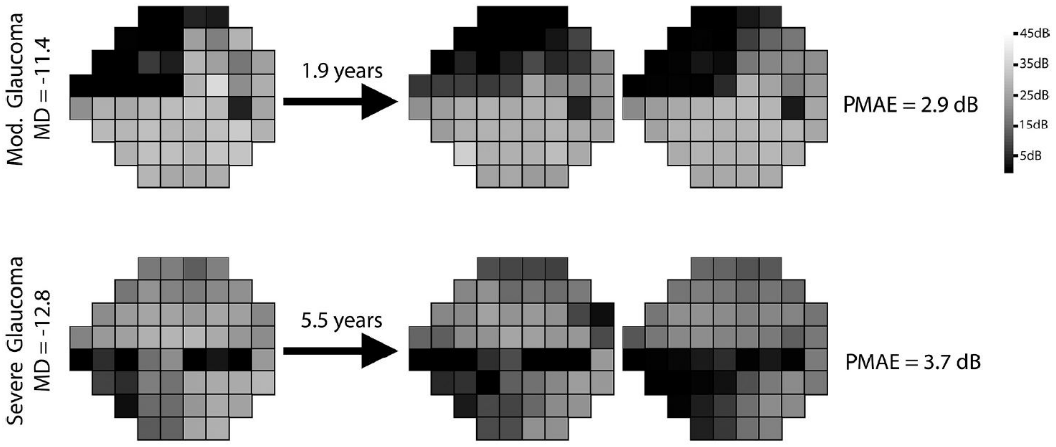


Figure 3.2: Two example results of the model in [25] for different glaucoma types and time periods.

sion models. However, it still requires more number of previous VF tests than other methods.

A second class of methods can be named as Point-wise Linear Regression models. The method in [26] requires 5 VF tests and minimum of 10 months of prediction duration. Authors, while outperforming previous linear methods, took test-retest variability into consideration. This test-retest variability is found by testing patients in periods which no progression is possible. For example, 10 tests in 10 weeks. This measurements provided a baseline for prediction accuracy, which showed a possible of 3.08dB mean error. In [6] authors used least absolute shrinkage and selection operator regression to decrease required number of VF tests from 5 to 3. Comparing different linear regression techniques, authors concluded that, with Lasso regression, prediction error varied between 2.0 and 2.2dB with 3 past VFs. Time between each test was nearly a year.

The first class of methods, different from the second class, can predict possible non-linear progression, since they use NN based methods. The limitation of linear progression prediction is discussed in [26].

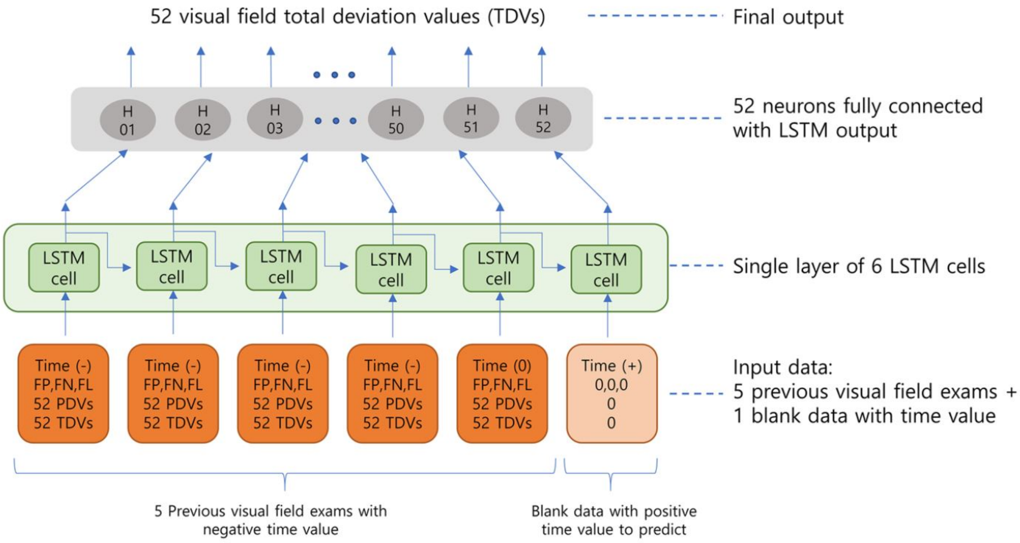

52 visual field total deviation values (TDVs)
Final output
H 01
H 02
H 03
H 50
H 51
H 52
52 neurons fully connected with LSTM output
LSTM cell
LSTM cell
LSTM cell
LSTM cell
LSTM cell
LSTM cell
Single layer of 6 LSTM cells
Time (-) FP,FN,FL 52 PDVs 52 TDVs
Time (-) FP,FN,FL 52 PDVs 52 TDVs
Time (-) FP,FN,FL 52 PDVs 52 TDVs
Time (-) FP,FN,FL 52 PDVs 52 TDVs
Time (0) FP,FN,FL 52 PDVs 52 TDVs
Time (+) 0,0,0 0 0
Input data:
5 previous visual field exams +
1 blank data with time value
5 Previous visual field exams with negative time value
Blank data with positive time value to predict

Figure 3.3: LSTM model in [16].

CHAPTER 4

PROPOSED METHODS

4.1 Increasing Available Test Locations by Using VR Headset Screens to Test Fewer Locations

VR headset screens are high resolution and this can be used in favor of Visual Field Testing. Proposed method here is, basically, showing stimuli in locations that normally fall between the conventional locations of a 24-2 VF test to represent those points with only a one single point. This is because, if there are similar valued points in the same area, all the points tested in that will likely yield similar values. The question here is whether a reinforcement learning agent learns to find those points. To achieve this, firstly, a VF is needed to be represented in an image form. Then it is assumed that each location in VF has a polynomial relationship with its neighboring points. With that assumption, by up-scaling the VF image representation, label values for training agent is generated. Now, RL agent's job is to find optimal locations to achieve the same VF by observing fewer points.

4.1.1 Interpolation of Missing Points

Originally, VF test points are located at specific locations in the hemisphere of a conventional VFA represented by angle values in x and y. However, for a VR based VFA, point locations are projected to a flat screen but, for the sake of this thesis, the point locations will be referred by their angle positions. The X and Y values of point locations can only be one of the unique values in,

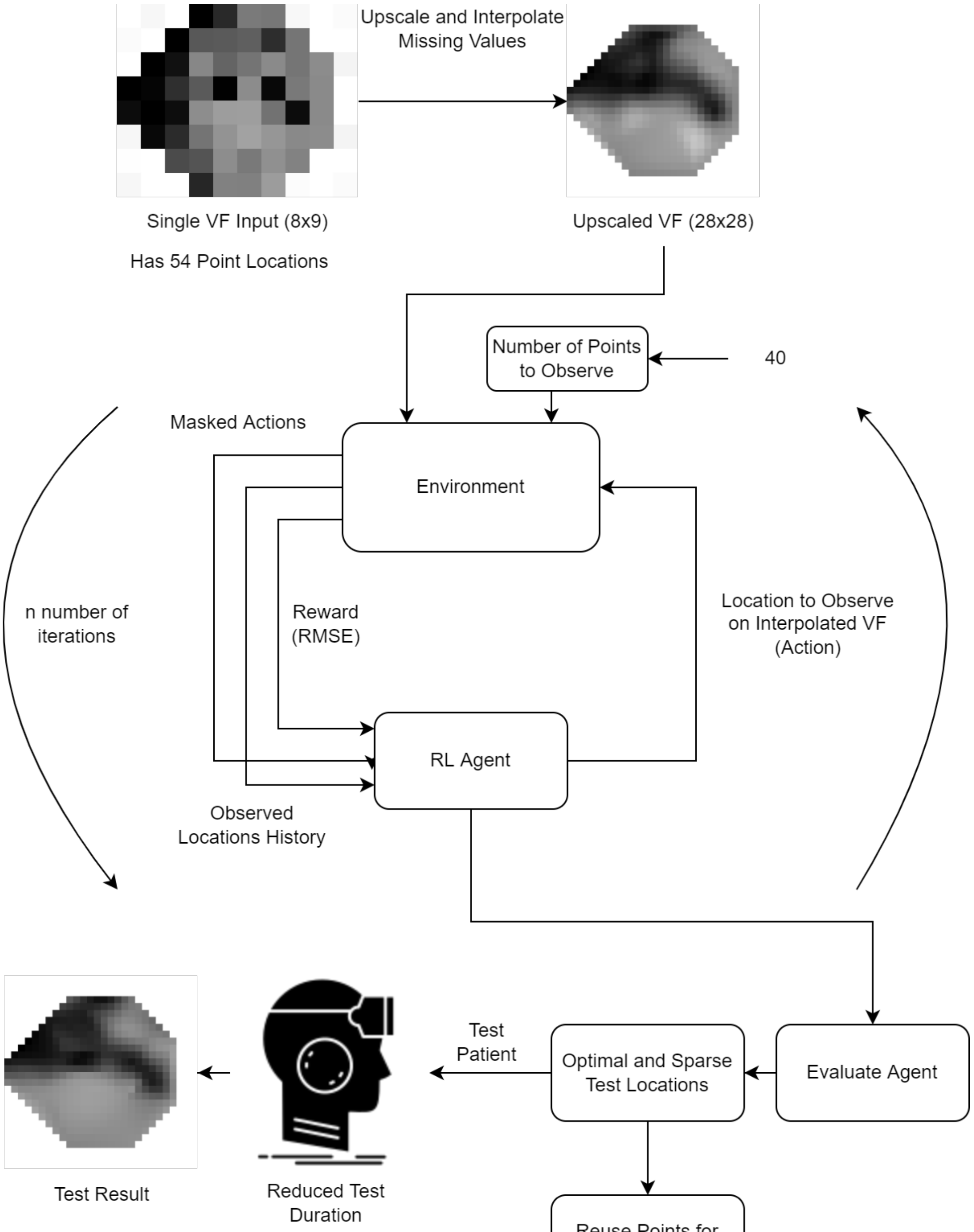
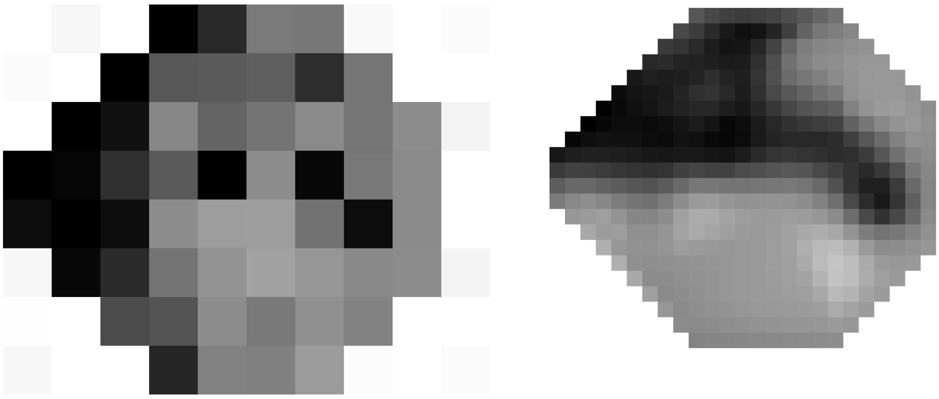

Upscale and Interpolate Missing Values
Single VF Input (8x9)
Has 54 Point Locations
Upscaled VF (28x28)
Number of Points to Observe
40
Masked Actions
Environment
n number of iterations
Reward (RMSE)
Location to Observe on Interpolated VF (Action)
RL Agent
Observed Locations History
Test Patient
Optimal and Sparse Test Locations
Evaluate Agent
Test Result
Reduced Test Duration
Reuse Points for Future Tests

Figure 4.1: Flow diagram of the method.



