DEVELOPMENT OF BAYESIAN NETWORK MODELS FOR OBSTRUCTIVE SLEEP APNEA SYNDROME ASSESSMENT

by

Anne Gisèle Fouron
B.Sc, McGill University, 1991

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
School
of
Computing Science

SIMON FRASER UNIVERSITY

Spring 2006

APPROVAL

Name: Anne Gisèle Fouron

Degree: Master of Science

Title of Thesis: **Development of Bayesian Network models for Obstructive Sleep Apnea Syndrome Assessment**

Supervisory Committee:

Chair: Dr. Richard Vaughan

Dr. M. Stella Atkins
Senior Supervisor
Professor of Computing Science

Dr. Oliver Schulte
Supervisor
Associate Professor of Computing Science

Dr. Najib Ayas
Supervisor
Assistant Professor of Medicine, University of British Columbia

---

Dr. William S. Havens
Internal Examiner
Associate Professor of Computing Science

Date Defended/Approved: December 19, 2005

SIMON FRASER UNIVERSITY library

STATEMENT OF ETHICS APPROVAL

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

(a) Human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

(b) Advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

(c) as a co-investigator, in a research project approved in advance,

or

(d) as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Bennett Library
Simon Fraser University
Burnaby, BC, Canada

ABSTRACT

Bayesian Belief networks have been used for diagnosis in some medical domains and in this thesis we provide a methodology for creating Bayesian Networks to predict Obstructive Sleep Apnea Syndrome severity. We build 3 Bayesian Network topologies: by knowledge engineering, Naïve Bayes configuration and a third topology is created using results of the Naïve network. All networks are trained on data from 652 patients referred for an overnight polysomnogram. Data is derived from multiple data sources and includes a mix of continuous and discrete variables. We investigate the impact of different topologies and discretizing continuous variables, adding nodes with large amounts of missing values, and removing nodes from networks.

Results show that performance is dependent on the interaction between topology and discretization. Node removal increases sensitivity while node addition decreases it.

Keywords: Bayesian Belief Networks, Medicine, Obstructive Sleep Apnea Syndrome, machine learning

DEDICATION

To my parents, who taught me never to stop learning and who can still help me with my homework.

To my children, who show me that it is the simple things in life that bring the most joy.

To Karan, without whom I would not be complete.

ACKNOWLEDGEMENTS

To Dr. Stella Atkins I offer my endless gratitude for her understanding, motivating influence and support of my work and its slow pace over all these years. I definitely would not have been able to complete this without her guidance and help. Many thanks also to Drs. Najib Ayas and Oliver Shulte for their help, input and support. Thank you to the staff in the computer science office, especially Heather Muckart and Val Galat, to the Library Staff, particularly Penny Simpson, and to the many wonderful people I have met at SFU over the last seven years. Special thanks to Mila Kwiatkowska for her inspiration, help, motivation and for being the supportive friend and collegue that she is, to Karan and my parents who kept Zahra and Ronan at bay while I wrote this, and to Surinder Samra for helping me for the last four years with the munchkins, for her parathas and saag and her "no pootchie pootchie" attitude that kicked me in the pants when I needed it. Finally, I am eternally grateful to all my family and friends for their unwaivering patience, humour and moral support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:
INTRODUCTION AND MOTIVATION

"Expert systems...involve the application of various logical and computational techniques of AI to the representation of human knowledge for automated inference" [1]. In general, the term expert system is used to refer to a computer program that can aid with decision-making on a tightly delineated problem. Expert systems attempt to model domain knowledge in different ways and include rule-based systems, constraint-based systems, semantic networks, neural networks and Bayesian belief networks. Early expert systems focused on symbolic reasoning where logic was used to represent knowledge and solve problem. However, when faced with complex, real world problems, it became apparent that first order logic was insufficient for dealing with the uncertainty inherent in such problems. Different methods of handling this uncertainty include fuzzy logic, certainty factors, belief functions and probability theory [2].

Bayesian belief networks (BBNs) are useful for several different reasons. Firstly, they permit a graphical modelling of a network, allowing experts to concentrate on building a qualitative representation of a diagnostic problem before even thinking about quantitative specification [2]. Nodes in a graph represent random variables and directed edges between them represent a direct probabilistic influence of one variable over another. Secondly, prior knowledge such the accumulated knowledge possessed by an expert, can be combined with observed data to determine the final probability of an expected hypothesis [3]. Learning methods exist that, from a graph, or network structure

and a set of data, can determine the parameters (prior probabilities and conditional probabilities) of the network that can then be used to perform diagnostic inference.

Obstructive Sleep Apnea Syndrome (OSAS) is a significant health problem worldwide and in North America, estimated to affect at least three percent of the adult population [4] [5]. Positive diagnosis for OSAS requires an overnight polysomnogram (PSG) in a sleep clinic. Due to the paucity of sleep clinics, even those individuals that are referred to a sleep clinic for suspected OSAS need to wait lengthy periods of time before an overnight PSG can be scheduled, due to long waiting lists. In addition, the overnight PSG is an expensive way to diagnose individuals with OSAS. It would be useful to have a diagnostic system that could either diagnose patients, or prioritise them in terms of likely severity of illness without the requirement of an overnight PSG. Such a system could take as input simple measurements that could be performed by any community health nurse, and responses to questions such as how sleepy the individual was. One of the most common symptoms of OSAS is excessive daytime sleepiness, which can result in significant performance decrements, costly mistakes and even death. The annual cost associated with sleep-related accidents in the US alone was estimated at $56.02 billion and is said to account for a total of 52,650,000 lost work days in 1988 [5].

A variety of sleepiness detection methods exist. The cheapest and easiest measures of sleepiness are subjective questionnaires, but these have been found to be not completely reliable. A more objective measure of sleepiness, the Multiple Sleep Latency Test (MSLT), considered the gold standard of sleep propensity measures, is time consuming and as expensive to administer as an overnight PSG. Also, it is not readily available to populations living away from large urban centres. The analysis of various

eye-related parameters, both physiological and autonomic (obtained from pupillometry), provides a simple, short and relatively inexpensive test for ascertaining excessive daytime sleepiness that could be attributable to OSAS. Pupillometric data, in particular, has been shown to be quite reliable in discriminating between sleep-deprived and non sleep-deprived states in individuals, with daytime variations paralleling those of the MSLT [6], [7], [8], [9], [10] and [11]. However, studies have not yet been conducted to ascertain if pupillometric variables can be used to differentiate between people with different severities of OSAS. It would be useful to have an expert system, that could combine this pupillometry data with data acquired from subjective questionnaires and simple clinical measurements to assist physicians, by inferring the likelihood of an individual having OSAS and inferring with what severity he or she is suffering.

In this thesis, we look at the application of Bayesian belief networks to the assessment of Obstructive Sleep Apnea Syndrome (OSAS) severity, as represented by an Apnea Hypopnea Index (AHI) variable in the network. AHI is defined as the number of apnea (complete blockage of the airway) or hypopnea (partial airway obstruction) events per hour. In Section 3.2 we discuss the creation of a database from three different sources, including pupillometry data that we collected in order to assess the usefulness of pupillography data as a measure of sleepiness in our networks. In Sections 3.3, 3.4 and 3.5 we discuss the selection and computation of our data variables and the creation of a final data set to be used for training our networks. Section 3.6 discusses how we evaluate our networks. In Section 3.7 we discuss building three different Bayesian belief network topologies ranging from a naïve Bayes topology to a knowledge engineered topology created from literature review and with the assistance of a sleep expert. We study the

impact of network topology on a network's ability to accurately predict moderate to severe OSAS. In Section 3.8 we investigate how different strategies for discretizing the networks' continuous variables affect their predictive abilities. Section 3.9 discusses investigating the impact of adding nodes with very sparse data values (large amounts of missing data) or moderate amounts of missing data to an existing network. In section 3.10 we then look at the effect of removing existing nodes representing variables with low correlation to AHI from the network. Chapter 4 shows how we evaluate these BBNs using cross validation and compare performance in predicting moderate to severe OSAS, using measures of sensitivity, specificity and positive predictive value.

CHAPTER 2:LITERATURE REVIEW AND BACKGROUND

2.1 Obstructive Sleep Apnea

The term apnea is derived from the Greek word for the absence of breath. First discovered by European scientists in 1965 [12], it comes in two varieties.

The first, and most rare, is called central apnea. This term includes several disorders which are characterized by the lack of effort to pull air into the lungs when sleeping, either due to the diaphragm, the brain or the nerve connection between the two [12]. Essentially when a sufferer sleeps, they stop breathing. A common form of central apnea is known as Cheyne-stokes respiration.

By far, the most common form of apnea is called Obstructive Sleep Apnea Syndrome (OSAS). It is caused by the laxity of the muscles that dilate the upper airway (pharynx) during sleep. Normally, the laxity of the muscles of the upper airway increases when a person is sleeping, resulting in a narrowing of the airway. In normal people, this does not impede beathing. However, in some individuals, this narrowing can be so extreme at some points that it effectively blocks the air passages to the lungs.

From this point forth, when the term apnea is used, it will be referring to obstructive sleep apnea syndrome (OSAS).

2.1.1 Symptoms

The typical OSAS sufferer is a middle aged, overweight man that snores loudly, although women, especially after menopause, and thin people can also suffer from the

disease. Risk factors include obesity, enlarged tonsils or lymph nodes and naturally small airways [12].

Symptoms indicating severe OSAS can include high blood pressure, cardiovascular problems, and most commonly, excessive daytime sleepiness (EDS). Other, less threatening symptoms include oesophageal reflux, frequent night time urination, heavy sweating at night, morning headaches, male impotence and a reduction in sex drive in both men and women [12].

2.1.2 Etiology and Pathogenesis

In OSAS, the negative pressure acts on an anatomically narrow airway (usually due to obesity, jaw and throat abnormalities etc.) resulting in the walls of the throat are pulled together by the suction created during inspiration.

As a result of the lack of air being sucked into the lungs, carbon dioxide builds up and oxygen levels drop sharply in the bloodstream. A normal individual's blood oxygen level during sleep is usually between 96% and 99%. Often OSAS sufferers are observed with blood oxygen levels as low as between 80 – 85% before his or her breathing resumes. Because cells, especially brain and nervous system cells, can die when blood oxygen levels drop below 90%, brain damage can be a result of OSAS [5], [12].

In addition to low blood oxygen, during a breath stoppage the heart may slow. In some cases, it stops beating for up to 11 seconds at a time. Also, the heart muscle itself is oxygen starved and as a result, can develop serious rhythm problems, the most serious of which is ventricular tachycardia: a wild uncoordinated rapid heartbeat. This condition is fatal if it continues [12].

