EXTENDING BAYESIAN NETWORK MODELS FOR MINING AND CLASSIFICATION OF GLAUCOMA

A thesis submitted for the degree of
Doctor of Philosophy

Stefano Ceccon

School of Information Systems, Computing and Mathematics
Brunel University

April 2013

Abstract

Glaucoma is a degenerative disease that damages the nerve fiber layer in the retina of the eye. Its mechanisms are not fully known and there is no fully-effective strategy to prevent visual impairment and blindness. However, if treatment is carried out at an early stage, it is possible to slow glaucomatous progression and improve the quality of life of sufferers. Despite the great amount of heterogeneous data that has become available for monitoring glaucoma, the performance of tests for early diagnosis are still insufficient, due to the complexity of disease progression and the difficulties in obtaining sufficient measurements.

This research aims to assess and extend Bayesian Network (BN) models to investigate the nature of the disease and its progression, as well as improve early diagnosis performance. The flexibility of BNs and their ability to integrate with clinician expertise make them a suitable tool to effectively exploit the available data. After presenting the problem, a series of BN models for cross-sectional data classification and integration are assessed; novel techniques are then proposed for classification and modelling of glaucoma progression. The results are validated against literature, direct expert knowledge and other Artificial Intelligence techniques, indicating that BNs and their proposed extensions improve glaucoma diagnosis performance and enable new insights into the disease process.

Acknowledgments

First and foremost I would like to thank my supervisor Allan Tucker, who accompanied me through this journey with great guidance and support. His enthusiasm and positivity were contagiuous and helped me throughout all the PhD. I am also thankful for the excellent example that he has provided me with as a successful man in life and academia.

I would also like to thank the members of the Glaucoma Research Unit at Moorfields Eye Hospital and City University, for their direct contributions to my work and interesting discussions. In particular, thanks to prof. David Garway-Heath for providing support, despite being very busy. He has shown me how to effectively guide a great research group with dedicatation and kindness.

I am also grateful to Luke Saunders, Ailbhe Finnerty and especially Richard Russell, for helping me in the writing process. Without their help this piece of work would not be intelligible to humans.

I would also like to thank my past and present colleagues, for interesting discussions and for keeping the lab fun and a pleasure to work in: Valeria, Fadra, Cici, Djibril, Ali, Amirahmad, Maciej, Sara, Panos, Chandrika, Stelios and Fotis.

A big thankyou to my friends and housemates. Their company and friendship have made the last three years a wonderful period of my life. A special mention goes to Stefani, who was there for me during all these years.

Last but definitely not least, thankyou to my family for the love and support, and especially for giving me the best tools to find my way in life.

”We are drowning in information and starving for knowledge”

Rutherford D. Roger

Publications

The following publications have resulted from the research published in this thesis:

1.

   Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2010). Investigations of clinical metrics and anatomical expertise with Bayesian Network models for classification in early glaucoma. In *Intelligent Data Analysis in Biomedicine and Pharmacology (IDAMAP) 2010.*

2.

   Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2010). Combining expertise-driven and semi-supervised Bayesian Networks for classification of early glaucoma. In *European Conference in Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).*

3.

   Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2011). The Dynamic Stage Bayesian Network: identifying and modelling key stages in a temporal process. In *Advances in Intelligent Data Analysis X - Lecture Notes in Computer Science*, 7014, 101-112.

4.

   Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2011). Ensembles of Bayesian Network Classifiers using glaucoma dat and expertise. In *Ensembles in Machine Learning Applications - Studies in Computational Intelligence*, 373, 133-148.

5.

   Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2012). Non-Stationary Clustering Bayesian Networks for glaucoma. In *International Conference on Machine Learning (ICML) 2012.*

6.

Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2012). Clustering post-trabeculectomy glaucoma patients using Non-Stationary Dynamic Bayesian Networks. In *Intelligent Data Analysis in Biomedicine and Pharmacology (IDAMAP) 2012.*

Item 1 results from the research described in Chapters 3 and 4. Items 2 and 4 result from the research presented in Chapter 4. Finally, Items 3, 5 and 6 are based on work discussed in Chapter 5. In addition, the work in Chapter 4 has been submitted as a journal article to *IEEE Journal of Biomedical and Health Informatics*, and is currently in the second cycle of the review process. The work in Chapter 5 has been submitted as an article to *Artificial Intelligence in Medicine.*

Contents

List of Figures

List of Tables

Key Acronyms

ML Machine Learning

AI Artificial Intelligence

MLC Machine Learning Classifier

BN Bayesian Network

VF Visual Field

RT Retinal

OD Optic Disc

NFL Nerve Fiber Layer

RNFL Retinal Nerve Fiber Layer

IOP Intraocular Pressure

ON Optic Nerve

ONH Optic Nerve Head

SAP Standard Automated Perimetry

OCT Optical Coherence Tomography

HFA Humphrey Visual Field Analyzer

HRT Heidelberg Retinal Tomograph

MRA Moorfields Regression Analysis

AGIS Advanced Glaucoma Intervention Study

MLP Multilayer Perceptron

KNN K-Nearest Neighbour

CPD Conditional Probability Distribution

CPT Conditional Probability Table

ROC Receiver Operator Characteristic

AUROC Area under the ROC Curve

SA Simulated Annealing

EM Expectation-Maximisation

SEM Structural Expectation-Maximisation

CSA Clustering Simulated Annealing

HMM Hidden Markov Model

DBN Dynamic Bayesian Network

DSBN Dynamic Stages Bayesian Network

NSDBN Non-Stationary Dynamic Bayesian Network

NSCBN Non-Stationary Clustering Bayesian Network

BIC Bayesian Information Criterion

Ts Temporal Superior Sector

Ns Nasal Superior Sector

Ni Nasal Inferior Sector

Ti Temporal Inferior Sector

N Nasal Sector

T Temporal Sector

Chapter 1

Introduction

Glaucoma is an optic neuropathy that affects the retinal nerve fibre layer of the eye. It is the second leading cause of blindness worldwide and is an irreversible disease. Its mechanisms are not fully known, but it has been shown that early medication can slow its progression. However, early diagnosis is often inaccurate due to the high inter- and intra-subject variability. In addition, tests of both healthy and glaucomatous patients are subject to a large degree of fluctuation, which makes it difficult to decide whether any observed change is due to inherent variation or real progression of the disease between examinations.

While technology provides several methods to monitor glaucoma and large amounts of data, there is no gold standard definition of glaucoma progression. Thus, in clinical practice, decisions are based on subjective interpretation of test results by expert clinicians, who are not able to fully exploit all the available data. Artificial Intelligence (AI) techniques suit this type of problem very well, being able to analyse and integrate large and heterogeneous datasets. In fact, several AI techniques have been successfully applied to glaucoma. Among these, Bayesian Networks (BNs) have shown promising results both for modelling and classification of glaucoma. Their flexibility and ability to integrate with clinician expertise make them a suitable tool for more effectively exploring the available data.

This research aims to assess and extend BNs to investigate the nature of glaucoma and its progression, as well as to improve early diagnosis performance. After presenting the problem, a series of BNs for cross-sectional data classification and integration are assessed; in the second phase of the work, novel techniques are proposed for classification and modelling of glaucoma progression. The methods and the results of this research are discussed throughout the thesis in relation to their performance and potential for use in clinical practice. The remainder of this chapter presents the motivations, aims and contributions of the thesis. An overview of glaucoma and the available data is presented in section 1.1. Following this, section 1.2 sets out the aims and research questions of this thesis, while its contributions are listed in section 1.3. The structure of the thesis is outlined in section 1.4.

1.1 Modelling Glaucoma

Glaucoma is a leading cause of irreversible blindness worldwide. The primary site of glaucomatous damage is the retina and optic nerve (optic disc), and raised intraocular pressure (IOP) is a major risk factor for the development and progression of the disease. No cure is available for glaucoma; however, reducing IOP may slow disease progression. IOP-reduction may be attained through the use of medication or by surgical means. Trabeculectomy is one such surgical procedure, which creates an alternative route for aqueous humour to drain from the anterior eye, thereby reducing IOP. The timing of intervention is a key factor in glaucoma treatment, so early diagnosis is highly desirable.

Glaucoma may be monitored through imaging of the retina over time to detect changes in its structure. Alternatively, the disease can be monitored by measuring eye functionality, specifically through the use of visual field (VF) testing, known as perimetry. Both approaches have recently undergone a technological revolution, making more data available. For example, in the past a photo of the retina was used to assess the optic disc conditions, while now hundreds of parameters are provided by advanced computerised tomographies of the retina. In the same fashion, automated VF tests are used to build a 52-point map of VF sensitivity of a patient, in less time than was possible before using manual perimetry. Despite the technological advances and the great amount of data available, the performances of glaucoma tests to diagnose the condition remain unacceptable. Furthermore, the rate of glaucoma incidence is snowballing so it is becoming increasingly necessary to utilise all of the available data in order to more efficiently identify glaucomatous patients as early as possible, and to understand why and how sufferers develop glaucoma, while other individuals do not.

To improve classification performance it is useful to recognise the inherent characteristics of the disease and problems associated with the data available. In fact, glaucoma is a set of pathological and physiological conditions, and thus develops from different starting conditions, at different rates and exhibits different patterns of damage. Several recurrent spatial patterns of VF defects have been identified, but there is no understanding as to why certain patients develop certain patterns. Furthermore, the degeneration of the optic nerve is a physiological process that occurs naturally with ageing, therefore there is an overlap between the condition of the structure in healthy individuals and patients with glaucoma. In this context, there is no gold standard definition of glaucoma. Instead, different studies have used different criteria and diagnosis metrics, in relation to the instrument used. Moreover, the results of functional tests often do not agree with those from structural tests, even though the functional manifestation of glaucoma is caused by structural damage. Finally, the data from all instruments in subject to very high ”noise” (variability). Patients exhibit large variability in their measurements even if tests are conducted on the same day. In the case of VF testing, differences can be explained by the complex nature of the task; the subjects reaction time, physical conditions and attitude to the test all greatly influence its measurements.

Many data analysis techniques have been proposed to exploit the available data in order to perform better classification and gain insights into glaucoma. Simple statistical techniques are implemented in screening instruments, but these tend to suffer from all the issues described in the previous paragraph. Advanced Machine Learning (ML) techniques have also been proposed, mainly only with cross-sectional data, showing promising results. Among these, the so called white-box models are of particular interest to clinicians because they are easy to understand and can give meaningful insights. This thesis focuses on one such model, called the Bayesian Network (BN). This is a graphical model that may be used in this context to model spatial and anatomical-functional relationships of glaucoma, as well as provide an efficient probabilistic parametrisation of the data.

1.2 Thesis Aims

The previous section introduced motivations for the use of data analysis techniques to help clinicians in making decisions and understanding the mechanisms behind glaucoma. It also discussed the characteristics of the data and problems associated with them.

With respect to cross-sectional data, the identified key aspects of the problem include data integration, inter-subject variability, noise and biases in the data. Regarding the progression of glaucoma, the problem of non-stationary time dependencies is appointed as a key aspect in the modelling.

The objective of the research presented in this thesis is to overcome these issues and build a set of models that are able to gain insights into the field and enhance classification performance. To achieve this, simple and extended BN models are explored.

The main contributions of this thesis can be described as a knowledge-building process. Starting with cross-sectional glaucoma modelling and classification, the research describes glaucoma metrics and data integration, before covering the less-explored field of modelling glaucoma progression. Using the cross-sectional data available, a set of ad-hoc BN models is built. The extension into semi-supervised and unsupervised models is proposed in order to avoid metric biases, while data integration is tackled by the use of ensembles of classifiers. To increase performances in classification, special learning algorithms are implemented. A novel approach to cluster the data is also presented. To tackle the more complicated issue of describing glaucoma progression, BN models are extended to deal with longitudinal data that account for non-stationarity in the data. The proposed techniques are also able to simultaneously cluster patients into different stages of disease.

While high diagnosis performance is pursued throughout this research, particular focus is given to gaining insight with regards to the glaucoma disease process and its manifestation. However, the proposed techniques represent a general framework that can be applied to a wide range of problems.

1.3 Thesis Contributions

The key contributions of this thesis are outlined below:

- **A set of supervised, semi-supervised and unsupervised BN to model glaucoma and increase classification performance.**

  A set of expertise-driven and data-driven supervised models is obtained using both anatomical and functional glaucoma data. To avoid metric biases, unsupervised and semi-supervised models are also proposed. Empirical results demonstrate an increase in performance and insights gained with respect to state-of-the-art approaches.

- **An ensemble of classifiers to model anatomical and functional metrics and combine different BNs.**

  A BN model is used as a stacked ensemble of classifiers to increase classification performance by integrating different base models. In particular, anatomical and functional metrics are integrated and evaluated by the model, allowing understanding of how metrics interact and impact on the final decision. Comparisons with expertise and empirical evaluations show the utility of this approach in glaucoma.

- **A metric-independent VF clustering BN-based model.**

  With regards to functional data, a new algorithm is proposed to model metric-independent variance and investigate clusters in glaucoma VF data. A novel algorithm that makes use of an heuristic search to perform clustering and classification is used. Results show insights are gained with respect to state-of-the-art algorithms.

- **An extension of BN models for classification and modelling of longitudinal data.**

  A new model is proposed as an extension of BNs for longitudinal data. The model is able to extract key disease stages and discover patterns in temporal data. It also can be used for classification with empirical results showing promise.

- **A new technique for modelling and clustering non-stationary temporal processes.**

A novel algorithm based on Non-Stationary Dynamic BNs is developed and assessed using the longitudinal functional data available. The model can capture non-stationary variation as well as cluster groups of patients into different temporal patterns. The process is independent of any clinical metric and allows key disease stages to be identified. Empirical evaluations and comparison with expert knowledge demonstrate that such models are of great interest for understanding glaucoma.

1.4 Thesis Outline

The remainder of the thesis is structured as follows:

- Chapter 2 illustrates the physiology of glaucoma and gives an overview of the process of monitoring the condition. Screening instruments and different types of data are also introduced, followed by a comparative analysis of state-of-the-art AI techniques for glaucoma.

- Chapter 3 introduces BN models in relation to other AI models. The issues related to the data are also outlined and simple BN models are evaluated.

- Chapter 4 assesses specific ad-hoc BN models. Supervised, semi-supervised and unsupervised BN models are built using clinical expertise as well as anatomical and functional data. The resulting models are then integrated using an ensemble of classifiers. A model to describe metric-independent variance and extract clusters in the data is also proposed. Part of the work included in this chapter has been published in (Ceccon et al., 2010b; Ceccon et al., 2010a; Ceccon et al., 2011a), and has also been submitted to *IEEE Journal of Health and Biomedical Informatics* (currently in the second cycle of the review process).

- Chapter 5 describes glaucoma progression and classification, and develops a model to take into account non-stationarities in the data. An extension of Non-Stationary Dynamic Bayesian Networks for clustering temporal processes is also explored. The work presented in this chapter has been published in (Ceccon et al., 2011b; Ceccon et al., 2012a; Ceccon et al., 2012b) and has also been submitted to *Artificial Intelligence in Medicine*.

- Chapter 6 presents contributions and conclusions of this research, and discusses limitations of the current work and potential paths for further research.

Chapter 2

Glaucoma and Machine Learning

Glaucoma is a major eye disease and is the second leading cause of blindness in the world. In recent years, new technologies to monitor glaucoma have become available, producing a great amount of data. In this context, Machine Learning (ML) techniques represent a tool that can be used to improve diagnosis performance and shed light on the disease process. This chapter forms the literature review of this thesis. In the first section, glaucoma and its mechanisms are introduced, followed by an overview of the instruments available. Section 2.3 presents and discusses the ML techniques that have been explored in the context of glaucoma, while in section 2.4 a comparative review of the results obtained using such techniques is presented. Finally, section 2.5 closes the chapter and draws conclusions that form the motivations for research in this thesis.

2.1 Glaucoma, the Sneak Thief of Sight

The World Health Organization (WHO) has estimated that the percentage of blindness due to glaucoma is around 20% in western countries (Resnikoff et al., 2004). Visual impairment can occur without a sufferer noticing, and it is projected that about 50% of individuals with glaucoma in North America (and most likely the western world) are not aware they have the disease (Quigley, 1996). For this reason, glaucoma is often referred to as *the sneak thief of sight*. In fact, glaucoma is often unnoticed until irreversible damage occurs in the retina.

Glaucoma is an increasing problem: in 2010, there were 60.5 million people with the disease, but according to (Quigley and Broman, 2006) and the WHO Vision 2020 (Pizzarello et al., 2004), in 2020 there will be 79.6 million people with glaucoma, given the growing longevity of the world's population.

The American Academy of Ophthalmology, in 2010, defined glaucoma (open-angle variant) as "a progressive, chronic optic neuropathy" in which there is a "characteristic acquired atrophy of the optic nerve and loss of retinal ganglion cells and their axons" (American Academy of Ophthalmology, 2010). The characteristics of glaucoma are based on "reliable and reproducible visual field (VF) abnormalities" or "optic disc or retinal nerve fibre layer structural abnormalities". Early glaucoma is defined as characteristic optic nerve abnormalities and a normal VF. However, among past and recent studies, there is a lack of consistency in how glaucoma is defined (Pizzarello et al., 2004). For example, also in 2010, the American Optometric Association (American Optometric Association, 2010) defined early glaucoma according to optic nerve and nerve fibre layer appearance, and an abnormal VF. Other recent studies also include functional measurements in the definition of early glaucoma (Artes and Chauhan, 2005; Bathija et al., 1998). In the past, increased intraocular pressure (IOP) was typically included in the definition of glaucoma. However, researchers demonstrated that IOP values may overlap considerably between normal and glaucomatous groups and that IOP is influenced by many factors (Kronfeld, 1952; Podos and Becker, 1973; Shields et al., 1996). Furthermore, several studies have indicated that a sizeable percentage of patients have a "low" IOP (Sommer and Doyne, 1996; Shiose et al., 1991); this condition is commonly referred to as Normal Tension Glaucoma.

Glaucoma is, in fact, a set of quite diverse eye conditions, and about 60 different types of glaucoma are known. The most common type of glaucoma is the Primary Open Angle Glaucoma (POAG). Major population studies found prevalence of POAG ranges from 0.9% to 2.4% in Caucasians (Dielemans et al., 1994; Leske et al., 1994; Mitchell et al., 1996; Bengtsson, 1981; Klein et al., 1992; Tielsch et al., 1991), and from 3,9% to 8,8% in African-Caribbeans and African-Americans (Tielsch et al., 1991; Leske et al., 1994; Mason et al., 1989; Wormald et al., 1994). In the same studies, Primary Angle Closure Glaucoma (PACG) prevalence has been quoted to range from 0.04% to 0.9%, and other secondary glaucomas from 0.09% to 0.27%. However, as already stated, it is estimated that 50% of those with glaucoma do not know they have it (Coffey et al., 1993; Mitchell et al., 1996; Sommer et al., 1991b), a percentage that rises to over 90% in developing nations with poor access to health-care. (Buhrmann et al., 2000; Ramakrishnan et al., 2003)

2.1.1 Mechanisms of the Disease

For most forms of glaucoma, IOP is a major causative risk factor for the development and progression of the disease (Yanoff and Duker, 2008). Elevation of IOP is caused by the production of aqueous humor by the non-pigmented epithelium of the ciliary body, indicated in Figure 2.1.

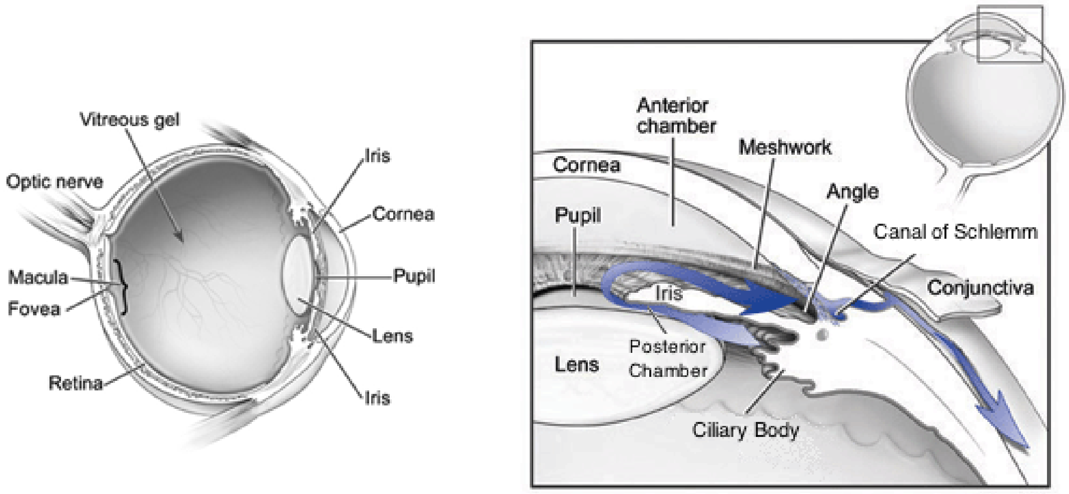


Figure 2.1: The aqueous humor outflow pathway. Adapted from (National Eye Institute, 2006).

This tissue actively transports ions and nutrients into the posterior chamber, which then flow into the anterior chamber through the pupil, to eventually reach the Schlemm's canal through the trabecular meshwork. From here, liquid diffuses into the aqueous veins and general circulation. Identification of the site (and nature) of any impediment to the aqueous flow is extremely useful to diagnose glaucoma. If the site is at the pupil (pupillary block), the increased size of the lens blocks the flow through the pupil, which results in a gradient of pressure that causes the iris to bow forward and mechanically cover the trabecular meshwork, increasing IOP. This is known as PACG. If the obstruction is at the level of the trabecular meshwork then the condition is known as open-angle glaucoma. In this case there are many different reasons for the block, which lead to numerous different subtypes of glaucoma. For example, if the obstruction is caused by abnormalities in the extracellular matrix of the trabecular meshwork, it is referred to as POAG.

The mechanism of optic nerve injury in glaucoma is not well-characterised, but there is evidence that elevated IOP damages the Optic Nerve Head (ONH) (Yanoff and Duker, 2003). This is the location where the optic nerve (ON) joins the retina. The ONH is also referred to as the Optic Disc (OD), because of its disc-like shape. Damage to the ONH causes injury to retinal ganglion cell axons; either their local blood supply is compromised (Quigley and Anderson, 1976), or they are mechanically pinched (Quigley and Addicks, 1981; Anderson, 1996). Recently, it has been argued that, given the non-uniform clinical phenotype of glaucoma, it is likely that multiple mechanisms account for initiation or progression of glaucomatous damage, each of them with a variable impact on the phenotype seen clinically (Yanoff and Duker, 2008). All these different stimuli eventually activate the programmed death of retinal ganglion cells. This process is thought to occur by apoptosis and results in a so-called cupping of the ONH. As highlighted in the left part of Figure 2.2, the cup is an empty space in the middle of the ON, surrounded by optic fibres (the so-called neuro-retinal rim). It resembles a cup if looked from above in 3-dimensions. When glaucoma progresses, the fibres of the external part of the ONH die (thinning of the neuro-retinal rim) making the cup larger. Functional impairment typically begins with loss of peripheral vision and progresses to so-called tunnel vision, where only a central spot of the VF is preserved. The loss of vision is also described as a loss of VF sensitivity. Figure 2.2 (right) shows a simulation of how driving appears to patients with moderate glaucoma.

2.1.2 Risk Factors for Glaucoma

The risk factors for developing glaucoma can be broadly grouped into demographic, ocular, genetic and other. The known demographic factors include age and ethnic origin. Increasing

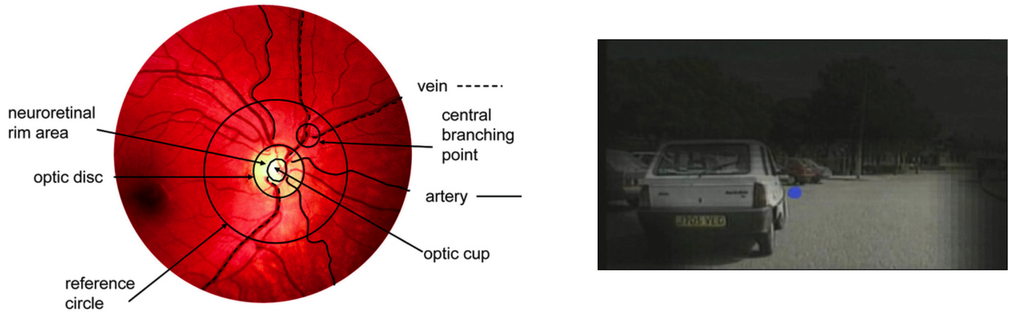


Figure 2.2: Left: Drawing of the retina illustrating the optic disc, cup, and neuroretinal rim. Source: (Wikstrand et al., 2010). Right: Simulation of impaired vision due to moderate glaucoma. Source: (Crabb, 2012)

age is a major risk factor for glaucoma that has been described in all population-based studies of glaucoma where age has been examined (Tielsch, 1991). Ethnic origin has recently been confirmed by large-population studies to be an important factor as well. In particular, individuals of African, African-American and African-Caribbean origin are at higher risk of POAG (Leske et al., 1994; Tielsch et al., 1991; Klein et al., 1992; Wormald et al., 1994). The ocular risk factors associated with glaucoma are IOP, ONH, hypermetropia and systemic hypertension. IOP is known to be strongly related with glaucoma, and strong evidence supports this: POAG incidence and prevalence increase with increasing IOP (Buhrmann et al., 2000; Ramakrishnan et al., 2003; Sommer et al., 1991b). Moreover, the rate of VF loss due to glaucoma is slowed or even prevented when IOP is lowered (Kass et al., 2002; Leske et al., 2003; Heijl et al., 2002) and, in eyes with asymmetric IOP, VF loss is more severe in the eye with higher IOP than in the other eye (Cartwright and Anderson, 1988; Chrichton et al., 1989). Although IOP is strongly related to glaucoma, it has been shown that it is a poor diagnostic tool. The cut-off value of 21 mmHg commonly used to separate normal tension glaucoma and high tension glaucoma is not a threshold level of any particular biological importance and has poor diagnostic performance. In (Quigley et al., 2001), a cut-off of 22 mmHg would result in only 80% of POAG patients being diagnosed. This is consistent with other studies demonstrating that among patients with characteristic POAG, 20-30% patients have lOPs consistently below 21 mmHg (Sommer and Doyne, 1996; Kamal and Hitchings, 1998). Although a very high IOP results in glaucomatous damage, this is not inevitable for lower IOP (Davanger et al., 1991). It is therefore suggested that the relationship between lOP and glaucoma is a continuum, much like the relationship between systemic hypertension and stroke. A realistic concept is to consider that an individual's ON has a level of IOP that it can or cannot withstand, manifested at a clinical level by the presence or absence of VF damage. Since other factors also contribute with IOP to produce characteristic glaucomatous changes, IOP is considered a risk factor but not a diagnostic requisite for glaucoma. Structural changes at the ONH are an important marker of glaucoma, and play a pivotal role in disease pathogenesis. Disc variation and vertical and horizontal cup-to-disc-disc ratios correlate with subsequent VF loss (Shiose and Tsukahara, 1990). In addition, African groups, who are more susceptible to glaucoma, have been noted to have larger discs and cup-to-discs ratios than Caucasians (Zangwill et al., 2004; Sommer and Doyne, 1996). Myopia has also been found to be associated with POAG, with rates 2 to 4 times higher for myopes (Ramakrishnan et al., 2003; Mitchell et al., 1999). In the Baltimore Eye Survey (Tielsch et al., 1995) a strong association between POAG prevalence and low diastolic perfusion pressure was found. The latter is a measure obtained from systemic blood pressure. A hypothesis is that ON damage may occur in these subjects because of poor optic nerve perfusion. Among the genetic factors, there is little doubt that glaucoma is more common if there is a positive family history; a 5 to 20 times increase in prevalence rate occurs in those who have a positive family history (Yanoff and Duker, 2003). Unbiased population studies, such as the Baltimore Survey, have also shown that family history is a significant risk factor for POAG (odds ratio of 3.69 for those with siblings with disease, 2.17 for those with parents and 1.12 for those with children).

The decision on how to treat glaucoma is a complex task that aims to preserve the vision and therefore the quality of life of patients, while taking into consideration the risk of side effects and complications. Lowering eye pressure using medication is the most effective treatment in glaucoma management, but surgery can also give excellent IOP control. An operation, called trabeculectomy, is one such surgical procedure which creates an alternative route for aqueous humor to drain from the anterior eye, thereby reducing IOP (Khaw, 2001).

2.2 Progression of Glaucoma: Structure and Function

Glaucoma onset is often unnoticed until later stages of the disease. Early diagnosis, treatment and follow up are key to the management of the disease, therefore understanding the mechanisms involved in the onset and progression of glaucoma is of great value for clinicians. In this context, it is necessary to consider the complex relationship between structure (i.e. the anatomical characteristics of the nerve fibre layer and optic disc) and function (i.e. quality or "sensitivity" of vision). The nature of relationship remains unclear and controversial, especially in the progression of glaucoma. However, it is known that the relative utilities of structural and functional tests vary at different stages of disease (Strouthidis et al., 2006), and in many cases VF test results or OD imaging results alone may not be sufficient for accurate diagnosis or disease monitoring. For example, the OD in a patient suspected of glaucoma may change, yet remain within the large "healthy" range of physiological between-subject variability. Likewise, detectable change over time in the VF can be highly suggestive of disease, even if distinct VF measurements may appear well within normal limits (Artes and Chauhan, 2005). It is frequently reported that OD examination has more utility during the earlier stages of the disease than VF analysis and can predict future VF progression (Chauhan et al., 2009). It is also purported that particular functional tests have greater utility than others in early glaucoma (Demirel and Johnson, 2000).

2.2.1 Investigating Glaucoma Progression

Current methods for determining glaucomatous progression can be grouped into four categories (Spry and Johnson, 2002): clinical judgement, defect classification systems, event analyses and trend-based analyses. Clinical judgement methods involve subjective grading of the OD and VF, and categorise progressing and non-progressing eyes without any standardisation of the criteria used for change. Consequently, inter-observer and intra-observer agreement are often poor (Kamal et al., 1999; Azuara-Blanco et al., 2003; Viswanathan et al., 2003), and classification is usually binary. Defect classification systems for ODs and VFs use specific criteria to stratify loss as a discrete score and define progression if the score changes over time. An example is the Advanced Glaucoma Intervention Study (AGIS) scoring system for VFs (Gaasterland et al., 1994). The AGIS defect scoring system depends on the number and depth of clusters of adjacent depressed VF points in the Humphrey Field Analyzer II (HFA II) VF instrument. The AGIS score ranges from 0 (no defect) to 20 (all test points deeply depressed). Another category of methods for classification of ODs and VFs is event analyses, in which a follow-up examination is compared to a baseline VF and changes are compared to variability limits defined from "stable" glaucoma patients; if the change exceeds this limits, progression is deemed to have occurred. The last category of methods for assessing change in VFs and ODs is trend-based analyses. These involve testing a dependent variable (such as VF sensitivity or OD rim area) sequentially over time with standard linear regression analysis. Linear regression accounts, to some extent, for variability within each subject, but often requires many examinations and/or follow up time to establish progression (Vesti et al., 2003). These approaches are available in the most common imaging and perimetry instruments, but determining progression in individual patients remains one of the most challenging aspects of glaucoma management.

One advantage of clinical judgement to define glaucomatous progression is that clinicians are trained to identify typical patterns of VF damage that are associated with glaucoma that are difficult to classify precisely using current scoring systems, events analyses or trend analyses. These patterns include sensitivity changes across the horizontal meridian, especially in the nasal portion of the field (i.e. the part of the VF that is closer to the nose) and shallow defects occurring in the Bjerrum area (an arcuate area in the superior half of the VF), generalised loss of retinal sensitivity, other typical arcuate scotomas (i.e. areas of degenerated sensitivity), enlargement of the blind spot, and selective loss of sensitivity in the nasal periphery (Drance, 1969; Hart Jr and Becker, 1982; Spry and Johnson, 2002; Artes and Chauhan, 2005; Yanoff and Duker, 2003). Characteristic OD changes over time are also well-known and are investigated using disc photography (Tuulonen and Airaksinen, 1991; Airaksinen et al., 1992) and more recently high-resolution scanners (Chauhan et al., 2001; Kamal et al., 1999). Typical structural changes observed are enlargement of the OD cup and Retinal Nerve Fibre Layer (RNFL) thinning.

2.2.2 The "Structure-Function" Relationship

The cross-sectional relationship between visual function measurements and structural measurements of the OD or nerve fibre layer has been studied for many years. In (Garway-Heath et al., 2000), an anatomical relationship between VF test points and regions of the ONH was established. The technique used was to superimpose a VF grid onto RNFL images, and calculate the proximity of test points to RNFL defects and/or prominent bundles. ODs were divided into six sectors, relating to six VF areas. Figure 2.3 shows the structure-function map proposed in the study. The results show that in 95% of cases the VF test point was associated with a position at the ONH within 14° either side of the mean. A good correlation with previous studies was also found (Wirtschafter et al., 1982; Weber et al., 1990). The structure-function relationship in glaucoma was also investigated in (Johnson et al., 2000; Strouthidis et al., 2006) and good correlations were found between VF locations and angular distance at the corresponding part of the ONH. However, (Artes and Chauhan, 2005) demonstrated that in the eyes of 84 patients with OAG and 41 healthy subjects, a remarkably poor correlation exists between VF and OD changes. According to the authors, this

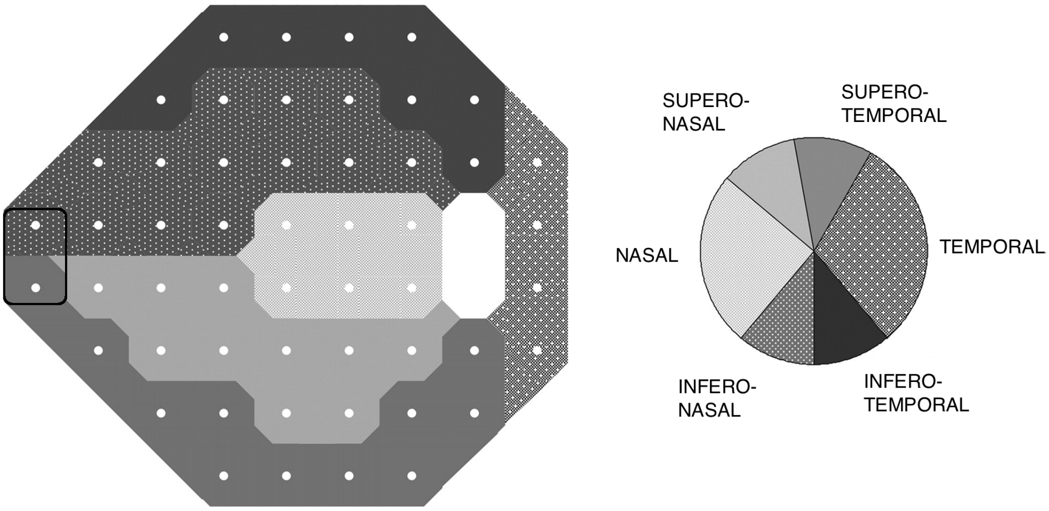


Figure 2.3: Anatomic relationship between Humphrey 24-2 VF test-points (left) and optic nerve head disc sectors (right). A black rectangle indicates the nasal step. Sectors are also referred to as T (Temporal), Ti (Temporal-Inferior), Ts (Temporal-Superior), N (Nasal), Ns (Nasal-Superior) and Ni (Nasal-Inferior). Source: (Garway-Heath et al., 2000)

does not demonstrate that the actual relationship is weak, but rather that structural and functional measurements are subject to high variability. Even between different functional tests, the correlation between measurements over time is weak. Thus, in clinical practice it is very unlikely that VF changes can be corroborated by OD changes and vice versa. In (Hood and Kardon, 2007), a linear model relating glaucomatous loss in RNFL thickness to the loss of sensitivity is proposed. Excellent agreement with the (Garway-Heath et al., 2000) map was found. The authors also found that initial large RNFL thickness measurements were associated with eventual abnormal RNFL, explaining the difficulty in understanding whether ON or VF damage occurs earlier. Other studies found strong correlations between inferior RNFL thickness measurements and the superior VF (Bowd et al., 2006; Sanchez-Galeana et al., 2001). Furthermore, in (Gardiner et al., 2005), a topographical map of sectors of the optic nerve and corresponding areas of the VF was found to support the map in (Garway-Heath et al., 2000). Although some areas (inferior and inferotemporal) presented good correlation with predicted VF locations, other areas obtained weaker or unexpected correlations. The authors argue that the regional structure-function relationships may vary between individuals more than previously assumed (Mackenzie and Cioffi, 2008). In (Sommer et al., 1991a), VF defects were found in corresponding RNFL sectors, amongst 1344 patients that were followed for 6 years; in many cases, structural change occurred before functional change. In (Kass et al., 2002), 55% of glaucoma patients presented with OD changes first, 35% presented with VF damage before OD change, while 10% presented with simultaneous structure-function change. However, different results were obtained when using different onset criteria. An interesting paper relevant to this issue is (Vesti et al., 2003), where 7 different methods of determining VF change were compared, obtaining very different results. In (Heijl et al., 2002), 86% of 255 patients observed presented with VF change alone, while 13% presented with simultaneous changes in the VF and OD. Only 1 patient of 136 presented with isolated OD change. Conversely, in (Johnson et al., 2003), 479 eyes were followed for 4 or more years, and the authors found that structural changes generally occurred in the absence of VF abnormality. Another longitudinal study (Chauhan et al., 2001) found that only 4% of the patients progressed in VFs only, and most of the cases progressed on OD only or on both OD and

VF. Histological studies have also been conducted to investigate the structure-function relationship looking at post-mortem retinal ganglion cell counts in glaucoma patients (Quigley et al., 1989; Harwerth et al., 1999). These studies found support for OD changes occurring before VF changes and for a linear relationship between VF sensitivity and cell count.

2.3 Screening for Glaucoma

Early screening for glaucoma provides a method to promptly intervene and possibly slow the glaucoma disease process (Heijl et al., 2002; Kass et al., 2002). To this purpose, several techniques have been proposed over the years, using several instruments with different results. In this section we will review the properties of different screening instruments, with particular focus on the most recent state-of-the-art technologies that have been adapted for glaucoma.

To compare performances of different techniques, the proportion of positive subjects identified by each test is usually evaluated. This is known as the sensitivity, which is not to be confused with VF sensitivity, which refers to the level of vision at each point in the VF. The specificity, on the other hand, is the proportion of healthy subjects correctly excluded by a test. Sensitivity and specificity are also known respectively as the True Positive rate and the True Negative rate, and can be combined for performance comparison, as will be discussed in the section 2.4.3.

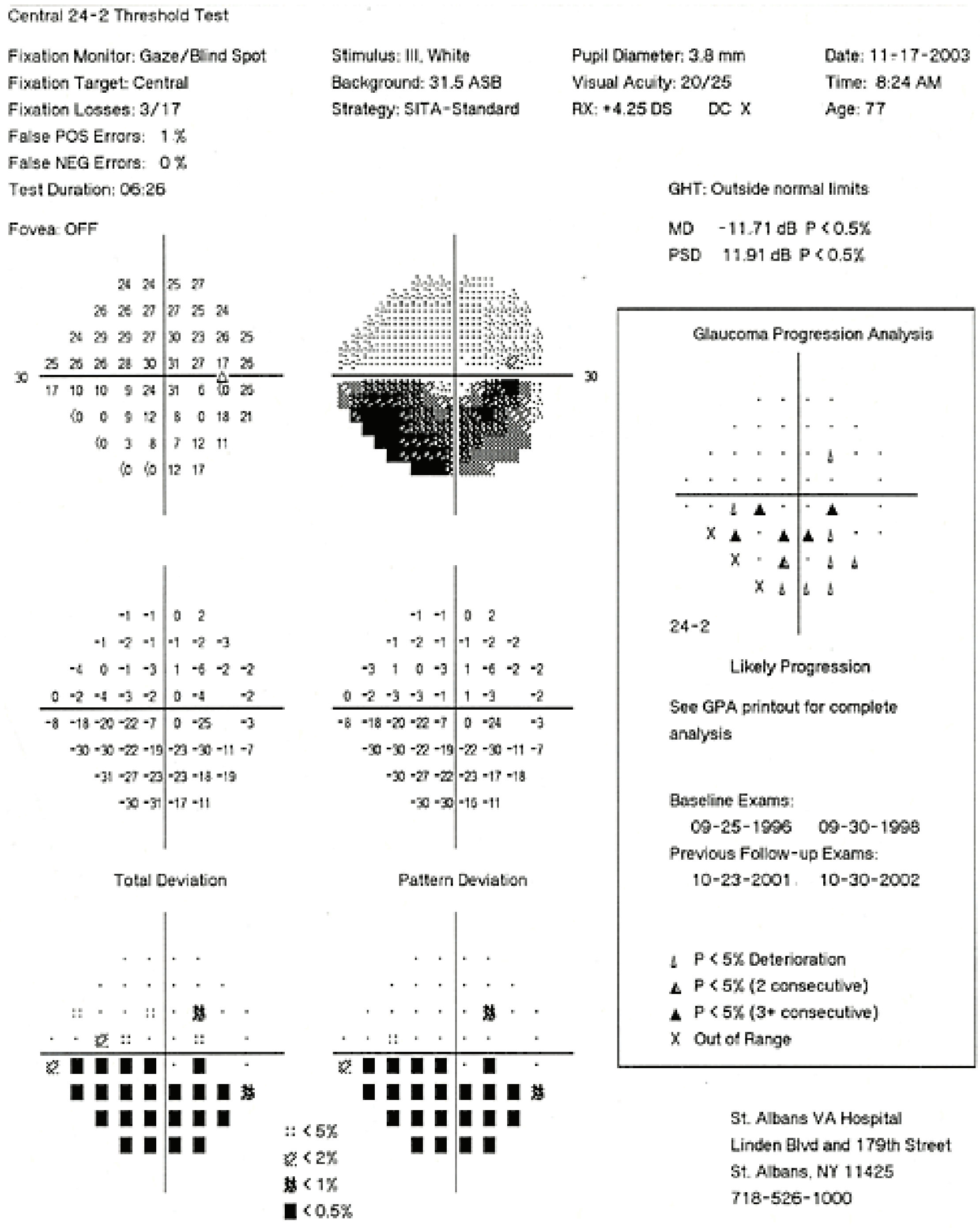
2.3.1 Visual Field Assessment

The most commonly used instruments for assessing a patient's VF are the Octopus Perimeter and the Humphrey VF Analyzer (HFA). These instruments perform the "Standard Automated Perimetry" (SAP), where illuminated targets are projected onto an illuminated background. The brightness of the target is varied and the patient is asked to respond when the target is seen. By presenting targets of different luminosity, the average brightness of the dimmest test object that can be seen is determined. This is called the threshold and is determined for different locations in the VF. The standard test-program used in glaucoma patients is the 24-2, in which the target is randomly presented to 54 locations spanning 24°

from the fixation point, i.e. the central focus of our vision. The time needed to perform the 24-2 test, depending on the test algorithm employed, is about 5 to 15 minutes. In VF testing, it is unlikely that a patient with a significant VF defect would have a normal result; however, it is common that individuals with healthy vision record abnormal VF test results. Many factors can contribute to an abnormal VF result, so the specificity of SAP is not as high as clinicians would like (Artes and Chauhan, 2005). At present, VF data must be combined with the ability of experienced clinicians to identify if there is a true abnormality due to glaucoma. The HFA provides reliability indexes (false negatives and false positives) and global indices, such as the mean deviation (MD) and the pattern standard deviation (PSD) to summarise VF test results. The MD is an index based on the size of a VF defect, and is sensitive to generalised loss of sensitivity. The PSD, instead, is sensitive only to localised VF defects. The Loss Variance (LV) index of the Octopus perimeter provides similar information. If indices are significantly different (P <0.05) from normal, this information is also provided by the instrument. The HFA also provides a map of the total deviation of each test point, which represents the difference between the measured threshold and the age-corrected normal value expected for that location. This map is accompanied with a probability plot indicating whether a total deviation value, at a particular test location, is likely to be found in a population of health individuals (four symbols indicate: P <5%, P <2%, P <1% and P <0.5%). Another index provided by the HFA is the Glaucoma Hemifield Test (Asman and Heijl, 1992), which evaluates VF sensitivity differences in either side of the horizontal midline, above or below the point of fixation (referred to as the superior and inferior hemifields). The absolute sensitivities of the VF is also presented as a grey scale interpolated map. A typical printout of VF test results from the HFA is presented in Figure 2.4.