(a) Sample VF result represented as an image. (b) A VF result up-scaled to 28x28 by cubic interpolation

Figure 4.2: Sample VF result as an image and the up-scaled one with cubic interpolation.

$$A = [-27, -21, -15, -9, -3, 3, 9, 15, 21, 27] \tag{4.1}$$

for the 24-2 test pattern. The VF is represented as a 2 dimensional array like in the Figure 4.6a. Positions of points are the index of the angle value in A for the x axis and the inverse for y axis. Resulting matrix is a 10x10 matrix. For this method, only the right eye's VF result shape is used, and left VF representations are flipped. For the right eye pattern, there is no point in the 27 degree, and in the 24-2 pattern there is no point in -27 and 27 degrees at the y axis, so it can be represented as a 8x9 matrix. Untested points are represented by a -1 value.

$$X = \begin{bmatrix} -1 & -1 & -1 & 0 & 10 & 24 & 22 & -1 & -1 \\ -1 & -1 & 0 & 18 & 16 & 20 & 11 & 21 & -1 \\ -1 & 0 & 3 & 23 & 24 & 22 & 27 & 24 & 26 \\ 0 & 0 & 10 & 19 & 0 & 28 & 0 & 25 & 29 \\ 0 & 0 & 0 & 27 & 32 & 31 & 27 & 0 & 28 \\ -1 & 0 & 10 & 27 & 29 & 30 & 31 & 28 & 28 \\ -1 & -1 & 18 & 15 & 24 & 27 & 31 & 26 & -1 \\ -1 & -1 & -1 & 8 & 27 & 26 & 28 & -1 & -1 \end{bmatrix} \tag{4.2}$$

However, we want to use the continuous nature of a digital screen. Therefore, more point locations are needed. Even though, virtually any angle value can be used, within the limits of the specific VR headset's capabilities of course, for the sake of simplicity, by adding 2 more angle locations between each available angle the scaled angle array A' is generated.

$$\begin{aligned} A' = [&-27, -25, -23, -21, -19, -17, -15, -13, -11, -9, -7, \\ &-5, -3, -1, 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27] \end{aligned}$$

Again, to construct a matrix the index of angle values in *A'* is used for x axis and the inverse for y axis. This results in a 28x28 matrix. Since, only the 54 points coming from the 24-2 test pattern is filled, this matrix is very sparse. Figure 4.3 shows the locations of points that have a value other than -1.

To use the matrices for training purposes, we need to fill the empty cells. The purpose for this operation will be clear in the following chapters of the thesis. To fill the empty cells, multivariate interpolation techniques are used. For interpolation, a python package called *scipy* [23] is used. Possible options for multivariate interpolation can be seen in Figure 4.4.

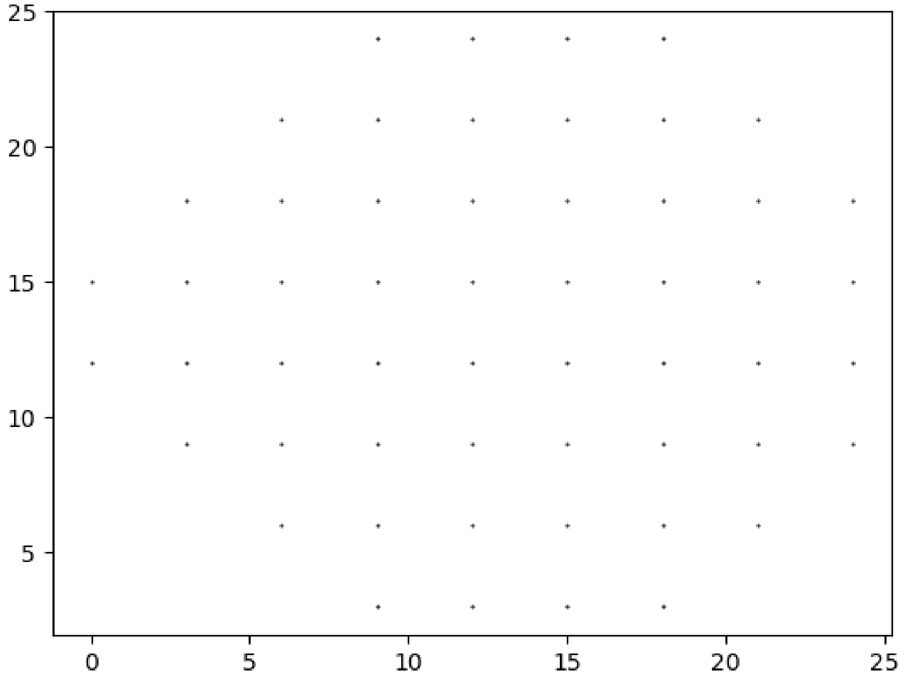


Figure 4.3: Point locations of 24-2 shown on the sparse up-scaled matrix sized 28x28.

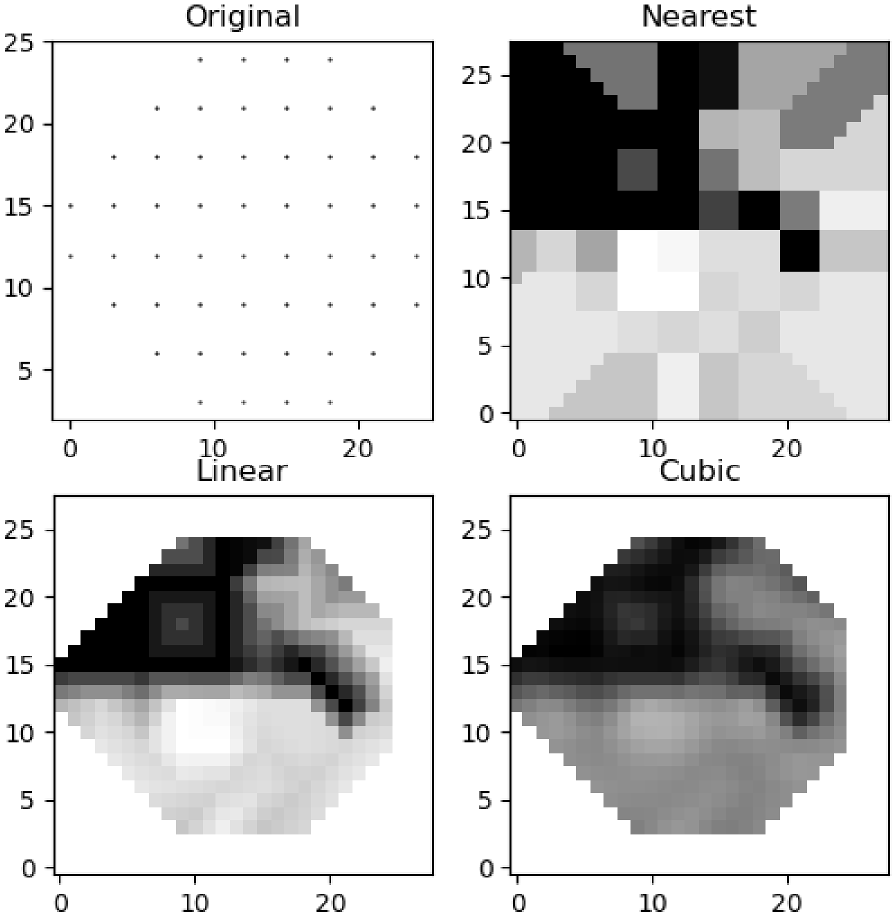


Figure 4.4: Possible options for multivariate interpolation.

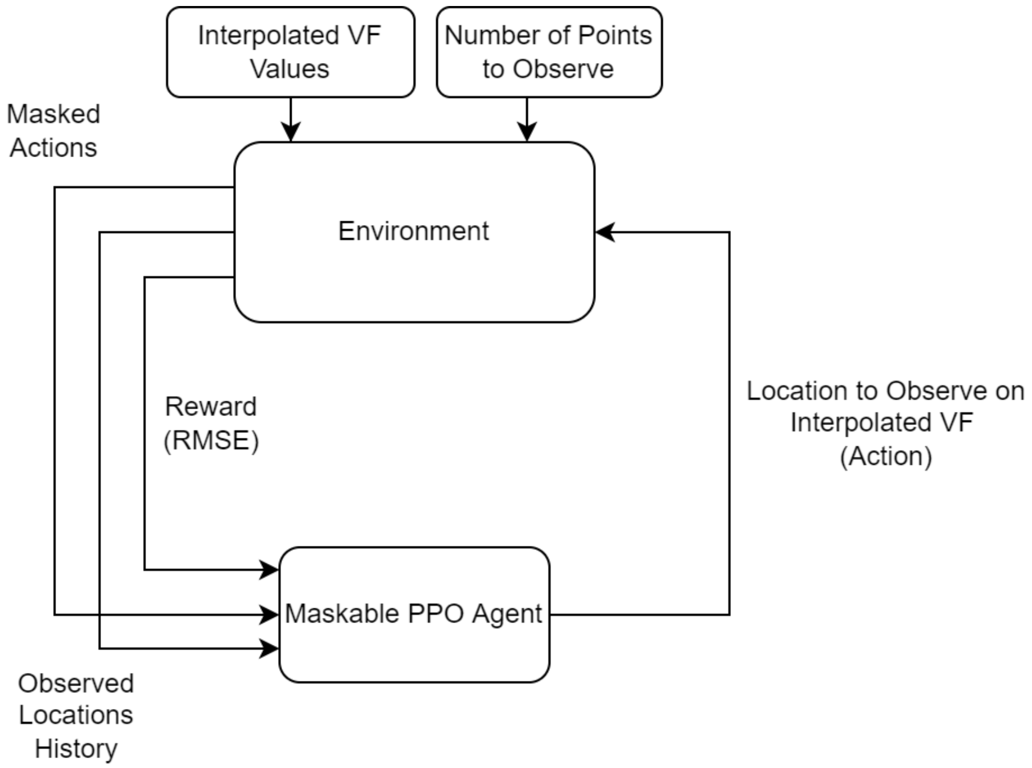

Interpolated VF Values
Number of Points to Observe
Masked Actions
Environment
Reward (RMSE)
Location to Observe on Interpolated VF (Action)
Maskable PPO Agent
Observed Locations History

Figure 4.5: Interactions of Maskable PPO Agent with the RL environment.

4.1.2 Training the Reinforcement Learning Agent

After the interpolation is done, the RL agent is trained to find the optimal points for estimating the values of 54 points of the 24-2 test pattern, but with fewer points. The number of points to observe is supplied to the agent at the beginning of the test.

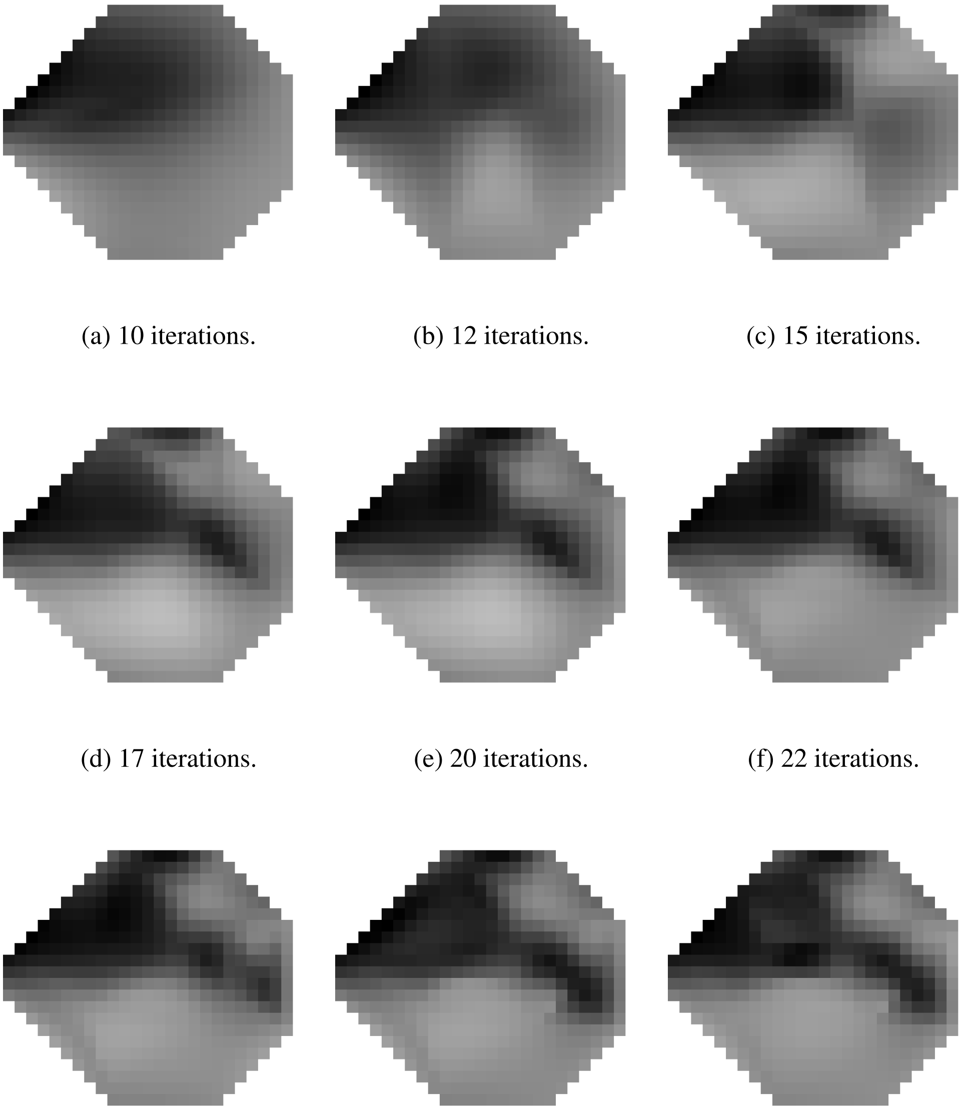
For each patient, a single VF test is supplied to the environment. While training, agent selects a location on the 28x28 matrix, meaning there are 784 available actions. After each action, environment interpolates the missing points and extracts the points that correspond to the original ones at the 24-2 test pattern. The difference between the extracted points on the newly interpolated map, and the original points are used to calculate RMSE as the reward. Equation 4.4 shows the RMSE calculation, x_i representing a point that belongs to the newly interpolated VF map at location i and y_i representing a point that belongs to the actual result at location i. n is the number of locations. Used action is added to the observations array and flagged at the masked actions array. Masked actions array is a boolean array with available actions as True and unavailable actions as False.

$$Mask = [True, True, False, ..., True, False, False] \tag{4.3}$$

After each step, action mask, observed locations history and reward is supplied to the agent. After the training reached to a reasonable RMSE value, training is stopped and a validation iteration is run. The output of the final run is the optimized test locations. Figure 4.6 shows the iterations of the algorithm while constructing the VF from the selected points.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - y_i)^2} \tag{4.4}$$

This method in its center assumes that the locations on the eye have a polynomial relationship between them. This assumption is exploited in other studies such as [12]. However, those studies have all used the standard 54 points of the 24-2 pattern, but here we are using locations not on but between those points.