It is not until oxygen levels drop to such low levels as those mentioned above that the stimulation becomes intense enough to awaken the brain momentarily. Then, the brain panics and the sleeper struggles awake. This arousal causes tongue and throat muscle activation resulting in the pharynx opening, which in turn allows oxygen back into the lungs in a series of gasping, snorting breaths. Patients with sleep apnea do not completely awaken after all this, but instead immediately fall asleep again. After a few seconds, when muscle tensility of the pharyx has again decreased, snoring resumes and the cycle begins again, repeating hundreds of times a night. The arousals from sleep, though life saving, result in surges in blood pressure (equivalent of what would be seen if the patient were lifting heavy weights) [12] that may occur hundreds or times per night, night after night. Surprisingly, most OSAS sufferers have no memory of their sleepless night.

There are numerous consequences to suffering from untreated obstructive sleep apnea. The surges in blood pressure that occur during the arousals from sleep may, over time lead to heart failure, strokes and cardiac ischemia. The National Commission on Sleep Disorders Research estimates that 38,000 heart attacks & strokes in United States per year are due to apnea [12].

In addition, there are consequences resulting from the disruption of sleep. Because sleep is so fragmented, it has little restorative value [12]. As such, extreme fatigue and sleepiness is one of the most common symptoms of OSAS. Other symptoms include a depressed mood, reduced work performance, loss of motivation, inability to concentrate, loss of short term memory, poor stamina, inefficient problem solving abilities, disorientation, depression, and increased risk of accidental injuries [5]. Patients

with OSAS frequently fall asleep while driving or eating, often without awareness. The most common danger is falling asleep while driving, with some studies suggesting that the accident rate of OSAS sufferers is ten times that of the general population [12].

2.1.3 Prevalence

OSAS is one of most serious general health problems in America. It is also one of the most under-diagnosed problems in medicine. OSAS affects approximately 4% of middle-aged men and 2% of middle-aged women. It has been estimated that 93% of women and 82% of men with moderate to severe SAS have not been clinically diagnosed [4].

2.1.4 Treatment

2.1.4.1 Continuous positive airway pressure (CPAP)

This therapy has been the most effective way of treating OSAS for the past 15 years. It entails putting a mask over the nose and ensuring the patient sleeps with his or her mouth closed. The CPAP machine gently blows air into the nose at a pressure slightly higher than the surrounding air pressure. The positive air pressure prevents the airway from collapsing during sleep. Patients often report dramatic increases of daytime alertness and energy after even just a few nights on CPAP [12].

2.1.4.2 Surgery

Multiple surgeries for the treatment of OSAS have been proposed, including uvulopalatopharyngoplasty (UPPP, a procedure that involves cutting away excess tissue at the back of the throat; the uvula, tonsils and parts of the soft palate), a procedure that combines standard UPPP with a procedure to pull the large tongue muscle forward and

away from the back of the throat, increasing the airway diameter without any visible external changes or a recent treatment that involves administering radio frequency waves to tissue in the upper airway, resulting in an overall reduction in tissue volume [12]. Whichever treatment is taken, often OSAS can be completely cured, unlike many other serious diseases.

Losing weight, quitting smoking, avoiding sedatives, nightcaps and allergens, sleeping on one's side instead of back and using a firm pillow and mattress can also all help to decrease the severity of the apnea experienced, although they cannot be considered treatment as they do not generally cure sleep apnea [5].

2.1.5 Diagnosis

One of the major problems with making a diagnosis of obstructive sleep apnea is that "many doctors do not consider the possibility unless patients present with sleepiness plus snoring as their prime complaints" [13]. General practitioners rarely take sleep histories and often, while every other possible cause for symptoms is explored, the possibility of a patient having OSAS is not. Often patients do not complain about sleep problems because they are not aware that their sleep is being compromised or they are not aware of their snoring or nocturnal apneas.

Presently, the diagnosis of OSAS depends on a polysomnography (PSG) test. This is a complex test that requires one night in hospital attended by a trained sleep technician. The PSG monitors a variety of physiological signals including air flow, thoraco-abdominal movement, airflow obstruction, snoring, sleep itself, oxygen saturation, heart rate and body position. From these various physiological signals an

Apnea Hypopnea Index (AHI) is calculated for the patient. This is the average number of total or partial blockages of the airway that occur per hour during the overnight study. AHI is often used to qualify how severe a given patient's condition is. In general, a patient with an AHI less than or equal to 15 is considered to have no OSAS up to mild OSAS, while an AHI over 15 indicates moderate to severe OSAS.

Traditionally, all of the above measurements are recorded during a sleep study for the diagnosis of OSAS and is too difficult to do in the patient's home. Unfortunately, this approach is expensive in terms of hospital space, equipment and staff. Often simpler and less expensive approaches are sufficient for diagnosis in many patients [13]. The most common is an at home overnight blood oxymetry recording.

2.2 Sleepiness Detection Methods

A major symptom of many sleep disorders, including OSAS, insomnia, narcolepsy and others is Excessive Daytime Sleepiness (EDS) ([14]. As a result, much research has been done to try and develop methods for detecting and quantifying daytime sleepiness. However, before entering into a discussion of these methods, it is important to note that the concept of sleepiness itself can be defined to mean different things. A few of these include: [15]

A state of languor or inertness

A subjective state of sleep need

A physiological drive resulting from sleep deprivation

A strong sleep propensity

2.2.1 Sleep Propensity Measures

The first class of sleepiness measures are based on a simple definition of sleep need: the greater an individual's sleep need, the sleepier they are [5]. The idea is that sleepiness is a state in which the propensity to fall asleep is urgent and therefore, the sleepier an individual is, the faster they will fall asleep [15].

The most famous such test is the Multiple Sleep Latency Test (MSLT). It polygraphically measures the time it takes a subject to fall asleep in a darkened room by using the EEG pattern to tell researchers when a subject's brain enters the first stage of sleep [5]. Each session of the test has a maximum duration of twenty minutes and it is administered at two-hour intervals throughout a day. The time a person takes to fall asleep is called their sleep latency. Sleep latency seems not only to be sensitive to circadian fluctuations of sleepiness, but also to sleep deprivation (both total and partial) and sleep extension. A normal person generally falls asleep within 10 to 15 minutes, while a sleep latency of less than five minutes usually indicates a medically significant sleep disorder [5] or a situation of significant sleep deprivation. This test is considered the gold standard for evaluating sleepiness. However, it is not useful in diagnosing insomniacs [15].

Other sleep propensity tests include the Polygraphic Index of Sleepiness and the Polygraphic Score of Sleepiness [15]. These evaluate not only the propensity to fall asleep, but also look at architecture of the sleep stages once a subject is asleep.

2.2.1.1 Disadvantages

Critiques of MSLT say that it measures only situational sleep propensity and that it does not clearly distinguish between normal and abnormal daytime sleepiness levels

[5]. Also, it has been shown that some people fall asleep very easily but are not really sleepy. In addition, the test is very expensive, time consuming, and most sleep clinics have significant waiting lists for the administration of such a test.

Based on the evidence that non-sleepy people sometimes can have very low sleep latencies, two other tests have been devised that test a subject's ability to stay awake in situations of low stimulus. These are the Repeated Test of Sustained Wakefulness. Subjects lie in bed with the lights off, and the Maintenance of Wakefulness Test in which subjects sit in an armchair in a non-stimulating environment [15].

In general, sleep propensity tests are considered optimal tests for evaluation of sleepiness, both for research purposes and for use in clinics, however, the invasiveness and lengthiness of such tests, as well as the equipment required, curtail its use in applied and field studies.

2.2.2 Subjective Sleepiness Detection Methods

The second class of measures for assessing sleepiness uses subjective methods. These methods, often in the form of scales or questionnaires, attempt to determine how sleepy people feel, by asking them to self-evaluate their own physical and cognitive symptoms, often by getting subjects to indicate which definition from a proposed set of definitions most closely matches their perceived state [16]. There are two categories of subjective measures. The first views sleepiness as a state-related condition that fluctuates as a function of time of day and can be induced by atypical situations such as sleep deprivation. The second views sleepiness as a steady and constant trait of a person [15].

2.2.2.1 Sleepiness as a state-related condition

There are a number of scales used to assess sleepiness as a state-related condition and how it functions throughout the day including the Stanford Sleepiness Scale, the Karolinska Sleepiness Scale, the Accumulated Time with Sleepiness Scale and the Visual Analogue Scale [15], [16]. The Stanford Sleepiness Scale, the most widely used, is a Lickert type scale, describing seven levels of vigilance. Subjects indicate which level describes their present state [15]. This scale has proven useful even for normal subjects and is currently the most widely used of all subjective sleepiness evaluation methods.

Unfortunately, such scales are not useful for patient diagnosis because of the fact that they are changeable and sensitive to circadian influences. For evaluation of sleepiness for the purposes of patient diagnosis, it is more useful to look at sleepiness as a permanent trait component.

2.2.2.2 Sleepiness as a permanent trait component

There are a number of scales and questionnaires that attempt to assess the overall sleepiness level of an individual, including the Epworth Sleepiness Scale, the Rotterdam Daytime Sleepiness Scale and the Sleep-Wake Activity Inventory [15].

The most commonly used measure of subjective sleepiness is the Epworth Sleepiness Scale (ESS) [17]. It is based on the hypothesis that all individuals are characterized by a constant level of sleepiness which is independent of their circadian or ultradian rhythms. This constant sleepiness level is evidenced by the propensity of an individual to fall asleep in low stimulus situations. A subject is therefore asked to record his or her probability of falling asleep under eight different situations. The ESS is often used to reveal daytime sleepiness and has been shown to discriminate between normal

and pathological subjects [17], [18]and [19]. The Rotterdam Daytime Sleepiness Scale and the Sleep-Wake Activity Inventory have been used but have not been as extensively validated as the ESS.

2.2.2.3 Drawbacks of Subjective Methods

Subjective sleepiness evaluation methods would be ideal if they were unequivocally proven to be reliable, valid and sensitive to certain pathologies. Unfortunately, this is not the case. These methods seem to be extremely vulnerable to environmental and/or motivational variables [15]. Some subjects do not report accurately. Certain subjects overestimate the severity of their disorder [15]. Still others negate their sleepiness partially or completely [5].

The ESS was originally designed to give a subjective report of objective sleep propensity in daily life. However, researchers have found that while the ESS should differentiate between different levels of sleepiness from no sleepiness to extreme sleepiness, most ESS items are located at the opposite extremes of this continuum. That is, that the ESS does not contain enough items representing situations of an intermediate sopophoric nature and hence that the sensitivity of ESS to detect intermediate levels of sleep propensity is limited [20]. When comparing ESS scores with objective sleepiness as measured by the MSLT in ten patients with OSAS it was found that there is no significant relationship between the ESS and the MSLT [21]. When 51 suspected OSAS patients were administered an ESS as well as an overnight PSG, it was found that while the ESS score correlated significantly with the overall arousal index, only a weak correlation was found between the ESS score and AHI [22]. This implies that while the ESS can be used to determine excessive daytime sleepiness, it cannot be used to diagnose

obstructive sleep apnea. For this, the AHI must also be used. In a study comparing ESS scores, results of overnight PSG and sleep latencies as measured by the MSLT on 225 subjects, it was found that the ESS was correlated with total sleep time, but not with sleep efficiency, nor with AHI. The authors concluded that the MSLT and the ESS are not interchangeable as the ESS is influenced by psychological factors, while the MSLT is not [23].