2.3.2 Optic Disc and Retinal Assessment

Optic disc analysis has been used for 150 years as an indicator of glaucomatous damage. OD appearance is included in the current definition of glaucoma and, as already discussed, constitutes the definition of early glaucoma for the American Academy of Ophthalmology (American Academy of Ophthalmology, 2010). To determine structural loss, imaging of the



```
Central 24-2 Threshold Test

Fixation Monitor: Gaze/Blind Spot     Stimulus: III, White         Pupil Diameter: 3.8 mm     Date: 11-17-2003
Fixation Target: Central              Background: 31.5 ASB         Visual Acuity: 20/25       Time: 8:24 AM
Fixation Losses: 3/17                 Strategy: SITA-Standard      RX: +4.25 DS    DC X       Age: 77
False POS Errors:  1 %
False NEG Errors:  0 %
Test Duration: 06:26                                               GHT: Outside normal limits

Fovea: OFF                                                         MD    -11.71 dB  P < 0.5%
                                                                   PSD    11.91 dB  P < 0.5%

            24  24 | 25  27
        26  26  27 | 27  25  24
    24  29  29  27 | 30  23  26  25
25  26  26  28  30 | 31  27  17  26
30 ----------------+----------------- 30
17  10  10   9  24 | 31   6  <0  26
    <0   0   9  12 |  8   0  18  21
        <0   3   8 |  7  12  11
            <0  <0 | 12  17

             -1  -1 |  0   2                        -1  -1 |  0   2
         -1  -2  -1 | -1  -2  -3                -1  -2  -1 | -1  -2  -2
     -4   0  -1  -3 |  1  -6  -2  -2        -3   1   0  -3 |  1  -6  -2  -2
  0  -2  -4  -3  -2 |  0  -4      -2     0  -2  -3  -3  -1 |  1  -3      -2
 -8 -18 -20 -22  -7 |  0 -25      -3    -8 -18 -20 -22  -7 |  0 -24      -3
    -30 -30 -22 -19 |-23 -30 -11  -7       -30 -30 -22 -19 |-22 -30 -11  -7
        -31 -27 -23 |-23 -18 -19               -30 -27 -22 |-23 -17 -18
            -30 -31 |-17 -11                       -30 -30 |-16 -11

        Total Deviation                         Pattern Deviation

:: < 5%
▨ < 2%
▩ < 1%
■ < 0.5%

Glaucoma Progression Analysis

24-2

Likely Progression

See GPA printout for complete
analysis

Baseline Exams:
  09-25-1996    09-30-1998
Previous Follow-up Exams:
  10-23-2001    10-30-2002

▵ P < 5% Deterioration
◭ P < 5% (2 consecutive)
▲ P < 5% (3+ consecutive)
X Out of Range

St. Albans VA Hospital
Linden Blvd and 179th Street
St. Albans, NY 11425
718-526-1000
```

Figure 2.4: Humphrey VF Analyzer output. Source: Zeiss Inc.

ONH is used. Stereoscopic ONH Photography is currently the most used method of ONH analysis, according to (Mardin and Jünemann, 2001). This study found that stereoscopic ONH photography is also the most sensitive single detection method for early glaucoma, although the comparison did not take account different definitions of glaucoma across studies. Clinicians are generally very experienced with evaluating such photographs, however any such assessment is clearly subjective. From ONH photographs, it is possible to extract ONH parameters, such as cup-to-disc ratio, disc and cup area and volume, neuro-retinal rim width and area. In the past, these were obtained by plotting the outline margins of the OD, optic cup and other features by hand, and then calculating areas according to their coordinates using a planimetry computer. In a prospective longitudinal study (Caprioli et al., 1996) on 193 eyes, detection rates of structural glaucomatous change was 15% using qualitative evaluation of ONH stereophotographs, and 3.6% using manual stereoplanimetry of the disc rim area. Computerised systems have emerged lately, sharing several basic features, such as Par Is 2000, Rodenstock ONH Analyzer and the Glaucoma Scope. In the latter, a lamp using near-infrared light is used to obtain a three-dimensional (3D) map of the ONH and a series of related parameters. The experience of the operator and pupil size are significant limitations of this test. Confocal Scanning Laser Ophthalmology (CSLO) offers real-time, 3D imaging of the ONH and RNFL, using the intensity of light reflected off the retinal surface. The Heidelberg Retina Tomograph (HRT) is the only commercially available confocal scanning laser ophthalmoscope. In the HRT, the final topography image is obtained by averaging 3 scans, and contours are traced automatically and presented to easily view the depth of the cup and other tissues, together with ONH stereographic parameters. Among its output, a reflectivity image showing the classification of the different sectors of the ONH (e.g. normal, borderline, etc.) is provided. The HRT also provides the Moorfield Regression Analysis (MRA) score, first proposed in (Garway-Heath, 2005). This is based on the 95% confidence interval of age-corrected differences in the optic disc and the neuro-retinal rim area. However, the method requires the operator to identify the optic disc. Several studies have investigated the performance of CSLO in detecting OD changes (Zangwill et al., 1995; Zangwill et al., 1996), finding that there was good agreement between HRT and clinicians in detecting early glaucomatous disc changes. Sensitivity for detecting patients with early glaucomatous VF loss was equal to approximately 88% at 78% specificity (Mikelberg et al., 1995). A longitudinal study collecting data with HRT II by (Chauhan et al., 2000; Chauhan et al., 2001), found that glaucomatous disc changes determined with the instrument occurred more frequently than VF changes. Most patients with field changes also had disc changes but less than half of those with disc changes had field changes. Similarly, (Kamal et al., 1999) found that sequential optic disc images on the HRT allowed the detection of glaucomatous change before confirmed VF change in a group of OHT patients converting to early glaucoma.

Optical Coherence Tomography (OCT) is a recent technology that uses light in the near infrared range to achieve high-resolution cross-sectional imaging of the eye (Huang et al., 1998). OCT, like CLSO, measures OD areas and RNFL thickness, but with higher accuracy and precision than CLSO. There is high correlation between OCT and HRT measurements and disease status (Schuman et al., 2003).

Retinal Nerve fibre Layer status (RNFL) can be indicative of early glaucoma that is undetected with OD imaging and VF testing (Tuulonen et al., 1993). The RNFL is composed of millions of retinal ganglion cells (fibres), which converge at the OD, and lie in the inner retina. The measurement of the height of the retinal surface can be obtained by tomographic scans using CSLO. This method has been found to have 73% sensitivity for the detection of glaucomatous ON damage (O'Connor et al., 1993), where diagnostic precision was defined as the total proportion of correct diagnoses for the presence or absence of VF loss. However, the low axial resolution and the fact that superior techniques exist make this technique less than optimal for RNFL analysis. OCT can obtain high resolution images of the fundus, from which thickness measurements can be determined by computer analysis. Good correlation of thickness has been found with structural and functional parameters (Schuman et al., 1995). Scanning Laser Polarimetry (SLP) is also used to investigate the RNFL, obtaining thickness measurements in a short time by measuring the summed retardation of a polarised scanning laser beam. The GDx is a scanning laser polarimeter that incorporates a database of normative data. A sensitivity of 58% at a given specificity of 80% was found for preperimetric RNFL defects (Horn et al., 1999), and a sensitivity of 89% and 87% for early to moderate glaucoma (Greaney et al., 2002), where glaucoma was defined as an open anterior chamber with an early to moderate reproducible glaucomatous VF defect. Sensitivity and specificity of 90% and 100% were found using expert operators' classification of VF defects as the gold standard in (Munkwitz et al., 2004) and of 96% and 93% in (Tjon-Fo-Sang and Lemij, 1997). Histological examinations on monkeys found good correlation between SLP retardation values and histological RNFL thickness measurements (Munkwitz et al., 2004). Assessment of the RNFL appears to be more sensitive and specific than estimations of parameters of the ONH. As measurement techniques become more refined and characteristics of the RNFL become better understood, this kind of structural measure may prove to be better than VF testing for diagnosis and follow up in glaucomatous patients and patients at risk of glaucoma (Yanoff and Duker, 2008).

The presented imaging techniques can be used in conjunction with information from clinical examination and functional tests. However, a comparative study concluded that, when used alone, imaging techniques do not provide sensitivities and specificities that justify implementing them as a primary population screening tool for early to moderate glaucoma (Sanchez-Galeana et al., 2001). In (Mardin et al., 1999), HRT sensitivity was just 42% at a specificity of 95% for early glaucoma, rising to 84% for moderate glaucoma. Moderate glaucoma was defined as changes in the OD (on clinical examination) and VF, while early glaucoma was defined as increased IOP, normal VFs, and glaucomatous appearance of the OD as evaluated by photographs. OCT has been shown to have sensitivity of 82% at a specificity of 84% for moderate glaucoma (Greaney et al., 2002), defined by VF testing showing reproducible defects. As observed by (Michelson and Groh, 2001), conventional photographic imaging of the OD by an experienced examiner remains the most sensitive method for detecting early glaucoma. However, experts have also been shown to frequently fail to diagnose early to moderate glaucoma among patients with pertinent risk factors, VF changes, and ON damage (Wong et al., 2003). In (Stein et al., 2005), the use of multiple imaging devices together improved the specificity of the diagnosis. The use of these new technologies arguably hold potential for glaucoma screening, especially in combination with functional testing.

2.4 Classification Techniques for Glaucoma

In clinical settings, the techniques most used to assess the onset of glaucoma are strictly related to clinical knowledge, and they mainly subjectively assess VF testing according to OD examinations. Other factors are also taken into account by the clinician, such as IOP and other ocular, systemic and demographic factors. However, the interpretation of quantitative imaging and functional test measurements is problematic, especially regarding early detection of glaucoma (Chauhan et al., 1999). The Glaucoma Advisory Committee of Prevent Blindness America states that devices used in screening for glaucoma should attain at least 85% sensitivity at 95% specificity (preferably 98%) for moderate to severe VF defects (Committee, 1996). In a study conducted in 1994 among 120 people, a sensitivity of 67% and 51% at specificities of 65% and 83% was found by 2 trained glaucoma specialists who categorised VFs based on sensitivity, patient information and other indices (Goldbaum et al., 1994). The reference was based on OD examination. Other experiments obtained different results: in (Chauhan et al., 1990), perimetry indices obtained a sensitivity of 64% at 74% specificity, using the VF inspection by three investigators as a reference. For OD assessment, as already described, sensitivities are generally higher. However, variability in measurements and populations must always be kept in mind when comparing different studies. The next section gives an overview of the ML approaches that have been used for glaucoma and provides a comparison of such techniques across the literature.

2.4.1 Machine Learning Approaches

Many studies have been conducted in order to find automated techniques to help clinicians in the classification of glaucoma. In particular, statistical and ML approaches have been explored in this context. Among the former there is the STATPAC package, which is integrated with the HFA and provides simple deviation from average ”normal” values. The HFA also computes the AGIS score metric. Regarding the OD assessment, the HRT provides MRA and other deviation scores. These metrics, described in the previous sections, are universally used to help clinicians in the interpretation of the VF and OD, but are still insufficient for diagnosis of early glaucoma. The ML approaches use sophisticated algorithms to obtain a classification output given an a set of measures from the patient (input). The algorithm learns the classification rules using training data to be able to map the input to the output correctly. The learning process can be supervised or unsupervised. In the first case the training data is previously associated with a state (healthy or glaucomatous eye), and the classifier tries to minimise mistakes. Then it can be used to classify a new set of measures in a different dataset, or even to predict outcomes. Unsupervised learning instead uses unlabelled data, thus the classifier tries to set the rules to discriminate between different subsets of the training data. It can be used to find data clusters and patterns, which can be compared to expert-knowledge or even explored in more depth. Unsupervised learning can be used for classification and prediction too. It must be noted that an important difference between these two approaches lies in the labelling process. Supervised classifiers are in fact dependent on the process of pre-classification of the training data, which can be biased. In particular, as already described, glaucoma data is often difficult to classify because no gold standard is available, thus the impact of the labelling process should be taken into account. Another important difference in Machine Learning Classifiers (MLC) concerns how to solve decision problems. In one case, the algorithm finds a function to map the input $\mathbf{x}$ directly to a class label. In this case, probability plays no role. Another approach is to first determine the posterior class probability $p(C_k \mid \mathbf{x})$ and then use decision theory to assign each new x to one of the classes. Approaches that model the posterior probability directly are called discriminative models. A more complex approach is to first determine the class-conditional densities $p(\mathbf{x} \mid C_k)$ individually for each class C_k and also infer the prior class probabilities $p(C_k)$. Then, using the Bayes theorem, the posterior class probabilities $p(C_k \mid \mathbf{x})$ can be found:

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)p(C_k)}{p(\mathbf{x})} \tag{2.1}$$

Approaches that explicitly or implicitly model the distribution of inputs and outputs are known as generative models, because by sampling from them it is possible to obtain synthetic data points in the input space (Bishop, 2006). Discriminative approaches are focused on classification accuracy but provide less insight into the structure of the data space, while generative approaches, by allowing the marginal density of data $p(\mathbf{x})$ to be estimated, focus on modelling the data completely and allow more insights to be gained. However, it can be wasteful of computational resources and excessively demanding to find the joint distribution $p(\mathbf{x}, C_k)$. Thus the debate is still open, and there has been much interest in exploring the merits of generative and discriminative approaches to machine learning, and in finding ways to combine them (Jaakkola and Haussler, 1999). Following the rationale of exploring the nature of glaucoma through data, this thesis focuses on probabilistic graphical models, which belong to the generative approach and allow both supervised and unsupervised learning. However, in the remainder of this chapter the main ML models that have been applied in the glaucoma context and their performances will be presented.

2.4.2 Machine Learning Techniques for Glaucoma

Several Machine Learning algorithms have been used with glaucoma data. The techniques will be briefly discussed here and their performances on published works will follow.

The Classification and Regression Trees (CART) algorithm (Breiman et al., 1984) is a widely used statistical technique for producing classification and regression models with a tree-based structure. For classification, CART uses a tree structure, consisting of a hierarchy of univariate binary decisions. Each internal node specifies a binary test on a variable which splits the data into two disjointed branches, and each leaf node specifies the class label. The structure of the tree is derived from the data in a complicated pruning process, using the misclassification loss function as a score function. The misclassification loss function is minimised using a cross-validation technique, which is a method that partitions the training data into a subset for building the tree (i.e. learning data) and then estimates the misclassification rate on the remaining validation set. CART uses a greedy local search method to identify good candidate tree structures, recursively expanding the tree and then pruning branches. Even if the CART technique can handle a large number of variables and contains mixed data types, its decision boundaries are constrained to be parallel to the input variable axes (Hand et al., 2001). Other data mining algorithms based on tree models exist, and a widely used alternative to CART is the C4.5 algorithm (Quinlan, 1993). Instead of using a cross-validated score function, it adjusts the estimated error rate on the training data to approximate the test error rate.

Multilayer Perceptron (MLP), also termed a feed-forward neural network, is a common supervised learning classifier applied to a wide class of problems such as face recognition (Samal and Iyengar, 1992) and optical character recognition (Baird, 1993). It is a particular class of neural networks (NNs), which in general constitute an approach to classification that fixes the number of the basis functions (i.e. classification functions on the input to obtain the output) but allows them to be adaptive during training, in an attempt to avoid the curse of dimensionality (Bellman, 1961). In particular, MLP architecture is a feed-forward network with input and output layers separated by hidden layers of nodes. The links between nodes represent weight parameters, which are calculated on the data and connections between layers. Learning involves multiple bidirectional passes through the layers of the network, and the error signal is passed backward (i.e. error-backpropagation) in order to reinforce or inhibit each weight assigned by the hidden layer. The cycle is repeated until the error rate is low compared with the labelled examples in the training set (Bowd and Goldbaum, 2008). The output value is obtained with non-linear transformation (activation function) on the inner products in the internal layers and, because of the non-linear nature of the model, the decision boundaries between classes produced can also be non-linear (Hand et al., 2001). The MLP is the most popular architecture in neural networks and have proven to be of greatest practical value (Bishop, 2006) even if it requires repeated training from different random initial conditions and convergence to a global solution is not guaranteed (Chan et al., 2002).

Self-Organizing Maps (SOMs) (Kohonen, 2001) is another type of NN used with unsupervised learning, which clusters high dimensional data thus decreasing its complexity. The SOM forms a nonlinear projection of a high-dimensional data manifold on a regular, low-dimensional (usually 2D) grid, in which the clustering of the data space is clearly visible. The variables investigated can be displayed separately using grey-level or pseudocolor coding on the SOM grid, so that it is possible to understand the mutual dependencies between them and the structures of the data set (Kohonen et al., 2004).

Support Vector Machine (SVM) algorithms (Vapnik, 2000) are used with success for solving a variety of classification and regression problems in different fields (Furey et al., 2000; Dumais et al., 1998). These models non-linearly map training data (i.e. labelled cases) to high-dimensional space where an optimal hyperplane (a linear boundary) can be found to minimise the separation error with lower computational burden. Further, just a subset of the training data containing the samples which are the most difficult to classify is used to find the hyperplane, instead of using all the samples as in the linear learning classifiers. Reflected back to the input space, the decision surface is not constrained to linear separation anymore, obtaining a curve which is well tuned to the samples in the teaching set (Bowd and Goldbaum, 2008).

The Relevance Vector Machine (RVM) classifier is also based on fewer sample vectors, and typically results in much sparser models (Chan et al., 2002). The benefit of the latter property is that its results are more generalisable, decreasing the chances of over-fitting. Further, the output of the RVM is the probability of class membership, in contrast to the non-probabilistic value obtained with SVM (Bowd et al., 2008), although this is at the expense of non-convex optimisation during training. A probability output allows a more intuitive expression of uncertainty in the prediction. Methods have been developed to obtain SVM probabilistic outputs (Chan et al., 2002).

The Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) classifiers use supervised learning in order to classify data into categorical values (e.g. healthy or glaucomatous). They divide the input space into decision regions whose boundaries are called decision surfaces (Bishop, 2006). In both LDA and QDA, Gaussian class-conditional densities $p(\mathbf{x} \mid C_k)$ are assumed. However, in the LDA these densities share the same covariance matrix, so the discriminant function obtained is a linear function of input $\mathbf{x}$, while in the QDA these assumptions are relaxed, obtaining a quadratic discriminant.

Mixture of Gaussians (MoG) and Mixture of Generalized Gaussian (MGG) are mixture of ”expert” networks (in this case Gaussian estimators) that provide a Bayesian-based probabilistic output based on labelled input (Tresp, 2001; Bowd et al., 2008). As for the generative approach, we need to model the individual class-conditional densities $p(\mathbf{x} \mid C_k)$

to obtain the posterior probability and build the classifier. Since the input of the glaucoma data contains continuous variables, we can model each $p(\mathbf{x} \mid C_k)$ by a normal multivariable density: this would result in a classical LDA or QDA. However, often the distributions of these variables are not well represented by a single Gaussian distribution, and a more flexible solution would be useful in this case. Using simple nonparametric methods such as the histogram method (Hand et al., 2001) is not the best choice, since it subjects the classifier to the curse of dimensionality, even if, according to (Bishop, 2006), the nonparametric Bayesian methods are attracting increasing interest. Instead, the Mixture of Gaussians approach is a semi-parametric model that uses mixtures of normal densities to describe more complex distributions, i.e. for each variable we will have

$$p(\mathbf{x}) = \sum_{m}^{M} p(\mathbf{x} \mid m) P(m) \tag{2.2}$$

where for each cluster m, $p(x \mid m)$ is a Gaussian distribution with parameters $P(m)$. The clusters could be thought as each modelling (i.e. fitting) different regions of the input space, with the goal of improving generalisation (Bowd et al., 2008). The MGG regards the case when two Gaussians are not the best choice to fit a density. It uses the same mixture model as in 2.2, but this time each distribution is now described by a linear combination of non-Gaussian random variables. The parameters are adapted during training, which is done by gradient ascent on the data likelihood (Lee and Lewicki, 2000). The Subspace Mixture of Gaussians (SSMoG) is a dimension-reducing MoG method that iteratively samples, with replacement, multiple random subsets of the full-dimensional dataset, allowing the use of MoG classifiers on high dimensional data (Bowd et al., 2008).

The Parzen Window is a kernel-based nonparametric approach to density estimation (Parzen, 1962), which has been used with a particular isotropic Gaussian Parzen Window function, in analogy with MoG, in order to classify glaucomatous and healthy data (Chan et al., 2002).

Another nonparametric method, which is widely used, is the K-nearest-neighbour method (KNN), which uses a smoothed Gaussian function with an optimal SD as a kernel function. To estimate the distribution on a point, a small sphere centered on the point is considered to grow until it contains a certain amount of data points K, which actually governs the degree of smoothing. To use KNN for classification, the density estimation is applied to each class separately and then the Bayes' theorem is used. The so called nearest-neighbour rule is the particular case of K=1, meaning that a test point is simply assigned to the same class as the nearest point from the training set.

An interesting and useful task in ML techniques for classification is the so called feature selection, i.e. the search for the subset of input variables that contribute most in the classification. Since exhaustive search with the combination of all possible subsets of variables is prohibitively expensive, other approaches are used. For example, in (Chan et al., 2002), forward selection and backward elimination are used to rank the variables; in the former variables are added one at a time, choosing the next variable that most increases classification. Feature selection may be a valuable way to speed up and enhance test procedures for different ML techniques, as well as improving classification accuracy and physiological understanding.

Another class of models that has been used in the glaucoma context is the probabilistic graphical models. Instead of just using purely algebraic manipulation to solve probabilistic models, probabilistic graphical models are useful to augment the analysis using diagrammatic representations of probability distributions. They offer several useful properties: first, they provide a simple visualisation of the structure that can be used to design and motivate new models. Second, many insights into the properties of the model, such as conditional independence properties, can be seen in the graph. Third, complex computations required can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly (Bishop, 2006). Among these models, the Bayesian Networks (BNs) are directed graphical models in which each node represents a variable, and a lack of arcs between nodes represents a conditional independence assumption. Once the structure and the parameters describing the variables in the model are set using a training dataset or arbitrarily, it is possible to infer about any value of any node. For classification, BNs can be used to infer the group of the patient by taking into account the value of the observed variables. Missing data and unobserved variables can also be included in BN models.

2.4.3 Performance of Machine Learning Classifiers

There have been many studies comparing the performances of different ML techniques applied to glaucoma. Most of them rely upon VF measurements, while some of them rely upon OD measurements or a combination of structural and function data. Comparisons are usually obtained by considering Receiver Operator Characteristic (ROC) curves (Bradley, 1997). The ROC curve consists in plotting the sensitivity of the test (in our case the proportion of correctly classified glaucomatous cases, or "True Positives") versus 1-specificity (in our case the proportion of normal eyes misclassified as glaucomatous, or "False Positives"). The plot shows these values for different choices of threshold, increasing from 0 to 1. The threshold is the value to which the estimated probability $p(C \mid x)$ is compared to when classification is made. To compare different ROC curves, the area under the ROC curve (AUROC) and its variance can be calculated, summarising the quality of classification over a wide range of misclassification costs (Bradley, 1997).

Regarding VF data, Table 2.1 reports the main studies on classification using ML techniques. Despite the different absolute performance results, which depend on several factors such as inclusion criteria or gold standard assumptions, ML techniques perform better than state-of-the-art solutions in all studies. MLPs, QDAs and MoGs perform particularly well with SAP data. Clustering was also assessed using SOM and more flexible models, obtaining interesting insights and promising results. MLCs have also been applied to imaging measurements of glaucoma. Imaging techniques and tomographies can provide enormous arrays of data and the use of MLC is then an obvious approach. The results of such studies on cross-sectional datasets are presented in Table 2.2. Again, the potential of ML techniques are shown in these studies. However, in this case, MLP, SVM and RVM techniques perform similarly. Note that the variability in glaucoma definition and number of subjects included in the studies make it difficult to directly compare classifiers across the studies. When OD and VF data are combined, the studies in Table 2.3 indicate that classification performances further increase. BNs have had relatively limited application to glaucoma data, despite their

Table 2.1: ML studies on cross-sectional SAP data (Supervised and Unsupervised).

Reference	ML Technique	Subjects	Criteria for glaucoma	Results (sensitivity and and specificity)
(Goldbaum et al., 1994)	MLP, Experts	60 G, 60 H	IOP > 22 mmHg and GON (cup-disk ratio > 0.7 and experts' evaluation)	MLP obtained 60%/72%, which was similar to experts results.
(Lietman et al., 1999)	MLP, STATPAC, AGIS, NTG score	106 G, 249 H	IOP > 21 mmHg, reproducible VF defect	MLP better at higher specificities but worse at lower specificities.
(Chan et al., 2002)	MLP, SVM, LDA, QDA, STATPAC, Parzen Window, MoG, MoGG	156 G, 189 H	Analysis of OD photographs and ocular history	Best QDA (78%/76%). Generative better than discriminative. All ML better than STATPAC.
(Goldbaum et al., 2002)	MoG, MoGG, MLP, SVM, Human	156 G, 189 H	Consensus masked evaluations of OD photographs	MoG highest, better than STATPAC. All ML better than human.
(Bizios et al., 2007)	MLP, PSD, STATPAC	100 G, 116 H	Expert evaluation of OD photographs	MLP accuracy of 93.5%, against 85.6% to 91.7% for global indexes.
(Henson et al., 1996)	SOM	560 G	VF deviation < -6 dB	Found 50 clusters with 7 typically observed patterns.
(Sample et al., 2004)	Bayesian Mixture of Factors	189 H, 156 G	Consensus masked evaluations of OD photographs	Found 5 clusters, 98% of healthy eyes were assigned to one cluster.
(Goldbaum et al., 2005)	Variational Bayesian-independent component analysis mixture model	189 H, 156 G	Consensus masked evaluations of OD photographs	31% of healthy eyes were assigned to one cluster.

Table 2.2: ML studies on cross-sectional OD data (Supervised and Unsupervised).

Reference	ML Technique	Subjects	Criteria for glaucoma	Results (sensitivity and specificity)
(Uchida et al., 1996)	MLP, Linear analysis (CSLO data)	53 G, 43 H	OD appearance consistent with early VF defects	MLP AUROC of 0.94, better than the others.
(Bowd et al., 2002)	MLP, SVM, LDA (CSLO data)	108 G, 189 H	VF defects (based on PSD and GHT)	Comparable (85% - 91%/90%), AUROC 0.94 - 0.97.
(Bowd et al., 2005)	SVM, RVM (SLP data)	92 G, 72 H	VF defects (based on PSD and GHT)	Same AUROC (0.94), better than standard SLP parameters (AUROC 0.51 - 0.87).
(Huang and Chen, 2005)	MLP (OCT data)	89 G, 100 H	IOP $>$ 22 mmHg and early VF defect	MLP (82%/90%), AUROC of 0.82.
(Burgansky-Eliash et al., 2005)	SVM (OCT data)	47 G, 42 H	VF defects (based on PSD and GHT) and GON (cup-disk ratio $>$ 0.6 and experts' evaluation)	SVM (92%/95%), AUROC of 0.98.
(Huang et al., 2007)	SOMs, decision trees (OCT data)	71 H, 64 G	IOP $>$ 22 mmHg and reproducible VF defect	SOM identified the two of five most informative clusters, derived decision tree (92%/83%).

characteristics that make them well-suited to glaucoma diagnosis. In (Tucker et al., 2005), a BN classifier was built in an attempt to give a deeper understanding of glaucoma, and it was then compared to traditional black-box MLCs. The network was learned using every point of the VF data as a variable, and applying a quasi-greedy searching technique, called Simulated Annealing (SA), to explore the space of structures. For every iteration of the algorithm, three spatial operators (add a link to a near node, change the parent of a link

Table 2.3: ML studies on cross-sectional OD and VF data.

Reference	ML Technique	Subjects	Criteria for glaucoma	Results (sensitivity and specificity)
(Brigatti et al., 1996)	MLP (CSLO data)	185 G, 54 H	Early VF defects	MLP obtained 87%/56% when trained only on OD data and 90%/84% when trained also on VF data.
(Goldbaum et al., 2004)	RVM, different types	SAP: 156 G, 189 H; HRT: 95 G, 135 H	SAP: evaluation of OD photographs, HRT: repeatable VF damage	SAP with HRT criteria (76%/90%) better than HRT with SAP criteria (83%/90%).
(Mardin et al., 2006)	LDA, classification tree	88 G, 88 H	VF defect (MD > 2.8 dB) and evaluation of OD appearance	Combination of HRT and VF data overall better (95%/91%).
(Bowd et al., 2008)	RVM, SSMoG, STATPAC	156 G, 69 H	Repeatable VF defects or GON	RVM and SSMog (75%/90%), better when OCT and SAP data combined (AUROC 0.84 and 0.86) than OCT alone (0.80 and 0.81) and SAP alone (0.81 and 0.84).
(Racette et al., 2010)	RVM, LDA	GON: 144 G, 68 H	VF defects or GON	RVM on combined HRT and SWAP data, better (AUROC 0.93) than LDA and than RVM on HRT (0.88) and SWAP (0.76) alone.

and swap the relative position of two nodes' parents) were applied to the existing structure. Subjects included 78 subjects with early VF loss (based on AGIS score) and 102 normal volunteers (AGIS score of 0) known not to be sufferers. After categorisation of the VF data into four states, 10-fold cross-validation was applied to the BN classifier learned and to a Naive Bayes classifier, a Tree Augmented Network (TAN), a linear regression classifier and a KNN classifier (where K was chosen using 10-fold cross validation). Results between Bayesian Classifiers found that the BN performed best with an AUROC of 0.94 (sensitivity was around 85% at 90% specificity), while the next best was the Naive Classifier with an AUROC of 0.9. The TAN performed worst (AUROC of 0.8). Compared to more traditional classifiers, the BN classifier performed better than the KNN method (AUROC of 0.91) and was comparable to linear regression (AUROC of 0.94). However, the great advantage given by the BN is the transparency of the model, which allowed an explicit modelling of relationships between variables. In order to assess the reliability of such learned relationships between points, expert knowledge concerning the anatomy of the eye was used: mean optic nerve-head angle distance between linked nodes, which is supposed to be low, was 15.28 degrees (over 180). Further, VF points linked directly to the class node were analysed, as a form of feature selection, and these reflected a common feature of the early stages of glaucoma (the nasal step VF defect, Figure 2.3). Other unexpected patterns were discovered, showing the potential of BN models in learning links that can teach interesting patterns in clinical data. In (Tucker et al., 2003), 623 VFs from 24 converters patients (i.e. changing from healthy to glaucomatous state, according to AGIS score) were analysed similarly to the above discussed work. The BN classifier obtained a sensitivity of 63 % at a specificity of 80%, the Naive Bayes classifier obtained 60% sensitivity at 81% specificity the TAN yielded 60% sensitivity at 79% specificity.

To our knowledge, MLCs have been used to detect glaucomatous progression using mostly SAP data. Results are reported in Table 2.4.

A predictive approach with MLCs has been also implemented. In (Bowd et al., 2004) an SVM classifier was used to predict future conversion to a glaucomatous VF in suspect eyes. Data used for baseline cut-off were CSLO measurements from a previous study, subsequently followed longitudinally. Suspected eyes classified as abnormal by SVM were 1.7 times more likely to develop three consecutive VF defects than eyes classified as normal, which was similar to predictive values of expert assessment of OD photographs and MRA in CSLO (Garway-Heath, 2005). In a recent study (Demirel et al., 2009), a classification tree (CART)

Table 2.4: ML studies on glaucoma progression.

Reference	ML Technique	Subjects	Criteria for glaucoma	Results (sensitivity and specificity)
(Brigatti et al., 1997)	NN	233 G	VFs evaluated by experts	Sensitivity of 73% at 88% specificity.
(Sample et al., 2002)	SVM, MoG, STATPAC	114 G, trained on 156 G (Goldbaum et al., 2002)	IOP > 23 mmHg, trained on GON	36 converters based on STATPAC. Agreement STATPAC-MLC 88% to 95%, but MLCs predicted abnormalities earlier.
(Lin et al., 2003)	NN	80 G over 7 years	AGIS	AUROC 0.92, (86%/80%) and (91%/90%).
(Sample et al., 2005)	VB-ICA-mm model	191 G over 6 year	OD appearance (GON) and VF defects	31% converters, better than AGIS.

was applied to VF age-adjusted data of 100 healthy individuals and 168 patients with high IOP or early glaucoma, in order to predict progressive glaucoma optic neuropathy, i.e. glaucomatous RNFL appearance. Classification tree models suggested that patterns of baseline VF findings were predictive of progressive glaucoma optic neuropathy with sensitivity of 65% at specificity of 87%, which may indicate that spatial MLCs can be used to determine which VF locations are more predictive of poorer prognosis of progressive glaucoma optic neuropathy.

2.4.4 Modelling Glaucoma Progression

Temporal relationships in progressive glaucoma data are seldom modelled. Although several studies (Heijl et al., 1990; Hood and Kardon, 2007) have used statistical models to explore and predict the temporal aspects of VF data, MLC classifiers have been applied in very few studies. In (Ibanez et al., 2007) a statistical spatio-temporal model of healthy eyes was built for forecasting and simulation of healthy VFs. In (Swift and Liu, 2002), multivariate time series were modelled with a novel algorithm for large variable datasets to predict and model VF deterioration. Regarding longitudinal data, the use of BNs has only been partially explored, although graphical models have been assessed. In (Fitzke et al., 1996), a new method of analysis using a graphical display of longitudinal field data is presented. In this method, linear regression models of sensitivity at different locations in the VF over time are coloured to illustrate the changes and aid the interpretation of field loss. The model was compared with STATPAC 2, obtaining a good level of agreement. In (Anderssen and Jeppesen, 1998), VF data were investigated with State Space Models (SSMs). SSMs are models where latent variables are introduced to obtain a more general framework, while still retaining tractability with sequential data. In fact, temporal sequences must be "limited" to ensure that the amount of data produced is manageable, as the sequence increases, and to take into account the correlation between subsequent values. This leads to the consideration of the Markov Models (MM), in which it is assumed that future predictions are independent of all but the most recent observations. However, while MMs are severely limited, SSMs provide more general models, and can be readily characterised using the framework of probabilistic graphical models. Examples of such State Space Models are the Hidden Markov Model (HMM), in which the latent variables are discrete, and Linear Dynamical Systems (LDS), in which the latent variables are Gaussian (Chan et al., 2002). In this context, Dynamic Bayesian Network (DBN) is an extension of BNs that can model sequential data (Friedman et al., 1998). Inference is very similar to standard inference in BN (Dagum et al., 1992), and different methods for learning from data have been proposed. DBN can be seen as a generalisation of HMM and LDS, by representing the hidden (and observed) state in terms of state variables, which can have complex interdependencies (Murphy, 1998). SSMs and in particular DBNs are likely to be valuable techniques for modelling glaucoma data, because of their spatial and temporal nature. In (Tucker et al., 2005), a BN classifier, already partly presented in this chapter, is combined with temporal data to produce a DBN classifier, which is able to model and classify data with spatial and temporal relationships. In this case, in addition to three spatial operators, three temporal operators were applied iteratively to find the best scored structure for the classifier. The three non-spatial operators were:

adding a link, removing one and changing the parent of an existing link. The learned spatio-temporal classifier (STC) was then compared with other BN classifiers (TAN and Naive Bayes), KNN and linear regression. Among BN classifiers, the worst performers were TAN and Naive Bayes (AUROC of 0.79 and 0.77), while STC with 3 parents was the best (AUROC of 0.88). ROC curves show that linear regression performs similar to STC and KNN when the misclassification cost is higher than 0.6. Generally, for diseases that have a low frequency and a slow progression, the misclassification cost lies on the higher values (Goldbaum et al., 2002); this is also the case for glaucoma. It must be noted that VF clinical classification requires three consecutive abnormal fields, but MLC do not use such conservative criteria. Further, VFs are not re-classified after conversion, as patients cannot become healthy. However, external interventions (e.g. medication) may result in negative output from MLCs, resulting in FNs. In this study, the graphical model of the STC is also evaluated, showing that the effect of noise in the data is more present in the STC than in the non-temporal BN. However, many autoregressive and some others temporal links were found, justifying the better performances of the STC over the BN. Interesting insights are also derived from the temporal structure, and the suggestion of data integration and metrics definition is provided in the study.

2.5 Summary

This chapter has shown that glaucoma, an irreversible form of blindness, is a major public health problem. This situation is likely to worsen in the future and further studies need to be carried out to ensure that as few sufferers as possible succumb to visual disability from this condition. As the American Optometric Association (American Optometric Association, 2010) highlights, the only risk factor that can be clinically modified is IOP. Although IOP is a poor diagnostic tool for glaucoma, lowering it through medication has been proven to be beneficial in preventing glaucoma progression (Collaborative Normal-Tension Glaucoma Study Group, 1998; The AGIS Investigators, 2000; Heijl et al., 2002; Leske et al., 2003; Kass et al., 2002). Early diagnosis is essential in this context.

With regard to the available screening techniques, considerable discordance exists been different results from structural and functional tests in glaucoma, especially in relation to its progression over time. Whether ON damage or VF damage appears first seems to be unclear, probably because it is a co-occurrence of both, and because many factors influence which is perceived to occur first. These results once again point out the importance of using both functional and structural techniques to monitor the onset of the disease and its progression (Mackenzie and Cioffi, 2008). In fact, combining different screening techniques seems to hold the potential to improve classification performances, as well as give insights into the disease.

In this context, ML approaches have shown excellent results with OD and SAP data in cross-sectional studies for classification of glaucoma. Their characteristics and performance with combined functional and structural cross-sectional data suggest that further studies concentrating on MLCs should be carried out. However, little modelling has been done on longitudinal data using ML techniques. Relevant to this, (Bowd and Goldbaum, 2008) suggest that future exploration should include large longitudinal datasets, with techniques able to generalise and combine structural and functional measurements. One approach that seems to have this potential is definitely the BN model, which allows large datasets of different data types and expertise knowledge to be combined. Its intuitive output and the ability to model causal relationship between variables explicitly is highly useful to clinicians. Extensions of BNs can be used to incorporate the temporal aspects of the data and improve classification performance as well as help researchers to learn more about the mechanisms behind glaucoma.

The rest of the thesis is focused on such models, which will be presented in more depth in the next chapter. Cross-sectional and longitudinal datasets will then be used in Chapters 4 and 5 to explore and assess BN extensions for glaucoma. In particular, ad-hoc models will be used to tackle data integration, variability, and biases in the data in order to improve diagnostic performance and gain insights in the nature of the disease.

Chapter 3

Exploring Glaucoma with Bayesian Networks

Bayesian Networks (BNs) are probabilistic graphical models that provide intuitive data representation and perform classification. As discussed in the previous chapter, they may be particularly appropriate for investigating many questions related to glaucoma diagnosis and monitoring. This chapter introduces the fundamentals of the BN framework and explores the potential of these models in the context of glaucoma.

In Section 3.1, the datasets that are investigated in the thesis are presented. Section 3.2 illustrates the basics of inference and learning using a simple BN. In Section 3.3, two BNs based on functional and anatomical data and one BN based on prior anatomical knowledge are built and assessed. Finally, Section 3.4 summarises the results and introduces following chapters of the thesis.

3.1 Data Investigated

For this research, three independent datasets were available (Table 3.1). Datasets A and B both consist of 52-point VF raw sensitivity values (i.e. not age-corrected) obtained with the Humphrey Field Analyser (HFA) II and retinal data from the Heidelberg Retinal Tomograph (HRT) (Fig. 3.1).

Table 3.1: Characteristics of the available datasets

	Dataset A	Dataset B	Dataset C
Healthy Subjects (values)	102 (102)	19 (155)	-
Converters (values)	78 (78)	43 (474)	-
Total Subjects (values)	180 (180)	62 (629)	51 (912)
Mean Age (Healthy)	67.6 (57.5)	65.7 (66.7)	66.9

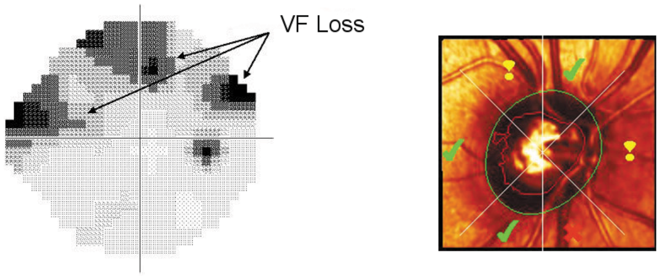


Figure 3.1: VF Test output for the right eye with VF Loss (left) and HRT output image showing six sectors of the optic disc with possible defects in three sectors (right).

Retinal data from the HRT consists of a number of measurements in six sectors of the Optic Disc (OD). Retinal data was pre-processed by applying the 95% prediction interval *Moorfield Regression Analysis* (MRA) indicated in (Garway-Heath, 2005; Wollstein et al., 1998). As described earlier, this regression equation is a linear combination of Age, Optic Disc Area (ODA) and Retinal Rim Area (RRA) into one single parameter. MRA is grouped into six sectors as suggested in (Garway-Heath et al., 2000) for computational and anatomical reasons.

Dataset A is a cross-sectional dataset of 78 early glaucomatous patients and 102 healthy control subjects. The definition of glaucoma was based on an Intraocular Pressure (IOP) greater than 21 mmHg and clinically evident early VF defects repeatable over at least three visits. Healthy people had IOP $<$ 21 mmHg and a normal VF defined according to the Advanced Glaucoma Intervention Study (AGIS) classification (Gaasterland et al., 1994).

Early VF defects were defined as those with a score of 5 or less in the AGIS classification. Patients included had visual acuity greater than 20/40, refraction less than $6D$ ametropia, no recent ocular trauma or surgery or other posterior segment ocular pathology, and no history of diabetes. Healthy patients also had to have no history of primary open-angle glaucoma in a first-degree relative.

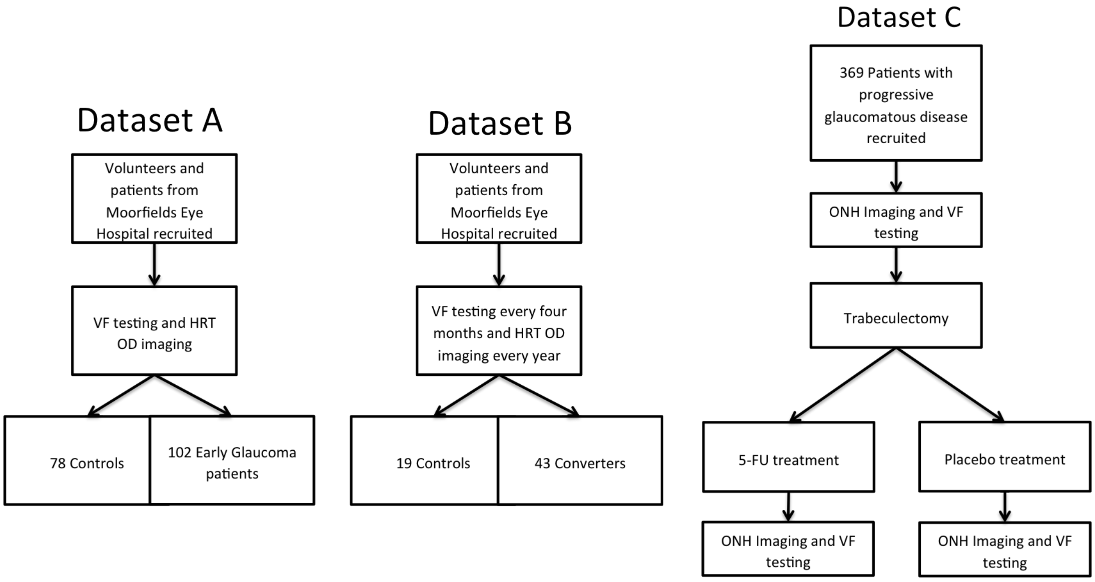

Dataset A
Volunteers and patients from Moorfields Eye Hospital recruited
VF testing and HRT OD imaging
78 Controls
102 Early Glaucoma patients
Dataset B
Volunteers and patients from Moorfields Eye Hospital recruited
VF testing every four months and HRT OD imaging every year
19 Controls
43 Converters
Dataset C
369 Patients with progressive glaucomatous disease recruited
ONH Imaging and VF testing
Trabeculectomy
5-FU treatment
Placebo treatment
ONH Imaging and VF testing
ONH Imaging and VF testing

Figure 3.2: Flowcharts that illustrates how each sample was created from each cohort for Datasets A, B and C.

Dataset B is a longitudinal dataset of 43 patients from an ocular hypertension treatment trial (Kamal et al., 1999), who developed glaucoma in the time span observed (”converters”) and 19 healthy subjects. Initial eligibility criteria for the ocular hypertensive subjects were IOP $>$ 21 mmHg on two or more occasions, and two normal consecutive VF test (AGIS score of 0), absence of any other ocular disease and age higher than 35 years. Conversion was defined as an AGIS $>$ 1 on three consecutive tests, but with no stipulation for the AGIS score in subsequent VFs. Healthy subjects had an AGIS score of 0 in two baseline tests and IOP $<$ 21 mmHg, no ocular disease, no family history of glaucoma or ocular hypertension and age higher than 35 years; there was no stipulation for the AGIS score in subsequent

VFs. Mean length of the series is 10.14 VFs, ranging from 4 to 18.

As highlighted in the thesis, datasets A and B used similar AGIS criteria to indicate glaucoma, and optic disc appearance was not a restriction criterion. IOP thresholds used were the same for both datasets.