(a) 10 iterations.

(b) 12 iterations.

(c) 15 iterations.

(d) 17 iterations.

(e) 20 iterations.

(f) 22 iterations.

(g) 25 iterations.

(h) 30 iterations.

(i) 35 iterations.

Figure 4.6: Iterations of agent estimating the VF map.

4.1.3 Gamified Environment

One of the major problems in this method is the action space size. Having a large action space increases the time to converge to an optimal solution. To overcome this bottleneck, a new type of environment is developed. Contrary to the previous method, which the actions are the locations on the matrix, we reduced the action space by

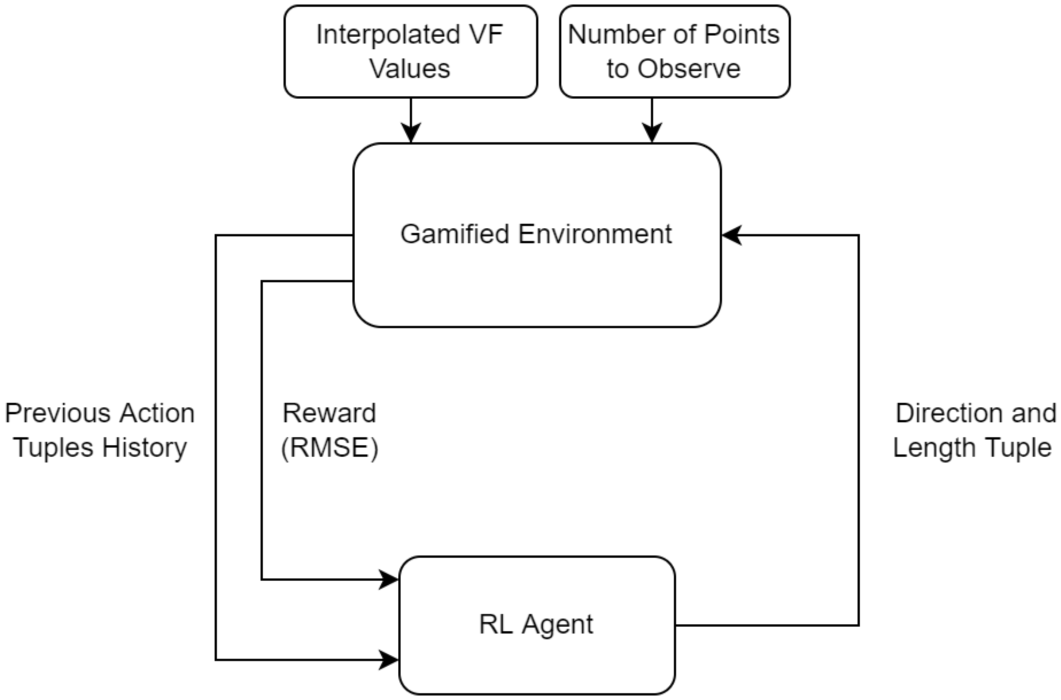

Interpolated VF Values
Number of Points to Observe
Gamified Environment
Previous Action Tuples History
Reward (RMSE)
Direction and Length Tuple
RL Agent

Figure 4.7: Interactions of the RL Agent with the gamified environment.

gamifying it.

What we mean by gamifying is the way the agent executes actions is changed. Rather than picking locations on the matrix, the agent is now deciding which direction to go and how many cells to go. Just like a maze-solving game. Agent, at each turn, chooses the direction and length tuple, and if the tuple is valid, meaning the destined location is not outside the grid, the destined location is marked as the next selected point. Agent always starts at the location nearest to the blind spot since it is the most significant point on most of the VFs due to it having the lowest luminosity threshold in most cases. Therefore, starting the agent at the blind spot location gives it a big starting advantage. Length of cells agent can go at once is limited to 8. Figure 4.7 show the interaction of the agent with the gamified environment.

However, gamified environment has more exception cases since it is more complicated than just selecting a number as action. One additional check is, as mentioned before, checking the validity of the action, meaning the resulting point is inside the matrix. Another implemented additional validation is checking whether *n* number of consecutive actions result in the same axis, then mark the action as invalid. This is

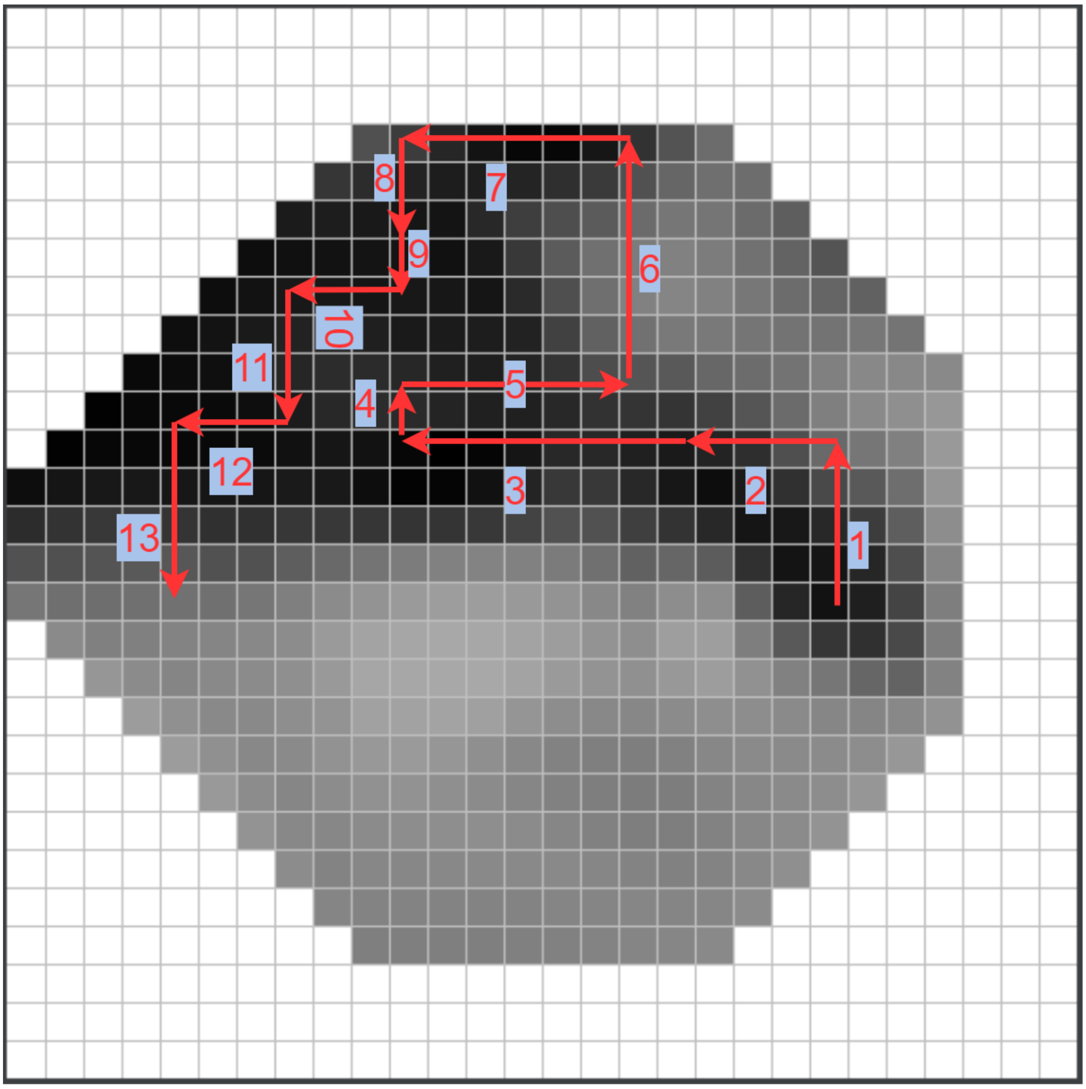


Figure 4.8: A partial sample run of the gamified agent.

because, too many same axis actions will result in a sub-optimal solution.

A sample run of the gamified agent can be seen in Figure 4.8.

Even though the gamified environment is a good improvement on paper to reduce action space and reduce conversion time, it actually limits the agent. Due to the fact that the agent can not go cross and more than 8 cells at a time, it finds sub-optimal solutions. There are cases in which selecting a point right across the matrix is the action that reduces the PMAE the most at that step. The comparison of the gamified environment and the former one is quantified in the Results section.

4.2 Using ML Based Forecasting Techniques to Decrease Test Duration

Another way of speeding up the VF testing is to adjust starting luminosities. This method will have significant impact on some algorithms and insignificant impact for others. This is because, some algorithms dynamically update starting values during testing, others not. Nevertheless, a good starting values set can be used to speed up most of the tests in a direct or indirect way. In this section, the impact on forecasting to find better starting points is investigated.

To use the forecasted VF values to speed up the test, first we need to forecast the VF values according to the elapsed time after the last test.

To forecast VF values, the model and method proposed in [25] is used, which is CascadeNet-5, since the authors had a very promising result with just a single test to predict future VF values. However, the scale of the dataset they used is significantly bigger than ours. Getting the same results would be a challenge but, getting results nearly accurate will also be useful to decrease test time.

Input to the NN is a tensor with shape 8x9x2, consisting of a 8x9 matrix filled with threshold sensitivity values of the patient as the first channel, and the second channel is again a matrix with values all filled with the age of the patient.

Output channel is the forecasted values in the shape of the VF map which is 8x9.

4.2.1 Feature Sets Used on Forecasting

Different from the dataset used in [25], the dataset used in this thesis has other features different from threshold values which are; intraocular pressure (IOP) and total deviation values for each point location. Total deviation values are derived from threshold values by subtracting the age normal values. Total deviation values show the deviation from the age normal values.