It seems clear that, unfortunately, there is not yet a subjective method of measuring sleepiness that has been unequivocally proven to be reliable and accurate. Therefore, one must look to objective measures of sleepiness: these are usually based on an evaluation of a state of decreased vigilance or on the basis of specific behavioural or physiological parameters [15].

2.2.3 Performance/ Behavioural measures of sleepiness

Performance-based or behavioural measures of sleepiness can be made using tasks in various categories including psychomotor tasks, cognitive tasks, attentional tasks and substitution tasks. Psychomotor tasks include the Wilkinson Auditory Vigilance Task, a variety of reaction time tests and tracking tasks [15]. Cognitive tests include addition, reasoning and memory tests while attentional tasks include tasks such as visual search [15].

2.2.3.1 Disadvantages

As with subjective measures of sleepiness, performance measures of sleepiness often suffer from the problem of being sensitive to and influenced by motivational and environmental variables. In addition they also suffer from methodological and statistical

problems [15]. That being said, they are often the only usable methods in field investigations.

2.2.4 Pupillometry and Other Arousal Decrease Measures

This class of sleepiness measures arose from an attempt to find more accurate measures of sleepiness based on physiological parameters and on the desire to assess sleepiness in awake and active subjects[15]. These include a number of different types of measures.

The first class of arousal decrease measures looks at electro-encephalography (EEG) parameters, such as alpha (8 – 12 Hz), and theta (4-8 Hz) band power increases (Alpha Attenuation Test [24]), and using evoked potentials (Eps) to evaluate vigilance variations.

Another class of arousal decrease measures uses data from electro-oculography (EOG). Hypovigilance, or sleepiness is characterized by changes in a variety of types of eye movement. Parameters that have been investigated as to their relation with sleepiness include smooth pursuit eye movements, saccades and blinks [25], [26], [27], [15], [28], [29], [30].

A third class of measures evaluates changes in the response of the Autonomic Nervous System (ANS) as a result of sleepiness. The most commonly used technique is called pupillometry. It involves quantifying parameters of both the Pupillary Light Reflex (PLR), the contraction of the pupil in response to visual stimuli and of the pupil's response to dark adaptation (sitting in a completely darkened room for an extended

period of time). These parameters have been found to be affected by fatigue and sleepiness [15], [11].

In dark adaptation, a subject is sitting in a darkened room; after the eye adjusts to the reduced light level a broad and steady pupil diameter is typical of a normal level of alertness while a contracted and changeable pupil is attributed to sleepiness and hypo activation of the autonomic nervous system ([31], as cited in [15]). Generally, in an alert subject, the pupil remains dilated in darkness, with an amplitude of pupil diameter change below 0.3 mm and a frequency of approximately 1 Hz. In sleepy subjects, not only does pupil diameter decrease with time, but in addition, the pupil oscillations can reach amplitudes of several millimeters and the frequency of such oscillations is generally lower (0.8 Hz or less) [11].

Researchers have also looked at measures such as the latency to pupil constriction and the amplitude of pupil constriction in response to a visual stimulus [32]. However, pupillometry results remain controversial as some studies have shown it is not clearly reliable [33],[34].

2.3 Review of Results of Objective Sleepiness Detection Using Pupillometry

2.3.1 Background

Pupillometers, devices that record the size of an individual's pupils over time, have been used for several decades. Lowenstein and Feinberg first observed that during pupillometry recording in the dark (dark adaptation), the pupils of sleepy subjects

oscillated widely in size, as seen in Figure 2.1 [35]. They named this phenomenon pupillary "fatigue waves." These fatigue waves can be quantified both in terms of the wave amplitude (the amount that the pupil diameter increases and decreases in size) and in terms of the frequencies of the waves. To analyze the frequencies of the fatigue waves, a Fast Fourier Transform is often performed. Another observation made by these researchers was that the pupils of sleepy subjects decreased in size over the period of dark adaptation with increasing sleepiness. This contrasted sharply with the pupillary behaviour of subjects deemed to be alert. In alert subjects, pupils maintained a stable size during dark adaptation, as seen in Figure 2.2. This research was followed by the development of the Alertness Level Test (ALT), a standardized pupillometry test which consists of 15 minutes of pupil size recording in the dark with infrared-sensitive video cameras while the person sits quietly with eyes open and staring at a stationary small red spot [36].

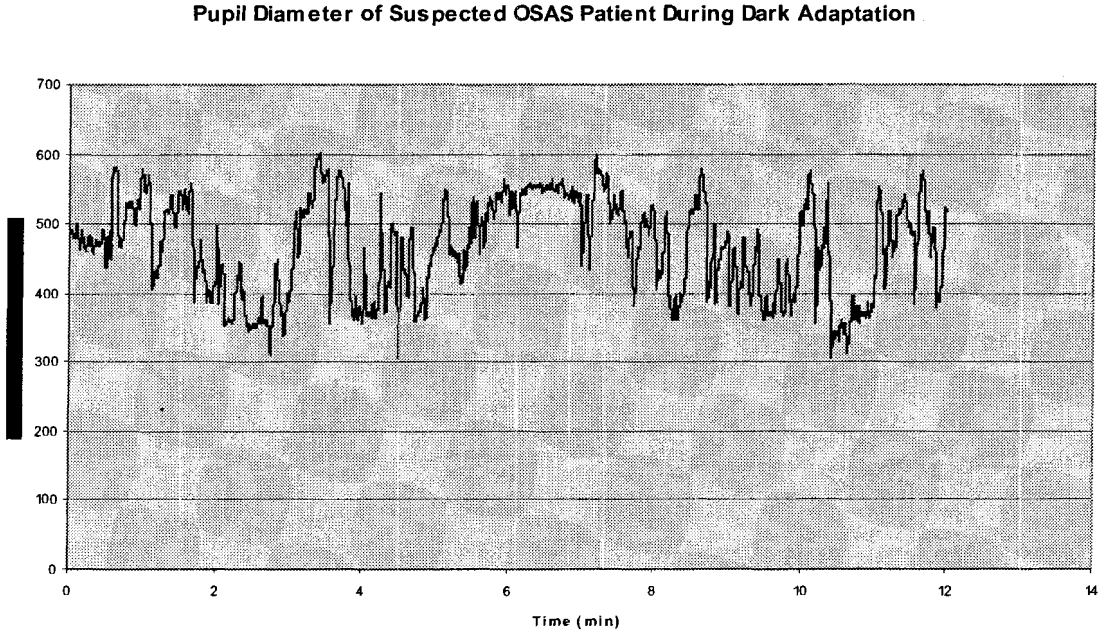

Figure 2.1 Pupil Diameter during an 11 minute ALT on an individual suspected of OSAS

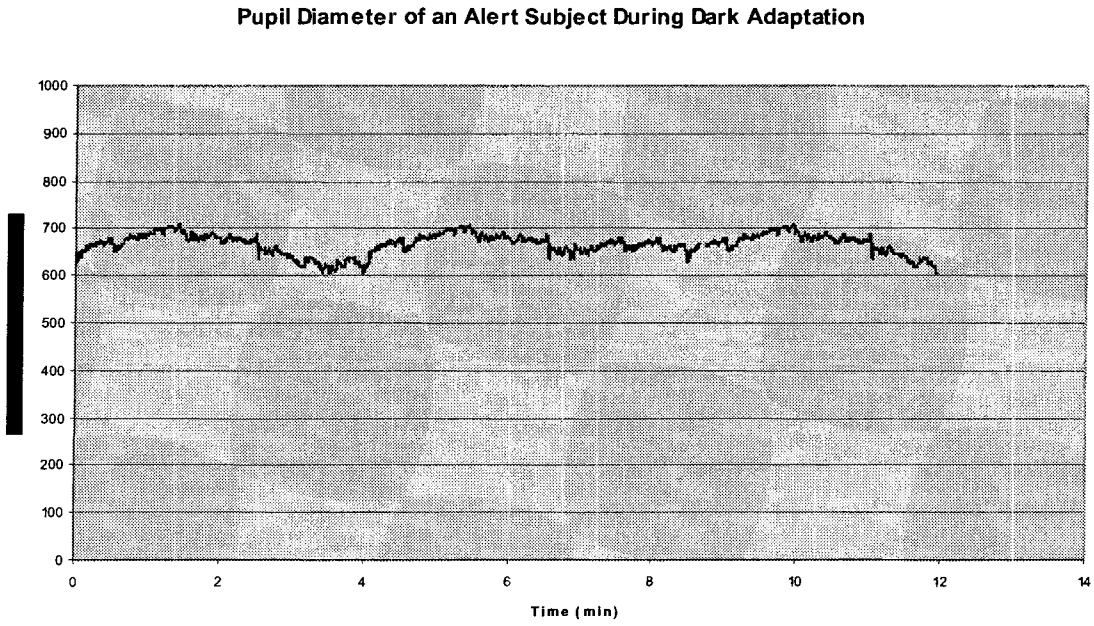

Figure 2.2 Pupil Diameter during an 11 minute ALT on an alert subject

In the past, researchers using the ALT have evaluated the level of sleepiness based on the visual inspection of graphs that depict changes in the pupil size over time; sleepiness was estimated by the expert doing the rating. The drawback to this was that ratings could vary widely between the individuals who were making a determination. However, in 1998, Barbara Wilhelm and her group at the pupil laboratory at Tubingen University published a paper detailing a set of techniques for mathematically analysing and quantifying the amount of pupil oscillation that occurred during an 11 minute ALT [37]. One such measure is the Pupillary Unrest Index (PUI). The PUI, because it allows a pupillary recording of dark adaptation to be objectively summarized in terms of the average amount of pupil oscillation over time, permits a mathematical comparison of the amount of oscillation between individuals as a measure of pupil size change in millimeters per minute. This opened the door for the use of pupillometry as an objective

measure of sleepiness. A summary of studies performed to investigate the relationship between parameters of dark adaptation and sleepiness can be found in Table 2.1, Table 2.2, Table 2.3 and Table 2.4. Table 2.1, Table 2.2 and Table 2.3 summarize studies investigating pupil dynamics during dark adaptation. Table 2.1 summarizes the data on Pupil diameter studies. Table 2.2 summarizes the data on pupil miosis studies. Table 2.3 and Table 2.4 summarize the data on studies examining the pupillary unrest index (PUI). Table 2.5 and Table 2.6 deal with studies looking at pupillary dynamics during the pupillary light reflex. Table 2.5 summarizes data from studies investigating pupil constriction latency and Table 2.6 summarizes data from studies investigating pupil constriction amplitude.