Dataset C is a longitudinal dataset of 51 patients selected from the MoreFlow Medical Research Council 5-Fluorouracil (5-FU) study (Kotecha et al., 2009). This was an 80-month prospective randomised controlled trial on the efficacy of pre-operative 5-FU treatment on primary trabeculectomy. Patients who showed progressive glaucomatous disease despite maximally tolerated medical therapy were listed for trabeculectomy and randomized to receiving either per-operative 5-FU or placebo. The study cohort consisted of 369 eyes of 369 patients. As part of the study protocol, the patients had ONH imaging performed before surgery, at 3-months after surgery, and then at annual intervals for 3 years. At this point, a change in the study protocol was introduced that allowed imaging every 6-months. VF examination was performed before surgery and at 3-month intervals for the first year and at 4-month intervals thereafter. Intraocular pressure measurement was performed before surgery and at every trial visit (five visits in the first 3 months after surgery, then at 3-month intervals for the first year and at 4-month intervals thereafter). The follow up period was 5 years. VF test data was filtered by excluding series with less than 10 VFs and without corresponding HRT data. The resulting data consisted of 912 VFs in total; the mean number of VFs per patient was 17.88, ranging from 10 to 28. Raw VF sensitivity data was grouped into six sectors according to the functional-structural map proposed in (Garway-Heath et al., 2000).

The selection of the functional and anatomical variables from the main datasets followed mainly the availability of such variables across all datasets, as well as the indications from consolidated knowledge in the field. VF sensitivity is the most used classification input data across the literature, while for anatomical data the most used variables are OD area, Rim size and Cup size, as described in sections 2.3 and 2.4. Regarding the anatomical measurements, the MRA was chosen as it combines in a convenient parameter OD area, rim size and age. Regarding the summarized sectorial VF variables, the sensitivity was averaged across the six sectors indicated in (Garway-Heath et al., 2000). This is a widely accepted functional map based on the anatomical characteristics of the retina, as described in section 2.2.2. The map is based on the anatomy of the retinal nerve fibre bundles, so that the points in each sector lie on the same nerve fibre bundle and are more likely to be correlated between each other. The values of each sectorial variable are obtained by averaging the pointwise sensitivity values, so that only six values are obtained for each VF test. In this work, this operation was performed to reduce complexity of the calculations needed in certain models. The retinal sectorial variables were provided directly by the HRT instrument. While IOP could also be included, as discussed in future work, it wasnt included in the analysis as this work focuses on structural and functional measurements.

Figure 3.2 shows the flowcharts that illustrates how each sample was created from each cohort. More details on the available data are reported in Appendix A.

A set of simulated datasets were also used for validation purposes. These were obtained using ad-hoc models that will be described in the relevant sections.

3.2 Introducing Bayesian Networks

This research aims to explore and extend the use of BNs for glaucoma modelling. As demonstrated in the previous chapter, BNs have excellent properties relevant to glaucoma modelling, and offer many advantages over other AI and graphical models (Heckerman, 1995; Friedman et al., 1997): first, they can readily handle incomplete datasets because they offer a natural way to encode correlations between input variables. Second, BNs allow one to learn about causal relationships, which is extremely useful in clinical research, such as glaucoma, when we are trying to gain an understanding about a problem domain. They also allow us to make predictions in the presence of interventions. Third, BNs combined with Bayesian statistical techniques facilitate the combination of domain knowledge and data, having causal semantics that makes the encoding of causal prior knowledge particularly straightforward. Fourth, BNs offer an efficient and principled approach for avoiding the over-fitting of data.

The BN is a graphical model of the probabilistic relationship between variables, which encodes the joint probability distribution for a large set of variables (Heckerman, 1995). BNs for a set of variables $\mathbf{X}$ consist of a network structure S that encodes a set of conditional independence assertions about variables in $\mathbf{X}$, and a set of local probability distributions associated with each variable. The network S is a graph with nodes and directed links (known as arcs), and must be a directed acyclic graph (DAG). In a DAG, there are no closed paths within the graph; thus we cannot move from node to node along links following the direction of the arrows and end up back at the starting node (i.e. there are no directed cycles). The nodes in S are in one-to-one correspondence with the variables $\mathbf{X}$. The links are connected from a parent node to a child node, and the lack of possible arcs in S encode conditional independences. Conditional independence between the random variables X and Y given Z is defined as

$$p(X \mid Y, Z) = p(X \mid Z) \tag{3.1}$$

Therefore, once the value of Z is known, X and Y are independent.

3.2.1 A Simple Bayesian Network for Glaucoma

A simple BN for glaucoma data is shown in Figure 3.3. This network includes a *Subject Condition* node that represents whether a patient is a glaucoma sufferer or not, and a set of leaf nodes that represent the output of three different tests.

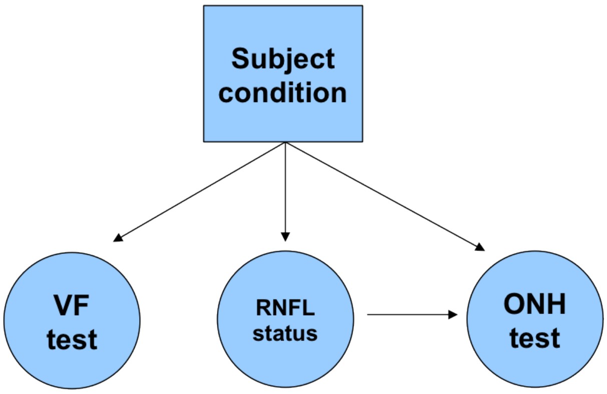


Figure 3.3: BN for Glaucoma with 3 binary leaf nodes representing tests' outputs and a binary Subject Condition class node, representing the patients' glaucoma condition.

The network assumes conditional independence between VF and ONH tests, given the underlying condition of the patient. In general, a BN can encode a set of conditional independence relationships between variables, and it can be demonstrated that a node is conditionally independent of its non-descendants given its parents (Pearl and Shafer, 1988). Also, in a BN a node is conditionally independent of all the other nodes in the graph given its *Markov blanket*, which is defined as the node's parents, children and children's parents.

The CPD associated with each variable X encodes the probability of observing its values given the values of its parents, and can be described by a continuous or a discrete distribution. In this case, the CPD is called a Conditional Probability Table (CPT). For example, in the above graph the CPTs are as reported in Tables 3.2-3.5.

Table 3.2: Subject Condition (SC) CPT.

Value	Probability
Healthy	0.6
Glaucoma	0.4

Table 3.3: VF Test CPT.

Sensitivity	Probability	Probability
	(PD = Healthy)	(PD = Glaucoma)
Low	0.1	0.7
Medium	0.2	0.2
High	0.7	0.1

Table 3.4: RNFL status CPT

Outcome	Probability	Probability
	(SC = Healthy)	(SC = Glaucoma)
Healthy	0.6	0.3
Impaired	0.4	0.7

The CPT in Table 3.2 indicates the prior probability of having glaucoma, as the *Subject Condition* node has no parents. All the CPTs together provide an efficient factorisation of

Table 3.5: ONH Test CPT (G = Glaucoma, H = Healthy).

RNFL status = Healthy		
Outcome	Probability	Probability
	(SC = H)	(SC = G)
Positive	0.3	0.8
Negative	0.7	0.2

RNFL status = Impaired		
Outcome	Probability	Probability
	(SC = H)	(SC = G)
Positive	0.2	0.95
Negative	0.8	0.05

the joint probability

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa_i}) \tag{3.2}$$

where $\mathbf{pa_i}$ are the parents of the node x_i (which denotes both node and variable). For the network shown in Figure 3.3, the joint probability of all nodes is:

$$p(SC, VF, ONH, RNFL) = p(SC)p(VF \mid SC)p(ONH \mid VF, SC)p(RNFL \mid ONH, SC, VF)$$

using the Chain Rule of probability. Fortunately, we can exploit the conditional independences encoded in the BN, obtaining a more compact representation of the joint probability:

$$p(SC, VF, ONH, RNFL) = p(SC)p(VF \mid SC)p(ONH \mid SC)p(RNFL \mid ONH, SC)$$

In general, for n binary nodes, the number of parameters to be estimated is $O(2^n)$, but the factored form needs $O(n2^k)$, where k is the maximum number of parents of a node.

3.2.2 BN Inference

Once the CPTs are defined, it is possible to make inference about the value of any target variable in a BN. For instance, if we observe a *Positive (P)* ONH test, the probability of the patient being *Glaucomatous (G)* can be calculated using Bayes' Theorem:

$$p(SC = G \mid ONH = P) = \frac{p(SC = G, ONH = P)}{p(ONH = P)}$$

$$= \frac{\sum_{VF,RNFL} p(SC = G, VF, RNFL, ONH = P)}{p(ONH = P)}$$

$$= \frac{\sum_{VF,RNFL} p(SC = G, VF, RNFL, ONH = P)}{\sum_{VF,RNFL,ONH} p(SC, VF, RNFL, ONH = P)}$$

The calculations are carried out using the Chain Rule of Probability and applying marginalisation, i.e. summing out all the irrelevant variables. The sums can be "grouped" while marginalising according to the conditional independences between the variables, in a process called *variable elimination*:

$$= \frac{\sum_{VF} \sum_{RNFL} p(SC = G)p(VF \mid SC = G)p(RNFL \mid SC = G)p(ONH = P \mid SC = G, RNFL)}{\sum_{VF,RNFL,ONH} p(SC, VF, RNFL, ONH = P)}$$

$$= \frac{p(SC = G) \sum_{RNFL} p(RNFL \mid SC = G)p(ONH = P \mid SC = G, RNFL) \sum_{VF} p(VF \mid SC = G)}{\sum_{VF,RNFL,ONH} p(SC, VF, RNFL, ONH = P)}$$

$$= \frac{p(SC = G)p(RNFL = H \mid SC = G)p(ONH = P \mid SC = G, RNFL = H)}{\sum_{VF,RNFL,ONH} p(SC, VF, RNFL, ONH = P)} +$$

$$+ \frac{p(SC = G)p(RNFL = I \mid SC = G)p(ONH = P \mid SC = G, RNFL = I)}{\sum_{VF,RNFL,ONH} p(SC, VF, RNFL, ONH = P)}$$

$$= \frac{0.4 \cdot 0.3 \cdot 0.8}{K} + \frac{0.4 \cdot 0.7 \cdot 0.95}{K} = \frac{0.096 + 0.266}{K} = \frac{0.362}{K} = 0.7137$$

where K is a normalising constant corresponding to the likelihood of the data. Note that when RNFL is also observed to be *Impaired*, the probability of glaucoma becomes 0.8471, but when it is observed alone the probability is just 0.5385. This allows assessment of the impact of different tests and, even if in a small exemplificative context, shows the potential of using BNs.

A similar approach is used to perform classification, i.e. finding the posterior probability given observations of the other variables:

$$p(C_k \mid \mathbf{x}) = p(C_k \mid x_1, x_2, x_3, ...) = p(C_k, x_1, x_2, x_3, ...)p(x_1, x_2, x_3, ...) \tag{3.3}$$

Exact inference in BNs can become a very demanding task; therefore approximated methods are often used in practice. These include methods based on Monte-Carlo sampling and variational methods. In this thesis, the junction tree engine was used (Murphy, 1998). This is an exact inference method that uses a greedy search procedure to find a good ordering for variable elimination.

3.2.3 Parameter Learning

The process of building a BN involves learning the structure and estimating the conditional probabilities of the nodes. Given the structure, the latter objective is to determine the parameters of a distribution (or a set of distributions, in relation to the dimension of the variables) for a node for every configuration of its parents. For a full Bayesian approach, the posterior distribution of the parameters $p(\boldsymbol{\theta} \mid \boldsymbol{D})$ should be computed. However, often the calculation of the full posterior is intractable, and approximate methods are used. For example, the parameters $\boldsymbol{\theta}$ can be estimated by maximising the posterior probability of the parameters given the data $\boldsymbol{D}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{D}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3.4}$$

This can be done only under the assumptions that there are no missing data and that parameters are mutually independent. Note that, with a uniform prior distribution of the parameters, the latter equation is simplified by maximising the marginal likelihood function, i.e. $\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{D} \mid \boldsymbol{\theta})$. In this case, the estimates are called Maximum Likelihood estimates, as opposed to Maximum A Posteriori (MAP) estimates. Often, the log-likelihood is considered, so that the log-likelihood of the dataset D_i, $i = 1..M$, is a sum of terms, one for each node k:

$$L_k = \frac{1}{M} \sum_{i=1}^{M} \log p(x_k | pa(x_k), D_i) \tag{3.5}$$

The log-likelihood contribution can be calculated for each node independently and, for discrete nodes, the CPD estimation is reduced to simple counting. For example, for the BN in Figure 3.3, the conditional probability parameter of the node *VF test* is obtained with:

$$\begin{aligned} p(VF = k \mid PD = j) &= \frac{n(VF = k, PD = j)}{n(PD = j)} \\ &= \frac{n(VF = k, PD = j)}{\sum_{m=1}^{M} n(VF = m, PD = j)} \end{aligned} \tag{3.6}$$

where n is the number of samples in the dataset and $m = 1..M$ is the set of values taken by the variable *VF test*.

When the learning data is incomplete or there are unobserved variables, the posterior distribution becomes intractable as the log-likelihood is not decomposable. Thus, approximation techniques such as Monte-Carlo methods, Gaussian approximation and the Expectation Maximisation (EM) Algorithm can be used to solve the problem (Dempster et al., 1977). EM is an iterative procedure in two steps that aims to estimate the parameters. In the first step, the hidden values for the unobserved data are inferred using the present model, while in the second step the estimated likelihood function is maximised. When the algorithm converges to a local maximum, the parameters are estimated. The ability to learn from incomplete or missing data is referred to as clustering or unsupervised learning, which were introduced in the previous chapter. In a BN, this extends to including unobserved or hidden variables, i.e. variables for which no instances of the dataset are specified.

3.2.4 Structure Learning

The structure of the network can be imposed or learned from the data. The first case is useful when external knowledge is available, such as from known causal relationships or expert advice. Learning the structure from data, on the other hand, often leads to interesting insights and performance improvements. Structure learning is a NP-hard problem, since the number of DAGs on N variables is super-exponential in N. Therefore, techniques to search for local optima in the structure space are typically employed, such as a search-and-score approaches. In this case, the neighbourhood of an initial guessed network is iteratively searched until a local optimum is reached. A widely used search technique is the Simulated Annealing (SA) algorithm (Kirkpatric et al., 1983), which performs a quasi-greedy search by allowing non-optimal solutions to be explored to escape local minima.

Algorithm 3.1 Simulated Annealing algorithm.

```
Input: Temperature, Cooling_Factor, Model, Data, Max_Iterations
Score Model and store in Best_Score
Loop on iterations until Max_Iterations is reached
    Apply OPERATION to Model
    Score Model and store in New_Score
    Diff_Score is New_Score - Score
    if (New_Score is higher than Score)
        Update Score and Model
        if (New_Score is higher than Best_Score)
            Store Model for Output
            Best_Score is New_Score
        end if
    elseif Random_Number is less than exp(Diff_Score / Temperature)
        Update Score and Model
    end if
    Temperature is Temperature · Cooling_Factor
end Loop
Output: Model stored for Output
```

The procedure is described in Algorithm 3.1. The input parameters *Temperature*, *Cooling_Factor* and *Max_Iterations* regulate the convergence of the algorithm, *Model* is the starting model and *Data* is the training dataset in use. *Best_Score* is used to store the current best model score, while *Score* and *New_Score* hold respectively the score of the current model and the new model at each iteration. The latter is evaluated after operations are carried out on the current model. If the new score is higher than the current score, the parameters of the current model and the current score are updated. If the new solution is the best among the explored ones, the results are also stored in *Best_Score* and for the final output. If the new solution is worse than the current one, a temperature factor *Temperature* regulates whether to accept it as a new current solution. Note that *Random_Number* represents a random number between 0 and 1. The acceptance of solutions which are not the best allows the algorithm to escape local minima, which are a common issue in structure learning of BNs. The temperature factor is lowered at each iteration proportionally to a cooling factor *Cooling_Factor*. After a specified number of iterations *Max_Iterations*, the algorithm returns the best explored model.

To evaluate the models, the full posterior distribution of the parameters for a structure can be estimated. As explained before, point estimates are usually considered, so that the task is to maximise the posterior distribution, which can be decomposed as in Equation 3.7:

$$p(M \mid \boldsymbol{D}) = \frac{p(\boldsymbol{D} \mid M)p(M)}{p(\boldsymbol{D})} \tag{3.7}$$

where $p(\boldsymbol{D} \mid M)$ is called the *marginal likelihood* of the model M and $p(M)$ is the prior probability of the model. Note that the same task can be solved by maximising the logarithm of these two functions, which can be approximated for efficient computation. For example, the Bayesian Information Criterion (BIC) score (Schwarz, 1978) for the model M can be calculated as follows:

$$BIC = \log p(M) + \log p(\theta) + \log L(\hat{\theta}, M \mid \boldsymbol{D}) - \frac{k}{2}\log n \tag{3.8}$$

Where $p(\theta)$ is the prior probability of the parameters θ, n is the number of cases in $\boldsymbol{D}$, k is the number of parameters in the model and L is the Maximum Likelihood configuration of θ, i.e. $L(\hat{\theta}, M \mid \boldsymbol{D}) = p(\boldsymbol{D} \mid \hat{\theta}, M)$ with fully observed $\boldsymbol{D}$. Often, a non-informative uniform prior is assumed for both the models and the parameters, obtaining a likelihood-based score adjusted for the model complexity. Since the likelihood function monotonically increases with the number of parameters, overfitting and estimation problems may arise if no correction factor was present to take into account the complexity of the model. Other approaches, such as the Akaike score (Akaike, 1974), penalise differently the dimension of the model but have the same rationale. Since the number of parameters in the model is the sum of the number of parameters in each node, the BIC score also decomposes.

Missing data, or unobserved variables, make the structure search problem more complicated as the computation of the score becomes very inefficient. Therefore, few solutions are available for this task. One solution, called Structural EM (SEM), extends the EM algorithm in order to search the joint space of parameters and structure (Friedman et al., 1997). At each iteration the algorithm can either find better parameters for the current structure or select a new structure.

3.2.5 Model Averaging

When learning BNs, resampling techniques can be used in order to reduce noise and obtain more robust results. Bootstrapping (Friedman et al., 1999), for example, involves iterative random sampling from the dataset (with replacement). The technique permits to obtain an unknown characteristic of an unspecified distribution by drawing subsets from the observed data iteratively, and computing a statistic for each subset. For a great number of iterations, the bootstrap distribution approximates the actual distribution. This technique has been applied in many fields (Zoubir and Boashash, 1998), and it can also be applied in searching the space of BN models using subsets of training data. Bayesian Model Averaging can be used on the set of learned structures to calculate the posterior probability for each arc as proposed in (Friedman and Koller, 2003), i.e. by weighting each arc f on the network score in which it is present. Since calculation of the posterior probability on the whole set of possible structures is feasible only for tiny domains, an approximation is made by enumerating only over the learned set of structures (i.e. a set of optimal structures), as proposed by (Madigan and Raftery, 1994). Thus, the formula used to estimate the relative mass of the structures in G that contains arc f given the data D is:

$$P(f|D) \approx \frac{\sum_{G \in G} P(G|D) f(G)}{\sum_{G \in G} P(G|D)} \tag{3.9}$$

where $f(G)$ is a binary function equal to 1 when the arc is present in the structure G, and $P(G|D)$ represents the posterior probability of each structure given the data D. The latter function can be calculated as the product of the likelihood of the data and a prior over the structures. The arcs with the highest posterior probability are then selected to obtain the final structure.

3.3 Bayesian Networks for Glaucoma

This section presents a set of experiments using BNs for glaucoma to assess the impact of the technique and the characteristics of the datasets, as well as highlighting some key concepts in BN inference and prediction. The BN models were implemented using ad-hoc functions and the Bayes Net Toolbox for Matlab (Murphy, 2001).

3.3.1 Naïve Bayesian Network

A first set of experiments was carried out by training a Naïve BN using Dataset A and testing it on Dataset B (Table 3.1). In a Naïve BN all variables are assumed to be conditionally independent, given the class. This model therefore easily allows the approximation of the full conditional distribution, being linear in the number of variables N rather than exponential. The assumption of independence is very limiting and almost always incorrect in real applications; nevertheless Naïve BNs have been shown to perform well in many classification applications (Tucker et al., 2005). The structure of the chosen Naïve BN is shown in Figure 3.4.

VF pointwise data was grouped into six variables corresponding to the nerve fibre layer structure, following (Garway-Heath et al., 2000) and Figure 2.3. Therefore, it was possible to obtain anatomically concordant HRT and VF measurements for the six sectors of the OD,

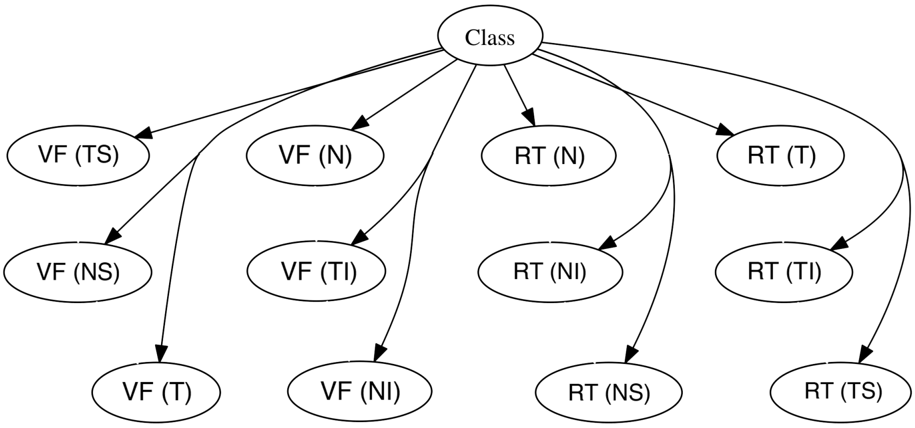


Figure 3.4: Naïve Sectorial Bayesian Network

i.e. temporal superior (Ts), nasal superior (Ns), temporal (T), nasal (N), temporal inferior (Ti) and nasal inferior (Ni) sectors.

The CPD of each variable was assumed to be a continuous Gaussian distribution with unknown mean and standard deviation. The training of the parameters was performed using ML estimation as already described for the discrete case. The AGIS metric was not used as an indicator of glaucoma in the testing dataset (Dataset B). The BN classifier was compared to a C4.5 Tree classifier, a KNN classifier, a Multilayer Perceptron and a Logistic Regression classifier. These classifiers are described in the previous chapter and were built using Weka software ("The WEKA Data Mining Software : An Update"). The ROC curves of the assessed classifiers are shown in Figure 3.5. The sensitivity of AGIS and MRA metrics is lower than other ML classifiers. Although at high specificity the BN does not perform better than the other ML classifiers; for specificities below 88%, the BN model outperforms all the other models. This is reflected in the overall AUROC values, which are shown in Table 3.6.

3.3.2 Anatomy-Based Bayesian Network

The independency between the variables that is imposed by the Naïve BN does not incorporate any external anatomical knowledge about glaucoma. To include such information, the structure presented in Figure 3.6 was considered. s Instead of the MRA linear combination of Age, Optic Disc Area (ODA) and Retinal Rim Area (RRA), raw ODA and RRA were

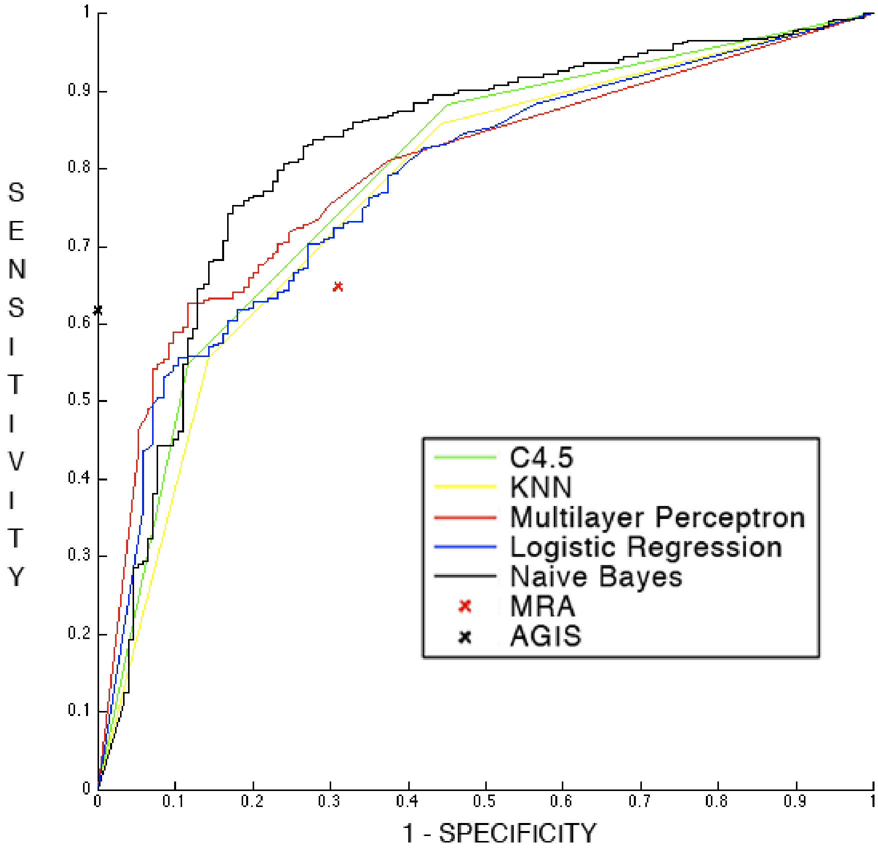


Figure 3.5: ROC curves for a set of ML classifiers tested on Dataset B.

Table 3.6: AUROC values for ML and traditional classifiers

Classifier	AUROC
KNN	0.70
C 4.5	0.72
Linear Log	0.78
Multilayer	0.79
BN	0.82

considered. This allows the inclusion of anatomical knowledge in the model, by imposing dependency between ODA and RRA and between nerve fibre bundles and VF sectors. The

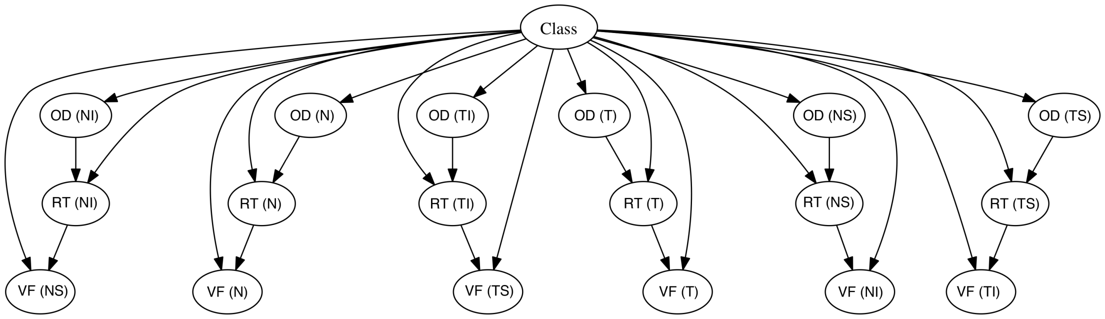


Figure 3.6: Anatomy-based BN

direct dependence between ODA and glaucoma has been proposed in the literature as an explanation for African-Caribbean subjects being at increased risk of developing glaucoma (Sommer and Doyne, 1996). The performance of this classifier was compared to the Naïve BN built using both raw and pre-processed HRT values, obtaining the AUROC reported in Table 3.7.

Table 3.7: AUROC values for ML techniques

Classifier	AUROC
Naïve BN	0.823
Raw Naïve BN	0.830
Anatomy-based BN	0.835

The overall performance of the anatomy-based BN is higher than the others. However, a better analysis can be obtained by looking at the full ROC curve, which shows the performance at different specificities (Figure 3.7). At high specificities, no classifier outperforms the others. However, the anatomy-based classifier outperforms the others for specificities lower than 80%, while the Naïve BN is the best classifier for specificities between 90% and 80%. The MRA pre-processing affects the performance of the classifier at higher specificities, but has a positive impact for lower specificities. Thus, each network has a particular utility at different specificities. The output provided by the three classifiers for two subjects is shown in Figure 3.8. The differences in the outputs can be explained by analysing how

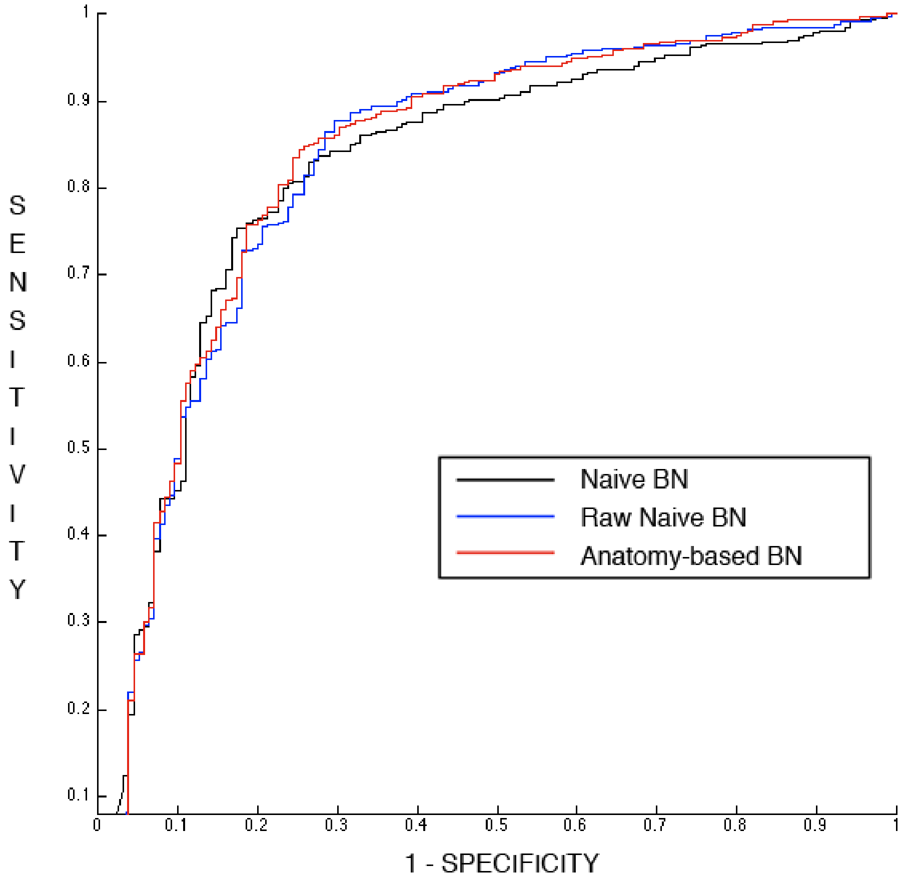


Figure 3.7: ROC curves for anatomy-based and Naïve BNs

conditional dependencies act on each BN. The Naïve BN, in fact, is composed of diverging arcs only, i.e. arcs from the class node to the leaf nodes. In this case, information may be transmitted through a diverging connection if the state of the parent variable in the connection is not known, e.g. in classification. Once the parent variable is observed, the two leaf nodes become independent. By introducing VF-RT arcs, converging connections are added to the model. In converging connections, information passes from one parent to the other if either information about the state of the variable in the connection or one of its descendants is available. In other words, the parents are conditionally dependent given the value of their common child, even if they are marginally independent. This effect is called *explaining away*, because the parent nodes compete to explain the observed value of the common child.

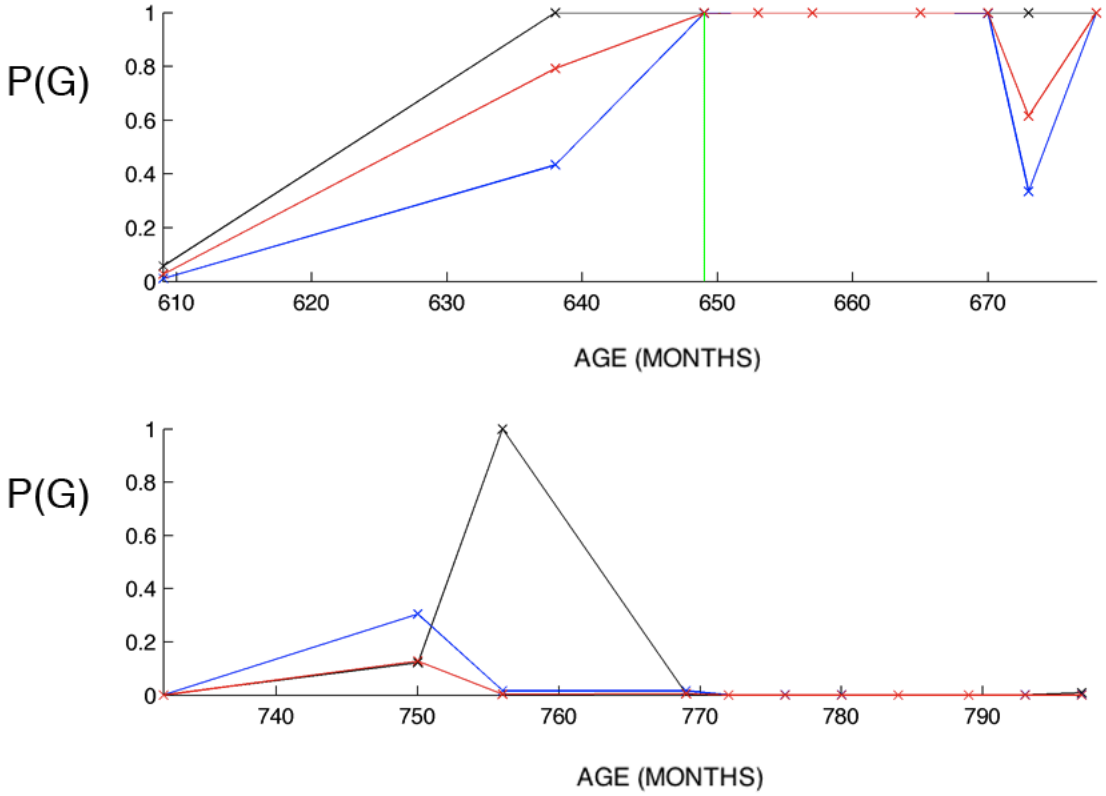


Figure 3.8: Probability of having glaucoma for a converter and a healthy subject. The colors are chosen according to Figure 3.7 and the green vertical line represents the AGIS onset of glaucoma.

For instance, having observed that *VF (NS)* is low (e.g. 15 dB), the probability of having Glaucoma is:

$$P(Class = Glaucoma \mid VF(NS) = 15)$$

However, if we also observe a thin neuroretinal rim *RT (NI)* (e.g. 5), the posterior probability of glaucoma becomes lower, as the rim size is ”explaining away” the observed low VF sensitivity:

$$P(Class = Glaucoma \mid VF(NS) = 15, RT(NI) = 5) < P(Class = Glaucoma \mid VF(NS) = 15)$$

Indeed, a thinner rim is indicative of glaucoma, therefore in this case a direct arc between the two parents *RT (NI)* and *Class* was included. Modelling causality is a debated subject and (Pearl, 2000) offers a wide discussion about the matter. From a general point of view, causality can be extracted from data in conformity with the Causal Markov Condition. The latter implies that if a causal relation is present, then there is a BN that correctly captures it. However, understanding whether the correct BN is found is often not feasible, considering the number of possible networks and the nature of the tests that should be carried out to assess causality between variables, as discussed in (Heckerman, 1995).

3.3.3 Data-Driven Bayesian Network

Among the techniques used to learn conditional independence between variables, BN structures can be learned from data using a search-and-score approach. SA, described in Algorithm 3.1, was carried out using a particular set of operational functions on VF and HRT sectorial variables. In particular, three structural operations were considered:

- Adding an arc: a random node among the non-connected nodes is chosen and connected to another node, with the maximum number of parents of a node set to 2.
- Deleting an arc: a random arc in the BN is chosen and removed.
- Moving an arc: the other two operations are called in sequence, so that a random arc is deleted and a new arc is added.

The following Figure shows possible effects of the operations on a simple BN.

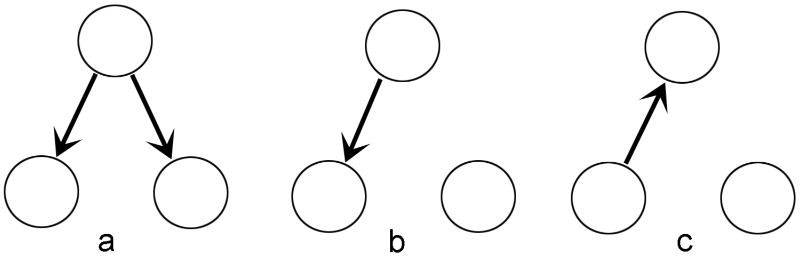


Figure 3.9: Operations on the structure of a BN. a to b: Deleting an arc, b to c: Moving an arc, b to a: Adding an arc.

The learning algorithm was run 100 times using bootstrapped data from Dataset A. In particular, 100 sets of records were randomly sampled with replacement from dataset A and used as training datasets for the algorithm. The resulting 100 networks were then weighted based on their likelihood score, according to (Friedman and Koller, 2003) methodology presented in Section 3.2.5. Typical parameters were 2100 iterations per run, temperature starting at 5 and a cooling factor equal to 0.99. These parameters were obtained by tuning and observing the score curves obtained during the learning. Parameters were accepted when curves consistently ended in a plateau, as illustrated in Figure 3.10. The figure shows how the number of iterations is chosen: the likelihood-based score typically improves dramatically during the first 500 iterations, therefore about 2000 iterations are set before stopping the algorithm. The figure also shows the amount of score variability obtained across the 100 runs, in terms of standard deviations from the mean.

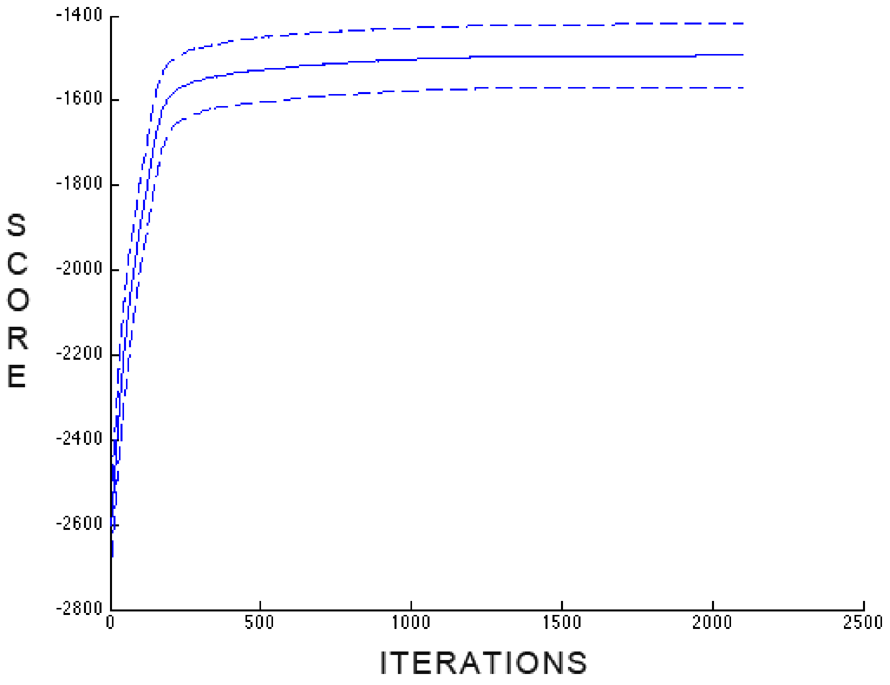


Figure 3.10: Likelihood-based mean score plus and minus Standard Deviation (dashed) for 100 runs of the SA algorithm on Dataset A.

As in the other experiments, each variable was modelled as a continuous Gaussian distribution. The resulting networks were combined using Bayes Model Averaging, as explained earlier. The resulting relationships are shown in Figure 3.11.

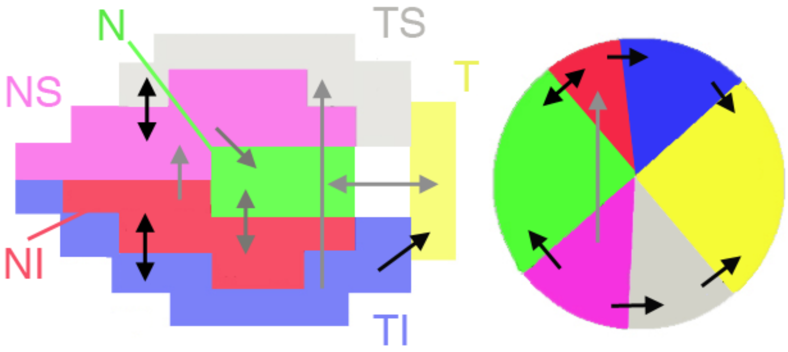


Figure 3.11: Relationships learned on Dataset A using Simulated Annealing. Lighter arcs represent lower probabilities (black: P>0.8, grey: 0.6<P<0.8).

By looking at the data-driven structure, several insights can be obtained. Most of the discovered arcs connect adjacent sectors in relation to the structure-function map. This is due to the pre-processing of the data and can be seen as a proof of the reliability of the BN and the implemented learning algorithm. Many arcs involve the inferior part of the OD, which is confirmed by several studies in which early damage in glaucoma is shown to occur earlier at the inferior segment of the OD in arcuate sectors of the VF (Yanoff and Duker, 2003; Johnson et al., 2003), although this hypothesis is not fully accepted (Artes and Chauhan, 2005). From this set of arcs, a BN was built by avoiding cycles while preserving conditional dependencies, obtaining the network in Figure 3.12.

The model obtained was then used for classification on Dataset B and compared to the previous investigated BNs (Figure 3.13). The new data-driven classifier obtained the highest AUROC among the models (0.853), with consistently higher sensitivities, especially at high specificity values. High specificities are more of interest for diseases such as glaucoma, because it is a relatively low frequency disease. This reflects the higher weight put to minimise the occurrence of false positives (FPs).

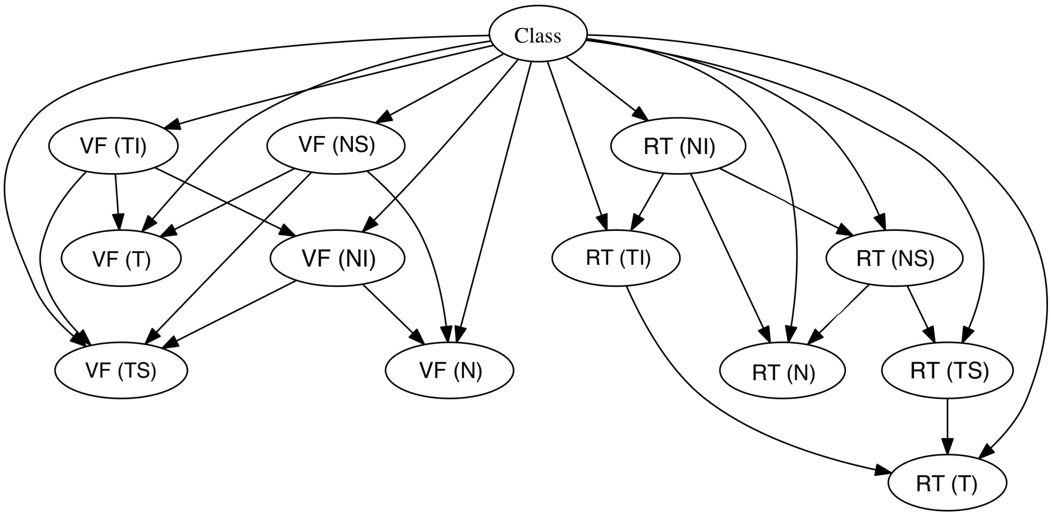


Figure 3.12: BN learned on Dataset A using Simulated Annealing.

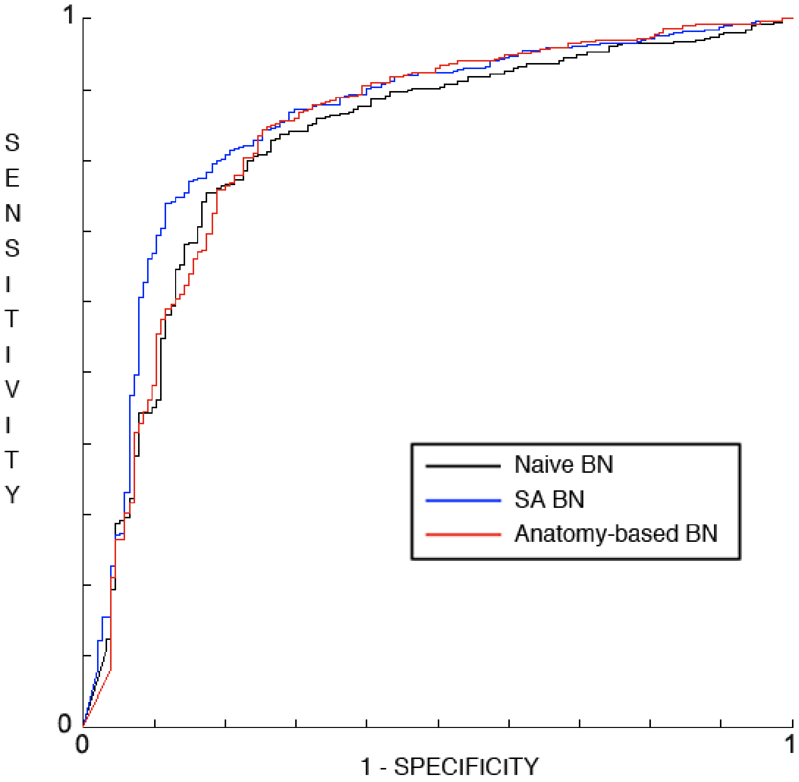


Figure 3.13: Comparison of ROC curves for SA derived BN, anatomy-based BN and Naïve BN.

