

# MACHINE LEARNING ALGORITHMS FOR IMPROVED GLAUCOMA DIAGNOSIS



**LUND**  
UNIVERSITY

by

Dimitrios Bizios

AKADEMISK AVHANDLING

som för avläggande av filosofie doktorsexamen i ämnet Medicin

vid medicinska fakulteten, Lunds universitet,

kommer att offentligen försvaras i MFC Lilla Aulan, Ingång 59, Universitetssjukhuset  
SUS, Malmö,

Lördagen den 15 okt 2011, kl. 9.15.

Fakultetsopponent:

Professor Anja Tuulonen,

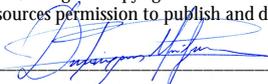
Head Department of Ophthalmology, University of Oulu, Finland

Organization LUND UNIVERSITY  Department of Clinical Sciences in Malmö	Document name DOCTORAL DISSERTATION	
	Date of issue 15 October 2011	
	Sponsoring organization	
Author(s) Dimitrios Bizios		
Title and subtitle Machine Learning Algorithms for Improved Glaucoma Diagnosis		
Abstract  <p>Primary open angle glaucoma, one of the leading causes of blindness in the world, constitutes a slow progressing condition characterized by damage to the optic nerve and retinal nerve fibre layer, and results in visual field defects afflicting the visual function. Highly specific and sensitive diagnostic tests able to detect the clinically significant glaucomatous changes in the structure of the nerve fiber layer and visual field are therefore required for the early detection and management of this disease. This thesis treats the application of advanced statistical techniques based on machine learning for automated classification of tests from visual field examinations and retinal nerve fibre measurements to detect glaucoma. Diagnostic performance of the applied machine learning classification algorithms was shown to depend primarily on the type of test information that was provided. Optimized parameters from standard automated perimetry tests and OCT measurements of the nerve fibre layer derived from statistical processing to highlight significant functional and structural changes, led to improvements in diagnostic accuracy. Moreover, the combination of structural and functional test information through incorporation of a priori knowledge about the anatomical relationship of the retinal nerve fibre layer and the visual field further increased the diagnostic performance of the automated classification algorithms. Machine Learning Classifiers based on optimized test input data could become useful decision support tools for more accurate glaucoma diagnosis.</p>		
Key words: glaucoma, machine learning, perimetry, optical coherence tomography, artificial neural networks		
Classification system and/or index terms (if any):		
Supplementary bibliographical information:		Language English
ISSN and key title: 1652-8220		ISBN 978-91-86871-20-8
Recipient's notes	Number of pages	Price
	Security classification	

## Distribution by (name and address)

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature



Date

2011 / 09 / 19

# MACHINE LEARNING ALGORITHMS FOR IMPROVED GLAUCOMA DIAGNOSIS



**LUNDS**  
UNIVERSITET

Dimitrios Bizios

ISBN 978-91-86871-20-8

ISSN 1652-8220

Copyright © Dimitrios Bizios

Cover Design: Tien Pham

Layout: Tien Pham

Medicinska fakultet, Avdelning för Oftalmologi

Lunds Universitet 2011

„Das Instrument, welches die Vermittlung bewirkt zwischen Theorie und Praxis, zwischen Denken und Beobachten, ist die Mathematik; sie baut die verbindende Brücke und gestaltet sie immer tragfähiger. Daher kommt es, daß unsere ganze gegenwärtige Kultur, soweit sie auf der geistigen Durchdringung und Dienstbarmachung der Natur beruht, ihre Grundlage in der Mathematik findet. Schon GALILEI sagt: Die Natur kann nur der verstehen der ihre Sprache und die Zeichen kennengelernt hat, in der sie zu uns redet; diese Sprache aber ist die Mathematik, und ihre Zeichen sind die mathematischen Figuren. KANT tat den Ausspruch: „Ich behaupte, daß in jeder besonderen Naturwissenschaft nur so viel eigentliche Wissenschaft angetroffen werden kann, als darin Mathematik enthalten ist.“ In der Tat: Wir beherrschen nicht eher eine naturwissenschaftliche Theorie, als bis wir ihren mathematischen Kern herausgeschält und völlig enthüllt haben. Ohne Mathematik ist die heutige Astronomie und Physik unmöglich; diese Wissenschaften lösen sich in ihren theoretischen Teilen geradezu in Mathematik auf. Diese wie die zahlreichen weiteren Anwendungen sind es, den die Mathematik ihr Ansehen verdankt, soweit sie solches im weiteren Publikum genießt. Trotzdem haben es alle Mathematiker abgelehnt, die Anwendungen als Wertmesser für die Mathematik gelten zu lassen. GAUSS spricht von dem zauberischen Reiz, den die Zahlentheorie zur Lieblingswissenschaft der ersten Mathematiker gemacht habe, ihres unerschöpflichen Reichtums nicht zu gedenken, woran sie alle anderen Teile der Mathematik so weit übertrifft. KRONECKER vergleicht die Zahlentheoretiker mit den Lotophagen, die, wenn sie einmal von dieser Kost etwas zu sich genommen haben, nie mehr davon lassen können. Der grosse Mathematiker POINCARÉ wendet sich einmal in auffallender Schärfe gegen TOLSTOI, der erklärt hatte, daß die Forderung „die Wissenschaft der Wissenschaft wegen“ töricht sei. Die Errungenschaften der Industrie, zum Beispiel, hätten nie das Licht der Welt erblickt, wenn die Praktiker allein existiert hätten und wenn diese Errungenschaften nicht von uninteressierten Toren gefördert worden wären. Die Ehre des menschlichen Geistes, so sagte der berühmte Königsberger Mathematiker JACOBI, ist der einzige Zweck aller Wissenschaft.

Wir dürfen nicht denen glauben, die heute mit philosophischer Miene und überlegenem Tone den Kulturuntergang prophezeien und sich in dem Ignorabimus gefallen. Für uns gibt es kein Ignorabimus, und meiner Meinung nach auch für die Naturwissenschaft überhaupt nicht. Statt des törichtigen Ignorabimus heiße im Gegenteil unsere Losung:

Wir müssen wissen,  
Wir werden wissen.“

“We must not believe those, who today with philosophical bearing and deliberative tone prophesy the fall of culture and accept the Ignorabimus. For us there is no Ignorabimus, and in my opinion none whatever in natural science. In opposition to the foolish Ignorabimus I offer our slogan:

We must know,  
We will know.“

David Hilbert

# CONTENTS

<b>1. Abstract</b> .....	8
<b>2. Original Publications</b> .....	9
<b>3. Abbreviations</b> .....	10
<b>4. Introduction</b> .....	11
4.1 Primary open angle glaucoma .....	11
4.2 Visual field testing – The perimetric examination .....	11
4.3 Testing the morphology of RNFL and ONH – OCT Imaging.....	12
4.4 Machine Learning Classifiers (MLCs) – Artificial Neural Networks & Support Vector Machines .....	13
4.5 Artificial Neural Networks – ANNs .....	14
4.6 Support Vector Machines – SVMs .....	17
4.7 Relevance Vector Machines – RVMs.....	18
<b>5. Aims</b> .....	19
5.1 General Aim.....	19
5.2 Specific Aims.....	19
<b>6. Methods</b> .....	20
6.1 Study Design .....	20
6.2 Subjects .....	21
6.3 Diagnostic Tests.....	24
6.4 MLCs .....	27
6.5 Analyses.....	34
<b>7. Results</b> .....	36

<b>8.</b> Discussion .....	42
<b>9.</b> Conclusions .....	46
<b>10.</b> Populärvetenskaplig Sammanfattning .....	47
<b>11.</b> Acknowledgements .....	49
<b>12.</b> References .....	52
<b>13.</b> Appendix .....	61

# 1. Abstract

Primary open angle glaucoma, one of the leading causes of blindness in the world, constitutes a slow progressing condition characterized by damage to the optic nerve and retinal nerve fibre layer, and results in visual field defects afflicting the visual function. Highly specific and sensitive diagnostic tests able to detect the clinically significant glaucomatous changes in the structure of the nerve fiber layer and visual field are therefore required for the early detection and management of this disease. This thesis treats the application of advanced statistical techniques based on machine learning for automated classification of tests from visual field examinations and retinal nerve fibre measurements to detect glaucoma. Diagnostic performance of the applied machine learning classification algorithms was shown to depend primarily on the type of test information that was provided. Optimized parameters from standard automated perimetry tests and OCT measurements of the nerve fibre layer derived from statistical processing to highlight statistically significant functional and structural changes, led to improvements in diagnostic accuracy. Moreover, the combination of structural and functional test information through incorporation of *á priori* knowledge about the anatomical relationship of the retinal nerve fibre layer and the visual field further increased the diagnostic performance of the automated classification algorithms. Machine Learning Classifiers based on optimized test input data could become useful decision support tools for more accurate glaucoma diagnosis.

## 2. Original Publications

- i. Bengtsson B, **Bizios D**, Heijl A: Effects of input data on the performance of a neural network in distinguishing normal and glaucomatous visual fields. *Invest Ophthalmol Vis Sci* 2005, 46: 3730-3736.
- ii. **Bizios D**, Heijl A, Bengtsson B: Trained artificial neural network for glaucoma diagnosis using visual field data: a comparison with conventional algorithms. *J Glaucoma* 2007, 16: 20-28.
- iii. **Bizios D**, Heijl A, Hougaard JL, Bengtsson B: Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by Stratus OCT. *Acta Ophthalmol Scand* 2010, 88: 44-52.
- iv. **Bizios D**, Heijl A, Bengtsson B. Integration and Fusion of Standard Automated Perimetry and Optical Coherence Tomography Data for Improved Automated Glaucoma Diagnostics. *BMC Ophthalmology* 2011, 11:20 (doi:10.1186/1471-2415-11-20).

### 3. Abbreviations

<b>ANN</b>	–	Artificial Neural Network
<b>AROC</b>	–	Area under the Receiver Operating Characteristic curve
<b>asb</b>	–	apostilbs
<b>dB</b>	–	decibel
<b>GHT</b>	–	Glaucoma Hemifield Test
<b>HFA</b>	–	Humphrey Field Analyzer
<b>LTSA</b>	–	Local Tangent Space Alignment
<b>MD</b>	–	Mean Deviation
<b>MLC</b>	–	Machine Learning Classifier
<b>MLP</b>	–	Multi-Layer Perceptron
<b>µm</b>	–	micrometres
<b>mW</b>	–	milliwatt
<b>NTG</b>	–	Normal Tension Glaucoma
<b>OCT</b>	–	Optical Coherence Tomography
<b>ONH</b>	–	Optic Nerve Head
<b>PCA</b>	–	Principal Component Analysis
<b>PD</b>	–	Pattern Deviation
<b>PEX</b>	–	Pseudoexfoliation Glaucoma
<b>PG</b>	–	Pigment Glaucoma
<b>POAG</b>	–	Primary Open Angle Glaucoma
<b>PSD</b>	–	Pattern Standard Deviation
<b>RNFL</b>	–	Retinal Nerve Fibre Layer
<b>RNFLT</b>	–	Retinal Nerve Fibre Layer Thickness
<b>ROC</b>	–	Receiver Operating Characteristic
<b>RVM</b>	–	Relevance Vector Machine
<b>SAP</b>	–	Standard Automated Perimetry
<b>SD-OCT</b>	–	Spectral Domain Optical Coherence Tomography
<b>SITA</b>	–	Swedish Interactive Threshold Algorithm
<b>SVM</b>	–	Support Vector Machine
<b>TD</b>	–	Total Deviation
<b>TD-OCT</b>	–	Time Domain Optical Coherence Tomography

## 4. Introduction

### 4.1 Primary open angle glaucoma

POAG comprises the most frequent type of glaucoma afflicting the visual function of individuals and is the third leading cause of blindness worldwide<sup>1</sup>. It is a progressive condition characterized by damage of the RNFL and optic nerve head, and resulting in visual field defects<sup>2-4</sup>. According to epidemiological data, the prevalence of POAG is increasing with age and in white populations<sup>5</sup> over 70 years of age is estimated at 6%. The incidence of the disease in mainly European populations is estimated to about 0.1% - 0.2% per year<sup>6</sup>. Main clinical signs of POAG are alterations of the optic nerve head topography and structural defects of the RNFL, visual field defects corresponding to the anatomical organization of the RNFL, and in circa 50% of cases increase in intraocular pressure<sup>7-10</sup>. The level of intraocular pressure, initially considered a diagnostic criterion, is currently viewed as the main independent risk factor for the onset of POAG<sup>11</sup> and its progression<sup>12,13</sup>. Structural alterations of the ONH and RNFL, as well as visual field defects, are the most important signs of the onset of POAG. Early diagnosis and management of POAG has gained support following the results of large clinical randomized trials indicating the positive effect of intraocular pressure lowering therapy on the progress of the disease<sup>12,13</sup>. Highly specific and sensitive diagnostic tests able to detect the clinically significant glaucomatous changes are therefore required for the early detection of POAG.

### 4.2 Visual field testing – The perimetric examination

Visual field loss in glaucoma – initially manifested as localized variability in perceived light sensitivity - may develop early and progress gradually over time, long before the patient perceives any abnormalities or loss of vision. Thus, examination of the visual field is important both for the diagnosis of open-angle glaucoma and for following its progression in order to construct an appropriate therapy plan. Perimetry is the method used to examine visual fields. Initially introduced in the 1800's<sup>14,15</sup> the first perimetric methods able to determine the extent of the visual field were further developed by

Bjårrum to become capable of discovering specific patterns of visual field defects<sup>16</sup>. The subsequent development of static perimetry<sup>17</sup> enabled quantitative measurements of light sensitivity in the visual field. The advent of computerization during the 1970's evolved the perimetric method into its current form (SAP) by automating both the presentation of light stimuli and the registration of patient responses. The automated static perimeter functions by determining the intensities of the dimmest light stimulus that can be seen in specific pre-selected test point locations across the visual field. In this way, threshold values of differential light sensitivity are measured. Statistical analysis of the raw threshold data can facilitate the identification of significant visual field changes<sup>18</sup>. Such analysis is based on the creation of a normative dataset, from tests on healthy individuals, and enables corrections that account for the effects that age and the presence of media opacities have on test measurements.

### 4.3 Testing the morphology of RNFL and ONH – OCT Imaging

Degenerative changes of the RNFL have been shown to occur at very early stages of glaucoma<sup>19 – 23</sup>. Defects of the peripapillary RNFL and alterations of ONH morphology can be observed by examining mono- or stereoscopic photographs of these structures. Evaluation of RNFL and ONH photographs by experts is of course subjective and dependent on the skill of the individual examiner<sup>24,25</sup>. Advances in imaging technology enabled the quantitative description of retinal structures and led to the development of diagnostic instruments that depict structural details at high resolutions and with good reproducibility<sup>26,27</sup>. Imaging methods like confocal scanning laser ophthalmoscopy, scanning laser polarimetry and optical coherence tomography, provide morphological information such as thickness measurements of the RNFL and ONH.

OCT is a technique used for the characterization of semi-transparent structures and in-vivo, non-invasive, imaging of biological tissues<sup>28,29</sup>. It is based on the principle of interferometry and utilizes a near-infrared, broad bandwidth light source with a short coherence length. The OCT instrument is able to provide cross-sectional images of tissue structures by analyzing the light that is reflected (scattered) back from the different tissue elements. Due to the short coherence length of the light source, the axial resolution of OCT images is very high. The initial OCT instruments used an interferometric setup where the interference pattern was resolved in time (TD-OCT). Recent iterations of OCT (SD-OCT) utilize instead a detector setup that registers the light at different frequencies and then analyses the different light spectra by Fourier transformation. This

approach leads to improved scan acquisition times and scan quality (i.e. higher signal to noise ratio).

Several studies investigated the ability of the TD-OCT parameters to discriminate between normal and glaucomatous eyes<sup>30–38</sup>. RNFLT parameters in particular, have shown better discriminative ability than OCT ONH measurements<sup>35,39,40</sup> and have exhibited sensitivities between 70% and 80% at specificity levels of more than 90% and AROCs of about 0.90. The best diagnostic performance was provided by RNFLT parameters derived from the inferior and superior peripapillary RNFL quadrants and their associated clock hour sectors, as well as from the average RNFLT value of the whole OCT scan circle test pattern.

#### 4.4 Machine Learning Classifiers (MLCs) – Artificial Neural Networks & Support Vector Machines

In the field of artificial intelligence, machine learning is an active area of research concerned with the development of computational methods that are able to learn to perform classification, clustering and regression tasks through a training process<sup>41</sup>. For classification tasks, machine learning algorithms are able to adapt their decision boundaries based on the data that is presented during the training process, in contrast to conventional statistical methods with explicitly defined functional parameters. MLCs have been successfully used in a variety of fields<sup>42</sup>, including medicine<sup>43–45</sup>, for automated interpretation of medical diagnostic tests<sup>45–48</sup> and modeling of biological systems<sup>44</sup>. In ophthalmology and the area of glaucoma diagnosis in particular<sup>45</sup>, MLCs have been used for classification of tests based on visual field data<sup>51–61</sup>, structural measurements of the RNFL and ONH<sup>62–69</sup>, as well as for detecting progression of glaucomatous visual field defects<sup>70–72</sup>, and combining functional and structural diagnostic parameters<sup>73–76</sup>. Moreover, the high diagnostic accuracy of MLC methods has been favorably compared against the performance of traditional linear discriminant analyses<sup>77–78</sup> and human experts<sup>58,79</sup>.

In order to broadly evaluate the diagnostic performance of MLCs, this thesis compares three supervised MLCs with different architectures and learning methods – artificial neural networks, support vector machines and relevance vector machines. Supervised classification techniques such as artificial neural networks were employed due to their efficient learning and well-documented performance in classification problems<sup>80</sup>. Recently developed statistical learning methods using kernels (support vector machines)

or a Bayesian framework (relevance vector machines) have been proposed as more optimally trained classifiers compared to artificial neural networks. We chose to use the above techniques as they have already been used as decision support systems in ophthalmology<sup>81,82</sup> and have demonstrated high diagnostic accuracies in various benchmark datasets<sup>83</sup>.

## 4.5 Artificial Neural Networks – ANNs

The ANNs constitute a class of machine learning algorithms whose creation was inspired by the attempts to mathematically model the function of biological neural networks. ANNs however are not exact functional representations of biological neural networks. Taking the structure of neural tissue as an analogy, ANNs are composed of a number of processing elements called artificial neurons, with their connection nodes approximating the axons and dendrites, their connection weights approximating the synapses, and their threshold functions approximating the activity in the biological neuron's soma. The learning process in biological neural networks is accordingly modelled by the ANNs through the incremental adjustment of the values of their connection weights.

Following the work of early researchers<sup>84</sup>, Rosenblatt provided a mathematical description of the function of the single artificial neuron and introduced the perceptron<sup>85</sup>. Learning of the perceptron on a set of training data occurs by changing the value of the perceptron weights ( $W$ ) in proportion to the difference (error) of the target (correct) output  $Y$ , and the perceptron output  $y$ , for each training example presented (Figure 1). The architecture and learning law of the perceptron, however, imposed limitations in its classification ability, resulting in accurate discrimination of objects that are only linearly separable.

# The Perceptron

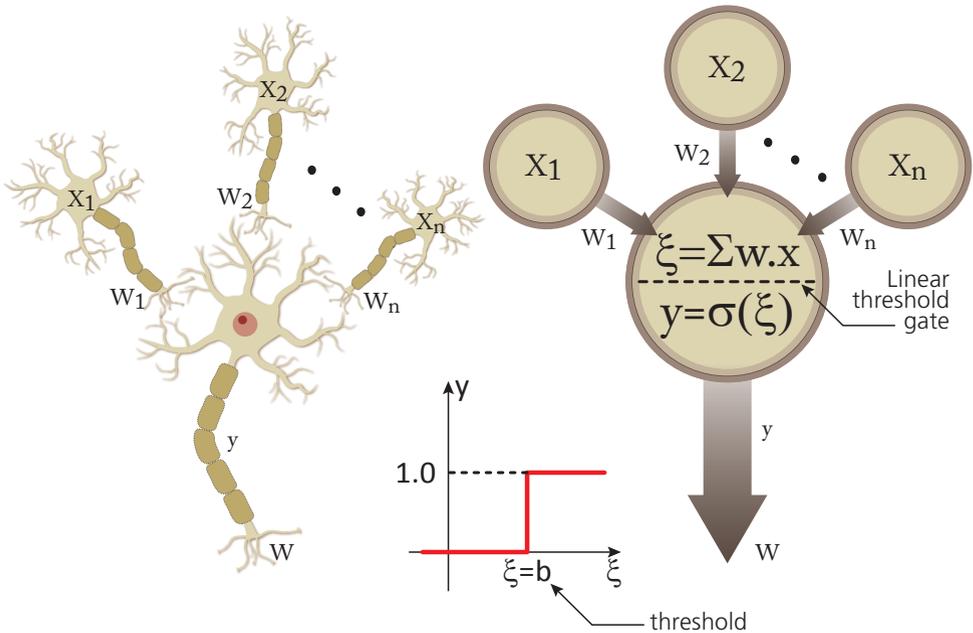


Figure 1. Illustration of the interaction between  $n$  biological neurons with signals of intensity  $X$  and synaptic strength  $W$  feeding into a neuron with a threshold  $b$  and producing an output  $y$ . As an analogy to the connections of biological neurons, the Perceptron receives input that is the product of the input value  $X$  and its connecting weight  $W$ . The transfer function allows the perceptron to produce an output  $y$  when the sum of weighted inputs is larger than the specified threshold  $b$ .

In order to cope with nonlinearly separable objects, additional layers of neurons can be placed between the input layer and the output neuron, leading to the widely used MLP architecture and its learning algorithm based on backpropagation of error<sup>86</sup>. The widespread use of MLPs is mainly due to their ability to conduct nonlinear classification tasks and their efficient learning algorithms. The general structure of an MLP (Figure 2) includes an input layer (representing the input data or variables of the problem into the network), one more hidden layers which form representations of the relevant information and enable the construction of the MLP decision boundaries during the training process, and the output layer consisting of one or more nodes that produces the solution (output) of the network to the specific problem modelled. The transfer function of each neuron in the network is usually a continuous differentiable sigmoid function (e.g. logistic or hyperbolic tangent).

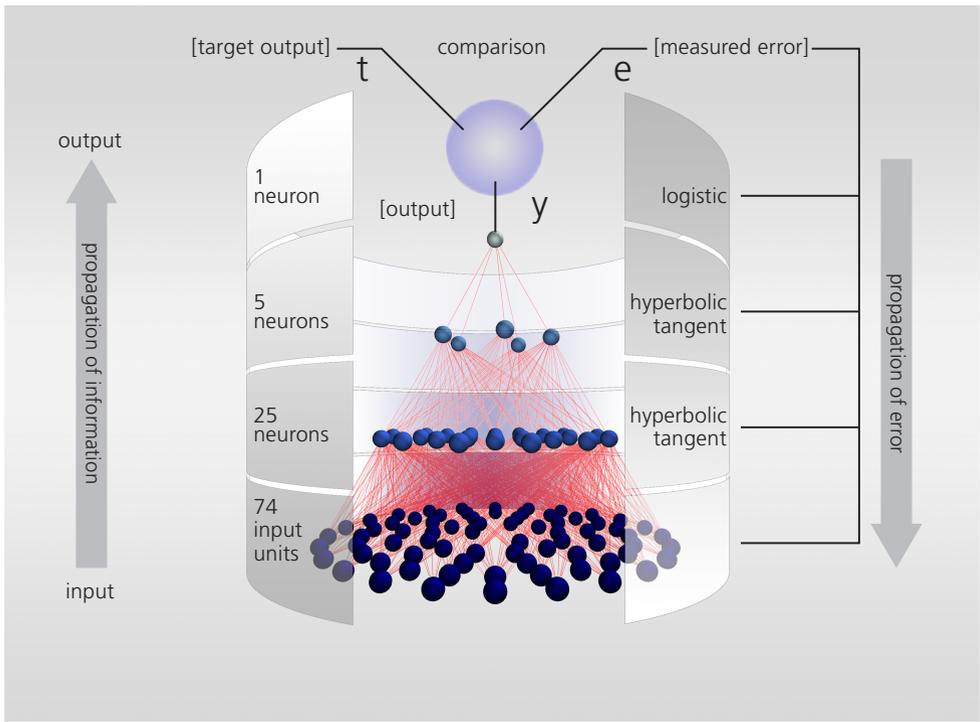


Figure 2. Example of an MLP with an input layer of 74 units, 2 hidden layers of 25 and 5 neurons respectively and 1 output neuron. During each training cycle (epoch), information is passed forward from the input to the output layer, and the MLP classification output ( $y$ ) is compared to the known label (target output,  $t$ ) of the training examples. The error ( $e$ ) of the MLP from this comparison is backpropagated to the connections of each layer in order to adjust the connection weights according to an error function that tries to minimize the MLP classification error during each training epoch. The MLP undergoes numerous training epochs and is considered to be trained when its error approaches zero or falls under a specified value.

In a backpropagation MLP, the data from the input layer nodes are weighted by the connections, summed, and transformed by the transfer function in order to be used as input into the next layer. The same process continues forward until it reaches the output node(s). The solution generated by the output layer, is compared with the desired (correct) output value of the example from the training data set. The measured error is passed in a backward fashion from the output layer to the hidden and input layers, in order to adjust the connection weights between the neurons. Each example is processed in the same way until the whole training set is presented (training epoch/cycle). This procedure is repeated until there is a sufficiently low error rate (convergence). The trained network's ability to generalize is tested with a set of data containing different examples from those of the training set.

## 4.6 Support Vector Machines – SVMs

Following developments in statistical learning theory<sup>87</sup>, new techniques for classification and regression were introduced within the group of machine learning algorithms. SVMs are kernel-based methods that, like ANNs, can be trained to recognize patterns in data and adapt their decision boundary to the training data. Unlike ANNs, these algorithms perform classification by using kernels to map the input data in a space of higher dimensionality and, with the help of constructed support vectors (from part of the training data), they create hyperplanes that maximize the separation between the classes while minimizing the generalization error (Figure 3).

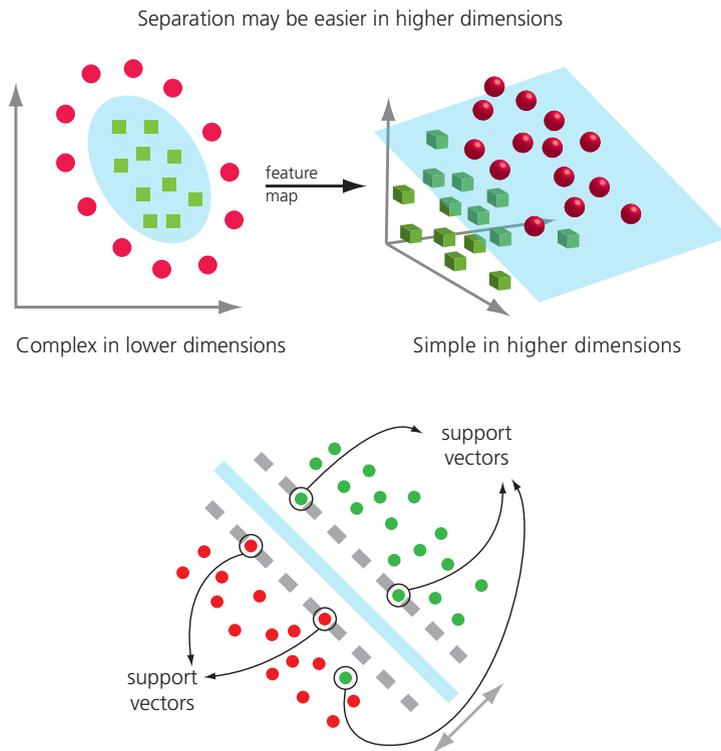


Figure 3. An illustration of the principle of SVM classification. By projecting data into a higher dimensional parameter space it can be easier to construct hyperplanes that can separate the different classes of data. The construction of the decision boundary for the SVM is based on a subset of training examples (support vectors) that lay close to the decision boundary (i.e. belong to a set of examples that are most difficult to correctly classify).

## 4.7 Relevance Vector Machines – RVMs

Another type of MLC that has produced promising results is the Relevance Vector Machine (RVM). RVM is a sparse linear model formulated within the Bayesian framework<sup>88</sup>. Even though its functional form resembles the SVM, the two algorithms are based on different principles. RVM provides a probabilistic output that is easier to interpret in the context of test classification compared to the SVM output of class membership. RVM also uses a sparse representation requiring fewer relevance vectors to create decision boundaries, compared to the number of support vectors used by the SVM. On the other hand, RVM training is highly non-linear, making optimal results hard to achieve.

## 5. Aims

### 5.1 General Aim

The aim of this thesis is the development of improved glaucoma diagnostic tests through the utilization of new automated techniques for interpretation of perimetric data and measurements of the RNFLT with OCT. The employed techniques are based on the MLC paradigm and constitute methodologies related mainly to ANNs.

### 5.2 Specific Aims

1. Investigate the effect of using raw perimetric measurements and processed SAP parameters on the performance of an ANN classifier. Identify the type of SAP parameter that optimizes the performance of our automated classifier for the visual field data set.
2. Confirm the classification accuracy of the previously trained best performing ANN classifier on an independent material of SAP tests from healthy persons and patients with mild and moderate glaucomatous visual field defects. Compare the performance of the ANN classifier with that of other algorithms commonly used for diagnosis of glaucoma based on visual field data.
3. Evaluate the ability of conventional and novel parameters, derived from OCT RNFLT measurements, to provide relevant diagnostic information when used as input data for MLCs. Optimize the MLC architecture and RNFLT parameters that provide the best results in terms of diagnostic accuracy.
4. Investigate ways to combine the best performing SAP and OCT RNFLT parameters. Construct and evaluate the diagnostic performance of an MLC-based method that can utilize the complementary information from the functional SAP data and structural measurements of OCT.

## 6. Methods

### 6.1 Study Design

#### Studies I and II

Both were case-control studies evaluating data collected prospectively from healthy individuals and retrospectively from patients with POAG. The glaucoma definition in the first two studies, where SAP tests were used as input data to an ANN, was based on glaucomatous changes of the optic nerve head evaluated by an expert from optic disc photographs and/or comprehensive descriptions of optic nerve head appearance found in patient records.

#### Study III

The third case-control study was a retrospective analysis of prospectively collected OCT test data, taken from randomly chosen healthy individuals and glaucoma patients followed at the Department of Ophthalmology at Skåne University Hospital, Malmö, Sweden. The definition of glaucoma required both functional and corresponding structural glaucomatous defects to be present. Similarly, the included healthy individuals had normal visual function and a healthy RNFL judged by RNFL and /or ONH photographs.

#### Study IV

This case-control study is based on analysis of prospectively collected data from a random population sample of healthy individuals residing in a defined catchment area of southern Sweden and glaucoma patients followed at the Department of Ophthalmology at Skåne University Hospital, Malmö Sweden. For both healthy persons and glaucoma patients, the definition of normality and glaucoma was based on evaluation of the optic disc during fundus examination.

All four studies were conducted according to the tenets of the Declaration of Helsinki and were approved by the Regional Ethical Review Board of Lund, Sweden. In the studies where clinical data were prospectively collected, all healthy individuals and glaucoma patients provided informed consent prior to any examinations.

## 6.2 Subjects

All glaucoma patients included in the four studies had POAG, normal tension glaucoma, pseudoexfoliation glaucoma or pigment glaucoma. Patients with other types of glaucoma such as angle-closure glaucoma, congenital, or other secondary forms of glaucoma, were excluded. Additionally, neuro-ophthalmic or other systemic disorders (except diabetes mellitus without retinopathy) as well as retinal diseases that affected the visual field or the RNFL constituted grounds for exclusion. All four studies were based on data derived from only one randomly chosen eye per included individual.

### Selection of individuals and test data

#### *Study I*

Since media opacities in the form of cataract are often present in patients with glaucoma it was deemed important to train our ANN with SAP tests belonging to healthy individuals and glaucoma patients with or without the presence of cataract. The following four groups were thus created:

#### 1. Healthy persons

SAP test data from healthy individuals, derived from a large normative database created by a multicenter study with the aim to establish the normal thresholds and limits for the SITA algorithms<sup>89</sup>, consisting of 335 healthy persons. The inclusion of individuals depended on their ophthalmic status after a clinical examination, and not on the results of their visual field examination. However, suspicious or pathologic visual fields consistent with the ocular status, or due to obvious test artifacts, were excluded. From this database, 213 tests from 213 subjects were randomly selected as the normal material for the first study. The mean age of the healthy group providing the 213 tests was 52 years (range 19 to 84 years) and the average MD was -0.02 dB, ranging from -6.11 to +3.07 dB.

#### 2. Patients with media opacities

We identified 55 patients with a diagnosis of cataract, and descriptions of a normal ONH as well as normal previous visual field examinations, in their medical records. After removal of those with unreliable field test results, mostly due to poor fixation, 41 eyes of 41 patients remained in this group. The mean age in this group was 77 years (range: 54 to 96), with MD ranging from -9.82 to -2.46 dB. The visual field tests from this group were grouped together with the 213 normal tests.