Different feature sets used to train the NN are;

- Threshold Values, Age

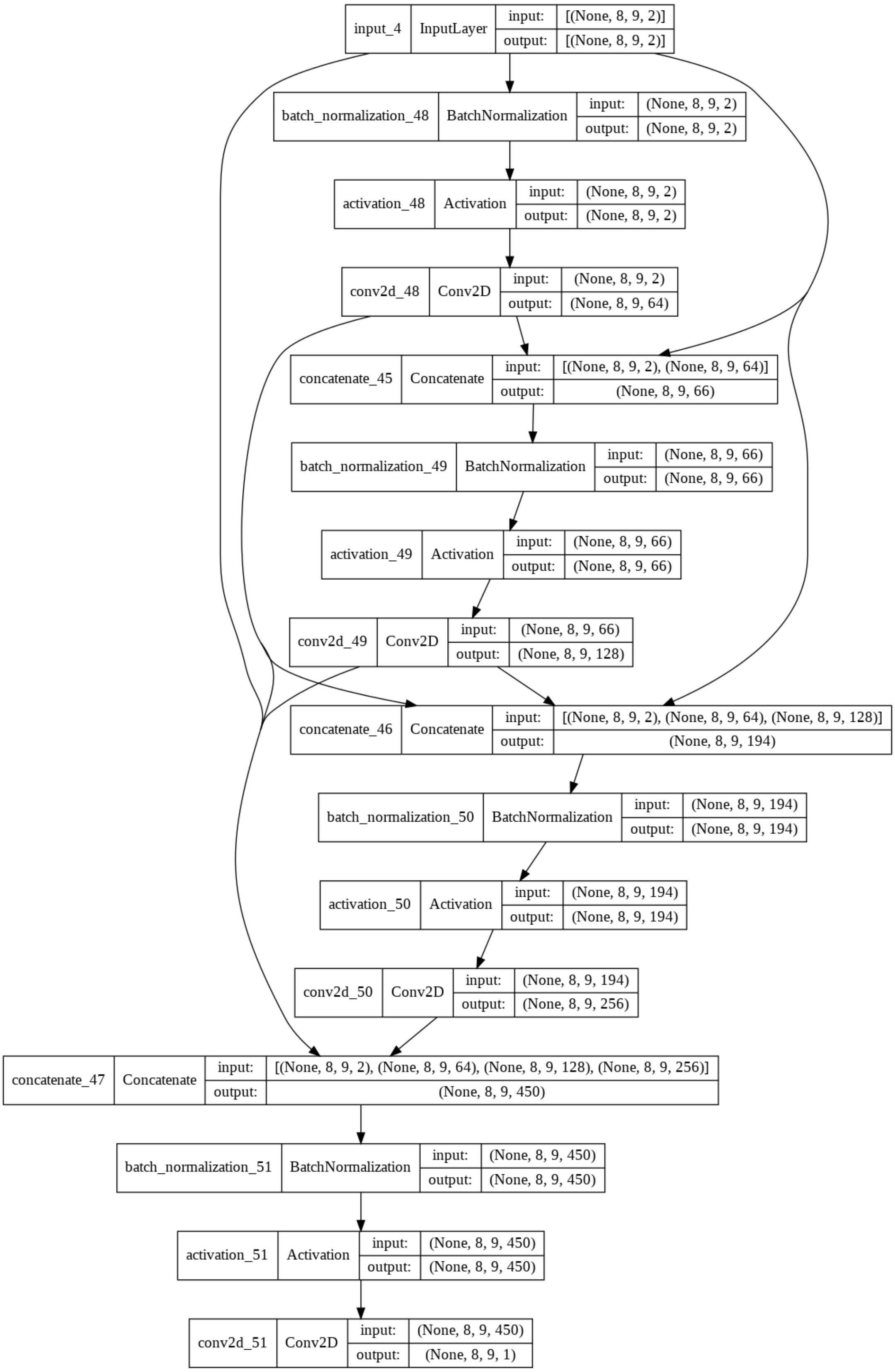

input_4 | InputLayer | input: [(None, 8, 9, 2)] | output: [(None, 8, 9, 2)]
batch_normalization_48 | BatchNormalization | input: (None, 8, 9, 2) | output: (None, 8, 9, 2)
activation_48 | Activation | input: (None, 8, 9, 2) | output: (None, 8, 9, 2)
conv2d_48 | Conv2D | input: (None, 8, 9, 2) | output: (None, 8, 9, 64)
concatenate_45 | Concatenate | input: [(None, 8, 9, 2), (None, 8, 9, 64)] | output: (None, 8, 9, 66)
batch_normalization_49 | BatchNormalization | input: (None, 8, 9, 66) | output: (None, 8, 9, 66)
activation_49 | Activation | input: (None, 8, 9, 66) | output: (None, 8, 9, 66)
conv2d_49 | Conv2D | input: (None, 8, 9, 66) | output: (None, 8, 9, 128)
concatenate_46 | Concatenate | input: [(None, 8, 9, 2), (None, 8, 9, 64), (None, 8, 9, 128)] | output: (None, 8, 9, 194)
batch_normalization_50 | BatchNormalization | input: (None, 8, 9, 194) | output: (None, 8, 9, 194)
activation_50 | Activation | input: (None, 8, 9, 194) | output: (None, 8, 9, 194)
conv2d_50 | Conv2D | input: (None, 8, 9, 194) | output: (None, 8, 9, 256)
concatenate_47 | Concatenate | input: [(None, 8, 9, 2), (None, 8, 9, 64), (None, 8, 9, 128), (None, 8, 9, 256)] | output: (None, 8, 9, 450)
batch_normalization_51 | BatchNormalization | input: (None, 8, 9, 450) | output: (None, 8, 9, 450)
activation_51 | Activation | input: (None, 8, 9, 450) | output: (None, 8, 9, 450)
conv2d_51 | Conv2D | input: (None, 8, 9, 450) | output: (None, 8, 9, 1)

Figure 4.9: CascadeNet model used, shown with a single cascade block for simplicity.

- Threshold Values, Age, IOP
- Threshold Values, Age, Time Between Tests
- Threshold Values, Age, IOP, Time Between Tests
- Threshold Values, Age, Time Between Tests, TD values
- Threshold Values, Age, Time Between Tests, TD values, IOP

4.2.2 Simulating Against Age Normal Values and Last Visual Field Test Values

To assess the ability of the method to reduce test duration, handful of simulation strategies are developed. A common method to prepare starting values for VF test strategy is using age normal values. Age normal values are what people with normal visual fields have as the luminosity threshold at each point. Age normal datasets are prepared by conducting a clinical trial to hundreds of people from different age groups and genders. The age normal dataset we use here is from [3], generated by testing 102 healthy patients with the Humphrey Visual Field Analyzer. Figure 5.9 shows the age normal values categorized by age.

Using age normal values as starting points for VF testing is a great approach for testing healthy people to check for any signs of possible illnesses like glaucoma. However, to monitor the progression of a glaucoma patient, it is an ineffective method since, at every test, additional stimuli presentations will be shown to the patient to check if the tested location is healthy, which as the clinician knows a priori, it is not. For this, using forecasted values will be a great alternative because, the forecasted results will be closer to the real sensitivity values of the patient.

Even though, as an idea, using forecasted values looks promising, it must have a clear lead on any naive method that comes to mind. For example, for progressing glaucomatous patients, a naive method like using the values from the patient's last conducted test. Therefore, the performance on forecasting against age normal values and previous test values should be examined.

Here we simulated aforementioned 3 methods using a computer program, simulating the Full Threshold [22] [4] test strategy, which is a staircase based strategy. Since, the possible false-negative and false-positive values will affect the test duration in a similar proportion, they are omitted.

The results of the simulation and the limitations of the method is given in the following chapter.

			29.2	28.3	27.6	27.8		
		29.1	29.8	30.2	29.8	30.9	30.4	
	29.6	31.2	32.6	32.3	31.7	31.2	30.6	30.4
28.9	30.4	32.1	32.8	33.7	33.5	32.2	24.7	31.3
29.4	31	32.6	33.3	33.8	33.8	32.7	9.2	31.5
	30.4	31.5	33.4	32.9	32.8	32.8	31.1	31.6
		30.6	31.2	32.1	31.3	30.9	32.4	
			29.8	30.9	31.4	31.8		

(a) 20-30

			28.5	29.5	28.4	28		
		28.2	30.1	24.6	29.2	24.3	28.7	
	29.6	30.9	32.1	31.6	31.4	30.1	29.8	30.5
28.4	30.2	31.8	32.2	33.2	31.9	31.4	27.2	30.9
24.8	30.8	32.5	32.6	32.9	32.8	31.9	3.7	30.7
	30.1	31.5	32.8	31.7	31.9	32.2	30.8	31.2
		30.3	31.5	31	31.1	31.3	31.5	
			29.9	30.3	30.8	30.4		

(b) 30-40

			26.8	28.1	26.9	26.7		
		27	27.8	27.9	27.2	27.9	26.8	
	27.8	29.6	30.3	29.8	29.8	28.7	28.2	29.1
27.4	28.9	30.3	31.2	31.8	32.2	30.6	25.1	29.8
28.1	29.2	30.9	31.6	31.7	31.9	31.1	8.1	30.1
	28.2	29.2	30.9	31.1	31.5	31.8	29.8	29.9
		28.6	29.3	29.8	29.8	30.1	30.6	
			28.5	29.3	29.1	29.7		

(c) 40-50

			28.4	28.2	27.3	28		
		29.2	29	29.3	28.8	29.2	29.4	
	29.2	30.3	31.6	31	30.9	30.1	29.9	29
28.1	29.5	20.8	31.7	32.3	31.5	30.6	26.7	30.1
28.6	29.9	31.4	31.8	32.9	32.4	32.1	1.7	30.3
	30.1	30.7	32.6	32.2	31.8	31.3	30.1	30.9
		29.7	31.3	30.6	30.9	30.9	31.1	
			29.6	30.2	31.1	30.3		

(d) 50-60

			24.8	26	25.1	25.2		
		25.9	27.2	27.8	25.8	26.5	25.3	
	26.8	28.6	29.4	30.1	28.3	27.8	27.2	27.1
25.1	27.1	29.3	29.8	30.8	30.2	29.7	23.2	27.9
26.4	27.4	29.7	31.2	31.1	31.2	30.4	2.8	28.2
	27.9	29.3	30.7	30.3	30.2	29.9	27.8	28.2
		27.8	29	29.2	28.9	28.8	28.3	
			27.1	27.4	27.6	27.7		

(e) 60-70

			24.8	25.2	26.6	25.2		
		26.1	27.1	26.8	26.5	26.2	24.2	
	27.2	28	28.6	29.1	28.2	26.9	26.9	26.5
25.8	28.1	28.6	29.8	30.2	29.2	28.6	22.2	26.6
25.2	28	29.9	29.6	30.8	29.7	29.8	5.8	27.4
	27.8	28.8	29.8	30.1	29.1	29.2	28.5	28
		27.5	28.6	28.2	28.6	28.8	28.5	
			26.6	27.5	28.5	28.4		

(f) 70+

Figure 4.10: Age normal values for each age group calculated in [3].

CHAPTER 5

RESULTS

5.1 Dataset

The dataset used in NN training and RL simulations is acquired at the Rotterdam Eye Institute (Netherlands) [5]. Dataset includes 5108 VF results from 139 patients with glaucoma. All VFs are acquired using the 24-2 test pattern of the Humphrey Visual Field Analyzer. Each VF consists of 54 point locations. Each patient has taken 34.9856 tests on average. The mean period of testing is 7 years. Extra to VF thresholds, dataset includes total deviation values, mean deviation values, intraocular pressure and age of the patient in days.

Figure 5.1 shows the distribution of test numbers by patients. One can see from the figure that majority of patients took more than 20 tests. Figure 5.2 shows the distribution of the duration between two VF tests. Figure 5.2 is constructed by generating test pairs which will be used for NN training in the following chapters. Each test is paired with all possible future tests that belongs to the same patient and the same eye. Table 5.1 summarizes the features of the dataset.

Table 5.1: Properties of Rotterdam Dataset

Property	Value
Number of VFs	5108
Number of VF pairs	37577
Number of Patients	139
Available Features	Thresholds, TD, IOP, MD

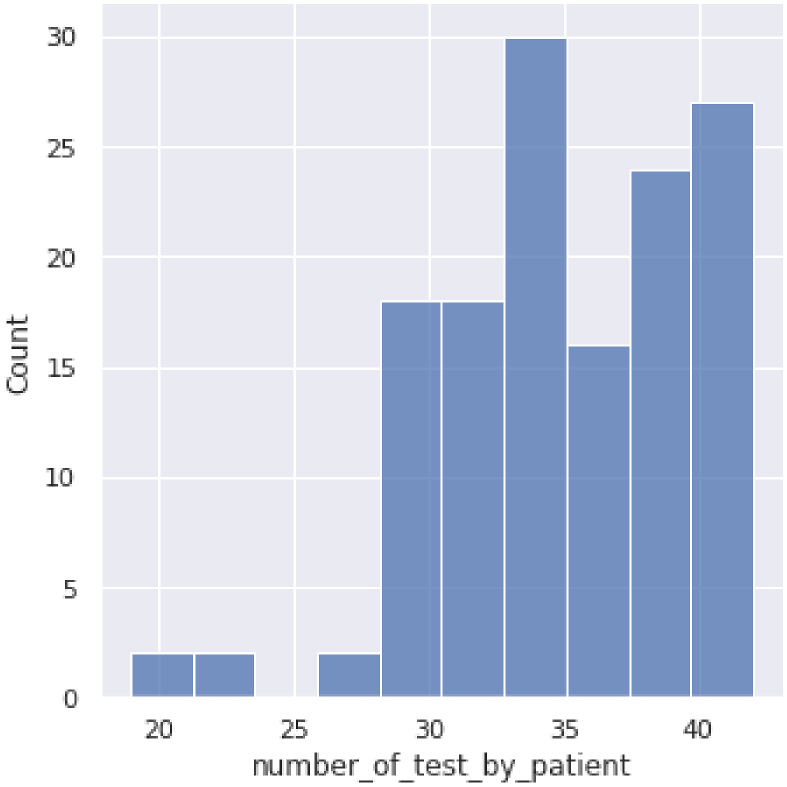


Figure 5.1: Distribution of the number of tests by patient.