Another test was also performed to take into consideration the different nature of the two datasets. Dataset A is a cross sectional dataset, while dataset B is a longitudinal dataset. Therefore, a test between the models explored was carried out by selecting only one visit from dataset B to obtain a cross-sectional testing dataset. The sampling was based on the AGIS score, consistently with the inclusion criteria for dataset A, so that the first positive visit for each patient was selected for testing. A random visit was sampled from each patient in the controls group. The training dataset was randomly split into 3 groups. Of these, 2 parts were used to train the model. The process was repeated three times, one for each combination of training subsets. This procedure was carried out to guarantee a ratio of training and testing datasets size of roughly 2:1. The results showed that, as in the full dataset tests, the SA data-driven BN classifier outperformed the other two, which behaved similarly. Average AUROC was 0.909 for the SA classifier, while the Naïve BN and the Anatomy-based classifier obtained 0.852 and 0.856 respectively. The ROC curves for each test are showed in Figure 3.14

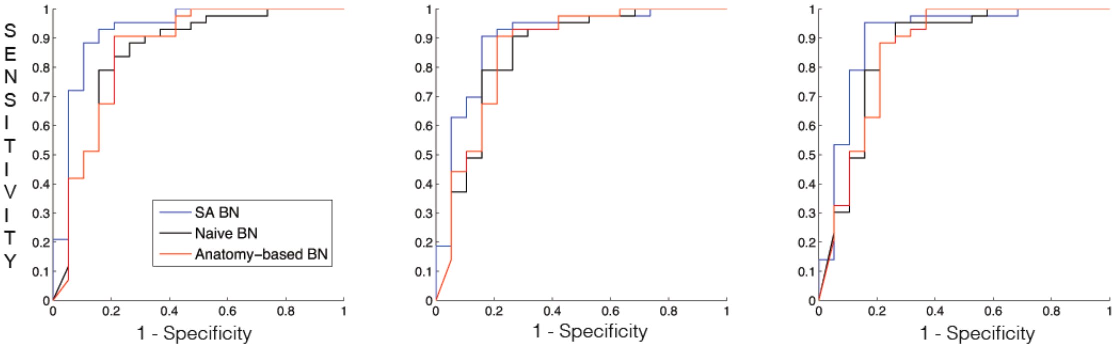


Figure 3.14: Test results of ROC curves for SA derived BN, anatomy-based BN and Naïve BN. The classifiers were tested on a cross-sectional dataset extracted from dataset B and with three different subsets of dataset A as training datasets.

3.4 Summary

This chapter has introduced the specific datasets used in this research and the basics of BN models. A small set of simplistic BNs were introduced to demonstrate the usefulness of these models in glaucoma. Early investigations on combined functional and anatomical data have shown that BNs obtain better performances over other ML techniques and classic metrics. The results of data- and anatomy-driven BNs also suggest that such models are effective and informative. In particular, the anatomy-based BN seems to perform well at lower specificities, but it is outperformed by the data-driven BN at higher specificities.

The utility of data integration and the potential of BNs will be further explored in following chapters, where more advanced models will be presented. The next chapter will look deeper into BNs for cross-sectional data and will also propose new techniques for tackling bias associated with metrics and inter-subject variability, while Chapter 5 will present BN models for application in longitudinal data.

Chapter 4

Bayesian Networks for Cross-Sectional Data

Supervised ML classifiers, such as the BNs presented in previous chapters, require a training dataset to model the relationships between inputs (observed variables) and outputs (outcome variables, e.g. diagnoses). Such training data needs to be labelled; in the case of diagnoses, labels would be having a disease or not having a disease. However, in glaucoma, the true condition is not assessable, since there is no gold standard test for the condition and because there is no univocal definition of glaucoma. Using a particular metric to label subjects in the training dataset introduces biases that should be taken into consideration.

The first section of this chapter aims to tackle this issue and explore the impact of different metrics. Section 4.2 is focused on combining different BNs. In particular, an unsupervised anatomy-based BN and a data-driven BN supervised on the AGIS metric are combined using a BN-based ensemble of classifiers, with the purpose of enhancing classification performance and better understanding their interaction. Section 4.3 presents a novel clustering algorithm that aims to tackle inter-subject variability in the data, and at the same time obtain more insights into glaucoma and the AGIS metric. The last section summarises the main conclusions of the previous sections and introduces the next chapter.

4.1 Supervised, Semi-Supervised and Unsupervised BNs

This section explores supervised and unsupervised BNs in conjunction with the metrics AGIS, MRA and a combination of the two. A semi-supervised approach, which is based on healthy subjects' data, is also described and assessed. The last part of the section compares and discusses the results obtained with the different approaches and summarises their characteristics.

4.1.1 AGIS and MRA based Bayesian Networks

In Datasets A and B, the AGIS functional metric was used to classify patients as having glaucoma, de facto introducing a bias towards this metric. The parameters of the models presented in the previous chapter were trained on VF AGIS-labelled data and ignored the condition of the OD of the subjects. MRA can be used as an OD diagnosis metric for glaucoma, with a significant negative deviation from the mean in MRA suggesting glaucoma. By re-labelling patients using this MRA classification metric, a new set of relationships, S_{MRA}, can be learned (Figure 4.1) using the SA algorithm and Bayesian Model Averaging, as was performed in the previous chapter for AGIS-labelled training data.

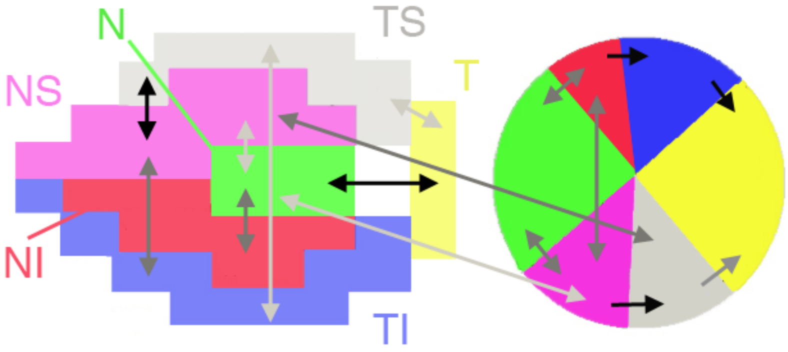


Figure 4.1: Relationships learned using MRA-labelled data. Lighter arcs represent lower probabilities (black: P>0.8, grey: 0.6<P<0.8).

From this, a BN structure can be extracted and used to make inference (Figure 4.2). The notation S_{MRA} indicates that the structure (S) of the BN is obtained using MRA-labelled data. The notation P indicates the data used to train the parameters of the BN. For example, a BN described by S_{MRA},P_{MRA} is a network trained using MRA-labelled data both for structure and parameters.

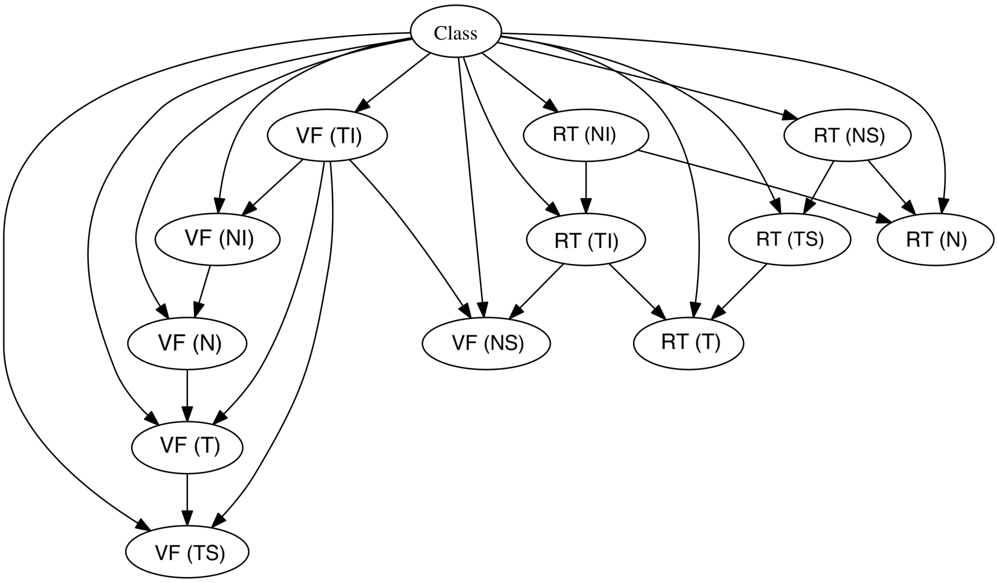


Figure 4.2: BN structure learned using MRA-labelled data (S_{MRA}).

The classification performance on the longitudinal dataset (Dataset B) for the BN fully-trained on MRA-labelled data is lower than the AGIS-based network, as shown in Table 4.1.

Table 4.1: Performance of AGIS and MRA combined BN classifiers. Subscripts S and P indicate the metrics used for structure learning and parameters training, respectively.

Classifier	Sensitivity at 90% specificity	Sensitivity at 80% specificity	AUROC
S_{AGIS}, P_{AGIS}	0.66	0.80	0.853
S_{MRA}, P_{MRA}	0.59	0.74	0.818
$S_{MRA+AGIS}, P_{MRA+AGIS}$	0.64	0.78	0.838

This table shows the AUROC values and the sensitivity values at 90% specificity and

80% specificity for the classifiers. The table also reports the performance of a BN learned and trained on MRA- or AGIS-labelled data, $S_{MRA+AGIS}, P_{MRA+AGIS}$, so that a point was considered positive in the dataset used for its training if either the MRA or AGIS metric was positive. Different metrics can be combined differently. For instance, the structure S_{MRA} can be trained using AGIS-labelled data, obtaining S_{MRA}, P_{AGIS}, which can be seen as a BN that combines the different metrics. The classification performance of the two possible combinations of structure learning and parameters training are compared in Figure 4.3.

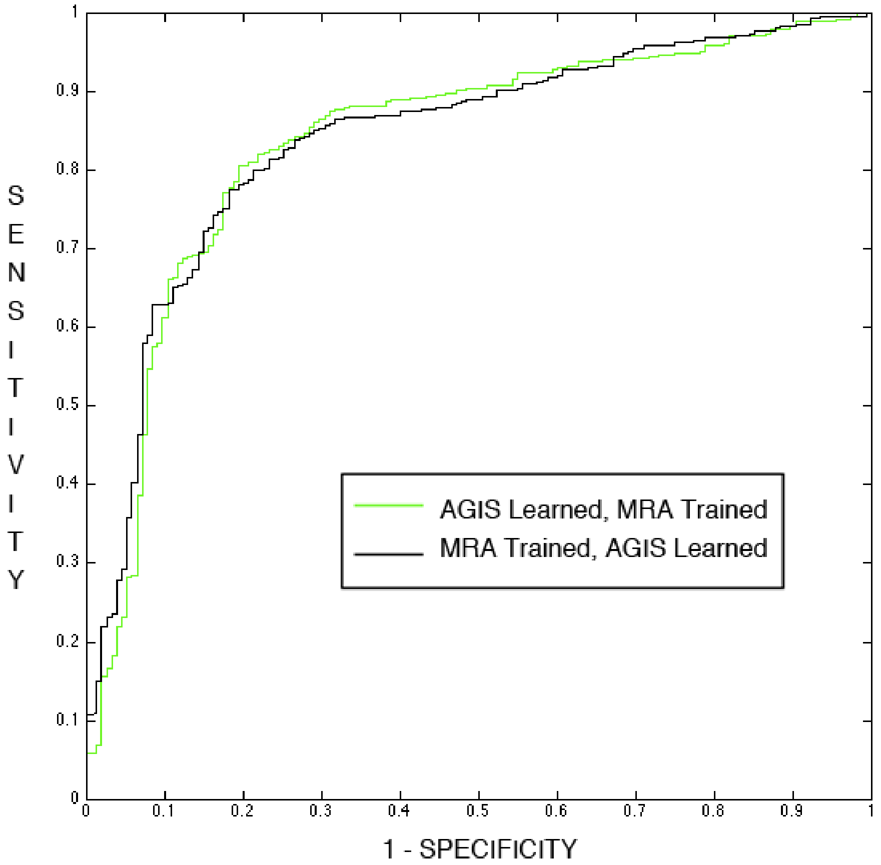


Figure 4.3: Performance of BNs learned using data labelled according to a metric and trained using data labelled according to another metric.

The classifiers perform differently at different specificities, with frequent crossings between the ROC curves but overall similar performance.

4.1.2 Semi-Supervised Bayesian Network

Ideally, it would be best to extract the structure of a BN from the data while being independent from a positive MRA or AGIS clinical metric. One solution is to isolate healthy data and use it to train a classifier. This leads to semi-supervised classification, i.e. using only labels belonging to a subset of the data. The control group is assumed to be glaucoma-free, which is likely given the low incidence of the condition and the inclusion criteria applied for control subjects. Thus, considering these subjects as healthy, it is possible to obtain higher specificity and make use of all the structural and functional information without the application of imperfect metrics or subjective clinical interpretation. In the BN framework, a semi-supervised model can be learned using only control-labelled data and then trained on labelled control and unlabelled glaucomatous data. In other words, the new label C_i^{New} for patient i is assumed to be equal to

$$C_i^{New} = \begin{cases} C_i^{Old} & \text{if } C_i^{Old} = Control \\ H & \text{if } C_i^{Old} = Glaucomatous \end{cases} \tag{4.1}$$

where C indicates controls and H is the unobserved value. The set of relationships obtained using only *Control* labelled data can be seen in Figure 4.4, followed by the derived BN $S_{CONTROLS}$.

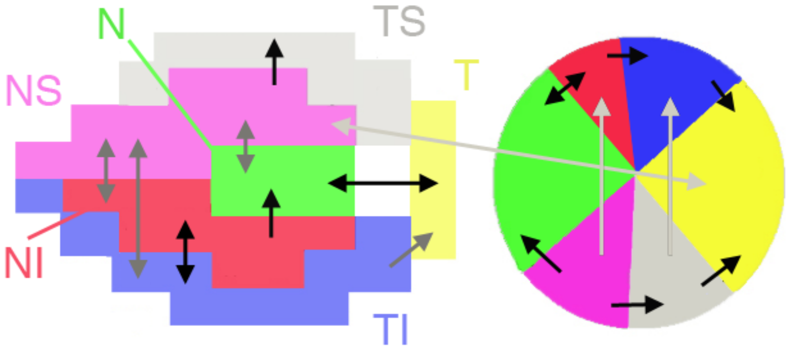


Figure 4.4: Relationships learned using only healthy subjects labelled data ($S_{CONTROLS}$). Lighter arcs represent lower probabilities (black: P>0.8, grey: 0.6<P<0.8).

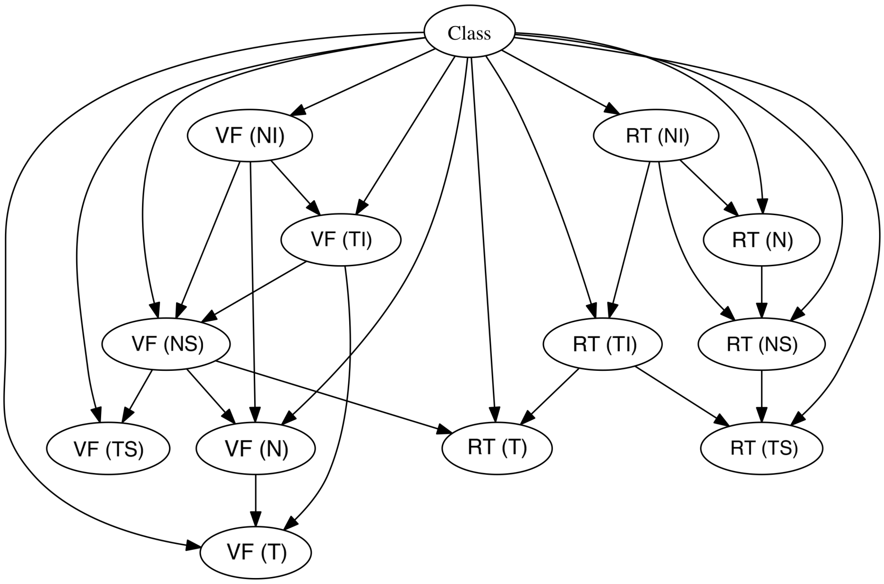


Figure 4.5: BN structure learned using *Controls* labelled data.

The combination of this structure with the AGIS metric, i.e. $S_{CONTROLS}, P_{AGIS}$, was tested on Dataset B, and obtained the best results observed among the classifiers introduced so far. In particular, it achieved 71% sensitivity at 90% specificity, and 80% sensitivity at 80% specificity, with an AUROC equal to 0.856.

4.1.3 Unsupervised Bayesian Network

As already introduced, the use of unlabelled data is referred to as *clustering*. In BNs, unknown labels are treated as missing values and algorithms such as the EM can be used to estimate the parameters of the model.

Unsupervised BNs performances on Dataset B are reported in Table 4.2. The $S_{CONTROLS}$, $P_{CONTROLS}$ network is indeed a semi-supervised network, as it makes use of the healthy patients' labels (Controls) both for the structure (S) and the parameters (P). The performances of unsupervised models were generally lower than supervised or semi-supervised techniques.

Table 4.2: Performance of AGIS and MRA unsupervised BN classifiers. The subscript of S and P indicate the metrics used for structure learning and parameters training, respectively.

Classifier	Sensitivity at 90% specificity	Sensitivity at 80% specificity	AUROC
$S_{CONTROLS}, P_{CONTROLS}$	0.66	0.80	0.851
$S_{AGIS}, P_{UNSUPERVISED}$	0.58	0.78	0.841
$S_{MRA}, P_{UNSUPERVISED}$	0.55	0.72	0.818

4.1.4 Discussion

Figure 4.6 shows the three structures obtained with AGIS-, MRA- and Healthy subjects labelled data. The best performing networks (i.e. AGIS-labelled and semi-supervised networks) have similar structures but are different to the MRA-based model, especially with regards to strengths and across-types relationships. This leads to the conclusion that being more conservative is more effective in this dataset in terms of classification. In fact, a triple out-of-range VF test is a conservative metric, whilst MRA at 95% P.I. is more prone to False Positives. However, from a qualitative point of view, the OD-based network more reliably captures different characteristics of glaucomatous eyes, as it is guided by OD appearance instead of VF results. For instance, the structure-function relationship may be overlooked by the S_{AGIS} because of the exclusion criteria of patients with impaired OD. The controls-based network captures more relationships between sectors compared to the AGIS-based network, because of the learning process: the presence of a defined class node in the AGIS-based network "explains away" the relationships between sectors by selecting the most informative relationships for classification in the feature selection process.

Parameters appear to play a large role in the learning process. When the controls-based BN was trained using AGIS-labelled data it obtained the best results observed among all the classifiers, i.e. 71% and 80% sensitivity at 90% and 80% specificity, respectively, and an AUROC of 0.856. The results obtained with the combined networks, S_{AGIS}, P_{MRA} and S_{MRA}, P_{AGIS}, also highlight the impact of the metric chosen; however, when compared with

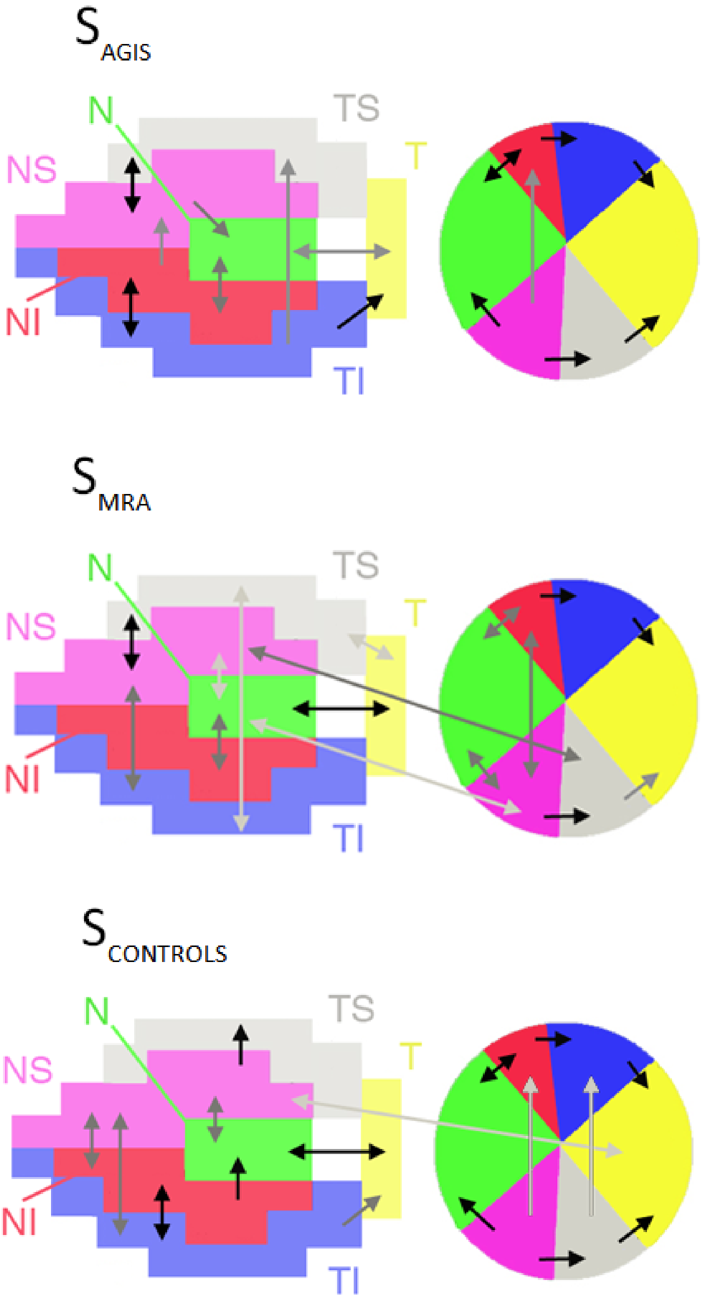


Figure 4.6: BN structures learned using SA approach. Lighter arcs represent lower probabilities (black: P>0.8, grey: 0.6<P<0.8).

the other networks, their performance was worse at higher specificities. The full unsupervised models performed least well, and the semi-supervised model performed best.

There are several possible reasons for these results. Firstly, the EM algorithm needs sufficient data to compute the probabilities iteratively. Another plausible reason is that the S_{MRA} and S_{AGIS} structures are learnt on a supervised dataset, introducing a bias that can lead to poor results if combined with unlabelled data for training. The use of metrics in the inclusion criteria should also be considered while looking at the performances of the models. In fact, glaucoma patients in Dataset A were defined on the basis of the AGIS score and for Dataset B the conversion was also based on three positive AGIS scores. Even if the effect of AGIS on the testing dataset was minimised by considering a series of visits as positive from the first visit, it may be argued that the subjects in this group are more likely to show AGIS score-related defects, and therefore be captured earlier by methods trained on the same metric. The results presented in this section suggest that further investigations in how to combine different metrics could lead to better performing classifiers.

To further compare the models introduced, performances are assessed at 90% specificity in Table 4.3. Errors are broken down into pre- and post- glaucoma diagnosis, where the cut-off was the AGIS-based diagnosis of glaucoma, after which medication was given in many cases.

Table 4.3: Errors of BN classifiers at 90% specificity. The subscripts of S and P indicate the metrics used for structure learning and parameters training, respectively.

Classifier	Total	Pre-Diagnosis	Post-Diagnosis
$S_{CONTROLS}, P_{AGIS}$	133	102	31
S_{AGIS}, P_{AGIS}	157	113	44
$S_{AGIS+MRA}, P_{AGIS+MRA}$	170	119	51
S_{MRA}, P_{AGIS}	176	124	52
S_{AGIS}, P_{MRA}	183	126	57
S_{MRA}, P_{MRA}	191	126	65

At 90% specificity, it is clear that using the semi-supervised structure leads to fewer errors in the post-treatment data, especially when combined with AGIS-labelled data. This is an unsurprising result, because all the networks are learned and trained on glaucomatous cases and so the controls-based network is the only one with an unbiased structure. This leads to more precision in discriminating healthy and non-healthy subjects, as the learning process is not biased by any imperfect metric.

Note that each variable in the BNs so far has been modelled with a continuous Gaussian distribution, but mixtures of Gaussians (MoGs) can also be used. MoGs, introduced in Chapter 2, can be easily implemented in BNs by linking a discrete node of size K to a continuous node. Given that a continuous Gaussian distribution is defined by two parameters, the resulting mixture of Gaussians will have size $2K$. These parameters need to be estimated, for instance using the EM algorithm, in a process that is not guaranteed to reach global optima. Therefore, performance improvement and ease of convergence should be taken into consideration when deciding which model to use to describe the data locally. For Dataset B, similar results are obtained with MoG distributions (Ceccon et al., 2010b).

4.2 Ensembles of Bayesian Networks

As shown in the previous chapter, data-driven networks appear to be better at excluding negative patients than anatomy-based BNs, whilst the latter seem to produce higher sensitivities at low specificities. This highlights the need of a classifier which is able to integrate different datasets follows the indication from consolidated knowledge in the field, as described in Chapter 2 and in the previous chapter. Among the available techniques, the ensembles of classifiers, and in particular the stacked ensembles, seems to fit the problem. Their ability to exploit the single classifiers outputs through a BN framework allows to retain the convenient characteristics and performances of BN base models and combine them selectively to build a single output. Therefore, this combining method was chosen. Non-generative ensemble methods try to combine existing base classifiers, without acting on the base classifiers structure. There is a broad literature on ensembles of classifiers (Kittler, 1998; Sharkey, 1996), and given the input available (i.e. probabilistic outputs from base BNs), a non-generative stacked structure ensemble is chosen (Duin and Tax, 2000). Stacked ensembles use Machine Learning techniques on the top of the base learners, using their output as a set of input data for the final combiner. This type of model has been shown to be an effective combiner in (Garg et al., 2002), especially if the individual classifiers disagree with each other (Hansen and Salamon, 1990). In particular, since the aim here was not just to increase the performance, but also to understand how data and anatomy-driven networks interact, a *Combining BN* model was chosen. This model combines the outputs of two BNs in one node, as shown in Figure 4.7. In order to use different classifiers, we selected the data-driven BN ($S_{CONTROLS}, P_{AGIS}$) and the anatomy-driven unsupervised BN ($S_{ANATOMY}, P_{UNSUPERVISED}$).

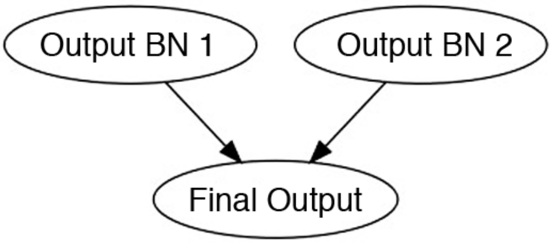


Figure 4.7: Ensemble Bayesian Network.

Each base classifier output was discretised into 8 states and then linked to a 2-states class node. The discretisation was performed using an equal frequency approach in order to make the examination of the CPT of the class node easy to assess. In fact, the CPT results in a 8x8 matrix showing the interactions between the two networks. As will be further discussed later in this section, the CPT is also optimised with a smoothing filter window of size 3. The procedure is presented in Algorithm 4.1.

Where *OutputBN1* and *OutputBN2* represent the two base classifiers' outputs and *Output_CPD* is the Conditional Probability Distribution of the node associated with the final output. The *Output_CPD* can be easily extracted after training the BN *Model*, and used to replace the non-smoothed CPD in the *Update* operation. An optimised BN was also set up by weighting its output on the accuracy of the base classifiers on the training dataset. The probabilistic outputs are, in this way, biased towards the output provided by the most

Algorithm 4.1 BN-based Combining Algorithm.

```
Input: Data, OutputBN1, OutputBN2
Discretise OutputBN1 and OutputBN2
Build and Train Model
Smooth Output_CPD
Update Model
Output: Model
```

accurate base classifier. Algorithm 4.2 was thus applied to the final output of the *combining BN*.

Algorithm 4.2 Accuracy Weighted Algorithm.

```
Input: Data, OutputBN1, OutputBN2, AccuracyBN1, AccuracyBN2
Align accuracyBN1 and accuracyBN2
Weighted_output = (outputBN1 · accuracyBN1 + outputBN2 · accuracyBN2) /
                / (accuracyBN1 + accuracyBN2)
Final_output = (output + Weighted_output) / 2
Output: Final_output
```

Where the prefix *Output* is the output of the raw *combining* BN and the suffixes $BN1$ and $BN2$ represent the two base classifiers' outputs and accuracies. For example, *OutputBN1* represents the raw output of the first network $BN1$. The *align* function aligns the base classifiers accuracies using the threshold values, in order to weight each base classifier's output on the corresponding accuracy for each threshold value on the testing dataset. *Weighted_output* is the new output weighted on the accuracies of the base classiifers, while *Final_output* is the final output of the classifier. For validation purposes, a combination of the outputs was also performed using a weighted voting approach (Woods et al., 1996). The weights were adjusted in relation to the accuracy of the base networks.

4.2.1 Experimental Results

Results were evaluated using 3-fold CV on Dataset A and 6-fold cross validation (CV) on Dataset B. K-fold CV randomly splits the data into K groups. Of these, $K-1$ parts are used to train the model and the remaining part is used to test model performance. The results from each K testing set are then arranged together to calculate the overall performance of the classifier. Different folds for the two datasets reflect the different sizes of the datasets, so that a consistent amount of data points was guaranteed to be tested at each iteration. The performances of the single and the *Combining Networks* are shown in Table 4.4 and Figures 4.8 - 4.9. The specificity cut-off of 90% was chosen for consistency with other tests and in relation to glaucoma clinical management. False positives in glaucoma lead to overtreatment and cost in clinical management, so methods with high false positive rates, over 15%, are generally considered as not clinically useful.

Table 4.4: Performances of base and combined BNs in terms of mean AUROC and total errors at maximum accuracy and at 90% specificity.

	Dataset A			Dataset B		
	AUROC	Errors	Errors at 90% spec	AUROC	Errors	Errors at 90% spec
Semi-Supervised BN	0.98	6	13	0.87	84	206
Anatomy-Driven BN	0.98	7	13	0.75	110	326
Weighted Vote Combined	0.98	5	13	0.84	90	252
BN-based Combined	0.98	8	12	0.87	93	186
Accuracy Weighted BN	0.98	5	12	0.85	93	118

Considering the overall performances on Datasets A and B, the semi-supervised data-driven BN performs better than the anatomy-driven BN, and is at least comparable to the combined ones. On Dataset A, however, the performances of all the classifiers are comparable. Among the combined classifiers, the Weighted Vote is outperformed only at high specificities, but in Dataset B it is outperformed by both the other combined networks.

The BN combined networks perform well in both datasets, and were the best classifiers at high specificities.

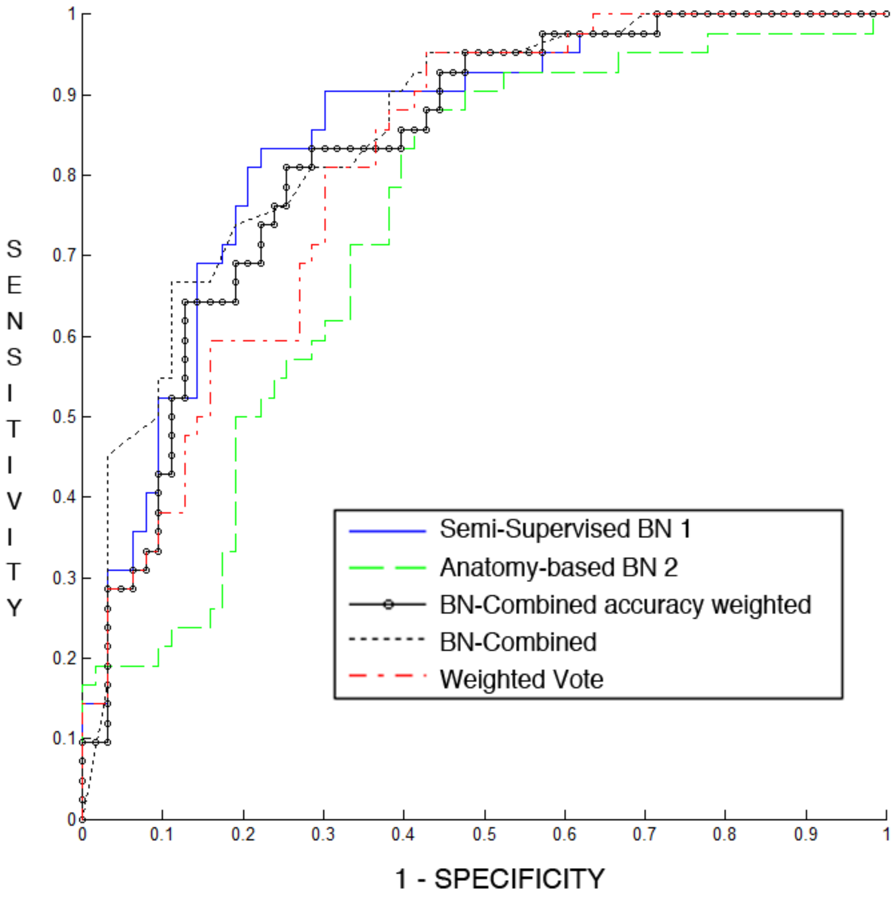


Figure 4.8: ROC curves of base and combined classifiers tested with 6-fold CV test on Dataset B.

Looking at the ROC curves in a particular test on Dataset B (Figure 4.8) it can be seen that the performances of the base semi-supervised data-driven BN are highest at mid-range specificities. However, at higher specificities, the BN combined networks are comparable or outperform it, as shown in Table 4.4. On Dataset A, the performance of the two base BNs are more similar and often the anatomy-driven BN outperforms the others (Figure 4.9). The Weighted Vote classifier performs better at low specificities, as pointed out considering the total errors.

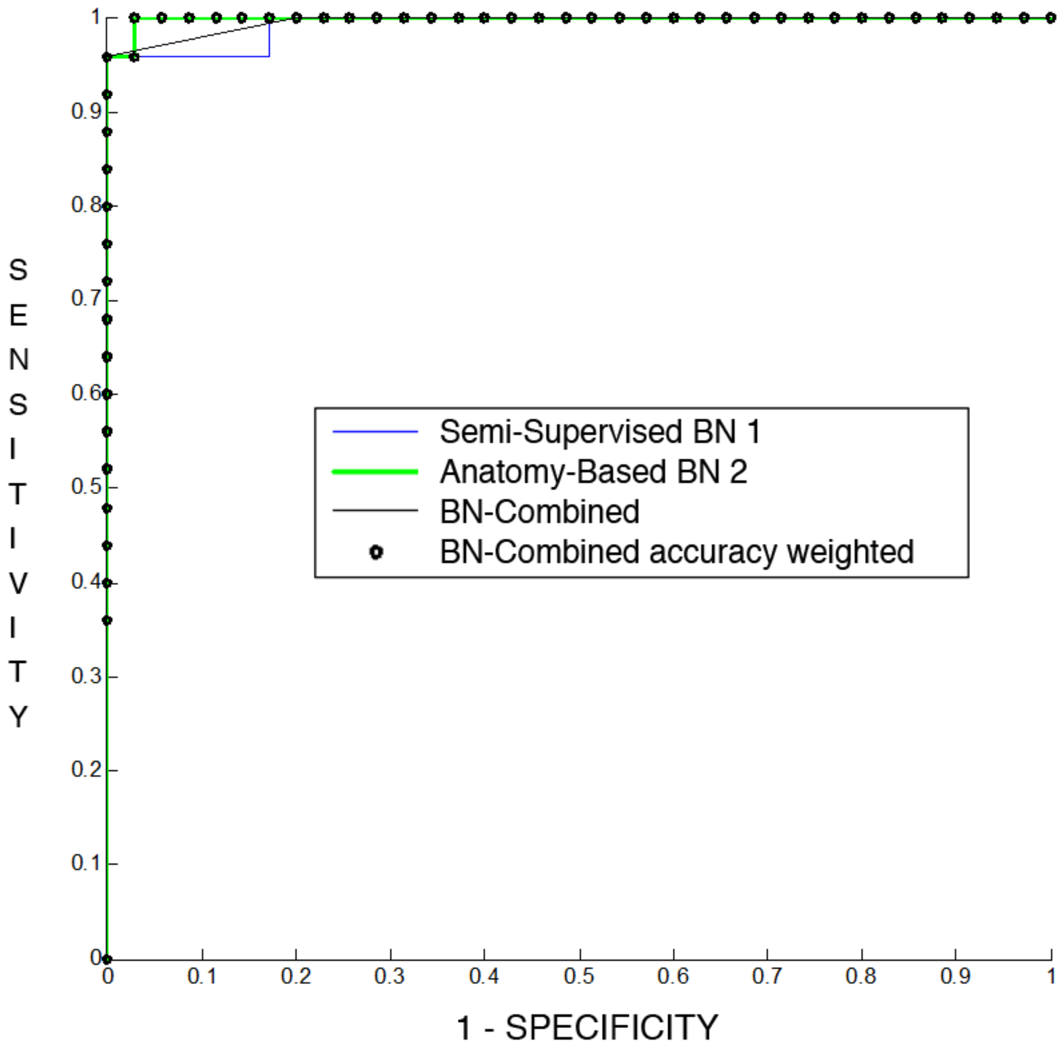
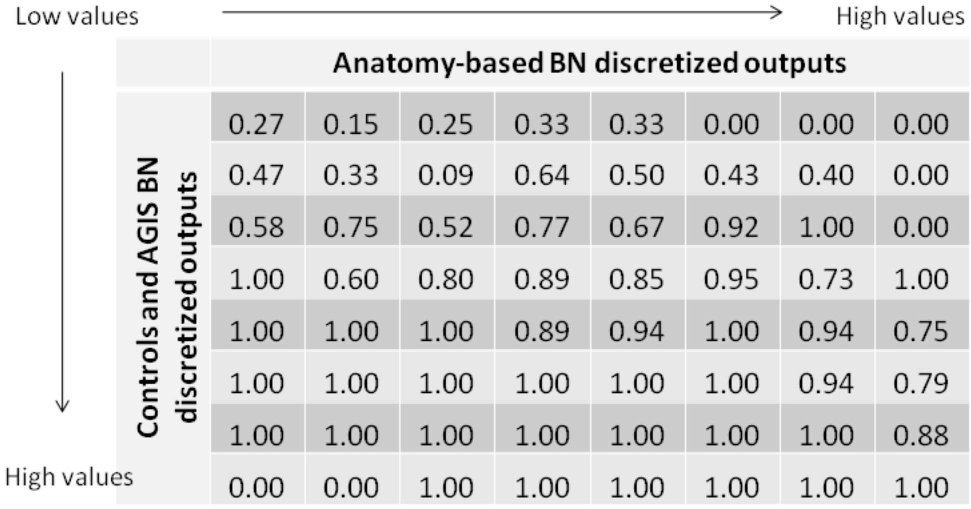


Figure 4.9: ROC curves of base and combined classifiers tested with 3-fold CV test on Dataset A.

4.2.2 Discussion

The semi-supervised network based on controls and AGIS is clearly performing very well at all specificities, confirming that using the conservative nature of AGIS metric score and the idea of modelling healthy subjects is effective for classification in glaucoma. However, the bias introduced by using the AGIS score in the inclusion criteria and its unreliability must be kept in mind. On the other hand, as already seen in independent dataset testing, the anatomy-driven network does not perform very well. However, the different results obtained with this model suggest that improvements in performance can be obtained by using as a base classifier in combination with others. In fact, diversity and accuracy of base classifiers are key factors for obtaining better performing ensembles of classifiers. Diversity was obtained by using an unsupervised approach on the anatomy-based network. This consequently decreased the performance of the base classifier but increased performance for the combined one, especially at higher specificities. The performance of the accuracy Weighted Vote classifier decreased, in relative terms, when the specificity increased. This is also the case in the BN combining network weighted on the accuracy, as expected. Nevertheless, the simple BN combining classifier is not biased towards high accuracy and this allows it to outperform all the other classifiers at high specificities. In fact, looking at the ROC curves it can be seen that, with an increase of accuracy of the anatomy-based network, there is a decrease in performance of the accuracy weighted ones; in this particular case this is due to the anatomy-driven network that does not outperform the other models at any specificity. However, on Dataset A, the opposite was observed (Figure 4.9), i.e. the BN Accuracy-Weighted classifier outperformed the non-weighted one. The higher number of total errors observed in Dataset A may be explained by the differences between this dataset and Dataset B: in Dataset A the diversity between the results of the base classifiers is lower than in Dataset B, leading to worse combined performance. This highlights the importance of choosing the most efficient base classifiers, and further studies on generative ensembles of classifiers (i.e. active learning of base classifiers to improve diversity and accuracy of base classifiers) (Jacobs et al., 1991) is required. It must also be pointed out that the structures of the base classifiers were learned and trained on Dataset A, introducing a bias with respect to the performance. It is worth reiterating that the AGIS score was used in the inclusion criteria for both datasets, making it difficult to compare performances of classifiers supervised on AGIS itself or unsupervised. The difference in performances obtained with the two base classifiers and the two datasets highlights the potential of combining data. The data of interest is very "noisy" due to the absence of a gold standard and the high variability of the measurements, therefore a generalised algorithm that accords itself to the best performing network independently from the given data is desirable. This was not possible using accuracy weight, as the training dataset and the testing dataset were different in nature: when using Dataset A to train an accuracy-based network and Dataset B to test it, results will be biased by the different accuracies for the same networks in each dataset. A solution would be to use an independent dataset of the same type (cross- sectional or longitudinal) for training and testing; using cross-sectional data to build models designed for longitudinal data is not advisable for glaucoma (Artes and Chauhan, 2005). Further, a broader search on both datasets for a network that shows more accurate and diverse performances than the other could lead in this case to better results for both datasets.


Low values
High values
High values

	Anatomy-based BN discretized outputs							
Controls and AGIS BN discretized outputs	0.27	0.15	0.25	0.33	0.33	0.00	0.00	0.00
	0.47	0.33	0.09	0.64	0.50	0.43	0.40	0.00
	0.58	0.75	0.52	0.77	0.67	0.92	1.00	0.00
	1.00	0.60	0.80	0.89	0.85	0.95	0.73	1.00
	1.00	1.00	1.00	0.89	0.94	1.00	0.94	0.75
	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.79
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88
	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 4.10: Instance of a raw CPD of the class node. Each value represents the probability of positive output for each combination of base classifiers inputs (i.e. the final output of the combining network).

An interesting advantage offered by the BN combining classifiers is the possibility to observe the CPD for the class node. This gives interesting insights about how the network works and how the two outputs combine to obtain better results. In Figure 4.10 an instance of a CPD is shown; for each combination of discretised outputs of the base classifiers, a probability is reported in the table. This reflects the output of the final classifier; in particular, it is the probability of diagnosis of glaucoma for each combination of inputs. For example, the probability of observing an output of 7 and 4 for the anatomy-based BN and the data-driven BN, respectively, is 0.73. Note that an output of 7 corresponds to a very high probability of glaucoma for the base classifiers, as the outputs were discretised in 8 states. Some 0 values are obtained in the corners of the matrix, due to the lack of data for these combinations of outputs: these cells are smoothed with the mean-filter application. By observing the probability values in relation to their position, it can be noted that if any of the base classifiers observe a low probability of glaucoma then the combined output will be low. Also, if any of the two observe a high probability, the output is likely to be high. However, the higher values in the matrix are skewed towards the bottom-left, so that the AGIS semi-supervised network is more reliable at high specificities. Looking at lower values of the semi-supervised network (e.g. value 3), the output of the anatomy network increases the probability of glaucoma sensibly, adding its knowledge to the result. Several different instances of this matrix have been found in this study, again indicating high variability between different datasets used. Thus, exploration of the CPD of the combined network confirms higher performances of semi-supervised networks at high specificities, and it also gives a quantitative measure of the improvement that each single base classifier brings to the final output.