## Selection of groups with glaucoma

We randomly selected 30-2 SITA Standard SAP tests from one of our HFA databases (containing 11,134 tests of 3,629 patients) and matched the patients corresponding to the selected tests to our glaucoma register. Only tests from patients with a diagnosis of glaucoma or suspect glaucoma were included for further analysis. Additionally, first-time visual field tests were not chosen, so as to avoid perimetric training effects<sup>90-92</sup>. This process yielded 643 SAP tests. Review of medical records of patients corresponding to the selected tests was carried out and only tests of patients having a glaucomatous ONH (either in photographs and/or comprehensively described in the patient records) prior to the SAP test date were included. Depending on the presence or absence of cataract as described on the patients' medical records, the following 2 groups were formed:

### 3. Patients with glaucoma

Patients with records having no description of cataract or noting a clear lens or the existence of intraocular lens implant (without posterior capsule opacification) were regarded as patients having glaucoma without cataract. In this way, 127 SAP tests of 127 patients were selected. The mean age of the patients in this group was 75 years (range: 40 to 96), and the MD ranged from -31.18 to +0.74 dB.

### 4. Patients with concomitant glaucoma and cataract

Patients with medical records indicating the presence of any type or stage of cataract were classified as having concomitant glaucoma and cataract. This group consisted of 68 tests. The mean age of this group was 77 years (range: 51 to 97), and the MD values ranged from -29.99 to -0.12 dB.

## *Study II*

### 1. Healthy persons

From the SITA normative database<sup>89</sup> consisting of 335 healthy individuals, after the exclusion of 213 tests used in the first study, 122 SAP tests of 122 individuals remained. From this set of tests, 6 were lost due to corrupted data leaving 116 SAP tests of 116 individuals to form the healthy group. The mean age of this group was 51 years (range: 19 to 83), and the MD ranged between -4.62 and +2.4 dB (mean: +0.08 dB).

## 2. Glaucoma patients

Randomly selected 30-2 SITA Standard SAP tests from one of our HFA databases (containing ca 25,000 tests of ca 6000 patients), and the corresponding patient records were retrieved. In this way, 588 visual field tests from 588 patients were extracted. Since we wanted to test our trained ANN classifier in patients with mild and moderate glaucomatous changes, only SAP tests with a mean deviation (MD) value (rounded to the nearest integer) better than or equal to -10 dB were included. SAP tests with MD values worse than -10dB were instead replaced with earlier SAP tests from the same patient if the MD value was within the specified range. Unreliable tests were excluded, as well as tests of patients that participated in the first study. None of the included SAP tests were the patients' first tests, so as to avoid erroneous test results due to lack of perimetric experience<sup>90-92</sup>. After application of all criteria, 100 SAP tests from 100 patients were included. The presence of cataract was not taken into account during the selection process. Of the included patients, 28% had media opacities (26% in the form of cataract and 2% in the form of postoperative opacities of the posterior capsule) at the time of test acquisition. The average age was 75 years (range: 41 to 95), and MD ranged from -10.42 to +0.31 dB (mean: -5.77 dB).

### *Study III*

#### 1. Healthy persons

This group was formed by persons recruited mainly by a random selection of presumably healthy individuals living in Malmö, Sweden. The collected OCT measurements formed a database that was divided into 2 parts. Two-thirds of the database (178 persons) was used for the construction of a normative RNFLT model with reference limits corrected for both age and refraction<sup>93</sup>. This normative database was used to correct for age and refractive status all TD-OCT RNFLT measurements of the remaining one-third (90 persons) that was subsequently used as the normal group for training our MLCs.

#### 2. Glaucoma patients

Glaucoma patients followed at the Department of Ophthalmology, Skåne University Hospital, Sweden during the last 3 years prior to this study were recruited. After review of the corresponding medical records, persons between 40 and 80 years of age, having POAG, NTG, PEX or PG and not involved in other ongoing studies were invited to participate. After application of all inclusion criteria, 62 patients with glaucoma were included. Four of those were newly diagnosed with glaucoma during recruitment of the presumably healthy persons. If both eyes of each patient were eligible, the eye with the better MD value on SAP tests was included. All included eyes had reproducible visual

field defects on SITA Standard 24-2 tests corresponding to glaucomatous changes in the ONH and / or the RNFL as judged by examination of photographs. The included eyes had SAP tests with MD better than -12 dB, PSD outside the 95% normal limit and classified by GHT as falling outside normal limits.

### *Study IV*

#### 1. Healthy Individuals

We performed a random selection from a population register containing 4,718 persons over 50 years, living in two primary care catchment areas of Scania, Sweden. This selection yielded a sample of 307 individuals who were invited to participate in the study. Of those, 170 individuals accepted the invitation and underwent a comprehensive ophthalmic examination including SAP testing and TD-OCT imaging. All included persons had VA > 0.5, refractive errors < 5 D sphere and/or < 3 D cylinder, and a healthy appearance of the ONH as judged by a trained physician during fundus biomicroscopy. In this way, 125 healthy persons were included.

#### 2. Glaucoma Patients

The initial recruitment was based on a random selection of 397 patients with a diagnosis of POAG, PEX, NTG or PG from a register of 2,174 visits of patients with these diagnoses, followed at the Department of Ophthalmology, Skåne University Hospital, Sweden between January 2<sup>nd</sup> 2007 and March 13<sup>th</sup> 2008. After review of the corresponding patient medical records and exclusion of patients with confounding ocular or systemic pathological conditions, 164 patients were invited to participate and underwent a comprehensive ophthalmic examination, including SAP testing and RNFL imaging with TD-OCT. After further exclusion of patients with VA < 0.5, refractive errors of > 5 D sphere and/or > 3 D cylinder, unreliable SAP tests and artefactous OCT images or errors on OCT RNFLT analysis, 135 patients remained. Eight of those patients were newly diagnosed with glaucoma, detected during recruitment of healthy individuals, and were included in the glaucoma group.

## 6.3 Diagnostic Tests

### 6.3a Standard Automated Perimetry – SAP

The most broadly used automated perimeter in both research and clinical praxis is the HFA (Carl Zeiss Meditec, Dublin CA). All visual field data in this thesis were collected with the HFA II.

## Perimetric threshold tests

SAP tests with the HFA use projected white stimuli of varying intensity over a range of 5.1 log units (51 dB) in the range of 0.08 to 10,000 asb, and usually with a stimulus size corresponding to the 0.43-degree Goldmann size III stimulus. Thresholds of differential light sensitivity are measured by projection of the standardized stimuli against an illuminated background of 31.5 asb brightness for 200 ms in specified locations of the visual field. The perimeter is also able to detect the point of fixation of the patient's gaze in order to ensure proper gazing during the test. This is achieved by either the blind spot monitoring technique or the gaze tracker monitoring system<sup>94</sup>.

The calculation of differential light sensitivity is based on repeated presentation of stimuli at varying intensities across a threshold level. This threshold level can be defined as the level of light intensity where each presented stimulus has a 50 % probability to be perceived. In order to provide a reduction of test time while maintaining or improving the level of measurement accuracy, the more recent SITA test strategy, employs advanced mathematical modelling of the visual field and statistical processing of patient responses<sup>95-97</sup>. The SITA algorithm allows for appropriate selection of both the intensity of stimulus presentation and the pace and length of the inter-stimulus interval based on analysis of patient responses and their statistical consistency. Moreover, by recording all test-related parameters, SITA can accurately estimate threshold values based on the whole pattern of patient responses. In studies I and II, we used the SITA Standard 30-2 program that measures light sensitivity at 76 locations of the central visual field within 30 degrees from the point of fixation. In study IV, we used the SITA Standard 24-2 program, which provides sensitivity measurements at 54 locations of the central visual field within 24 degrees from the point of fixation.

## Statistical analysis of the threshold tests – the deviation plots

Interpretation of SAP test results by examination of the numerical threshold sensitivity measurements can be a difficult task for a physician. The main reason is that the range of normal threshold values in each test point varies by a different amount, without following any theoretical Gaussian distribution. Thus in order to correctly distinguish normal from pathological visual fields, every physician should possess knowledge of the normal sensitivity ranges at each test point in the visual field. STATPAC is a statistical analysis package that incorporates this type of knowledge in the interpretation of SAP tests in the HFA<sup>18</sup>. In the Single Field Analysis format of the standard threshold test, STATPAC highlights any sensitivity values that deviate from normal, by comparing the threshold measurements with age-corrected measurements from pre-constructed

databases containing tests from perimetrically healthy individuals. Apart from the raw threshold numerical values with the corresponding greyscale representation, STATPAC's output includes test reliability parameters, global indices for the visual field, deviation plots, and the GHT labelling<sup>98</sup>.

The TD deviation numerical plots indicate the deviation of each measured threshold value after comparison with the age-corrected values from the normative database. The significance of this deviation compared to the instrument's normative database is illustrated on the TD probability maps. The PD analysis (based on the TD highest sensitivity values) denotes deviation of sensitivity at each test point location after adjustment to remove any generalized depression of light sensitivity from the hill of vision. The PD probability plot and associated probability maps highlight in this way localized loss of sensitivity and are able to detect visual field defects earlier than the greyscale printouts<sup>99</sup>, and de-emphasize common artifactual patterns<sup>100</sup>.

### 6.3b Optical Coherence Tomography – OCT

OCT data in this thesis were obtained with the only commercially available TD-OCT instrument (StratusOCT, Carl Zeiss Meditec, Dublin CA), which has been extensively used in both research and the clinical environment since its introduction in 2003.

In TD-OCT, the standard examination for the detection of glaucomatous structural changes entails the scanning of the peripapillary RNFL in a circular pattern with a diameter of 3.4 mm centred on the ONH, in axial sections with a resolution of circa 10 $\mu$ m. The acquired raw reflectivity measurements from each A-scan are then processed by the instrument in order to correct for signal noise and motion generated signal artifacts. A segmentation algorithm provides the boundaries of the RNFL based on the reflectance intensity profiles of the image, and RNFLT measurements of the scan are calculated. In our third and fourth studies, OCT scans were obtained at a beam power of 750 mW with the "Fast RNFLT scan protocol", which produces average thickness values of three circumpapillary scans, each with 256 A-scan measurement points. The OCT instrument provides parameters based on the average RNFLT measurements of the whole scan circle or sectors of it (4 quadrants and 12 clock hour sectors). Even some ratios between average values and the highest measurements in specific RNFL scan sectors are being presented in the analysis output of the OCT examination. In study III, we used the best performing of the commercial RNFLT parameters and constructed novel parameters based on the lowest measured thickness values and their percentiles, and on A-scan

RNFLT values transformed by the LTSA algorithm. In study IV, we utilized corrected OCT A-scan measurements after processing by PCA.

## 6.4 MLCs

### 6.4a MLC architectures and training

In the following section, some more technical details concerning the architecture and training method of our MLCs are presented.

#### *ANN structure and training*

The ANNs in the first three studies were fully connected feed forward MLPs with an input layer consisting of a vector of the data values and their input weights, two hidden layers with hyperbolic tangent transfer functions and an output layer of one neuron with a logistic transfer function. The number of neurons in the hidden layers was chosen based on the number of input parameters, and for the first two studies was 25 and 5 neurons respectively. In the third study the ANN configuration was 12 and 6 neurons in each hidden layer respectively. In the fourth study we used a variation of the feed forward MLP, the cascade forward neural network. The general ANN structure is the same with the exception of the input layer, which provides connections to all other layers instead of only the first hidden layer. All ANNs were trained with the scaled conjugate gradient training algorithm developed by Møller<sup>101</sup>. We used a subset of the training data as an early stopping dataset, to avoid overfitting of the networks during training. All ANNs were programmed and run in the Neural Network toolbox of MATLAB (The MathWorks Inc., Natick, MA, USA).

#### *ANN ensembles and training with bagging*

The ensemble approach can be used to decrease the classification error of an ANN classifier by combining the prediction of a number of ANNs<sup>102</sup>. The decomposition of the ANN classification error into the factors of bias (i.e. the classification accuracy on the training data) and variance (i.e. the stability of ANN classification with respect to the variability of the training data) reveals an inverse relationship (i.e. a trade-off) between these two measures. It can be shown that the classification error of the ensemble equals the averaged classification error made by an individual ANN minus the averaged variance (a.k.a. diversity) of the individual ANNs in the ensemble. Training the individual ANNs with a resampling algorithm such as Bootstrap aggregating (bagging) on slightly different subsets of the training data increases the diversity of individual ANNs and thus

decreases the ensemble classification error. The bagging algorithm generates training subsets by uniformly sampling examples from the training data with replacement<sup>103</sup>. The created bootstrap samples (expected to contain 63.2% of unique examples) are then used for training the individual ANNs. This approach can also be incorporated into a cross-validation testing setup.

### *SVM structure and training*

The SVMs utilized a radial basis function kernel and were trained by a variation of Platt's sequential minimal optimization algorithm<sup>104</sup>. Programming, testing and training of the SVM classifiers were done in Python (Python Software Foundation) and MATLAB (The MathWorks Inc., Natick, MA, USA) with the Libsvm software<sup>105</sup>. We used a global optimization technique based on simulated annealing<sup>106</sup> to determine the values for the C and g parameters of the SVM algorithm.

### *RVM structure and training*

The RVMs used a Gaussian kernel with bias and were trained with the first version of the SparseBayes software package for MATLAB (The MathWorks Inc., Natick, MA, USA). The width of the kernel was chosen as the value that provided the best results on 10-fold cross-validation.

### 6.4b MLC training and testing by cross-validation

We employed the 10-fold cross-validation procedure<sup>107</sup> in order to maximize the use of our collected data without the need of a completely separate test subset, and avoid the bias of simultaneous training and testing on the same individual tests. In this cross-validation setup, all test data were randomly divided into 10 subsets, with each subset containing approximately the same amount of healthy and glaucomatous tests. One subset was used for testing classification performance, and the remaining 9 subsets were used for training the MLCs. In training of the ANNs, one of the 9 training subsets was used for early stopping of network training in order to avoid overfitting. The remaining 8 subsets provided the training data. In the ANN ensemble, the training data were created with bagging from these 8 subsets. ANN training was repeated by keeping the same test subset and changing the early stopping set, until all training data was used both in training and early stopping of the ANNs and the classification error of the networks was averaged. The training and testing process for the MLCs was iterated, each time with a different test set, and the results were merged to produce a single average output for each type of machine classifier.

## 6.4c Input Data to MLCs

### *SAP data*

In the first study we used the raw threshold numerical values, as well as the TD and PD numerical plots and probability maps. In order to use the TD and PD probability maps as input data, we represented each probability level by numerical values based on a scoring procedure. The scoring scheme was adopted from the process of calculating the GHT output<sup>98</sup>. The SAP data used as MLC input in the second and fourth study were based on the scored probabilities of the PD probability maps. The SITA Standard 30-2 and 24-2 test patterns provided 74 and 52 scored PD probability values respectively (excluding 2 test point measurements falling on the blind spot).

### *OCT data*

In the third and fourth study, all OCT RNFLT measurements were derived by the instruments' peripapillary RNFL scan circle examination protocol. We performed corrections for age and refraction on all the collected A-scan data prior to any analyses. These corrections were accomplished by linear regression analysis on a model of the relationship of age, refraction and measured RNFLT on a normative database<sup>93</sup>, and use of the derived coefficients to calculate the corrected values of measured RNFLT. In the third study, we used the A-scan values to calculate mean, highest and lowest RNFLT values for the whole scan circle as well as for different RNFL sectors (quadrants and clock hour sectors). We then derived the commercially available OCT RNFLT parameters in the same way that these are produced by the StratusOCT instrument. The performance of the age- and refraction corrected A-scan measurements was also investigated after processing to reduce their complexity (i.e. their high number of parameters).

### *Dimensionality reduction of OCT A-scan data*

The A-scan data of OCT acquired with the FAST protocol require 256 parameters (i.e. dimensions) to be represented. The problem with high-dimensional data lays in the fact that one needs very large number of tests in order to successfully train MLCs, something that is impractical in medical research studies. These difficulties can be ameliorated by simplifying the representation of high-dimensional data using techniques able to map the large number of values into a set of fewer parameters (i.e. a lower dimensional parameter space). We have examined the application of both linear (study IV) and non-linear (study III) dimensionality reduction methods on the A-scan measurements.

### *Linear Dimensionality Reduction – PCA*

PCA<sup>108</sup> is a well-described, widely used technique for reducing complexity in datasets. The PCA method is essentially an orthogonal linear transformation of data. It transforms the dimensions of the data into a new dimensional space where the first dimension maps the largest variance of any projection of the data; the second dimension maps the second largest variance of the projected data, etc. Parameters that contain useful and relevant information, add to the variance of data in a dataset, in contrast to redundant parameters. The function of PCA allows for the maximum possible variance in the data to be mapped in the first few dimensions (principal components). During dimensionality reduction with PCA, the few initial principal components describe the largest part of variance of the original dataset and the remaining parameters can be overlooked without significant loss of information. Since PCA is depended on the multidimensional mean, it is very sensitive to the scale of each parameter. This is important to consider when having data from different sources with different measurement scales. In the fourth study, we used PCA to reduce the parameters of OCT A-scan measurements, and included the principal components that retained a large amount of relevant information (99.9% of variation) from the original measurements.

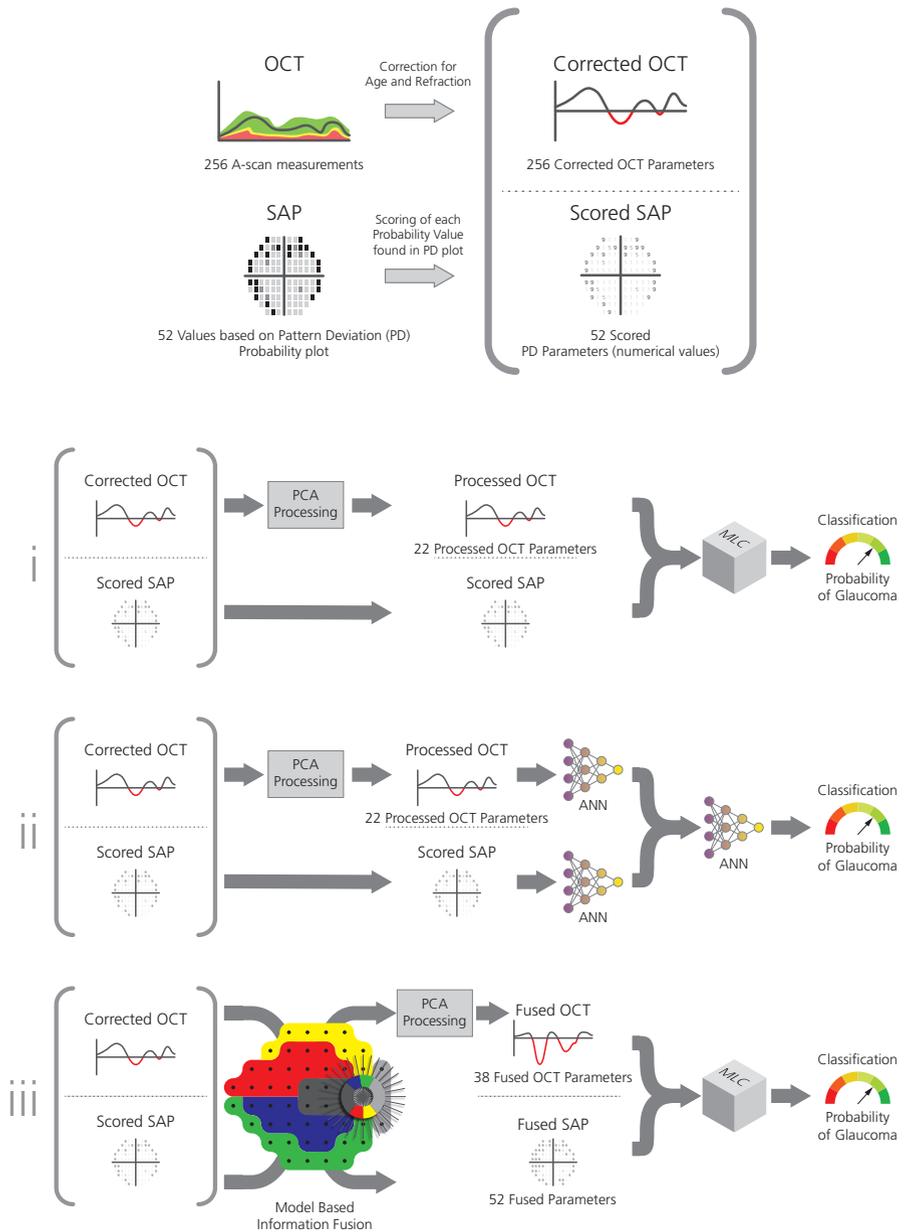
### *Nonlinear dimensionality reduction – LTSA*

In study III, we used the LTSA algorithm<sup>109</sup>. LTSA is a non-linear dimensionality reduction method that has performed very well in other datasets<sup>110</sup> and is more robust to the choice of its function parameters compared to similar techniques. It belongs to the group of sparse spectral dimensionality reduction methods and is a local embedding technique. It functions by linearly mapping the high-dimensional measurement points to their local tangent space and finding low-dimensional representations whose aligned linear mappings reconstruct the same local tangent space. The number of reduced A-scan parameters after the application of LTSA was based on the estimation of the intrinsic dimensionality of the OCT A-scan data, calculated by a maximum likelihood estimator.

### *Combinations of SAP and OCT data*

In the fourth study, we investigated three different ways of combining SAP test data and OCT RNFLT measurements (Figure 4) in an attempt to further increase the ability of our MLCs to diagnose glaucoma. Based on combinations of the best performing SAP and OCT parameters that were discovered in the previous studies, these approaches were:

- i.** The simple combination of 52 SAP PD probability scores from SITA Standard 24 – 2 tests with 22 OCT parameters derived from PCA processing of the RNFLT A-scan measurements, to form an input vector of 67 parameters. The 67 values were then used as input to an ANN ensemble and an RVM classifier.
- ii.** The construction of a 2-stage ANN classifier consisting of 2 ANNs in the first stage with each network receiving input from either SAP or OCT data, and providing output used as input for the second stage ANN.
- iii.** The fusion of SAP and OCT measurements based on a model relating sectors of the peripapillary RNFL to areas of the visual field. We tested the performance of the 52 fused SAP (F-SAP) parameters and the 38 fused OCT (F-OCT) derived from PCA, as well as their combination (i.e. 90 fused parameters), by using them as input to an ANN ensemble and an RVM classifier.



OCT: Optical Coherence Tomography SAP: Standard Automated Perimetry PCA: Principal Component Analysis MLC: Machine Learning Classifier

Figure 4. Three different approaches to combination of SAP- and OCT derived data: The first approach is based on simple combination of measurements for the creation of the input data vector to the MLC (i), the second approach is based on a 2-stage ANN classifier (ii), while the third approach is based on a model driven data fusion of OCT and SAP before the integration of the fused measurements as input data to the MLC (iii).

### Model-based fusion of data

The fusion of OCT RNFLT measurements and SAP PD scored probability values was based on the map constructed by Garway-Heath et al<sup>111</sup> to represent the topographical relationship between sectors of the peripapillary RNFL and areas of the visual field. Accordingly, the OCT RNFL scan circle was divided into 6 sectors, with the A-scan measurements of each sector corresponding to SITA Standard 24-2 test point locations of a specified area in the visual field (Figure 5). All OCT A-scan thickness measurements were translated into probabilities, based on the calculated normal distribution of RNFLT values derived from a separate database<sup>93</sup>. The fusion process represented a weighting scheme of the age- and refraction-corrected OCT A-scan measurements in each test with the corresponding scored PD values from SAP.

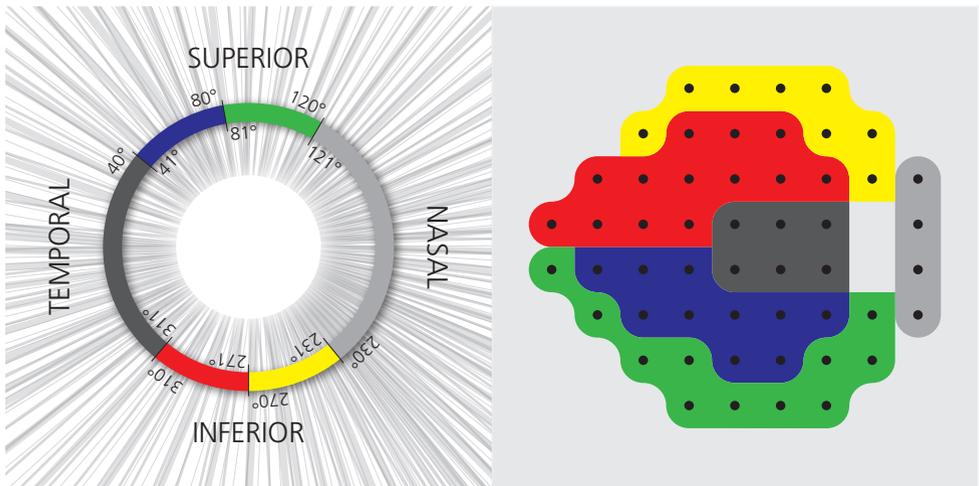


Figure 5. The map of Garway-Heath et al<sup>111</sup> illustrating the relationship between sectors of the peripapillary RNFL and areas of the visual field tested in SAP.

### Fusion of OCT data

For every A-scan position in each of the 6 OCT sectors, the corresponding A-scan probability values falling below the fifth percentile of the RNFLT distribution from our normative database were multiplied with an exponential factor. This factor was the mean pattern deviation probability score (i.e. the sum of all PD scores divided by the number of test points) of the visual field sector corresponding to the OCT scan circle sector. PCA was subsequently applied to the fused OCT A-scan values and provided 38 principal components that were used as input to the MLCs.

### *Fusion of SAP data*

For every visual field sector, the PD probability score value of each test point was transformed by an additive factor that was derived from the A-scan probability values of the corresponding OCT sector. The A-scan probability values were subsequently scored in a manner similar to the calculation of the GHT<sup>98</sup>. The lowest scored probability value below the fifth percentile or the highest scored probability value above the ninety-fifth percentile of our normal RNFLT distribution from each OCT sector was used as the factor in the fusion process. The fused SAP parameters were obtained by adding this factor to the SAP PD probability score value of each SAP test point in the corresponding area of the visual field. In the event that scored probability values outside both the fifth and ninety-fifth percentile of our normative RNFLT database existed in the same OCT sector, only the lowest probability value was used as the additive factor.

## 6.5 Analyses

Using the number of correct classifications (true-positive and true-negative results), as well as the number of incorrect classifications (false-positive and false-negative results) we calculated the sensitivity and specificity of each MLC. Both measures depend on the position of the cut-off limit for defining a field as glaucomatous or normal (i.e. a value between 0 and 1) over the range of the MLC output. Plotting the sensitivity and specificity pairs for all cut-off limits produces a ROC curve<sup>112</sup>. The AROC is a measure of the diagnostic accuracy of a classifier since it represents the probability that a randomly selected test from either the normal or the glaucoma group will be accurately classified<sup>113</sup>. The largest possible AROC has a value of 1, indicating perfect accuracy of classification, whereas an AROC of 0.5 indicates classification accuracy no better than chance. Comparison between AROCs for examining significant differences in the performance of our MLCs is accomplished in all four studies with DeLong's non-parametric method<sup>114</sup>.

Calculation of confidence intervals at the 95% significance level was based on a normal approximation of a binomial distribution, according to the score method<sup>115</sup>. Diagnostic accuracy values were calculated by dividing the sum of true positive and true negative responses with the sum of true and false, positive and negative responses<sup>116</sup>. Significance testing for differences in gender distribution between the healthy persons and patients with glaucoma was accomplished with the Chi square test. The Mann-Whitney test was used for significance testing on the variables of age, visual acuity and refractive error between the healthy and glaucoma groups (studies III and IV). Diagnostic accuracy

of the SAP and OCT parameters was compared with the McNemar test for correlated proportions (study IV).

# 7. Results

## Studies I & II

We found significant performance differences between the different SAP parameters used as input data. The largest AROC was produced with the PD probability scores (0.988), while the smallest AROCs belonged to TD probability scores and numerical values (0.943 and 0.942 respectively). Our ANN trained on the PD probability scores performed significantly better ( $p < 0.001$ ) compared to the ANN using raw threshold sensitivities as input data (Figure 6).

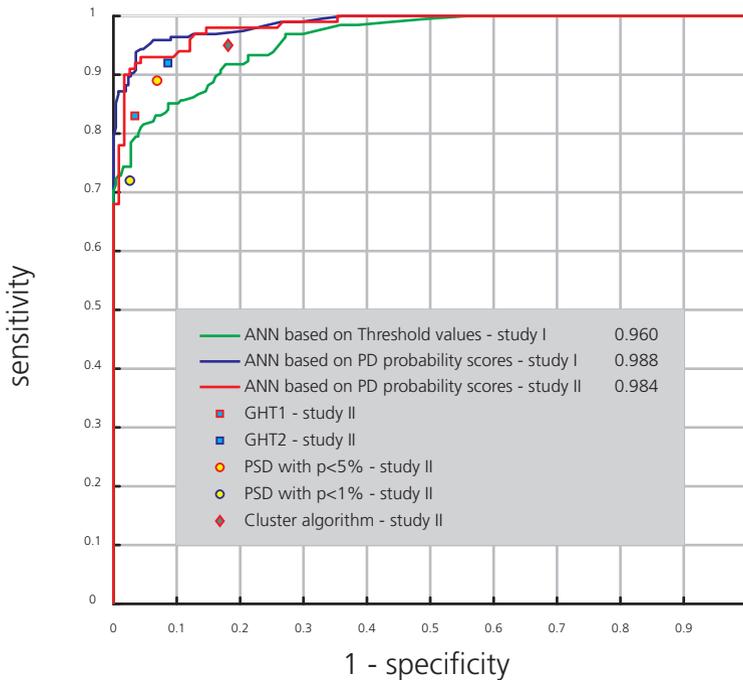


Figure 6. AROCs of ANN based on the raw threshold values and the best performing ANN based on PD probability scores in the first study. The latter ANN was subsequently tested on a new set of SAP tests in study II and achieved similar performance. The specificity and sensitivity of other SAP interpretation algorithms (Glaucoma Hemifield Test – GHT, Pattern Standard Deviation – PSD and Cluster of defects algorithm) for interpretation of SAP tests is shown for comparison.

In study II, our ANN previously trained on PD probability scores was tested on an independent dataset and achieved similar performance (AROC: 0.984). With a diagnostic accuracy of 93.5% and sensitivity and specificity of 93.0% and 94.0% respectively, it provided the best trade-off between sensitivity and specificity compared to other commonly used interpretation algorithms.

### Study III

The performance of MLCs based on all conventional and new parameters is presented in Table 1. The novel input formed from A-scan values transformed by LTSA provided the largest AROCs of all tested parameters for both MLCs (Figure 7).

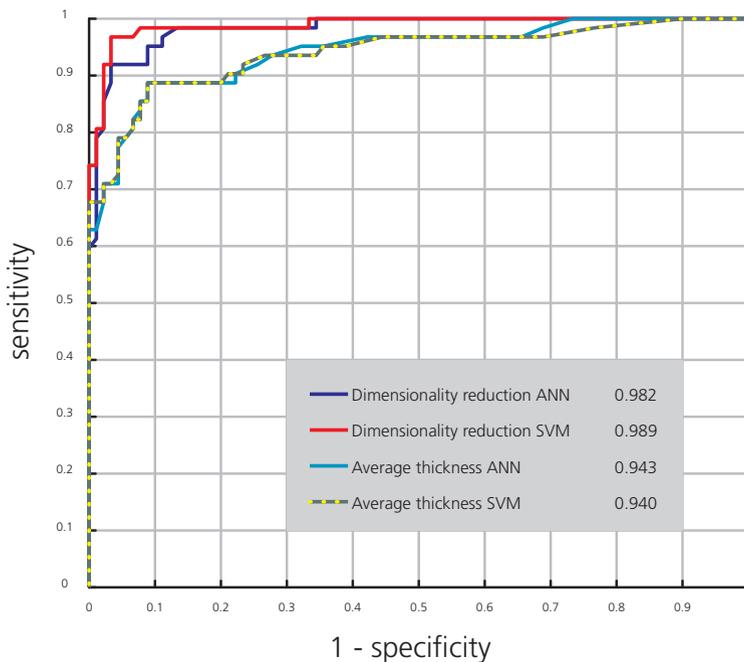


Figure 7. AROCs of MLCs based on the LTSA transformed A-scans, and the best performing single value conventional RNFLT parameter (average RNFLT).

The SVM trained on the LTSA transformed data (sensitivity of 96.8% at specificity of 96.7%) performed significantly better than MLCs trained on the best single commercial parameter (full circle average thickness,  $p = 0.028$ ). SVM performed also significantly better than average RNFLT of the full scan circle without the use of MLCs ( $p = 0.013$ ). Novel parameters based on the thinnest measurements of RNFLT or on percentiles of

measured thickness performed at least as good as the commercially available parameters. Comparison of AROCs for all studied RNFLT parameters, revealed no significant differences between the ANN and SVM classifiers.