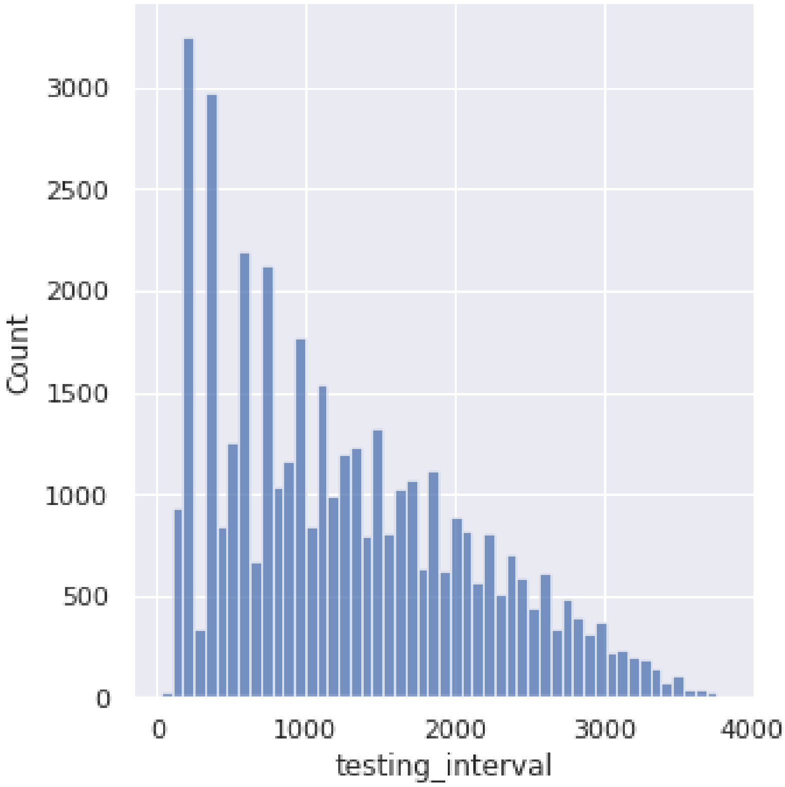


Figure 5.2: Distribution of the test intervals.

5.2 Finding Optimal Test Locations on VR Headset Screen

5.2.1 Comparison of Different Environment Types

Two types of environments are developed for RL agent to interact. In the first one, as action, agent chooses a point on the matrix, which means agent has 28x28 options at the start, decreasing by one at each step.

In the second one, the gamified environment, agent, at each step, chooses a direction in 4 options and a length to travel from 1 to 8 cells, which gives the agent 32 options as the action. However, as mentioned in the previous section, this while reducing the action space, limits the agent's ability to find the most optimal sequence of locations.

Both environments are simulated with the same data for 10 runs. The mean of the RMSE and PMAE scores are recorded for 100,000 training steps and 500,000 training steps.

Table 5.2: Simulation comparisons of standard and gamified environment for 100,000 training steps for 10 runs.

	PMAE	RMSE
Standard Environment	1.02	1.78
Gamified Environment	3.56	4.86

Table 5.3: Simulation comparisons of standard and gamified environment for 500,000 training steps for 10 runs.

	PMAE	RMSE
Standard Environment	0.99	1.75
Gamified Environment	3.43	4.36

As can be seen from Table 5.2, gamified environment has a significantly higher

PMAE and RMSE which means it has a lower accuracy than the standard environment. To check the affect of increasing the number of training steps, same simulation is run 10 times for 500,000 training steps. As can be seen from Table 5.3, increasing the number of training steps did not increase the accuracy at both of the environments. Therefore, one can conclude that, agents on both environments had no benefit from increasing training time, and agent on gamified environment, due to the limitations mentioned above had lower performance.

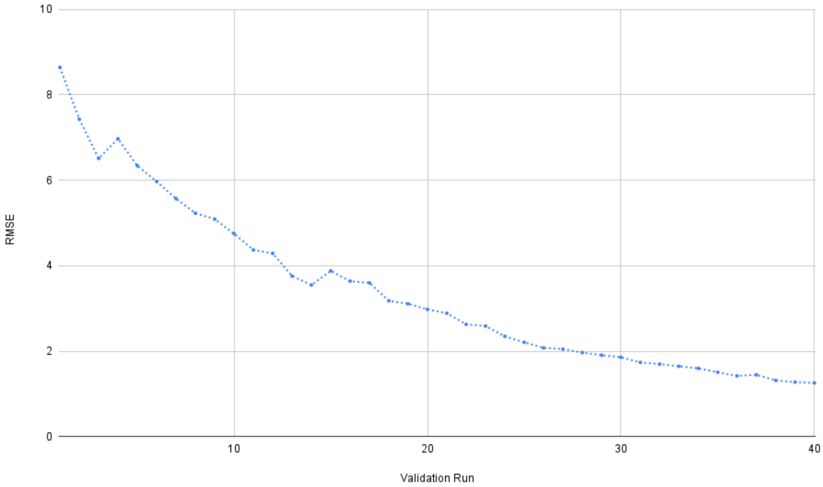


Figure 5.3: Validation error measured periodically for 2500 training steps.

Figure 5.3 shows the validation error through 100,000 training steps in 40 measurements.

5.2.2 Simulations

The main promise on using such a method is to find optimal values that are less in quantity than the original ones, but can produce the same quality as them while leveraging the continuousness of the VR headset screens. Moreover, the point locations that are extracted by the algorithm should decrease the test timing on future tests while keeping an acceptable accuracy. To verify the effectiveness on the method on future tests, a patient's first test on the dataset was given to the RL agent and extracted points from the agent after training are used to estimate the VF on future tests.

To simulate, VFs are interpolated to 28x28 images and the resulting up-scaled images are used as the data. After the agent finds the optimal points, values corresponding to the locations are extracted. Then the VF is reconstructed by interpolation using only those points. That way a sub-optimal representation of the VF is acquired. By comparing the original 54 points at the sub-optimal and the original image, one can find the accuracy of reconstruction. Therefore, the performance of the model is validated.

To compare representations, two metrics is used. One is the RMSE and the other is the PMAE. Equation 4.4 shows the RMSE calculation and the Equation 3.1 shows the PMAE calculation.

The mean number of training steps is 100.000. With a simulation setup using Intel Core i7-9850H and 32GB of RAM, 100.000 steps of training approximately takes 1.5 hours. Due to the computational limitations, all the available data can not be used in simulation. To validate the method, a number of patients' first test data is simulated and the accuracy metrics are recorded for the future tests.

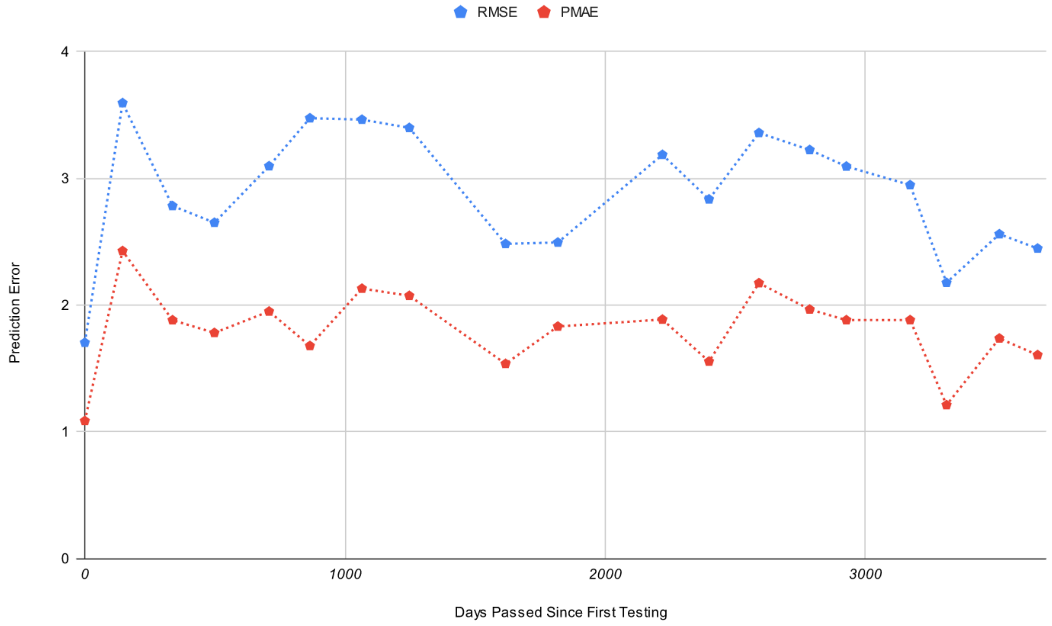


Figure 5.4: The change of accuracy metrics on future tests for Study_ID 50 of the dataset.

Figure 5.4 shows the results of one of the simulations. The data belongs to a patient labeled with Study_ID 50 on the Rotterdam dataset. X axis shows the number of

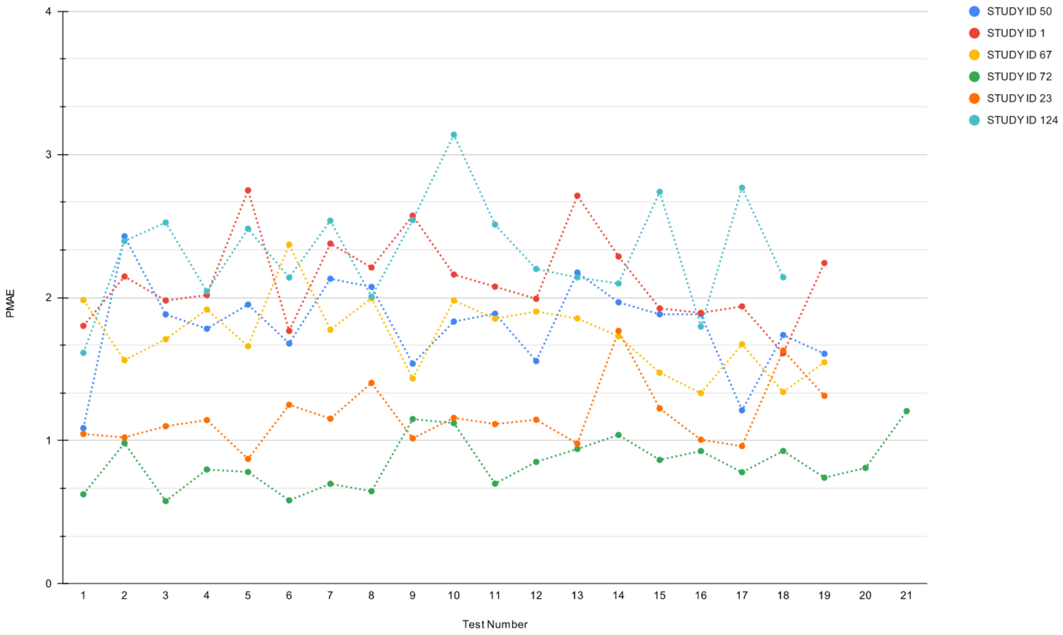


Figure 5.5: PMAE values of 6 patients against conducted tests spanning 10 years.

days passed after the first testing. The RL agent extracts the point locations using the test at x=0. All the following VF maps are created using the extracted points at x=0. Accuracy metrics show that even with using the same points of the first test, future VFs are constructed with point-wise error of below 2.5 which in the context of a highly subjective test is very good. The number of points extracted at the first test is 40 which is 25% less than the 24-2 pattern, meaning a test strategy like Full Threshold will be completed 25% faster.

Same simulation is run for 5 other Study_IDs. Both the RMSE and PMAE values are recorded for all the tests of a single patient. Average errors for each patient are shown. Errors are plotted against the conducted test number in Figure 5.5 and Figure 5.6.

Moreover, the different error values for different number of optimal points can be seen in Figure 5.7. After 40 points PMAE becomes larger than 3dB.