4.3 Clustering Bayesian Networks

Clustering is another technique that can be used to improve the performance of BNs and gain insights into data. When performed with BNs, it represents a model-based clustering technique that inherits all the qualities of BNs, including efficient computations, high performances in classification, the intrinsic ability to deal with noisy data and the clear graphical representation of the model. In this section, pointwise functional data from Datasets A and B were used, i.e. 52-point VF raw sensitivity values (i.e. not age-corrected) obtained with the HFA II.

In glaucoma, the standard approach is to perform clustering on glaucomatous subjects only or to discard class information and perform clustering on the whole dataset (Goldbaum et al., 2002). However, clustering only on glaucoma patients is not useful for classification because it does not take into account healthy subjects. On the other hand, clustering on the whole dataset may result in discovering the artificial differences between healthy and glaucomatous subjects, reflecting the inclusion criteria for the dataset. Our approach, instead, aims to gain performance and insights by applying clustering and classification simultaneously. In BNs this can be obtained by including an unobserved variable in the model that is related to the classification variable. Figure 4.11 shows the network used for this task. A class node (C) is linked to all variables, which are in turn dependent on the clustering node (i.e. H hidden variable).

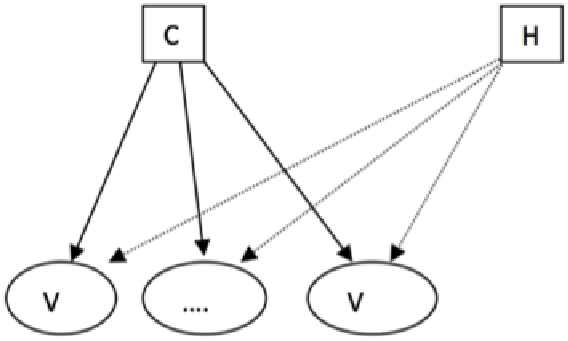


Figure 4.11: Clustering Naive Bayes Network. Squares represent Class (C) and Hidden (H) nodes, while ellipses represent VF sensitivity nodes (V).

Considering the dependencies between the variables in the model, the clustering and the class nodes are marginally independent. However, it can be demonstrated that they are conditionally dependent on the observed variables V (Wellman and Henrion, 1993). This effect has been described in Chapter 3 where one or more variables (i.e. parents) converge to an observed one (i.e. child). Given the observed child, observing the value of one parent will act on the parameters of the other parents. Intuitively, it can be explained by the fact that the parents are competing to explain the observed value. In other words, knowing that a variable belongs to a cluster acts on the probabilistic output for the classification node. In the same way, knowing the class value of the data changes the clustering as well. Since, in the training dataset, the class node represents the AGIS-dependent glaucoma label, the variance associated with the AGIS label is explained away by the class node. The clustering uses the remaining variance to separate the data without regards to the AGIS score, which in turn has an impact when the classification is performed on new data. The rationale here is based on the fact that there might be other factors, either shared, or not, by healthy and glaucomatous subjects, competing to explain the observed values. The hidden variable describes these factors and uses them to perform a better classification. For example, there may be particular VF patterns that are shared by certain people which are particularly helpful to indicate glaucoma in that group. Capturing and using this difference to calculate the classification probabilistic output could lead to better classification performance. To summarise, this model is able to characterise the different clusters of people who share patterns and relationships not directly dependent on the scoring metric in use, while at the same time exploiting this information in the classification process. Moreover, it is possible to quantify the added utility by building a ”confidence” index for each of the clusters of data, which can be directly used in clinical practice.

4.3.1 Clustering Simulated Annealing

Regarding the structure learning process, the Structural Expectation-Maximization (SEM) algorithm can be used. This extends the Expectation-Maximization (EM) algorithm to structure searching, as introduced in Chapter 3. The algorithm iterates over two steps: in the first step, the model that maximises the score with the present instantiation of the parameters is chosen, while in the second step the parameters that maximise the score with the present model are selected. In analogy with the EM algorithm, SEM uses the expected values for the unobserved parameters while performing the model search. This allows one to learn a set of relationships, and the parameters of the model, without using any clinical metric. SEM is one of the most commonly used techniques to deal with missing data and structure learning for BNs, although it is not the only method that can be used. This section introduces a technique which performs clustering and structure learning using a SA framework (CSA). As already shown, SA has been shown to perform particularly well with BNs. The differences of CSA with respect to SEM lie in the single-stepwise nature of the SA algorithm and in its convergence qualities. In particular, the unobserved data are treated as observed, starting with a random grouping of the data. At each further iteration, the cluster of one randomly picked patient is changed, until convergence is found. It should be noted that the EM step is skipped in favour of a traditional scoring of the present complete data. This leads to a faster model evaluation, although more iterations are needed to stabilise the groups of patients. While the application of such a technique to BN-based clustering is novel, to our knowledge, the rationale has been investigated in (Selim and Alsultan, 1991) and it has been shown that clustering benefits from the annealing algorithm, obtaining many advantages with respect to other methods, e.g. to K-means clustering. The pseudocode of the algorithm is described in Algorithm 4.3. Where the input parameters *Temperature*, *Cooling_Factor* and *Max_Iterations* regulate the convergence of the algorithm, *Model* is the starting model and *Data* is the training dataset in use. *Best_Score* is used to store the current best model score, while *Score* and *New_Score* hold respectively the score of the current and the model explored at each iteration. The latter is evaluated after either an operation on the model structure (adding an arc, deleting an arc and moving an arc) or a random reassignment of a data point to a cluster. Only arcs between leaf nodes are permitted when peforming operations. The probability of carrying out one operation or the other is regulated by the *Selection_Factor*. Note that *Random_Number* represents a random number between 0 and 1. If the new score is higher than the current score, the parameters of the current model and the current score are updated. If the new solution is the best among the explored ones, the results are also stored in *Best_Score* and for final output. If the new solution is worse than the current one, a temperature factor *Temperature* regulates whether to accept it as a new current solution. The temperature factor is lowered at each iteration through a cooling factor *Cooling_Factor*. After a number of iteration *Max_Iterations*, the algorithm returns the best explored network and clusters configuration. Typical parameters used in the experiments were 15000 *Max_Iterations*, a starting temperature of 15, a cooling factor of 0.99 and *Selection_Factor* equal to 0.15.

4.3.2 Experimental Results

The ROC curves of the clustering BNs implemented are presented in Figure 4.12. As suggested by (McNeil and Hanley, 1984), in cases where ROC curves cross, global AUROC comparisons may not be indicative. The partial AUROC values can instead be considered

Algorithm 4.3 Clustering Simulated Annealing.

```
Input: Temperature, Selection_Factor, Cooling_Factor, Model, Data, Max_Iterations
Score Model and store in Best_Score
Loop on iterations until Max_Iterations is reached
    if (Random_Number is lower than Selection_Factor)
        Apply OPERATION to Model
    else
        Apply CLUSTER to Data
    end if
    Score Model and store in New_Score
    Diff_Score is New_Score - Score
    if (New_Score is higher than Score)
        Update Score, Data and Model
        if (New_Score is higher than Best_Score)
            Store Model and Data for Output
            Best_Score is New_Score
        end if
    elseif Random_Number is less than exp(Diff_Score / Temperature)
        Update Score, Data and Model
    end if
    Temperature is Temperature · Cooling Factor
end Loop
Output: Model, Data
```

by using the more clinically meaningful part of the ROC curve (McClish, 1989). Table 4.5 presents the partial AUROC values for specificities above 80%. The table shows a consistent improvement in performances at high specificities with the inclusion of the clustering process and structural learning when compared to the Naive Bayes classifier. Even if no significant difference was found over the entire ROC curve, the left part of Figure 4.12 seems to confirm the higher

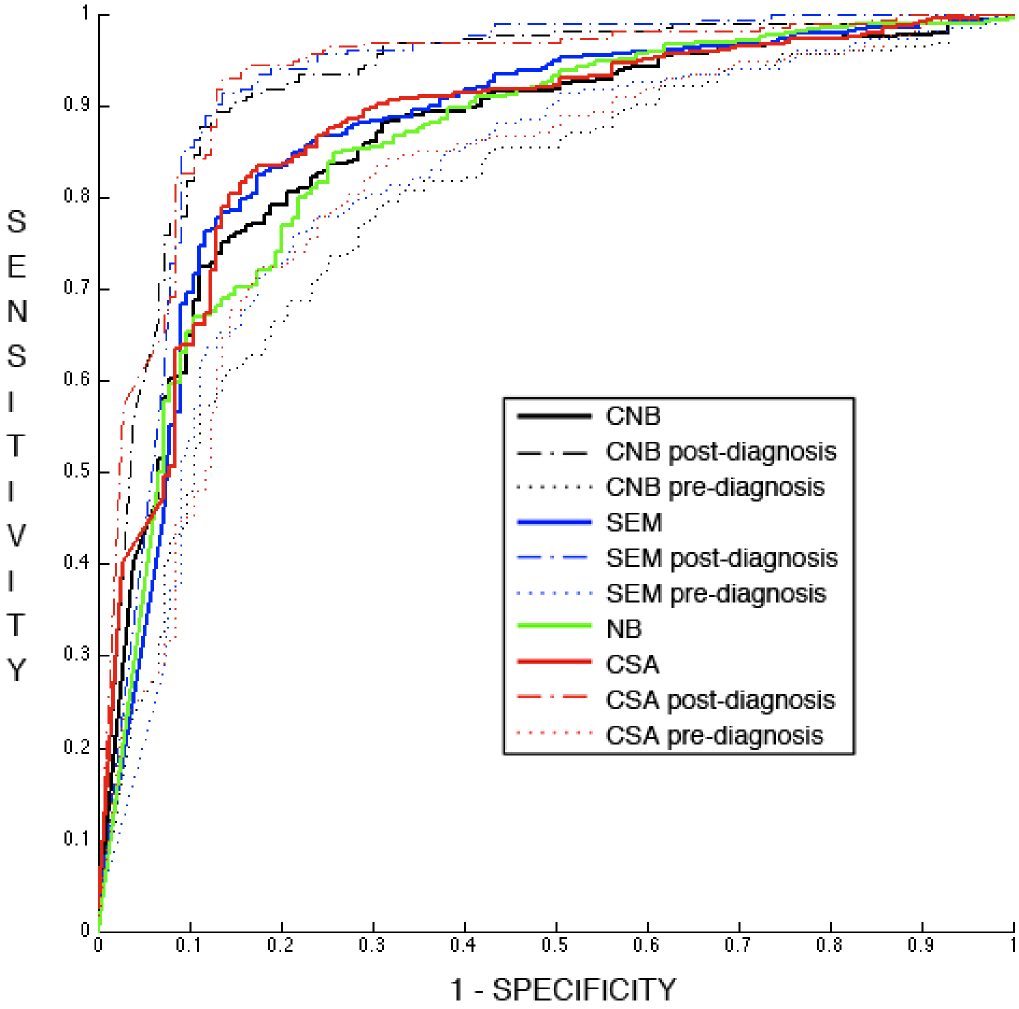


Figure 4.12: ROC curves for clustering BNs on pre-, post- and full testing dataset. The models included are Clustering Naive Bayes (CNB), Structural Expectation-Maximization (SEM), Naive Bayes (NB) and Clustering Structural Annealing (CSA). Pre- and post- diagnosis curves (thin lines) are above and below the correspondent overall performance curves (thick lines), respectively.

performances at high specificities of CSA, SEM, and CNB over NB. Among all, the best classifiers are SEM and CSA, which perform similarly to each other.

Regarding the structures learned using SEM and CSA, the respective results are shown

Table 4.5: Normalised partial AUROC Curve for Specificity > 80% for Bayesian classifiers. The models included are Clustering Naive Bayes (CNB), Structural Expectation-Maximization (SEM), Naive Bayes and Clustering Structural Annealing (CSA).

Classifier	Overall	Pre-Diagnosis	Post-Diagnosis
Naive Bayes	0.53	0.39	0.66
CNB	0.57	0.43	0.72
SEM	0.56	0.44	0.67
CSA	0.59	0.45	0.74

in Figure 4.13. The networks present several similarities, such as dense first order (adjacent location) spatial relationships and fewer distant relationships. However, CSA obtained a sparser network structure with arcuate, or partial arcuate, patterns in the central and peripheral VF, with a similar distribution to the known anatomical distribution of the retinal nerve fibre layer.

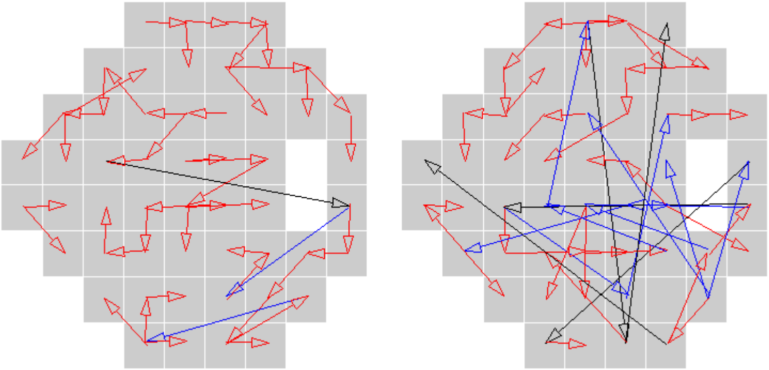


Figure 4.13: Structure learned with CSA (left) and SEM (right) superimposed to the VF map. Different colours help to highlight different spatial order relationships (red: < 2, blue: < 5, black: ≥ 5).

4.3.3 Clusters Analysis

The number of clusters in the analysis was chosen arbitrarily after a discussion with clinicians. Higher number of clusters may give more insightful results, although the lower number of elements in each cluster may affect the robustness of the results. To quantitatively measure the quality of the clusters, the silhouette mean index is used. This measure was first proposed by (Rousseeuw, 1987), and is based on the comparison of clusters tightness and separation. The mean silhouette index is obtained by averaging over the silhouette index of each element i, which is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4.2}$$

In the above formula, $a(i)$ is the dissimilarity between the element i and the elements in the same clusters and $b(i)$ is the dissimilarity from the elements in the other clusters. A silhouette mean value is close to 1 when elements are appropriately clustered and close to -1 when they are assigned to the less appropriate clusters. The difference between sensitivity values is used to measure dissimilarity between elements, and results for the assessed classifiers can be found in Appendix A (Table A.4). The table is not reported because silhouette values were moderate and similar for all BN models, but higher for the K-Means clustering. This result reflects the fact that since the silhouette score measures the distance between clusters, a technique that only separates healthy subjects from glaucomatous subjects would be expected to obtain a higher value. Therefore, the K-Means clustering performed similarly to a simple discrimination task, being unable to separate the data into groups based on variance other than that carried by the inclusion criteria used in the data collection, i.e. the AGIS metric.

The similarity between different clustering techniques was assessed using the Cohens Kappa index (Carletta, 1996). This index can be used to measure the agreement between two clustering techniques using the following formula:

$$Kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{4.3}$$

Where $P(a)$ is the observed agreement between two techniques (i.e. elements assigned to the same cluster) and $P(e)$ is the probability of chance agreement. Kappa value is 1 if there is complete agreement and 0 if there is no agreement. It must be noted that such a measure does not take into account the qualitative similarity between the clusters obtained. The Kappa values for the investigated models are reported in Table 4.6.

Table 4.6: K-Values for Bayesian classifiers. The models included are Clustering Naive Bayes (CNB), Structural Expectation-Maximization (SEM) and Clustering Structural Annealing (CSA).

	CNB	SEM	CSA
CNB	-	0.29	0.30
SEM	-	-	0.34
CSA	-	-	-

The clusters obtained with different BN clustering techniques are rather similar, showing good robustness in the data. This is reflected in a fairly high kappa value and in a relatively smooth distribution of the data across clusters, as will be shown later. K-Means clustering obtained qualitatively different clusters, therefore it was not included in Table 4.6.

A qualitative representation of the clusters obtained with CSA is shown in Figure 4.14. The upper section shows the VF maps learnt for each cluster. These were obtained by sampling from the conditional distribution $p(\mathbf{X}|C_i)$, where C_i is the observed cluster label ($i = 1..4$) and $\mathbf{X}$ is a vector of variables representing each point in the VF map. In the lower section, the distribution of glaucomatous and healthy subjects over clusters in the training dataset is shown for each technique. The clusters are aligned to maximise the Kappa value of SEM, CNB and K-Means with respect to the clusters obtained with the CSA technique. While relatively uniform distributions were obtained using BN clustering techniques, K-Means obtained sparse clusters effectively separating glaucomatous and healthy subjects. Figure 4.15 shows the results of the CSA algorithm on the testing dataset. The VF maps for each cluster are shown on the left side of the figure. Each point value in the VF map was obtained by averaging over the tested patients assigned to each cluster by the algorithm, separately for glaucomatous and healthy subjects. Age distributions are also included in the graphic, together with histograms above VF maps that converted, showing the age distribution for patients in that group. The red vertical lines in the histograms are the

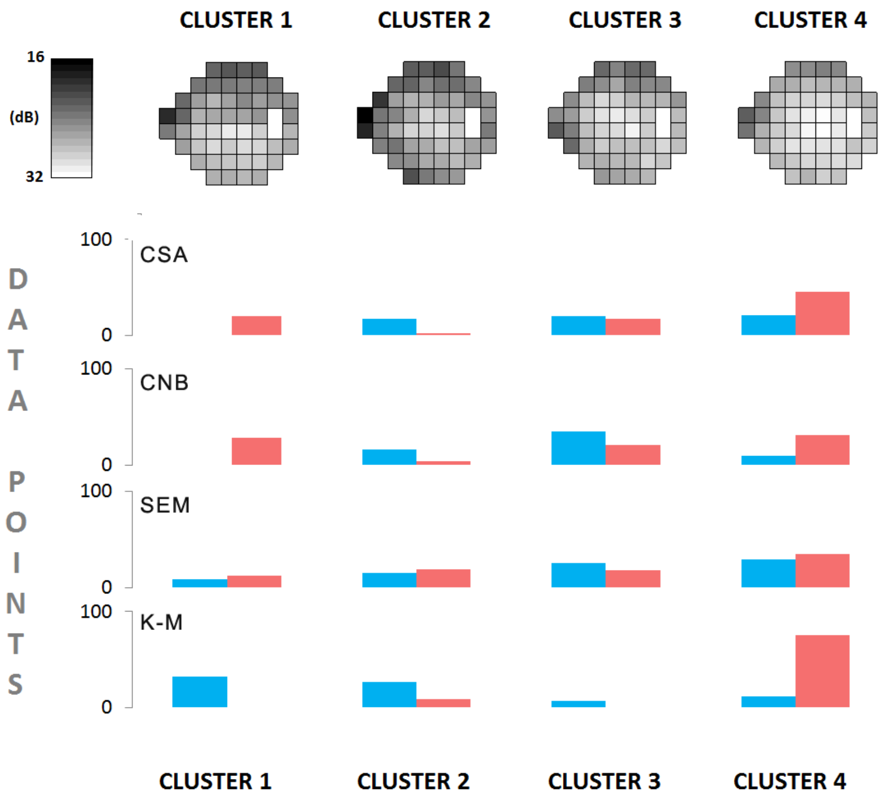


Figure 4.14: Models obtained from the training dataset. Above: Clusters obtained with CSA. Below: Distribution of data points for each cluster obtained with Clustering Naive Bayes (CNB), Structural Expectation- Maximization (SEM), K-Means (K-M) and Clustering Structural Annealing (CSA). For each histogram, left bar (blue) represents the number of healthy data points and right bar (red) represents the number of glaucomatous data points in a given cluster.

AGIS classifications, so that it is possible to compare the ability of CSA to capture late or early glaucoma with respect to this metric. For example, distributions skewed to the right indicate presence of late glaucoma in the cluster, which were mostly correctly captured by the AGIS metric. The right section of the figure shows the number of VFs assigned to each cluster, separately for healthy subjects, pre-diagnosis and post-diagnosis. Cluster 1, on which the highest discrimination power was found, shows a strong difference in sensitivity in VF hemifields (i.e. the difference in VF sensitivity between the upper and lower half of the VF). This is a characteristic pattern of glaucoma, and it is the rationale behind the Glaucoma

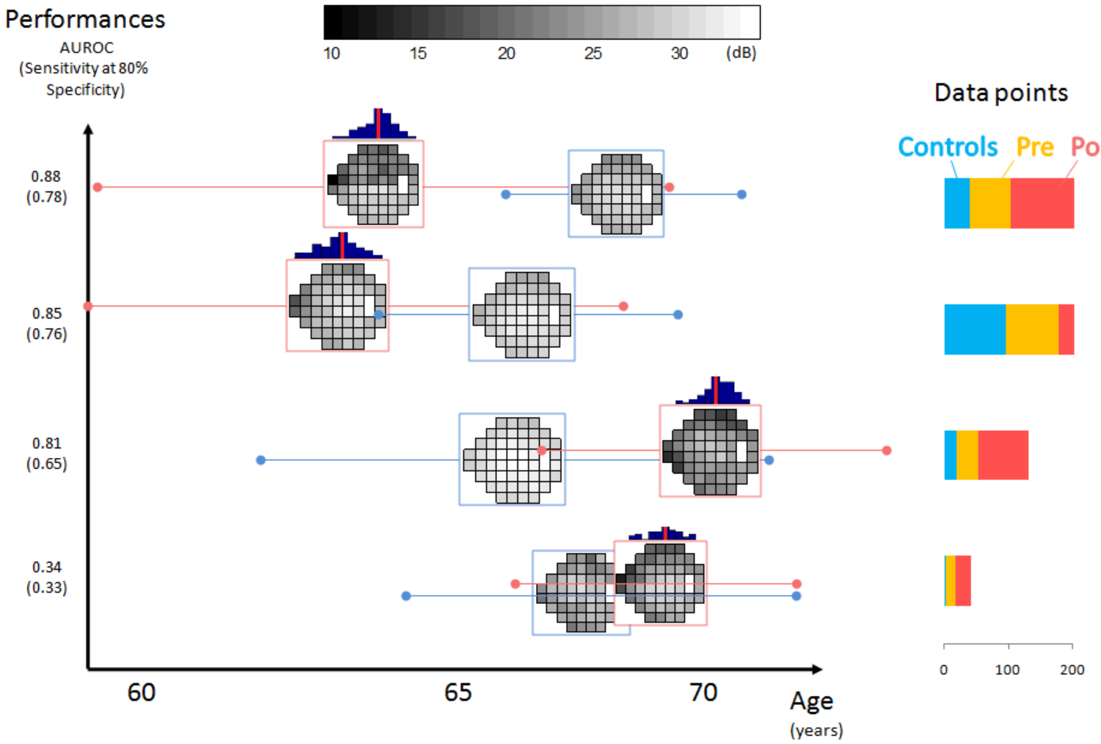


Figure 4.15: Results of CSA algorithm on the testing dataset. The four clusters are presented in four rows. The left section of the figure shows the VF maps for each cluster for Glaucomatous (red) and Healthy subjects (blue). The x-axis in the graph is the mean age for each group. The deviation from the mean age is also reported with whiskers for each VF map. Above each glaucomatous VF map, a histogram showing the distribution of the data points relative to the AGIS diagnosis is shown. The diagnosis is represented with a vertical red line in the centre of the histogram. The right section of the figure shows the number of data points assigned to each cluster for healthy subjects (Controls), pre-diagnosis (Pre) and post-diagnosis (Post).

Hemifield Test (Asman and Heijl, 1992), currently implemented in several VF screening machines. Cluster 4, which also performed well in classification, presents a nasal defect pattern, which is often described in literature as an early indicator of glaucoma (Drance, 1969; Spry and Johnson, 2002). Cluster 3 shows general functional impairment in the peripheral VF, which is another early sign of glaucoma development. Cluster 2 performed rather poorly as a discriminator between converters and healthy subjects, and included healthy subjects with ”bad” (unreliable) VF tests and glaucomatous patients with a non-uniform pattern of depressed points.

4.3.4 Discussion

By considering the clusters presented in Figures 4.14 and 4.15, several insights into glaucoma can be obtained that are of interest for clinical practice. The best discriminating cluster (Cluster 1) shows the greatest age difference between healthy and glaucomatous subjects, as well as the largest proportion of cases in the post-diagnosis group. This indicates that a strong hemifield difference is typical of later stages of glaucoma, and when occurring at a younger age likely indicates glaucoma. The higher proportion of post-diagnosis patients in a cluster means that the AGIS score is able to capture these cases too. This is also reflected in a right- skewed histogram above the VF map in Figure 4.15. The second best performing cluster (Cluster 4) characterises an early nasal defect, and it is also placed on the left side of the figure, meaning that if this pattern is observed in younger subjects also suggests glaucoma. In this case, AGIS seems to perform less well than the proposed model, with a high proportion of pre-diagnosis patients falling in this group. This may be due to the inclusion criteria (conversion defined as three successive positive AGIS scores) but also due to the nature of AGIS scoring. The AGIS score reflects the pointwise VF sensitivity deviation relatively to age-corrected healthy individuals and labels points showing a depression below a certain threshold. For example, in the nasal area, the threshold is an 8 dB loss. The threshold criterion and the reference model may contribute to the overlooked cases, i.e. false negatives, obtained by AGIS classification of cases in Cluster 4. Our technique, instead, uses absolute pointwise values with different parameterisation for each cluster and was able to capture the defect earlier. Cluster 3 obtained acceptable classification performance, modelling mainly post-diagnosis glaucoma developing peripheral VF glaucomatous damage at an older age. Cluster 2, instead, is difficult to characterise because of the small number of cases allocated and because of its inhomogeneous appearance, which is probably due to unreliable VF tests.

In comparison to SEM, CSA obtained a sparser and ”cleaner” structure. This is probably due to the broader search space that the algorithm explores in comparison with SEM. The CSA network structure learned on the training data shows strong agreement with typical progression paths in the VF and stronger similarity to the distribution of the RNFL. Similar dependencies were found in (Gardiner et al., 2004), in which a VF filter was found to be in accordance with the accepted physiological shape of the RNFL. In (Strouthidis et al., 2006), higher correlations between VF and ONH measures were found in peripheral VF areas, which is confirmed by the higher density of the CSA structure on corresponding sectors.

4.4 Summary

This chapter explored the impact of metrics on BNs and combinations of BN models to improve classification performance and gain new insights. With regards to metric bias and data integration, the positive impact of combinations of anatomy and traditional metrics was measured, showing that even if the use of classic metrics alone is not sufficiently precise in early glaucoma classification, BNs built using both metrics obtain good results. In particular, if combined with semi-supervised unbiased structures, the AGIS score performed the best for high specificities. If combined with MRA, it obtained among the best results at lower specificities. A new combining algorithm, using a BN-based ensemble of classifiers, demonstrated that the performance of the combined classifier is better where results for the base classifiers are different. It was also possible to investigate the nature of anatomy and data-driven interactions, obtaining interesting insights and some performance improvements. Further investigations into glaucoma mechanisms and variability of VF data were carried out using clustering approaches, and in particular a novel clustering technique. This allowed us to understand inter- and intra-subject variability, and the impact of metrics on diagnosis. Several patterns of relationhips between the variables were found to be supported by previous clinical research, while others represent new interesting insights for clinicians.

The incorporation of time-dependent relationships in the model represents the next natural step to explore glaucoma classification and understand its progression. To this aim, the next chapter will propose several BN-based models that deal with longitudinal datasets.

Chapter 5

Bayesian Networks for Longitudinal Data

The BN models presented in the previous chapters explicitly take into account relationships between variables but not changes in the variables over time. In other words, time-dependent relationships are not modelled, even though progression is a key aspect of glaucomatous disease. We have already seen that functional and structural measurements are related in glaucoma, as indicated by the cross-sectional BNs constructed and discussed so far. Furthermore, evidence in clinical literature shows that the RNFL deteriorates simultaneously as the VF is impaired with the progression of the disease. In this context, assessing how relationships between variables in BNs change over time is of great interest for understanding glaucoma progression.

This chapter presents three BN models for longitudinal datasets, with the purpose of improving classification performance and exploring the mechanisms of glaucoma progression. In Section 5.1, the widely used Dynamic Bayesian Network model is presented. The limitations of this model in glaucoma are discussed and addressed using novel techniques proposed in Section 5.2, called the Dynamic Stages Bayesian Network. This BN model allows key stages to be extracted from the data and exploited for classification. Section 5.3 presents a novel BN for longitudinal data, called the Non-Stationary Clustering Bayesian Network, which allows clustering of temporal processes. Finally, Section 5.4 discusses the overall conclusions from this chapter.

5.1 Dynamic Bayesian Networks

To model time-dependent relationships in the BN framework, an extension called Dynamic BNs (DBNs) (Dean and Kanazawa, 1989) was presented in Chapter 2. Intuitively, DBNs are BNs in which directed arcs flow forward in time. Nodes are used to represent variables at different time points; therefore arcs can occur at the same time point and between different time points. In DBNs, learning and inference methods are very similar to those described for BNs (Dagum et al., 1992). A simple univariate example is shown in Figure 5.1.

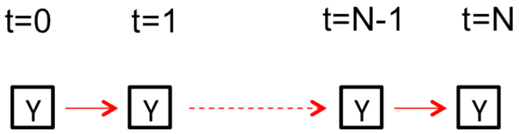


Figure 5.1: Simple Dynamic Bayesian Network.

Red arcs in the figure flow from left to right through the chain, i.e. from $t = 1, .., N$. The model does not represent dependencies between variables over more than one step so it is called a first-order Markov Process. In these models, each variable is directly influenced only by the last variable, i.e. $P(Y_1, Y_2, .., Y_N) = P(Y_1)P(Y2 \mid Y_1)...P(Y_N \mid Y_{N-1})$. In the BN framework, once the CPD of the nodes in the first two time points and the relationships within and between them are defined, the model is fully parameterised and it can be replicated for the remaining time points. An extension of these models is to make the observations dependent on a hidden variable, called *state* (H) so that the sequence of states is a Markov Process. This defines the so-called State Space Models (SSMs), shown in Figure 5.2. If the state variable assumes a set of discrete values, the model is called a Hidden Markov Model (HMM). HMMs were introduced in Chapter 2 and are one of the most commonly used DBNs. In general, DBNs represent a general framework that include several configurations of variables with continuous and discrete distributions, such as Factorial HMM, Kalmann

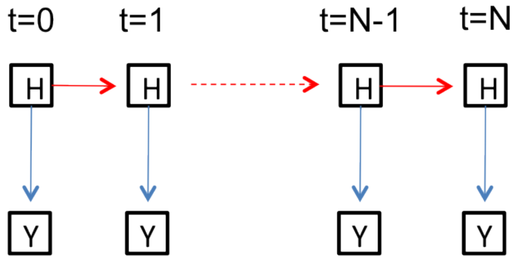


Figure 5.2: State Space Model.

Filters, and many other widely used models.

Although DBNs can capture the time-dependent nature of relationships in the data, they are models for stationary processes, therefore parameters and relationships between variables are not allowed to change over time. Once the relationships are set for two time points, the model is replicated over time. In many medical contexts, however, dependency between variables can change over time. The results reported in the previous chapters have shown that relationships between variables in glaucoma vary at different disease severities, suggesting that an approach based on *stages of disease* is necessary.

5.2 Dynamic Stages Bayesian Networks

This section presents a novel BN model, called a Dynamic Stages Bayesian Network (DSBN) that can extract the key stages of glaucoma progression, and can be generalised to any temporal process. In the DSBN, time points are aligned and the pattern is extracted. For diseases such as glaucoma, where there is high variability in terms of rate and onset of progression, the idea of stretching and shrinking the time dimension of the individual time series to align them and identify common stages of disease may be useful for classification and gaining insights about the temporal process of progression.

It must be noted that the term *time series* typically refers to sets of data samples collected in time spaced at uniform time intervals (Chatfield, 1996; Brillinger, 1981). Instead, real-world glaucoma observations are irregularly sampled, being therefore *irregular time series.* For the remaining Chapters, however, the term *irregular* was dropped for simplicity and *time* *series* was used to refer also to series of samples collected at not-uniform time intervals.

5.2.1 Stages of Disease

In the DSBN, a *stage* represents a phase in the disease progression defined by a particular set of values of the variables observed. For example, a healthy VF may be considered the first stage of glaucoma, while a deep depression in the sensitivity of all sectors of the VF may be considered the last stage. In between the two extremes, VF disease stages are characterised by patterns and relationships between variables, such as early superior hemifield defects or general mild depression. In the DSBN, the stages are obtained from the data. Sequences of stages can also be considered, for instance by identifying common sets of consecutive stages. Sequences of stages are independent of time, but have only one direction of flow. The sequences of stages are also referred to as *temporal patterns*, which are not to be confused with spatial patterns in VF tests. A model that is able to extract and identify patients' sets of stages across time would address the issue of speed of glaucomatous progression. The irregular sampling of the data, the bias and noise in the data are also tackled in the DSBN by selecting certain data points in the learning process.

Modelling the stages of glaucoma in a BN framework can be obtained by explicitly assigning each stage to a variable in a static BN model. The idea is to link a set of stage-nodes to a classification node, as shown in Figure 5.3. The number of leaf nodes is arbitrarily chosen.

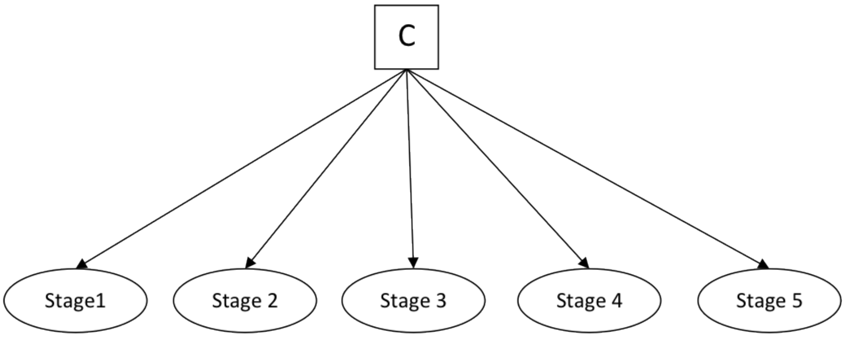


Figure 5.3: Dynamic Stages Bayesian Network.

Intuitively, each time series contributes to the model with a number of data points equal to the number of leaf nodes, so that for a dataset of K time series and a model with N leaf nodes, the size of the training data used is KxN. Time series with less time points than leaf nodes will be excluded from the process. N data points are sampled from each series so that only the most informative data points are used.

5.2.2 Learning Dynamic Stages Bayesian Networks

For a model with N leaf nodes, the pseudocode for the learning algorithm is reported in Algorithm 5.1. This is based upon the SA search described in Chapter 3 (Algorithm 3.1), so that several variables correspond to those described previously. In particular, $Temperature$, *Cooling_Factor* and *Max_Iterations* are parameters that regulate the SA search.

The algorithm starts with a random DSBN model $Model$, which undergoes a set of operations at each search iteration. The first of these is the $WARPING$ operation on one randomly selected time series. This operation samples N random data points from the full time series, so that there is one point from each time series at each stage. The operation $UPDATE$ then updates the data matrix with the new sampled N points for the selected time series. Note that in the selected subset of data points from each time series the ordering of the data is preserved, i.e. the order of the data points in a time series is a constraint as subsequent events cannot cross each other going backwards in time. The rest of the algorithm procedure is similar to the SA algorithm already described in Algorithm 3.1, which compares the explored solution with the current solution and the best solution explored so far. If a better solution is found, the model is updated, i.e. reparameterised using the data matrix updated in the $UPDATE$ operation. Intuitively, sampling N data points from each time series leads to an 'alignment' of the time series, because the marginal likelihood of each node is higher for more similar data points. In other words, from each time series the subset of points that gives a better total likelihood of the model is chosen, i.e. the N most similar stages through which all time series pass will be identified by the model. An example of the alignment process can be seen in Figure 5.4. When convergence is reached, the best alignment of either the glaucoma or controls data is obtained. The points that are left out from the time series are those that are worse at discriminating between the two classes and at

Algorithm 5.1 Learning the DSBN Algorithm.

```
Input: Temperature, Cooling_Factor, Model, Data,
          Max_Iterations
Score Model and store in Best_Score
Loop on iterations until Max_Iterations is reached
     Select Random Time_Series
     Apply WARPING to Time_Series
     UPDATE Data
     Score Model and store in New_Score
     Diff _Score is New_Score - Score
     if (New_Score is higher than Score)
          Update Score, Data and Model
          if (New_Score is higher than Best_Score)
               Store Model and Data for Output
               Best_Score is New Score
          end if
     elseif Random_Number is less than exp(Diff_Score / Temperature)
          Update Score, Data and Model
     end if
     Temperature is Temperature  ·  Cooling_Factor
end Loop
Output: Model, Data
```

characterising the patterns of progression. This set of discarded points is likely to constitute deviations due to noise or variability.

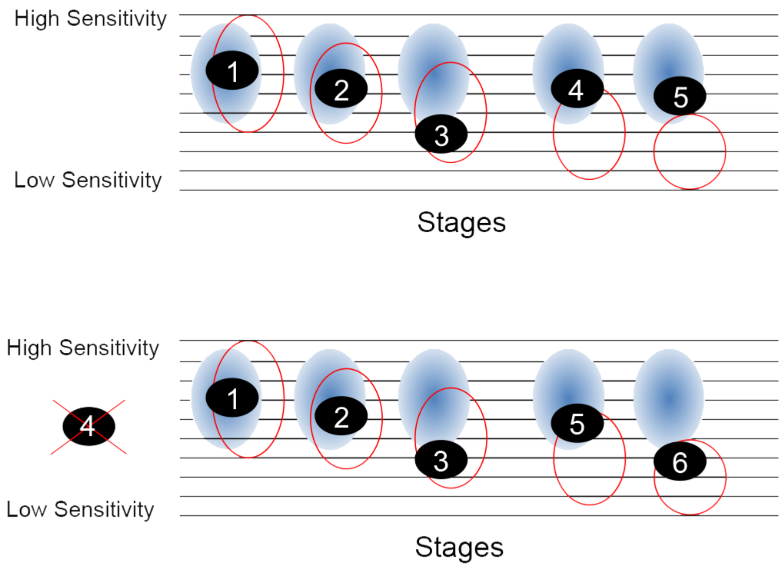


Figure 5.4: Scheme of the warping algorithm in two steps (from top to bottom). Data points are represented with incrementing numbers indicating their cardinal position in the time series. Red and blue circles represent glaucomatous and controls parameter distributions for each stage, i.e. continuous Gaussian distributions, respectively. Data points 1 to 5 (top) are all fitted in the best set of parameter distributions of DSBN (up), in this case controls. When data point number 6 is considered (bottom), data point number 4 is left out because a higher likelihood is obtained with this new alignment of the data. Notice that the order of subsequent visits is kept.

5.2.2.1 Experimental Results

The DSBN model was tested on Dataset B using only one global sensitivity value, obtained by averaging sensitivity over all VF points. SA was run 10 times with 3000 iterations (initial temperature of 7 and cooling factor of 0.95) and the best scoring solution was chosen. Random restart was used at each run. The results obtained are shown in Figure 5.5. A clear descending pattern in sensitivity values can be seen for glaucomatous patients, which shows their progressive loss of functionality. The key stages are meaningful as part of this pattern. On the other hand, healthy subjects seem to keep stationary high functionality, even though fluctuation is present throughout the pattern. As observed in literature, even healthy subjects are subject to a physiological loss of functionality and high variability. However, patients can worsen suddenly, and lose their vision at different rates (Haas et al., 1986; Spry

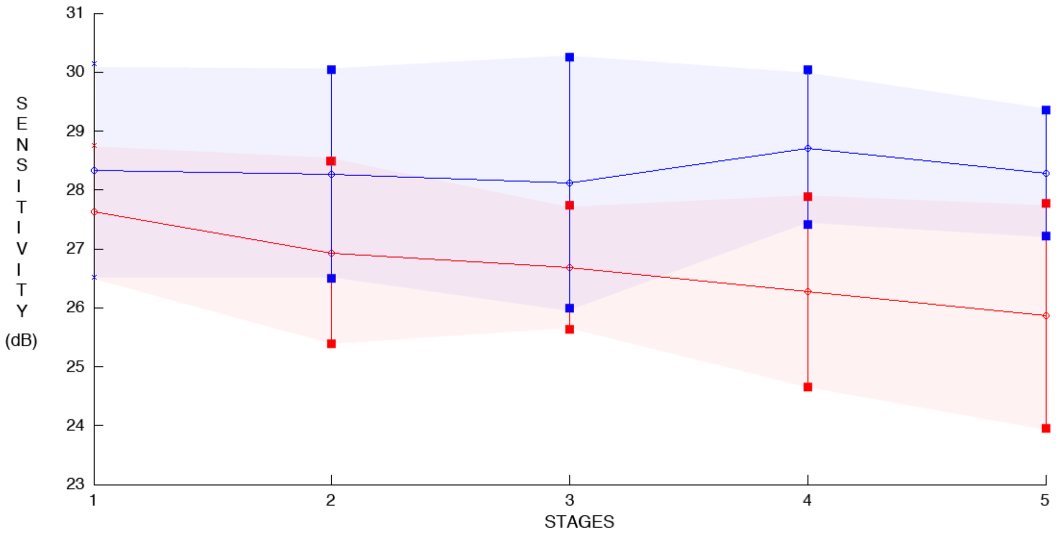


Figure 5.5: Key stages identified using DSBN for glaucoma (red) and control (blue) subjects, in terms of sensitivity (dB) mean and standard deviation (areas) values for each stage.

and Johnson, 2002). The key stages identified confirm this, as time points are aligned and the pure temporal pattern is left. The physiological deterioration of visual function was taken out of the model, because it was not a useful discriminator between controls and glaucomatous patients. This explains why no progression was observed for control subjects, but there was mild fluctuation and a slight increase in sensitivity at Stage 4. The standard deviation (SD) for different stages is also of interest, being different for the controls and glaucomatous patterns. For controls it reduces over the stages, while for glaucomatous subjects it increases. Higher variability for glaucomatous subjects is also confirmed in literature (Flammer et al., 1984; Russell et al., 2012). For controls, the SD decreases over the stages, so that a subject with a good test performance is more likely to be correctly diagnosed as healthy at later stages than at early ones, most likely due to learning effects (Heijl and Bengtsson, 1996). For glaucoma patients, the mean values are also more variable in later stages. This is due to the nature of glaucoma and the fact that the model ”learns” that at later stages it is easier to discriminate between controls and glaucoma patients, i.e. it exploits time dependent information. However, for the glaucomatous group, the distributions have longer tails and this reflects the higher variability of VF measurements in the later stage of the disease.

5.2.3 Dynamic Stages Bayesian Networks for Classification

While the learning algorithm uses a SA-like approach to sample from the original full dataset, in the testing phase an exhaustive search can be used, as the search space is limited to one time series. Also, during the testing phase, as in clinical practice, the whole time series was not made available in full from the beginning. In fact, it may be needed to investigate if a patient has glaucoma at his first appointment, then at the second, and so on. Therefore, in the testing phase the exhaustive search is carried out by adding one data point at a time. At each iteration, the complexity of an exhaustive search follows the combination complexity of $C(n, N)$, where C is the combination of n data points taken N at a time without repetition. N is the number of stages and n the number of data points in the series. The complexity of the combination search is $O(n)$, but since this search is performed at each new point available, the complexity of the full search is $O(n^2)$. The complexity in calculating the model score at each iteration is not included in the calculation, and can make the algorithm particularly slow when increasing the number of variables included in the model.

5.2.3.1 Experimental Results

The performances and robustness of the algorithm were tested on Dataset B using CV, described earlier. Results for the DSBN, DBN (simple HMM model) and BN are shown in Figure 5.6 for 2-, 3-, 5-, 10- and 15- fold CV. K-fold CV randomly splits the data into K groups and then uses K 1 parts to train the model and 1 part to test model performance iteratively. The results from each K testing set are then arranged together to calculate the overall performance of the classifier. In the testing phase, for each node, the SD was set at the value obtained on the whole dataset (i.e. 25), in order to take into account the variability of the points discarded by the model during the learning phase.