Table 1. AROCs of the MLCs for all conventional and novel parameters

RNFLT Parameters	ANN	SVM
full circle average	<b>0.943</b>	0.940
temporal quadrant	<b>0.766</b>	0.757
superior quadrant	<b>0.926</b>	0.922
nasal quadrant	<b>0.789</b>	0.783
inferior quadrant	<b>0.930</b>	0.922
kl 8	<b>0.778</b>	0.777
kl 9	<b>0.640</b>	0.601
kl 10	<b>0.771</b>	0.713
kl 11	<b>0.935</b>	0.933
kl 12	0.833	<b>0.844</b>
kl 1	<b>0.825</b>	0.821
kl 2	0.788	<b>0.796</b>
kl 3	0.703	0.666
kl 4	0.747	<b>0.758</b>
kl 5	0.806	<b>0.811</b>
kl 6	<b>0.929</b>	0.912
kl 7	<b>0.887</b>	0.877
all 17 parameters	<b>0.977</b>	<b>0.977</b>
all clock hour sectors	<b>0.977</b>	<b>0.977</b>
all quadrants	<b>0.959</b>	0.955
best 2 hours	0.970	<b>0.976</b>
best 2 quadrants	<b>0.961</b>	0.959
Smax	<b>0.876</b>	0.861
lmax	<b>0.898</b>	0.896
90% of Smax (S_90)	<b>0.885</b>	0.865
90% of lmax (l_90)	<b>0.91</b>	0.904
Smin	<b>0.919</b>	0.908
lmin	<b>0.916</b>	0.906
10% over Smin (S_10)	<b>0.916</b>	0.909
10% over lmin (l_10)	<b>0.915</b>	0.909
Max – Min	<b>0.942</b>	0.927
Max_90 – Max_10	<b>0.946</b>	0.940
L TSA-transformed A-scans	0.982	<b>0.989</b>

Smax: The highest measured RNFLT in the superior quadrant of the OCT scan circle

Smin: The lowest measured RNFLT in the superior quadrant of the OCT scan circle

Max – Min: The difference between the highest and lowest measured RNFLT of the OCT scan circle

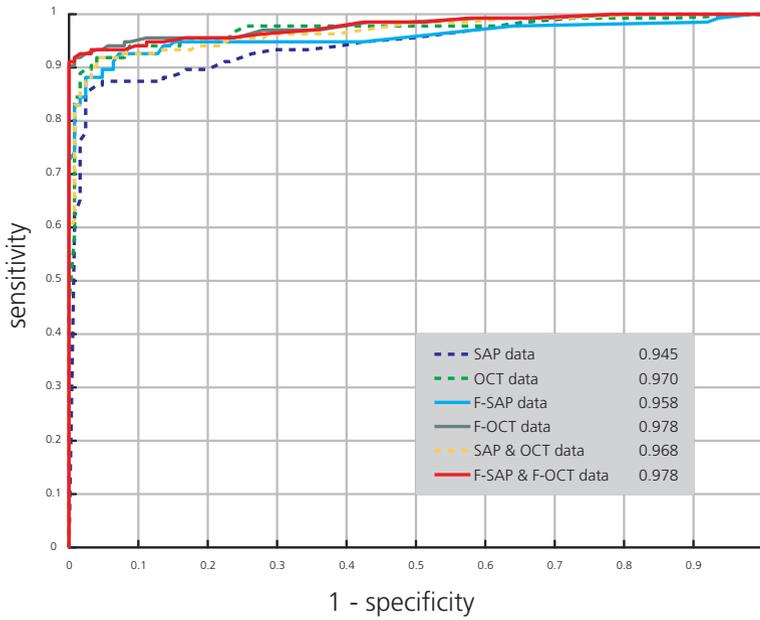
Max\_90 – Min\_10: The difference between the highest and lowest 10<sup>th</sup> percentile of measured RNFLT of the OCT scan circle

## Study IV

The simple combination of SAP and OCT measurements did not lead to significant improvements in the performance of MLCs. The 2-stage ANN model provided very similar results to the simple combination of data. The data fusion approach provided the best results. ANNs based on the fused OCT and the combined fused OCT and SAP data respectively provided almost identical AROC values of 0.978, performing better than the ANN based on SAP measurements alone ( $p=0.047$ ). RVM produced results similar to the ANN classifier. The AROCs of ANNs and RVMs based on the fused and non-fused parameters are shown in Figure 8 (page 40).

The use of fused parameters as input, improved the agreement in classification (reflected by the odds ratios) between SAP- based and OCT-based ANNs. This improvement led to a larger number of tests correctly classified by both function-and structure-based MLCs (Figure 9, page 41).

## AROCs for ANN



## AROCs for RVM

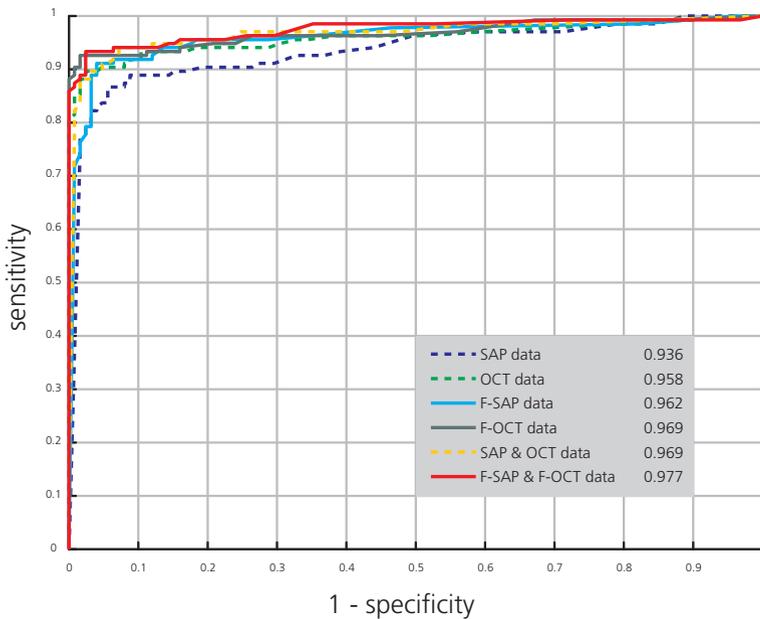
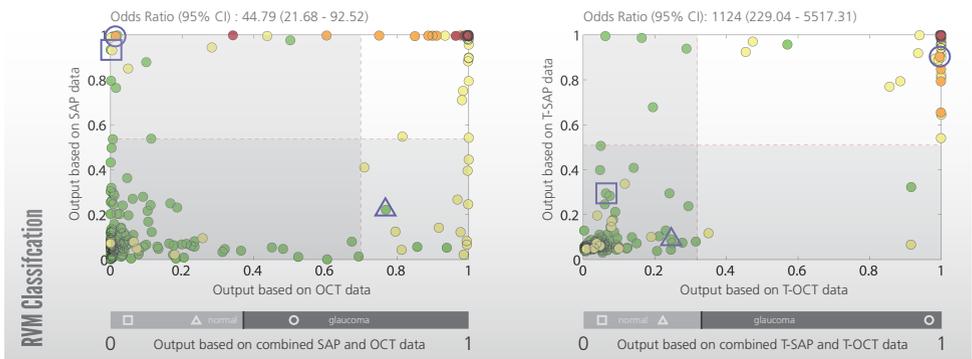
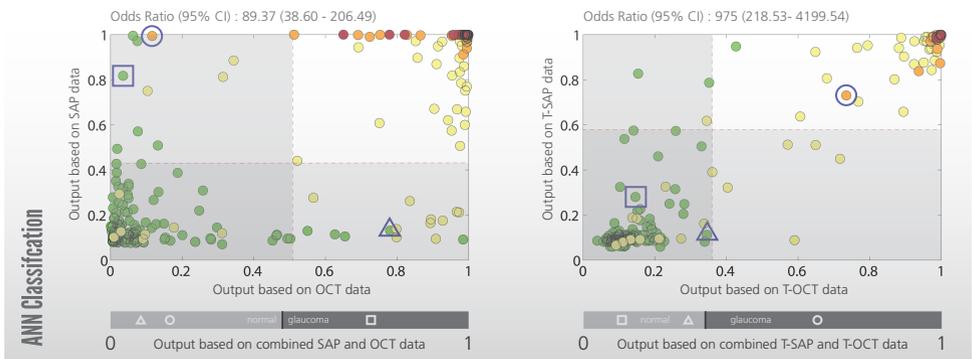
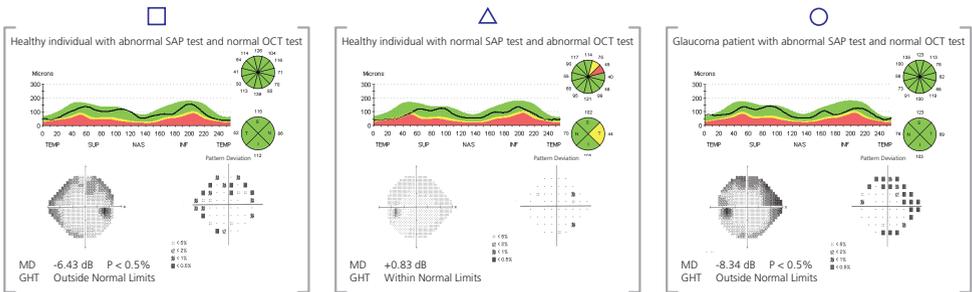


Figure 8. AROCs for both MLCs based on the fused and non-fused SAP and OCT data.



● Normal  
 ● Early glaucoma  
 ● Moderate glaucoma  
 ● Advanced glaucoma

SAP data: Standard Automated Perimetry data, based on Pattern Deviation (PD) probability scores  
 F-SAP data: Fused SAP data, based on weighted transformation of PD probability scores with OCT-derived probability scores  
 OCT data: Age and refraction-corrected Optical Coherence Tomography A-scan data, optimized by principal component analysis (PCA)  
 F-OCT data: Fused OCT data, based on weighted transformation of A-scan data with PD probability scores and optimized by PCA

Figure 9. Three examples (two healthy individuals and one glaucoma patient) with disagreement between the SAP-based and OCT-based ANN and RVM classification results, which not evident when fused OCT (F-OCT) and fused SAP (F-SAP) parameters are used as input data. The odds ratios signify the chance that a test will be classified as normal or abnormal by both SAP- and OCT based ANNs. ANN and RVM classification results using combined OCT and SAP as well as combined F-OCT and S-SAP data are also shown under each diagram.

## 8. Discussion

The primary aim of this thesis was to investigate the potential of automated diagnostic algorithms based on machine learning to detect glaucomatous changes, thus assisting clinicians in recognizing the onset of glaucoma earlier and with higher diagnostic accuracy. To this extent, we investigated the effect in classification performance of different types of input data and different MLC-based architectures. Our results show that MLCs are capable of detecting glaucomatous defects in the visual field and the RNFL with high diagnostic accuracy. MLC performance depends much more on the type of input parameters used and their optimization, than the type of employed MLC architecture.

*ANNs based on statistically processed data can diagnose glaucoma from perimetric test measurements of the visual field (studies I and II)*

We have shown that ANN classifiers can discriminate between healthy and glaucomatous visual field tests with high degree of accuracy. Our results were based on a large sample of 449 persons. We intended to investigate the effect that different types of SAP input had on the performance of automated classifiers, since all previous studies utilized only the raw unprocessed threshold sensitivity values as input data. We focused on the deviation plots of SAP since other indexes such as MD or PSD only provide a summary description of the visual field status and omit spatial information that could be important for the recognition of glaucoma-related patterns of visual field defects.

The benefits of selecting STATPAC parameters from SAP tests depend on their statistical processing that highlights significant changes while accounting for factors affecting SAP test measurements. Thus, the ability of scored PD probability values to provide age-corrected significance limits that highlight localized depressions of VF sensitivity while accounting for the presence of media opacities, could explain the performance improvement exhibited by our ANN in the first two studies.

Since media opacities in the form of cataract are often present in the population of patients with glaucoma, it is important for the practical applicability of MLC methods to be able to detect those patterns of glaucomatous defects despite the presence of other confounding conditions affecting the status of the visual field.

*Evaluation on an independent material showed the generalization ability of trained ANNs in discriminating between glaucomatous and healthy visual fields (study II)*

In the second study, our goal was to validate the trained ANN classifier on a new independent set of data from healthy persons and patients with mild and moderate glaucoma. The majority of glaucoma diagnostic studies investigating the performance of MLCs conduct the training and testing of the algorithms on the same groups of subjects. Although resampling techniques like the cross-validation procedure maximize the use of available data and decrease the bias of training and testing the algorithms on the same group of individuals, separate testing on a completely independent material is a better indicator of the MLC generalization ability.

Our study, the first in the area of glaucoma diagnosis to test a trained automated classifier on a completely separate set of data, showed that the ANN was able to generalize very well, achieving the same level of performance as in the cross-validation setting during training and initial testing (study I). The best threshold for the ANN found from cross-validation was also shown to be the best threshold for the network even on the new material, indicating that there was no significant overfitting during the training process.

Circa one-third of the glaucomatous SAP data were derived from patients with media opacities. The performance of our ANN did not degrade during testing on this new set of SAP fields, irrespective of the presence or absence of media opacities, probably due to our use of optimized input in the form of PD probability scores. Moreover, the ANN was tested on a group of patients having only mild and moderate visual field defects. Even though it is important to provide the ANN with examples representing the whole spectrum of glaucomatous defects during training, the clinically relevant situations where MLCs could function as decision support systems do not encompass cases of advanced glaucoma with obvious visual field defects.

*The used of novel RNFLT parameters enabled MLCs to accurately detected glaucoma-related changes of the RNFL measured with TD-OCT (study III)*

Both types of MLCs (ANNs and SVMs) were able to accurately distinguish between the normal and glaucomatous OCT tests based only on RNFLT information, providing very similar results. Our results showed that when examining the OCT derived RNFLT parameters, the input selection and optimization of parameters significantly affected the performance of MLCs. The commercially available measurements from StratusOCT such as mean-RNFLT measured over the whole scan circle or sectors of it are only sum-

mary values of all RNFLT measurements associated to every individual A-scan location, and can lead to loss of relevant information. On the other hand, the use of all A-scans without pre-processing, introduces a very large number of parameters for each test to be classified. This large number of parameters would necessitate a very large numbers of training examples in order to avoid overfitting during the training process. Overfitting occurs when MLCs memorize the pattern of training examples instead of learning the underlying rules characterizing the different classes where these examples belong. Utilizing all available information found on the A-scan RNFLT values and optimizing their representation by reducing the number of parameters through dimensionality reduction techniques significantly improved the diagnostic accuracy of our classifiers. Another reason for the high diagnostic accuracy of our MLCs was the implementation of an ensemble structure for the ANN, and a global optimization scheme for selection of the SVM training parameters.

*Knowledge based fusion of visual field and OCT data provide parameters that can increase the accuracy of MLCs in diagnosis of glaucoma (study IV)*

There has been interest in investigating the functional and structural relationship in glaucoma<sup>117-127</sup>. Recent studies have attempted to combine information from structural and functional tests in order to improve the classification accuracy of glaucoma diagnostic systems<sup>73-76</sup>. The results from these studies were mixed, but indicated possible advantages of integrating information from different test modalities in order to achieve higher diagnostic performance. In all previous attempts, the integration of information was carried out by simple combination of data from visual field tests and structural data of the ONH and RNFL. Our approach was instead based on utilization of a priori knowledge about the anatomical relationship between the RNFL structure and the visual field. We thus used a morphological model representing this relationship in order to fuse the available structural and functional information in the test measurements, instead of relying on the MLCs to internally construct the same representation based on the limited set of patient data.

Utilizing a fusion process for the individual test measurements, in contrast to creating a global probabilistic estimate of the structure – function effects on test measurements<sup>128</sup>, enables the incorporation of additional knowledge related to the specific advantages of each test modality. This can be accomplished by applying different weighting schemes on the fusion of OCT and SAP data. Accordingly, significantly low OCT RNFLT values could only be accentuated during fusion, by the existence of corresponding defects on the visual field, whereas a normal visual field did not attenuate low RNFLT values due to

the fact that very early glaucoma can be evident on OCT tests without any detectable visual field defects. However, visual field defects shown by the PD probability scores could be either accentuated or attenuated during the fusion procedure, depending on whether a significantly low or high RNFLT existed on the corresponding OCT sector, since discreet visual field defects (e.g. due to learning effects) can occur in healthy persons with a normal RNFL.

Furthermore, the use of PCA on the OCT data enabled the reduction of complexity of the dataset while preserving the relevant information used by the MLCs for classification of the SAP and OCT tests. PCA was chosen because it can be extended to include new data (out of sample extension) without the need to reapply the algorithm to the whole dataset.

## 9. Conclusions

- ANNs can successfully identify glaucomatous defects from SAP tests. The type of SAP parameters, chosen as input data, significantly affected the performance of the classifier. Refined input data, based on PD, resulted in significantly higher accuracy compared to the raw threshold values.
- The high diagnostic performance of ANN based on refined input visual field data was confirmed in an independent sample. The ANN performed at least as good or better than other SAP interpretation algorithms.
- MLCs were able to detect glaucomatous changes on OCT RNFLT measurements with high accuracy, exhibiting similar diagnostic performance. RNFLT input parameters had a much larger impact on diagnostic performance than the type of MLC used. Optimized MLC input based on A-scan RNFLT measurements provided the highest diagnostic accuracy of all tested parameters.
- The utilization of optimized parameters through data fusion of OCT and SAP measurements significantly increased the performance of MLCs, compared to the use of SAP parameters only. Integrating parameters by including a priori relevant information through data fusion could improve MLC classification accuracy compared to currently available methods.

## 10. Populärvetenskaplig Sammanfattning

Grön starr (glaukom) är en relativt vanlig sjukdom som påverkar synfunktionen och utgör den andra vanligaste orsaken till blindhet. Den vanligaste varianten av grön starr, s.k. öppenvinkelsglaukom, är en kronisk sjukdom som drabbar ca 6 % av befolkningen över 70 år. Den kännetecknas av långsamt tilltagande skador på synnerven och näthinnans nervfiberlager med åtföljande bortfall i ögats synfält. Förhöjt ögontryck är en stark riskfaktor för att utveckla glaukom och för försämring av sjukdomen. Stora kliniska studier har visat att sänkning av ögontrycket kan bromsa utvecklingen av sjukdomen, även hos patienter med normalt ögontryck, och därmed förhindra en hastig utveckling av skador i synfältet. Därför är det viktigt att kunna upptäcka glaukom på ett tidigt stadium.

Synfältsundersökningen är en viktig metod för att upptäcka, kartlägga utbredningen och uppskatta allvarlighetsgraden av skadorna i ögat. Synfältsundersökning är även nödvändig för att kunna följa sjukdomens utveckling och effekten av behandlingen. Den statistiska bearbetningen av mätningarna som utförs vid en modern datorstyrd synfältsundersökning hjälper läkarna att tolka undersökningsresultat.

Synfältsundersökningen visar den funktionella skadan, medan avbildning av strukturer visar skadan på ögats vävnader. En modern metod, Optical Coherence Tomography (OCT), använder ljussignaler för att avbilda näthinnans nervfiberlager med mycket hög upplösning och möjliggör upptäckten av diskreta förändringar orsakad av glaukom. Tidiga förändringar i nervfiberlagret kan ofta föregå synfältsskador.

Syftet med detta projekt var att utveckla mer känsliga metoder för diagnostik av glaukom, genom användning av avancerade statistiska metoder baserad på artificiell intelligens, s.k. maskininlärningsklassificerare (MLC), som kan tränas att känna igen sjukliga förändringar i synfältet och i ögats nervfiberlager.

Projektet består av fyra delstudier.

I första studien har vi undersökt hur väl en typ av maskininlärningsklassificerare, Artificial Neural Networks (ANN), kan upptäcka glaukomsador i synfältet. Vi använde olika mer eller mindre förbehandlade data, dvs. råa mätvärden och mätvärden som korrigerades för olika förvillande faktorer såsom ålder och förekomst av lättare grå starr, samt även mätvärden relaterade till normalgränserna, som inmatningsinformation till ANN. Vi fann att en viss typ av förbehandlade data som använder information om hur normala synfält ser ut, gav ett statistisk bättre resultat, dvs. var bättre på att känna igen glaukom i synfält från patienter, och bättre på att känna igen normala synfält från friska individer. I denna studie tränades ANN för detta ändamål.

I andra studien provade vi det färdigtränade ANN på synfält från helt andra personer (både patienter och friska) än de som ingick i den första studien. Vi jämförde också resultatet från ANN med de andra tolkningsmetoder som används idag för att diagnostisera glaukom med hjälp av synfältstest. Resultatet bekräftade den höga diagnostiska träffsäkerheten av ANN som vi såg i den första studien, samt att ANN gav minst lika bra (och antydande till bättre) resultat som andra etablerade tolkningsmetoder.

I tredje studien tittade vi på avbildning av nervfiberlager med OCT. Vi använde två typer av maskininlärningsklassificerare; ANN och Support Vector Machines (SVM) och provade olika typer mätvärden, korrigerade för olika förvillande effekter, precis som vid synfältsundersökningarna. Vi tog också fram nya typer av mätvärden som vi tror skulle kunna vara bättre för diagnos än de som finns kommersiellt tillgängliga i OCT-instrumentet. Resultaten visade att våra nya mätvärden ökade den diagnostiska förmågan av både ANN och SVM jämfört med de gängse mätvärdena, men att typ av maskininlärningsklassificerare inte hade någon, eller väldigt liten, betydelse.

I den fjärde studien använde vi oss av den kunskap vi fått från studier I – III för att kombinera synfältstester och OCT mätningar till en ännu bättre diagnostisk metod. Vi använde de bästa synfälts- och OCT- mätvärdena, och kombinerade dem med hjälp av en modell som beskriver anatomiska samband mellan delar av nervfiberlagret och områden i synfältet. Resultaten visade att de nya kombinerade mätvärdena ökade den diagnostiska träffsäkerheten jämfört med enbart bästa synfälts- eller bästa OCT resultat.

Sammanfattningsvis har vi kunnat visa att maskininlärningsklassificerare förbättrar tolkningen av både synfälts- och OCT undersökningar, samt att en kombinerad analys ytterligare förbättrar den diagnostiska träffsäkerheten. Förbättringar i resultaten förklarar mestadels av typen av mätvärden som används som indata i MLC, och i betydlig mindre grad av typen av MLC metod som används.

# 11. Acknowledgements

The making of this thesis has been an exciting and rewarding journey, made possible by the contributions of many people whom I sincerely thank. I would like to thank in particular:

**Boel Bengtsson**, my tutor. With an exceptional scientific intellect and a sense of humor to match, you taught me how to take my first steps into the research world, while giving me the freedom to explore new ideas and satisfy my scientific curiosity. Your active involvement and enthusiasm (and kicks when needed) has helped me to come this far, and for this I will always be grateful. Thank you for a thoroughly enjoyable ride!

**Anders Heijl**, my co-tutor. With an unparalleled ability to solve problems and come up with great ideas, you always helped move things forward. Your humor and energy were certainly contagious and always welcomed. You have been a source of knowledge and inspiration. Many thanks for providing both the practical means and the support that allowed me to finish this project. It has been a pleasure to work alongside you!

**Ola Engwall**, my friend / programmer / mathematician. This thesis would not have been possible, if not for all the hours of programming and solving the occasional mathematical problems that came up. Your incredible mathematical intellect, programming skills and dedication to this project were instrumental in its completion and for this you have my gratitude. I am really glad that this project gave me the opportunity to work with you and get to know you!

**Johnny Ring**, for your help with all things computer-related. Thank you for helping me get through all the system crashes, hard drive failures, kernel panics, and for preparing so many great posters! I am also grateful for your help with the data collection in Skurup and Anderslöv; you made working there great fun!

**Christina Gustavsson**, for all your administrative help. You are the one that kept everything running smoothly in the background. Thank you also for the company during occasional 5 minute breaks for fresh air!

**Sabina Andersson**, a colleague and friend. Thank you for all the hard work you've put during the data collection in Skåne. Most importantly though, I would like to thank you for the laughs and the support (both the verbal and chocolate variety) when things got rough!

All the doctors of the glaucoma team, especially **Jesper Hougaard**, **Dorothea Peters**, **Johan Aspberg**, for the interesting discussions, the laughs and the support. We'll always have Berlin!

**Elisabeth Agardh**, for all the good advice I have gotten from you, and for the philosophical discussions during coffee breaks!

**Carin Gustavsson**, for all the good times and the stimulating discussions!

**Lillian Tindberg**, **Monica Wollmer**, **Bodil Ölund** and **Gunilla Lundskog** for teaching me the intricacies of conducting visual field tests, helping with patients and being so nice.

The administration at the Eye Clinic, and especially **Anette Lindström** for making sure I had time to do research and always solving any schedule-related problems.

All the doctors and nurses at the Eye Clinic in Malmö for always helping me become a better doctor and making it a fun working environment!

The primary health care centers in Anderslöv and Skurup for their hospitality and all those individuals who accepted our invitation to be examined in our fourth study.

**Tien Pham**, for everything. Without your patience and amazing skills in graphic design this book would not look as good as it does! You are the best! Azazu!

My parents **Nikolaos and Vasiliki**, and my brother **Georgios** for always backing me up. My extended family, **Despina Tzimoula**, **Dimos Chatzoglakis**, **Elli Kaltsidou**, thank you for your support. All our great late-night discussions on subjects not even remotely associated to medicine, have helped me to put things in better perspective.

All my good friends for their encouragement; Special thanks to the "Zukunft!" team members **Isak Johnsson**, **Benny Liberg** and **Georgios Dimiroudis**. Our discussions have always been a source of inspiration to me. I am glad to have met you all!

The work in this thesis has received financial support from the Swedish research council, the foundation of Crown Princess Margareta for visually handicapped, the foundation for visually impaired in the former Malmöhus län, the Järnhardt foundation and through departmental funding from Carl Zeiss Meditec.



## 12. References

1. World Health Organization: Visual Impairment and Blindness: fact sheet no. 282. 2011 April (online; verified 2011-08-31), URL: <http://www.who.int/media-centre/factsheets/fs282/en/index.html>.
2. American Academy of Ophthalmology: Preferred Practice Pattern. Primary open angle glaucoma. American Academy of Ophthalmology 2000, page 3.
3. European Glaucoma Society: Terminology and Guidelines for Glaucoma. 3rd Edition. Editrice DOGMA, Savona Italy 2008, page 95.
4. Tuulonen A, Airaksinen PJ, Erola E, Forsman E, Friberg K, Kaila M, et al: The Finnish evidence-based guideline for open angle glaucoma. *Acta Ophthalmol Scand* 2003, 8:3-18.
5. Rudnicka AR, Mt-Isa S, Owen CG, Cook DG, Ashby D: Variations in primary open-angle glaucoma prevalence by age, gender, and race: A Bayesian meta-analysis. *Invest Ophthalmol Vis Sci* 2006, 47:4254-4261.
6. Leske MC: Open-angle glaucoma - An epidemiologic overview. *Ophthalmol Epidemiol* 2007, 14:166-172.
7. Hollows FC, Graham PA: Intra-Ocular Pressure, Glaucoma, and Glaucoma Suspects in a Defined Population. *Brit J Ophthalmol* 1966, 50:570-586.
8. Armaly MF: Visual Field Defect and Ocular Pressure Level in Open Angle Glaucoma. *Invest Ophth Visual* 1969, 8:105-124.
9. Dielemans I, Vingerling JR, Wolfs RCW, Hofman A, Grobbee DE, Dejong PTVM: The Prevalence of Primary Open-Angle Glaucoma in a Population-Based Study in the Netherlands - the Rotterdam Study. *Ophthalmology* 1994, 101:1851-1855.
10. Grodum K, Heijl A, Bengtsson B: A comparison of glaucoma patients identified through mass screening and in routine clinical practice. *Acta Ophthalmol Scand* 2002, 80:627-631.
11. Gordon MO, Beiser JA, Brandt JD, Heuer DK, Higginbotham EJ, Johnson CA, Keltner JL, Miller JP, Parrish RK, Wilson MR et al: The Ocular Hypertension Treatment Study - Baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol* 2002, 120:714-720.
12. Gordon MO, Kass MA, Heuer DK, Higginbotham EJ, Johnson CA, Keltner JL, Miller JP, Parrish RK, Wilson MR, Stud OHT: The Ocular Hypertension Treatment Study - A randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol* 2002, 120:701-713.
13. Heijl A, Leske MC, Bengtsson B, Hyman L, Bengtsson B, Hussein M, Grp EMGT: Reduction of intraocular pressure and glaucoma progression - Results from the early manifest glaucoma trial. *Arch Ophthalmol* 2002, 120:1268-1279.

14. Von Graefe A: Ueber die Untersuchung des Gesichtsfeldes bei amblyopischen Affectionen. *Archiv für Ophthalmologie* II 1856, 258-298.
15. Aubert H, Foester R: Untersuchungen über den Raumsinn der Retina. *Archiv für Ophthalmologie* III 1857, 1-37.
16. Bjerrum J: Om en tillføjelse til den sædvanlige synsfeltundersøgelse samt om synsfeltet ved glaukom. *Nordisk Ophthalmologisk Tidsskrift* 1889, 2:141-185.
17. Sloan L: Instruments and techniques for the clinical testing of light sense. III. An apparatus studying regional differences in light sense. *Arch Ophthalmol* 22: 233-251, 1939.
18. Heijl A, Lindgren G, Olsson J: A package for statistical analysis of visual fields. In: Greve & Heijl (eds.) *Doc Ophthalmol proc series* 49: 154-168. Proceedings of the 7th International Perimetric Society Meeting 1986. Junk Publishers, Dordrecht.
19. Sommer A, Katz J, Quigley HA, et al: Clinically detectable nerve fiber atrophy precedes the onset of glaucomatous field loss. *Arch Ophthalmol* 1991, 109:77-83.
20. Quigley HA, Katz J, Derick RJ, Gilbert D, Sommer A: An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. *Ophthalmology* 1992, 99:19-28.
21. Tuulonen A, Lethola J, Airaksinen PJ: Nerve fiber layer defects with normal visual fields: do normal optic disc and normal visual field indicate absence of glaucomatous abnormality? *Ophthalmology* 1993, 100:587-598.
22. Harwerth RS, Carter-Dawson L, Shen F, Smith EL III, Crawford ML: Ganglion cell losses underlying visual field defects from experimental glaucoma. *Invest Ophthalmol Vis Sci.* 1999, 40:2242-2250.
23. Kerrigan-Baumrind LA, Quigley HA, Pease ME, Kerrigan DF, Mitchell RS: Number of ganglion cells in glaucoma eyes compared with threshold visual field tests in the same persons. *Invest Ophthalmol Vis Sci.* 2000, 41:741-748.
24. Abrams LS, Scott IU, Spaeth GL, Quigley HA, Varma R: Agreement among Optometrists, Ophthalmologists, and Residents in Evaluating the Optic Disc for Glaucoma. *Ophthalmology* 1994, 101:1662-1667.
25. Sommer A, Quigley HA, Robin AL, Miller NR, Katz J, Arkell S: Evaluation of Nerve-Fiber Layer Assessment. *Arch Ophthalmol* 1984, 102:1766-1771.
26. Schuman JS, Townsend KA, Wollstein G: Imaging of the retinal nerve fibre layer for glaucoma. *Brit J Ophthalmol* 2009, 93:139-143.
27. Lin SC, Singh K, Jampel HD, Hodapp EA, Smith SD, Francis BA, Ducker DK, Fechtner RD, Samples JS, Schuman JS et al: Optic nerve head and retinal nerve fiber layer analysis - A report by the American Academy of Ophthalmology. *Ophthalmology* 2007, 114:1937-1949.
28. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA et al: Optical Coherence Tomography. *Science* 1991, 254:1178-1181.
29. Swanson EA, Izatt JA, Hee MR, Huang D, Lin CP, Schuman JS, Puliafito CA, Fujimoto JG: In-Vivo Retinal Imaging by Optical Coherence Tomography. *Opt Lett* 1993, 18:1864-1866.

30. Budenz DL, Michael A, Chang RT, McSoley J, Katz J: Sensitivity and specificity of the StratusOCT for perimetric glaucoma. *Ophthalmology* 2005, 112:3-9.
31. Chang R, Budenz DL: New developments in optical coherence tomography for glaucoma. *Curr Opin Ophthalmol* 2008, 19:127-135.
32. Hougaard JL, Heijl A, Bengtsson B: Glaucoma detection by Stratus OCT. *J Glaucoma* 2007, 16:302-306.
33. Leung CK, Chong KK, Chan WM, Yiu CK, Tso MY, Woo J, Tsang MK, Tse KK, Yung WH: Comparative study of retinal nerve fiber layer measurement by StratusOCT and GDx VCC, II: structure/function regression analysis in glaucoma. *Invest Ophthalmol Vis Sci* 2005, 46:3702-3711.
34. Medeiros FA, Zangwill LM, Bowd C, Weinreb RN: Comparison of the GDx VCC scanning laser polarimeter, HRT II confocal scanning laser ophthalmoscope, and stratus OCT optical coherence tomograph for the detection of glaucoma. *Arch Ophthalmol* 2004, 122:827-837.
35. Medeiros FA, Zangwill LM, Bowd C, Vessani RM, Susanna R, Jr., Weinreb RN: Evaluation of retinal nerve fiber layer, optic nerve head, and macular thickness measurements for glaucoma detection using optical coherence tomography. *Am J Ophthalmol* 2005, 139:44-55.
36. Shah NN, Bowd C, Medeiros FA, Weinreb RN, Sample PA, Hoffmann EM, Zangwill LM: Combining structural and functional testing for detection of glaucoma. *Ophthalmology* 2006, 113:1593-1602.
37. Sihota R, Sony P, Gupta V, Dada T, Singh R: Diagnostic capability of optical coherence tomography in evaluating the degree of glaucomatous retinal nerve fiber damage. *Invest Ophthalmol Vis Sci* 2006, 47:2006-2010.
38. Wollstein G, Ishikawa H, Wang J, Beaton SA, Schuman JS: Comparison of three optical coherence tomography scanning areas for detection of glaucomatous damage. *Am J Ophthalmol* 2005, 139:39-43.
39. Leung CKS, Chan WM, Yung WH, Ng ACK, Woo J, Tsang MK, Tse RKK: Comparison of macular and peripapillary measurements for the detection of glaucoma - An optical coherence tomography study. *Ophthalmology* 2005, 112:391-400.
40. Schuman JS, Guedes V, Hertzmark E, Wollstein G, Correnti A, Mancini R, Lederer D, Voskanyan S, Velazquez L, Pakter HM et al: Optical coherence tomography measurement of macular and nerve fiber layer thickness in normal and glaucomatous human eyes. *Ophthalmology* 2003, 110:177-189.
41. Mitchell T: *Machine Learning*. WCB/McGraw-Hill 1997, pages 1-2.
42. Kotsiantis SB, Zaharakis ID, Pintelas PE: Machine learning: a review of classification and combining techniques. *Artif Intell Rev* 2006, 26:159-190.
43. Baxt WG: Application of Artificial Neural Networks to Clinical Medicine. *Lancet* 1995, 346:1135-1138.
44. Kononenko I: Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* 2001, 23:89-109.
45. Drew PJ, Ramesh AN, Kambhampati C, Monson JRT: Artificial intelligence in medicine. *Ann Roy Coll Surg* 2004, 86:334-338.
46. Andersson R, Bartosch-Harlid A, Andersson B, Aho U, Nilsson J: Artificial neural networks in pancreatic disease. *Brit J Surg* 2008, 95:817-826.