Table 2.1 Pupil Diameter Studies

Research Study	Subject Group	Parameter evaluated: pupil diameter
Wilhelm, H et al, 1998	Normal (7 subjects)	always larger, morning test: (6.04mm (+-0.72), afternoon test: 5.99mm (+- 0.61) pm test
	hypersomniac (3 narcoleptic subjects, 4 OSA subjects)	always smaller, but diff not statistically significant (only 7 subjects) morning test: 5.85mm (+- 0.7) , afternoon test: 5.62mm (+- 0.67)
Merritt, SL et al 1999 and 1998	normal pupil measurements taken repeatedly over 5 days at the same time of day(10 subjects)	(from 1994 study looking at avg. over 5 days: 6.5 +/- .89, 6.8 +/- .93, 6.7 +/- .87, 6.6 +/- 1.09, 6.84 +/- 1.25)
Danker-Hopfe, Kraemer et al 2001	12 Normal, healthy subjects tests every two hours from 7 am until 11:00 pm	found a decreased pupil diameter was an indicator of sleepiness, as assessed by the MSLT, but did not correspond to subjective sleepiness
Russo, Thomas et al 2003,	Control period (57 subjects)	the means of the four groups were between 5.4mm and 5.9mm
	7 days partial sleep deprivation (four groups restricted to 3, 5, 7 or 9 hours in bed) (57 subjects)	after 7 days, average initial pupil diameters: 3hr: 5.3mm, 5 hr: 5.7mm, 7 hr: 5.5mm, 9 hr 5.1mm, only the 9 hour group showed a significant decrease in pupil diameter ($p<0.05$, these were not sleep deprived)

Table 2.2 Pupil Miosis During Dark Adaptation Studies

Research Study	Subject Group	Parameter evaluated: pupil diameter miosis during dark adaptation
Richmond, Jack w EyeCheck	Normal (48 subjects)	Note: specific numbers not given
	sleep deprived for 24 hours (48 subjects)	Greater decrease correlate significantly ($p<.01$) with people sleep deprived. Sensitivity: 81.82%. Specificity: 94.29%
Merritt, SL et al 1999 and 1998	Normal subjects (9 subjects) measurements taken 4 times during day to observe circadian effect	Pupil miosis significantly greater during ALT given at 2 pm, than those given at 10 am, 12 noon and 4 pm. This parallels the results of the MSLT and Maintenance of wakefulness test, (no specific numbers given)

Table 2.3 Pupillary Unrest Index During Dark Adaptation Studies

Research Study	Subject Group	Parameter evaluated: pupillary unrest index (PUI) (variability of pupil diameter during dark adaptation)
Wilhelm, H et al, 1998	Normal (7 subjects)	morning test: 3.92 mm/min, (+- 0.95), afternoon test: 4.90 mm/min (+- 1.75)
	hypersomniac (3 narcoleptic subjects, 4 OSA subjects)	much higher morning test: 9.07 mm/min (+- 1.81), afternoon test: 9.74 mm/min (+- 3.59)
Merritt, SL et al 1999 and 1998	Normal subjects (9 subjects) measurements taken 4 times during day to observe circadian effect	increase in pupil size oscillation at 2 pm 10 am = 5.91 mm/min (+- 2.5), noon: 7.32 mm/min (+- 3.5), 2 pm: 11.24 mm/min (+- 7.9), 4pm: 7.52 mm/min (+- 2.4)
Danker-Hopfe, Kraemer et al 2001	12 Normal, healthy subjects tests every two hours from 7 am until 11:00 pm	- increased coeff of variation of pupil diam, - increased square root of power w/in 0.1-0.8Hz in pupillary oscillations(fatigue waves) - increased pupillary unrest index
[11]Wilhelm, B., Wilhelm, H. et al. 1998	Normal (13 subjects, 5 females, 8 males)	PUI increases significantly with increased sleep deprivation
[38]Wilheml, B., Wilhelm, H. et al. 1998	35 male OSAS patients before and after 3 months of nCPAP treatment	Significant decrease in PUI after 3 months (one sample t-test, $p<0.12$) of nCPAP

Table 2.4 Relative Pupillary Unrest Index During Dark Adaptation Studies

Research Study	Subject Group	Parameter evaluated: relative PUI (PUI / baseline pupil diam) during dark adaptation
Wilhelm, H et al, 1998	Normal (7 subjects)	morning test: 0.66, (+- 0.13), afternoon test: 0.81, (+- 0.23)
	hypersomniac (3 narcoleptic subjects, 4 OSA subjects)	much higher morning test: 1.57, (+- 0.39), afternoon test: 1.79 (+- 0.70)

Table 2.5 Pupil Constriction Latency During Pupillary Light Reflex Studies

Research Study	Subject Group	Parameter evaluated: latency to pupil constriction during pupillary light reflex
Dal Santo, Tousman et al 1997	Normal level of alertness (self-assessed)	
Note: total of 166 subjects, not known how many in what group)	fatigue (self assessed) in truck drivers	smaller latency (faster constriction velocity) in self-assessed fatigued
Krichmar, Thomas et al 1997	Normal (12 subjects)	decrease (faster)
	sleep deprived (total 64 hours) (same 12 subjects)	increase (i.e. response slower), is second largest contributor to index score of sleepiness, which is significant
Russo, Thomas et al 2003,	Control period (57 subjects)	baseline average 294 msec
	7 days partial sleep deprivation (four groups restricted to 3, 5, 7 or 9 hours in bed) (57 subjects)	after 7 days of sleep deprivation, group means are: 3 hr group: 302 msec, 5 hr group: 296 msec, 7 hr group: 294 msec, 9 hr group: 288 msec

Table 2.6 Pupil Constriction Amplitude During Pupillary Light Reflex Studies

Research Study	Subject Group	Parameter evaluated:amplitude of pupil constriction during pupillary light reflex
Dal Santo, Tousman et al 1997	Normal level of alertness (self-assessed)	
Note: total of 166 subjects, not known how many in what group)	fatigue (self assessed) in truck drivers	larger in self-assessed fatigued
Richmond, Jack w EyeCheck	Normal subjects (48)	
	24 hour sleep deprived (48 same subjects)	
Russo, Thomas et al 2003,	Control period (57 subjects)	baseline: 1.24, 1.26, 1.245, 1.305
	7 days partial sleep deprivation (four groups restricted to 3, 5, 7 or 9 hours in bed) (57 subjects)	increased signif in 3 hr group after 7 days: 3hr: 1.2, 5 hr: 1.245, 7 hr: 1.243, 9 hr: 1.33 mm

Pupillometry can also be used to quantify the eye's response to a visual stimulus. After a period of dark adaptation, the eye is subjected to a series of light flashes (the number of flashes, and time over which the flashes are shown varies depending on researchers). When light is focused upon the eye, the pupil constricts. This is called the direct light reflex. At the same time, the pupil of the other eye constricts too, and that is the consensual light reflex. There are a number of parameters that can be measured in this pupillary light reflex. These include the following:

- Constriction latency (time to initiation of pupil constriction)
- Time to minimum (time for the pupil to get to its position of maximum constriction
- Maximal constriction of pupil size during the response
- Reflex amplitude percentage (pupil constriction as a percentage)

The Pupillary Light Reflex parameters have also been investigated as to their sensitivity to sleepiness and have been summarized in Table 2.5 and Table 2.6.

2.3.2 Pupillometry as an Objective Measure of Sleepiness

Krichmar, Thomas et al. [7] found that the second largest contributor to an index score of sleepiness based on oculomotor measures was an increase in the latency of response of the pupillary light reflex as shown in Table 2.5, Row 3.

Danker-Hopfe, Kramer and their group [8] examined mean pupil diameter, its coefficient of variation and the pupillary unrest index during dark adaptation. They also looked at the square root of the power within the frequency band 0.1 – 0.8 Hz in pupillary oscillations and how it varied with time of day. They then analysed how these variables inter-related with the MSLT and the Stanford Sleepiness Scale. What they found was that the MSLT and pupillometric variables seem to measure the same dimension of sleepiness, whereas the Stanford Sleepiness Scale showed an almost opposite pattern of time of day variation. A decreased pupil diameter, increased coefficient of the variation of pupil diameter, increased Pupillary Unrest Index and an increase in fatigue waves of frequencies below 0.8 Hz were all deemed to be indicators of increased sleepiness, as measured by the MSLT as shown in Row 4 of Table 2.1 and Table 2.3. However, the variation of such indicators did not correspond with subjective sleepiness as measured by the Stanford Sleepiness Test.

Dr. Sharon Merritt and her group at the Center for Narcolepsy Research have done many studies using pupillometry. One study validated the reproducibility of Pupillometry measurements by testing a group of normal controls at the same time of day for five consecutive days. No difference in mean pupil diameter and mean PUI was found between days, indicating that individuals who maintained a consistent sleep schedule also had a consistent pupil diameter and pattern of pupil oscillation. Subsequent

studies have found that there is a significantly greater decrease in pupil size and significantly more oscillation in pupil size during the Alertness Level Test (ALT) administered at two p.m. as compared to those assessed at ten am, noon and four p.m. This parallels sleep latency as assessed by MSLT and indicates that both the magnitude of the decrease in pupil size and increase in pupil size oscillation during the Alertness Level Test are sensitive to circadian changes in the level of alertness [10] [9] as shown in Row 3 of Table 2.1. This same group also looked at increases in theta brain wave activity and how this correlated with decreases in pupil size in untreated narcoleptics, untreated obstructive sleep apnea and normal control subjects, as shown in Row 3 of Table 2.2. They found that for both narcoleptics and OSAS subjects the amount of theta activity was significantly greater for pupil stages corresponding to the largest decrease in pupil size, whereas in controls, the amount of theta activity did not increase significantly by pupil stage [39].

The Pupil Research Group at the Tubingen University in Germany has also conducted many pupillometry studies. One early study looked at the dark adaptation of the pupils of healthy subjects forced to remain awake from 7 pm to 7 am. During this period, their pupillary behaviour was recorded every two hours. It was found that the power of slow pupillary oscillations (<= 0.8 Hz) and PUI increased significantly, as did subjective sleepiness as assessed with the Stanford Sleepiness Scale as shown in Row 4 of Table 2.3 [11]. The group performed a similar study of dark adaptation and found that the pupillary unrest index, as well as the mean value of summed power values of frequencies below 0.8 Herz were significantly different when comparing a group of normal subjects with a group of hypersomniacs (narcoleptics and OSAS patients) [40], as

shown in Row 1 and 2 of Table 2.3. They also performed studies using pupillometry to investigate the effects of nCPAP treatment in obstructive sleep apnea patients. The first such study examined 35 patients with OSAS the day before therapy and again after after three months nCPAP treatment. They found comparison of morning values (10 am) in PST showed a significant reduction (mean 21%) of the PUI, whereas before the nCPAP therapy, the morning PST had been the same as those at the afternoon PST as shown in Row 5 of Table 2.3. There was very little change in the PST results for the afternoon and evening tests [38].