The DSBN outperforms the other models for every K-fold CV tested, while the DBN and BN perform similarly to each other. Note that different Ks correspond to different number of elements in the training and testing datasets. For the experiments reported with

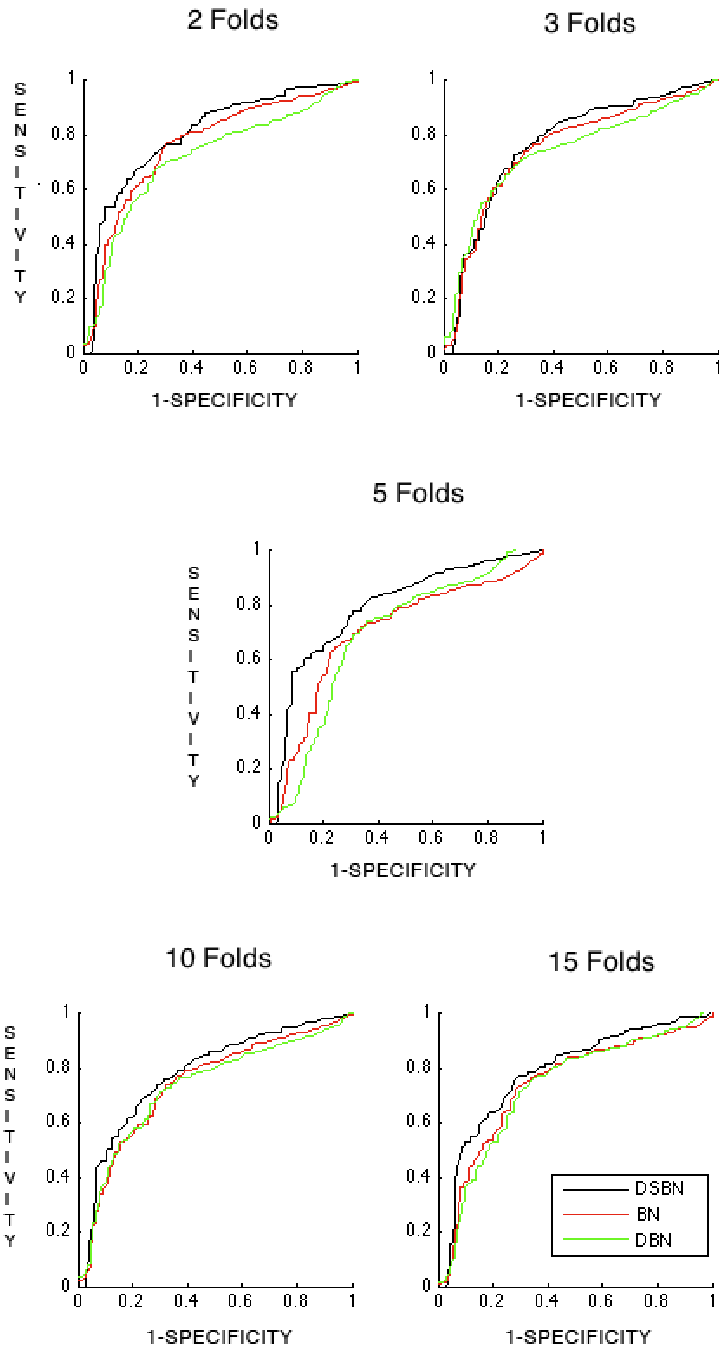


Figure 5.6: ROC curves obtained on different K-fold CV for the DSBN, DBN and BN.

increasing K, the sizes of the training datasets were 31, 41, 49, 55, and 56, respectively. The average AUROC values for the CV tests were 0.73 for the BN, 0.71 for the DBN and 0.78 for the DSBN. An ANOVA test confirmed the difference in AUROC values ($p = 0.03$). To assess normality of the AUROCs, the Shapiro-Wilk normality test was carried out (Shapiro and Wilk, 1965). This test, which is appropriate for small sample sizes, did not reject the null hypothesis of normality for all sets of AUROC ($p > 0.05$). Therefore, DSBN models discriminate between glaucomatous patients and controls better than the other approaches implemented. Figure 5.7 shows an example of how the algorithm can be used to characterise different patients by looking at their disease progression in terms of stages. For instance, in the top of Case 1, it is possible to observe that a patient's visual function decreases smoothly over time, and this corresponds to an increased probability of having glaucoma (bottom of Case 1). However, a different patient shown in the top of Case 2, remains at the same stage of glaucoma for a long time, and then quickly deteriorates to stages 3-5. It is also very interesting to look at the discarded data, i.e. the noise component of the time series; we see that this patient's VF sensitivity dramatically improves from the last time point on two occasions. This noise, due to short term variability, is appropriately discarded by the algorithm, and the corresponding glaucoma probability value (bottom of Case 2 5.7) remains at 100%. This differs from static BNs, where each visit is treated independently from previous visits.

5.2.4 Discussion

There are several advantages of the DSBN model over static BNs and DBNs. Even if constructed with a single average variable, the results shown above suggest that the DSBN identified meaningful stages of disease and obtained better classification performances than DBNs and static BNs. The data can be combined further by exploiting the ability of BNs to model underlying relationships in the data when multiple variables are observed. The pattern obtained with one variable can be compared to multiple variables and other data types can easily be added to the model. In this context, post-processing of aligned data can show interesting insights. For example, mapping the aligned data back to the VF sensitivity

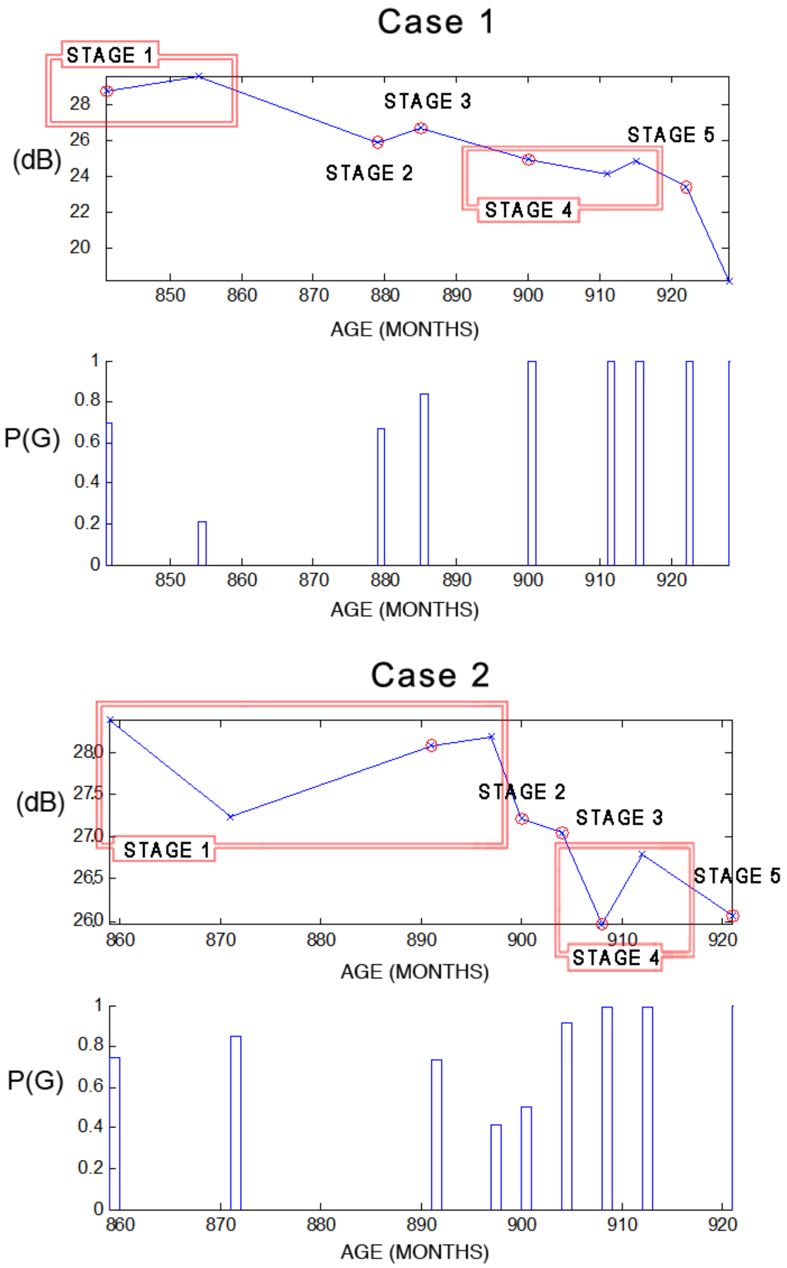


Figure 5.7: Sensitivity values in dB over time in months (top) and probability of having glaucoma according to DSBN over time in months (bottom) for two patients (left and right). Red circled data represent the selected data for each stage after the last visit.

in the original data produces sectorial pattern profiles as shown in Figure 5.8. The numbers

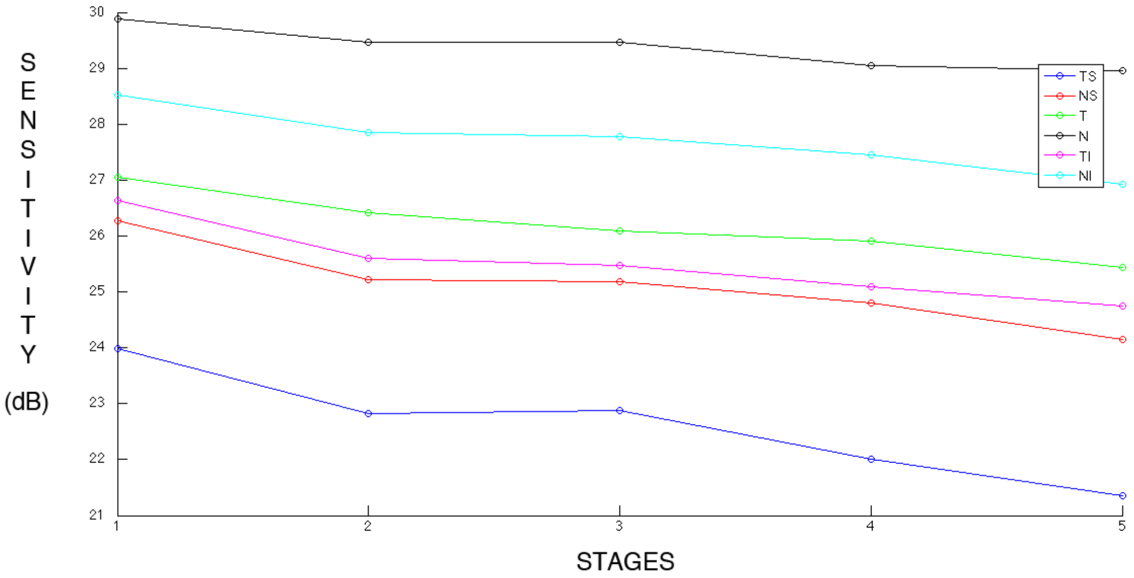


Figure 5.8: Key stages on sectorial data post-processed from the DSBN results obtained on one average variable. Sectorial variables are Temporal-Superior (TS), Nasal-Superior (NS), Nasal (N) , Temporal (T), Temporal-Inferior (TI) and Nasal-Inferior (NI).

represent the actual rate of decrease in terms of dB difference between states. Note that these results come entirely from the glaucomatous data without any external conversion indicator or metric. The progression between the first and second stages shows a decrease in functionality mainly in the arcuate sectors (TS, NS and TI), while the next disease step is mainly related with a decrease in the temporal sector. The last stage seems to occur when all the sectors decrease, which corresponds to the final stage of the disease.

The relationship between discarded and aligned data can also be used to gain insights. When N points are selected from each series, the remaining points can be analysed. The histograms of the discarded glaucomatous data between each stage are shown on the right hand side of Figure 5.9, while the data kept in the model for glaucomatous patients is presented on the left hand side. Note the distribution of noise, showing long tails towards the final stages.

Among the possible extensions of DSBNs, adding variables to each stage and accounting

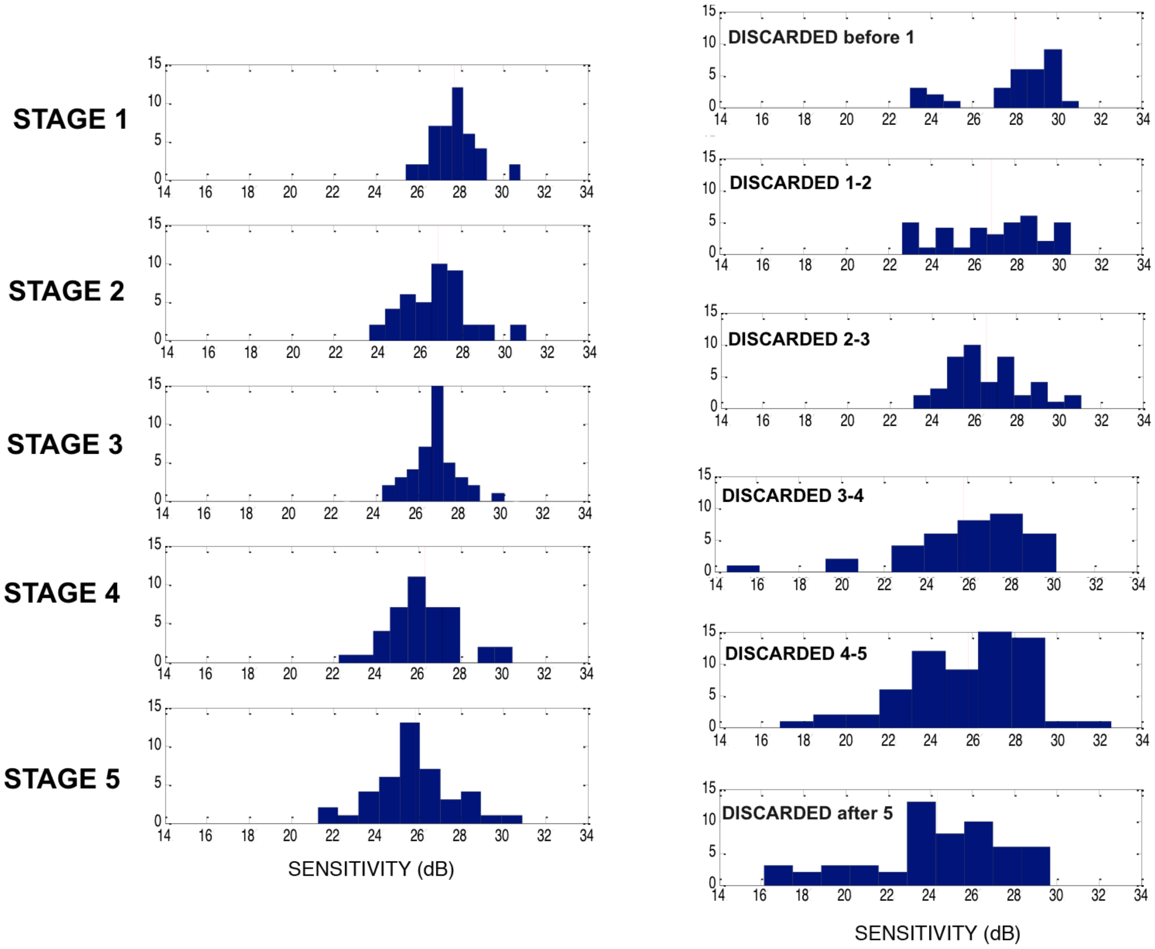


Figure 5.9: Histograms for informative and non-informative sectorial data after DSBN learning on one variable.

for relationships between variables in separate stages can be used to capture non-stationary relationships, although a structure-learning framework must be put in place in this case. This has only partially been investigated thus far, so represents its natural extension. Another natural extension of the model is to use a hidden class variable. In this case the model can be used to perform clustering of the time series. As with static BNs, the EM algorithm can be used to assign each time series to a group while maximizing the likelihood of the data in an iterative process. Such a configuration was tested using a simple off-the-shelf dataset, provided by Transport for London. The experiment is reported in Section A.2 of Appendix

A. The results show meaningful clusters and interesting insights into the data.

Another limitation with the DSBN that should be addressed in future work is the ability of time series to skip stages. In the DSBN, all time series must provide N data points to fill all stages, which implies that all time series progress through all the stages. However, in practice, it is very likely that patients present with the disease at a later stage or are not observed in middle stages. Another limitation lies in the fact that it uses only a small minority of the total data points, and it may be better to use all information available, even noise, in the decision process. The need to set beforehand the number of stages of the model may also represent a limitation of the DSBN.

5.3 Non-Stationary Clustering Bayesian Networks

This section presents a novel algorithm, called the Non-Stationary Clustering Bayesian Network (NSCBN). This is based on the Non-Stationary Dynamic Bayesian Network (NSDBN) framework, and resolves key limitations of the DSBN model. In particular, no data is omitted from the NSCBN model and stages can be skipped. Furthermore, the number of stages is learned from the data. After introducing the NSDBN, this section presents the NSCBN and discusses some experimental results.

5.3.1 Non-Stationary Dynamic Bayesian Networks

The Non-Stationary Dynamic Bayesian Network (NSDBN) is a class of models that shares several features with DSBN models, including the concept of *stages*. However, unlike DSBN models and many other temporal models, NSDBNs do not assume that the underlying process originating the data is a stationary process. Furthermore, in DSBNs, each stage is modelled using one data point from each time series, while each stage in the NSDBN consists of a subset of the data.

Intuitively, in the NSDBN learning process both model parameterisation and segmentation are performed. Figure 5.10 shows a NSDBN with 4 stages, $G1, .., G4$. In NSDBNs, a *stage* is represented as a unique BN with independent parameters and conditional dependen-

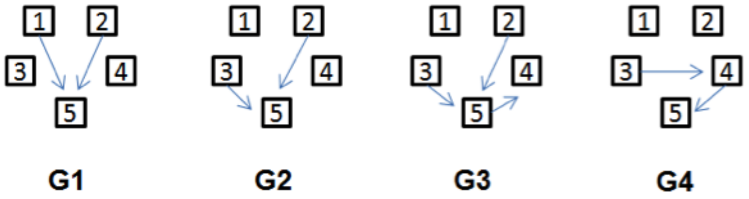


Figure 5.10: Non-Stationary Dynamic Bayesian Networks.

cies. Each stage models a subset of the data points available, i.e. a *segment* of data from each time series. This is different from DSBNs, where one data point is selected from each time series for each stage. In both models, each stage has the same number of variables. In the example shown in Figure 5.10, each stage has 5 variables. The learning process for NSDBN models typically corresponds to sampling over the posterior distribution of the model. However, the search space is usually limited by constraints on one or more degrees of freedom; in particular, the segmentation points of the time series, the parameters of the variables, the dependencies between the variables and the number of segments. To our knowledge, no studies have attempted to resolve all of these problems simultaneously, perhaps because of the size of the search space. Among the most recent and complete work, (Talih and Hengartner, 2005) used a Markov chain Monte Carlo approach to estimate the variance structure of the data, but the search space was limited to a fixed number of segments and undirected edges only. (Xuan and Murphy, 2007) proposed an approach to model changing dependency structures from multivariate time series, but the search was also limited to undirected edges. (Robinson and Hartemink, 2008) formalised the concept and proposed a solution for dealing with the aforementioned degrees of freedom bar the parameters. (Grzegorczyk and Husmeier, 2009) instead retained the stationarity of the structure in favour of parameter flexibility, arguing that structure changes lead almost certainly to over-flexibility of the model in short time series. While the first approach may only capture parameter changes that are strong enough to give rise to a structural change, the latter may not correctly model underlying conditional dependencies over the stages. The ability to both assess weak and strong changes in variable distributions and explicitly model the evolution of their relationships would be extremely informative, especially in unknown processes such as glaucoma.

5.3.2 Degrees of Freedom in Non-Stationary Clustering Bayesian Networks

The NSCBN is a novel algorithm that tackles all elements of the degree of freedoms simultaneously. In particular, the number of segments and the segmentation points of the time series are allowed to vary. Also, the parameters of the variables and the dependencies between them are not set beforehand. Moreover, the NSCBN adds a further degree of freedom for the clustering of patterns. Each cluster can be defined as a sequence of stages, i.e. a *temporal pattern* as defined in Section 5.2.1. As in DSBNs, temporal patterns do not reflect the time dimension of the data, but have only one direction of flow. The clustering of the temporal patterns is obtained by allowing an unconstrained segmentation of the data. In fact, when performing segmentation in the NSCBN, time series are allowed to skip stages, which translate into an automatic separation of the data into different temporal clusters. The ability to separate data into temporal clusters can be an extremely useful tool when exploring temporal processes and is a novelty of NSCBNs over NSDBNs.

5.3.3 Learning Non-Stationary Clustering Bayesian Networks

The algorithm used to obtain a NSCBN is a two-step SA technique that switches between *WARPING* operations on the data and *STRUCTURAL* operations on the model. This is analogous to the SA and the DSBN algorithm previously described (Algorithms 3.1 and 5.1). The pseudocode for the learning algorithm is shown in Algorithm 5.2. The input parameters *Temperature*, *Cooling_Factor*, *Model*, *Data* and *Max_Iterations* were described in Algorithm 3.1 and 5.1 and have the same functions in learning a NSCBN. The new inputs are the following: *Structural_Iterations*, *Warping_Iterations* and *Operation*. The first two regulate the number of *STRUCTURAL* and *WARPING* operations carried out, while the *Operation* is a binary variable that defines which of the two operations is being carried out. This is typically set to perform warping operations at the start of the algorithm, but switches to structural operations after a number of iterations. As with DSBNs, the *WARPING*

Algorithm 5.2 Learning NSCBN Algorithm.

```
Input: Temperature, Cooling_Factor, Model, Data, Max_Iterations,
          Structural_Iterations, Warping_Iterations, Operation
Score Model and store in Best_Score
Loop on iterations until Max_Iterations is reached
     if (Operation is Warping)
          if iterations < Warping_Iterations
               With Probability P
                    Select Random Time_Series
                    Apply WARPING to Time_Series
                    UPDATE Data
               With Probability 1-P
                    Apply STAGE to Model
          else
               Switch Operation from Warping to Structural
          end if
     end if
     if (Operation is Structural)
          if iterations < Structural_Iterations
               Apply STRUCTURAL to Model
          else
               Switch Operation from Structural to Warping
          end if
     end if
     Score Model and store in New_Score
     Diff_Score is New_Score - Score
     if (New_Score is higher than Score)
          Update Score, Data and Model
          if (New_Score is higher than Best_Score)
               Reset iterations
               Store Model and Data for Output
               Best Score is New Score
          end if
     elseif Random_Number is less than exp(Diff_Score / Temperature)
          Update Score, Data and Model
     end if
     Temperature is Temperature  ·  Cooling_Factor
end Loop
Output: Model, Data
```

operation performs the segmentation of the time series. This corresponds to (i) randomly selecting a time series and (ii) randomly picking a set of segmentation points. Unlike in the DSBN, the function may segment the selected time series over a subset of the total set of stages so that time series are allowed to skip stages. With a lower probability, in this configuration the $STAGE$ function is carried out. This corresponds to (i) randomly picking a stage and (ii) removing it from the model or (iii) adding a new stage after having selected one. Note that in the former case the data is rearranged in the remaining stages. As in the other SA approaches, after each iteration the model is scored and the solution is accepted according to the current $Temperature$ and the difference between the new and the old score. When a solution is accepted, the structure and the parameters of the model are updated and the old data segmentation is replaced. Once a better solution is not found for a certain number of *Warping_Iterations*, the algorithm switches to the structural operations by changing the *Operation* binary variable. The structural operation is performed by the $STRUCTURAL$ function. After randomly selecting a stage, this function carries out one of the following operations on the selected stage: (i) adding an arc between two randomly selected nodes, (ii) removing a randomly selected arc or (iii) rearranging the arcs by removing a randomly selected subset of arcs and adding them back at a randomly selected position. This is analogous to the structural operations described in the previous chapters. When no more improvements are obtained with structural operations for *Structural_Iterations*, the algorithm switches back to the warping operations and so forth until the best score does not improve for a given number *Max_Iterations*. Note that convergence of the algorithm was found typically after about 10000 total iterations, with an initial temperature of 10 and a cooling factor of 0.9. *kbic* was set to 0.9, P to 0.05. In all the experiments, the best scoring solution among 5 to 10 runs was selected.

Intuitively, since the parameters depend on the segmentation of the data and the score of a new segmentation depends on the current parameters and the structure of the model, the algorithm tends to converge by grouping together similar data. Given that data is not forced to pass through all the stages, if clusters of data are present they will tend to aggregate into separate stages, forming the temporal clusters described above. The process is illustrated in

Figure 5.11. In part A of the figure, a new time series D_k is to be allocated in the model.

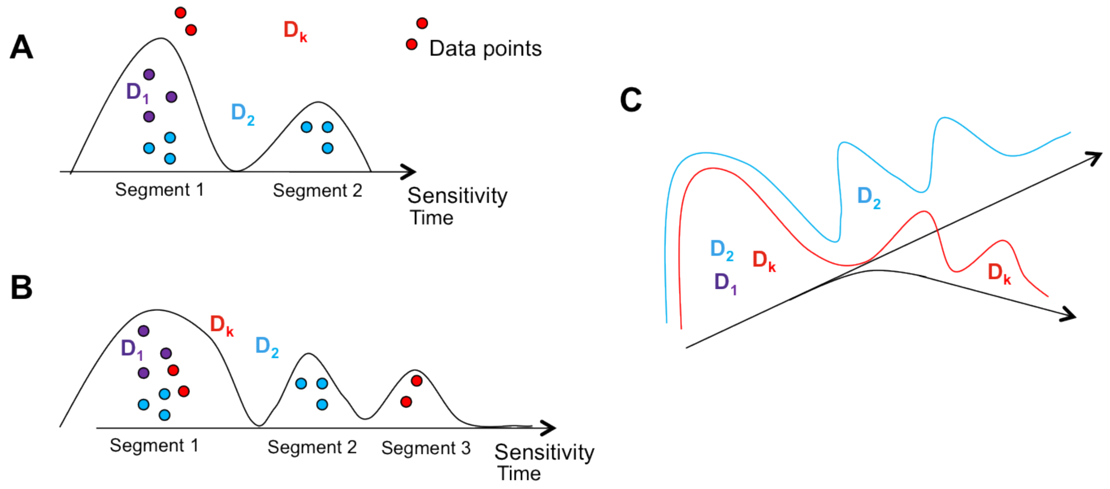


Figure 5.11: A. New time series D_k to be allocated in the model. B. The time series is allocated and a new stage is added to the model. C. The clusters are highlighted by observing the distribution of the model and the allocation of the data in each stage.

The exemplificative model shown in the figure has two stages, represented by a continuous line that shows two peaks in the CPD. Time series are represented by sets of dots of the same color, i.e. D_1 in purple and D_2 in cyan. Data is segmented in two segments to fit the two stages. When a third time series D_k (red dots) is considered, a third stage is added to the model to collocate the new points. The stage is added if the increase of the likelihood of the data $p(D \mid \theta)$ is higher than the increase of the complexity factor of the model, which is regulated by the complexity parameter in the BIC score. Part C of Figure 5.11 illustrates the clustering of the temporal patterns for the same example. Although all three time series share Stage 1, the D_2 progression oath spans Stage 2 while D_k skips this stage and progresses to stage 3. In the figure, the different characteristics of Stages 2 and 3 are highlighted by the direction of the arrows, so that it is clear that the selective allocation of the data points into certain stages divide the data into different paths of progression, performing in fact a clustering of the temporal patterns in the data.

5.3.4 Experimental Results

The NSCBN algorithm was tested using three simulated datasets and two real-world datasets (Datasets B and C, Table 3.1). The real data experiments elicit interesting insights from a clinical perspective, but a systematic testing of the algorithm using simulated datasets enables assessment of the performance and the robustness of the algorithm. This section presents and discusses the results obtained from these experiments.

5.3.4.1 Simulated Data

The simulated datasets were generated using four BNs with 5 variables, shown in Figure 5.10. After a random initialisation of the parameters, the four BNs (i.e G1, G2, G3 and G4) were repeatedly sampled and the obtained data was concatenated to build three different datasets to assess performance of the algorithm with different degrees of freedom. The first dataset consisted of 800 data points distributed in three time series with only one cluster, i.e. all time series were obtained using G1, G2 and G3. The second simulated set was made up from 650 data points in four time series. Of these, two were obtained by sampling from G1 and G2 and the other two by sampling from G3 and G4, i.e. two separate clusters were simulated. The third dataset was obtained by sampling 650 points distributed in 40 time series. Of these, 20 were obtained using only G1 and G4 and 20 using G2 and G3. For the real-world datasets analysed (Datasets B and C), the variables were grouped into six sectors as described in Section 3.1. The datasets were categorised into five values using a frequency based approach to ensure that an equal number of discrete states were observed.

The simulated datasets were used to analyse how the algorithm performed when increasing the degrees of freedom. Thus, four experiments were carried out. The first (see Figure 5.12, UTKNKC) was carried out using the first simulated dataset and imposing Unknown segmentation Timepoints, Known Number of segments (i.e. three segments) and Known Clusters (i.e. one cluster). The second (see Figure 5.12, UTUNKC) was obtained using the same dataset, but Unknown segmentation Timepoints, Unknown Number of segments and Known Clusters (i.e. one cluster) were imposed. The third experiment (see Figure 5.12, UTUNUC) was carried out using the second simulated dataset made of two clusters of two time series each. All degrees of freedom were allowed in the experiment, i.e. Unknown segmentation Timepoints, Unknown Number of segments and Unknown Clusters. The fourth experiment (see Figure 5.12, UTUSUK, bottom right) was carried out on the third simulated dataset, made of two clusters of 20 time series each. All degrees of freedom were allowed in this experiment. For all experiments, the best scoring solution over five runs was selected. The results for each class of experiments follow:

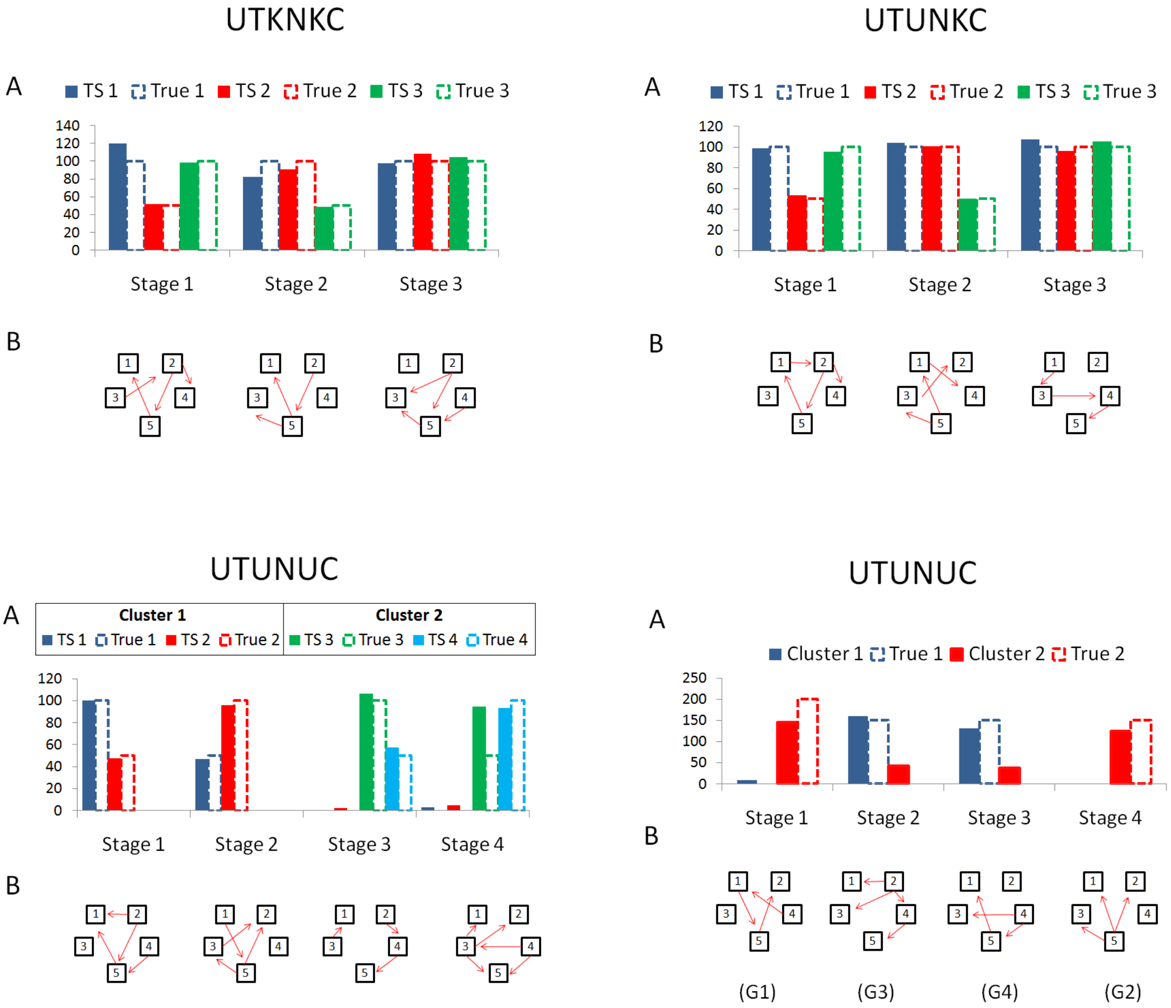


Figure 5.12: Results on the three simulated datasets. For each experiment, Section A shows the distribution of the data points across the stages (dashed: true distribution) and Section B presents the structure learned at each stage.

UTKNKC The results for this experiment are presented in Figure 5.12 (top left). In section A, the segmentation obtained by the model is compared with the true one for each time series. The difference between the two is minimal. The structural dependencies present in G1, G2 and G3 (Figure 5.10) were captured by the algorithm, although a further relationships was also obtained.

UTUNKC Results when the number of stages was allowed to vary during the learning are presented on the top right of Figure 5.12. The number of stages learned by the algorithm is correct. While the learned segmentation of the time series almost matches the true one, the structural relationships learned by the algorithm are less precise in comparison with the first experiment. Stage 1 captures all the relevant dependencies in G1 (and more), but Stage 2 and Stage 3 capture only 50% of the direct relationships in G2 and G3. However, considering the Markov Blankets at Stages 2 and 3 of the original model (Figure 5.10), conditional dependency should be considered between nodes 2 and 3 through node 5, while in Stage 1 the same applies for nodes 1 and 2 through node 5. Both of these relationships were captured by the algorithm, which increases the number of conditional dependencies captured by the algorithm.

UTUNUC Two experiments were carried out with full degrees of freedom. In the first such experiment four time series were correctly clustered into two groups, as shown on the bottom left section of Figure 5.12. The learned number of stages was correct. All the relationships were captured by the algorithm, except for Stage 3 (i.e. G3, the first stage of the second cluster). The segmentation obtained is very close to the original. The second experiment was carried out with a simulated set of 40 short time series distributed in two clusters. The results are shown on the bottom right section of Figure 5.12. The algorithm correctly identified four stages and found most of the conditional dependencies present in the original model when we consider the Markov Blankets described above. Notice that a different alignment of the stages into the clusters was found, i.e. Stages 2 and 3 are assigned to Cluster 1 and Stages 1 and 4 are assigned to Cluster 2. The stages were remapped by the algorithm because there is no constraint on which stages each cluster has to take up. This rearrangement does not affect the final result and can be seen as a proof of the flexibility of the model. In fact, as shown in Section A, data was aligned correctly to each stage in most of the cases. In Section B, the corresponding structures in relationships with Figure 5.10 are reported in brackets for each stage.

The results obtained in the simulated dataset are acceptable in terms of segmentation and modelling of the data. Even though minimal segmentation error was present in all experiments, the number of stages and clusters were correctly identified. The separation into clusters was obtained by considering the distribution of the patients and the data points, in particular in relationships to peculiar starting and ending stages. Conditional independence relationships in the original model were captured by the algorithm, although several spurious relationships were also found. However, data were generated after random parameterisation, so that parameter distributions may have acted on the conditional independence manifestation. In addition, structures were chosen to be partially shared by stages, as expected in real world problems. In particular, 50% of the conditional dependencies in G1 and G2 are the same (Figure 5.10). G3 also shares 66% of its arcs with G2, and 33% with G1. In fact, in the clustering experiment shown in the bottom right of Figure 5.12, a small amount of data belonging to G1 was placed also in Stages 2 and 3. However, the total number of misplaced time points between Stages 1, 2 and 3 in Cluster 2 comprised 13% of the data. This has a relatively low impact on trends, clusters extraction and analysis, but it shows that to obtain an exact model in these conditions is difficult. As for the practicality of these models from a clinical standpoint, our view is pragmatic and aims to find a trade-off between informative power (e.g. variables trends, dependencies between variables and clusters of time series), acceptable precision and processing time. The algorithm has been shown to be quite robust and fast on both simulated and real-world datasets. Among different runs, in fact, several models were found to be similar and the highest score was a fair indicator for the most similar model with respect to the original one. The running time of the algorithm was less than two hours on a modern desktop machine (dual-core, 2.8 GHz, 4GB RAM) for all the datasets. The impact of the dataset size and type was only partially investigated, although the algorithm obtained acceptable results for both using a few long time series and many short ones.

5.3.4.2 Glaucoma Progression Data

Results of using the NSCBN algorithm on Dataset B are presented in Figure 5.13. The distribution of the time points over the stages is shown, grouped into clusters. Each cluster presents different distribution over the stages, in particular regarding starting and ending stages. At each stage, a VF sensitivity map is presented. The map was obtained through averaging the original sensitivity values for each stage and variable and mapping them in greyscale onto their corresponding spatial locations. This view conforms with traditional VF visualisation, where darker values represent lower sensitivities. The conditional relationships found at each stage (red arrows) are superimposed on the sectorial VF map. Considering all clusters, the first four stages and Stages 8 and 9 hold the higher number of visits, while the middle stages hold less data. In the first analysis there seemed to be one main group progressing from Stages 1, 2, 3 and 4 and ending in Stages 8, 9 and 10, and another relatively ”stable” group starting in Stage 5 and ending in Stage 6. However, extra care should be taken when drawing conclusions from this visualisation, since it is dependent on the sampling of the

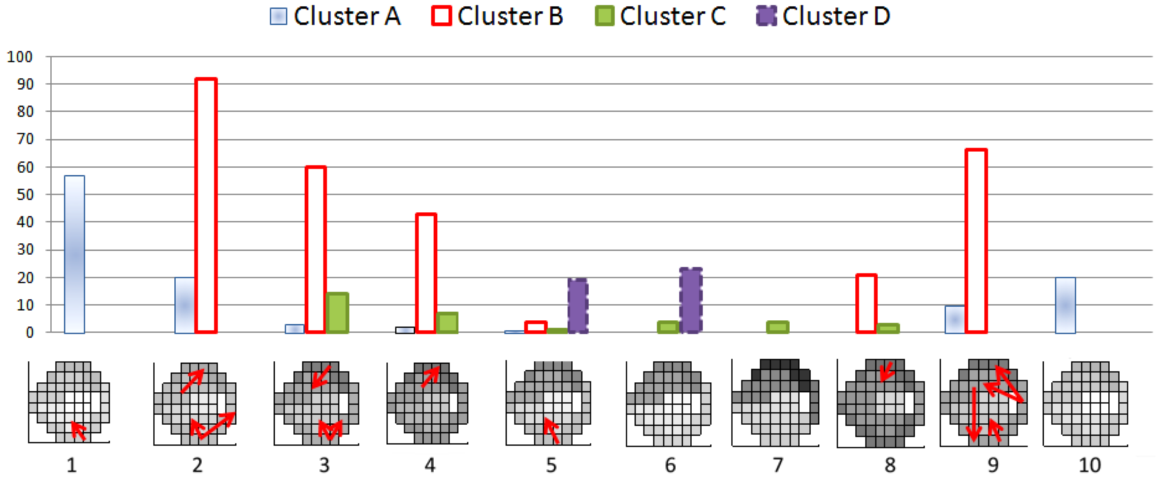


Figure 5.13: Results on the real Dataset B: distribution of data across the stages for the four clusters identified. The VF map associated with each stage is also shown.

patients. For example, a patient who is in a stable condition for a long time or who is tested more frequently will contribute many points to a number of stages and many fewer points to other stages. Still, the histogram can be used in conjunction with the distinct patients graph (see Figure 5.14) to explore the temporal clusters obtained. Patient transitions for each cluster is shown in this graph. The groups can be further characterised through inspection

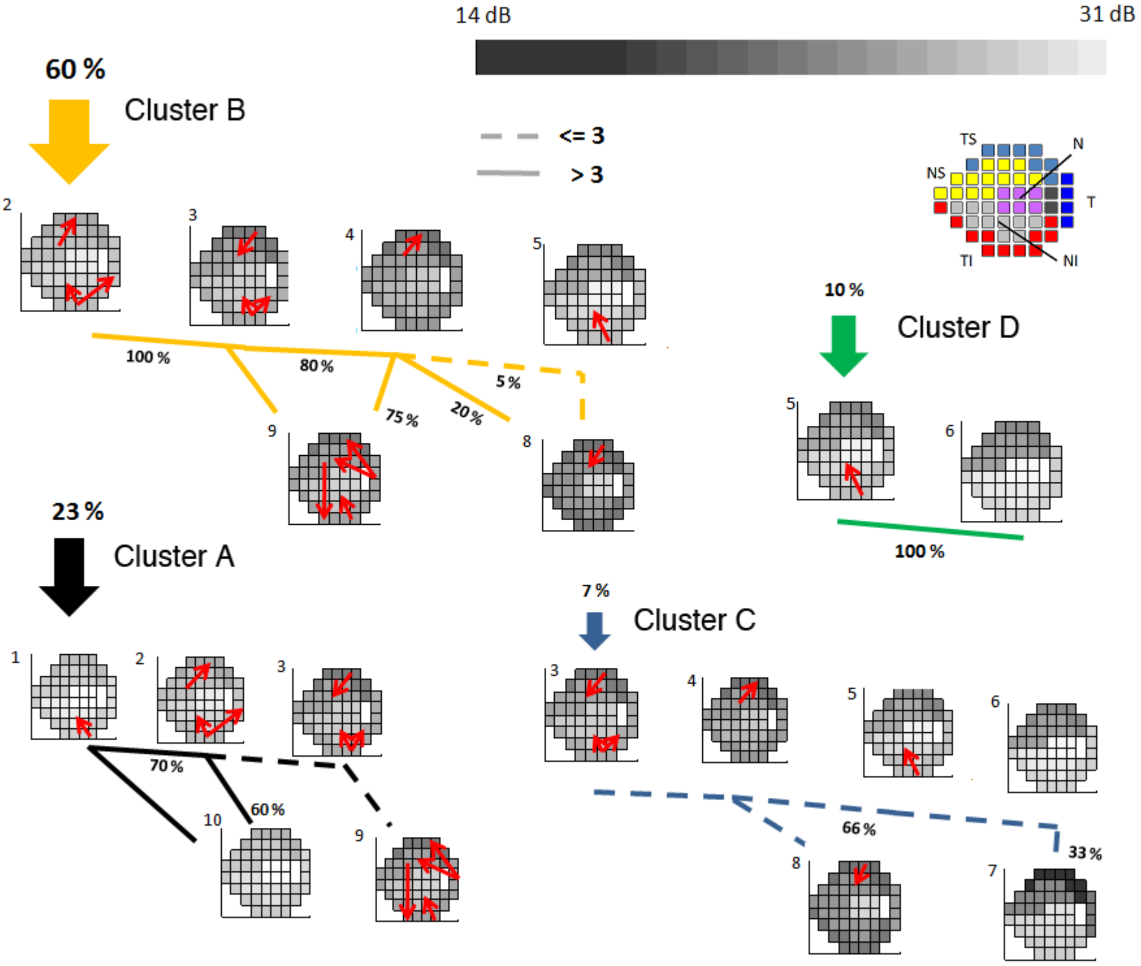


Figure 5.14: Results on Dataset B: patients transition across stages for the four clusters identified. Stages are again represented as VF maps. Vertical arrows represent the number of distinct patients ”entering” each stage, as well as describing the distinct transitions across stages. Solid lines represent strong transitions between stages in terms of the number of patients, while dashed lines represent lines which carry three or fewer patients. This threshold was chosen arbitrarily for clear visualisation purposes. The percentages of patients at each transition are expressed in percentages relative to the data present at the previous level.

of this figure. Cluster D, starting at Stage 5 and ending in Stage 6, seems to model a rather stable condition, which can be described as an asymmetric upper hemifield defect. In the other clusters, instead, glaucomatous progression occurs. Stage 1 seems to be a key stage for patients ending at Stage 10. Both these stages show early defects and are characteristic of Cluster A. Therefore we could associate this group (Cluster A) of patients with stable mild glaucoma. Note that data in this cluster may have been misplaced in Stage 2 because of its similarity with Stage 1, as was seen in the clustering simulated experiment described above. Stage 4 is a key stage for patients ending at Stage 8, which shows general depression with a severe peripheral VF endpoint. Stage 9 is the end stage for most of the patients (55% of the whole dataset) and can be reached from Stages 3 and 4. Its characteristics are also severe, but in comparison with Stage 8 it shows a depression in the upper hemifield, i.e. an asymmetric pattern. Cluster C is a small subset of three patients who were found to cover all stages from 3 to 7, progressing from mild diffuse to strong upper hemifield defects. Note that Stage 2 is strongly represented in another cluster (Cluster B), being the starting stage for 60% of the patients. This stage represents mild glaucoma worsening to Stage 3. Here the upper hemifield shows impairment and loses its dependency with adjacent sectors. This is also a key stage for separation, as progression splits into two paths: most of the patients continue to develop asymmetrical defects ending in Stage 9, while others also develop peripheral defects in the lower hemifield, leading to the most severe state described by Stage 8. Notice that Stage 8 is reached only through Stages 2, 3 and 4, which are thus key stages for understanding progression. In particular, this shows that according to this dataset glaucoma typically acts initially on the superior hemifield. Afterwards, while asymmetrical VF progression is likely, Stage 9 and also Stages 5 and 6 show that as long as only upper hemifield defects are present, they tend generally to be stable, especially in terms of spatial localisation. Strikingly, if inferior hemifield VF defects occur, even in the presence of diffuse loss, the progression to peripheral severe defects is reached in 20% of cases. Therefore, this might be seen as an important indicator for clinicians that the patient should be carefully monitored. As discussed, the appearance of functional defects in the peripheral sectors of VF in early glaucomatous patients is strongly supported in literature, although it is not fully accepted. Typical other changes during progression are change in sensitivity across the horizontal meridian and generalised loss of retinal sensitivity.