47. Ohlsson M, Haraldsson H, Edenbrandt L: Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks. *Artificial Intelligence in Medicine* 2004, 32:127-136.
48. Zhu YN, Williams S, Zwiggelaar R: Computer technology in detection and staging of prostate carcinoma: A review. *Med Image Anal* 2006, 10:178-199.
49. Wysoski SG, Benuskova L, Kasabov N: Evolving spiking neural networks for audiovisual information processing. *Neural Networks* 2010, 23:819-835.
50. Bowd C, Goldbaum MH: Machine learning classifiers in glaucoma. *Optometry & Vision Science* 2008, 85:396-396.
51. Keating D, Mutlukan E, Evans A, McGarvie J, Damato B: A backpropagation neural network for the classification of visual field data. *Phys Med Biol* 1993, 38:1263-1270.
52. Madsen EM, Yolton RL: Demonstration of a neural network expert system for recognition of glaucomatous visual field changes. *Mil Med* 1994, 159:553-557.
53. Mutlukan E, Keating D. Visual field interpretation with a personal computer based neural network. *Eye* 1994, 8:321-323.
54. Goldbaum MH, Sample PA, White H, et al: Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci* 1994, 35:3362-3373.
55. Spenceley SE, Henson DB, Bull DR: Visual field analysis using artificial neural networks. *Ophthalmic Physiol Opt* 1994, 14:239-248.
56. Henson DB, Spenceley SE, Bull DR: Spatial classification of glaucomatous visual field loss. *Br J Ophthalmol* 1996, 80:526-531.
57. Lietman T, Eng J, Katz J, et al: Neural networks for visual field analysis: how do they compare with other algorithms? *J Glaucoma* 1999, 8:77-80.
58. Goldbaum MH, Sample PA, Chan K, et al: Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Invest Ophthalmol Vis Sci* 2002, 43:162-169.
59. Sample PA, Chan K, Boden C, et al: Using unsupervised learning with variational Bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci* 2004, 45:2596-2605.
60. Goldbaum MH, Sample PA, Zhang Z, et al: Using unsupervised learning with independent component analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci* 2005, 46:3676-3683.
61. Goldbaum MH: Unsupervised learning with independent component analysis can identify patterns of glaucomatous visual field defects. *Trans Am Ophthalmol Soc* 2005, 103:270-280.
62. Swindale NV, Stjepanovic G, Chin A, Mikelberg FS: Automated analysis of normal and glaucomatous optic nerve head topography images. *Invest Ophthalmol Vis Sci* 2000, 41:1730-1742.
63. Poinosawmy D, Tan JC, Bunce C, et al: The ability of the GDx nerve fibre analyser neural network to diagnose glaucoma. *Graefes Arch Clin Exp Ophthalmol* 2001, 239:122-127.
64. Mardin CY, Hothorn T, Peters A, Junemann AG, Nguyen NX, Lausen B: New glaucoma classification method based on standard Heidelberg retina tomo-

- graph parameters by bagging classification trees. *J Glaucoma* 2003, 12:340-346.
65. Hothorn T, Lausen B: Bagging tree classifiers for laser scanning images: a data- and simulation-based strategy. *Artif Intell Med* 2003, 27:65-79.
  66. Zangwill LM, Chan K, Bowd C, Hao J, Lee TW, Weinreb RN, Sejnowski TJ, Goldbaum MH: Heidelberg retina tomograph measurements of the optic disc and parapapillary retina for detecting glaucoma analyzed by machine learning classifiers. *Invest Ophthalmol Vis Sci* 2004, 45:3144-3151.
  67. Huang ML, Chen HY: Development and Comparison of Automated Classifiers for Glaucoma Diagnosis Using Stratus Optical Coherence Tomography. *Invest Ophthalmol Vis Sci* 2005, 46:4121-4129.
  68. Burgansky-Eliash Z, Wollstein G, Chu T, et al: Optical Coherence Tomography Machine Learning Classifiers for Glaucoma Detection: A Preliminary Study. *Invest Ophthalmol Vis Sci* 2005, 46:4147-4152.
  69. Bowd C, Medeiros FA, Zhang Z, Zangwill LM, Hao J, Lee TW, Sejnowski TJ, Weinreb RN, Goldbaum MH: Relevance vector machine and support vector machine classifier analysis of scanning laser polarimetry retinal nerve fiber layer measurements. *Invest Ophthalmol Vis Sci* 2005, 46:1322-1329.
  70. Tucker A, Vinciotti V, Liu X, Garway-Heath D: A spatio-temporal Bayesian network classifier for understanding visual field deterioration. *Artif Intell Med* 2005, 34:163-177.
  71. Brigatti L, NouriMahdavi K, Weitzman M, Caprioli J: Automatic detection of glaucomatous visual field progression with neural networks. *Arch Ophthalmol* 1997, 115:725-728.
  72. Sample PA, Boden C, Zhang Z, Pascual J, Lee TW, Zangwill LM, Weinreb RN, Crowston JG, Hoffmann EM, Medeiros FA et al: Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields. *Invest Ophthalmol Vis Sci* 2005, 46:3684-3692.
  73. Brigatti L, Hoffman D, Caprioli J: Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol* 1996, 121:511-521.
  74. Bowd C, Chiou C, Hao J, Racette L, Zangwill LM, Medeiros FA, Lee TW, Weinreb RN, Goldbaum MH, Sample PA: Relevance vector machine for combining HRT II and SWAP results for discriminating between healthy and glaucoma eyes. *Acta Ophthalmol Scand* 2006, 84:569.
  75. Bowd C, Hao J, Tavares IM, Medeiros FA, Zangwill LM, Lee TW, Sample PA, Weinreb RN, Goldbaum MH: Bayesian machine learning classifiers for combining structural and functional measurements to classify healthy and glaucomatous eyes. *Invest Ophthalmol Vis Sci* 2008, 49:945-945.
  76. Racette L, Chiou CY, Hao J, Bowd C, Goldbaum MH, Zangwill LM, Lee TW, Weinreb RN, Sample PA: Combining functional and structural tests improves the diagnostic accuracy of relevance vector machine classifiers. *J Glaucoma* 2010, 19:167-175.
  77. Bowd C, Chan K, Zangwill LM, Goldbaum MH, Lee TW, Sejnowski TJ, Weinreb RN: Comparing neural networks and linear discriminant functions for glaucoma

- detection using confocal scanning laser ophthalmoscopy of the optic disc. *Invest Ophthalmol Vis Sci* 2002, 43:3444-3454.
78. Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN, Sejnowski TJ: Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng* 2002, 49:963-974.
  79. Uchida H, Brigatti L, Caprioli J: Detection of structural damage from glaucoma with confocal laser image analysis. *Invest Ophthalmol Vis Sci* 1996, 37:2393-401.
  80. Zhang GP: Neural Networks for Classification: A survey. *IEEE Trans Syst Man Cybern* 2000, C30:451-462.
  81. Da Pozzo S, Iacono P, Marchesan R, Fantin A, Ravalico G: Scanning laser polarimetry with variable corneal compensation and detection of glaucomatous optic neuropathy. *Graefes Arch Clin Exp Ophthalmol* 2005, 243:774-779.
  82. Coops A, Henson DB, Kwartz AJ, Artes PH: Automated analysis of heidelberg retina tomograph optic disc images by glaucoma probability score. *Invest Ophthalmol Vis Sci* 2006, 47:5348-5355.
  83. Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B: An introduction to kernel-based learning algorithms. *IEEE Trans Neural Netw*, 12:181-201.
  84. McCulloch WS, Pitts W: A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 1943, 5:115-133.
  85. Rosenblatt F: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958, 65:386-408.
  86. Rumelhart DE, Hinton GE, Williams RJ: Learning representations by back-propagating errors. *Nature* 1986, 323:533-536.
  87. Vapnik VN: An overview of statistical learning theory. *IEEE Trans Neural Netw* 1999, 10:988-999.
  88. Tipping ME: Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res* 2001, 1:211-244.
  89. Bengtsson B, Heijl A: Inter-subject variability and normal limits of the SITA standard, SITA fast, and the Humphrey full threshold computerized perimetry strategies, SITA STATPAC. *Acta Ophthalmol Scand* 1999, 77:125-129.
  90. Heijl A, Lindgren G, Olsson J: The effect of perimetric experience in normal subjects. *Arch Ophthalmol* 1989, 107:81-86.
  91. Wild JM, Dengler-Harles M, Searle AE, et al: The influence of the learning effect on automated perimetry in patients with suspected glaucoma. *Acta Ophthalmol Scand* 1989, 67:537-545.
  92. Heijl A, Bengtsson B: The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol* 1996, 114:19-22.
  93. Hougaard JL, Ostensfeld C, Heijl A, Bengtsson B: Modelling the normal retinal nerve fibre layer thickness as measured by Stratus optical coherence tomography. *Graefes Arch Clin Exp Ophthalmol* 2006, 244:1607-1614.
  94. Heijl A, Patella VM: Essential Perimetry. *The Field Analyzer Primer*. Third edition. Carl Zeiss Meditec Inc, Aug 2002, pp 17-21.
  95. Bengtsson B, Olsson J, Heijl A, Rootzén H: A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmol Scand* 1997, 75:368-375.

96. Bengtsson B, Heijl A, Olsson J: Evaluation of a new threshold visual field strategy, SITA, in normal subjects. *Acta Ophthalmol Scand* 1998, 76:165-169.
97. Bengtsson B, Heijl A: Evaluation of a new perimetric threshold strategy, SITA, in patients with manifest and suspect glaucoma. *Acta Ophthalmol Scand* 1998, 76:268-272.
98. Asman P, Heijl A: Glaucoma Hemifield Test - Automated Visual-Field Evaluation. *Arch Ophthalmol* 1992, 110:812-819.
99. Heijl A, Bengtsson B: Early visual field defects in glaucoma: a study of eyes developing field loss. In: Bucci MG, ed. *Glaucoma: Decision Making in Therapy*. Springer Verlag, Milan Italy 1996, pp 75-78.
100. Bengtsson B, Lindgren A, Heijl A, Lindgren G, Asman P, Patella M: Perimetric probability maps to separate change caused by glaucoma from that caused by cataract. *Acta Ophthalmol Scand* 1997, 75:184-188.
101. Møller MF: A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks* 1993, 6:525-533.
102. Hansen LK, Salamon P: Neural Network Ensembles. *Ieee T Pattern Anal* 1990, 12:993-1001.
103. Breiman L: Bagging predictors. *Mach Learn* 1996, 24:123-140.
104. Fan RE, Chen PH, Lin CJ: Working set selection using second order information for training support vector machines. *J Mach Learn Res* 2005, 6:1889-1918.
105. Chang CC, Lin CJ: LIBSVM: a library for support vector machines, 2001. Software available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
106. Imbault F, Lebart K: A stochastic optimization approach for parameter tuning of Support Vector Machines. In *17th International Conference on Pattern Recognition (ICPR'04) 2004*, vol. 4:597-600.
107. Stone M: Cross-validation choice and assessment of statistical predictions. *J R Stat Soc.* 1974, B36:111-147.
108. Jolliffe T: *Principal Component Analysis*. Springer Series in Statistics Springer-Verlag New York Inc 2002, page 9.
109. Zhang Z, Zha H: Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM J Sci Comput* 2004, 26:313-338.
110. L. Teng, H. Li, X. Fu, W. Chen, and I.-F. Shen: Dimension reduction of microarray data based on local tangent space alignment. In *Proceedings of the 4th IEEE International Conference on Cognitive Informatics 2005*, pp 154-159.
111. Garway-Heath DF, Poinoosawmy D, Fitzke FW, Hitchings RA: Mapping the visual field to the optic disc in normal tension glaucoma eyes. *Ophthalmology* 2000, 107:1809-1815.
112. Zweig MH, Campbell G: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993, 39:561-577.
113. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982, 143:29-36.
114. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988, 44:837-845.
115. Altman DG: Comparing groups—categorical data. In: Altman DG, ed. *Practical Statistics for Medical Research*. Chapman and Hall, London 1991, pp 229-276.

116. Friis RH, Sellers TA: *Epidemiology for Public Health Practice*. 3rd ed. Jones and Bartlett Publishers, Sudbury MA 2004, page 384.
117. Ajtony C, Balla Z, Somoskeoy S, Kovacs B: Relationship between Visual Field Sensitivity and Retinal Nerve Fiber Layer Thickness as Measured by Optical Coherence Tomography. *Invest Ophthalmol Vis Sci* 2007, 48:258-263.
118. Sato S, Hirooka K, Baba T, Yano I, Shiraga F: Correlation between retinal nerve fibre layer thickness and retinal sensitivity. *Acta Ophthalmol Scand* 2008, 86:609-613.
119. Harwerth RS, Wheat JL, Fredette MJ, Anderson DR: Linking structure and function in glaucoma. *Prog Retin Eye Res* 2010, 29:249-271.
120. Gardiner SK, Johnson CA, Cioffi GA: Evaluation of the Structure-Function Relationship in Glaucoma. *Invest Ophthalmol Vis Sci* 2005, 46:3712-3717.
121. Strouthidis N, Vinciotti V, Tucker AJ, Gardiner SK, Crabb DP, Garway-Heath DF: Structure and Function in Glaucoma: The relationship between a Functional Visual Field Map and an Anatomic Retinal Map. *Invest Ophthalmol Vis Sci* 2006, 47:5356-62.
122. Leung CK, Chong KK, Chan WM, Yiu CK, Tso MY, Woo J, Tsang MK, Tse KK, Yung WH: Comparative study of retinal nerve fiber layer measurement by StratusOCT and GDx VCC, II: structure/function regression analysis in glaucoma. *Invest Ophthalmol Vis Sci* 2005, 46:3702-3711.
123. El Beltagi TA, Bowd C, Boden C, Amini P, Sample PA, Zangwill LM, Weinreb RN: Retinal nerve fiber layer thickness measured with optical coherence tomography is related to visual function in glaucomatous eyes. *Ophthalmology* 2003, 110:2185-2191.
124. Hood DC, Kardon RH: A framework for comparing structural and functional measures of glaucomatous damage. *Prog Retin Eye Res* 2007, 26:688-710.
125. Ferreras A, Pablo LE, Garway-Heath DF, Fogagnolo P, García-Feijoo J: Mapping standard automated perimetry to the peripapillary retinal nerve fiber layer in glaucoma. *Invest Ophthalmol Vis Sci* 2008, 49:3018-3018.
126. Turpin A, Sampson GP, McKendrick AM: Combining Ganglion Cell Topology and Data of Patients with Glaucoma to Determine a Structure-Function Map. *Invest Ophthalmol Vis Sci* 2009, 50:3249-3256.
127. Zhu H, Crabb DP, Schlottmann PG, Lemij HG, Reus NJ, Healey PR, Mitchell P, Ho T, Garway-Heath DF: Predicting Visual Function from the Measurements of Retinal Nerve Fiber Layer Structure. *Invest Ophthalmol Vis Sci* 2010, 51:5657-5666.
128. Boland MV, Quigley HA: Evaluation of a combined index of optic nerve structure and function for glaucoma diagnosis. *BMC Ophthalmol* 2011, 11:6.



## **13.** Appendix

Papers I – IV



# Paper I





# Effects of Input Data on the Performance of a Neural Network in Distinguishing Normal and Glaucomatous Visual Fields

Boel Bengtsson, Dimitrios Bizios, and Anders Heijl

**PURPOSE.** To compare the performance of neural networks for perimetric glaucoma diagnosis when using different types of data inputs: numerical threshold sensitivities, Statpac Total Deviation and Pattern Deviation, and probability scores based on Total and Pattern Deviation probability maps (Carl Zeiss Meditec, Inc., Dublin, CA).

**METHODS.** The results of SITA Standard visual field tests in 213 healthy subjects, 127 patients with glaucoma, 68 patients with concomitant glaucoma and cataract, and 41 patients with cataract only were included. The five different types of input data were entered into five identically designed artificial neural networks. Network thresholds were adjusted for each network. Receiver operating characteristic (ROC) curves were constructed to display the combinations of sensitivity and specificity.

**RESULTS.** Input data in the form of Pattern Deviation probability scores gave the best results, with an area of 0.988 under the ROC curve, and were significantly better ( $P < 0.001$ ) than threshold sensitivities and numerical Total Deviations and Total Deviation probability scores. The second best result was obtained with numerical Pattern Deviations with an area of 0.980.

**CONCLUSIONS.** The choice of type of data input had important effects on the performance of the neural networks in glaucoma diagnosis. Refined input data, based on Pattern Deviations, resulted in higher sensitivity and specificity than did raw threshold values. Neural networks may have high potential in the production of useful clinical tools for the classification of visual field tests. (*Invest Ophthalmol Vis Sci.* 2005;46:3730–3736) DOI:10.1167/iovs.05-0175

Perimetry is one of the most important examinations for diagnosis and monitoring of glaucoma. Static computerized threshold perimetry in which white stimuli are shown on an evenly illuminated white background has been the most common type of perimetry in clinical glaucoma management for a long time. The way in which perimetric findings are analyzed and presented is important in the interpretation of test results. Reading fields by looking only at maps of numerical threshold sensitivities or gray-scale representation of such values, is difficult even for experts. Programs such as the Humphrey Stat-

pac (Carl Zeiss Meditec, Inc., Dublin, CA)<sup>1</sup> for computer-assisted interpretation were developed in the mid- to late 1980s. The probability maps included in the Statpac program are often able to highlight early glaucomatous field defects before they become visible in gray-scale representations of raw threshold values<sup>2</sup> and can also reduce effects caused by cataract.<sup>3</sup> This probability map concept has enjoyed wide acceptance and has subsequently been applied in most new perimetric devices and perimetric modalities, such as frequency-doubling perimetry<sup>4</sup> and short-wavelength automated perimetry.<sup>5,6</sup>

The Glaucoma Hemifield Test (GHT)<sup>7</sup> included in Statpac, is a rather simple expert system based on up-and-down hemifield differences between probability scores calculated from Pattern Deviation probability maps. The GHT was one of the first computerized systems that was able to classify field test results reliably as normal or abnormal and improved the ability of ordinary clinicians to assess visual field test results.<sup>8</sup>

In the beginning of the 1990s artificial neural networks (ANNs), one of many algorithms in the machine learning classifier concept, were tested as a tool for the interpretation of perimetric results (Goldbaum MH, et al. *IOVS* 1990;31:ARVO Abstract 2471; Keating D, et al. *IOVS* 1992;33:ARVO Abstract 1394).<sup>9</sup> ANNs were reported to be able to differentiate between glaucoma and normal visual field status at least as well as trained readers.<sup>10</sup> In other papers, it was also reported that machine learning classifiers discriminate better between normal and glaucomatous fields than do global visual field indices.<sup>11,12</sup> Global visual field indices are far from ideal as diagnostic tools, however, because they condense all threshold data into one number, resulting in loss of valuable spatial information, and visual field indices are not particularly sensitive to early localized glaucomatous visual field loss.<sup>13-15</sup>

The performance of ANNs has also been compared with that of other types of field interpretation criteria based on localized loss.<sup>11</sup> Disc topography data have also been added to visual field data to improve the diagnostic ability of ANNs.<sup>16</sup>

We hypothesized that it may be possible to enhance the diagnostic performance of ANNs further by using input data from which the effects of age and media opacities have been eliminated or reduced and in which measured sensitivities have already been compared to the range of age-corrected normal sensitivities and subsequently translated into probabilities. The Statpac program provides two important analyses: (1) Numerical Total Deviations represent the deviation at each tested point of the measured threshold from age-corrected normal values. (2) Numerical Pattern Deviations represent a modification of the Total Deviation results in which a correction has been applied to account for any general elevation or depression of the field caused by media opacities or changes in pupil size. Total and Pattern Deviation probability maps are graphic presentations of the significances of the numerical deviations, relative to the known ranges of normal values at each test point location.

The purpose of this study was to test our hypothesis by comparing sensitivities and specificities achieved by ANNs for glaucoma diagnosis by using different types of perimetric in-

From the Department of Ophthalmology, Malmö University Hospital, Lund University, Malmö, Sweden.

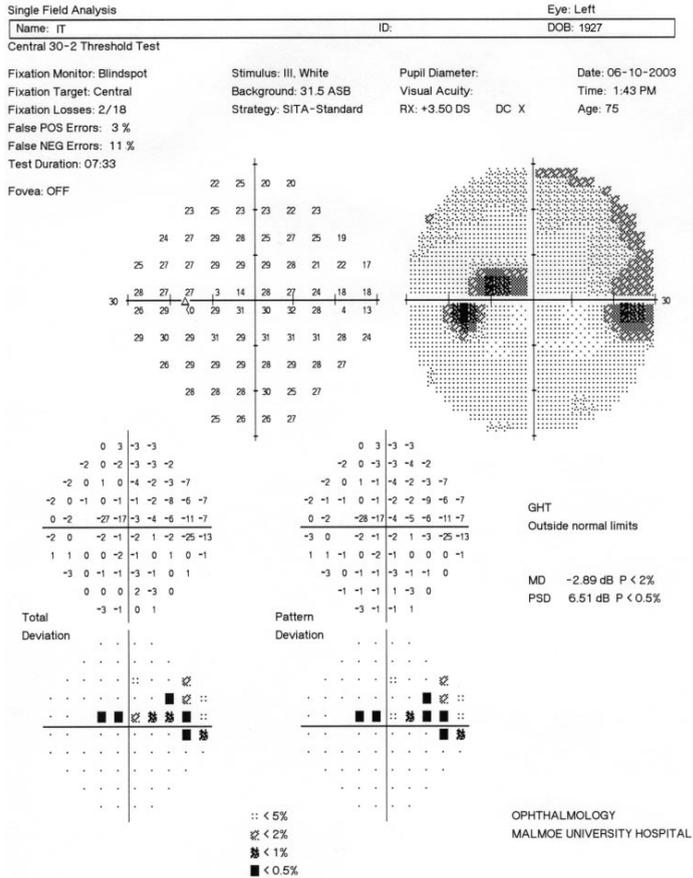
Supported by Grant K2002-74X-10426-10A from the Swedish Research Council, by the Järnhardt Foundation, and by funds administered by Malmö University Hospital.

Submitted for publication February 10, 2005; revised April 14 and May 12, 2005; accepted July 1, 2005.

Disclosure: **B. Bengtsson**, None; **D. Bizios**, None; **A. Heijl**, None

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Boel Bengtsson, Department of Ophthalmology, Malmö University Hospital, Lund University, SE-205 02 Malmö, Sweden; boel.bengtsson@ofal.mas.lu.se.



**FIGURE 1.** A 30-2 SITA Standard visual field test with thresholds and Statpac Total and Pattern Deviations and corresponding probability maps. This field was correctly classified as glaucomatous by all five types of networks.

puts: numerical threshold values in decibels, and Statpac numerical Total and Pattern Deviations and probabilities.

**METHODS**

**Visual Fields**

All visual fields included were obtained with the 30-2 SITA Standard program of the Humphrey Field Analyzer II (model 750; Carl Zeiss Meditec, Inc.). All test point locations in the 30-2 test point pattern were included, except two located in the area of the physiological blind spot. The input data from each field test were:

1. Raw threshold sensitivity values in decibels.
2. Statpac numerical Total Deviations (i.e., deviations in decibels from age-corrected threshold).
3. Numerical Pattern Deviations, which are those same deviations adjusted for the general height of the field and probability scores computed from the Total Deviation and Pattern Deviation probability maps (Fig. 1). The probability scoring scale was

identical with that used in the calculation of the GHT<sup>7</sup>—that is, all test points were assigned a score according to significance level of the deviation from the normal value.

Unreliable field test results, defined as a frequency of false-positive answers exceeding 15% or a fixation loss larger than 20% were excluded from the analysis. False-negative rates were not included in the exclusion criteria, because they have been found to correlate highly with the degree of legitimate and reliable glaucomatous field loss measurements.<sup>17,18</sup>

**Subjects**

Because patients with glaucoma often have concomitant cataract, it is desirable that methods designed to recognize glaucomatous visual field loss not be affected by ocular media changes. Therefore, it was necessary to train the neural networks with fields from healthy subjects, patients with cataract, patients with glaucoma, and patients with both glaucoma and cataract. Patients with glaucoma had primary open-angle glaucoma (POAG), including normal tension, exfoliation, and pigment

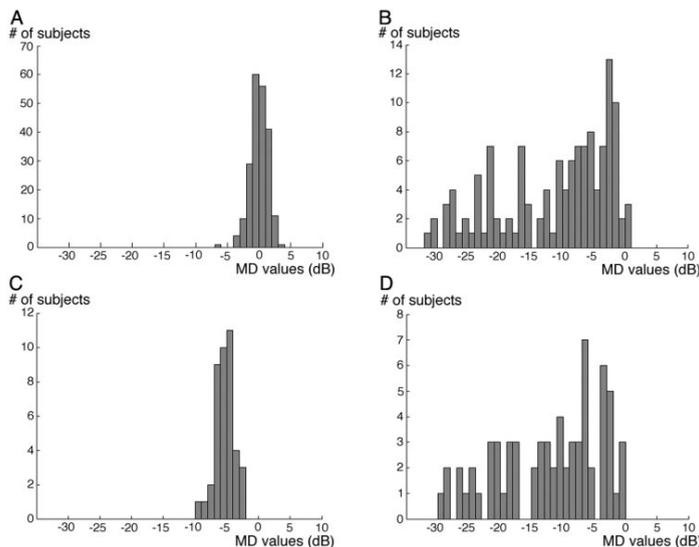


FIGURE 2. Distribution of MDs in (A) healthy subjects (average MD,  $-0.02$  dB), (B) patients with glaucoma (average MD,  $-9.84$  dB), (C) patients with cataract (average MD,  $-5.27$  dB), and (D) patients with concomitant glaucoma and cataract (average MD,  $-12.13$  dB).

glaucoma. Other types of glaucoma, such as angle-closure, secondary, and congenital forms were not included. Glaucomatous eyes were defined as those having typical glaucomatous changes in the optic disc: notches, thin or absent neural rims or marked vertical optic cup asymmetry, combined with glaucomatous visual field defects. Glaucomatous field defects were those that were compatible with glaucoma and not explained by other disease. The visual field classification was subjective, including all information available on the single-field printouts. However, we also included seemingly normal fields from eyes with pathologic disc topography, if field defects were found in later visual field tests. Patients with macular or retinal changes and neurologic or endocrinological disorders or other conditions likely to cause field defects were excluded, whereas patients with diabetes mellitus without retinopathy were included. No first field test results of any subjects were considered, to avoid patterns caused by lack of perimeter experience.<sup>19–21</sup> The study was conducted according to the tenets of the Declaration of Helsinki and was approved by the Ethics Committee of Lund University.

**Healthy Subjects.** Two hundred thirteen tests results of 213 subjects were randomly selected from an existing large normal database originally collected to establish normal thresholds and normal limits for the SITA thresholding strategies.<sup>22</sup> The mean age of these subjects was 52 years, ranging from 19 to 84. Most fields in the normative database appeared quite normal, although normality was not a criterion for inclusion; Average Mean Deviation (MD) was  $-0.02$  dB, ranging from  $-6.11$  to  $+3.07$  dB (Fig. 2A).

**Patients with Media Opacities.** These patients had normal disks and normal visual fields and a notation of cataract in their record. We identified 55 such patients. After removing those with unreliable field test results, mostly due to poor fixation, 41 eyes of 41 patients remained in this group. The mean age of these 41 patients was 77 years, ranging from 54 to 96. The MD ranged from  $-9.82$  to  $-2.46$  dB (Fig. 2C). Forty of these patients had cataract and one had postsurgical opacification of the posterior capsule. These fields were regarded as normal and were included among the 213 normal tests.

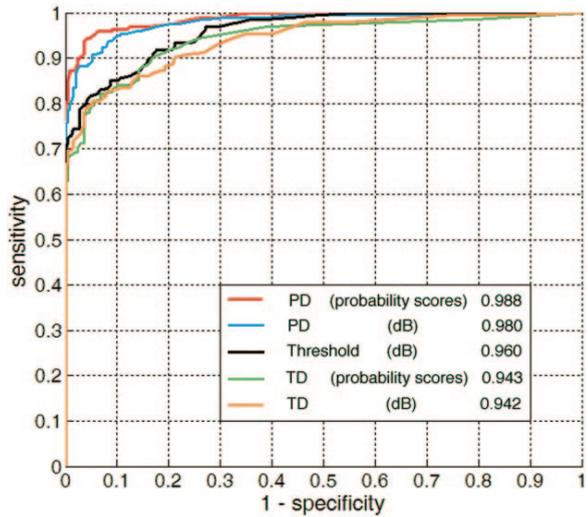
**Patients with Glaucoma.** The field tests of patients with glaucoma were randomly selected from the directory fields included in the database in one of our Humphrey Field Analyzers. This database

consisted of 11,134 tests of 3,629 patients, almost all assessed by the 30-2 SITA Standard program. The directory was sorted in alphabetic order according to the patient's surname. Starting with the letter A, one field test was randomly selected from every fifth patient; no first field results were selected, to avoid patterns of learning. The selected patients were then matched to our glaucoma register. Only patients with a diagnosis of glaucoma or suspected glaucoma were eligible, and patient records were retrieved. In this way, 643 SITA Standard 30-2 test results were selected to be evaluated for inclusion. At this point the only information available was that the patient had undergone 30-2 SITA Standard visual field testing at least twice, and that the patient had a diagnosis of suspected glaucoma or glaucoma. After retrieving patient records disc photographs obtained before the selected field test were inspected. Fields of all eyes with glaucomatous disc appearance were deemed usable. A comprehensive description of disc topography was required in patient records lacking disc photographs. A description of lens status was also required. The absence of such a description or a notation of a clear lens or pseudophakic eyes was regarded as glaucoma without cataract, whereas data indicating the presence of any type or stage of cataract classified the eyes as having glaucoma plus cataract. After exclusion of eyes according to these criteria, 127 tests of 127 eyes with glaucoma and 68 tests of 68 eyes with concomitant glaucoma and cataract remained.

The mean age of the 127 patients with glaucoma was 75 years, ranging from 40 to 96. MDs ranged from  $-31.18$  to  $+0.74$  dB (Fig. 2B). The group with both glaucoma and cataract averaged 77 years of age, ranging from 51 to 97 and had MDs ranging from  $-29.99$  to  $-0.12$  dB (Fig. 2D). In some eyes, the selected field test results appeared normal, but then the disc appeared suspicious or pathologic, and later field tests, not included in the analysis, showed glaucomatous field loss.

### Neural Network Design

Our networks were fully connected feed-forward multilayer perceptrons built using commercial software (Neural Network Toolbox, ver.4.0 of MatLab; The MathWorks Inc., Natick, MA). This network architecture, consisting of an input layer, two hidden layers, and an output layer, was the same for the different sets of input data. There were 74 units in the input layer, each unit corresponding to one test



**FIGURE 3.** ROC curves for each of the five different types of input data. Pattern Deviation (PD) probability scores had the largest area under the curve (0.988), whereas numerical Total Deviations (TD) had the smallest area (0.942).

point in the 30-2 test point pattern. The number of processing elements in the two hidden layers was 25 and 5. The output layer, one neuron with a logistic transfer function, provided the network's output: glaucoma or normal.

**Network Training.** The networks were trained in batch mode by using an optimization of the back propagation algorithm developed by Moller.<sup>23</sup> This algorithm has been shown to have a fast convergence rate (i.e., relatively few iterations are needed to achieve a small classification error calculated from the network output). An early stopping technique was applied to terminate the training procedure to prevent overfitting of the data. A glaucomatous field classified with 100% certainty was assigned an output of 1 and a 100% normal field an output of 0. Fields falling between were assigned values between 0 and 1. Outputs close to the endpoints 0 or 1 indicated high confidence in the classification, whereas those close to 0.5 indicated uncertainty of the output. Because the task of the network was to identify glaucomatous field loss, patients with cataract only were included in the normal group and patients with concomitant cataract and glaucoma in the glaucoma group. During network training, classification errors were calculated and used to adjust weights in the neural network. The number of necessary iterations, arbitrarily set to a maximum of 300, was also determined by the size of the classification error. Eighty percent of all fields were used in the training procedure.

**Validation.** A validation procedure was applied, using half the fields not used in initial training, to prevent overfitting of data. Over-

fitting of data hampers the network's generalization ability and effective classification of previously unseen data.