To validate the assumption that the proposed method will reduce test duration, 6 patients and their 115 tests are simulated with the Full Threshold algorithm. For each test, 54 points and estimated 40 points versions are used. As the starting points, age

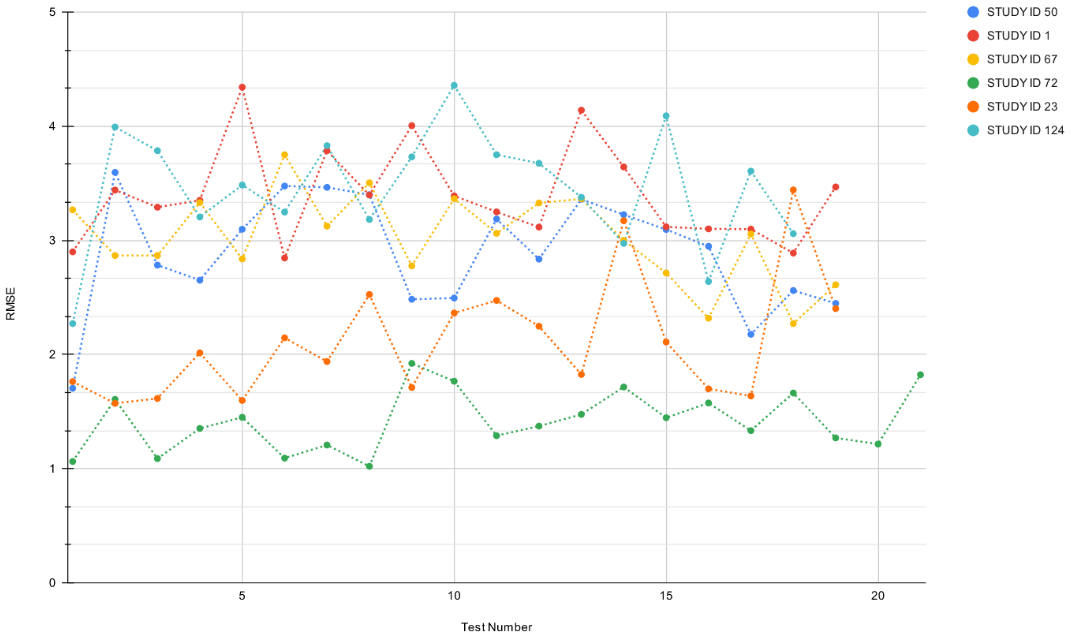


Figure 5.6: RMSE values of 6 patients against conducted tests spanning 10 years.

normal values are used which can be seen in Figure 5.8. To find the age-normal values of the estimated points falling between the locations of original points, same interpolation strategy is used. Figure 5.8 shows the interpolated age-normal values for the age groups. The results of the simulation is given in Table 5.4. Since, inter-response times are dynamically changing throughout the testing duration and depend on the patient itself, to compare the duration number of shown stimuli is used.

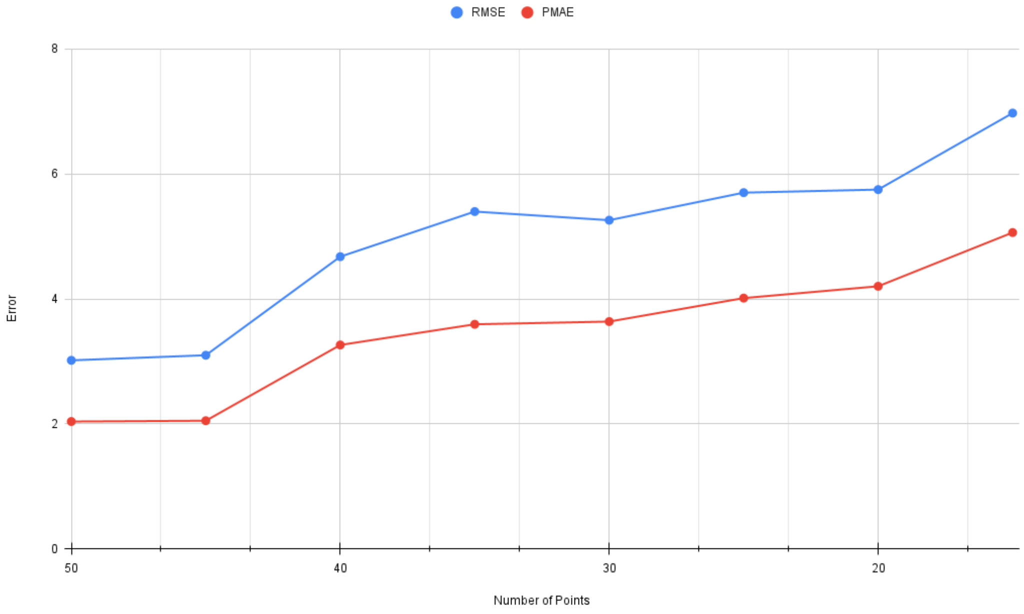
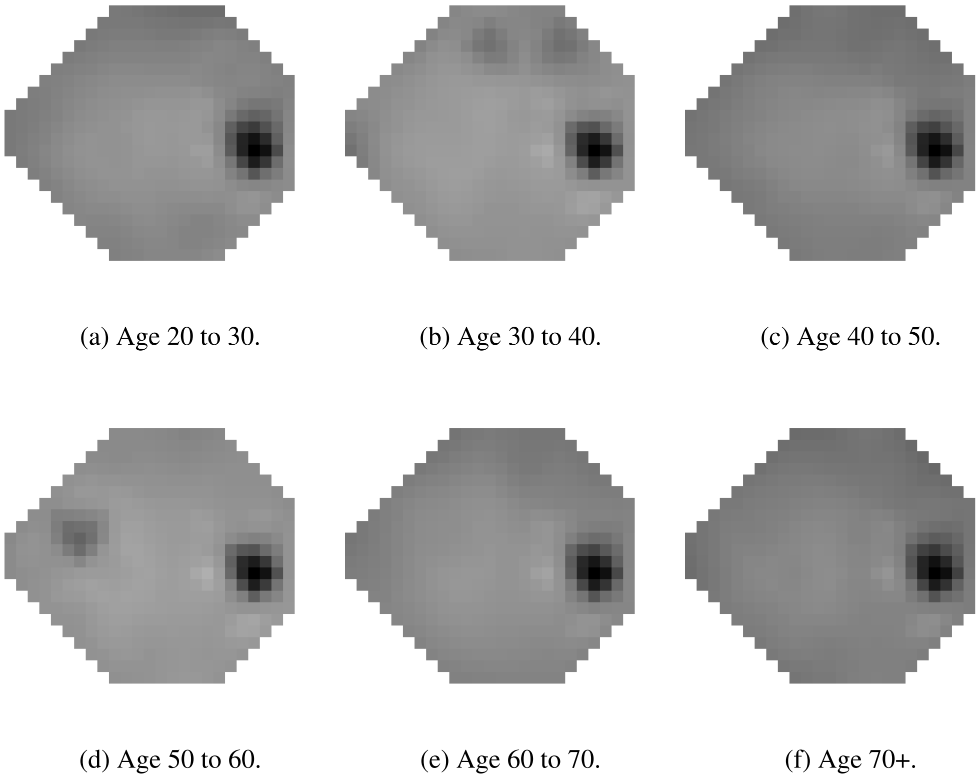


Figure 5.7: Error versus the number of points.



(a) Age 20 to 30.

(b) Age 30 to 40.

(c) Age 40 to 50.

(d) Age 50 to 60.

(e) Age 60 to 70.

(f) Age 70+.

Figure 5.8: The interpolated age-normal values for the age groups.

$$MNOSS = \frac{1}{m} \Sigma_{j=1}^{m} \Sigma_{i=1}^{n} a_{ji} \tag{5.1}$$

Equation 5.1 shows the computation of Mean Number of Stimuli Shown (MNOSS), a metric to compare test durations as the mean of the number of total stimuli shown for a test. m is the number of tests in the input test set, n is the number of point locations in a VF test, a_{ji} is the number of stimuli shown at location i for test number j.

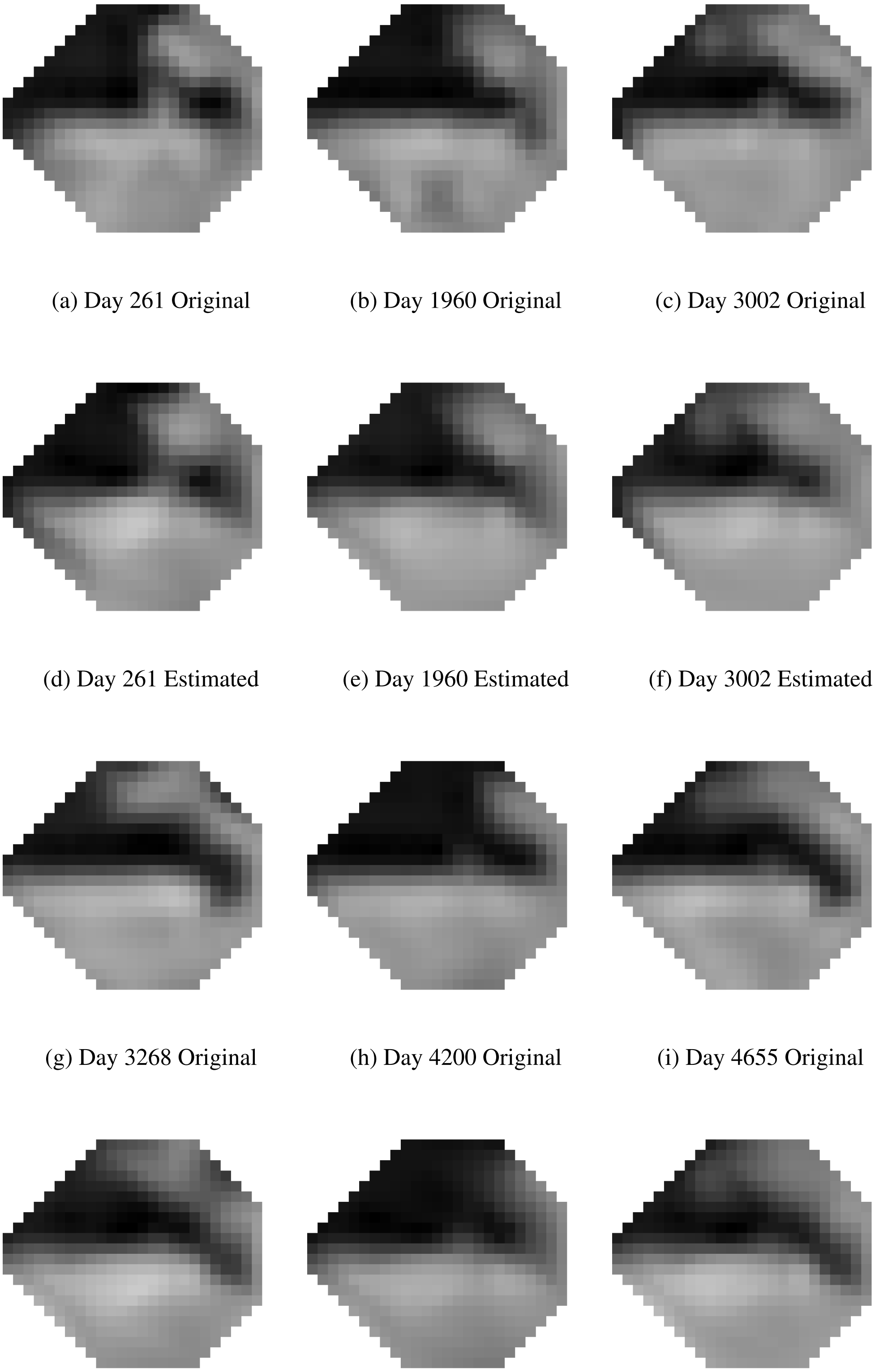
Table 5.4: Simulation comparisons of standard and gamified environment for 500,000 training steps for 10 runs.

Number of Points	Mean Number of Stimuli Shown (MNOSS)
54	184.86
40	138.22

As can be seen from the results, showing less number of points directly decreases the test duration proportionally in a deterministic test strategy like Full Threshold. However, in real world scenarios patients make mistakes such as responding to seen stimuli as not seen and responding to not shown stimuli as seen. This can vary the test duration for a patient but, for a population average this variable will have an insignificant effect.

Lastly, to justify an error of 1.5dB to 4dB as acceptable, previous studies can be examined. In [24], [8] and [2], authors calculated intertest variability for each position in 30-2 test pattern which encloses the 24-2 test pattern. Point-wise intertest variability can take values from 2dB to 5dB.

To further validate the effectiveness of the method, a clinical trial including glaucomatous eyes can be conducted.



(a) Day 261 Original

(b) Day 1960 Original

(c) Day 3002 Original

(d) Day 261 Estimated

(e) Day 1960 Estimated

(f) Day 3002 Estimated

(g) Day 3268 Original

(h) Day 4200 Original

(i) Day 4655 Original

(j) Day 3268 Estimated

(k) Day 4200 Estimated

(l) Day 4655 Estimated

Figure 5.9: Comparison of the original and estimated VFs for Study ID 50.