2.4 Bayesian Networks

Knowledge-based expert systems involve using various techniques of Artificial Intelligence in order to represent human knowledge for automated inference; that is, to formalize human expert knowledge in an attempt to improve human decision by allowing computer-based reasoning to do much of the work.

Especially when modelling human disease, its symptoms, diagnosis and co-variables, the goal of researchers is to investigate causal connections, the relative strengths of those connections and how to infer them from real, noisy observations. Disease pathologies are often rife with exceptions and lack of regularity. Not all patients suffering from a given disease exhibit all the usual diagnostic symptoms. Almost all symptoms can be manifested in response to multiple diseases. The connection between diseases and symptoms cannot therefore be easily summarized by first order logic for two reasons:

> Abduction confounds diagnosis. Abduction is the logical process of inferring a 'best explanation', or in our case, the most likely disease, from

a set of known facts (or symptoms). The problem is that there can be many explanations for a given observation (test result, or symptom). It is unusual that the manifestation of a symptom is proof-positive of the existence of a specific disease. Instead, it is usually the observation of several symptoms together that make it 'highly likely' that a patient had a certain disease.

Even if all the rules were known, there may still be uncertainty about a particular patient because not all possible tests have been run. [41]

Our thesis develops Bayesian networks for the purpose of OSAS assessment. An example of a Bayesian network can be seen in Figure 2.3. This network displays a number of diseases (AHI, as a stand-in for OSAS, PLM and Depression), a number of symptoms (High Blood Pressure, Oxygen Desaturation) as well as a number of variables that are thought to impact OSAS prevalence(Snoring, resulting from a narrow airway which is thought to increase the chances of having OSAS).

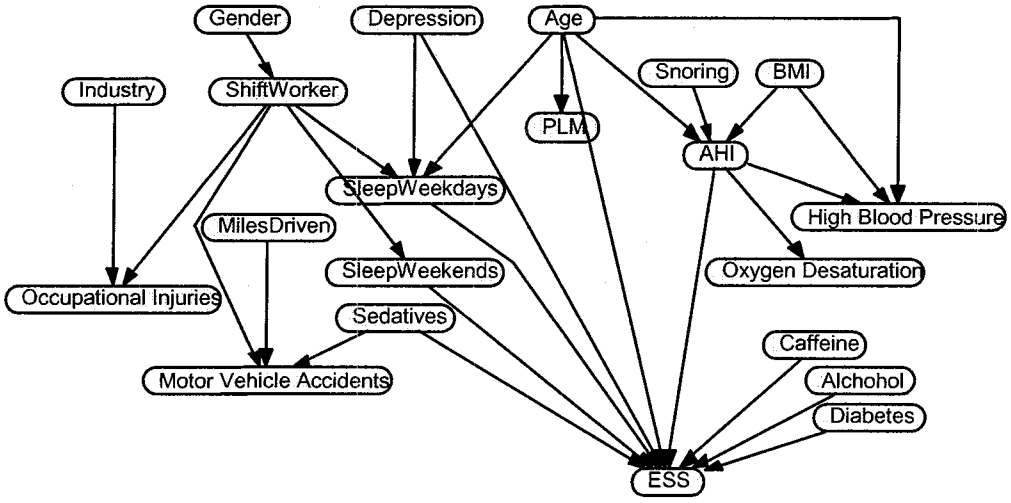

Figure 2.3 Bayesian Network for Prediction of OSAS Severity

Any expert system that hopes to aid in diagnosis must take these factors into account and deal with the uncertainty due to incomplete or conflicting data and/or

models. "Probability is a language for expressing uncertainty about propositions and quantities in terms of degrees of belief" [1, p.65]. It is a language that distinguishes various shades of likelihood and provides a method of summarizing the uncertainty inherent in a given domain, such as medicine.

2.4.1 Overview of Probability Theory

A probability is "simply a number expressing the chance that a proposition is true or that some event has occurred, with a value in the range from 0 (certainly false) to 1 (certainly true)." [1, p.65] There are three different views of probability. The first is the propensity or objectivist view; where probability is a physical property of something, a propensity of an object to behave in a certain way – for example the tendency of a given coin to land heads up in a sequence of coin tosses. The second is the frequency view, where probability is a property of a population of similar events – for example, the fraction of head's in a series of coin tosses. In this view, probability numbers can only come from experiments. The third view of probability is the subjective view, where a probability is defined as an expression of a person's degree of belief in a proposition or in the occurrence of a given event, based on the person's current information. It does not necessarily have any external physical significance. Using the coin example, a subjectivist would start with some prior belief about the tendency of a given coin to land head's up. As data is collected (by doing many coin tosses), this belief is updated based on the data being observed. After much data is gathered, in general the belief of the subjectivist will tend converge to the actual fraction of head's landing up as the data overwhelms the prior belief, and thus the subjectivist will agree with the frequentist. The major difference between the two, is that subjectivists are "willing to assign probabilities

to events that are not members of any obvious repeatable sequence" [1, p.65] whereas the frequentist is not [41], [1].

Problems in the medical domain, as with most other judgmental domains, almost always involve events or quantities for which empirical data is either unavailable or too expensive to collect, making the frequentist view of probability impractical for the construction of useful expert systems. However, when one takes the subjectivist view of probability, expert opinion and knowledge, built up over years of experience, can be used to help create useful systems that can be applicable even when there is little data available [1, p.66]. There are also techniques to combine data, when it becomes available, with judgement to revise these subjective probabilities and refine a knowledge base.

2.4.1.1 Definitions

A few preliminary definitions will be useful for subsequent discussion. These definitions are taken from [41].

Proposition: a sentence that can take on the values true or false

Elementary proposition: a proposition where one random variable takes on a value. E.g. Weather = sunny. Elementary propositions can be combined using standard logical connectives to form complex propositions.

Random variable: the basic element of probability language, which refers to a part of the world, whose value is initially unknown. There are three kinds of random variables:

Boolean random variables: random variables that have the domain <true, false>. An example of this in Figure 2.3 is the variable "High Blood Pressure", which can take on the values of true or false.

Discrete random variables: random variables that take on values from a countable domain. These values must be mutually exclusive and exhaustive. An example of this in Figure 2.3 is Gender, whose states can be <male, female>.

Continuous random variables: random variables that take on values from real numbers. Examples of this in Figure 2.3 include Age, Snoring, AHI, BMI and ESS. These are often measurements of some kind.

Domain: a set of values that a given random variable can take on. For example the domain of gender is {male, female}.

Atomic Event: a complete specification of the state of the world, an assignment of particular values to all the random variables of which the world is composed. Atomic events are mutually exclusive and exhaustive.

Prior/ unconditional probability: the degree of belief accorded to a proposition in the absence of any other information. E.g. P(a)

Posterior/ Conditional probability: the degree of belief accorded to a proposition given a certain piece of information. E.g. $P(a|b)$ means the probability of a, given that all we know is b.

If $P(a \mid b) = P(a)$, we say that propositions a and b are *absolutely independent.* [42]. This can also be expressed as $P(a,b) = P(a)P(b)$ and $P(b \mid a) = P(b)$.

If $P(a \mid b,c) = P(a \mid c)$, we say that a and b are *conditionally independent given* c [42], meaning that once c is known, learning b does not change our belief in a.

Probability distribution: a vector of values for the probabilities of each individual state/ value of a given random variable. For continuous random variables, these are instead referred to as probability density functions because these variables can take on infinitely many values. Instead, the probability that a random variable takes on some value x is usually a parameterised function of x.

Joint Probability Distribution: a vector denoting the probabilities of all possible combinations of values of a set of random variables. When the set of random variables is the complete set used to describe the world, this is referred to as the full joint probability distribution.

2.4.1.2 Notation

Capital letters will be used for random variable names (X, Y, Z, ...) and lowercase letters (x, y, z, ...) will be used for specific values taken by the variables X, Y, Z, ... respectively. Propositions (X = x) will be represented by lowercase italicised letters (*a, b,c,* ...). When discussing sets of variables, capital, italicised letters will be used (*X, Y, Z,* ...)

2.4.1.3 Rules and Axioms of Probability

There are three basic axioms of probability calculus. These are often called Kolmogorov's axioms.

I. All probabilities are between 0 and 1

II. Necessarily true propositions have probability 1. Necessarily false propositions have probability 0

III. P(*a* OR *b*) = P(*a*) + P(*b*) – P(*a, b*) where P(*a, b*) is short for P(*a* AND *b*)

[42], [41]

Assuming that all variables have finitely many values (are not continuous), other rules can be derived from these axioms. For example, any probability distribution on a single random variable must sum to 1. This follows from axiom II). This can be extended to saying that any joint probability distribution on any set of variables must sum to 1 [41].

Also, axiom II) and III) can be used together with the knowledge that any proposition is equivalent to the disjunction of all the atomic events (which are mutually exclusive) in which it holds (referred to as e(*a*)) to derive the following relationship: The

probability of a proposition is equal to the sum of the probabilities of the atomic events in which it holds [41]. What this means is that if we have a full joint distribution specifying the probabilities of all atomic events, it is possible to compute the probability of any individual proposition within the joint distribution.

Another useful and common calculation is, given a full joint distribution, to calculate the probability distribution of a single variable or a subset of variables. This is called the **marginal** or **unconditional probability** of a variable (or subset of variables). For any sets Y and Z, the marginal probability of set Y is given as:

$\mathbf{P}(Y) = \Sigma_z \mathbf{P}(Y,z)$ [41], [42]

When one has conditional probabilities instead of a full joint probability distribution, **P**(Y) can be calculated using a rule called **conditioning** which is applied in conjunction with the product rule governing conditional probabilities.

The **product rule** is given as: $P(a|b) = P(a \text{ AND } b) / P(b)$, which can also be written as:

$$P(a \text{ AND } b) = P(a|b)P(b) = P(b \mid a)\, P(a)$$

The product rule can also be applied to distributions. **P**(X | Y) is defined as the probability distribution representing the set of probabilities $P(X = x_i \mid Y = Y_j)$ for each possible i, j and is represented as: **P**(X , Y) = **P**(X | Y) **P**(Y).

Conditioning is defined as: $\mathbf{P(Y)} = \Sigma_z \mathbf{P(Y \mid z)}\, P(z)$

2.4.2 Bayes' Theorem

In many cases, expert systems are interested in computing the probability of seeing a specific value of some random variables, given evidence about others. This is

referred to as computing the conditional or posterior probability of proposition *a* given *b* denoted P(*a*|*b*) and is defined above in section 2.4.1.1.

As we saw previously, the product rule is defined in terms of conditional probabilities. Bayes' theorem is derived from the product rule and given by the equation:

$$P(b \mid a) = \frac{P(a \mid b)\, P(b)}{P(a)} \qquad [41]$$

Bayes rule is useful because it provides a method of computing a posterior probability of a certain proposition based upon its prior probability and the conditional probability of seeing certain evidence given the proposition is true; a way for updating beliefs in response to evidence [42].