**Evaluation.** The performance of the network was evaluated with a 10-fold cross-validation procedure, in which all fields were randomly divided into 10 subgroups each containing 10% of the full data set.<sup>12,24</sup> The number of subgroups used in training, early stopping, and test procedures was 8, 1, and 1, respectively. With this procedure, each subgroup was used for training, validation, and evaluation, while ensuring that the network was trained and evaluated, by using different sets of visual fields to avoid confounding.

**Analyses**

Network receiver operating characteristic (ROC) curves<sup>25</sup> were produced by adjusting the network threshold. The network threshold, ranging from 0 to 1, was used to define patient classification or diagnosis. For each network threshold, fields with outputs larger than the threshold were classified as glaucomatous, and outputs lower than the network threshold were classified as normal. The areas under the ROC curves, one for each type of input data, were compared by a nonparametric method described by Delong et al.,<sup>26</sup> and the Bonferroni correction was applied to adjust for effects of multiple comparisons on the type I error—that is, to reject falsely the null hypothesis stating no difference between ROC curves.

**TABLE 1.** Performance of Neural Network in Classifying Standard Automated Perimetric Visual Fields, using Different Input Data

	Pattern Deviation				Threshold Sensitivity		Total Deviation			
	Prob. Scores		dB		dB		Prob. Scores		dB	
Network threshold	0.50	0.30 (best)	0.50	0.37 (best)	0.50	0.43 (best)	0.50	0.42 (best)	0.50	0.47 (best)
Sensitivity (%)	89.7	93.9	86.7	90.8	81.5	85.1	79.5	82.1	79.5	80.5
Specificity (%)	97.6	96.5	98.0	94.9	95.3	91.3	94.9	93.3	94.9	94.9
Area under ROC curve	0.988*		0.980†		0.960		0.943		0.942	

\* Significantly better than threshold sensitivity (dB), Total Deviation (dB) and Total Deviation prob. scores ( $P < 0.001$ ).

† Significantly better than Total Deviation (dB) and Total Deviation prob. scores ( $P < 0.001$ ).

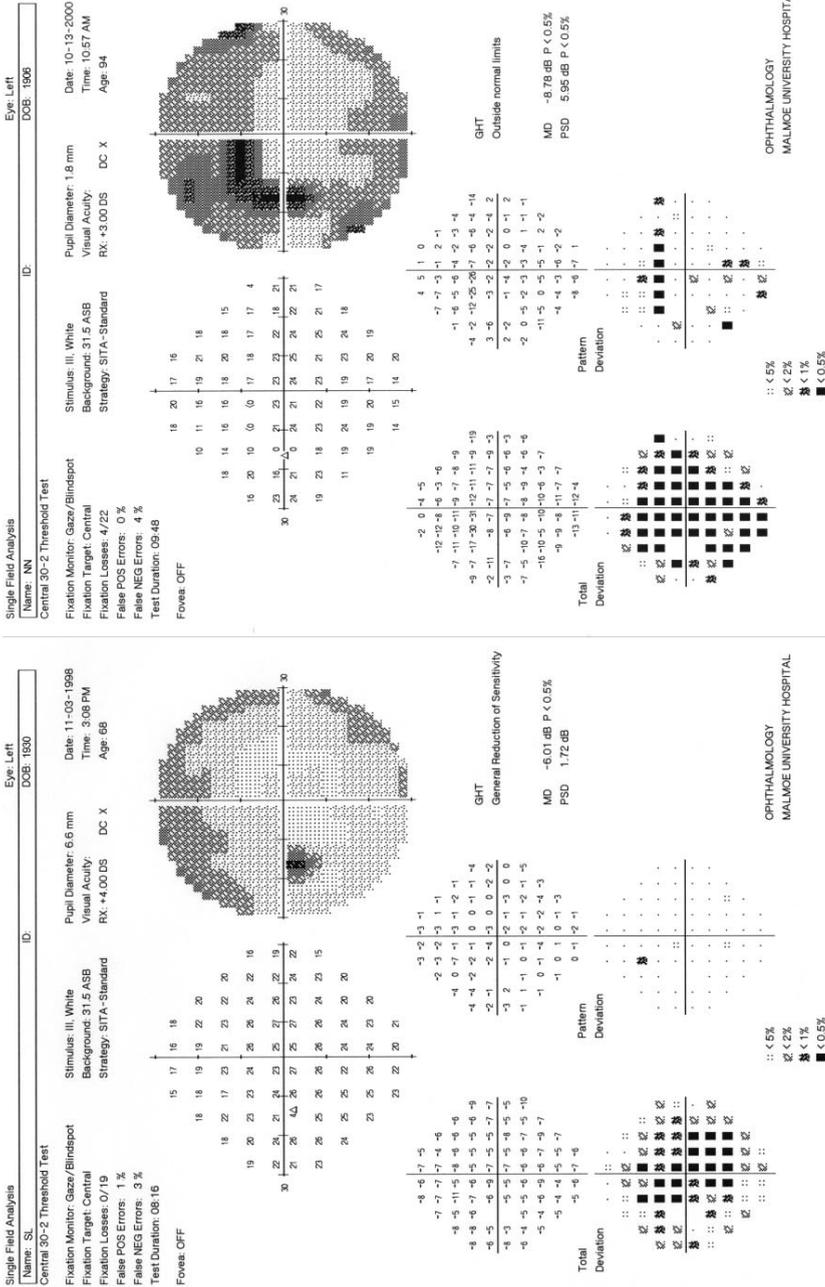


Figure 4. Two visual fields correctly classified by the Pattern Deviation numerical and probability score networks, but not with the three networks trained with thresholds and numerical

Figure 4. Total Deviation and scores. Left: a field of an eye with healthy disc appearance and cataract; right: a field of an eye with glaucomatous disc appearance and cataract.

## RESULTS

ROC curves showing combinations of sensitivities and specificities for the different types of data input are shown in Figure 3. All types of input data formed areas under the ROC curve larger than 0.9, but the different data inputs yielded quite different results. The best results, defined as the largest area under the ROC curve, was achieved with the Pattern Deviation probability scores, followed by the numerical Pattern Deviations. Pattern Deviation probability scores had a significantly larger area under the ROC curve than did threshold sensitivities and numerical Total Deviations and Total Deviation probability scores ( $P < 0.001$ ). Numerical Pattern Deviations were significantly better ( $P < 0.001$ ) than both numerical Total Deviations and Total Deviation probability scores, but not better than raw threshold sensitivity values. Threshold sensitivities performed slightly, but nonsignificantly, better than both numerical Total Deviations and Total Deviation probability scores (Fig. 3).

In the set of normal fields MDs differed between healthy subjects without (Fig. 2A) and with (Fig. 2C) cataract, whereas the proportion of significantly depressed points in Pattern Deviation probability maps was more similar. For example, the relative number of points depressed at the  $P < 0.5\%$  level (black box in the probability map) was 0.3% in eyes without cataract and 0.6% in eyes with cataract. The corresponding proportion of such depressed points in Pattern Deviation probability maps in glaucomatous eyes was 26.8% in eyes without cataract and 26.4% in eyes with cataract.

A best threshold for the network was determined by the best combination of sensitivity and specificity, simply defined as the product of the two. Best network thresholds differed for the various types of input data (Table 1).

## DISCUSSION

ANNs have been suggested as tools for interpretation of automated visual field test results in patients with glaucoma.<sup>10,11</sup> Other types of machine learning classifiers, such as support vector machines or committee machines, have also been reported to interpret visual fields adequately.<sup>12</sup> In all studies that we have been able to find, however, the inputs have been trained and tested with unprocessed threshold sensitivities. There is no reason to believe that different types of machine learning classifiers would yield different results when different types of input data are compared. We found that using the more refined input data available from a program for computer-assisted interpretation (i.e., Statpac data) could significantly enhance sensitivity and specificity. Pattern Deviation probability scores based on the Pattern Deviation probability maps produced the largest area under the ROC curve, indicating high performance in discrimination between normal and glaucomatous fields.

The improved results obtained when field data were entered as Pattern Deviations is probably explained by the reduction of the influence of cataract on Pattern Deviations. Both Pattern Deviation numerical displays and probability maps were designed to reduce the effect of media opacities. Pattern Deviation misclassified only 2 normal eyes with cataract, whereas 13 were misclassified when Total Deviation was used. The network was designed to identify the absence or presence of glaucomatous visual field loss. Thus, we included subjects with cataract in the normal group and patients with concomitant cataract and glaucoma in the glaucoma group. We used this approach because cataract frequently occurs in the age groups where glaucoma is most prevalent.

The normal fields obtained in healthy subjects without cataract were randomly selected from a larger multicenter database used for calculation of Statpac normal values and

normal limits for SITA fields. We do not believe that this has biased our results. A large database including data from multiple centers is probably more representative of a normal population than a smaller sample collected at one center only. We did not use the full database; 66% of the records were randomly selected for the purpose of this study. We also included normal fields of patients with media opacities in our set of normal fields. The results, as presented in ROC curves, depended considerably more on the network output than on the Statpac normal limits. Further, our purpose was to compare different input derived from the same normal and pathologic fields and the conclusion pertaining to that comparison would not be expected to cause any bias, as the effects of the selection of the normal data would be equal in all five parameters.

The five different ANNs correctly classified most fields; but, as expected, normal eyes with substantial cataract were more often classified correctly by the two Pattern Deviation-based ANNs compared with the Total Deviation and unprocessed threshold ANNs (Fig. 4). In fields with severe damage, Pattern Deviation-based ANNs did not perform as well as ANNs trained with Total Deviation and threshold sensitivities. This was also anticipated, as the Pattern Deviation concept cannot presently be successfully used in end-stage fields.<sup>27,28</sup>

The selection of subjects is crucial when evaluating diagnostic methods. Testing the method in only patients with obvious moderate to severe field defects would give results suggesting better discrimination than would be found in patients with early defects. We randomly selected our glaucoma fields from the directory of tests on the hard disk in one of our perimeters. This resulted in a representative selection of patients with a wide range of visual field defects, including glaucomatous eyes without apparent field loss. With this method, 39% had MDs better than  $-5$  dB and thus could be considered to have mild loss. If only fields with clear-cut reproducible defects were selected, one would expect higher sensitivities for all types of input data. Our selection of fields including a random sample of glaucomatous eyes has advantages, but the selection, in principle, should not be critical when comparing performance of neural networks all using different input data from the same normal and glaucomatous visual fields.

Our results suggest that the ability of artificial neural networks to classify visual fields can be further improved if refined input data based on Pattern Deviations is used. Such input data resulted in higher sensitivity and specificity than did raw threshold sensitivity values, probably because of the former's ability to separate field loss caused by glaucoma from that caused by cataract. Further studies including independent visual field data not used for training of network data are needed to evaluate a more general applicability of ANNs for classification of visual field test results. Neural networks and other machine classifiers seem to have a great potential to become a useful clinical tool in the diagnosis of glaucomatous visual field loss, and it may be of value in the study of the performance of a range of types of data inputs with different machine classifiers.

## Acknowledgments

The authors thank Ola Engwall, MSc (Lund, Sweden), for performing the necessary programming in the MatLab environment.

## References

- Heijl A, Lindgren G, Olsson J. A package for statistical analysis of computerized fields. *Doc Ophthalmol Proc Ser*. 1987;49:153-168.
- Heijl A, Bengtsson B. Early visual field defects in glaucoma: a study of eyes developing field loss. In: Bucci MG, ed. *Glaucoma: Deci-*

- sion Making in Therapy. Milan, Italy: Springer Verlag; 1996:75-78.
3. Bengtsson B, Lindgren A, Heijl A, Lindgren G, Åsman P, Patella VM. Perimetric probability maps to separate change caused by glaucoma from that caused by cataract. *Acta Ophthalmol Scand.* 1997; 75:184-188.
  4. Johnson CA, Wall M, Fingeret M, Lalle P. *A Primer for Frequency Doubling Technology.* Skaneateles, NY: Welch-Allyn Inc.; 1998:9-11.
  5. Johnson CA, Adams AJ, Casson EJ, Brandt JD. Blue-on-yellow perimetry can predict the development of glaucomatous visual field loss. *Arch Ophthalmol.* 1993;111:645-650.
  6. Heijl A, Patella VM. *Essential Perimetry. The Field Analyzer Primer.* 3rd ed. Dublin, CA: Carl Zeiss Meditec, Inc.; 2002:37.
  7. Åsman P, Heijl A. Glaucoma Hemifield Test: automated visual field evaluation. *Arch Ophthalmol.* 1992;110:812-819.
  8. Katz J, Sommer A, Gaasterland DE, Anderson DR. Comparison of analytic algorithms for detecting glaucomatous visual field loss. *Arch Ophthalmol.* 1991;109:1684-1689.
  9. Mutlukan E, Keating D. Visual field interpretation with a personal computer based neural network. *Eye.* 1994;8:321-323.
  10. Goldbaum MH, Sample PA, White H, et al. Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci.* 1994;35:3362-3373.
  11. Lietman T, Eng J, Katz J, Quigley HA. Neural networks for visual field analysis: how do they compare with other algorithms? *J Glaucoma.* 1999;8:77-80.
  12. Goldbaum MH, Sample PA, Chan K, et al. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Invest Ophthalmol Vis Sci.* 2002;43:162-169.
  13. Harrington DO, Drake MW. Computerized perimeters. *The Visual Fields: Text and Atlas of Clinical Perimetry.* 6th ed. St. Louis: CV Mosby; 1990:54-55.
  14. Chauhan BC, Drance SM, Lai C. A cluster analysis for threshold perimetry. *Graefes Arch Clin Exp Ophthalmol.* 1989;27:216-220.
  15. Asman P, Heijl A, Olsson J, Rootzen H. Spatial analyses of glaucomatous visual fields; a comparison with traditional visual field indices. *Acta Ophthalmol Scand.* 1992;70:679-686.
  16. Brigatti L, Hoffman D, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol.* 1996;121:511-521.
  17. Katz J, Sommer A. Reliability indexes of automated perimetric tests. *Arch Ophthalmol.* 1988;106:1252-1254.
  18. Bengtsson B, Heijl A. False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? *Invest Ophthalmol Vis Sci.* 2000;41:2201-2204.
  19. Heijl A, Lindgren G, Olsson J. The effect of perimetric experience in normal subjects. *Arch Ophthalmol.* 1989;107:81-86.
  20. Wild JM, Dengler-Harles M, Searle AE, O'Neill EC, Crews SJ. The influence of the learning effect on automated perimetry in patients with suspected glaucoma. *Acta Ophthalmol Scand.* 1989;67:537-545.
  21. Heijl A, Bengtsson B. The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol.* 1996;114:19-22.
  22. Bengtsson B, Heijl A. Inter-subject variability and normal limits of the SITA Standard, SITA Fast, and the Humphrey Full Threshold computerized perimetry strategies, SITA STATPAC. *Acta Ophthalmol Scand.* 1999;77:125-129.
  23. Møller MF. A scaled conjugate gradient algorithm for fast supervised training. *Neural Networks.* 1993;6:525-533.
  24. Stone M. Cross-validation choice and assessment of statistical predictions. *J R Stat Soc.* 1974;B36:111-147.
  25. Sweets J. Measuring the accuracy of diagnostic systems. *Science.* 1988;240:1285-1293.
  26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837-845.
  27. Heijl A, Patella VM. *Essential perimetry. The Field Analyzer Primer.* 3rd ed. Dublin, CA: Carl Zeiss Meditec, Inc.; 2002:50.
  28. Blumenthal E, Sapir-Pichhadze R. Misleading statistical calculations in far-advanced glaucomatous visual field loss. *Ophthalmology.* 2003;110:196-200.



# Paper II





# Trained Artificial Neural Network for Glaucoma Diagnosis Using Visual Field Data

## A Comparison With Conventional Algorithms

Dimitrios Bizios, MD, Anders Heijl, MD, PhD, and Boel Bengtsson, PhD

**Purpose:** To evaluate and confirm the performance of an artificial neural network (ANN) trained to recognize glaucomatous visual field defects, and compare its diagnostic accuracy with that of other algorithms proposed for the detection of visual field loss.

**Methods:** SITA Standard 30-2 visual fields, from 100 glaucoma patients and 116 healthy participants, formed the data set. Our ANN was a previously described fully trained network using scored pattern deviation probability maps as input data. Its diagnostic accuracy was compared to that of the Glaucoma Hemifield Test, the Pattern Standard Deviation index at the  $P < 5\%$  and  $< 1\%$ , and also to a technique based on the recognizing clusters of significantly depressed test points.

**Results:** The included tests had early to moderate visual field loss (median MD =  $-6.16$  dB). ANN achieved a sensitivity of 93% at a specificity level of 94% with an area under the receiver operating characteristic curve of 0.984. Glaucoma Hemifield Test attained a sensitivity of 92% at 91% specificity. Pattern Standard Deviation, with a cut off level at  $P < 5\%$  had a sensitivity of 89% with a specificity of 93%, whereas at  $P < 1\%$  the sensitivity and specificity was 72% and 97%, respectively. The cluster algorithm yielded a sensitivity of 95% and a specificity of 82%.

**Conclusions:** The high diagnostic performance of our ANN based on refined input visual field data was confirmed in this independent sample. Its diagnostic accuracy was slightly to considerably better than that of the compared algorithms. The results indicate the large potential for ANN as an important clinical glaucoma diagnostic tool.

**Key Words:** glaucoma diagnostics, perimetry, visual field, artificial neural network

(*J Glaucoma* 2007;16:20-28)

Perimetry is important for the detection of glaucoma and for monitoring disease progression. Interpretation of the quantitative measurements of differential light sensitivity provided by static automated perimetry is facilitated by prior knowledge of the normal sensitivity range at each tested location. The Statpac program in the Humphrey Field Analyzer<sup>1,2</sup> includes deviation plots simplifying the interpretation of visual field test results. Pattern deviation probability maps emphasize localized visual field loss after adjustment for age and for any general depression of light sensitivity, and are often capable of detecting early glaucomatous defects not yet visible in gray-scale representations of the measured threshold values,<sup>3</sup> while reducing diffuse sensitivity loss caused by media opacities.<sup>4</sup>

Over the past decade, several studies have investigated the ability of neural networks and other machine classifiers, trained using both supervised and unsupervised techniques, to detect the presence or progression of glaucomatous defects based on data from perimetric testing and other diagnostic tests.<sup>5-15</sup> Techniques based on artificial neural networks (ANNs) have been successful in diagnosing glaucoma, both with functional<sup>16</sup> and structural<sup>17</sup> tests. ANNs using data from standard automated perimetry have produced results comparable to those of other methods of discriminant analysis<sup>16,18,19</sup> and glaucoma experts.<sup>20</sup>

The Glaucoma Hemifield Test (GHT),<sup>21</sup> part of Statpac, is a plain-text interpretation tool based on probability scores derived from the pattern deviation probability maps, and functions by detecting threshold sensitivity differences across the horizontal meridian of the visual field.<sup>21,22</sup> GHT has high sensitivity and specificity,<sup>23</sup> and the simplicity of its output can be of help to physicians lacking detailed knowledge in the assessment of perimetric test results.

Other types of statistical analyses, such as the global index Pattern Standard Deviation (PSD) included in Statpac, and cluster analyses based on the pattern deviation probability maps,<sup>24</sup> have also been proposed as criteria for recognizing glaucomatous defects, both in experimental settings and clinical practice.<sup>23,25</sup> Studies

Received for publication March 6, 2006; accepted October 10, 2006.  
From the Department of Clinical Sciences, Ophthalmology, Malmö University Hospital, Lund University, SE-205 02 Malmö, Sweden. Supported by grants K2005-74X-1426-13A and K2005-74BI-15375-01A from the Swedish research council [vetenskapsrådet], by the foundation of Crown Princess Margareta for visually handicapped [Stiftelsen Kronprinsessan Margaretas arbetsnämnd för synskadade (KMA)], and by the Järnhardt foundation (Järnhardts stiftelse). Reprints: Dimitrios Bizios, MD, Department of Clinical Sciences, Ophthalmology, Malmö University Hospital, Lund University, SE-205 02 Malmö, Sweden (e-mail: dimitrios.bizios@med.lu.se). Copyright © 2007 by Lippincott Williams & Wilkins

have shown that these analytical approaches may have high degrees of specificity at acceptable levels of sensitivity.<sup>26,27</sup>

Both GHT and cluster recognition approaches are based on the concept of pattern deviation probability maps. We have previously demonstrated that the use of refined input data based upon pattern deviation, rather than on total deviation plots or on the raw numerical values of measured threshold sensitivity, can result in better performance of neural network classifiers.<sup>28</sup> The aim of the present study was to assess the diagnostic accuracy of a trained neural network with its input based on pattern deviation probability maps, using an independent data set of both healthy and glaucomatous eyes with early to moderate damage, and to compare the performance of this neural network to that of other interpretation methods proposed for the detection of glaucomatous field loss.

## METHODS

### Visual Fields

The study material consisted of SITA Standard 30-2 field tests from the Humphrey perimeter (Humphrey Field Analyzer II 750, Carl Zeiss Meditec, Dublin, CA). The visual field interpretation algorithms used in our analysis, included:

1. ANN, using probability scores derived from the pattern deviation probability maps as input. The ANN employed here has been described earlier and is a feed-forward multilayer perceptron (MLP).<sup>28</sup> It uses a scoring process, in which all test points were assigned a score according to the significance level of the deviation from the normal value, adapted from the calculation of the GHT results.<sup>21</sup>
2. PSD expressed in decibel (dB), with 2 cut-off values defining pathology.
  - PSD with  $P < 5\%$  and
  - PSD with  $P < 1\%$ .
3. GHT classification: Two definitions of normality were used:
  - GHT1: Normal fields were indicated by both the "within normal limits" and the "borderline" test outputs.
  - GHT2: Only visual field tests with a "within normal limits" result were considered normal.
 Tests showing only a "general reduction" of sensitivity result were considered normal.
4. Clusters of test points exhibiting reduced differential light sensitivity at the  $P < 0.05$  significance level with at least one point depressed at the  $P < 0.01$  significance level, as seen in the pattern deviation probability maps. Following the definition found in Anderson<sup>25</sup> and suggested by Katz et al.,<sup>23</sup> we defined a cluster as 3 or more nonedge adjacent depressed test points in the same hemifield.

### Eligibility

Patients with primary open angle glaucoma, normal tension glaucoma, exfoliation, and pigmentary glaucoma

were eligible. Tests from patients with angle closure, secondary and congenital glaucoma or with retinal disease, neuro-ophthalmologic and metabolic disorders, and systemic diseases affecting the visual field, were not eligible for the study. Tests from glaucomatous eyes in diabetic patients without retinopathy and from patients with cataract were also eligible.

Visual field-based information was not used as reference standard. The definition of glaucomatous damage was instead based on optic disc evaluation performed by a glaucoma expert. The information used by the expert to determine the presence of glaucomatous optic nerve damage included structural changes of the optic disc from fundus photographs and/or comprehensive descriptions of the optic disc appearance as found in patient records. A number of tests with normal visual fields were included based on the appearance of the corresponding optic discs.

### Selection of Glaucomatous Fields

A database search was performed in one of the Humphrey perimeters at the Department of Ophthalmology of Malmö University Hospital containing approximately 25,000 visual field tests from about 6000 individuals, and the corresponding patient records were retrieved. Starting alphabetically and choosing either a right or a left eye in an alternating order, the most recent SITA Standard visual field test with the 30-2 test pattern was retrieved from 1 out of every 5 individuals found in the database. In this manner, 588 visual field tests from 588 patients were extracted from the perimetric database.

As we wanted to explore the diagnostic ability of the compared methods in participants with mild and moderate glaucomatous changes, only tests with a mean deviation (MD) value (rounded to the nearest integer) better than or equal to  $-10$  dB were included. If a visual field test had an MD value worse than that, we used an earlier test from the same patient if damage was then within the acceptable range. Tests with fixation losses exceeding 20% or false positive response rates of more than 15% were excluded. None of the included visual fields were the patients' first tests, to avoid false defects due to lack of perimetric experience.<sup>29-31</sup> We also excluded any patients who had provided visual field tests previously used in training of the neural network.

### Selection of Tests From Healthy Individuals

This group was part of an existing normative database, created by applying a multicenter design and consisting of 335 visual field tests from 335 healthy individuals recruited in several different ways, as spouses or accompanying persons of patients, friends or relatives of personnel, etc. Inclusion of participants was based on criteria pertaining to the ocular status after an ophthalmologic examination. Visual field test results did not constitute grounds for exclusion, except where suspicious or pathologic visual fields explicable by ocular status or clearly artefactual test results were present.<sup>32</sup> From this database, the 213 tests randomly chosen to train and test

the ANN<sup>28</sup> were excluded. Of the remaining tests, 6 were lost due to corrupted data leaving 116 to form the healthy group in the current study.

### Neural Networks

The network structure and its training process have been previously described in detail.<sup>28</sup> Briefly, this was a fully connected feed-forward MLP with one 74 unit input layer, two hidden layers with 25 and 5 processing elements respectively, and one output layer, created with the Neural Network toolbox version 4.0 of MATLAB (The MathWorks Inc, Natick, MA). The single neuron output layer with a logistic transfer function provided the network's output: glaucoma or normal.

### Analyses

The output layer provides a network response in the form of any number in the range between 0 and 1, indicating normality and glaucoma, respectively. Setting a cut-off value (threshold) for the network responses enables separation of the 2 classes (normal and glaucoma), and measurement of the correct and incorrect responses. Sensitivity was defined as the ability of ANN to detect glaucomatous visual fields from eyes with evidence of possible or certain glaucoma, whereas specificity as the ability to classify visual fields from healthy eyes as normal. Sensitivity-specificity pairs were measured through adjustment of the network threshold over the spectrum of output values. We plotted the resulting receiver operating characteristic (ROC) curve and calculated the area under the curve.<sup>33,34</sup> For the PSD index, the GHT and cluster analysis, sensitivity and specificity values were calculated. The statistical significance of the difference in sensitivity and specificity between the compared methods (probability, *P*) along with the confidence intervals, based on a normal approximation of a binomial distribution, at the 95% significance level according to the score method, was also computed.<sup>35</sup> After application of the Bonferroni correction, statistical significant differences were considered those having a probability of  $< 0.003$ . Diagnostic accuracy was calculated by dividing the sum of true positive and true negative responses with the sum of true and false, positive and negative responses.<sup>36</sup>

This study was conducted according to the tenets of the Declaration of Helsinki and was approved by the Ethics Committee of Lund University.

## RESULTS

The proportion of normal and glaucomatous 30-2 SITA Standard visual field tests in the test data was 46.3% and 53.7%, respectively. The glaucoma group included 100 tests from 100 patients. The selection process is schematically presented in Figure 1. Patient age ranged between 41 and 95 years, with an average of 75 years. MD ranged from  $-10.42$  to  $+0.31$  dB (median:  $-6.16$  dB; mean:  $-5.77$  dB). According to the patient records, approximately one-third (28%) of the eyes had media opacities in the form of cataract (26%) or

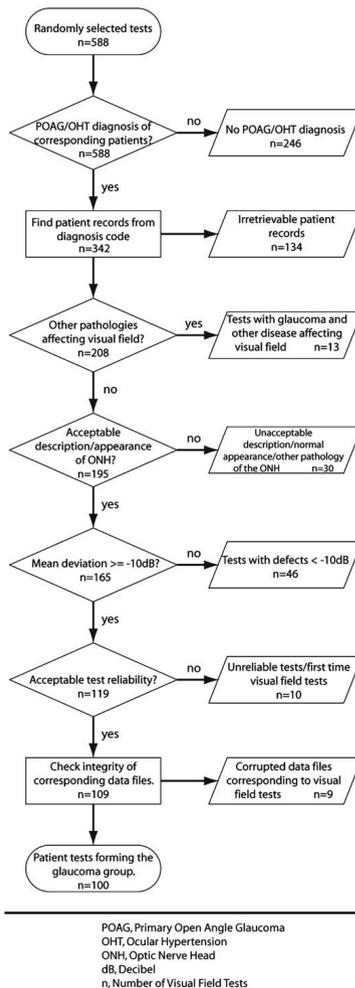
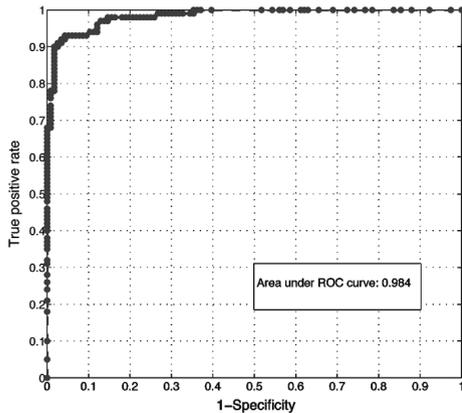


FIGURE 1. Flow chart illustrating the selection process of the glaucomatous visual field tests included in the study.

postoperative opacities of the posterior capsule (2%) at the time of the visual field examination.

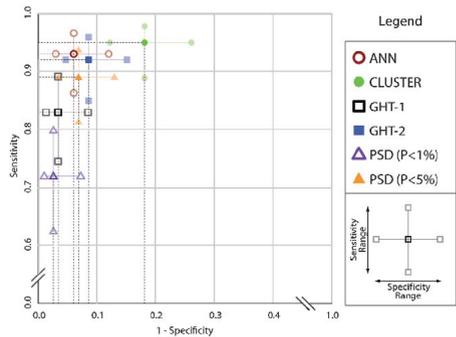
The 116 individuals in the healthy group had a mean age of 51 years (range 19 to 83). In this group MD ranged between  $-4.62$  and  $+2.4$  dB (median:  $+0.22$  dB; mean:  $+0.08$  dB).



**FIGURE 2.** Area under the ROC curve for the neural network output. A test with perfect discrimination would achieve an area of 1.0 under the ROC, whereas discrimination no better than random selection would produce an area of 0.5.

The area under the ROC curve for the ANN classification was 0.984 (Fig. 2). The network achieved a sensitivity of 93.0% at a specificity of 94.0% and had a diagnostic accuracy of 93.5%, with the network threshold set at 0.3. This level provided the best results with respect to specificity and sensitivity during previous training of the network.<sup>28</sup>

The value of diagnostic accuracy for the ANN was slightly but not significantly higher than that achieved with both GHT approaches, and the PSD  $P < 5\%$ . For these methods diagnostic accuracy exceeded 90% (Table 1). The accuracy of both the most sensitive cluster procedure and the most specific PSD  $P < 1\%$  was 87.9% and 85.6%, respectively. The ANN was more sensitive than the PSD  $< 1\%$  and more specific than the cluster approach ( $P < 0.0001$ ). The differences in sensitivity and specificity between ANN, PSD  $P < 5\%$  and GHT methods were not statistically significant. GHT1, where “borderline” test results were regarded as normal,



**FIGURE 3.** Sensitivity and specificity of classification for the compared methods, along with the 95% confidence intervals (sensitivity and specificity range). In this graph, the ANN approach is located closest to the point exhibiting perfect discrimination (sensitivity and specificity value of 1.0).

revealed relatively low sensitivity, 83%, but high specificity, 96.6%, whereas GHT2, where “borderline” test results were regarded as abnormal, revealed both sensitivity and specificity above 90%, as expected (Table 1).

**DISCUSSION**

Our ANN was able to distinguish between normal and glaucomatous fields slightly or considerably better than the other interpretation algorithms even though most differences did not reach statistical significance (Fig. 3 and Table 1). Our results were obtained using visual field tests from patients with moderate and early glaucoma, independent of the presence or absence of cataract.

There was rather good agreement in classification between the ANN and GHT. Among the 7 glaucoma patients misclassified by the network, 4 patients had GHT “within normal limits” and 2 patients had “borderline” test results. The best GHT algorithm, GHT2, misclassified 8 glaucoma patients whereas 9 had borderline test results. Comparison of the GHT1 and GHT2 showed, as

**TABLE 1.** Diagnostic Accuracy and Sensitivity With Specificity, Including the 95% Confidence Intervals, for the Different Interpretation Algorithms

	ANN	GHT1	GHT2	PSD $P < 5\%$	PSD $P < 1\%$	CLUSTER
Accuracy (%)	93.5	90.3	91.7	91.2	85.6	87.9
Sensitivity (%)	93.0	83.0	92.0	89.0	72.0	95.0
95% confidence interval	86.3 to 96.6	74.5 to 89.1	85 to 95.9	81.4 to 93.7	62.5 to 79.9	88.8 to 97.8
Specificity (%)	94.0	96.6	91.4	93.1	97.4	81.9
95% confidence interval	88.0 to 97.0	91.5 to 98.7	84.9 to 95.3	87.0 to 96.5	92.7 to 99.0	73.9 to 87.8

CLUSTER indicates approach based on identification of clusters of test points exhibiting reduced sensitivity, based on pattern deviation probability maps; GHT1 indicates Glaucoma Hemifield Test with “borderline” results indicating normal visual fields; GHT2, Glaucoma Hemifield Test with “borderline” results indicating glaucomatous visual fields; PSD  $P < 5\%$ , Pattern Standard Deviation outside the 5% normal limit; PSD  $P < 1\%$ , Pattern Standard Deviation outside the 1% normal limit.