5.3 Forecasting Methods for Reduction in Test Duration

5.3.1 Forecasting Model Training

To forecast future VF results, an accurate forecasting method is needed. There are several models used in the literature, which are compared in Figure 5.10. Using the findings in [25], we decided to use CascadeNet-5 in this thesis for forecasting task.

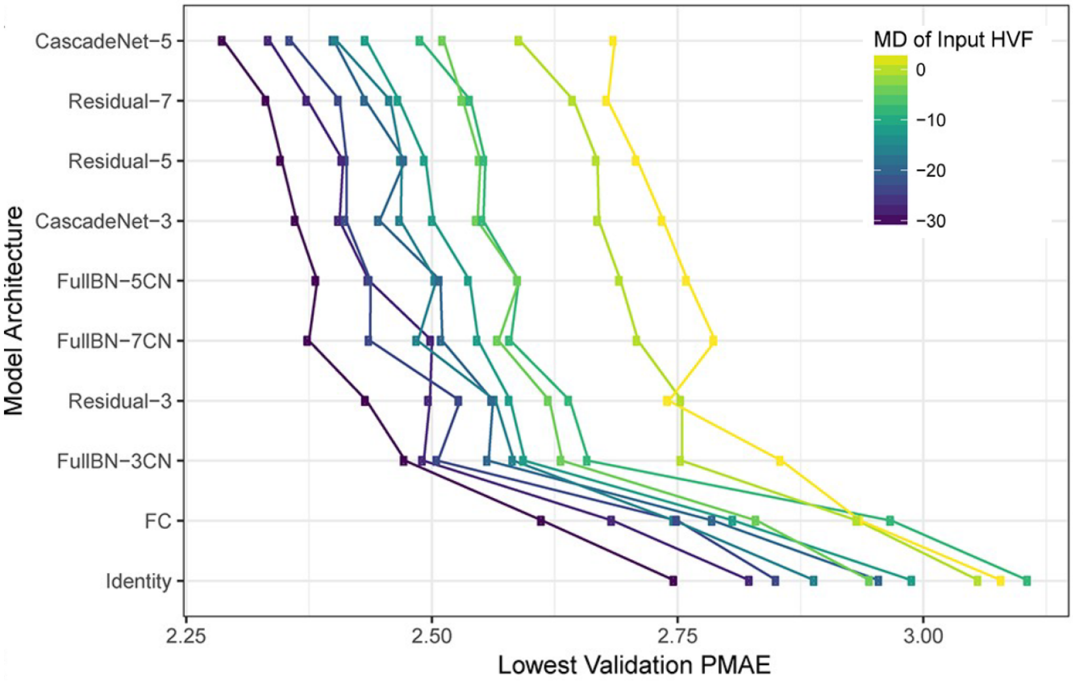


Figure 5.10: Comparison of different NN models for the forecasting task. [25]

To train the model, different feature sets are used which are given in Chapter 3. Training with different feature sets resulted in different forecasting accuracies. Feature sets can be divided into two groups; fixed interval and variable interval. In fixed interval, intertest durations are fixed and training-label tuples are constructed with a fixed time between them, for example, half a year. In variable interval, intertest durations are given as the 3rd channel of the input as a matrix with all values filled with the intertest duration as days. 2nd channel is always the patient's age as years, again in a matrix form with all cells equal to age.

For the feature sets that do not use the "Time Between Tests" feature, the data and label are prepared in a way that conducted test and the test used as the label have a fixed time between them. Therefore, each feature set used had a unique training set. For the fixed interval feature sets, tests between `t.0,75` and `t.1,25` are used. For example, for a 1 year period, pairs are constructed between 274 days and 456 days.

Table 5.5: Dataset sizes for different types of feature sets.

Time Between Pairs	Number of Pairs
Variable Time	37577
0.5 Year (Fixed)	3615
1 Year (Fixed)	3632
1.5 Year (Fixed)	4143
2 Years (Fixed)	6442

For the fixed interval, the number of training pairs are low in number since, for each interval a separate model needs to be trained. Because of that, model accuracy is lower.

Table 5.6: Test set PMAE values of different feature sets after training for different year constants.

Difference in Years	Threshold Values	Threshold Values + IOP	Threshold Values + TD values	Threshold Values + IOP + TD values
0.5	3.26	3.14	2.91	3.01
1.0	2.96	2.97	3.02	3.14
1.5	3.02	3.13	2.97	3.03
2.0	3.27	3.14	3.08	3.09
2.5	3.32	3.35	3.25	3.35
3.0	3.55	3.54	3.45	3.39

Table 5.6 shows the PMAE values of different models trained with pairs of different time differences and Table 5.7 shows the RMSE values of different models trained with pairs of different time differences. As can be seen from Table 5.6, even though with a little difference, supplying the TD values as an extra feature gives the least error on the test set.

Table 5.7: Test set RMSE values of different feature sets after training for different year constants.

Difference in Years	Threshold Values	Threshold Values + IOP	Threshold Values + TD values	Threshold Values + IOP + TD values
0.5	4.48	4.34	4.20	4.28
1.0	4.18	4.21	4.25	4.32
1.5	4.31	4.38	4.27	4.29
2.0	4.49	4.39	4.39	4.37
2.5	4.68	4.67	4.61	4.60
3.0	4.96	5.02	4.78	4.81

For the variable interval features, the number of training pairs are high in number. This results in the amplification of the effect of other features like IOP and TD values. For variable time training set, a single model is trained. The test set error values of model are categorized by the difference in years between training data and labels. For all the feature sets, limiting the difference in years to 5.5 years reduced the error on all time difference categories. This limitation was found empirically and also authors in [25] had the same limitation and the same effect on training error. Table 5.8 shows the improvement of PMAE by limiting the pair test difference to 5.5.

Table 5.8: Difference of test set PMAE values comparison for time difference upper limit and no upper limit.

Time Difference Category (Years)	No Upper Limit	Upper Limit of 5.5 Years
0.5	3.18	3.11
1.0	3.30	3.18
1.5	3.42	3.28
2.0	3.57	3.41
2.5	3.81	3.63
3.0	4.01	3.85
3.5	4.07	3.89
4.0	4.23	4.09
4.5	4.52	4.35
5.0	4.54	4.36
5.5	4.58	4.20
5+	5.18	

Table 5.9 shows the PMAE value of the test set for different groups of time differences between training and label data. As can be seen from the table, feature set that yields the lowest error value on test data is the one including Threshold Values, Intraocular Pressure and Total Deviation values. Each feature set also includes a channel that

Table 5.9: Test set PMAE values of different feature sets against different test pair time difference categories.

Time Difference Category (Years)	Threshold Values	Threshold Values + IOP	Threshold Values + TD values	Threshold Values + IOP + TD values
0.5	3.11	3.08	3.00	2.98
1.0	3.18	3.15	3.08	3.04
1.5	3.28	3.25	3.17	3.12
2.0	3.41	3.44	3.30	3.27
2.5	3.63	3.64	3.49	3.45
3.0	3.85	3.86	3.65	3.65
3.5	3.89	3.87	3.72	3.69
4.0	4.09	4.09	3.93	3.90
4.5	4.35	4.30	4.19	4.15
5.0	4.36	4.35	4.20	4.11
5.5	4.20	4.19	4.07	4.00

holds the age value in years. Therefore resulting input shape is 8x9x4. By comparing the fixed interval and variable interval results, one can conclude that variable interval results only have a slight advantage.

5.3.2 Algorithm Simulation

To understand the advantage of using different starting values for the visual field testing, Full Threshold algorithm is simulated with using 3 different starting values; age-normal, previous test and forecasted values.

Simulating with age-normal values resulted in comparably worst test duration than the other methods. Average duration of the Full Threshold test with starting values as age-normal values is given in Table 5.4. Since, this value is comparably worse, it will not be included in the comparison tables below. Instead of this, difference between the simulation result of forecasted value start and last test start will be used as the metric of comparison. Equation 5.2 gives the used metric for comparing test speed up. *fvs* is the forecasted staring values and *lts* is the last test starting values.

$$Duration\ Improvement = lts - fvs \tag{5.2}$$

All VF pairs are simulated with the 3 starting value options. Forecasted values are obtained using fixed interval and variable interval models.The models with feature

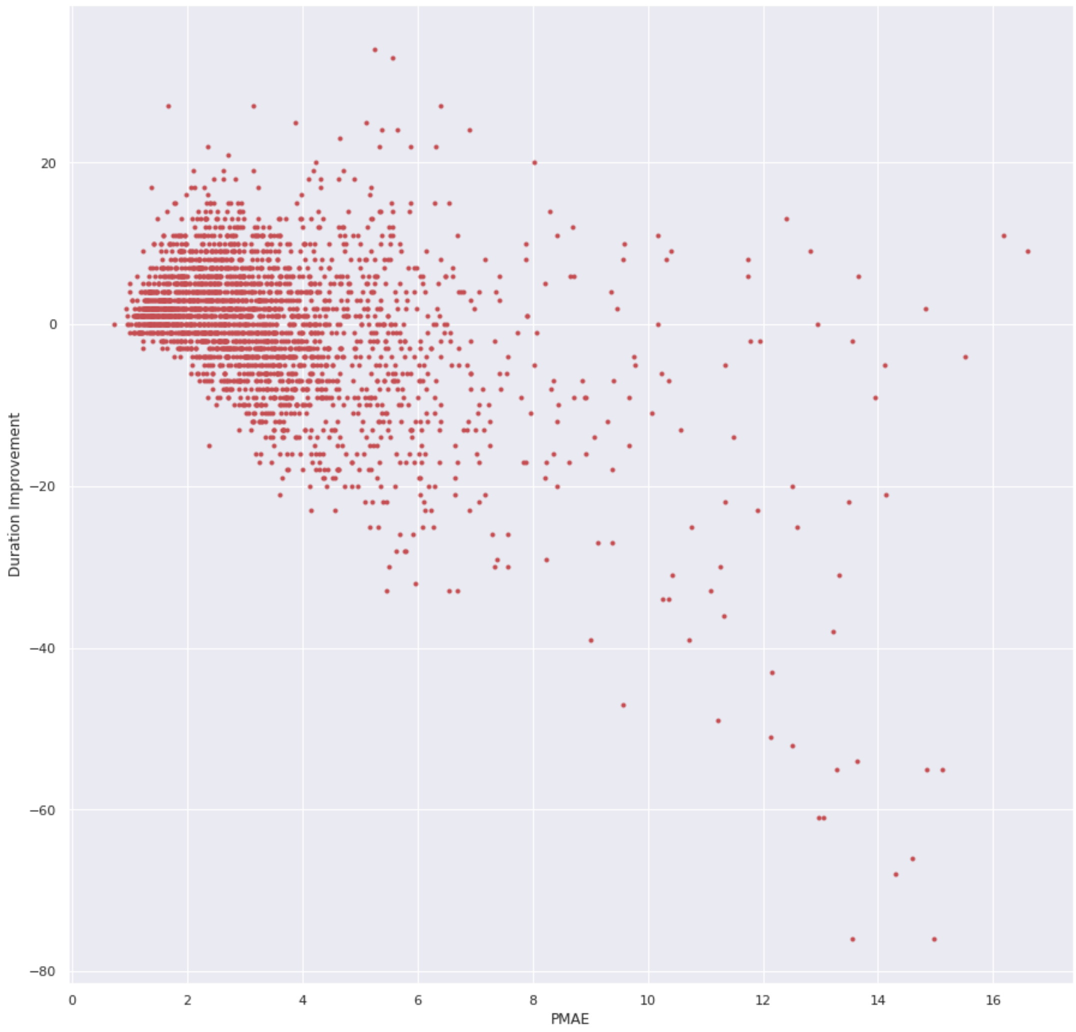


Figure 5.11: Scatter graph of PMAE values against Duration Improvements

sets that result in the lowest errors are used. For the variable interval model, duration improvement average is 0.74, which is not significant. However, by plotting the duration improvement distribution we can see that some test results benefited greatly while others not. Figure 5.11 shows the PMAE values against the duration improvements. From the figure, one can see that even though at the extremity of PMAE values duration improvement steadily gets worse, there is no clear relation between the prediction accuracy and duration improvement.