It is also possible to use Bayes' theorem to derive the probability of a proposition given two or more pieces of evidence. However, this requires, for n pieces of evidence, knowing the conditional probabilities for 2^n possible combinations of observed values, which quickly becomes infeasible. What is needed are ways of simplifying expressions. Independence allows such simplification.

If proposition *a* is absolutely independent of propositions *b,c* and *d* then the joint probability distribution is given by: $\mathbf{P}(a, b,c,d) = \mathbf{P}(a)\mathbf{P}(b,c,d)$. That is, the full joint probability distribution can be factored into two smaller distributions. Unfortunately, many random variables are not absolutely independent. However, many random variables are conditionally independent of each other, given another random variable (see section 2.4.1.1 for a definition). This is given by the equation:

$$\mathbf{P}(a,b \mid c) = \mathbf{P}(a \mid c)\, \mathbf{P}(b \mid c)$$

In this equation, both a and b are affected by c, but neither has a direct effect on the other. That is to say, that the probability distribution governing a is independent of the value of b, given a value for c [3]. Conditional independence is useful because it simplifies probability calculations. The information required to compute the probability of posterior probability of proposition c given two pieces of evidence is the same as that required to compute the posterior probability of c for each piece of evidence individually if those two pieces of evidence are conditionally independent given c [41]. This is given by the formula:

$$\mathbf{P}(c \mid a,b) = \frac{\mathbf{P}(a \mid c)\,\mathbf{P}(b \mid c)\,\mathbf{P}(c)}{\mathbf{P}(a,b)}$$

Conditional independence also allows a decomposition of the full joint distribution into smaller components. For n pieces of evidence all conditionally independent given a certain proposition, the size of the representation of the full joint distribution does grow, but only linearly as opposed to exponentially as we had before. "Conditional independence assertions can allow probabilistic systems to scale up; moreover, they are much more commonly available than absolute independence assertions" [41]. This is especially useful when trying to model medical knowledge because it is often based on the notion that a symptom is a stable characteristic of a given disease and is fairly independent of other factors [42]. Also, combinations of disease do not occur very often [43]. If one assumes that diseases are mutually exclusive, it is possible to model different diseases as different values of a single disease variable or node in the network. If, on the other hand, one wishes to allow for multiple diseases to be present at the same time, and possibly sharing the same symptoms, one can have one variable or node in the network for each disease. In our network, Figure 2.3, the nodes

PLM (Periodic Limb Movement), AHI (a stand-in for the OSAS diagnosis) and Diabetes, all represent different pathologies, and each one affects the ESS node, which represents subject sleepiness in the network.

2.4.3 Bayesian Networks

2.4.3.1 Introduction

A Bayesian network (also known as a belief network or a probabilistic network) is a data structure used to give a compact graphical representation of the full joint probability distribution of a set of random variables. It is a directed graph where [44]

- a node represents a random variable, either discrete or continuous
- a set of directed arrows connects pairs of nodes. If there is an arrow from node X to node Y, X is called the parent of Y.
- each node Xi has a conditional probability distribution $P(X_i \mid \text{Parents } (X_i))$ that quantifies the effect of its parents on it. For discrete variables, this is represented as a conditional probability table, where each row contains the conditional probability of each node value for a possible combination of values of its parent nodes.
- The graph is directed and acylic.

These graphs play a key role in the decomposition of large probability distribution functions because they provide a visual representation of the sets of random variables that are relevant to each other in any given state of knowledge [42]. Bayesian networks allow conditional independence statements that apply to subsets of variables, as opposed to all variables [3]. The topology of the Bayesian network specifies the conditional independence relationships that hold within that world. Combined with a conditional probability distribution for each child node given its parents and prior probability

distributions for source variables (nodes with no parents), the topology of the Bayesian network is sufficient to specify the full joint probability distribution for all of its component variables [44].

Given a distribution P defined on n discrete variables X_1, X_2, ... Xn, the probability of a conjunction of particular assignments to each variable $P(x_1, \ldots x_n)$ is given by:

$$P(x_1, \ldots x_n) = \Pi_{\text{from } i = 1 \text{ to } n} P(x_i \mid \text{parents}(X_i))$$

where parents(X_i) denotes specific values of the variables in Parents(X_i). This implies that each entry in the full joint probability distribution can be calculated by the product of the appropriate elements of the conditional probability tables in the Bayesian network, and thus that the Bayesian network can answer any query about the given domain [44].

One of the main advantages of Bayesian networks is that they are often much more compact than the full joint distribution. If a network contains n Boolean variables and each variable can be influenced by at most k other nodes, the amount of information needed to specify each conditional probability table for each node is at most 2^k and the complete network can be specified using $n2^k$ numbers. By contrast, a full joint probability distribution contains 2^n numbers. However, it is often possible to reduce the numbers needed to specify a Bayesian network even further. Deterministic nodes (nodes whose value is exactly specified by the values of their parents) often require no conditional probability tables because their values can directly be calculated from their parents' values using a formula. Noisy-OR logical relationships can also be used to reduce the size of the conditional probability table of a random variable which depends

on k parents from 2^k numbers to k numbers [44]. A noisy-OR represents the situation where each boolean parent of a boolean node has some probability of being sufficient to cause the child node to be true, and the event of a given parent P_i being true is independent from the event of each other parent P_j being true. The noisy-OR relationship is often used to represent causal relationships such as those where several different diseases can each cause a common symptom [1, p.76]. When using the noisy-OR relationship, it is often useful to introduce a *leak* node which can be used to represent 'all other unknown causes'. It is used to encode the probability that a given effect/ symptom can occur in the absence of any cause explicitly represented as a random variable in the Bayesian network topology [1, p.76].

Bayesian networks allow the decomposition of complex subjective judgments into simpler subjective judgments about the probabilities of component events. The components of the model are then reassembled and the Bayesian network used to infer probabilities implied by these simpler judgments in order to facilitate the making of complex subjective judgments [1, p.67]. In medicine these systems are used for aiding in diagnosis: inferring the most probable cause of an observed problem given a set of symptoms, patient history, physical signs and test results. They are especially useful in the medical domain because they allow the creation of a probabilistic network using expert knowledge of causal dependencies in a given domain but can then be used for diagnostic inference (predicting probabilities in the reverse direction from effect to cause). Bayesian networks also easily support intercausal inference (when the increased belief in one possible cause of an observed effect decreases the belief in another possible cause of the same effect) [1, p.71].

2.4.3.2 Constructing Bayesian Network Topology

When building the topology of a causal Bayesian network (one whose links represent represent causation between variables), it is important to add nodes in the correct order; nodes representing 'root causes' (diseases in the medical domain) are input first, then the variables those 'root cause' nodes influence are added and so on until leaf nodes are added (nodes which have no direct causal influence on any other variables in the network) [44]. If instead, a diagnostic model is built with links from symptoms to causes/ diseases, the resulting network not only has far more links (dependencies), but often the numbers representing the conditional probability tables for these additional links will also be more difficult to obtain.

Properly constructed causal Bayesian networks satisfy the following specifications:

- a node is conditionally independent of its predecessors, given its parents
- a node is conditionally independent of its non-descendants, given its parents
- a node is conditionally independent of all other nodes in the network, given its Markov blanket (this consists of a node's parents, a node's children and the node's childrens' parents) [44]

Another equivalent criterion for a properly constructed Bayesian network is called d-separation and is discussed in more detail in [42].

2.4.4 Learning with Bayesian Networks

There are four major types of problems that Bayesian networks are often used for. The first is to infer the probabilities of seeing specific values of some target variable(s)

given the known values of the random variables in the network. In this situation, the network topology, as well as prior and conditional probabilities are all specified. The three other types of problems involve learning Bayesian networks from data. Once this is successfully accomplished, the resulting Bayesian network can then be used to answer problems of the first type. In the simplest situation, a network topology is given in advance and prior probabilities as well as conditional probabilities have to be inferred from a set of training data. The second is to infer the probabilities of seeing specific values of some target variable(s) in networks with hidden variables. These are variables that are not directly observable in the data that is available for learning or from expert opinion as a prior probability. These hidden variables often represent, in the medical domain, the disease itself; while it isn't directly measurable, it is affected by other random variables in the network and can affect other random variables in the network. The third type of problem involves learning the actual structure of the Bayesian network from data.

2.4.4.1 Inference in completely specified Bayesian networks

This represents the situation where the Bayesian network is used to infer the value of some target variable(s) based on some evidence (a set of observed values for some other variables in the network). This task involves the probability that a random variable will take on each of its possible values given the observed values of other variables [3].

This is very straightforward when dealing with a full joint probability distribution. If X is the query variable, **E** the set of evidence variables with **e** the observed values for them and **Y** the set of unobserved variables, the probability distribution for X is given by:

$$\mathbf{P}(X \mid \mathbf{e}) = \frac{\mathbf{P}(X, \mathbf{e})}{} = \frac{1}{} \text{ SUM}_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}) \qquad [41]$$

P(e) **P(e)**

However, as mentioned previously, this requires an input table of size $O(2^n)$ and takes $O(2^n)$ time. Bayesian networks can be used instead to infer the probabilities.

Exact inference can be performed using a Bayesian network by computing the sums of products of conditional probabilities of the nodes in the network [44]. This has been shown to be NP hard [45]. Methods for exact inference include the variable elimination algorithm and join tree algorithms (a form of clustering algorithms) [44].

Alternatively, approximation methods can be used in an attempt to be more efficient. Randomized sampling algorithms, also called Monte Carlo algorithms generate a random sample of network instantiations and estimate probabilities from this sample [1, p.80]. Another class of approximate inference algorithms uses a heuristic search to find hypothesis that best explain observed findings [1, p.80].

2.4.4.2 Learning parameters with complete data

This problem involves inferring prior probabilities as well as conditional probabilities from a set of training data when a network topology is given in advance. One standard approach is to do maximum likelihood parameter learning. This involves deriving an expression for the likelihood of the data as a function of the unknown parameters of the network (the probabilities). Then, the parameter values are those for which derivative of the log likelihood with respect to each parameter is zero. Essentially there is a separate learning problem for each parameter of the network [46]. However, the problem with this method is that when the data set is small and certain events have not yet been observed, this method assigns a zero probability to those events [46].

Another approach is to estimate the conditional probability table entries using a naïve Bayes classifier. In this model, the parameter to be predicted is the class variable C and is the root. Attribute variables are the leaves and are assumed to be conditionally independent of each other, given the class variable. With these assumptions, the model is trained using maximum likelihood parameter values as described above. Once this is accomplished, the model can be used to classify new examples for which the class variable C is unobserved. This naïve Bayes classifier works fairly well, scales well to large problems and deals easily with noisy data [46].