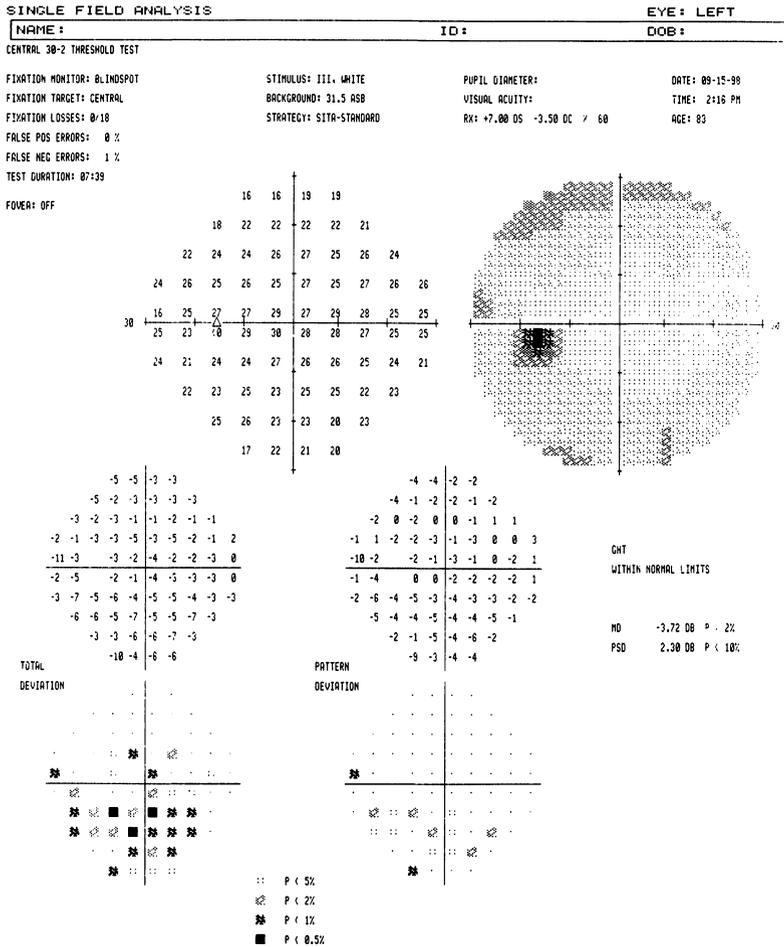


FIGURE 6. A visual field test from a glaucoma patient showing a shallow inferior defect on the pattern deviation probability map. This test was misclassified by all the compared algorithms.

the SITA normative database. Further, our comparison analysis should not be affected because any eventual bias would be equal to all the compared methods.

We used structural information, that is, optic disc appearance, as a reference standard to define normality and disease. By excluding visual field-related parameters in the reference standard, we avoided bias toward any particular diagnostic criterion derived from perimetric measurements.

There was an average age difference of 24 years between the normal and glaucoma groups. Although this difference was large, it should not affect the results of our analysis as all of the compared algorithms make use of age-corrected perimetric data.

Even if the results with the ANN were not statistically significant, the classification of visual fields by the ANN revealed the highest sensitivity at specificity values over 90%, providing an acceptable trade-off

between these 2 parameters of diagnostic accuracy (Fig. 3).

This is, to our knowledge, the first study to test the performance of a trained neural network for glaucoma diagnosis using an independent data set. In this sense, the results can also be viewed as a measure of the generalization ability of the previously proposed classification method. Despite the use of an independent set of patient data, our results on the performance of the ANN were very similar to those previously obtained.<sup>28</sup> This may imply a good level of reproducibility for our classifier. Previous studies evaluating ANN-based techniques for glaucoma diagnosis employed partitioning of the patient data set to obtain training and early-stopping subsets, testing the accuracy of classification in only a fraction of the patients. This necessitated resampling techniques to counteract the small size of the training set. The use of all available data solely for testing purposes generates a large and independent test set, increasing the strength of our study.

This study evaluated the applicability of ANN-based systems for classification of glaucomatous visual fields. The diagnostic accuracy of our ANN can be attributed in part to the use of preprocessed data based on probability maps. It is desirable for a glaucoma diagnosis test, operating in a clinical environment, to be able to account for cataract associated loss of sensitivity without compromising its ability to recognize glaucomatous visual field loss.

Further studies are needed to investigate the ability of classification methods based on ANNs to discern patterns of glaucomatous damage from visual field loss caused by other ophthalmic conditions. It has been suggested that the ability of machine classifiers to use as input combined data from structural and functional tests may lead to better tools for glaucoma diagnosis.<sup>7</sup> Incorporation of different test modalities such as structural measurements<sup>37</sup> and subsequent optimization might further facilitate the applicability of ANNs as clinical diagnostic tools.

## REFERENCES

- Heijl A, Lindgren G, Olsson J. A package for statistical analysis of computerized fields. *Doc Ophthalmol Proc Ser.* 1987;49:153-168.
- Heijl A, Patella VM. *Essential Perimetry. The Field Analyzer Primer.* 3rd ed. Dublin, CA: Carl Zeiss Meditec Inc; 2002:44-69.
- Heijl A, Bengtsson B. Early visual field defects in glaucoma: a study of eyes developing field loss. In: Bucci MG, ed. *Glaucoma: Decision Making in Therapy.* Milan: Springer Verlag; 1996:75-78.
- Heijl A, Lindgren G, Olsson J, et al. Visual field interpretation with empiric probability maps. *Arch Ophthalmol.* 1989;107:204-208.
- Mutlukan E, Keating D. Visual field interpretation with a personal computer based neural network. *Eye.* 1994;8:321-323.
- Lietman T, Eng J, Katz J, et al. Neural networks for visual field analysis: how do they compare with other algorithms? *J Glaucoma.* 1999;8:77-80.
- Zangwill LM, Chan K, Bowd C, et al. Heidelberg retina tomograph measurements of the optic disc and parapapillary retina for detecting glaucoma analyzed by machine learning classifiers invest. *Ophthalmol Vis Sci.* 2004;45:3144-3151.
- Bowd C, Medeiros FA, Zhang Z, et al. relevance vector machine and support vector machine classifier analysis of scanning laser polarimetry retinal nerve fiber layer measurements. *Invest Ophthalmol Vis Sci.* 2005;46:1322-1329.
- Burgansky-Eliash Z, Wollstein G, Chu T, et al. Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. *Invest Ophthalmol Vis Sci.* 2005;46:4147-4152.
- Huang ML, Chen HY. Development and comparison of automated classifiers for glaucoma diagnosis using stratus optical coherence tomography. *Invest Ophthalmol Vis Sci.* 2005;46:4121-4129.
- Spenceley SE, Henson DB, Bull DR. Visual field analysis using artificial neural networks. *Ophthalmic Physiol Opt.* 1994;14:239-248.
- Sample PA, Chan K, Boden C, et al. Using unsupervised learning with variational Bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci.* 2004;45:2596-2605.
- Goldbaum MH, Sample PA, Zhang Z, et al. Using unsupervised learning with independent component analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci.* 2005;46:3676-3683.
- Sample PA, Boden C, Zhang Z, et al. Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields. *Invest Ophthalmol Vis Sci.* 2005;46:3684-3692.
- Tucker A, Vinciotti V, Liu X, et al. A spatio-temporal Bayesian network classifier for understanding visual field deterioration. *Artif Intell Med.* 2005;34:163-177.
- Goldbaum MH, Sample PA, Chan K, et al. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Invest Ophthalmol Vis Sci.* 2002;43:162-169.
- Poinoosawmy D, Tan JC, Bunce C, et al. The ability of the GDx nerve fibre analyser neural network to diagnose glaucoma. *Graefes Arch Clin Exp Ophthalmol.* 2001;239:122-127.
- Bowd C, Chan K, Zangwill LM, et al. Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc. *Invest Ophthalmol Vis Sci.* 2002;43:3444-3454.
- Chan K, Lee TW. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng.* 2002;49:961-973.
- Goldbaum MH, Sample PA, White H, et al. Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci.* 1994;35:3362-3373.
- Åsman P, Heijl A. Glaucoma Hemifield Test. Automated visual field evaluation. *Arch Ophthalmol.* 1992;110:812-819.
- Åsman P, Heijl A. Evaluation of methods for automated Hemifield analysis in perimetry. *Arch Ophthalmol.* 1992;110:820-826.
- Katz J, Sommer A, Gaasterland DE, et al. Comparison of analytic algorithms for detecting glaucomatous visual field loss. *Arch Ophthalmol.* 1991;109:1684-1689.
- Chauhan BC, Drance SM, Lai C. A cluster analysis of threshold perimetry. *Graefes Arch Clin Exp Ophthalmol.* 1989;27:216-220.
- Anderson DR. *Automated Static Perimetry.* St. Louis: Mosby-Year Book Inc; 1992:123.
- Åsman P, Heijl A. Arcuate cluster analysis in glaucoma perimetry. *J Glaucoma.* 1993;2:13-20.
- Åsman P, Heijl A, Olsson J, et al. Spatial analyses of glaucomatous visual fields; a comparison with traditional visual field indices. *Acta Ophthalmol.* 1992;70:679-686.
- Bengtsson B, Bizios D, Heijl A. Effects of input data on the performance of a neural network to separate normal and glaucomatous visual fields. *Invest Ophthalmol Vis Sci.* 2005;46:3730-3736.
- Heijl A, Lindgren G, Olsson J. The effect of perimetric experience in normal subjects. *Arch Ophthalmol.* 1989;107:81-86.
- Wild JM, Dengler-Harles M, Searle AE, et al. The influence of the learning effect on automated perimetry in patients with suspected glaucoma. *Acta Ophthalmol Scand.* 1989;67:537-545.
- Heijl A, Bengtsson B. The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol.* 1996;114:19-22.
- Bengtsson B, Heijl A. Inter-subject variability and normal limits of the SITA Standard, SITA Fast, and the Humphrey Full Threshold

- computerized perimetry strategies, SITA STATPAC. *Acta Ophthalmol Scand*. 1999;77:125–129.
33. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39:561–577.
  34. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36.
  35. Altman DG. Comparing groups—categorical data. In: Altman DG, ed. *Practical Statistics for Medical Research*. London: Chapman and Hall; 1991:229–276.
  36. Friis RH, Sellers TA. *Epidemiology for Public Health Practice*. 3rd ed. Sudbury, MA: Jones and Bartlett Publishers; 2004:384.
  37. Brigatti L, Hoffman D, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol*. 1996;121:511–521.



# Paper III





# Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by Stratus OCT

Dimitrios Bizios, Anders Heijl, Jesper Leth Hougaard and Boel Bengtsson

Department of Clinical Sciences, Ophthalmology, Malmö University Hospital, Lund University, Malmö, Sweden

## ABSTRACT.

**Purpose:** To compare the performance of two machine learning classifiers (MLCs), artificial neural networks (ANNs) and support vector machines (SVMs), with input based on retinal nerve fibre layer thickness (RNFLT) measurements by optical coherence tomography (OCT), on the diagnosis of glaucoma, and to assess the effects of different input parameters.

**Methods:** We analysed Stratus OCT data from 90 healthy persons and 62 glaucoma patients. Performance of MLCs was compared using conventional OCT RNFLT parameters plus novel parameters such as minimum RNFLT values, 10th and 90th percentiles of measured RNFLT, and transformations of A-scan measurements. For each input parameter and MLC, the area under the receiver operating characteristic curve (AROC) was calculated.

**Results:** There were no statistically significant differences between ANNs and SVMs. The best AROCs for both ANN (0.982, 95%CI: 0.966–0.999) and SVM (0.989, 95% CI: 0.979–1.0) were based on input of transformed A-scan measurements. Our SVM trained on this input performed better than ANNs or SVMs trained on any of the single RNFLT parameters ( $p \leq 0.038$ ). The performance of ANNs and SVMs trained on minimum thickness values and the 10th and 90th percentiles were at least as good as ANNs and SVMs with input based on the conventional RNFLT parameters.

**Conclusion:** No differences between ANN and SVM were observed in this study. Both MLCs performed very well, with similar diagnostic performance. Input parameters have a larger impact on diagnostic performance than the type of machine classifier. Our results suggest that parameters based on transformed A-scan thickness measurements of the RNFL processed by machine classifiers can improve OCT-based glaucoma diagnosis.

**Key words:** artificial neural networks – glaucoma diagnosis – machine learning – optical coherence tomography – retinal nerve fibre layer

## Introduction

Degeneration of the retinal nerve fibre layer (RNFL) can occur at early stages of glaucoma (Sommer et al. 1991; Quigley et al. 1992; Tuulonen et al. 1993; Harwerth et al. 1999; Kerrigan-Baumrind et al. 2000). Optical coherence tomography is an interferometric noninvasive imaging technique that enables high-resolution cross-sectional images of tissues and provides quantitative measurements of ocular structures such as the RNFL (Chang & Budenz 2008), with good reproducibility (Schuman et al. 1996; Paunescu et al. 2004; Budenz et al. 2005a).

The commercially available RNFL thickness (RNFLT) parameters implemented in the software of Stratus OCT (Carl Zeiss Meditec Inc., Dublin, CA, USA) include measurements of average, or highest and lowest thickness values, at defined sectors of the scanned RNFL. This software assists interpretation of RNFLT measurements by highlighting parameters outside age-corrected reference values.

Several studies investigated the ability of the provided OCT RNFLT parameters to differentiate between normal and glaucomatous eyes (Medeiros et al. 2004, 2005; Budenz et al. 2005b; Leung et al. 2005; Wollstein et al. 2005; Shah et al. 2006; Sihota et al. 2006; Hougaard et al. 2007;

Chang & Budenz 2008). OCT thickness parameters reached sensitivities of 70–80% at specificity values of more than 90% and areas under the receiver operating characteristic curve (AROC) of about 0.90, although these results might be better than the actual performance of the studied parameters because in some studies (Budenz et al. 2005b; Leung et al. 2005; Wollstein et al. 2005) no corrections were made for the significant factor of age, which is correlated to measured RNFLT. Parameters representing measurements of the inferior and superior peripapillary RNFL sectors and the average RNFLT provided the best diagnostic performance.

In an effort to improve the diagnostic accuracy of OCT measurements, previous studies have also explored the implementation of linear discriminant functions and machine learning algorithms using RNFLT measurements as input parameters (Essock et al. 2003; Hougaard et al. 2004; Burgansky-Eliash et al. 2005; Huang & Chen 2005; Chen et al. 2006; Manassakorn et al. 2006; Naithani et al. 2007). Artificial neural networks (ANNs) and support vector machines (SVMs) are machine learning algorithms able to perform nonlinear classification. In contrast to conventional statistical techniques, such as linear discriminant analysis, these methods do not fit the data into a pre-existing set of model variables; instead they learn to nonlinearly adapt the classification decision based on the data that is presented. Provided a representative and adequately large training dataset in relation to the input parameters, this approach produces robust classifiers that are able to generalize well even in the absence of a model of the underlying process and are insensitive to noise and outliers in the data (Haykin 1999).

In the field of ophthalmology and glaucoma diagnosis, machine classifiers have been used for classification of visual field data (Brigatti et al. 1996; Chan et al. 2002; Goldbaum et al. 2002; Bengtsson et al. 2005; Bizios et al. 2007) and structural measurements of the RNFL and optic disc [Huang & Chen 2005; Burgansky-Eliash et al. 2005; Naithani et al. 2007; Bowd et al. 2002, 2005; Poinoosawmy et al. 2001; Zangwill et al. 2004], as well as for detecting visual field pro-

gression (Sample et al. 2005; Tucker et al. 2005). These methods produced AROCs larger than 0.95 and were more effective in discriminating between normal and glaucomatous eyes than the commercially available RNFLT parameters (Burgansky-Eliash et al. 2005; Huang & Chen 2005; Naithani et al. 2007). ANNs have a long history of development and successful deployment in many scientific fields, including medicine. SVMs, despite their recent conception, have shown performance improvements compared to more conventional approaches. Even though ANNs and SVMs have different theoretical principles outlining their function, both techniques have shown high classification accuracy. Although their diagnostic performance based on OCT measurements has been investigated (Burgansky-Eliash et al. 2005; Huang & Chen 2005; Naithani et al. 2007), none has compared the performance of both classifiers on the same OCT dataset.

The primary aim of our study was thus to compare the ability of ANN and SVM classifiers to detect differences in age-corrected RNFLT measurements between healthy persons and patients with predominantly early glaucomatous visual field loss. Performance analysis of the ANN and SVM classifiers should also provide a measure of the diagnostic ability of the standard and novel OCT RNFLT parameters.

## Methods

The study was conducted according to the tenets of the Declaration of Helsinki and was approved by the Committee for Research Ethics at Lund University, Sweden. Informed consent was obtained from all participants. Inclusion and exclusion criteria for healthy individuals and patients with glaucoma examined in this study have been described in detail elsewhere (Hougaard et al. 2007) and are summarized below.

### Healthy individuals

The normal OCT measurements were part of a larger collection of OCT data acquired from healthy individuals. The individuals were mainly recruited by random selection of pre-

sumably healthy persons living in Malmö, Sweden.

About two-thirds of this database (178 healthy subjects) were used for the construction of a normative RNFL model with reference limits corrected for both age and refractive status (Hougaard et al. 2006). All OCT thickness data (i.e. all conventional and novel RNFL thickness parameters) for both normal and glaucoma groups included in this study were corrected for age and refractive status (spherical equivalent) based on this normative database. From the remaining one-third of healthy individuals (90 subjects) whose measurements were not part of the normative database, one randomly selected eye was included in this study as the healthy group.

Subjects with visual acuity (VA)  $\geq 0.5$ , intraocular pressure (IOP)  $< 22$  mmHg, and refractive error  $\leq 5$  diopters (D) sphere and  $< 3$  D cylinder were eligible. Healthy subjects were also required to have normal visual fields (VF) as examined by frequency doubling perimetry screening program C20 – 1 (Welch Allyn, Inc, Skaneateles Falls, NY, USA), as well as a normal optic nerve head (ONH) and RNFL, as judged from photographs by two examiners.

### Glaucoma patients

The patients with glaucoma were mainly recruited among patients who had visited the department of Ophthalmology at Malmö University Hospital during the last 3 years prior to inclusion. All records of patients with glaucoma were retrieved and reviewed, and those who fulfilled the inclusion criteria and were not involved in other ongoing studies were invited to participate. Four of the included patients were newly diagnosed with glaucoma and had been detected during recruitment of the presumably healthy persons.

Eligible patients were between 40 and 80 years of age, with VA  $\geq 0.5$  and a diagnosis of primary open angle glaucoma (POAG), normal tension glaucoma, pigmentary glaucoma or pseudo-exfoliation glaucoma. The included patient's eyes had reproducible visual field defects (SITA Standard 24-2 tests), with mean deviation (MD)  $\geq -12$  dB, pattern standard

deviation outside the 95% normal limit and classified by the Glaucoma Hemifield Test as falling outside normal limits. The visual field defects measured by standard automated perimetry (SAP) corresponded to glaucomatous changes in the optic nerve head (ONH) and/or the RNFL judged by examination of photographs.

Patients with a history of serious ocular trauma, findings of systemic or retinal pathology affecting the visual field and refractive errors  $> 5$  D sphere or  $\geq 3$  D cylinder were not eligible, as were patients with cortical (LOCS C  $>$  II) and subcapsular cataract (LOCS p  $>$  0).

One eye of each patient was included in the study. If both eyes were eligible, the eye with the better visual field as measured by the MD value was included.

#### OCT tests

The OCT scans were obtained, at 750  $\mu$ W, with the 'Fast RNFL Thickness' scan protocol and software version A4.0.3 of the Stratus OCT. We used average thickness values of the three circumpapillary scans, each with 256 measurement points. Two OCT scan series were performed in each eye by the same experienced examiner. OCT tests with overt image artefacts or obvious software errors in the definition of the RNFL inner border were not considered eligible for further analysis. To be included, single scans had to be of good quality (i.e. having signal strength higher than 8 on the 10-grade scale). If both scan series were eligible, the first one was used in our analysis. The effect of age and refraction on OCT measurements of the RNFLT is already known (Hougaard et al. 2006). We have thus corrected all OCT thickness data for both age and refractive status (spherical equivalent) based on our normative database, in a way analogous to that of Statpac for perimetric tests.

#### Conventional RNFLT parameters

Tested parameters included mean RNFLT of the full scan circle; mean RNFLT of the temporal (315–45°, right eye terminology), superior (45–135°), nasal (135–225°, right eye terminology) and inferior (225–315°)

quadrants and average thickness values of each of the 12 clock hr sectors. The thickest (maximum) points of the superior and inferior quadrants (superior maximum or 'Smax' and inferior maximum or 'Imax') and the difference in thickness between the thickest and thinnest measurement points of the RNFLT scan (Max–Min) were also included in our analyses.

#### New RNFLT parameters

We evaluated the single A-scan value of the thinnest RNFLT measurement in the superior and inferior quadrants (superior minimum or 'Smin' and inferior minimum or 'Imin'). We conducted this evaluation based on the assumption that the thinnest RNFLT measurements should be more sensitive to localized, deep defects of RNFL compared to the thickest measurement points. The 'Smin' and 'Imin' parameters are not automatically defined and are not presented in the test output of the commercially available interpretation tool.

We also calculated and evaluated thickness values corresponding to the higher and lower 10th percentile (i.e. the 90th and 10th percentile of measured thickness) derived from the 64 RNFLT measurements for each of the superior and inferior quadrants, and of the whole RNFL scan circle. Parameters based on defined percentiles of measured thickness instead of highest and lowest values could conceivably be more robust and less likely to be affected by test error measurements and outliers in the data. The resulting new parameters based on percentiles were as follows:

- (1) The superior quadrant 90th and 10th percentile thickness values ('Smax\_90' and 'Smin\_10'),
- (2) The inferior quadrant 90th and 10th percentile thickness values ('Imax\_90' and 'Imin\_10') and
- (3) The difference in thickness between the 90th and 10th percentile of the full scan circle ('Max\_90–Min\_10').

The use of A-scan thickness values as input data to automated classifiers is a difficult task because of computational problems dealing with the large number of parameters (related to the number of A-scans). Novel techniques

for processing information as derived from A-scans could efficiently reduce the amount of data while preserving relevant diagnostic information (Bizios et al. 2007; Invest Ophthalmol Vis Sci 48: ARVO E-Abstract 525). To examine whether such an approach would improve classification, we created a novel parameter by transforming the age and refraction corrected thickness values of the 256 A-scans of each OCT image into a set of six values that we then used as input data in our machine learning classifiers. These values do not correspond to any particular A-scan, but reflect all 256 A-scan measurements of each OCT test. Data transformation was accomplished by the local tangent space alignment (LTSA), a dimensionality reduction algorithm based on manifold learning (cf. Appendix) (Zhang & Zha 2004).

We thus tested a total of 28 different OCT RNFLT parameters, consisting of 20 conventional and 8 new parameters, using both ANNs and SVMs.

#### Artificial neural networks

In this study, we used an ANN ensemble (cf. Appendix) of multilayer perceptrons (MLPs). The ensemble output was the averaged output of thirty-five MLPs, each consisting of a 12-neuron input layer, a hidden layer with six neurons and an output layer with one neuron and a logistic transfer function. The ANNs were programmed and trained in MATLAB's neural network toolbox version 5.0 (The MathWorks Inc, Natick, MA, USA) with the scaled conjugate gradient algorithm (Møller 1993).

#### Support vector machines

Our SVM (cf. Appendix) used a radial basis function kernel and was trained by a variation of Platt's sequential minimal optimization algorithm (Fan et al. 2005). Programming, testing and training of the SVM were performed in Python (Python Software Foundation) and MATLAB with the Libsvm software (Chang & Lin 2001). The C and  $\gamma$  parameters of the algorithm were determined by a global optimization technique based on simulated annealing (Imbault & Lebart 2004).

**Training of machine learning classifiers**

Training and testing were performed with each RNFLT parameter separately, as well as with all 17 conventional RNFLT parameters used conjointly. To maximize the use of our collected data and avoid bias, the 10-fold cross-validation resampling method was employed. Accordingly, all data were randomly divided into 10 subsets, each containing approximately the same number of healthy and glaucomatous OCT thickness measurements. Nine subsets were used for training the classifiers while the remaining subset was used for testing classification performance. In training the ANN ensemble, one of the training subsets was used for early stopping of network training to avoid overfitting of the MLPs. Training data for the thirty-five MLPs were created with bagging (cf. Appendix) from the remaining eight subsets. ANN training was repeated by keeping the same test subset and changing the early stopping set, until all training data were used both in training and early stopping of the ANNs and the classification results were averaged. The training and testing process for the ANN ensemble and the SVM were iterated, each time with a different test set, and the results were merged to produce a single output for each classifier.

**Analyses**

All OCT thickness data, i.e. all A-scans and the derived novel and conventional RNFLT parameters, were corrected for age and refraction prior to any analyses. By modelling the relationship of age, refraction and measured RNFLT on our normative database, we conducted a linear regression analysis and used the derived coefficients to calculate the corrected values of measured RNFLT. ROC curves were constructed from the specificity and sensitivity values of each MLC. AROCs were used as a measure of the classification performance of ANN and SVM classifiers, as well as of the diagnostic ability for the examined OCT RNFLT parameters. Confidence intervals for the AROCs at the 5th percentile (95% CI) were derived by the SPSS statistical software (version 16.0, SPSS Inc,

Chicago Ill, USA). Testing for significant differences between AROCs was performed with DeLong's nonparametric method (DeLong et al. 1988). For testing significant differences in the distribution of gender between the healthy and glaucoma groups, the Chi square test was used, whereas significance testing for differences in the distributions of age, visual acuity and refractive error was performed with the Mann-Whitney test.

**Results**

OCT data of 62 patients with glaucoma and 90 healthy individuals were used as input for the MLCs. Subject demographics are summarized in Table 1. Included OCT tests from the healthy group had an average signal strength score of 9.96, while the average signal strength score for the glaucoma group was 9.69.

Of the 28 RNFLT parameters tested, the temporal and nasal quad-

rants, the Smax, Smax\_90 and Imax parameters as well as 10 of the 12 clock hr sectors produced AROCs < 0.9. Consequently, we chose not to present those results in detail. The performance of MLCs based on conventional and new parameters with AROCs of more than 0.90 is presented in Tables 2 and 3, respectively. Comparison of the AROCs for all studied RNFLT parameters, revealed no statistically significant differences between the ANN and SVM machine classifiers.

**Conventional RNFLT parameters**

The average RNFLT of the full circle scan, the superior and inferior quadrants, the 6th and 11th clock hr sectors and the Max-Min thickness measurements provided good classification performance. The diagnostic ability of the conventional OCT RNFLT parameters without the use of MLCs has been previously reported

**Table 1.** Demographic data of included healthy subjects and patients with glaucoma. All values except for gender are represented as median and min-max values.

	Healthy (n = 90)	Glaucoma (n = 62)	p - value (Mann-Whitney test)
Gender (female/male)	60/30	45/17	NS ( $\chi^2$ test)
Age (years)	61.5 (22-79)	71.5 (52-78)	< 0.0001
Visual Acuity (decimal scale)	1.0 (0.7-1.0)	1.0 (0.6-1.0)	NS
Refractive error (spherical equivalent)	+ 0.1 (-5.5 to + 4.4)	+ 0.5 (-4.0 to + 4.3)	NS
Visual Field (MD)	-	-4.5 (-11.77 to 0.00)	-

NS = nonsignificant, MD = mean deviation value in decibel, as measured by 24-2 SITA Standard program.

**Table 2.** Classification performance based on the area under the receiver operating characteristic (AROC) curve, with the corresponding 95% confidence intervals (95% CI), for artificial neural networks (ANN) and support vector machines (SVM) trained on the conventional retinal nerve fibre layer thickness (RNFLT) parameters.

Parameters	ANN	SVM
	AROC (95% CI)	AROC (95% CI)
Full circle average	0.943 (0.905-0.981)	0.940 (0.901-0.982)
Superior quadrant	0.926 (0.872-0.963)	0.922 (0.876-0.968)
Inferior quadrant	0.930 (0.878-0.972)	0.922 (0.870-0.976)
Clock hr 11	0.935 (0.895-0.975)	0.933 (0.893-0.977)
Clock hr 6	0.929 (0.886-0.970)	0.912 (0.870-0.969)
Max-Min*	0.942 (0.905-0.982)	0.927 (0.881-0.979)
17 parameters†	0.977 (0.958-0.995)	0.977 (0.959-0.999)

\* Max-Min: difference in thickness between maximum and minimum RNFLT measurements.

† 17 parameters: average RNFLT of the full circle scan as well as RNFLT of the four quadrants and 12 clock hr sectors.

**Table 3.** Performance, as measured by the area under receiver operating characteristic curve (AROC) with the corresponding 95% confidence intervals (95% CI), for artificial neural networks (ANN) and support vector machines (SVM) trained on the new retinal nerve fibre layer thickness (RNFLT) parameters.

Parameters	ANN	SVM
	AROC	AROC
	(95% CI)	(95% CI)
Transformed A-scan thickness measurements	0.982 (0.966–0.999)	0.989 (0.979–1.0)
Imax_90*	0.91 (0.857–0.961)	0.904 (0.850–0.959)
Smin†	0.919 (0.876–0.962)	0.908 (0.858–0.958)
Imin‡	0.916 (0.862–0.970)	0.906 (0.850–0.969)
Smin_10§	0.916 (0.872–0.958)	0.909 (0.862–0.956)
Imin_10¶	0.915 (0.859–0.969)	0.909 (0.852–0.971)
Max_90–Min_10**	0.946 (0.907–0.983)	0.940 (0.898–0.986)

\* Imax\_90: inferior quadrant 90th percentile RNFLT.  
 † Smin: superior quadrant minimum RNFLT.  
 ‡ Imin: inferior quadrant minimum RNFLT.  
 § Smin\_10: superior quadrant 10th percentile RNFLT.  
 ¶ Imin\_10: inferior quadrant 10th percentile RNFLT.  
 \*\* Max\_90–Min\_10: difference between 90th and 10th percentiles of the full circle RNFLT.

on this study population (Hougaard et al. 2007). The full circle average RNFLT parameter measured with the FAST protocol exhibited at the 5th normative percentile a sensitivity of 76% at a specificity of 99%, which was not significantly different than the

performance achieved by MLCs using the same parameter as input data. Using 17 conventional RNFLT parameters (average thickness of the full circle scan as well as RNFLT of the four quadrants and 12 clock hr sectors) as input resulted in improved

classification accuracy for both the ANN (0.977, 95% CI: 0.958–0.995) and SVM (0.977, 95% CI: 0.959–0.999) as expected.

**New RNFLT parameters**

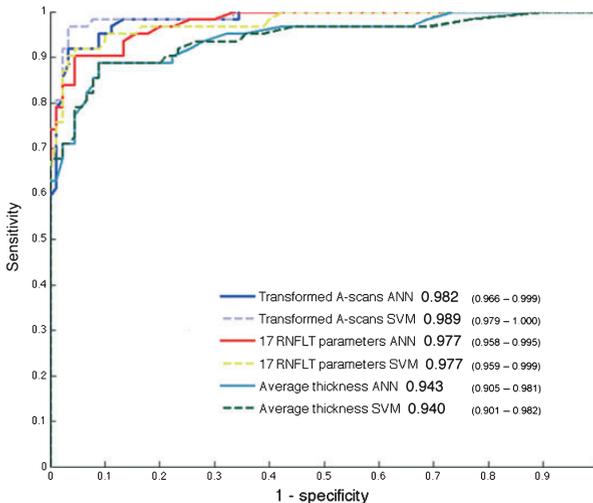
The A-scan thickness values transformed by the LTSA algorithm provided the largest AROCs of all tested conventional and new parameters for both ANN (0.982, 95%CI: 0.966–0.999) and SVM (0.989, 95% CI: 0.979–1.0) classifiers (Fig. 1).

Only our SVM trained on the transformed A-scan thickness values (sensitivity of 96.8% at specificity of 96.7%) performed significantly better than ANNs or SVMs trained on any of the single RNFLT parameters ( $p \leq 0.038$ ) (Table 4). SVM performance based on this input was also better than the performance of the average RNFLT of the full scan circle without the use of MLCs (McNemar’s test:  $p = 0.013$ ).

The difference in measured thickness between the highest and lowest tenth percentiles (Max\_90–Min\_10) gave the best classification results compared to the other single OCT parameters, even though for the parameters with AROCs > 0.9 the differences did not reach statistical significance. ANN and SVM performance of the lowest percentile of superior and inferior quadrants, and the Smin and Imin parameters (AROCs over 0.90, Table 3), was slightly but not significantly better than the performance of the conventional Smax and Imax parameters.

**Discussion**

This study is the first to our knowledge to apply a nonlinear dimensionality reduction method based on manifold learning for the transformation of OCT A-scan measurements, and to apply optimized MLCs such as an ANN ensemble for the classification of OCT test data. Our ANNs and SVMs, two of the most widely used classifiers in the machine learning paradigm, exhibited high performance with input based on RNFLT measurements only, without the use of any ONH data. Despite differences in the way ANNs and SVMs perform classification tasks, both methods were successful in recognizing pertinent



**Fig. 1.** Input based on the average RNFLT thickness produced good results for both classifiers. The input of 17 conventional RNFLT parameters further increased classification accuracy. The best results for ANN and SVM were achieved by using transformed A-scans as input.

**Table 4.** Significance testing (probability value) of the performance between two multivalued parameters (transformed A-scan thickness values and the conjointly used 17 conventional RNFLT parameters) and the best single-valued parameter (Max\_90–Min\_10); performance is measured as the area under the receiver operating characteristic (AROC) curve and the corresponding confidence intervals at 95% confidence level (95% CI) are presented in parenthesis.

	17 RNFLT parameters		Transformed A-scan thickness values	
	ANN	SVM	ANN	SVM
AROCs (95% CI)	0.977 (0.958–0.995)	0.977 (0.959–0.999)	0.982 (0.959–0.999)	0.989 (0.959–0.999)
Max_90–Min_10*				
ANN 0.946 (0.907 – 0.983)	0.156	0.170	0.084	0.038†
SVM 0.940 (0.898 – 0.986)	0.143	0.073	0.082	0.018†

ANN = artificial neural network; SVM = support vector machine.