The structures learned at different stages are consistent in each cluster and often cover adjacent sectors. While Stages 1 and 2 share same conditional independences between sectors (i.e. TS with NS, TI with NI and TI with T), Stage 3 lacks relationships in the upper hemifield. Its main successors, Stages 9 and 4, lose also the other arcs. Stages 8 and 5 present relationships between sectors reflecting their absolute values: lower to upper hemifields for Stage 5 and TS to NS for Stage 8. In general, arcs are most commonly found on arcuate sectors and less arcs are found towards the end stages. Thus, structural relationships learned by the model interestingly seem to relate with the cluster and severity of progression. This finding is supported by the real underlying structural impairment that occurs in the retina. In fact, while VF progression takes place, relationships between different sectors of the RNFL decay, and this may lead to a ”flatness” in the relationships between the variables that is well captured by the model. Late stages sensibly have fewer dependencies. This observation can be useful for characterizing patients and could be used as an indicator for the level of health of patients' retinas. Also, it confirms the utility of VF metrics, such as the Glaucoma Hemifield Test, which evaluates the relative sensitivity between hemifields. Clustering in this sense acts as a filter for variance, because it enables investigation of separate ”paths” of progression. A small group of three patients was found to cover Stages 3, 4, 5, 6, 7 and 8, confirming that Stage 3 is pluripotent as it can lead to both severe peripheral or asymmetrical defects. Out of three patients starting at Stage 3, one progressed to Stage 8, while two developed strong asymmetrical defects (Stage 7). Given the low proportion of patients belonging to this cluster, the time series were visually inspected. As shown in Figure 5.15, both cases present strong sensitivity fluctuations. In the case presented in section A, a progression on the temporal (T) and superior sectors is interrupted by a deep negative peak in several sectors, which acts as a trigger for a switch from Stage 3 to Stage 4. Once in Stage 4, however, the patient's condition unexpectedly improves for all sectors, including T, which triggers, in turn, a switch to Stage 5. The rest of the seris follow a more common pattern between Stages 5-6-7, so in this case the series seem to be affected by large fluctuations early

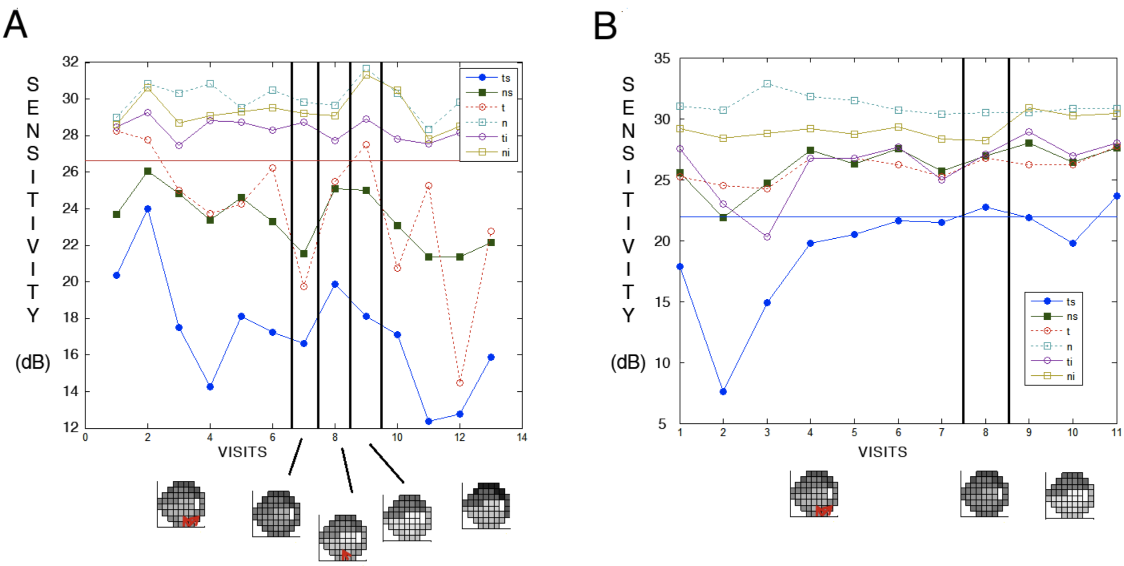


Figure 5.15: Case studies from Dataset B. Sensitivity sectorial values are shown for each visit. Vertical black lines represent changes in stages, represented by the corresponding VF maps in the lower section.

on, which may be due to other conditions of the patient or another type of test-dependent error. Section B shows a patient who presented with great depression in the TS sector at the beginning of the series and subsequently stabilised with high sensitivity levels. This appears to have affected their starting positioning, which was Stage 3, instead of Stage 1. This may be due to the introduction of medication after Stage 4, although this data was not available to us and so cannot be assimilated by the model. As it is impossible to know the underlying truth for this dataset, extra care should be taken when interpreting information from a small groups of patients as they may belong to other groups or represent large measurements error.

5.3.4.3 Post-Trabeculectomy Data

The algorithm was also tested on Dataset C, which consists of a series of patients who underwent glaucoma surgery (trabeculectomy). The clusters found are presented in Figure 5.16. Four temporal clusters were found in the analysis (A, B C and D). Clusters B and C share two stages, highlighted by arrows crossing the boundaries of the two clusters, but have different start and end stage characteristics; the other clusters show weak overlap.

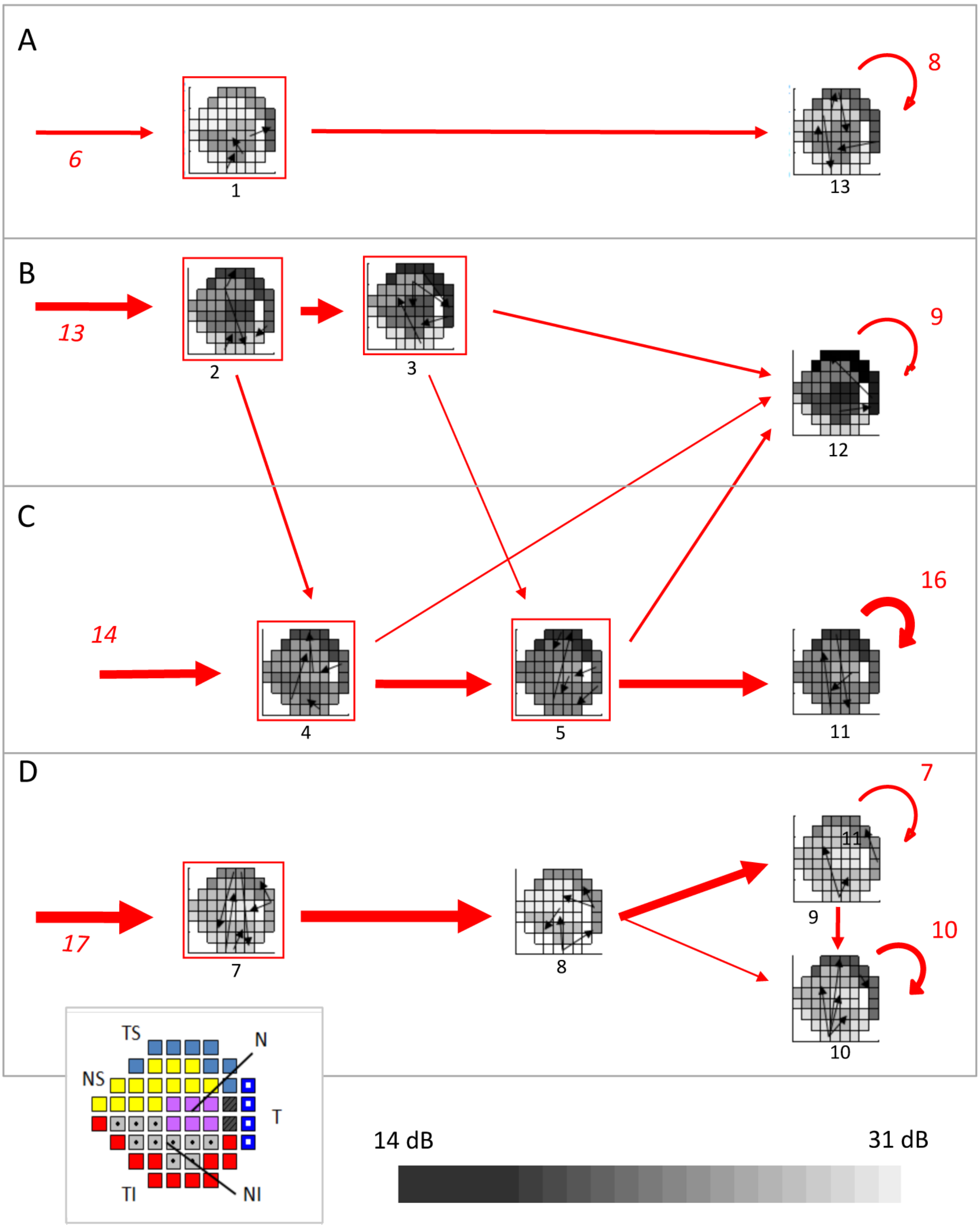


Figure 5.16: Results on the Dataset C: patients' transition across stages for the four clusters. The thickness of the arrows is proportional to the number of patients transitioning. For the start and end stages, the number of patients is reported. Borders are added to stages to highlight when surgery was carried out.

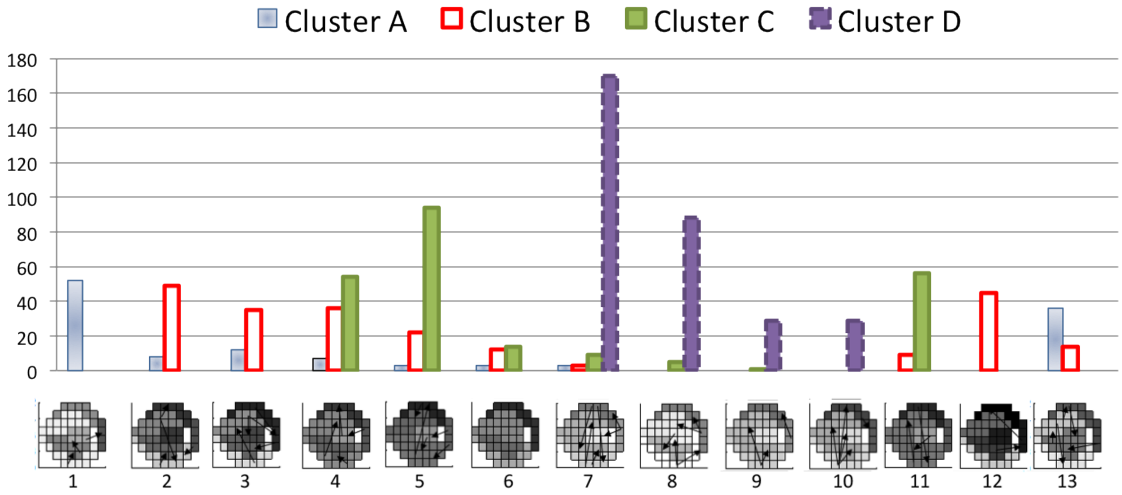


Figure 5.17: Results on Dataset C: distribution of data across the stages for the four clusters identified. The VF map associated with each stage is also shown.

This is confirmed in Figure 5.17, which shows the time point distribution over all stages for each cluster. While Clusters A and D do not present time points in the same stages, there is overlap between data from Clusters B and C on Stages 4 and 5. This confirms the higher diversity of Clusters A and D with respect to the others. Notice that Stage 6 is not characteristic of any cluster and therefore it is not included in Figure 5.16. One patient was also excluded from the analysis, having uncommon start and end stages.

The structures learned using the algorithm are consistent within the clusters and do not differ strongly across clusters. Arcs are concentrated in the nasal sector (N) for Cluster A and in the peripheral sectors (TS, TI, T) in the other clusters. VF sensitivity values seem characteristic of each cluster, showing consistent and different progression paths over clusters. Cluster A shows early damage and progression only in N, T, TS and NI sectors. Cluster B presents severe damage in the superior hemifield in all stages, progressing mostly in NI, T and TS sectors. Cluster C may be considered the only stable cluster, showing diffuse damage across all sectors and strong TS damage. The patients in this cluster constitute roughly a third of all the patients analysed. Cluster D shows mild diffuse damage in the first stages, but progression in the T and TS sectors. Looking at the histograms in Figure 5.17, Cluster D presents most of the visits in its first stage, i.e. Stage 7, indicating high stability. Therefore, observing a diffuse loss similar to Stage 7 may indicate a low rate of progression. Again, it must be noticed that the histograms do not take into account those patients who are checked less frequently after one year from the date of surgery, resulting in less time points towards final stages. Clusters B and C present different characteristics. The former starts with strong N, NI and TS damage and seems not to progress in the lower hemifield, while the latter has diffuse loss but weak N and TI damage. Despite their differences, several patients are found to pass from Cluster B to Cluster C, typically ending in Cluster C again. Even if several patients follow this pattern, the amount of time points for Cluster B in Stages 4 and 5 is low, as shown in Figure 5.17. Therefore, these VF tests are likely to represent errors, fluctuations or temporal abrupt conditions due to surgery. One example of this kind is presented in Figure 5.18, in which the sensitivity values of the six sectors of the VF of a particular patient are presented. The Figure shows five tests before surgery and ten tests after. Vertical lines on

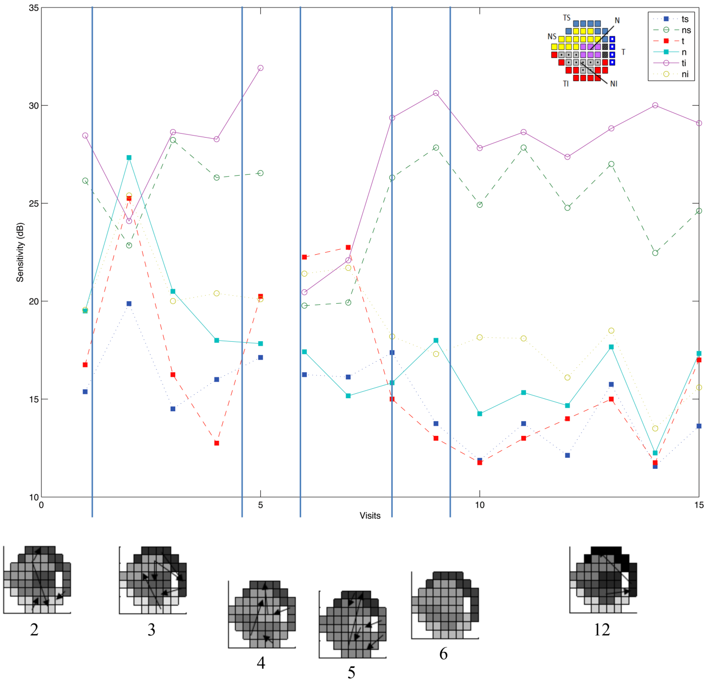


Figure 5.18: Results on Dataset C: case study. Sectorial sensitivity values are shown for each visit. Vertical blue lines represent changes in stages, which are represented by the corresponding VF maps in the lower section.

the graph separate different stages, which are represented as VF maps. Data before surgery is consistent with the first two stages of Cluster B, i.e. weak or no damage in TI and NS sectors, and progression in the other sectors. Following surgery, the patient performed two VF tests showing general diffuse loss, typical of Cluster C. However, progression resumed and the patient was placed again in Cluster B. In our analysis we found that many patients followed this pattern, so this may be considered by clinicians when making decisions. Other indications can be extracted from the model. For example, surgery on patients in Cluster C was more effective because patients did not progress after surgery in the cluster. Therefore, these patients may need less monitoring than patients in Cluster A or D, for which surgery is still effective (considering the large amount of time points in the starting stage) but for which a sudden onset of further VF progression is possible at anytime. In this case, key sectors for that particular cluster should be monitored more carefully.

5.4 Summary

In this chapter longitudinal glaucoma data was explored and exploited to perform classification of glaucoma. DBNs were introduced and their limitations for modelling glaucoma context were discussed. In particular, a lack of support for non-stationarities in the data was tackled by introducing DSBNs, which are BN models based on *stages of disease*. The new DSBN model outperformed DBN for classification and provided a framework capable of identifying key stages of progression. Only averaged data was tested, obtaining a descending pattern with increasing variability towards later stages. Although promising, DSBNs also have important limitations, such as the need to use data covering all stages of disease. Another novel technique was proposed to address this and other issues: the NSCBN is a non-stationary BN that models longitudinal data using all measurements available. The algorithm automatically learns the number of stages, the parameters and the relationships between variables. Moreover, it is able to extract and model temporal patterns in the data, i.e. sets of stages in groups of patients. The NSCBN was first tested on simulated data using incremental levels of complexity, which confirmed that the model provides robust and sufficiently reliable results with a number of variables and data points similar to those of the real data available. The use of the NSCBN algorithm on Datasets B and C led to interesting insights on the progression and treatment of glaucoma. In particular, early glaucoma defects were found to most often occur in arcuate sectors of the VF, while the inferior hemifield was generally found to be affected at later stages. Damage in the inferior hemifield was also associated with a key stage for future progression, suggesting close monitoring of the patient is necessary when this occurs. When the NSCBN was employed on the trabeculectomy dataset, the patterns obtained showed little new damage following surgery. Furthermore, for patients showing superior VF defects at the time of intervention, trabeculectomy was found to be more effective than for other groups. This finding has important implications on the actual utility and effectiveness of this operation with regards to whether patients have damage in their superior or inferior hemifield, as well as provide useful information for clinicians on progression of glaucoma.

Chapter 6

Conclusions

6.1 Contributions

Glaucoma is a major disease and its mechanisms are not fully understood. Early diagnosis is key to preventing blindness but this is not fulfilled by current diagnosis techniques. In this context, the thesis' contributions can be grouped as follows.

6.1.1 Diagnosis Performance

In this thesis a set of BN classifiers were tested against current techniques (e.g. AGIS) and other state-of-the-art ML techniques for glaucoma diagnosis. BNs were shown to perform well on all datasets investigated. In the cross sectional data, performance was highest at higher specificities using a semi-supervised BN (71% sensitivity at 90% specificity). The application of semi-supervised BNs allowed us to avoid the biases associated with an imperfect diagnosis metric. In longitudinal data, although tested on one average variable, DSBNs obtained better performances than BNs, indicating the potential of considering non-stationarity for classification of glaucoma progression. Across all the experiments, the use of Simulated Annealing allowed us to learn meaningful relationships between the variables and to further increase the classification performances of BN classifiers.

6.1.2 Data Integration

Data integration was carried out by exploiting BNs' flexibility and intuitive representation. This allowed easy integration of anatomical and functional data into a single network, and exploration of the relationships between the variables using different metrics. While the combination of both AGIS and MRA metrics did not improve the overall results, the use of ensemble of classifiers to combine anatomy-based and functional data-driven BNs led to fewer errors than considering the base classifiers alone on both cross-sectional and longitudinal datasets. By investigating the ensemble model, it was also possible to understand how the base classifiers combine to reach better results, indicating that the impact of the AGIS-based classifier is higher than the anatomy-based one on the final output.

6.1.3 Disease Exploration

BNs represent relationships between variables graphically, which allows clinicians to assess the results and contribute to the modelling process. In this study, Simulated Annealing learning algorithms obtained several interesting relationships between structural and functional measurements, and these were confirmed in clinical literature, particularly regarding VF defects occurring as arcuate patterns.

Pointwise clustering using BN models allowed us to confirm, and find new relationships, in the glaucomatous VF, as well as identify and group patients based on their condition. For example, subjects showing superior hemifield defects are more easily identified as glaucomatous with respect to patients showing other patterns, especially when patients are younger than average. Therefore, such patients should be followed more carefully and perhaps treated more intensively than others. The idea of characterising groups of patients is an interesting aspect of glaucoma treatment and is supported by results presented in this thesis.

Regarding longitudinal data, the exploration of patterns of disease by NSCBN models also led to interesting insights being gained from the data. These models are able to identify a number of key stages in glaucoma progression and characterise them, as well as extract relationships between variables and expose paths of glaucoma progression across patients. Although no post-processing implementation is presented in this work, NSCBNs proved useful for visually predicting glaucoma progression and identifying key stages. In particular, lower hemifield VF defects were found to be a key risk factor for progression. Applying NSCBNs to the trabeculectomy dataset suggests that surgery is effective in many cases. Moreover, the model was able to suggest how to maximise the effect of trabeculectomy by exposing different paths of disease.

6.2 Limitations

Structural data and functional data are hugely important for monitoring glaucoma; however, literature shows that interpreting these data is difficult. For instance, the inclusion of IOP and RNFL thickness measurements in conjunction with VF data could improve the performance of classifiers. In this context, BNs represent a good modelling choice given their flexibility in combining data; however, the number of variables must be taken into consideration. Including a relatively high number of variables may make inference and learning computationally demanding. This issue is common to several ML techniques, and in many cases a solution can be obtained by aggregating data or performing feature selection.

The use of Simulated Annealing in this thesis follows previous work on the field of glaucoma research and BNs. Although other structure-learning algorithms could be used, literature on BNs is focused on gradient search algorithms and investigating the impact of other searching techniques was not the aim of this work. In general, other techniques such as evolutionary algorithms may in fact be efficient and provide good results.

Several models described in Chapter 4 use different criteria for parameters and structure learning. However, the relationships learned with a certain dataset may not be well supported in the training phase, or in the inference, using another dataset. This issue affects the performance of any model, but also provides us with interesting insights on the nature of the metrics. Other techniques may be used to better exploit different metrics for classification.

NSCBNs proved very useful to better understand glaucoma and the effects of treatment on the progression of the disease. Although their intuitive representation allows one to qualitatively assess progression and the efficacy of treatment, no quantitative robust method is proposed. This may be represented by an extension to classification in the form of DSBNs; however, several differences make ad-hoc learning algorithms necessary. A proposed solution is shown in the next section. However, full posterior sampling techniques to find the best configuration of the model were not explored.

6.3 Further Work

Further work could extend the proposed modelling techniques in a number of ways. Referring to the limitations of the datasets presented, a breakthrough would be to include all available data types in one model. The ability to relate results from different sources is a very important aspect for clinicians, and the framework provided in this thesis would perfectly suit such an extension. In this sense, the main issue is data availability; however, further data collection is being carried out, in particular with regards to new imaging techniques to measure the RNFL. Preliminary results of a DSBN with VF sensitivity, OD MRA-combined parameters and RNFL thickness data are shown in Figure 6.1.

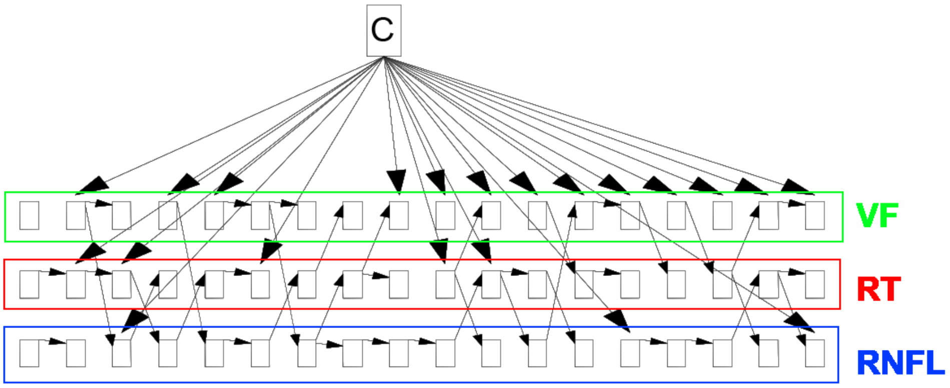


Figure 6.1: DSBN with 18 stages and 3 data types: VF sensitivity (VF), OD MRA-combined parameters (RT) and RNFL thickness (RNFL). This model is a preliminary experiment of DSBN with no discarded data and relationships allowed across data types and stages.

These results were obtained using severe discretisations, yet they represent potentially very useful extensions of BN-like models for data combination and investigation of relationships between data types and over different stages of disease. The model uses all data with one visit per node, so the number of nodes $N = 1 + MK$, where M is the length of the longest time series in the dataset and K is the number of variables per slice. An example of the insights that can be extracted by such multivariate models is shown in Figure 6.2, which was obtained by looking at the transitions of the aligned time series in the model for each of the three data types.

For VF data, controls seem less stable than converters, which may reflect the high variability in perimetry results caused by learning effects. More robust results are obtained for OD data, although arguably this also means less flexibility. For RNFL measurements, controls also seem more variable but the transitions are generally less interpretable. Perhaps this is due to the data being measured from HRT and not OCT, i.e. with a low resolution for such kind of analysis.

With regards to metrics and data combination, additional work on ensembles of classifiers could focus on the joined CPD and apply different prior knowledge to smooth the parameters, as well as using different discretisation and base classifiers. This would allow further understanding of the nature of the combination between networks and may improve classification performances. Also, the use of ensembles of classifiers to combine the base classifiers trained on different metrics presented in Section 4.1 was not assessed, although it must be noted that the different results obtained at high specificity for the MRA and AGIS based classifiers, showed in Section 4.1.1, may fit well the ensembles framework and may be object of future analysis.

The identification of reliable data-driven standards to define conversion or diagnosis is also a key aspect in the management of glaucoma, which in turn is the driving factor of research into the disease. Future work could look at generalising and tuning the results presented here, so that objective and comparative definitions of glaucoma may be extracted from data.

With regards to the clustering of the patients, a refinement of the technique involving both clinician expertise and other datasets with different number of clusters would allow one to define and explore more groups and conditions, opening and expanding the paths indicated by the present work. While some results will likely lead to well known patterns, new insights

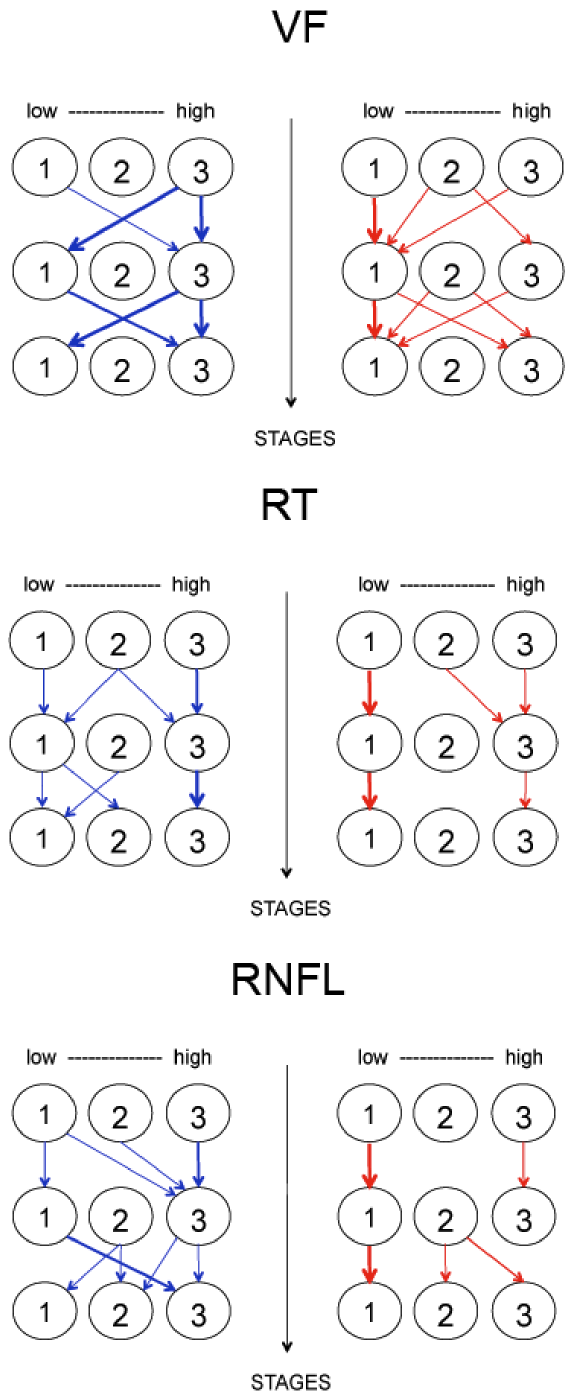


Figure 6.2: Main transitions in the data obtained from model in Figure 6.1.

could also be gained that could finally shed some light on inter-patient variability, and ultimately glaucoma management.

The use of non-stationary models for longitudinal data analysis is a promising approach that should be further investigated. In particular, testing DSBNs using multiple variables for clustering and classification seems to hold good potential. NSCBNs represent a broad platform for investigating glaucoma progression patterns. Their ability to automatically identify the number of key stages and characterise these stages with regards to progression is an important advantage over DSBNs. Limitations that should be addressed in future include structural learning across stages and classification ability. Other algorithms to search the best model configuration could also be implemented to allow faster and more robust clustering of the patients using different data types. An example of a NSCBN for classification is shown in Figure 6.3.

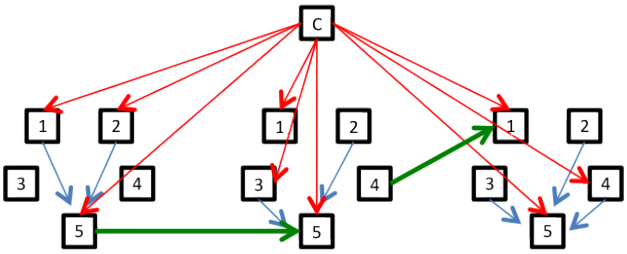


Figure 6.3: NSCBN with structural learning across stages and classification node.

The problem of calculating the score in such models, however, needs further investigation. In the DSBN each time series has one point per node; however, in NSCBN each node may have a different number of data points. This has to be taken into account when calculating the likelihood of nodes with parents in other stages. A solution may be to calculate the score by averaging the likelihood over all possible parent-child combinations.

It should be noted that all the models presented are easily generalisable to other fields and data types. The methods provided in this thesis may be used to explore other datasets without the need for many adjustments, apart from setting up the parameters in the learning algorithms. Regarding glaucoma, which is the focus of this thesis, future work could look at the integration of different data types and characterisation of the clusters of progression observed in the disease. Also, further datasets should be explored and compared to the results presented in this work. NSDBNs could also be extended for prediction, a key issue that relates with glaucoma management and can therefore be of great value. The ability to identify glaucoma patients and give them the right treatment at the right moment to minimise future progression is a target that can be pursued using the models presented in this thesis.

Appendix A

Additional Tables and Results

A.1 Additional Details on the Datasets

The listing of the variables and their basic distributional features is reported in this section.

Table A.1: Dataset A.

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Patient ID	Nominal	180	-	-
Condition	Nominal (Glaucoma, Healthy)	180	-	-
Gender	Nominal (M, F)	180	-	-
Eye	Nominal (Left, Right)	180	-	-
Age	Continuous (Months)	180	248, 991	774, 743, 147
Disc Area (Total)	Continuous	180	1.15, 3.01	1.95, 1.96, 0.34
Disc Area (Temporal)	Continuous	180	0.25, 0.75	0.44, 0.46, 0.09
Disc Area (Temporal-Superior)	Continuous	180	0.15, 0.38	0.26, 0.25, 0.04

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Disc Area (Temporal-Inferior)	Continuous	180	0.16, 0.42	0.27, 0.27, 0.05
Disc Area (Nasal)	Continuous	180	0.25, 0.76	0.45, 0.46, 0.09
Disc Area (Nasal-Superior)	Continuous	180	0.15, 0.38	0.25, 0.25, 0.04
Disc Area (Nasal-Inferior)	Continuous	180	0.14, 0.37	0.25, 0.24, 0.04
Rim Area (Total)	Continuous	180	0.47, 2.53	1.29, 1.30, 0.39
Rim Area (Temporal)	Continuous	180	0.02, 0.64	0.20, 0.21, 0.10
Rim Area (Temporal-Superior)	Continuous	180	0.02, 0.33	0.15, 0.15, 0.06
Rim Area (Temporal-Inferior)	Continuous	180	0.01, 0.36	0.17, 0.16, 0.07
Rim Area (Nasal)	Continuous	180	0.06, 0.70	0.38, 0.37, 0.10
Rim Area (Nasal-Superior)	Continuous	180	0.04, 0.33	0.20, 0.19, 0.05
Rim Area (Nasal-Inferior)	Continuous	180	0.06, 0.33	0.20, 0.19, 0.05
MRA score (Total)	Continuous	180	-0.38, 0.27	0.08, 0.06, 0.13
MRA score (Temporal)	Continuous	180	-0.68, 0.64	0.26, 0.22, 0.22
MRA score (Temporal-Superior)	Continuous	180	-0.75, 0.38	0.14, 0.11, 0.17
MRA score (Temporal-Inferior)	Continuous	180	-0.85, 0.35	0.12, 0.05, 0.22
MRA score (Nasal)	Continuous	180	-0.72, 0.20	0.11, 0.07, 0.13
MRA score (Nasal-Superior)	Continuous	180	-0.51, 0.23	0.09, 0.07, 0.13
MRA score (Nasal-Inferior)	Continuous	180	-0.44, 0.18	0.07, 0.03, 0.12
RNFL thickness (Total)	Continuous	180	0.02, 0.40	0.20, 0.20, 0.06
RNFL thickness (Temporal)	Continuous	180	-0.03, 0.13	0.07, 0.07, 0.02
RNFL thickness (Temporal-Superior)	Continuous	180	0.03, 0.52	0.24, 0.23, 0.08
RNFL thickness (Temporal-Inferior)	Continuous	180	-0.11, 0.45	0.20, 0.19, 0.09
RNFL thickness (Nasal)	Continuous	180	0.04, 0.52	0.23, 0.24, 0.10
RNFL thickness (Nasal-Superior)	Continuous	180	0.02, 0.60	0.27, 0.28, 0.09
RNFL thickness (Nasal-Inferior)	Continuous	180	-0.04, 0.57	0.28, 0.28, 0.11
VF sensitivity (Temporal)	Continuous	180	7.6, 30.1	25.1, 24.0, 4.5

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
VF sensitivity (Temporal-Superior)	Continuous	180	12.2, 33.1	27.3, 26.2, 3.9
VF sensitivity (Temporal-Inferior)	Continuous	180	7.0, 33.0	27.7, 27.1, 3.3
VF sensitivity (Nasal)	Continuous	180	7.6, 35.1	30.6, 29.9, 3.3
VF sensitivity (Nasal-Superior)	Continuous	180	12.9, 32.4	27.5, 26.8, 3.1
VF sensitivity (Nasal-Inferior)	Continuous	180	13.4, 33.0	29.3, 28.6, 3.0
Pointwise VF Sensitivity 1	Continuous	180	0, 33	23, 20.7, 8.91
Pointwise VF Sensitivity 2	Continuous	180	0, 32	23, 20.7, 8.72
Pointwise VF Sensitivity 3	Continuous	180	0, 32	22, 19.9, 8.82
Pointwise VF Sensitivity 4	Continuous	180	0, 32	23.5, 19.9, 9.06
Pointwise VF Sensitivity 5	Continuous	180	0, 32	24, 22, 8.81
Pointwise VF Sensitivity 6	Continuous	180	0, 33	25, 22.3, 8.76
Pointwise VF Sensitivity 7	Continuous	180	0, 34	27, 23.4, 8.87
Pointwise VF Sensitivity 8	Continuous	180	0, 34	26, 22.8, 8.84
Pointwise VF Sensitivity 9	Continuous	180	0, 32	25, 22.3, 8.84
Pointwise VF Sensitivity 10	Continuous	180	0, 31	25, 22.3, 8.95
Pointwise VF Sensitivity 11	Continuous	180	0, 34	24, 21, 9.16
Pointwise VF Sensitivity 12	Continuous	180	0, 34	26, 23.7, 8.91
Pointwise VF Sensitivity 13	Continuous	180	0, 35	29, 25.6, 9.11
Pointwise VF Sensitivity 14	Continuous	180	0, 35	29, 25.1, 9.49
Pointwise VF Sensitivity 15	Continuous	180	0, 34	28, 24.3, 9.51
Pointwise VF Sensitivity 16	Continuous	180	0, 34	27.5, 24.6, 9.01
Pointwise VF Sensitivity 17	Continuous	180	0, 36	26, 23.7, 8.94
Pointwise VF Sensitivity 18	Continuous	180	0, 32	26, 23.2, 8.8
Pointwise VF Sensitivity 19	Continuous	180	0, 31	23, 18.3, 10.3
Pointwise VF Sensitivity 20	Continuous	180	0, 35	25, 21.7, 9.65

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Pointwise VF Sensitivity 21	Continuous	180	0, 33	29, 24.5, 9.64
Pointwise VF Sensitivity 22	Continuous	180	0, 38	30, 25.9, 9.68
Pointwise VF Sensitivity 23	Continuous	180	0, 35	31, 26.4, 10.5
Pointwise VF Sensitivity 24	Continuous	180	0, 36	30, 26.8, 10.5
Pointwise VF Sensitivity 25	Continuous	180	0, 34	30, 25.3, 10.1
Pointwise VF Sensitivity 26	Continuous	180	0, 33	23, 20, 9.56
Pointwise VF Sensitivity 27	Continuous	180	0, 35	27, 24.4, 9.27
Pointwise VF Sensitivity 28	Continuous	180	0, 32	24, 19.8, 9.44
Pointwise VF Sensitivity 29	Continuous	180	0, 32	26, 22.9, 9.27
Pointwise VF Sensitivity 30	Continuous	180	0, 34	29.5, 25.7, 9.43
Pointwise VF Sensitivity 31	Continuous	180	0, 35	31, 26.7, 9.62
Pointwise VF Sensitivity 32	Continuous	180	0, 36	31, 27.8, 9.81
Pointwise VF Sensitivity 33	Continuous	180	0, 35	31, 28.2, 9.65
Pointwise VF Sensitivity 34	Continuous	180	0, 35	29, 26.6, 9.53
Pointwise VF Sensitivity 35	Continuous	180	0, 32	0, 3.31, 6.43
Pointwise VF Sensitivity 36	Continuous	180	0, 34	28, 24.9, 9.06
Pointwise VF Sensitivity 37	Continuous	180	0, 34	26, 22.6, 9.21
Pointwise VF Sensitivity 38	Continuous	180	0, 34	28, 24.7, 9.27
Pointwise VF Sensitivity 39	Continuous	180	0, 36	30, 26.6, 9.58
Pointwise VF Sensitivity 40	Continuous	180	0, 33	29, 26.3, 9.53
Pointwise VF Sensitivity 41	Continuous	180	0, 36	30, 26.6, 9.64
Pointwise VF Sensitivity 42	Continuous	180	0, 36	30, 26.3, 9.35
Pointwise VF Sensitivity 43	Continuous	180	0, 34	28, 25.4, 9.01
Pointwise VF Sensitivity 44	Continuous	180	0, 34	28, 24.8, 9.01
Pointwise VF Sensitivity 45	Continuous	180	0, 34	27, 23.8, 8.88
Pointwise VF Sensitivity 46	Continuous	180	0, 33	27, 24.8, 8.83
Pointwise VF Sensitivity 47	Continuous	180	0, 36	28, 25.4, 9.17
Pointwise VF Sensitivity 48	Continuous	180	0, 33	29, 25.5, 9.14
Pointwise VF Sensitivity 49	Continuous	180	0, 36	29, 26, 9.23

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Pointwise VF Sensitivity 50	Continuous	180	0, 36	28, 25.5, 9.23
Pointwise VF Sensitivity 51	Continuous	180	0, 34	26, 23.2, 8.88
Pointwise VF Sensitivity 52	Continuous	180	0, 33	27, 23.7, 8.71
Pointwise VF Sensitivity 53	Continuous	180	0, 36	28, 24.8, 8.95
Pointwise VF Sensitivity 54	Continuous	180	0, 34	28, 24.7, 8.84

Table A.2: Dataset B.

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Patient ID	Nominal	629	-	-
Condition	Nominal (Glaucoma, Healthy)	629	-	-
Gender	Nominal (M, F)	629	-	-
Eye	Nominal (Left, Right)	629	-	-
Age	Continuous (Months)	629	248, 991	820, 792, 115
Disc Area (Total)	Continuous	629	1.22, 2.92	1.80, 1.86, 0.43
Disc Area (Temporal)	Continuous	629	0.27, 0.72	0.42, 0.44, 0.11
Disc Area (Temporal-Superior)	Continuous	629	0.15, 0.38	0.24, 0.24, 0.06
Rim Area (Total)	Continuous	629	0.63, 2.43	1.11, 1.16, 0.32
Rim Area (Temporal)	Continuous	629	0.03, 0.55	0.17, 0.186 0.08
Rim Area (Temporal-Superior)	Continuous	629	0.04, 0.33	0.14, 0.13, 0.05
Rim Area (Temporal-Inferior)	Continuous	629	0.01, 0.29	0.14, 0.13, 0.05
Rim Area (Nasal)	Continuous	629	0.11, 0.66	0.34, 0.35, 0.10
Rim Area (Nasal-Superior)	Continuous	629	0.07, 0.33	0.17, 0.17, 0.05
Rim Area (Nasal-Inferior)	Continuous	629	0.09, 0.31	0.17, 0.17, 0.04
MRA Score (Total)	Continuous	629	-0.28, 0.27	0.05, 0.03, 0.10

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
MRA Score (Temporal)	Continuous	629	-0.50, 0.20	0.08, 0.048, 0.12
MRA Score (Temporal-Superior)	Continuous	629	-0.27, 0.17	0.03, 0.01, 0.08
MRA Score (Temporal-Inferior)	Continuous	629	-0.40, 0.23	0.06, 0.03, 0.13
MRA Score (Nasal)	Continuous	629	-0.50, 0.61	0.22, 0.20, 0.19
MRA Score (Nasal-Superior)	Continuous	629	-1.09, 0.34	0.05, 0.02, 0.18
MRA Score (Nasal-Inferior)	Continuous	629	-0.37, 0.39	0.09, 0.08, 0.15
RNFL thickness (Total)	Continuous	629	0.01, 0.46	0.19, 0.19, 0.07
RNFL thickness (Temporal)	Continuous	629	-0.32, 0.31	0.07, 0.07, 0.05
RNFL thickness (Temporal-Superior)	Continuous	629	-0.12, 0.62	0.22, 0.22, 0.10
RNFL thickness (Temporal-Inferior)	Continuous	629	-0.2, 0.46	0.19, 0.18, 0.09
RNFL thickness (Nasal)	Continuous	629	-0.02, 0.61	0.22, 0.23, 0.11
RNFL thickness (Nasal-Superior)	Continuous	629	-0.05, 0.72	0.27, 0.27, 0.10
RNFL thickness (Nasal-Inferior)	Continuous	629	-0.02, 0.59	0.28, 0.28, 0.10
VF sensitivity (Temporal)	Continuous	629	1.6, 30.4	24.5, 23.4, 4.2
VF sensitivity (Temporal-Superior)	Continuous	629	13.5, 32.2	26.4, 26.0, 2.8
VF sensitivity (Temporal-Inferior)	Continuous	629	8.0, 33.8	27.2, 26.7, 3.1
VF sensitivity (Nasal)	Continuous	629	10.3, 34.6	30.5, 30.0, 3.0
VF sensitivity (Nasal-Superior)	Continuous	629	11.2, 31.7	26.8, 26.3, 3.0
VF sensitivity (Nasal-Inferior)	Continuous	629	15.1, 33.5	28.8, 28.5, 2.3
Pointwise VF Sensitivity 1	Continuous	180	0, 32	24, 22.2, 5.6
Pointwise VF Sensitivity 2	Continuous	180	0, 32	23, 22.3, 5.7
Pointwise VF Sensitivity 3	Continuous	180	0, 31	23, 21.8, 5.94
Pointwise VF Sensitivity 4	Continuous	180	0, 33	23, 21.4, 6.54
Pointwise VF Sensitivity 5	Continuous	180	0, 32	25, 24.7, 4.22

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Pointwise VF Sensitivity 6	Continuous	180	2, 37	26, 25.3, 4.23
Pointwise VF Sensitivity 7	Continuous	180	0, 33	26, 25.6, 4.26
Pointwise VF Sensitivity 8	Continuous	180	0, 34	26, 25.1, 4.32
Pointwise VF Sensitivity 9	Continuous	180	0, 36	26, 25.1, 4.15
Pointwise VF Sensitivity 10	Continuous	180	0, 33	25, 23.8, 5.38
Pointwise VF Sensitivity 11	Continuous	180	0, 34	26, 24.8, 4.3
Pointwise VF Sensitivity 12	Continuous	180	9, 32	26, 26.4, 3.18
Pointwise VF Sensitivity 13	Continuous	180	10, 37	28, 27.9, 2.96
Pointwise VF Sensitivity 14	Continuous	180	9, 35	29, 28, 3.23
Pointwise VF Sensitivity 15	Continuous	180	10, 36	28, 28.1, 2.93
Pointwise VF Sensitivity 16	Continuous	180	15, 34	28, 27.5, 2.93
Pointwise VF Sensitivity 17	Continuous	180	10, 32	26, 26, 3.61
Pointwise VF Sensitivity 18	Continuous	180	0, 32	26, 24.1, 5.17
Pointwise VF Sensitivity 19	Continuous	180	0, 35	25, 23.7, 6.37
Pointwise VF Sensitivity 20	Continuous	180	0, 35	27, 26.4, 4.65
Pointwise VF Sensitivity 21	Continuous	180	7, 37	29, 28.9, 3.32
Pointwise VF Sensitivity 22	Continuous	180	0, 36	30, 29.3, 3.85
Pointwise VF Sensitivity 23	Continuous	180	0, 36	30, 29, 5.12
Pointwise VF Sensitivity 24	Continuous	180	0, 43	30, 28.6, 5.39
Pointwise VF Sensitivity 25	Continuous	180	0, 34	28, 26.1, 6.46
Pointwise VF Sensitivity 26	Continuous	180	0, 0	0, 0, 0
Pointwise VF Sensitivity 27	Continuous	180	0, 33	27, 23.8, 6.83
Pointwise VF Sensitivity 28	Continuous	180	0, 34	26, 24, 6.25
Pointwise VF Sensitivity 29	Continuous	180	0, 35	28, 27, 4.4
Pointwise VF Sensitivity 30	Continuous	180	5, 36	30, 29.6, 3.47
Pointwise VF Sensitivity 31	Continuous	180	6, 37	31, 30.3, 2.94
Pointwise VF Sensitivity 32	Continuous	180	16, 39	31, 30.7, 2.69
Pointwise VF Sensitivity 33	Continuous	180	0, 40	31, 30.4, 3.01
Pointwise VF Sensitivity 34	Continuous	180	3, 34	29, 28.3, 3.78

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Pointwise VF Sensitivity 35	Continuous	180	0, 0	0, 0, 0
Pointwise VF Sensitivity 36	Continuous	180	0, 35	26, 25.2, 5.26
Pointwise VF Sensitivity 37	Continuous	180	0, 34	26, 25.7, 4.77
Pointwise VF Sensitivity 38	Continuous	180	5, 36	28, 27.7, 3.31
Pointwise VF Sensitivity 39	Continuous	180	15, 36	30, 29.5, 2.67
Pointwise VF Sensitivity 40	Continuous	180	16, 39	30, 29.5, 2.51
Pointwise VF Sensitivity 41	Continuous	180	17, 38	30, 29.7, 2.4
Pointwise VF Sensitivity 42	Continuous	180	19, 36	30, 29.5, 2.52
Pointwise VF Sensitivity 43	Continuous	180	4, 34	28, 27.7, 3.74
Pointwise VF Sensitivity 44	Continuous	180	1.5, 38	28, 26.5, 3.81
Pointwise VF Sensitivity 45	Continuous	180	6, 33	27, 26.5, 3.73
Pointwise VF Sensitivity 46	Continuous	180	10, 35	29, 27.6, 3.2
Pointwise VF Sensitivity 47	Continuous	180	8, 37	28, 27.8, 3.27
Pointwise VF Sensitivity 48	Continuous	180	10, 36	28, 28.1, 3.16
Pointwise VF Sensitivity 49	Continuous	180	6, 35	29, 27.9, 3.23
Pointwise VF Sensitivity 50	Continuous	180	8, 36	28, 27.2, 3.83
Pointwise VF Sensitivity 51	Continuous	180	0, 34	26, 25.9, 3.88
Pointwise VF Sensitivity 52	Continuous	180	0, 35	27, 26.4, 4.07
Pointwise VF Sensitivity 53	Continuous	180	0, 33	27, 26.6, 3.98
Pointwise VF Sensitivity 54	Continuous	180	0, 48	26, 26.3, 4.25

Table A.3: Dataset C.