2.4.4.3 Learning in the presence of hidden variables

Many real problems are faced with the situation where only a subset of a given Bayesian networks' variables are observable from data. This is very common in the medical domain. While symptoms observed are included in the data, there is often no direct observation of the disease itself. One could attempt to construct a Bayesian network that omits such variables, however, this often dramatically increases the number of parameters required to specify the network as well as the amount of data required to learn the parameters [46]. These unobserved variables can be referred to as hidden variables, latent variables or simply unobserved variables. The problem of fully specifying a Bayesian network now involves learning the conditional probabilities of the hidden variable given its parent values and learning the conditional probabilities of the hidden variable's child nodes given its values. The EM algorithm (short for expectation-maximization) can be used to train Bayesian networks in the presence of hidden variables. In this situation, although their values are not observed (not present in the

training data), the EM algorithm is told that they do exist and must find a place for them in the network. The basic idea of the EM involves two steps [3], [46]:

The E-step (expectation step) involves pretending that the parameters of the model are known (i.e. take a hypothesis representing a completely specified Bayesian network) and using this hypothesis to estimate the hidden variables (computing their 'expected values'). Because we know the structure of the network (we have accepted the completely specified network hypothesis as being true for this step), these probabilities can be computed by any inference algorithm for Bayesian networks.

The M-step (maximization step) uses the expected values for the hidden variables just estimated in the E-step to find an improved hypothesis; i.e. to find a new network model (topology and parameters) that maximize the log likelihood of the data that is available, given the expected values that have been computed for the hidden variables.

The E and M steps constitute a loop that is iterated over. Each E step finds expected values for the hidden variables that better fit the data than the iteration previous. Each M step involves finding an improved network (topology and parameters) that better fits the data than the iteration before [46]. Under certain circumstances, the algorithm has been shown to converge to a local maximum likelihood hypothesis [3].

2.4.4.4 Learning the topology of Bayesian networks from data

In many situations, a causal model for a given domain is either not available or disputed. In such circumstances, it is of use to be able to learn a Bayesian network from data when the network structure is not given. Some algorithms involve heuristic searches for a good model. These algorithms often follow one of two strategies. This first

involves starting with a model containing no links and begin adding parents for each node, fitting parameters and measuring the accuracy of the resulting model. The second strategy involves making an initial guess as to the network topology and then using simulated annealing search to make modifications to the topology, retuning the network parameters after each change. Such searching algorithms are often fed an initial ordering of variables as input [3, 46]. Constraint-based approaches also exist which infer independence and dependence relationships from data and then use these learned relationships to construct a network topology [3].

Once network topologies have been created it is important to be able to judge when a good network structure has been found. One method involves testing whether the conditional independence assertions implicit in the network structure are satisfied in the data [46]. A second method evaluates the degree to which the proposed topologies explain the data. However, both these methods need to balance this accuracy over training data with network complexity, otherwise the resulting network will contain far too many connections and be impractical to use. As a result, model complexity is penalized in the scoring functions that determine the best of the proposed network topologies [3] [42].

2.4.5 Bayesian Networks in Medicine

Bayesian networks have been used in medicine for over a decade[47], [48], [49], [50], [51] , [52]. They are particularly well suited to medicine as it is a domain in which there is still an incomplete understanding of many of the processes at work in the progress of disease. Acquiring an understanding of these processes is rendered difficult by the fact that their characteristics vary widely, often only a fraction of the factors that

affect them can be observed and that they are subject to individual and random variation. In short, it is a domain full of uncertainty [43].

The types of tasks Bayesian networks are used for in biomedicine and health-care include diagnostic reasoning, prognostic reasoning, treatment selection and discovering functional interactions in the underlying physiological processes themselves [43]. Early medical diagnostic systems were constructed based on Bayesian networks constructed using two simplifying assumptions; first that hypotheses (diseases) were mutually exclusive and collectively exhaustive and the second that individual pieces of evidence (symptoms or test results) were conditionally independent of each other given a particular diagnosis. These included systems for diagnosing heart disease and acute abdominal pain [53]. Even in spite of these simplifications such systems often outperformed experts in terms of diagnosis [54]. More recent examples include the PATHFINDER project, a diagnostic system for lymph node pathology [48]. It became one of the first commercially successful expert systems for medical diagnosis and was called INTELLIPATH [1, p.84]. Its creators explored a variety of rule-based and non-probabilistic schemes before they settled on using a Bayesian probabilistic scheme, which they found was noticeably better than other schemes. Another system created in the late 1980's is called MUNIN and is a belief network for the diagnosis of neuromuscular disorders [1, p.86].

More recently, Bayesian networks have been used to calculate a prognosis of patients with severe bacterial or fungal infections, dependent on the choice of antibiotics [50]. In addition, Bayesian networks have been used to assist in the determination of patient-specific therapy selection for patients with oesophageal cancer. The system

models the presentation characteristics of the tumor, pathophysiological processes underlying its invasion into the oesophageal wall and its metastasis. It also includes characteristics of diagnostic tests and possible effects of different treatment therapies. The network is then used not only to predict the most likely stage of the patient's cancer, but also to asses the most likely outcomes to the different treatment therapies [52]. Another network has been developed to assist intensive care unit clinicians in diagnosing and selecting treatment for patients with pneumonia in intensive-care units [51].

Many of the afore-mentioned networks were constructed manually using human experts. Because they encode a career's worth of expertise, such networks can be quite accurate because the knowledge encoded in them "is more robust than the knowledge embedded in a data set of limited size" [43]. In addition, they can be created, even in the absence of data sets that specifically contain all the data required by the network. The downside to manually constructed Bayesian networks is that they are quite time consuming to build and don't use to its full advantage the vast amounts of clinical and biological data that is currently accessible to the scientific community. A more recent trend involves using Bayesian learning methods to estimate network topologies or structures from sets of data; for example, to construct models of metabolic and physiological processes using metabolic data [55]. Other examples include discovering gene interactions based on microarray expression data [56]. Another, very recently published paper looks at applying information retrieval techniques for obtaining prior probability information from World Wide Web to be used for learning Bayesian networks when available clinical data sets are too small to be exploited for learning. This particular paper used these techniques to construct a Bayesian network for the

classification of ovarian tumors in patients [57]. While learning both network topology and parameters from data is extremely attractive because it can illuminate variable interactions and connections previously undiscovered, it can also be problematic. The learning algorithms require a fairly large amount of data in order to reliably determine the probabilistic relationships between the networks variables and most learning algorithms assume that there are no missing values in the data set [43]. This is usually not the case in medical data sets. As a result, missing values have to be filled in, possibly with the help of a domain expert or by substituting a missing value with an average value for that variable. These can change the relationships that are present in the data set before it is altered to accommodate missing data. In the middle ground between completely manually constructed networks, and completely automated learning of networks is the construction of Bayesian networks using a mixed methodology. This can involve knowledge engineering (either with a subject matter expert or with literature review) a network structure or topology (creating it manually) and then, once a network topology is specified, it is possible to learn the parameters of the network from a real data set, even in the presence of missing data (if using the EM algorithm mentioned above in Section 2.4.4.3). This mixed methodology benefits from many of the advantages of both methods as it incorporates knowledge both from an expert and from existing data sets.

2.4.6 Bayesian Networks in Sleep Medicine

In spite of the wealth of active research using Bayesian networks in medicine in general, there is a paucity of such research in the domain of sleep medicine. The literature that does exist, focuses primarily upon the automated analysis of various PSG signals using Bayesian approaches in order to facilitate either sleep staging or diagnosis.

Only one article found allowed for the differential diagnosis of different sleep disorders from a set of symptoms that were not primarily based on PSG data [58].

The SIESTA project is a European project that is endeavouring to build an automatic classification system for sleep analysis. One subproject involves the automatic detection of sleep spindles in a EEG signal. These are defined as 0.5 to 2.0 second bursts of activity in the 12 – 16 Hz range. Using the EEG channels F4, C4 and P4, a spindle detector was implemented using a Bayesian approach, trained on a sample database. When used to detect sleep spindles, the spindle detector seemed to find many false positive spindles (segments of EEG not scored as spindles by an expert). However, when the signals were manually reviewed by an expert, the review showed that in many cases, these false positive segments were very similar to spindles, even though they hadn't been scored as such [59]. Another group within the SIESTA predicts the probability that a subject is either awake, in deep sleep (stage 4) or in rapid eye movement (REM) sleep using features extracted from 6 EEG channels as inputs, at a temporal resolution of one second. Their analyser consists of 3 major building blocks: pre-processing stage, a classification stage and a sensor fusion stage. Bayesian techniques were used in each of the blocks. The probabilities obtained by the sleep analyser showed less aging effects than corresponding manual Rechtschaffen and Kales scoring [60] as done by three experts. However, it is known that this is a known problem of the Rechtschaffen and Kales scoring rules [61]. Another research study used noseflow, diaphragm and thoracic signals from overnight PSGs to classify obstructive apnea, central apnea, paradoxical respiration and normal respiration events. Events were classified by combining neural networks (Multi-layer perceptrons and Kohonen networks) with classical Bayes theory.

The neural networks were used to estimate prior probabilities and the Bayes classifier to calculate posterior probabilities given these priors and the observed signal patterns. Using PSG signals from 3 patients, experiments to detect apneas and separate them from artifacts were promising in spite of the small set of data used [62].

Another study used Bayesian networks to perform sleep stage classification using features derived from EEG (sleep spindles, K complexes, and episodes of α, β, δ, σ and θ activity) and AOG (eyes) as input. The system used three different Bayesian networks. The first Bayesian model was used to detect sleep spindles in 2 second segments of EEG signal. The presence of absence of a sleep spindle was based upon 6 child nodes representing six features (σ activity and EEG power in three successive intervals). A second model, of identical structure was used to detect K-complexes (obviously parameters were different). Lastly, a network model was used to calculate the probability of a patient being in a given sleep state in a given 30 second window, given values in 9 child nodes. Nodes representing α, β, δ, $\sigma$and θ activity and features were quantified as relative values between 0 and 1. Nodes representing sleep spindles and K-complexes could take on two possible states (yes/no) and their values were determined from their corresponding Bayesian networks. The networks were trained using a learning set of data created from expert scoring and the corresponding values of the nodes. Results found up to 70.7% agreement between the system and two experts, which is quite high given that inter-expert agreement was 71.4% over six subjects [63].

Only one study was found that used Bayesian networks for the diagnosis of sleep disorders using Bayesian networks[58]. This study focussed more on the development of a Bayesian network-based development tool (software system) specifically aimed at

creating systems for medical diagnosis. This abstracts the concepts of Bayesian network by presenting variables as diseases, symptoms and test results and facilitates domain knowledge introduction using a graphical interface. A web-page can be used to interface with a constructed system in order to perform diagnosis using the constructed Bayesian Network. The example of such a system developed with their software tool was given as the Sleep-Disorders Diagnostic System (SDDS). It is a system for the differential diagnosis of four sleep disorders (Psychophysiological Insomnia, Idiopathic Insomnia, Obstructive Sleep Apnea and Narcolepsy) based on a set of symptoms observed in subjects. Most of these symptoms are not taken from overnight polysomnograph signals. The prior and conditional probabilities used as parameters in the system were specified using medical literature and from consulting with experts. However, as the sleep-medicine application of the development tool was not the focus of this article, there is no mention of how the system performs for diagnosis. No results are presented or discussed in the paper [64]. Because of this, there is no data or even descriptions with which we can compare the performance of the networks we develop in this thesis.