\* Max\_90–Min\_10: difference between the highest and lowest tenth percentiles of the full circle RNFLT measurements.

† Statistical significant difference.

patterns in the training data and applying this knowledge to classify new data, achieving similar performance. We found no significant differences in classification of normal and glaucomatous OCT tests between the two machine classifiers trained on the same input parameters.

When evaluating, however, the effect different input parameters had on ANN and SVM classification accuracy, we found significant differences in performance depending on the type of input used by the classifiers. Studies (Burgansky-Eliash et al. 2005; Huang & Chen 2005) have shown that reduction in the number of OCT RNFLT parameters used by machine learning classifiers can yield improvements in classification performance. Instead of using only the conventional OCT RNFLT parameters, we decided to use all 256 A-scan thickness values and reduce their complexity by means of a dimensionality reduction technique. This novel approach significantly improved SVM performance by enabling extraction of useful information from a larger set of unprocessed thickness data (A-scans).

Huang & Chen (2005) employed principal component analysis (PCA) to linearly reduce the number of 25 ONH and RNFLT input parameters, while Burgansky-Eliash et al. (2005) chose 8 of 38 ONH and RNFLT parameters that were closest related to visual field MD value. Using visual field information for reducing the number of OCT parameters could introduce bias, because these parameters would be correlated to information used in the selection of the study groups. For this reason, our data

reduction approach utilized structural OCT data only.

In a study by Huang & Chen (2005), an MLP trained with another technique (the steepest gradient descent with momentum algorithm) achieved an AROC of 0.874. Learning in our ANN was accomplished with the scaled conjugate gradient method, which has generally better training performance (Haykin 1999). Moreover, the ensemble approach provides more reliable results than any single ANN classifier, by decreasing the overall variability of ANN responses (Haykin 1999). The best SVM classifier presented in the study by Burgansky-Eliash et al. (2005) achieved an AROC of 0.981, using as input both RNFL thickness and ONH measurements not corrected for age or refraction. Our results suggest that the high performance of our MLCs can be attributed to the choice of input type, as well as the choice of classifier architecture and algorithms employed during the training phase.

Age differed significantly between the glaucoma and the healthy groups in our data set. Knowing the effect of age on OCT estimation of the RNFLT, we would expect the age difference to affect the performance of our classifiers. Training the classifiers on RNFLT parameters not corrected for age, as expected resulted in larger AROCs (AROCs > 0.99). Because we have quantified the effect of age on RNFLT measurements, we chose to correct all data prior to introducing them as input to the classifiers. In this way, we did not have to include age as a separate input to the classifiers. The inclusion of age would create input data with different measurement

units (age and thickness values) and would not allow any insight on the way classifiers use age information in their classification decision. We would thus be unable to utilize our prior knowledge concerning the influence of age in estimation of RNFLT. Hougaard et al. (2006) have previously shown a significant influence of refractive status, in addition to that of age, on normal RNFL thickness measurements by the Stratus OCT. We utilized our own normative database (Hougaard et al. 2006) to accomplish the appropriate corrections of age and refraction.

Despite the sole use of OCT-derived information for training the classifiers, our selection criteria required both functional and corresponding structural glaucomatous defects to be present in the glaucoma group, as well as normal visual field function and a normal RNFL in the healthy group. The definition of a healthy RNFL was, however, based on assessment of RNFL and/or ONH photographs, which does not have perfect agreement with OCT classification. Our definition of glaucoma and normality may result in higher specificity and sensitivity results, but that should not affect the performance comparison between ANN and SVM because all analyses are based on the same material. Thus, our classifier performance with AROCs approaching 1.0 (i.e. perfect discrimination) should be viewed in the context of the primary aim of this study (i.e. the comparison of different MLCs in classification of OCT test data).

It is also possible that our classifiers have a certain degree of bias towards the data used in their training and

deployment, and one would expect their performance to slightly decrease when tested on a different population. Although we used 10-fold cross-validation to minimize this bias, our results have to be verified in an independent sample. Other limitations in the current study include the use of an unbalanced dataset, because normal tests comprised more than half of all OCT tests. The ensemble method can counteract some of the potential problems with training classifiers on unbalanced data. Moreover, it is conceivable that other types of machine classifiers could be equally successful in classification of OCT RNFLT values. Similarly, the LTSA algorithm is only one of several techniques that could be applied to reduce the complexity and extract features from OCT data. We did not conduct an exhaustive comparison between all available machine learning algorithms and complexity reduction techniques, because that would be practically not feasible within the constraints of the study design and would adversely affect any significance testing.

The similar performance of our two machine learning algorithms and the high classification accuracy of the SVM based on the transformed A-scan measurements further supports the argument that the type of input data seem to be more important than the type of classification method. Our results also suggest that data-processing techniques like LTSA might better represent relevant diagnostic information found in the A-scan measurements compared to current RNFLT parameters. Performance improvements from methods such as LTSA, leading to more effective representation of structural information, may be even more prominent in future generations of OCT instruments that provide increasing numbers of test parameters.

### Acknowledgements

This study is supported by grants K2005-74X-1426-13A and K2005-74BI-15375-01A from the Swedish research council [vetenskapsrådet], by the foundation of Crown Princess Margareta for visually handicapped [Stiftelsen Kronprincessan Margaretas arbetsnämnd för synskadade (KMA)], Alcon and by the Järnhardt foundation [Järnhardts stiftelse]. This study

and all attached material is an original work, and no other person's work has been used without due acknowledgement. The Department of Clinical Sciences, Ophthalmology at Malmö University Hospital receives research funding from Carl Zeiss Meditec. No conflicts of interest exist between the authors and other parties.

### References

Bengtsson B, Bizios D & Heijl A (2005): Effects of input data on the performance of a neural network in distinguishing normal and glaucomatous visual fields. *Invest Ophthalmol Vis Sci* **46**: 3730–3736.

Bizios D, Heijl A & Bengtsson B (2007): Trained artificial neural network for glaucoma diagnosis using visual field data: a comparison with conventional algorithms. *J Glaucoma* **16**: 20–28.

Bowd C, Chan K, Zangwill LM, Goldbaum MH, Lee TW, Sejnowski TJ & Weinreb RN (2002): Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc. *Invest Ophthalmol Vis Sci* **43**: 3444–3454.

Bowd C, Medeiros FA, Zhang Z et al. (2005): Relevance vector machine and support vector machine classifier analysis of scanning laser polarimetry retinal nerve fiber layer measurements. *Invest Ophthalmol Vis Sci* **46**: 1322–1329.

Brigatti L, Hoffman D & Caprioli J (1996): Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol* **121**: 511–521.

Budenz DL, Chang RT, Huang X, Knighton RW & Tielsch JM (2005a): Reproducibility of retinal nerve fiber thickness measurements using the stratus OCT in normal and glaucomatous eyes. *Invest Ophthalmol Vis Sci* **46**: 2440–2443.

Budenz DL, Michael A, Chang RT, McSoley J & Katz J (2005b): Sensitivity and specificity of the StratusOCT for perimetric glaucoma. *Ophthalmology* **112**: 3–9.

Burgansky-Eliash Z, Wollstein G, Chu T, Ramsey JD, Glymour C, Noecker RJ, Ishikawa H & Schuman JS (2005): Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. *Invest Ophthalmol Vis Sci* **46**: 4147–4152.

Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN & Sejnowski TJ (2002): Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng* **49**: 963–974.

Chang R & Budenz DL (2008): New developments in optical coherence tomography for glaucoma. *Curr Opin Ophthalmol* **19**: 127–135.

Chang CC & Lin CJ (2001): LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Chen HY, Huang ML & Hung PT (2006): Logistic regression analysis for glaucoma diagnosis using Stratus Optical Coherence Tomography. *Optom Vis Sci* **83**: 527–534.

DeLong ER, DeLong DM & Clarke-Pearson DL (1988): Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**: 837–845.

Essock EA, Sinai MJ, Bowd C, Zangwill LM & Weinreb RN (2003): Fourier analysis of optical coherence tomography and scanning laser polarimetry retinal nerve fiber layer measurements in the diagnosis of glaucoma. *Arch Ophthalmol* **121**: 1238–1245.

Fan RE, Chen PH & Lin CJ (2005): Working set selection using second order information for training SVM. *J Mach Learn Res* **6**: 1889–1918.

Goldbaum MH, Sample PA, Chan K et al. (2002): Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Invest Ophthalmol Vis Sci* **43**: 162–169.

Harwerth RS, Carter-Dawson L, Shen F, Smith EL 3rd & Crawford ML (1999): Ganglion cell losses underlying visual field defects from experimental glaucoma. *Invest Ophthalmol Vis Sci* **40**: 2242–2250.

Haykin SS (1999): *Neural networks: a comprehensive foundation*. Upper Saddle River: N.J. Prentice Hall Pages 2–4, 234–245, 251, 353–355.

Hougaard JL, Heijl A & Krogh E (2004): The nerve fibre layer symmetry test: computerized evaluation of human retinal nerve fibre layer thickness as measured by optical coherence tomography. *Acta Ophthalmol Scand* **82**: 410–418.

Hougaard JL, Ostefeld C, Heijl A & Bengtsson B (2006): Modelling the normal retinal nerve fibre layer thickness as measured by Stratus optical coherence tomography. *Graefes Arch Clin Exp Ophthalmol* **244**: 1607–1614.

Hougaard JL, Heijl A & Bengtsson B (2007): Glaucoma detection by Stratus OCT. *J Glaucoma* **16**: 302–306.

Huang ML & Chen HY (2005): Development and comparison of automated classifiers for glaucoma diagnosis using Stratus optical coherence tomography. *Invest Ophthalmol Vis Sci* **46**: 4121–4129.

Imbault F & Lebart K (2004): A stochastic optimization approach for parameter tuning of support vector machines. *Pattern Recognition, ICPR, Proceedings of the 17th International Conference* **4**: 597–600.

Kerrigan-Baumrind LA, Quigley HA, Pease ME, Kerrigan DF & Mitchell RS (2000): Number of ganglion cells in glaucoma eyes compared with threshold visual field tests in the same persons. *Invest Ophthalmol Vis Sci* **41**: 741–748.

Leung CK, Chong KK, Chan WM et al. (2005): Comparative study of retinal nerve fiber layer measurement by StratusOCT and GDx VCC, II: structure/function

- regression analysis in glaucoma. *Invest Ophthalmol Vis Sci* **46**: 3702–3711.
- Manassakorn A, Nouri-Mahdavi K & Caprioli J (2006): Comparison of retinal nerve fiber layer thickness and optic disk algorithms with optical coherence tomography to detect glaucoma. *Am J Ophthalmol* **141**: 105–115.
- Medeiros FA, Zangwill LM, Bowd C & Weinreb RN (2004): Comparison of the GDx VCC scanning laser polarimeter, HRT II confocal scanning laser ophthalmoscope, and stratus OCT optical coherence tomograph for the detection of glaucoma. *Arch Ophthalmol* **122**: 827–837.
- Medeiros FA, Zangwill LM, Bowd C, Vesani RM, Susanna R Jr & Weinreb RN (2005): Evaluation of retinal nerve fiber layer, optic nerve head, and macular thickness measurements for glaucoma detection using optical coherence tomography. *Am J Ophthalmol* **139**: 44–55.
- Moller MF (1993): A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks* **6**: 525–533.
- Naithani P, Sihota R, Sony P, Dada T, Gupta V, Kondal D & Pandey RM (2007): Evaluation of optical coherence tomography and heidelberg retinal tomography parameters in detecting early and moderate glaucoma. *Invest Ophthalmol Vis Sci* **48**: 3138–3145.
- Paunescu LA, Schuman JS, Price LL, Stark PC, Beaton S, Ishikawa H, Wollstein G & Fujimoto JG (2004): Reproducibility of nerve fiber thickness, macular thickness, and optic nerve head measurements using StratusOCT. *Invest Ophthalmol Vis Sci* **45**: 1716–1724.
- Poinosawmy D, Tan JC, Bunce C & Hitchings RA (2001): The ability of the GDx nerve fibre analyser neural network to diagnose glaucoma. *Graefes Arch Clin Exp Ophthalmol* **239**: 122–127.
- Quigley HA, Katz J, Derick RJ, Gilbert D & Sommer A (1992): An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. *Ophthalmology* **99**: 19–28.
- Sample PA, Boden C, Zhang Z et al. (2005): Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields. *Invest Ophthalmol Vis Sci* **46**: 3684–3692.
- Schuman JS, Pedut-Kloizman T, Hertzmark E et al. (1996): Reproducibility of optic fiber layer thickness measurements using optical coherence tomography. *Ophthalmology* **103**: 1889–1898.
- Shah NN, Bowd C, Medeiros FA, Weinreb RN, Sample PA, Hoffmann EM & Zangwill LM (2006): Combining structural and functional testing for detection of glaucoma. *Ophthalmology* **113**: 1593–1602.
- Sihota R, Sony P, Gupta V, Dada T & Singh R (2006): Diagnostic capability of optical coherence tomography in evaluating the degree of glaucomatous retinal nerve fiber damage. *Invest Ophthalmol Vis Sci* **47**: 2006–2010.
- Sommer A, Katz J, Quigley HA, Miller NR, Robin AL, Richter RC & Witt KA (1991): Clinically detectable nerve fiber atrophy precedes the onset of glaucomatous field loss. *Arch Ophthalmol* **109**: 77–83.
- Tucker A, Vinciotti V, Liu X & Garway-Heath D (2005): A spatio-temporal Bayesian network classifier for understanding visual field deterioration. *Artif Intell Med* **34**: 163–177.
- Tuulonen A, Lehtola J & Airaksinen PJ (1993): Nerve fiber layer defects with normal visual fields. Do normal optic disc and normal visual field indicate absence of glaucomatous abnormality?. *Ophthalmology* **100**: 587–597; Discussion 597–8.
- Wollstein G, Ishikawa H, Wang J, Beaton SA & Schuman JS (2005): Comparison of three optical coherence tomography scanning areas for detection of glaucomatous damage. *Am J Ophthalmol* **139**: 39–43.
- Zangwill LM, Chan K, Bowd C, Hao J, Lee TW, Weinreb RN, Sejnowski TJ & Goldbaum MH (2004): Heidelberg retina tomograph measurements of the optic disc and parapapillary retina for detecting glaucoma analyzed by machine learning classifiers. *Invest Ophthalmol Vis Sci* **45**: 3144–3151.
- Zhang Z & Zha H (2004): Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM Journal of Scientific Computing* **26**: 313–338.

Received on January 26th, 2009.  
Accepted on September 20th, 2009.

*Correspondence:*  
Dimitrios Bizios  
Department of Clinical Sciences  
Ophthalmology  
Malmö University Hospital  
Lund University  
SE-205 02 Malmö  
Sweden  
Tel: + 46 40 337071  
Fax: + 46 40 336212  
Email: dimitrios.bizios@med.lu.se

## Appendix

### Nonlinear dimensionality reduction based on manifold learning

OCT A-scan data can be characterized as multidimensional, because they require multiple parameters (dimensions) to be represented. One approach to simplify the representation of data is to use techniques that are able to map the data into a set of fewer parameters (i.e. a lower dimensional space). Nonlinear dimensionality reduction techniques based on manifold learning, function by constructing parameters with values that are assumed to be measurement points embedded in a specific type of topological space (a nonlinear manifold), existing within the high-dimensional space of the complete dataset. In contrast to linear methods, such as principal component analysis, these

techniques retain information on nonlinear relationships between measurement points and are able to represent nonlinear structures in the data. The local tangent space alignment (LTSA) algorithm that is used in this study is considered a local embedding technique, and functions by constructing an approximation for the tangent space at each data point, and aligning these tangent spaces to provide the global coordinates of the data points. The choice for the number of transformed parameters that describes the A-scan data is based on estimations of the intrinsic dimensionality of the dataset.

### Machine learning classifiers – artificial neural networks & support vector machines

Machine learning is a field of artificial intelligence concerned with the

development of computational and statistical methods that are able to learn through a training process. ANNs comprise a group of machine learning methods used for both classification and function approximation tasks. They have been successfully used in a variety of fields, including medicine, for data mining, automated interpretation of tests and feature extraction from large datasets. Multi-layer perceptrons (MLPs), consisting of multiple layers of artificial neurons, are a type of ANN developed from the original perceptron model (the first type of ANN, created in the 1940s). The widespread use of MLPs is partly because of their ability to conduct nonlinear classification tasks and because of their efficient supervised learning algorithm based on back-propagation of error. Following developments in statistical learning theory, new techniques for classification and

regression were introduced within the group of machine learning algorithms. SVMs are kernel-based methods that, like ANNs, can be trained to recognize patterns in data and adapt their decision boundary to the training data. Unlike ANNs, these algorithms, perform classification by using kernels to map the input data in a space of higher dimensionality and, with the help of constructed support vectors (from part of the training data), they create hyperplanes that maximize the separation between the classes while minimizing the generalization error.

#### **Artificial neural network ensembles**

One important aspect of trained ANNs is their ability to generalize their adaptively created decision rules

on previously unseen data. It has been shown that the generalization error of ANNs can be decomposed into the factors of bias (i.e. a measure of the classification accuracy on the training data) and variance (i.e. a measure of the stability of the given classification solution provided by the ANN, depending on the variability of the training data). The two factors of bias and variance have an inverse relationship and there exist a trade-off between these two. To decrease the generalization error of a classifier, one can combine the prediction of a number of classifiers in an ensemble structure. It can be shown that the generalization error of the ensemble equals the averaged generalization error made by an individual ANNs *minus* the averaged variance (a.k.a.

diversity) of the individual ANNs in the ensemble.

#### **Training with bagging**

One approach to increase the diversity of individual ANNs in an ensemble (and thus decrease the ensemble generalization error) is to train the networks on slightly different subsets of the training data. This can be accomplished with a resampling algorithm such as Bootstrap aggregating, also called bagging. The bagging algorithm generates training subsets by uniformly sampling examples from the training data with replacement. The created bootstrap samples (expected to contain 63.2% of unique examples) are then used for training by the individual ANNs.



# Paper IV





RESEARCH ARTICLE

Open Access

# Integration and fusion of standard automated perimetry and optical coherence tomography data for improved automated glaucoma diagnostics

Dimitrios Bizios\*, Anders Heijl and Boel Bengtsson

## Abstract

**Background:** The performance of glaucoma diagnostic systems could be conceivably improved by the integration of functional and structural test measurements that provide relevant and complementary information for reaching a diagnosis. The purpose of this study was to investigate the performance of data fusion methods and techniques for simple combination of Standard Automated Perimetry (SAP) and Optical Coherence Tomography (OCT) data for the diagnosis of glaucoma using Artificial Neural Networks (ANNs).

**Methods:** Humphrey 24-2 SITA standard SAP and StratusOCT tests were prospectively collected from a randomly selected population of 125 healthy persons and 135 patients with glaucomatous optic nerve heads and used as input for the ANNs. We tested commercially available standard parameters as well as novel ones (fused OCT and SAP data) that exploit the spatial relationship between visual field areas and sectors of the OCT peripapillary scan circle. We evaluated the performance of these SAP and OCT derived parameters both separately and in combination.

**Results:** The diagnostic accuracy from a combination of fused SAP and OCT data (95.39%) was higher than that of the best conventional parameters of either instrument, i.e. SAP Glaucoma Hemifield Test ( $p < 0.001$ ) and OCT Retinal Nerve Fiber Layer Thickness  $\geq 1$  quadrant ( $p = 0.031$ ). Fused OCT and combined fused OCT and SAP data provided similar Area under the Receiver Operating Characteristic Curve (AROC) values of 0.978 that were significantly larger ( $p = 0.047$ ) compared to ANNs using SAP parameters alone (AROC = 0.945). On the other hand, ANNs based on the OCT parameters (AROC = 0.970) did not perform significantly worse than the ANNs based on the fused or combined forms of input data. The use of fused input increased the number of tests that were correctly classified by both SAP and OCT based ANNs.

**Conclusions:** Compared to the use of SAP parameters, input from the combination of fused OCT and SAP parameters, and from fused OCT data, significantly increased the performance of ANNs. Integrating parameters by including a priori relevant information through data fusion may improve ANN classification accuracy compared to currently available methods.

## Background

Glaucoma is an optic neuropathy resulting in characteristic visual field defects. Investigating the relationship between development of functional damage in the visual field and structural glaucomatous changes of the retinal nerve fiber layer (RNFL) has been the purpose of numerous studies [1-5].

Diagnostic instruments providing quantitative analyses in glaucoma assess either functional or structural aspects of the disease. Imaging and quantitative analysis of RNFL measurements can be accomplished with Optical Coherence Tomography (OCT). OCT is a noninvasive interferometric technique that provides cross sectional images and thickness measurements of the RNFL (RNFLT) with high resolution [6] and good reproducibility [7-9]. Standard Automated white-on-white Perimetry (SAP) is the standard for examining the visual field. Perimetric tests are able to provide quantitative measurements of differential

\* Correspondence: [dimitrios.bizios@med.lu.se](mailto:dimitrios.bizios@med.lu.se)  
Department of Clinical Sciences Malmö, Ophthalmology, Skåne University Hospital, Lund University, SE-205 02 Malmö, Sweden

light sensitivity at many test point locations in the visual field, and commercially available statistical analysis packages help clinicians in identifying significant visual field loss [10,11]. The diagnostic performance of both OCT and SAP in glaucoma as well as the correlation between SAP and OCT measurements has been investigated [12-15].

It is conceivable that integration of functional and structural test measurements could provide more relevant information and thus improved diagnostic performance for classification systems when used as input data. The relevance of integrated diagnostic information is dependent on the underlying relationship between structural and functional measurements. Statistical approaches such as the linear model constructed by Hood et al related RNFLT values to sensitivity losses in SAP [16]. Other studies trying to map the individual visual field test points in SAP to areas of the peripapillary RNFL through different models, showed moderate correlations between visual field sensitivity values and structural measurements [17,18]. More recent attempts to model the function - structure relationship in glaucoma demonstrated that machine learning algorithms, such as radial basis function artificial neural networks (ANNs), improved the modelling accuracy compared to linear methods [19].

The use of machine learning classifiers (MLCs) in glaucoma diagnosis using either functional or structural measures has been previously explored [20]. MLCs like ANNs have been used for classification of tests based on structural or functional measurements [21-28] and for detection of glaucoma progression [29,30]. ANN-based classification demonstrated better accuracy than linear methods [23,24,31] and performed at least as well as human experts [32].

Recent attempts to provide a combined evaluation of structural and functional tests showed promising results [33,34], though few studies have examined the diagnostic performance of combining functional and structural data with MLCs for glaucoma diagnosis [35,36]. One of the main advantages of MLCs is their ability to learn a classification task by training on given examples. Such adaptive classification based on the available data is useful, since a complete analytic theory of the structure-function relationship in glaucoma does not yet exist. The performance of MLCs can be influenced by a number of factors including data selection bias, choice of input and classifier architecture.

The purpose of this study was to investigate whether the integration of information from SAP and OCT data could improve the accuracy of glaucoma diagnosis, by using the data as input in ANN based classifiers. We evaluated the performance of simple combination of OCT and SAP data as well as novel approaches based on data fusion by utilizing a priori knowledge about the

physiologic relationship between the RNFL and visual function in glaucoma.

## Methods

This study is based on analysis of prospectively collected data from randomly selected healthy individuals from a defined catchment area and glaucoma patients followed at the Department of Ophthalmology at Skåne University Hospital, Malmö Sweden. The study was conducted according to the tenets of the Declaration of Helsinki and was approved by the Regional Ethical Review Board of Lund, Sweden. All healthy individuals and clinical glaucoma patients included in the study provided informed consent prior to any examinations.

## Healthy Individuals

We performed a random selection from a population register containing 4,718 persons over 50 years, living in two primary care catchment areas of Scania, Sweden. This selection yielded a sample of 307 individuals who were invited to participate in the study. Of those, 170 individuals accepted the invitation and underwent a comprehensive ophthalmic examination.

## Clinical Glaucoma Patients

We randomly selected 397 patients with a diagnosis of primary open angle glaucoma, normal tension glaucoma or pseudo-exfoliation glaucoma, from a register of 2,174 visits of patients having these diagnoses, followed at the Department of Ophthalmology, Malmö University Hospital, Sweden between January 2<sup>nd</sup> 2007 and March 13<sup>th</sup> 2008. After review of the 397 patient medical records we excluded patients with any history or additional diagnoses of ocular or systemic pathology affecting the visual field or the RNFL except glaucoma (e.g. neurological disorders or retinal disease). Our reference for the diagnosis of glaucoma was based on optic nerve head (ONH) topography and/or examination of available ONH photographs. After application of our exclusion criteria, 164 patients that fulfilled our diagnostic reference for glaucoma were invited to participate and underwent an ophthalmic examination.

## Examinations

Upon examination a detailed medical history was taken, including current medical conditions and treatments. Individuals with systemic or non-glaucomatous ocular diseases that could affect the ONH, RNFL and visual field were excluded. Persons with lens opacities or intraocular lenses were not excluded from the study. One randomly selected eye from each eligible healthy individual was chosen for inclusion in the study. For glaucoma patients only the affected eye was included in the study. In patients with bilateral disease, the eye with the best Mean Deviation value (i.e. the less negative

value denoting milder glaucomatous damage) on the most recent SAP examination was chosen.

The clinical ophthalmic examination consisted of the following parts:

1. Visual acuity was measured using an autorefractor (Humphrey model 595 - Carl Zeiss Meditec, Dublin, CA, USA). Manual refraction was performed when the autorefractor-measured visual acuity values were  $< 0.8$ . All participants were required to have visual acuity  $\geq 0.5$  and refractive error  $\geq 5$  dioptres (D) sphere and  $< 3$  D cylinder in order to be included in our subsequent analyses.

2. Intraocular pressure was measured by a Goldmann applanation tonometer.

3. Fundoscopy was performed by a trained clinician with a slit-lamp biomicroscope after the use of mydriatic agents (tropicamide 0.5% and phenylephrine hydrochloride 2.5%).

All examined individuals underwent a battery of functional and imaging tests, including:

1. Standard Automated Perimetry (SAP) with the Humphrey Field Analyzer (Carl Zeiss Meditec, Dublin, CA, USA) using the 24 - 2 SITA Standard program. Healthy participants underwent a second SAP examination of the study eye and the results of the second test were subsequently chosen. Perimetric tests were required to have reliable fixation as assessed by the perimetrist and  $< 15\%$  of false positive answers to be included.

2. Time domain OCT examination after pupil dilation with Stratus OCT (Carl Zeiss Meditec, Dublin, CA, USA) using the Fast RNFL thickness protocol, which derives the RNFL values by averaging three 3.4 mm circumpapillary scans, each with 256 measurement points (A-scans). All included OCT tests were required to be of good quality as defined by the manufacturer specifications (signal strength  $> 5$ ) and free of obvious artifacts from incorrect delineation of the RNFL by the instruments segmentation algorithm. The same experienced ophthalmic photographer performed all OCT examinations.

Inclusion of tests from healthy individuals and glaucoma patients in further analyses was based on evaluation of the optic disc during fundus examination. Among healthy individuals only subjects having a normal appearance of the optic disc were included. For inclusion of tests from the glaucoma patients, the eligibility criteria of a glaucomatous optic disc described on their records and/or present in previous optic disc photographs, had to be confirmed during fundus examination.

#### Structural and functional test Parameters

Our analyses in this study are based on the following OCT and SAP parameters:

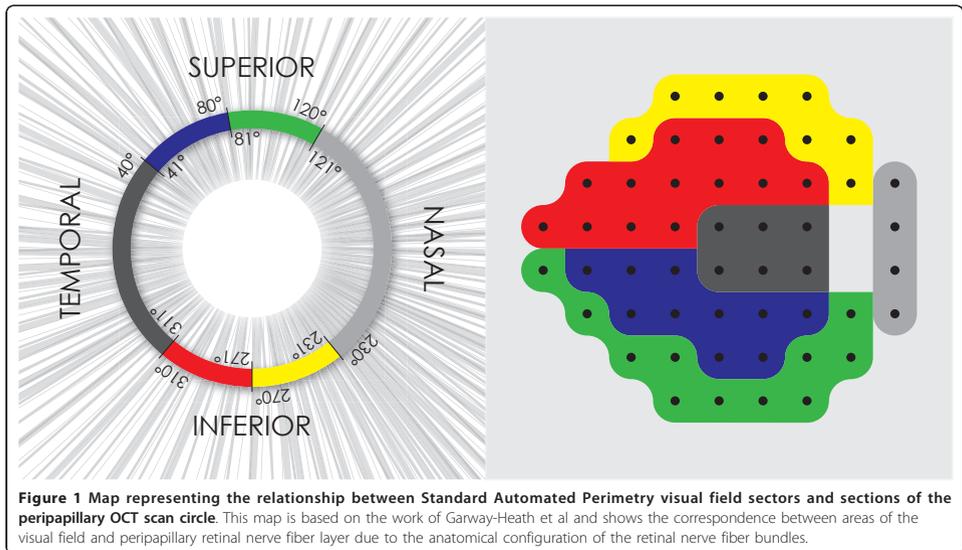
##### OCT RNFL parameters

- **RNFL standard parameters** The StratusOCT RNFL analysis printout provides average thickness

measurements for the whole scan circle, the four quadrants and the 12 clock hour sectors of each scan, while highlighting the values that fall below the 5% and 1% significance level, based on comparison to the instrument's normative database. Diagnostic accuracy of the best performing from these parameters was compared to ANN classification performance.

- **A-scan measurements and PCA processing** We used the 256 averaged A-scan values of the 3 peripapillary scan circles for each OCT test, after we decreased their complexity by means of principal component analysis (PCA). We adjusted PCA to maintain 99.9% of the variation in the data. This was achieved by the first 22 principal components, which were then used as input to the ANN classifiers. All OCT RNFLT data were corrected for age and refractive status (spherical equivalent) based on a separate normative database [37]. We have previously treated the use of A-scan derived parameters as input in automated classifiers [25].

- **Fused OCT parameters** The fused OCT parameters were derived by weighting the OCT A-scan measurements of each test with the corresponding scored pattern deviation (PD) values from SAP. In the fusion process for the OCT data we used the map constructed by Garway-Heath et al [38] to represent the relationship between RNFLT of OCT scan circle sectors and differential light sensitivity in specific areas of the visual field, and divided the OCT scan circle and the SITA standard 24-2 SAP test points into 6 sectors accordingly (Figure 1). In each OCT sector and for every A-scan position, we calculated the distribution of RNFLT values based in a separate normative database described elsewhere [37]. The probability values for each age-and refraction corrected OCT A-scan measurement of our test dataset, were then calculated. A-scan values falling below the fifth percentile of the distribution in our normal reference material were transformed through multiplication with an exponential factor. This factor was constructed by calculating the average pattern deviation probability scores (i.e. the sum of all pattern deviation scores divided by the number of SAP test points) of the visual field sector corresponding to each of the OCT scan circle sectors. The fused A-scan values depended on the decrease in RNFLT for the specific A-scan position relative to the distribution of the normative reference material, and the status of the visual field sector corresponding to that location. PCA was subsequently applied on the fused OCT A-scan values. In order to simplify the comparison between non-fused and fused OCT data we included the principal components that retained the same level of variation in the data (99.9%) as in the processing of the previously described non-fused A-scan measurements. In this way PCA provided 38 principal components that were then used as input to the ANN classifiers.

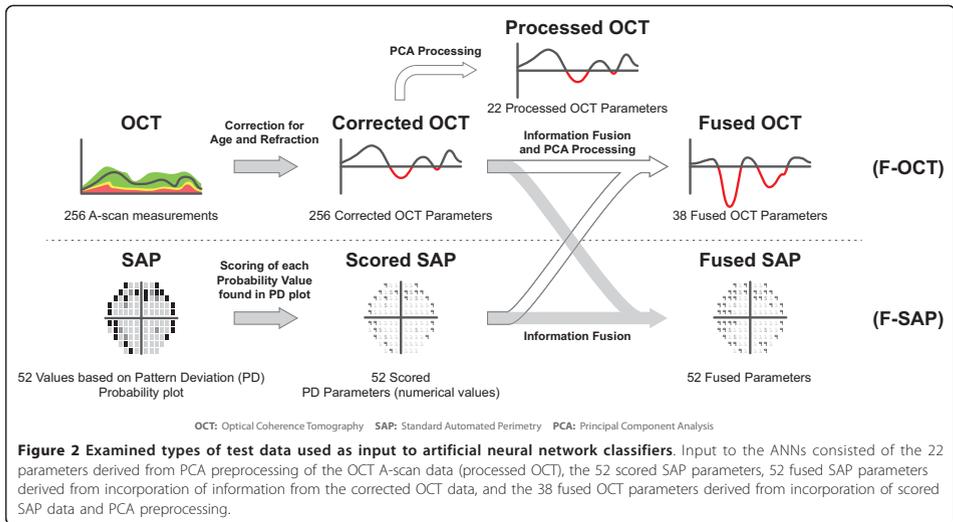


#### SAP test parameters

- **Glaucoma Hemifield Test (GHT)** The GHT index is available in the standard analysis printout of SAP tests. It is an expert system that classifies SAP tests as within normal limits, borderline or outside normal limits, based on the differences of PD values between test points in mirror image areas of the upper and lower hemispheres of the visual field. We measured the specificity and sensitivity of GHT and compared it to the ANN classifiers.
- **Pattern deviation probability scores from each of the 52 SAP test point locations (52 parameters)** For each SAP test point, we provided numerical values to the pattern deviation probability map values using a probability scoring scale identical to that used in calculating the GHT [39]. We have previously demonstrated the performance benefits of using pattern deviation probability scores as input to ANNs [21].
- **Fused SAP parameters (52 parameters)** Fused SAP parameters were derived by weighting each SAP PD scored value with the corresponding OCT A-scan measurements. In the fusion process for SAP data the OCT scan circle and the SITA standard 24-2 SAP test points were divided into six sectors based on the map by Garway-Heath et al [38] (Figure 1). For every visual field sector, the pattern deviation probability score at each test point was transformed by an additive factor. This factor was derived from the age- and refraction-corrected A-scan measurements of the corresponding OCT sector. All A-scan measurements were identified in each OCT

sector, and their probability values were assigned a score according to significance level of the deviation from the values of our separate normative database. The probability scoring scale was similar with that used in the calculation of the GHT [39]. The lowest scored probability below the fifth percentile or the highest scored probability above the ninetyfifth percentile of our normal RNFLT distribution from each OCT sector was used as the factor in the fusion process. The fused SAP parameters were obtained by adding this factor to the SAP pattern deviation probability score of each SAP test point in the corresponding visual field sector. In the event that both high and low scored probability values (outside the fifth or ninetyfifth percentile of our normative RNFLT database) existed in the same OCT sector, only the low value was used in the summation process. The fused SAP measurements thus depended on both the status of the visual field sector reflected by the pattern deviation probability scores, and the thickness of the corresponding OCT sector. Visual field defects as indicated by the pattern deviation probability scores could be either accentuated or attenuated during the fusion procedure, depending on the factor of scored probability from the corresponding OCT sector.