A negative difference between two MD values shows that the light sensitivity of the eye is getting worse. Equation 5.3 show the MD calculation. th_i is the measured threshold value of a location i and an_i is the age normal value of the location i. n is

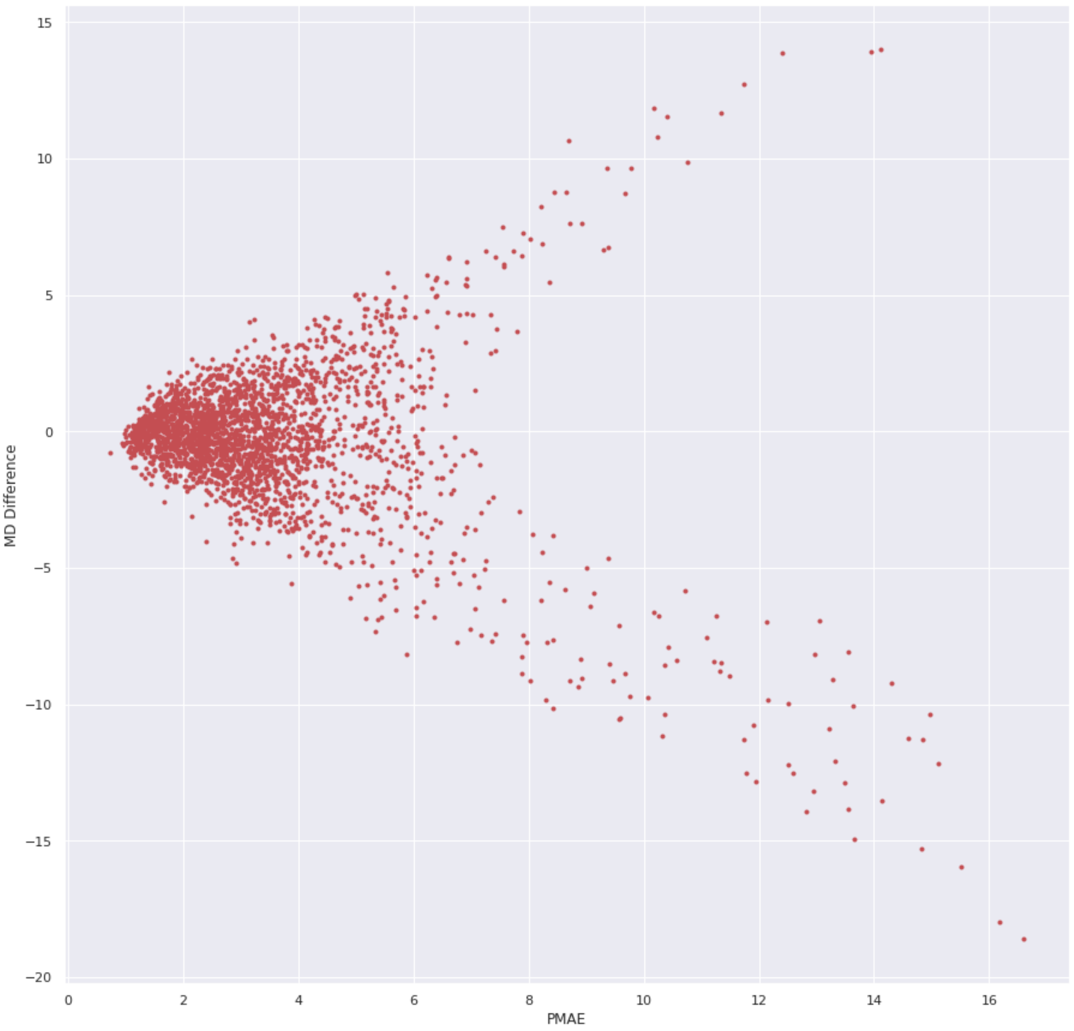


Figure 5.12: Scatter graph of PMAE values against MD differences.

the number of locations. a is the age of the patient and it is supplied to the equation to use the appropriate age-normal values.

$$MD(a) = \frac{1}{n}\sum_{i=1}^{n} an_i(a) - th_i \tag{5.3}$$

Figure 5.12 shows the PMAE values against the MD differences. One can see from the figure that as the difference between training VF MD and label VF MD is getting larger the PMAE value tends to get larger. However, as can be seen from the figure this can be in any direction. Therefore, the progression direction and the prediction error has a week correlation so the model has no information about the progression direction. To exploit this, MD differences are split into 8 sub-categories and not surprisingly with worsening eyes, in other words, with negative MD difference, duration improvement is mostly positive. This is due to the fact that while the threshold values between the last test and the new test increases, the effect of using forecasted values becomes apparent because, the last test threshold values are no longer similar to the real sensitivities of the eye, but the forecasted values are more similar to the real sensitivities. However, after 2.5 years the effectiveness of using the forecasted values, decrease. The results of the simulations, and the resulting duration improvements are displayed against different MD difference groups are given in Table 5.10 and 5.11.

Table 5.10: MD difference categories against duration improvements for different MD differences of the model with no extra features.

MD Difference Categories	0.5	1.0	1.5	2.0	2.5	3.0
-12 to -9			1.33	0.67	-2.00	-24.00
-9 to -6	19.75	15.25	5.50	5.14	6.18	-13.87
-6 to -3	5.75	5.20	7.16	5.00	4.82	-4.91
-3 to 0	2.28	2.23	1.50	2.10	0.84	-0.38
0 to 3	0.71	2.32	-1.10	2.11	-3.36	2.00
3 to 6	-3.10	3.00	-3.75	1.50	-9.53	-2.20
6 to 9	-6.00	-3.00		-5.00	-21.33	-12.83
9 to 12		1.00	-9.00	-7.00	-4.50	-6.67

Table 5.11: MD difference categories against duration improvements for different MD differences of the model with TD and IOP features.

MD Difference Categories	0.5	1.0	1.5	2.0	2.5	3.0
-12 to -9			5.33	2.50	-16.00	-22.10
-9 to -6	20.50	13.7	13.50	8.00	0.45	-11.67
-6 to -3	3.75	0.40	6.11	5.03	-0.03	-3.82
-3 to 0	2.49	0.83	2.22	2.16	1.38	0.37
0 to 3	1.71	4.18	2.17	1.43	2.56	3.22
3 to 6	0.20	8.50	0.92	0.05	1.83	3.77
6 to 9	-8.50	5.00		-6.00	-2.67	-4.50
9 to 12		4.00	- 9.00	-2.00	-2.00	-0.33

It is apparent from the tables that if the progression direction is known the proposed method of using forecasted values is beneficial and can result in the decrease of up to 20 stimuli presentations. However, predicting VF progression is not a trivial task. On the other hand, there are numerous methods that successfully predict VF progression like in [20]. Combining any such method that predicts VF progression with the proposed method will result in decreased VF test duration for progressing fields.

CHAPTER 6

CONCLUSIONS

There are many approaches to decrease VF test duration. Two of those approaches are decreasing the number of tested locations and changing the starting luminosities of test locations. In this thesis, new methods for both approaches are proposed. For decreasing the number of tested locations, a Reinforcement Learning approach that uses Maskable PPO algorithm is employed. By leveraging the continuousness of the Virtual Reality Headset screens, locations falling between the conventional 24-2 test pattern locations made available to the model. By selecting the optimal points from the available point locations, RL model successfully estimated the same VF with only an average point-wise error between 1dB and 4dB and using the same points extracted from a single VF test, estimated up to 20 future VF test with the same error margin.

For the second method of changing the starting luminosities, a convolutional neural network based model is used to forecast future VFs up to 3 years is used as the starting luminosity values to decrease the number of shown stimuli, and simulated using the Full Threshold algorithm. It is shown that, for the progressing visual fields, up 20% decrease in duration is observed compared to using the threshold values of the last conducted test. Against age-normal values the decrease in duration is up to 40%.

In future work:

- Found optimal testing locations can be tested on real glaucomatous patients in a clinical trial to measure real-world test duration decrease,
- Found optimal testing locations can be simulated using more state-of-the art testing methods which due to their complexity will need their own paper,
- Combining forecasting with a progression prediction algorithm to make a frame-

work to decrease test duration.

- Trying a lighter optimization method rather than modeling as a reinforcement learning environment to reduce convergence time.

REFERENCES

[1] https://www.zeiss.com/meditec/us/product-portfolio/perimetry/humphrey-visual-field-analyzer-3-with-sita-faster.html. Accessed: 2022-1-25.

[2] B. Bengtsson and A. Heijl. Normal intersubject threshold variability and normal limits of the SITA SWAP and full threshold SWAP perimetric programs. *Investigative Opthalmology & Visual Science*, 44(11):5029, Nov. 2003.

[3] R. S. Brenton and C. D. Phelps. The normal visual field on the humphrey field analyzer. *Ophthalmologica*, 193(1-2):56–74, 1986.

[4] A. Chandrinos. Methods of threshold estimation (algorithms) and new techniques in perimetry. a review. *SSRN Electronic Journal*, 2020.

[5] N. S. Erler, S. R. Bryan, P. H. C. Eilers, E. M. E. H. Lesaffre, H. G. Lemij, and K. A. Vermeer. Optimizing structure–function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Opthalmology & Visual Science*, 55(4):2350, Apr. 2014.

[6] Y. Fujino, H. Murata, C. Mayama, and R. Asaoka. Applying “lasso” regression to predict future visual field progression in glaucoma patients. *Investigative Opthalmology & Visual Science*, 56(4):2334, Apr. 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[8] A. Heijl, G. Lindgren, and J. Olsson. Normal variability of static perimetric threshold values across the central visual field. *Archives of Ophthalmology*, 105(11):1544–1549, Nov. 1987.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[10] A. Karakawa, H. Murata, H. Hirasawa, C. Mayama, and R. Asaoka. Detection of progression of glaucomatous visual field damage using the point-wise method with the binomial test. *PLoS ONE*, 8(10):e78630, Oct. 2013.

[11] Ş. S. Kucur, P. Márquez-Neila, M. Abegg, and R. Sznitman. Patient-attentive sequential strategy for perimetry-based visual field acquisition. *Medical Image Analysis*, 54:179–192, May 2019.

[12] Ş. S. Kucur and R. Sznitman. Sequentially optimized reconstruction strategy: A meta-strategy for perimetry testing. *PLOS ONE*, 12(10):e0185049, Oct. 2017.

[13] X. Li, W. Li, X. Xu, and Q. Du. Cascadenet: Modified resnet with cascade blocks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 483–488, 2018.

[14] @oculera. Oculera. `https://www.oculera.health/img/main_page_header/pico-neo2-eye.png`. Accessed: 2022-1-25.

[15] N. O'Leary, B. C. Chauhan, and P. H. Artes. Visual field progression in glaucoma: Estimating the overall significance of deterioration with permutation analyses of pointwise linear regression (PoPLR). *Investigative Opthalmology & Visual Science*, 53(11):6776, Oct. 2012.

[16] K. Park, J. Kim, and J. Lee. Visual field prediction using recurrent neural network. *Scientific Reports*, 9(1), June 2019.

[17] H. A. Quigley. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology*, 90(3):262–267, Mar. 2006.

[18] L. Racette, M. Fischer, H. Bebie, H. Gabor, C. A. Johnson, and C. Matsumoto. *Visual field digest: A guide to perimetry and the Octopus Perimeter*. Haag-Streit AG, 2018.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[20] K. Shon, K. R. Sung, and J. W. Shin. Can artificial intelligence predict glaucomatous visual field progression? a spatial–ordinal convolutional neural network model. *American Journal of Ophthalmology*, 233:124–134, Jan. 2022.

[21] C.-Y. Tang, C.-H. Liu, W.-K. Chen, and S. D. You. Implementing action mask in proximal policy optimization (ppo) algorithm. *ICT Express*, 6(3):200–203, 2020.

[22] A. Turpin, A. M. McKendrick, C. A. Johnson, and A. J. Vingrys. Properties of perimetric threshold estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer simulation. *Investigative Opthalmology & Visual Science*, 44(11):4787, Nov. 2003.

[23] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[24] M. Wall, C. K. Doyle, K. D. Zamba, P. Artes, and C. A. Johnson. The repeatability of mean defect with size III and size v standard automated perimetry. *Investigative Opthalmology & Visual Science*, 54(2):1345, Feb. 2013.

[25] J. C. Wen, C. S. Lee, P. A. Keane, S. Xiao, A. S. Rokem, P. P. Chen, Y. Wu, and A. Y. Lee. Forecasting future humphrey visual fields using deep learning. *PLOS ONE*, 14(4):e0214875, Apr. 2019.

[26] H. Zhu, D. P. Crabb, T. Ho, and D. F. Garway-Heath. More accurate modeling of visual field progression in glaucoma: ANSWERS. *Investigative Opthalmology & Visual Science*, 56(10):6077, Sept. 2015.

[27] H. Zhu, R. A. Russell, L. J. Saunders, S. Ceccon, D. F. Garway-Heath, and D. P. Crabb. Detecting changes in retinal function: Analysis with non-stationary

weibull error regression and spatial enhancement (ANSWERS). *PLoS ONE*, 9(1):e85654, Jan. 2014.