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Patient ID	Nominal	912	-	-
Condition	Nominal (Pre-Surgery,Post-Surgery)	912	-	-

Variable Name	Type/Domain	Entries	Min, Max	Median, Mean, SD
Gender	Nominal (M, F)	912	-	-
Eye	Nominal (Left, Right)	912	-	-
Age	Continuous (Months)	912	521, 1003	800, 803, 102
VF sensitivity (Temporal)	Continuous	629	10, 28.6	21.6, 20.8, 4.4
VF sensitivity (Temporal-Superior)	Continuous	629	10, 35.8	26, 26, 5.1
VF sensitivity (Temporal-Inferior)	Continuous	629	10, 30.8	23.8, 22.5, 4.8
VF sensitivity (Nasal)	Continuous	629	10.7, 33.8	27.3, 26.3, 4.9
VF sensitivity (Nasal-Superior)	Continuous	629	10.7, 37	27.9, 27.7, 5.4
VF sensitivity (Nasal-Inferior)	Continuous	629	10.3, 31.5	25, 24.4, 3.8

A.2 Chapter 4 Additional Tables and Results

Silhouette average index for clustering techniques can be found in Table A.4.

A.3 Dynamic Stages Bayesian Network Clustering Experiment

This section reports the results obtained using the DSBN model on a non-glaucoma dataset, provided by Transport for London. The model was trained using an unsupervised approach. The longitudinal dataset is a set of counts of London Underground passengers during a weekday in 2009. Counts are the number of entries in 30 stations during two-hour intervals over 24 hours. Therefore, each station presents 11 ordered values. Data was discretized in 20 states using a frequency based approach to obtain uniform values and expose the qualitative patterns in the data. A discrete variable was added to each time series, representing the unobserved group of belonging for each station. The results are shown in Figure A.1.

Table A.4: Silhouette average index for clustering techniques. Values are presented for healthy, converters and with all subjects (Mean). The models included are Clustering Naiive Bayes (CNB), Structural Expectation-Maximization (SEM), K-Means (K-M) and Clustering Structural Annealing (CSA).

Classifier (Class)	Silhouette	Mean Silhouette
K-M (Healthy)	0.26	0.19
K-M (Glaucomatous)	0.12	
CNB (Healthy)	0.12	0.10
CNB (Glaucomatous)	0.08	
SEM (Healthy)	0.10	0.11
SEM (Glaucomatous)	0.13	
CSA (Healthy)	0.15	0.08
CSA (Glaucomatous)	0.02	

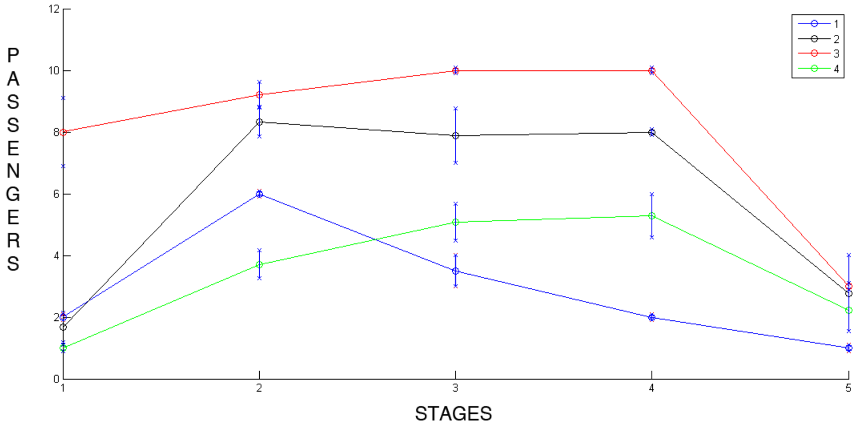


Figure A.1: Mean and SD values for each clusters over the 5 key stages identified for London Underground dataset.

Each of the four clusters is shown as a continuous line passing through the mean values estimated for each stage. Typical patterns are observed, including peaks in certain hours corresponding to the beginning and end of office hours (morning and evening) and high values throughout the day. Furthermore, underground stations located near to train stations (e.g. Euston, assigned to Cluster 2) present double peak patterns because of commuters, while stations associated with nightlife (e.g. Covent Garden, assigned to Cluster 4) present a peak at night time (before closure of the station at midnight). All-day busy stations are also identified, such as Kings Cross (which is home to six different tube lines and is next to St. Pancras International rail station) and Green Park (which is located near to many tourist attractions) do not show clear peaks, and are instead busy for most of the day almost uniformly. These stations were assigned to Cluster 3.

References

Airaksinen, P. J., Tuulonen, A and Alanko, H. I. (1992). “Rate and pattern of neuroretinal rim area decrease in ocular hypertension and glaucoma”. *Archives of Ophthalmology* 110.2, p. 206.

Akaike, H. (Dec. 1974). “A new look at the statistical model identification”. English. *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

American Academy of Ophthalmology (2010). *Preferred practice pattern guidelines: Primary Open-Angle Glaucoma.* Tech. rep. San Francisco: American Academy of Ohpthalmology.

American Optometric Association (2010). *Optometric clinical practice guideline: care of the patient with open angle glaucoma.* Tech. rep. American Optometric Association.

Anderson, D. R. (1996). “Glaucoma, capillaries and pericytes. 1. Blood flow regulation”. *Ophthalmologica* 210.5, pp. 257–262.

Anderssen, K. E. and Jeppesen, V (1998). “Classifying visual field data”. PhD thesis.

Artes, P. H. and Chauhan, B. C. (May 2005). “Longitudinal changes in the visual field and optic disc in glaucoma”. *Progress in retinal and eye research* 24.3, pp. 333–354.

Asman, P and Heijl, A (June 1992). “Glaucoma Hemifield Test. Automated visual field evaluation.” *Archives of ophthalmology* 110.6, pp. 812–9.

Azuara-Blanco, A et al. (2003). “Clinical agreement among glaucoma experts in the detection of glaucomatous changes of the optic disk using simultaneous stereoscopic photographs”. *American Journal of Ophthalmology* 136.5, pp. 949–950.

Baird, H. S. (1993). “Recognition technology frontiers”. *Pattern Recognition Letters* 14.4, pp. 327–334.

Bathija, R, Gupta, N, Zangwill, L and Weinreb, R. N. (1998). “Changing definition of glaucoma”. *Journal of glaucoma* 7.3, p. 165.

Bellman, R (1961). *Adaptive control processes: a guided tour.* Princeton, New Jersey: Princeton University Press.

Bengtsson, B (1981). “The prevalence of glaucoma.” *British medical journal* 65.1, p. 46.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer New York:

Bizios, D, Heijl, A and Bengtsson, B (2007). “Trained artificial neural network for glaucoma diagnosis using visual field data: a comparison with conventional algorithms”. *Journal of glaucoma* 16.1, p. 20.

Bowd, C et al. (2002). “Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc”. *Investigative ophthalmology & visual science* 43.11, p. 3444.

Bowd, C et al. (2004). “Confocal scanning laser ophthalmoscopy classifiers and stereophotograph evaluation for prediction of visual field abnormalities in glaucoma-suspect eyes”. *Investigative ophthalmology & visual science* 45.7, p. 2255.

Bowd, C et al. (2005). “Relevance vector machine and support vector machine classifier analysis of scanning laser polarimetry retinal nerve fiber layer measurements”. *Investigative ophthalmology & visual science* 46.4, p. 1322.

Bowd, C et al. (2006). “Structure-function relationships using confocal scanning laser ophthalmoscopy, optical coherence tomography, and scanning laser polarimetry”. *Investigative ophthalmology & visual science* 47.7, p. 2889.

Bowd, C. and Goldbaum, M. H. (June 2008). “Machine Learning Classifiers in Glaucoma”. *Optometry & Vision Science* 85.6, p. 396.

Bowd, C. et al. (Mar. 2008). “Bayesian machine learning classifiers for combining structural and functional measurements to classify healthy and glaucomatous eyes”. *Investigative ophthalmology & visual science* 49.3, p. 945.

Bradley, A. P. (1997). “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. *Pattern Recognition* 30.7, pp. 1145–1159.

Breiman, L, Friedman, J., Olshen, R and Stone, C (1984). *Classification and regression trees.* Washington, DC: Chapman & Hall/CRC.

Brigatti, L, Hoffman, D and Caprioli, J (May 1996). “Neural networks to identify glaucoma with structural and functional measurements.” *American journal of ophthalmology* 121.5, pp. 511–21.

Brigatti, L, Nouri-Mahdavi, K, Weitzman, M and Caprioli, J (1997). “Automatic detection of glaucomatous visual field progression with neural networks”. *Archives of Ophthalmology* 115.6, p. 725.

Brillinger, D. R. (1981). *Time series: data analysis and theory (Vol. 36).* SIAM.

Buhrmann, R. R. et al. (Jan. 2000). “Prevalence of glaucoma in a rural East African population.” *Investigative ophthalmology & visual science* 41.1, pp. 40–8.

Burgansky-Eliash, Z et al. (2005). “Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study”. *Investigative ophthalmology & visual science* 46.11, p. 4147.

Caprioli, J, Prum, B and Zeyen, T (1996). “Comparison of methods to evaluate the optic nerve head and nerve fiber layer for glaucomatous change”. *American Journal of Ophthalmology* 121.6, pp. 659–667.

Carletta, J. (June 1996). “Assessing agreement on classification tasks: the kappa statistic”. *Computational Linguistics* 22.2, pp. 249–254.

Cartwright, J. M. and Anderson, D. R. (1988). “Correlation of asymmetric damage with asymmetric intraocular pressure in normal-tension glaucoma (low-tension glaucoma)”. *Archives of Ophthalmology* 106, pp. 898–900.

Ceccon, S., Garway-Heath, D. F., Crabb, D. P. and Tucker, A. J. (2010a). “Combining Expertise-Driven and Semi-Supervised Bayesian Networks for classification of early Glaucoma”. *ECML PKDD 2010 (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases).* Barcellona, Spain.

— (2010b). “Investigations of Clinical Metrics and Anatomical Expertise with Bayesian Network Models for Classification in Early Glaucoma”. *IDAMAP 2010 (Intelligent Data Analysis in Biomedicine and Pharmacology).* Washington, USA.

— (2011a). “Ensembles of Bayesian Network Classifiers Using Glaucoma Data and Expertise”. *Ensembles in Machine Learning Applications - Studies in Computational Intelligence.* Berlin, Germany: Springer, (373)133–148.

— (2011b). “The Dynamic Stage Bayesian Network: identifying and modelling key stages in a temporal process.” *Advances in Intelligent Data Analysis X - Lecture Notes in Computer Science.* Berlin, Germany: Springer, (7014) 101–112.

Ceccon, S., Kotecha, A., Khaw, P. T. and Tucker, A. J. (2012a). “Clustering post-trabeculectomy glaucoma patients using non-stationary dynamic bayesian networks”. *IDAMAP 2012 (Intelligent Data Analysis in Biomedicine and Pharmacology).* Pavia, Italy.

Ceccon, S., Garway-Heath, D. F., Crabb, D. P. and Tucker, A. J. (2012b). “Non-Stationary Clustering Bayesian Networks for Glaucoma”. *ICML 2012 (International Conference on Machine Learning).* Edinburgh, Scotland.

Chan, K et al. (2002). “Comparison of machine learning and traditional classifiers in glaucoma diagnosis”. *IEEE Transactions on Biomedical Engineering* 49.9, pp. 963–974.

Chatfield, C. (1996). *The analysis of Time Series Fifth Edition.* London, UK: Chapman & Hall/CRC.

Chauhan, B. C., Drance, S. M. and Douglas, G. R. (1990). “The use of visual field indices in detecting changes in the visual field in glaucoma”. *Investigative ophthalmology & visual science* 31.3, p. 512.

Chauhan, B. C., House, P. H., McCormick, T. A. and LeBlanc, R. P. (1999). “Comparison of conventional and high-pass resolution perimetry in a prospective study of patients with glaucoma and healthy controls”. *Archives of Ophthalmology* 117.1, p. 24.

Chauhan, B. C., Blanchard, J. W., Hamilton, D. C. and LeBlanc, R. P. (2000). “Technique for detecting serial topographic changes in the optic disc and peripapillary retina using scanning laser tomography”. *Investigative ophthalmology & visual science* 41.3, p. 775.

Chauhan, B. C., McCormick, T. A., Nicolela, M. T. and LeBlanc, R. P. (2001). “Optic disc and visual field changes in a prospective longitudinal study of patients with glaucoma: comparison of scanning laser tomography with conventional perimetry and optic disc photography”. *Archives of Ophthalmology* 119.10, p. 1492.

Chauhan, B. C., Nicolela, M. T. and Artes, P. H. (2009). “Incidence and rates of visual field progression after longitudinally measured optic disc change in glaucoma”. *Ophthalmology* 116.11, p. 2110.

Chrichton, A, Drance, S. M., Douglas, G. R. and Schulzer, M (1989). “Unequal intraocular pressure and its relation to asymmetric visual field defects in low-tension glaucoma”. *Ophthalmology* 96, pp. 1312–1314.

Coffey, M et al. (1993). “Prevalence of glaucoma in the west of Ireland”. *British Journal of Ophthalmology* 77, p. 17.

Collaborative Normal-Tension Glaucoma Study Group (1998). “Comparison of glaucomatous progression between untreated patients with normal-tension glaucoma and patients with therapeutically reduced intraocular pressures”. *American Journal of Ophthalmology* 126, pp. 487–497.

Committee, P. B. A. G. A. (1996). *Criteria for adjunctive screening devices.* Tech. rep. Schaumburg, IL.

Crabb, D. P. (2012). *Optometry and Visual Science, City University London.*

Dagum, P, Galper, A and Horvitz, E (1992). “Dynamic network models for forecasting”. *Proceedings of the Eighth Workshop on Uncertainty in Artificial Intelligence*, pp. 41–48.

Davanger, M, Ringvold, A and Blika, S (Oct. 1991). “The probability of having glaucoma at different IOP levels.” *Acta ophthalmologica* 69.5, pp. 565–8.

Dean, T. and Kanazawa, K. (Feb. 1989). “A model for reasoning about persistence and causation”. *Computational Intelligence* 5.2, pp. 142–150.

Demirel, S. and Johnson, C. A. (Feb. 2000). “Isolation of short-wavelength sensitive mechanisms in normal and glaucomatous visual field regions”. *Journal of glaucoma* 9.1, pp. 63–73.

Demirel, S. et al. (Feb. 2009). “Predicting progressive glaucomatous optic neuropathy using baseline standard automated perimetry data”. *Investigative ophthalmology & visual science* 50.2, p. 674.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). “Maximum Likelihood from Incomplete Data via the EM Algorithm”. *Journal of the Royal Statistical Society* 39.1, pp. 1–38.

Dielemans, I et al. (1994). “The prevalence of primary open-angle glaucoma in a population-based study in the Netherlands”. *Ophthalmology* 101.11.

Drance, S. M. (Feb. 1969). “The Early Field Defects in Glaucoma”. *Invest. Ophthalmol. Vis. Sci.* 8.1, pp. 84–91.

Duin, R. P. W. and Tax, D. M. J. (2000). “Experiments with Classifier Combining Rules”. *Multiple Classifier Systems* 31.15, pp. 16–29.

Dumais, S, Platt, J, Heckerman, D and Sahami, M (1998). “Inductive learning algorithms and representations for text categorization”. *Proceedings of the seventh international conference on Information and knowledge management.* ACM New York, NY, USA, pp. 148–155.

Fitzke, F. W., Hitchings, R. A., Poinoosawmy, D, McNaught, A. I. and Crabb, D. P. (1996). “Analysis of visual field progression in glaucoma.” *British medical journal* 80.1, p. 40.

Flammer, J., Drance, S. M. and Zulauf, M. (1984). “Differential Light Threshold”. *Archives of ophthalmology* 102.

Friedman, N and Koller, D (2003). “Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks”. *Machine Learning* 50.1, pp. 95–125.

Friedman, N, Geiger, D and Goldszmidt, M (1997). “Bayesian network classifiers”. *Machine Learning* 29.2, pp. 131–163.

Friedman, N, Murphy, K and Russell, S (1998). “Learning the structure of dynamic probabilistic networks”. *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI98).* Citeseer, pp. 139–147.

Friedman, N, Goldszmidt, M and Wyner, A (1999). “On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks”. *AI&STAT VII.*

Furey, T. S. et al. (2000). “Support vector machine classification and validation of cancer tissue samples using microarray expression data”. *Bioinformatics* 16.10, p. 906.

Gaasterland, D. E., Ederer, F, Sullivan, E. K., Caprioli, J and Cyrlin, M. N. (1994). “Advanced glaucoma intervention study. 2. Visual field test scoring and reliability”. *Ophthalmology* 101.8, pp. 1445–1455.

Gardiner, S. K., Crabb, D. P., Fitzke, F. W. and Hitchings, R. A. (Apr. 2004). “Reducing noise in suspected glaucomatous visual fields by using a new spatial filter.” *Vision research* 44.8, pp. 839–48.

Gardiner, S. K., Johnson, C. A. and Cioffi, G. A. (Oct. 2005). “Evaluation of the structure-function relationship in glaucoma.” *Investigative ophthalmology & visual science* 46.10, pp. 3712–7.

Garg, A., Pavlovic, V and Huang, T. S. (2002). “Bayesian Networks as ensemble of Classifiers”. *Pattern Recognition* 51, p. 61801.

Garway-Heath, D. F. (2005). “Moorfields regression analysis”. *The Essential HRT Primer.San Ramon, California: Jocoto Advertising Inc*, pp. 31–39.

Garway-Heath, D. F., Poinoosawmy, D, Fitzke, F. W. and Hitchings, R. A. (2000). “Mapping the visual field to the optic disc in normal tension glaucoma eyes”. *Ophthalmology* 107.10, pp. 1809–1815.

Goldbaum, M. H. et al. (1994). “Interpretation of automated perimetry for glaucoma by neural network ”. *Investigative ophthalmology & visual science* 35.9, p. 3362.

Goldbaum, M. H. et al. (2002). “Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry”. *Investigative ophthalmology & visual science* 43.1, p. 162.

Goldbaum, M. H. et al. (2004). “Probability of Glaucoma Determined from Standard Automated Perimetry and from Optic Disk Topography using Relevance Vector Machine Classifiers”. *Investigative Ophtalmology and Visual Science* 45.5, p. 2137.

Goldbaum, M. H. et al. (2005). “Using unsupervised learning with independent component analysis to identify patterns of glaucomatous visual field defects”. *Investigative ophthalmology & visual science* 46.10, p. 3676.

Greaney, M. J. et al. (2002). “Comparison of optic nerve imaging methods to distinguish normal eyes from those with glaucoma”. *Investigative ophthalmology & visual science* 43.1, p. 140.

Grzegorczyk, M and Husmeier, D (2009). “Non-stationary continuous dynamic Bayesian networks”. *Advances in Neural Information Processing Systems.* Curran Associates, pp. 682–690.

Haas, A, Flammer, J and Schneider, U (Feb. 1986). “Influence of age on the visual fields of normal subjects.” *American journal of ophthalmology* 101.2, pp. 199–203.

Hall, M. et al. “The WEKA Data Mining Software : An Update”. 11.1, pp. 10–18.

Hand, D. J., Mannila, H and Smyth, P (2001). *Principles of Data MIning.* The MIT Press.

Hansen, L. K. and Salamon, P (1990). “Neural network ensembles”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10, pp. 993–1001.

Hart Jr, W. M. and Becker, B (Mar. 1982). “The onset and evolution of glaucomatous visual field defects”. *Ophthalmology* 89.3, pp. 268–279.

Harwerth, R. S., Carter-Dawson, L, Shen, F, Smith III, E. L. and Crawford, M. L. J. (1999). “Ganglion cell losses underlying visual field defects from experimental glaucoma”. *Investigative ophthalmology & visual science* 40.10, p. 2242.

Heckerman, D (1995). *A tutorial on learning with Bayesian networks.* Tech. rep. Redmond, WA.

Heijl, A et al. (1990). “Extended empirical statistical package for evaluation of single and multiple fields in glaucoma: Statpac 2”. *Perimetry update* 91, pp. 303–315.

Heijl, A, Leske, M. C., Bengtsson, B, Hyman, L and Hussein, M (2002). “Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial”. *Archives of Ophthalmology* 120.10, p. 1268.

Heijl, A. and Bengtsson, B. (Jan. 1996). “The Effect of Perimetric Experience in Patients With Glaucoma”. *Archives of Ophthalmology* 114.1, p. 19.

Henson, D. B., Spenceley, S. E. and Bull, D. R. (1996). “Spatial classification of glaucomatous visual field loss.” *British medical journal* 80.6, p. 526.

Hood, D. C. and Kardon, R. H. (2007). “A framework for comparing structural and functional measures of glaucomatous damage”. *Progress in retinal and eye research* 26.6, pp. 688–710.

Horn, F. K., Jonas, J. B., Martus, P, Mardin, C. Y. and Budde, W. M. (1999). “Polarimetric measurement of retinal nerve fiber layer thickness in glaucoma diagnosis”. *Journal of glaucoma* 8.6, p. 353.

Huang, D et al. (1998). “Optical coherence tomography”. *SPIE MILESTONE SERIES MS* 147, pp. 324–327.

Huang, M. L. and Chen, H. Y. (2005). “Development and comparison of automated classifiers for glaucoma diagnosis using Stratus optical coherence tomography”. *Investigative ophthalmology & visual science* 46.11, p. 4121.

Huang, M. L., Chen, H. Y. and Lin, J. C. (2007). “Rule extraction for glaucoma detection with summary data from StratusOCT”. *Investigative ophthalmology & visual science* 48.1, p. 244.

Ibanez, M. V., Simo, A, Ibáñez, M. V. and Simó, A (Dec. 2007). “Spatio-temporal modeling of perimetric test data”. *Statistical methods in medical research* 16.6, p. 497.

Jaakkola, T. S. and Haussler, D (1999). “Exploiting generative models in discriminative classifiers”. *Advances in neural information processing systems*, pp. 487–493.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). “Adaptive mixtures of local experts”. *Neural computation* 3.1, pp. 79–87.

Johnson, C. A. et al. (2000). “The relationship between structural and functional alterations in glaucoma: a review”. *Seminars in Ophthalmology.* Vol. 15. 4. Informa UK Ltd UK, pp. 221–233.

Johnson, C. A. et al. (Feb. 2003). “Structure and function evaluation (SAFE): II. Comparison of optic disk and visual field characteristics”. *American Journal of Ophthalmology* 135.2, pp. 148–154.

Kamal, D. and Hitchings, R. A. (July 1998). “Normal tension glaucoma - A practical approach”. *British Journal of Ophthalmology* 82.7, pp. 835–840.

Kamal, D. S. et al. (Mar. 1999). “Detection of optic disc change with the Heidelberg retina tomograph before confirmed visual field change in ocular hypertensives converting to early glaucoma”. *British Journal of Ophthalmology* 83.3, pp. 290–294.

Kass, M. A. et al. (2002). “The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma”. *Archives of Ophthalmology* 120.6, p. 701.

Khaw, P. T. (2001). “Advances in glaucoma surgery: evolution of antimetabolite adjunctive therapy”. *Journal of glaucoma* 10.5, pp. 81–4.

Kirkpatric, S, Gelatt, C. D., Vecchi, M. P., Kirkpatrick, S and Gelatt Jr, C. D. (1983). “Optimization by simulated annealing”. *Science* 220.4598, p. 671.

Kittler, J (1998). “Combining classifiers: A theoretical framework”. *Pattern Analysis & Applications* 1.1, pp. 18–27.

Klein, B. E. et al. (Oct. 1992). “Prevalence of glaucoma. The Beaver Dam Eye Study”. *Ophthalmology* 99.10, pp. 1499–1504.

Kohonen, T (2001). *Self-Organizing Maps, Third Edition.* New York: Springer.

Kohonen, T et al. (2004). “Self-organizing map”. *Neural Networks Research Centre BIENNIAL REPORT 2002 2003.* Hut, Finland: Helsinki University of Technology, pp. 113–122.

Kotecha, A. et al. (Oct. 2009). “Optic disc and visual field changes after trabeculectomy.” *Investigative ophthalmology & visual science* 50.10, pp. 4693–9.

Kronfeld, P. C. (1952). “Tonography”. *Archives of Ophthalmology* 48.4, p. 393.

Lee, T. W. and Lewicki, M. S. (2000). “The generalized Gaussian mixture model using ICA”. *International Workshop on Independent Component Analysis (ICA00)*, pp. 239–244.

Leske, M. C., Connell, A. M. S., Schachat, A. P. and Hyman, L (1994). “The Barbados Eye Study: prevalence of open angle glaucoma”. *Archives of Ophthalmology* 112.6, p. 821.

Leske, M. C., Heijl, A and Hussein, M (2003). “Factors for glaucoma progression and the effect of treatment: the early manifest glaucoma trial”. *Archives of Ophthalmology* 121, pp. 48–56.

Lietman, T, Eng, J, Katz, J and Quigley, H. A. (1999). “Neural networks for visual field analysis: how do they compare with other algorithms?” *Journal of glaucoma* 8.1, p. 77.

Lin, A, Hoffman, D, Gaasterland, D. E. and Caprioli, J (2003). “Neural networks to identify glaucomatous visual field progression”. *American Journal of Ophthalmology* 135.1, pp. 49–54.

Mackenzie, P. J. and Cioffi, G. A. (2008). *Measuring Structure and Function in Patients with Glaucoma.* New York.

Madigan, D and Raftery, A. E. (1994). “Model selection and accounting for model uncertainty in graphical models using Occam’s window”. *Journal of the American Statistical Association* 89.428, pp. 1535–1546.

Mardin, C. Y. and Jünemann, A. G. (2001). “The diagnostic value of optic nerve imaging in early glaucoma”. *Current opinion in ophthalmology* 12, pp. 100–104.

Mardin, C. Y., Horn, F. K., Jonas, J. B. and Budde, W. M. (1999). "Preperimetric glaucoma diagnosis by confocal scanning laser tomography of the optic disc". *British Journal of Ophthalmology* 83.3, p. 299.

Mardin, C. Y., Peters, A, Horn, F, Jünemann, A. G. and Lausen, B (2006). "Improving glaucoma diagnosis by the combination of perimetry and HRT measurements". *Journal of glaucoma* 15.4, p. 299.

Mason, R. P. et al. (Sept. 1989). "National survey of the prevalence and risk factors of glaucoma in St. Lucia, West Indies. Part I. Prevalence findings". *Ophthalmology* 96.9, pp. 1363–1368.

McClish, D. K. (1989). "Analyzing a portion of the ROC curve." *Medical decision making : an international journal of the Society for Medical Decision Making* 9.3, pp. 190–5.

McNeil, B. J. and Hanley, J. A. (Jan. 1984). "Statistical approaches to the analysis of receiver operating characteristic (ROC) curves." *Medical decision making : an international journal of the Society for Medical Decision Making* 4.2, pp. 137–50.

Michelson, G and Groh, M. J. M. (2001). "Screening models for glaucoma". *Current opinion in ophthalmology* 12.2, p. 105.

Mikelberg, F. S. et al. (1995). "Ability of the Heidelberg Retina Tomograph to detect early glaucomatous visual field loss". *Journal of glaucoma* 4.4, p. 242.

Mitchell, P, Smith, W, Attebo, K and Healey, P. R. (Oct. 1996). "Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study". *Ophthalmology* 103.10, pp. 1661–1669.

Mitchell, P, Hourihan, F, Sandbach, J and Wang, J. J. (1999). "The relation between glaucoma and myopia: the Blue Mountains Eye Study". *Ophthalmology* 106, pp. 2010–2015.

Munkwitz, S, Funk, J, Loeffler, K. U., Harbarth, U and Kremmer, S (2004). “Sensitivity and specificity of scanning laser polarimetry using the GDx”. *British Journal of Ophthalmology* 88.9, p. 1142.

Murphy, K (1998). *A brief introduction to graphical models and Bayesian networks.* http://www.cs.ubc.ca/murphyk/Bayes/bnintro.html.

— (2001). “The Bayes Net Toolbox for Matlab”. *Computing Science and Statistics*, p. 33.

National Eye Institute (2006). *Glaucoma: what you should know.* Tech. rep. Bethesda, US: US National Institutes of Health.

O’Connor, D. J., Zeyen, T and Caprioli, J (Oct. 1993). “Comparisons of methods to detect glaucomatous optic nerve damage.” *Ophthalmology* 100.10, pp. 1498–503.

Parzen, E (1962). “On estimation of a probability density function and mode”. *The annals of mathematical statistics* 33.3, pp. 1065–1076.

Pearl, J (2000). *Causality: Models, Reasoning, and Inference.* New York, New York, USA: Cambridge University Press.

Pearl, J and Shafer, G (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann.

Pizzarello, L et al. (2004). “VISION 2020: The Right to Sight: a global initiative to eliminate avoidable blindness”. *Archives of Ophthalmology* 122.4, p. 615.

Podos, S. M. and Becker, B (Apr. 1973). “Tonography - current thoughts”. *American Journal of Ophthalmology* 75.4, pp. 733–735.

Quigley, H. a. (May 1996). “Number of people with glaucoma worldwide”. *British Journal of Ophthalmology* 80.5, pp. 389–393.

Quigley, H. A. and Addicks, E. M. (1981). “Regional differences in the structure of the lamina cribrosa and their relation to glaucomatous optic nerve damage”. *Archives of Ophthalmology* 99.1, p. 137.

Quigley, H. A. and Anderson, D. R. (1976). “The dynamics and location of axonal transport blockade by acute intraocular pressure elevation in primate optic nerve”. *Investigative ophthalmology & visual science* 15.8, p. 606.

Quigley, H. A. and Broman, A. T. (Mar. 2006). “The number of people with glaucoma worldwide in 2010 and 2020”. *British Journal of Ophthalmology* 90.3, p. 262.

Quigley, H. A., Dunkelberger, G. R. and Green, W. R. (May 1989). “Retinal ganglion cell atrophy correlated with automated perimetry in human eyes with glaucoma”. *American Journal of Ophthalmology* 107.5, pp. 453–464.

Quigley, H. A., West, S. K. and Rodriguez, J (2001). “The prevalence of glaucoma in a population-based study of Hispanic subjects: Proyecto VER”. *Archives of Ophthalmology* 118, pp. 1105–1111.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* San Matteo, CA: Morgan Kaufmann.

Racette, L. et al. (Mar. 2010). “Combining functional and structural tests improves the diagnostic accuracy of relevance vector machine classifiers.” *Journal of glaucoma* 19.3, pp. 167–75.

Ramakrishnan, R et al. (Aug. 2003). “Glaucoma in a rural population of southern India: the Aravind comprehensive eye survey.” *Ophthalmology* 110.8, pp. 1484–90.

Resnikoff, S et al. (2004). “Global data on visual impairment in the year 2002”. *Bulletin of the World . . .* 012831.04, pp. 844–851.

Robinson, J. W. and Hartemink, A. J. (2008). “Non-stationary dynamic Bayesian networks”. *Advances in Neural Information*, pp. 1–2.

Rousseeuw, P. J. (Nov. 1987). “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. *Journal of Computational and Applied Mathematics* 20.1, pp. 53–65.

Russell, R. A., Malik, R., Chauhan, B. C., Crabb, D. P. and Garway-Heath, D. F. (May 2012). “Improved estimates of visual field progression using bayesian linear regression to integrate structural information in patients with ocular hypertension.” *Investigative ophthalmology & visual science* 53.6, pp. 2760–9.

Samal, A and Iyengar, P. A. (1992). “Automatic recognition and analysis of human faces and facial expressions: a survey”. *Pattern Recognition* 25.1, pp. 65–77.

Sample, P. A. et al. (2002). “Using machine learning classifiers to identify glaucomatous change earlier in standard visual fields”. *Investigative ophthalmology & visual science* 43.8, p. 2660.

Sample, P. A. et al. (2005). “Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields”. *Investigative ophthalmology & visual science* 46.10, p. 3684.

Sample, P. A. et al. (Aug. 2004). “Using unsupervised learning with variational Bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects”. *Investigative ophthalmology & visual science* 45.8, p. 2596.

Sanchez-Galeana, C et al. (2001). “Using optical imaging summary data to detect glaucoma”. *Ophthalmology* 108.10, pp. 1812–1818.

Schuman, J. S. et al. (1995). “Quantification of nerve fiber layer thickness in normal and glaucomatous eyes using optical coherence tomography: a pilot study”. *Archives of Ophthalmology* 113.5, p. 586.

Schuman, J. S. et al. (2003). “Comparison of optic nerve head measurements obtained by optical coherence tomography and confocal scanning laser ophthalmoscopy”. *American Journal of Ophthalmology* 135.4, pp. 504–512.

Schwarz, G (1978). “Estimating the dimension of a model”. *The annals of statistics* 6.2, pp. 461–464.

Selim, S. Z. and Alsultan, K. (Jan. 1991). “A simulated annealing algorithm for the clustering problem”. *Pattern Recognition* 24.10, pp. 1003–1008.

Shapiro, S. and Wilk, M. (1965). “An analysis of variance test for normality (complete samples)”. *Biometrika.*

Sharkey, A. J. C. (Dec. 1996). “On combining artificial neural nets”. *Connection Science* 8.3, pp. 299–314.

Shields, M., Ritch, R and Krupin, T (1996). “Classifications of the glaucomas”. *The Glaucomas, Vol. 2 Clinical Science.* St. Louis, pp. 717–25.

Shiose, K. Y. and Tsukahara, S (1990). “A collaborative glaucoma survey for 1988 in Japan”. *Jpn J Clin Ophthalmol* 44, p. 653.

Shiose, K. Y., Tsukahara, S and Kitazawa, Y (1991). “Epidemiology of glaucoma in Japan - a nationwide glaucoma survey”. *Jap J Ophthalmol* 35, pp. 133–155.

Sommer, A and Doyne, L (1996). “Glaucoma: facts and fancies”. *Eye* 10.Pt 3, pp. 295–301.

Sommer, A et al. (1991a). “Clinically detectable nerve fiber atrophy precedes the onset of glaucomatous field loss”. *Archives of Ophthalmology* 109.1, p. 77.

Sommer, A et al. (Aug. 1991b). “Relationship between intraocular pressure and primary open angle glaucoma among white and black Americans. The Baltimore Eye Survey.” *Archives of ophthalmology* 109.8, pp. 1090–5.

Spry, P. G. D. and Johnson, C. A. (2002). “Identification of Progressive Glaucomatous Visual Field Loss”. *Survey of ophthalmology* 47.2, pp. 158–173.

Stein, J. D., Girkin, C. A., Harizman, N and al, E. (2005). “Comparison of false-positive test results among the Stratus OCT 3, the GDx-VCC, and the HRT II”. *2005 Association for Research in Vision and Ophthalmology Annual Meeting.* Ft. Lauderdale, FL.

Strouthidis, N. G. et al. (Dec. 2006). “Structure and function in glaucoma: The relationship between a functional visual field map and an anatomic retinal map”. *Investigative ophthalmology & visual science* 47.12, p. 5356.

Swift, S and Liu, X (2002). “Predicting glaucomatous visual field deterioration through short multivariate time series modelling”. *Artificial Intelligence in Medicine* 24.1, pp. 5–24.

Talih, M. and Hengartner, N. (June 2005). “Structural learning with time-varying components: tracking the cross-section of financial time series”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.3, pp. 321–341.

The AGIS Investigators (2000). “The Advanced Glaucoma Intervention Study (AGIS):7. The relationship between control of intraocular pressure and visual field deterioration”. *American Journal of Ophthalmology* 130, pp. 429–440.

Tielsch, J. M. (1991). “The epidemiology of primary open angle glaucoma”. *Ophthalmol Clin North Am* 4, p. 649.

Tielsch, J. M. et al. (1991). “Racial variations in the prevalence of primary open-angle glaucoma. The Baltimore Eye Survey”. *Jama* 266.3, p. 369.

Tielsch, J. M., Katz, J, Sommer, A and al, E. (1995). “Hypertension, perfusion pressure, and primary open-angle glaucoma. A population-based assessment”. *Archives of Ophthalmology* 113, pp. 216–221.

Tjon-Fo-Sang, M. J. and Lemij, H. G. (Jan. 1997). “The sensitivity and specificity of nerve fiber layer measurements in glaucoma as determined with scanning laser polarimetry”. *American Journal of Ophthalmology* 123.1, pp. 62–69.

Tresp, V (2001). “Mixtures of Gaussian processes”. *Advances in Neural Information Processing Systems*, pp. 654–660.

Tucker, A., Liu, X., Garway-Heath, D and Unit, G (2003). “Bayesian classification and forecasting of visual field deterioration”. *Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP).* Citeseer.

Tucker, A., Vinciotti, V., Liu, X. and Garway-Heath, D. (June 2005). “A spatio-temporal Bayesian network classifier for understanding visual field deterioration.” *Artificial intelligence in medicine* 34.2, pp. 163–77.

Tuulonen, A and Airaksinen, P. J. (Apr. 1991). “Initial glaucomatous optic disk and retinal nerve fiber layer abnormalities and their progression”. *American Journal of Ophthalmology* 111.4, pp. 485–490.

Tuulonen, A, Lehtola, J and Airaksinen, P. J. (May 1993). “Nerve fiber layer defects with normal visual fields. Do normal optic disc and normal visual field indicate absence of glaucomatous abnormality?” *Ophthalmology* 100.5, pp. 587–588.

Uchida, H, Brigatti, L and Caprioli, J (Nov. 1996). “Detection of structural damage from glaucoma with confocal laser image analysis.” *Investigative ophthalmology & visual science* 37.12, pp. 2393–401.

Vapnik, V. N. (2000). *The nature of statistical learning theory.* Berlin, Germany: Springer Verlag.

Vesti, E, Johnson, C. A. and Chauhan, B. C. (Sept. 2003). “Comparison of different methods for detecting glaucomatous visual field progression”. *Investigative ophthalmology & visual science* 44.9, p. 3873.

Viswanathan, A. C. et al. (June 2003). “Interobserver agreement on visual field progression in glaucoma: a comparison of methods.” *The British journal of ophthalmology* 87.6, pp. 726–30.

Weber, J, Dannheim, F and Dannheim, D (1990). “The topographical relationship between optic disc and visual field in glaucoma”. *Acta Ophthalmol* 68, pp. 568–574.

Wellman, M. P. and Henrion, M (1993). “Explaining ”Explaining Away””. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.3, pp. 187–92.

Wikstrand, M. H., Hå rd, A.-L., Niklasson, A. and Hellström, A. (Mar. 2010). “Birth weight deviation and early postnatal growth are related to optic nerve morphology at school age in children born preterm.” *Pediatric research* 67.3, pp. 325–9.

Wirtschafter, J. D., Becker, W. L., Howe, J. B. and Younge, B. R. (Mar. 1982). “Glaucoma visual field analysis by computed profile of nerve fiber function in optic disc sectors”. *Ophthalmology* 89.3, pp. 255–267.

Wollstein, G, Garway-Heath, D. F. and Hitchings, R. a. (Aug. 1998). “Identification of early glaucoma cases with the scanning laser ophthalmoscope”. *Ophthalmology* 105.8, pp. 1557–1563.

Wong, E. Y., Keeffe, J. E., Rait, J. L. and al, E. (2003). “Detection of undiagnosed glaucoma by eye health professionals”. *Ophthalmology* 111, pp. 1508–1514.

Woods, K, Bowyer, K and Kegelmeyer Jr, W. P. (1996). “Combination of multiple classifiers using local accuracy estimates”. *1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1996. Proceedings CVPR’96*, pp. 391–396.

Wormald, R. P., Basauri, E, Wright, L. A. and Evans, J. R. (1994). “The African Caribbean Eye Survey: risk factors for glaucoma in a sample of African Caribbean people living in London”. *Eye* 8.Pt 3, pp. 315–320.

Xuan, X. and Murphy, K. (2007). “Modeling changing dependency structure in multivariate time series”. *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007).* Omnipress, pp. 1055–1062.

Yanoff, M. and Duker, J. S. (2003). *Ophthalmology 2nd Edition.* St. Louis: Mosby, pp. 1,652.

— (2008). *Ophthalmology 3rd edition.* St. Louis.

Zangwill, L, Shakiba, S, Caprioli, J and Weinreb, R. N. (Apr. 1995). “Agreement between clinicians and a confocal scanning laser ophthalmoscope in estimating cup/disk ratios”. *American Journal of Ophthalmology* 119.4, pp. 415–421.

Zangwill, L. M., Van Horn, S, de Souza Lima, M, Sample, P. A. and Weinreb, R. N. (1996). “Optic nerve head topography in ocular hypertensive eyes using confocal scanning laser ophthalmoscopy”. *American Journal of Ophthalmology* 122.4, pp. 520–525.

Zangwill, L. M. et al. (2004). “Racial Differences in Optic Disc Topography”. *Archives of Ophthalmology* 122.1, pp. 22–28.

Zoubir, A. M. and Boashash, B (1998). “The bootstrap and its application in signal processing”. *IEEE Signal Processing Magazine* 15.1, pp. 56–76.