The ANNs were trained and tested using the perimetric pattern deviation probability scores and the age- and refraction corrected OCT A-scan measurements after PCA preprocessing, as well as the fused SAP and OCT parameters (Figure 2). We also evaluated the



integration of the above parameters by simply combining them. We thus tested:

- Integration of the non-fused SAP and OCT parameters (74 parameters), and
- Integration of the fused OCT and SAP parameters (90 parameters)

#### Artificial Neural Networks (ANN)

Our ANN classifier consisted of an ensemble of thirty-five fully connected cascade-forward Multi Layer Perceptrons (MLPs). The number of neurons in the input layer of each cascade-forward MLP was equal to the number of parameters used as input data. All MLPs consisted of 2 hidden layers with tangent hyperbolic transfer functions and an output layer of one neuron with a logistic transfer function that provided the MLP output. The number of neurons in the hidden layers was chosen based on the type of input used in order to achieve the best performance as judged by the results derived from the 10-fold cross validation procedure. Our ANNs were constructed with the MATLAB neural network toolbox version 7 (The MathWorks Inc, Natick, MA, USA) and trained with the scaled conjugate gradient algorithm described by Möller [40].

#### Training the Artificial Neural Networks

ANNs were trained and tested with the 10-fold cross-validation procedure, to reduce bias from training and testing on the same individuals, while fully utilizing our

data set. Data were randomly divided into ten subsets, each containing test data from an approximately equal proportion of glaucoma patients and healthy individuals. One subset was used to test classification performance while the remaining nine subsets were used for training purposes. In our ANN ensemble, one out of the nine training subsets was reserved for early stopping of the ANNs in order to avoid overfitting. We additionally used bagging [41] of the remaining eight subsets to create the training sets used by the ANN ensemble. During training, this process was iterated, each time using a different subset as the early stopping set, until all the data subsets had been used to both train and stop the training of the ensemble. We further iterated the training process using each time a different test subset, so that all data could be used both for training and testing the classifiers, and averaged the test results in order to produce a single performance measure for each ANN.

#### Analyses

To measure the classification performance of the ANN ensemble and the diagnostic ability of the compared parameters, we calculated the area under the ROC curve (AOC). The cut off values for all ANNs were calculated based on the best performing specificity-sensitivity pairs (i.e. pairs that provided the largest area under ROC when their values were multiplied) from the 10-fold cross-validation procedure. Significance testing between the AOCs was conducted with DeLong's non-parametric method [42]. We evaluated the agreement in classification between the OCT and SAP based ANNs

by calculating odds ratios, which in this case signified the odds that tests classified by the SAP-based ANN receive the same classification by the OCT-based ANN, based on the same classification threshold of 0.5 for both OCT and SAP based classifiers.

The Chi square test was used to find significant differences in the distribution of gender between the healthy individuals and patients with glaucoma, whereas the Mann-Whitney test was used for the continuous variables age, visual acuity and refractive error. Diagnostic accuracy of the SAP and OCT parameters was compared using the McNemar test for correlated proportions.

## Results

Thirty-four healthy individuals and 36 glaucoma patients were excluded because of ophthalmic, neurological and metabolic disorders affecting the visual field and/or the retina, refractive errors and visual acuity outside the defined range for inclusion, erroneous estimation of the RNFL by the OCT segmentation algorithm, and due to inability to complete the examination. Only three healthy persons were excluded based on optic disc criteria (one with optic disc drusen, one with an optic disc hemorrhage and one with a peripapillary membrane). One patient with glaucoma but normal ONH appearance was also excluded. After application of all exclusion and inclusion criteria, OCT and SAP data from 125 healthy individuals and 135 patients with glaucoma were used in our analyses. Eight of the 135 glaucoma patients were initially part of the healthy population invited to participate in the study, but were diagnosed with glaucoma during the clinical examination and were included in the glaucoma group. The diagnosis for these 8 patients was based on a second clinical examination that included imaging of the ONH and RNFL, SAP testing and evaluation of the fundus and ONH. All OCT tests were of good quality, with mean ( $\pm$ SD) signal strength of 9.95 ( $\pm$ 0.23) and 9.36 ( $\pm$ 0.97) for the healthy and glaucoma groups

respectively. Based on the Mean Deviation of the SAP visual fields, the glaucoma group consisted of 49 patients (ca 36%) with early, 32 patients (ca 24%) with moderate and 54 patients (ca 40%) with advanced glaucomatous visual field loss. The demographic characteristics of the healthy subjects and glaucoma patients can be seen in Table 1. The significantly lower visual acuity of the glaucoma group could be attributed to the higher incidence of lens opacities in this group.

The integration of fused data offered higher diagnostic accuracy compared to the best performing SAP and OCT algorithms that exist in the available analysis packages of each instrument. For SAP, the GHT (with borderline results signifying glaucoma) provided an accuracy of 86.92%. For OCT, RNFLT abnormally depressed in at least one quadrant at the 5% significance level, had an accuracy of 91.54%. The combination of non-fused data led to a diagnostic accuracy of 93.85%, significantly better than GHT (McNemars test:  $p = 0.006$ ), whereas the accuracy of combined fused data was 95.39%, higher than the accuracy of both the GHT (McNemars test:  $p < 0.0001$ ) and OCT RNFLT algorithm (McNemars test:  $p = 0.031$ ).

The two ANNs with input based on the fused OCT and the combined fused OCT and SAP data respectively provided almost identical AROC values of 0.978, performing significantly better than the ANN based on the SAP measurements alone. Utilizing input based on the combined non-fused OCT and SAP measurements did not lead to similar significant improvements. The AROCs of ANNs based on the fused and non-fused parameters are shown in Figure 3. Significance testing between the compared parameters is shown in Table 2.

The ANN with input based only on the processed OCT A-scan data (AROC: 0.970) had larger AROC compared to the ANN based on PD probability scores (AROC: 0.945). Even though this difference was not statistically significant, the higher performance of the

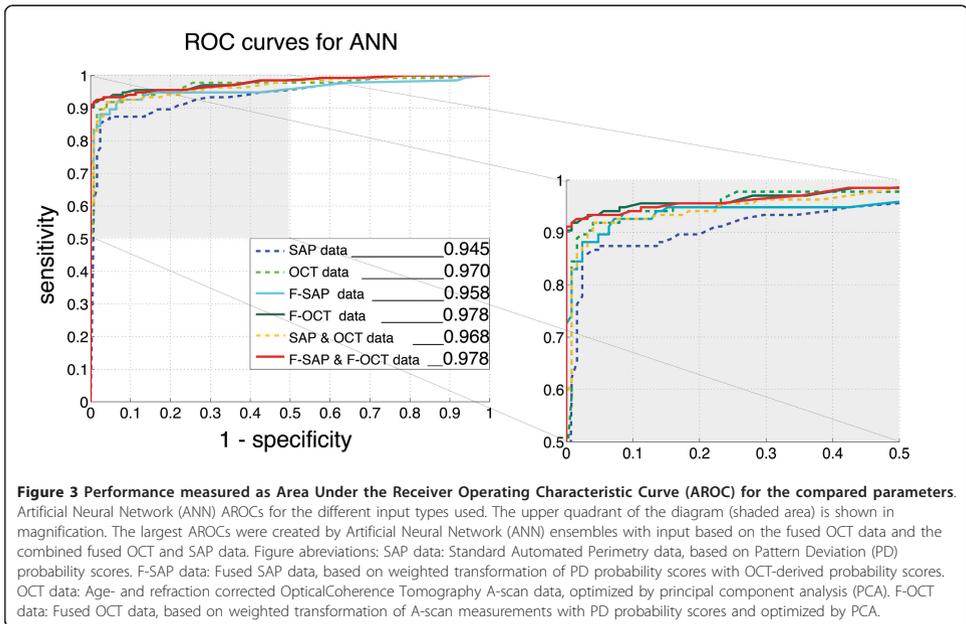
**Table 1 Demographic data of included healthy subjects and patients with glaucoma**

	Healthy (n = 125)	Glaucoma (n = 135)	p - value (Mann-Whitney test)
Gender (female/male)	66/59	79/56	NS*( $\chi^2$ test)
Age (years)	64.65 $\pm$ 8.11	73.36 $\pm$ 7.81	< 0.0001
Visual Acuity (decimal scale)	1.00 $\pm$ 0.15	0.86 $\pm$ 0.19	< 0.0001
Refractive error (spherical equivalent)	+0.53 $\pm$ 1.74	-0.15 $\pm$ 1.82	0.0015
Visual Field (MD)†	-0.66 $\pm$ 1.77	-11.04 $\pm$ 8.21	< 0.0001

All values except for gender are represented as mean  $\pm$  standard deviation.

\* NS: non significant

† MD: mean deviation value in decibel, as measured by 24-2 SITA Standard program



OCT-based ANN prevented any differences from reaching statistical significance when comparing the OCT-based ANN to ANNs trained on the integrated or fused input data. At high specificities fused OCT parameters provided the highest sensitivity values (Figure 3)

The agreement in classification (reflected by the odds ratios) between ANNs based on SAP and OCT measurements improved when using the fused parameters as input. The improved agreement led to a larger number of individuals correctly classified by both function- and structure-based ANNs (Figure 4). Examples on such

classification improvements are given in Figure 5. The misclassified tests belonging to glaucoma patients, did not exhibit discernable visual field or RNFLT defects in neither the SAP nor the OCT tests.

### Discussion

We evaluated the effect of combining SAP and OCT measurements on the ability of ANN classifiers to discriminate between normal and glaucomatous tests. We have previously demonstrated that the use of pre-processed RNFLT measurements based on A-scans improved the

**Table 2 Performance Comparison between Artificial Neural Networks based on fused, combined and single types of data**

	F-SAP data (AROC:0.958)	F-OCT data (AROC:0.978)	SAP & OCT Data (AROC:0.968)	F-SAP & F-OCT data (AROC:0.978)
SAP data (AROC: 0.945)	0.502	<b>0.047</b>	0.147	<b>0.047</b>
OCT data (AROC: 0.970)	0.431	0.576	0.879	0.562

Significance (p) values of Area under Receiver Operating Characteristic (AROC) curves were calculated by DeLong's non-parametric method.

SAP data: Standard Automated Perimetry data, based on Pattern Deviation (PD) probability scores

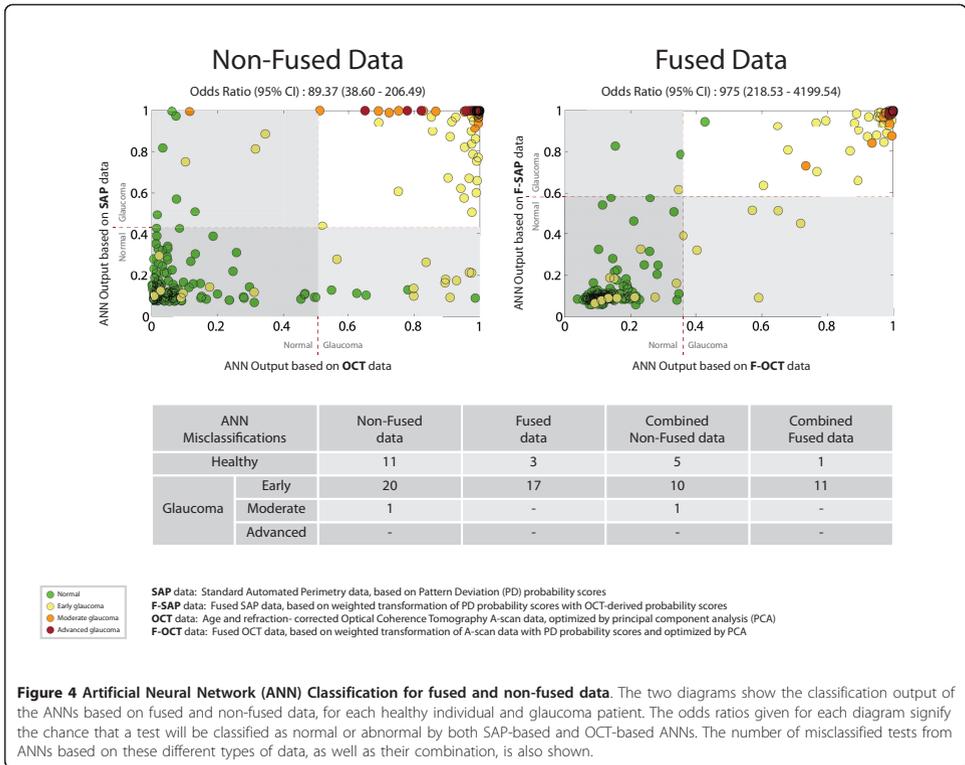
F-SAP data: Fused SAP data, based on weighted transformation of PD probability scores with OCT-derived probability scores

OCT data: Age- and refraction corrected Optical Coherence Tomography A-scan data, optimized by principal component analysis (PCA)

F-OCT data: Fused OCT data, based on weighted transformation of A-scan measurements

with PD probability scores and optimized by PCA

Bold indicates statistical significance (i.e.  $p < 0.05$ )

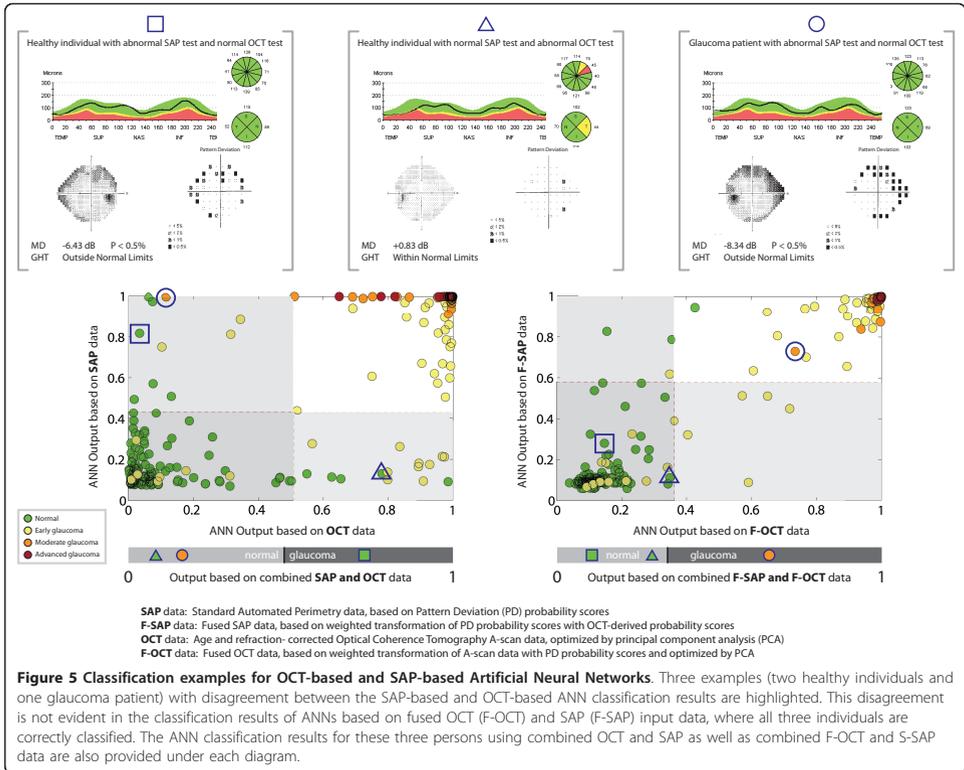


diagnostic performance of MLCs compared to the conventional RNFLT parameters presented by the instrument [25]. For SAP, the Pattern Deviation probability plots and maps provide probability values of all test points, highlighting those points with values falling outside the age corrected normal limits and also account for effects of media opacities on light sensitivity across the visual field. The performance benefits of pattern deviation score - based input data have been shown [21].

The combination of structural and functional information contained in the OCT and SAP test data respectively, can be viewed as a type of information integration. The simplest way to integrate the different types of data is to construct a vector that consists of all OCT and SAP measurements. We additionally attempted to construct and evaluate the performance of novel input parameters that fuse both structural and functional measurements. Integrating information about the structure-function relationship of glaucomatous damage through data fusion, presents some advantages over the simple combination of

the two different types of data. Instead of relying on MLCs to learn about the structure-function relationship based on limited training data, the fusion process allows for direct incorporation of prior knowledge obtained in other independent large datasets about the topographic relationship between structural and functional measurements into the classification problem. Controlling the incorporation of knowledge into MLCs can also counteract the lack of insight on the way stochastic processes like ANNs represent and use the acquired knowledge in their classification decisions. Our ANNs with input based on the novel parameters showed a high degree of agreement in their classification decisions, reflected on the presented odds ratio values (Figure 4). The higher odds ratios for the ANNs based on fused input data could indicate that these classifiers are more robust since the likelihood of a false positive or false negative test result by both fused OCT and SAP based ANNs was significantly lower.

Bowd et al has previously shown that MLCs trained on combinations of OCT and SAP derived input performed



at least as well as MLCs trained on each input type alone, while the use of data with reduced complexity (by means of the backward elimination technique), further improved MLC performance [35]. Our results did not show significant improvement using input that simply combined OCT and SAP measurements compared to when using SAP or OCT measurements separately. However, the combination of fused OCT and SAP parameters showed significant improvement compared to the use of ANNs based on SAP parameters alone, and to the best performing commercially available algorithms in both the SAP and Stratus OCT instruments. This improvement was not specific to our ANN, but could be also seen with another MLC, a relevance vector machine (RVM) classifier, that we constructed and tested for comparison purposes. We did not report the results of our RVM since its performance was very similar to that of our ANN.

The use of principal component analysis for dimensionality reduction of the OCT and fused OCT data

instead of a non-linear dimensionality reduction algorithm could have affected the results. Even though non-linear dimensionality reduction techniques might provide better representations of complex data, their extensions to new data are iterative in nature without exact numerical solutions in most cases.

The performance of Machine learning classifiers is dependent on their training process. During training, it is important to present learning examples with a known outcome (i.e. 'true' normal and 'true' glaucoma cases) and with all disease stages in order for the MLC to create representative classification decision boundaries. The inclusion of cases with an uncertain condition (i.e. patients characterized as glaucoma suspects) would adversely affect the false positive and negative rates of classification and our evaluation of specificity and sensitivity rates of the classifier.

The recruitment of healthy persons was based on a random population sample with the majority of individuals

having no previous experience in ophthalmic examinations. In our attempt to include healthy individuals that do not represent supernormal subjects, we did not exclude persons with cataract since it is a condition often seen in older population groups and in patients with glaucoma. The rates of misclassified tests could be partly explained by our choice of reference standard based on ONH morphology, which did not exclude patients with normal SAP and OCT test results. The bias in selecting a structure-or-function-related reference standard, affects the accuracy of combinatorial analyses by erroneous estimations of specificity, sensitivity and correlation measures of the examined structural and functional parameters. We did not base the definition of normality and glaucoma on either SAP or OCT test indices. Our choice of reference standard was instead based on clinical examination of ONH morphology. Even though this structure based reference standard relates more to RNFL morphology than function as measured by the visual field, it has not shown a high degree of correlation with OCT measurements [43]. The significant differences in age and refraction between healthy individuals and glaucoma patients are accounted for both in the pattern deviation probability based SAP input and the age- and refraction-corrected OCT input. Even though the 10-fold cross-validation process can account for certain bias pertaining to sample variability, further evaluation on an independent group of subjects is needed to support the general applicability of our findings. Future studies should also evaluate the fusion process with data based on the new generation of spectral domain OCT that provide higher spatial resolution and improved algorithms for detecting and analyzing the RNFL.

The incorporation of knowledge about known rules into black box classifiers could enable the construction of ANN-based systems that are more closely related to grey box models (i.e. models with known general structure but also unknown parameters), allowing for greater insight into the classification process and more effective sensitivity analyses of the test input parameters. Such advantages could facilitate the practical deployment of ANNs as decision support systems in glaucoma diagnostics.

## Conclusions

Our study showed that the combination and fusion of data from OCT and SAP has the potential to increase the accuracy of glaucoma diagnostics compared to parameters from either instrument alone. Moreover, fusion of test measurements could lead to test parameters that better reflect both the structural and functional glaucomatous changes that occur during the course of the disease, providing more relevant information to glaucoma diagnostic systems.

## Acknowledgements

The authors acknowledge the contributions of Sabina Andersson, MD (Dept of Ophthalmology, Skåne University Hospital, Lund University) for data collection including clinical examination of recruited healthy persons and glaucoma patients, Jesper Hougaard, MD PhD (Dept of Ophthalmology, Skåne University Hospital, Lund University) for creation of the OCT normative database and Ola Engwall Msc (Chalmers University of Technology) for aid with statistical analysis and programming. This study is supported by grants K2005-74X-1426-13A and K2005-74BI-15375-01A from the Swedish research council [vetenskapsrådet], by the foundation of Crown Princess Margareta for visually handicapped [Stiftelsen Kronprincessan Margaretas arbetsnämnd för synskadade (KMA)], by the foundation for visually impaired in the former Malmöhus län [Stiftelsen för synskadade i fd Malmöhus län], and by the Järnhardt foundation [Järnhardts stiftelse]. The Department of Clinical Sciences, Ophthalmology at Skåne University Hospital receives research funding from Carl Zeiss Meditec.

## Authors' contributions

DB conceived the study and the study design, collected the study material, conducted programming, performed statistical analyses and drafted the manuscript. AH participated in the study design, its coordination, and aided in drafting the manuscript. BB participated in the study design and its coordination, aided in statistical analyses and in drafting the manuscript. All authors read and approved the final manuscript.

## Competing interests

No conflicts of interest exist between the authors and other parties. AH and BB are both consultants for Carl Zeiss Meditec. All authors (DB, AH, BB) have applied for a patent pertaining to the presented methodology.

Received: 12 May 2011 Accepted: 4 August 2011

Published: 4 August 2011

## References

1. Ajtony C, Balla Z, Somoskeoy S, Kovacs B: Relationship between Visual Field Sensitivity and Retinal Nerve Fiber Layer Thickness as Measured by Optical Coherence Tomography. *Invest Ophthalmol Vis Sci* 2007, **48**(1):258-263.
2. Sommer A, Katz J, Quigley HA, Miller NR, Robin AL, Richter RC, Witt KA: Clinically detectable nerve fiber atrophy precedes the onset of glaucomatous field loss. *Arch Ophthalmol* 1991, **109**(1):77-83.
3. Sato S, Hirooka K, Baba T, Yano I, Shiraga F: Correlation between retinal nerve fibre layer thickness and retinal sensitivity. *Acta Ophthalmol Scand* 2008, **86**(6):609-613.
4. Hanwerth RS, Wheat JL, Fredette MJ, Anderson DR: Linking structure and function in glaucoma. *Prog Retin Eye Res* 2010, **29**(4):249-271.
5. Gardiner SK, Johnson CA, Cioffi GA: Evaluation of the Structure-Function Relationship in Glaucoma. *Invest Ophthalmol Vis Sci* 2005, **46**(10):3712-3717.
6. Chang R, Budenz DL: New developments in optical coherence tomography for glaucoma. *Curr Opin Ophthalmol* 2008, **19**(2):127-135.
7. Schuman JS, Pedut-Kloizman T, Hertzmark E, Hee MR, Wilkins JR, Coker JG, Pulliafto CA, Fujimoto JG, Swanson EA: Reproducibility of nerve fiber layer thickness measurements using optical coherence tomography. *Ophthalmology* 1996, **103**(11):1889-1898.
8. Paunescu LA, Schuman JS, Price LL, Stark PC, Beaton S, Ishikawa H, Wollstein G, Fujimoto JG: Reproducibility of nerve fiber thickness, macular thickness, and optic nerve head measurements using StratusOCT. *Invest Ophthalmol Vis Sci* 2004, **45**(6):1716-1724.
9. Budenz DL, Chang RT, Huang X, Knighton RW, Tielsch JM: Reproducibility of retinal nerve fiber thickness measurements using the stratus OCT in normal and glaucomatous eyes. *Invest Ophthalmol Vis Sci* 2005, **46**(7):2440-2443.
10. Bengtsson B, Olsson J, Heijl A, Rootzén H: A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmol Scand* 1997, **75**(4):368-375.
11. Heijl A, Lindgren G, Olsson J: A package for statistical analysis of computerized fields. *Doc Ophthalmol Proc Ser* 1987, **49**:153-168.
12. Budenz DL, Michael A, Chang RT, McSoley J, Katz J: Sensitivity and specificity of the StratusOCT for perimetric glaucoma. *Ophthalmology* 2005, **112**(1):3-9.

13. Hougaard JL, Hejili A, Bengtsson B: **Glaucoma detection by Stratus OCT.** *J Glaucoma* 2007, **16**(3):302-306.
14. Leung CK, Chong KK, Chan WM, Yiu CK, Tso MY, Woo J, Tsang MK, Tse KK, Yung WH: **Comparative study of retinal nerve fiber layer measurement by StratusOCT and GDx VCC, II: structure/function regression analysis in glaucoma.** *Invest Ophthalmol Vis Sci* 2005, **46**(10):3702-3711.
15. El Beltagi TA, Bowd C, Boden C, Amini P, Sample PA, Zangwill LM, Weinreb RN: **Retinal nerve fiber layer thickness measured with optical coherence tomography is related to visual function in glaucomatous eyes.** *Ophthalmology* 2003, **110**(11):2185-2191.
16. Hood DC, Kardon RH: **A framework for comparing structural and functional measures of glaucomatous damage.** *Prog Retin Eye Res* 2007, **26**(6):688-710.
17. Ferreras A, Pablo LE, Garway-Heath DF, Fogagnolo P, Garcia-Feijoo J: **Mapping standard automated perimetry to the peripapillary retinal nerve fiber layer in glaucoma.** *Invest Ophthalmol Vis Sci* 2008, **49**(7):3018-3018.
18. Turpin A, Sampson GP, McKendrick AM: **Combining Ganglion Cell Topology and Data of Patients with Glaucoma to Determine a Structure-Function Map.** *Invest Ophthalmol Vis Sci* 2009, **50**(7):3249-3256.
19. Zhu H, Crabb DP, Schlottmann PG, Lemij HG, Reus NJ, Healey PR, Mitchell P, Ho T, Garway-Heath DF: **Predicting Visual Function from the Measurements of Retinal Nerve Fiber Layer Structure.** *Invest Ophthalmol Vis Sci* 2010, **51**(11):5657-5666.
20. Bowd C, Goldbaum MH: **Machine learning classifiers in glaucoma.** *Optom Vis Sci* 2008, **85**(6):396-396.
21. Bengtsson B, Bizios D, Hejili A: **Effects of input data on the performance of a neural network in distinguishing normal and glaucomatous visual fields.** *Invest Ophthalmol Vis Sci* 2005, **46**(10):3730-3736.
22. Bizios D, Hejili A, Bengtsson B: **Trained artificial neural network for glaucoma diagnosis using visual field data: a comparison with conventional algorithms.** *J Glaucoma* 2007, **16**(1):20-28.
23. Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN, Sejnowski TJ: **Comparison of machine learning and traditional classifiers in glaucoma diagnosis.** *IEEE Trans Biomed Eng* 2002, **49**(9):963-974.
24. Goldbaum MH, Sample PA, Chan K, Williams J, Lee TW, Blumenthal E, Girkin CA, Zangwill LM, Bowd C, Sejnowski T, et al: **Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry.** *Invest Ophthalmol Vis Sci* 2002, **43**(1):162-169.
25. Bizios D, Hejili A, Hougaard JL, Bengtsson B: **Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fiber layer thickness parameters measured by Stratus OCT.** *Acta Ophthalmol Scand* 2010, **88**(1):44-52.
26. Burgansky-Eliash Z, Wollstein G, Chu T, Ramsey JD, Glymour C, Noecker RJ, Ishikawa H, Schuman JS: **Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study.** *Invest Ophthalmol Vis Sci* 2005, **46**(11):4147-4152.
27. Huang ML, Chen HY: **Development and comparison of automated classifiers for glaucoma diagnosis using Stratus optical coherence tomography.** *Invest Ophthalmol Vis Sci* 2005, **46**(11):4121-4129.
28. Poinoosawmy D, Tan JC, Bunce C, Hitchings RA: **The ability of the GDx nerve fibre analyser neural network to diagnose glaucoma.** *Graefes Arch Clin Exp Ophthalmol* 2001, **239**(2):122-127.
29. Sample PA, Boden C, Zhang Z, Pascual J, Lee TW, Zangwill LM, Weinreb RN, Crowston JG, Hoffmann EM, Medeiros FA, et al: **Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields.** *Invest Ophthalmol Vis Sci* 2005, **46**(10):3684-3692.
30. Tucker A, Vincicotti V, Liu X, Garway-Heath D: **A spatio-temporal Bayesian network classifier for understanding visual field deterioration.** *Artif Intell Med* 2005, **34**(2):163-177.
31. Bowd C, Chan K, Zangwill LM, Goldbaum MH, Lee TW, Sejnowski TJ, Weinreb RN: **Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc.** *Invest Ophthalmol Vis Sci* 2002, **43**(11):3444-3454.
32. Goldbaum MH, Sample PA, White H, et al: **Interpretation of automated perimetry for glaucoma by neural network.** *Invest Ophthalmol Vis Sci* 1994, **35**:3362-3373.
33. Boland MV, Quigley HA: **Evaluation of a combined index of optic nerve structure and function for glaucoma diagnosis.** *BMC ophthalmology* 2011, **11**(1):6-6.
34. Horn FK, Mardin CY, Bendschneider D, Junemann AG, Adler W, Tornow RP: **Frequency doubling technique perimetry and spectral domain optical coherence tomography in patients with early glaucoma.** *Eye* 2011, **25**(1):17-29.
35. Bowd C, Hao J, Tavares IM, Medeiros FA, Zangwill LM, Lee TW, Sample PA, Weinreb RN, Goldbaum MH: **Bayesian machine learning classifiers for combining structural and functional measurements to classify healthy and glaucomatous eyes.** *Invest Ophthalmol Vis Sci* 2008, **49**(3):945-945.
36. Racette L, Chiou CY, Hao J, Bowd C, Goldbaum MH, Zangwill LM, Lee TW, Weinreb RN, Sample PA: **Combining functional and structural tests improves the diagnostic accuracy of relevance vector machine classifiers.** *J Glaucoma* 2010, **19**(3):167-167.
37. Hougaard JL, Ostensfeld C, Hejili A, Bengtsson B: **Modelling the normal retinal nerve fibre layer thickness as measured by Stratus optical coherence tomography.** *Graefes Arch Clin Exp Ophthalmol* 2006, **244**(12):1607-1614.
38. Garway-Heath DF, Poinoosawmy D, Fitzke FW, Hitchings RA: **Mapping the visual field to the optic disc in normal tension glaucoma eyes.** *Ophthalmology* 2000, **107**(10):1809-1815.
39. Asman P, Hejili A: **Glaucoma Hemifield Test: Automated Visual Field Evaluation.** *Arch Ophthalmol* 1992, **110**(6):812-819.
40. Møller MF: **A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning.** *Neural Networks* 1993, **6**:525-533.
41. Haykin SS: **Neural networks: a comprehensive foundation.** Upper Saddle River, NJ: Prentice Hall; 2 1999.
42. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988, **44**(3):837-845.
43. Arthur SN, Aldridge AJ, León-Ortega JD, McGwin G, Xie A, Girkin CA: **Agreement in Assessing Cup-to-Disc Ratio Measurement among Stereoscopic Optic Nerve Head Photographs, HRT II, and Stratus OCT.** *J Glaucoma* 2006, **15**(3):183-189.

#### Pre-publication history

The pre-publication history for this paper can be accessed here:  
<http://www.biomedcentral.com/1471-2415/11/20/prepub>

doi:10.1186/1471-2415-11-20

**Cite this article as:** Bizios et al.: Integration and fusion of standard automated perimetry and optical coherence tomography data for improved automated glaucoma diagnostics. *BMC Ophthalmology* 2011 **11**:20.

#### Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