CHAPTER 3:MATERIALS AND METHODS

3.1 Bayesian Network Software

Belief network development requires the use of software that is capable of probabilistic inference. Several such software systems currently exist, including commercial products such as HUGIN, Netica™, Baron and Ergo. Some software is available free of charge to researchers including BAYES, BELIEF, and TETRAD.

Our data set contains many missing values. If one discounts all records with a missing value for any one variable, there are 43% records which could not be used. Many software packages that implement Bayesian networks stipulate that there cannot be any missing values. Our data set is only 652 records to begin with and cutting this almost in half would result in a data set that was very small. Instead, we chose to find a software package which would allow us to retain records that contain missing values for one or more variables. The software used for this research was Netica™ (version 3.05). Netica™'s main advantages for our research include the fact that it implements the Expectation Maximization (EM) algorithm for parameter estimation [46]. This algorithm allows for the presence of missing values, has a very user-friendly interface, generates presentation quality graphics and has functions for easily performing statistical tests and sensitivity measurements on belief networks using test data. However, while it handles missing values, Netica™ does not do network structure learning. Therefore, we need to supply a network topology to the software in order for it to determine network parameters from the training data provided.

All parameter learning done by our networks from the training data is done with EM learning selected as the mode of learning. This is due to the high level of missing data. We have chosen to leave missing values as unknown values and to use the EM algorithm to handle this situation.

3.2 Data

The data used to train our Bayesian network was gathered from three sources: two from existing databases, and one (pupillometry) from data we gathered and analysed.

3.2.1 Data from Existing Databases

3.2.1.1 Vancouver Coastal Health Subjective Questionnaire

The first source of data was gathered on six hundred and fifty-two (652) consecutive patients referred to the Sleep Disorders Clinic, at the University of British Columbia (UBC) hospital. These patients were referred to the clinic for a standard in-clinic overnight polysomnography (PSG) for the assessment of suspected OSAS between May 2003 and March 2005. At the time of their initial visit, having given informed consent, these patients completed a subjective questionnaire for the Vancouver Coastal Health Authority containing questions relating to demographics, sleep habits, sleep-related symptoms, health habits, occupation, medical history, mood, occupational accidents, driving and driving accidents, and an ESS questionnaire. Overall there are 57 questions in this questionnaire.

3.2.1.2 Overnight polysomnography (PSG) report

The second source of data is a report of the results of an overnight PSG study done at the Vancouver Sleep Disorders Clinic at UBC Hospital. These patients are the

same as those who filled out the subjective sleep questionnaire mentioned above. However, of the 653 patients, only five hundred and twenty-one (521) patients have digitized records of the overnight PSG.

3.2.2 Pupillometry Data

The third source of data consists of forty-four (44) patients who were at the Vancouver Sleep Disorders clinic at UBC hospital for an overnight PSG study, and who, in addition to filling out the subjective questionnaire, agreed to have their pupil size recorded during dark adaptation and the Pupillary Light Response (PLR) on the morning following the in-clinic overnight PSG study. All studies were conducted between the hours of 6:30 am and 8:00 am in the morning after their overnight stay at the clinic. The pupil size was recorded at 100 Hz using an Eyecheck™ pupillometer. For details on the Eyecheck™ pupillometer and how it is used, please refer to Appendix 1.

3.2.2.1 Dark adaptation

Pupil size was recorded at 100 Hz for 11 consecutive minutes while the patient sat in darkness, looking into the Eyecheck™ pupillometer at a set of red cross hairs. The device then averages every 10 data points and outputs to a text file with a data point every 10 Hz. The pupil size data gathered during dark adaptation is used to compute the Pupillary Unrest Index (PUI), a measure of pupil response during dark adaptation. An algorithm to compute PUI from pupil size data (during the dark adaptation phase) was coded in C++ by this author, in accordance with the criteria established by [37]. The code can be found in Appendix 2. When run on dark adaptation pupil size data, the code computes average PUI during the first 4 minutes of the dark adaptation period.

3.2.2.2 Pupillary light response (PLR)

Pupil dynamics in response to a flash of light are recorded at 100 Hz by the Eyecheck™ pupillometer. This data is used to internally (within the operating software of the device) to calculate several metrics that are output to a text file. Those of interest to us include Time To Minimum (TTM), the time it takes for the pupil to constrict maximally in response to the light, Time To Initiation of PLR response(TTI) and the Reflex Amplitude Percentage(RA%), the difference, as a percentage between the pupil size before the PLR and after five (5) seconds.

3.2.2.3 Limitations of pupillometry data

As mentioned above in section 3.2.2, forty-four subjects agreed to have their pupil size recorded the morning after their overnight study. Two of these did not have digitized questionnaires and overnight PSG reports available. As for the remaining 42 subjects, unfortunately, only fourteen subjects had useable data for the entire eleven minutes of dark adaptation. While we are not sure of the cause, in quite a few subjects, after a varying number of minutes (from 2 to 8), the pupillometer indicated that pupil size began increasing consistently up until a certain point and then indicated a special value (891) indicating the instrument could not determine the pupil size. This is thought to be in part due to the Eyecheck™ pupillometer, which according to the manufacturer, was not intended to be used for such long periods of dark adaptation. Another possibility is that some patients were extremely sleepy and could not keep their eyes open enough that the pupillometer could accurately measure their pupil size. We then categorized patients according to how long their dark adaptation pupil data was valid (the whole 11 minutes, at least 4 minutes, not long enough to be useful) and looked at the mean age, AHI and

ESS score in order to see if these measures offered any clues at to the reason the pupillometry did not work. The results of this analysis can be seen in Table 3.1. Other than the fact that mean ESS Score is slightly elevated in the category of patients for whom less than 4 minutes of useable dark adaptation pupil data was obtained, there seems to be no explanation from these variables as to a possible cause of pupil recording problems.

Table 3.1 Mean AHI, ESS, Age and Gender Organized by Length of Dark Adaptation period possible

	Mean AHI	Mean ESS Score	Mean Age	Gender
Full 11 minute ALT (14)	29.26(m) 16.32(f)	7.67(m) 11(f)	54.1(m) 51.8(f)	9 males 5 female
Only 4 minute ALT possible (17)	29.37(m) 29.88(f)	8.38(m) 7.67(f)	43(m) 44.75(f)	13 males 4 females
Pupil data collection not possible (11)	25.97(m - only 4 patients had data) 12.7(f – only 1 patient had data)	8.8(m – only 5 patients had data) 8.5(f – only 2 patients had data)	58(m) 46.3(f)	8 males 3 females

In order to use as many pupil size recordings as possible, we plotted the Average PUI on a per minute basis, from one up to the full eleven minutes on those data records that were valid for the whole period. In contrast to the studies by the Tubingen group [11], on average the second half of the dark adaptation period (minute 5.5 to 11) did not show higher average PUI values than those calculated for the first half of the dark adaptation period. Instead, we found that after four (4) minutes, the average PUI did not change significantly. As a result, our PUI is a measure of the average change in pupil size per minute over the first four (4) minutes of dark adaptation. This allowed us to use pupil data that could not have been used if we had calculated PUI over the full 11 minute ALT test. This allowed us to use thirty-one pupillometry recordings. However, for some

of these, their questionnaire and overnight PSG data was not available, leaving us with only 28 pupillometry recordings which could be incorporated into the data set.

3.3 Variable Selection

From the subjective questionnaire (data source one), out of the initial 57 questions, 18 variables were chosen based on literature review and the advice of our medical expert. Some of these variables are taken directly from the responses to the subjective questionnaire (data source one) while others are calculated from data fields present. From the results of the overnight PSG (data source two), four variables were chosen or calculated. From the pupil size data (data source three) four variables were taken or computed. Of these, only PUI was used in our networks for the purposes of this thesis. These variables can be seen in Figure 3.1. The details of each variable, its definition and corresponding states as well as its corresponding data source and how it was computed (if required) can be found in Appendix 3.

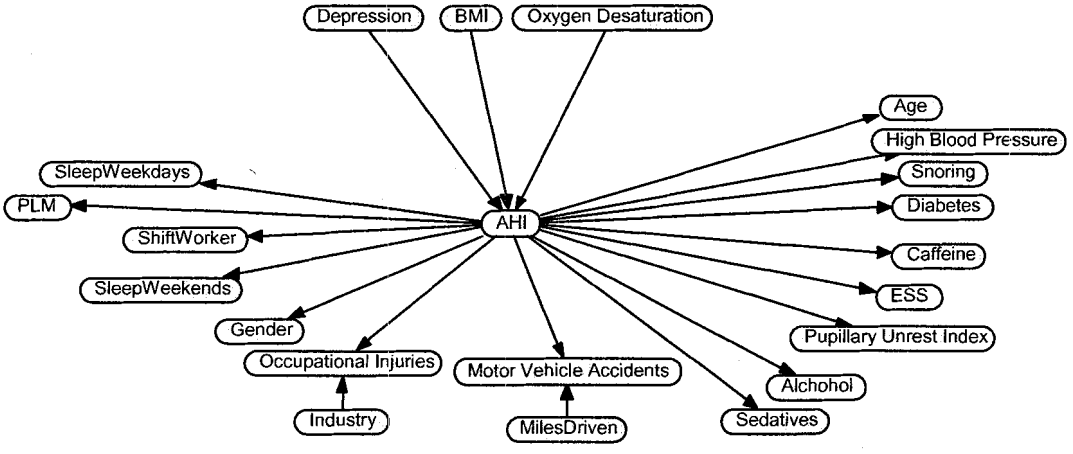

Figure 3.1 Variables in our OSAS Bayesian Network

3.4 Treatment of Missing values

In medical data incomplete patient records are a very common occurrence. This is also the case for our data sources. In our data there are two kinds of missing values: Source missing values and user-defined missing values. Source missing values occur when a value for a given patient/variable cannot be found or interpreted. These missing values can be the result of incomplete responses to questionnaires, problems with the digitisation of data sources or simply missing data sources for certain patients. User-defined missing values are responses to questions from data source where the patient selected a response such as "not sure" or "unknown". While it is meaningless to incorporate these missing values in the calculation of statistics, the data records containing such missing values should not simply be eliminated. This is due to the fact that the total amount of remaining data may not be sufficient and, on the other hand, the remaining values in the data record may still contain very useful information.

There are different strategies for dealing with missing data. One is to completely remove data records containing missing values for some variables. The second is to model missing values as a new value, namely "missing". This can permit researchers to discover if being missing, as a state, is related to other states in the network. However, this extra state makes each conditional probability table in the network more costly to calculate and more difficult to parameterize. Another strategy used can be to replace missing values with the mean. For the purposes of our belief network, all sources of missing values are treated equally and set to "*". This is a character code used by Netica^{TM} to indicate a data record for which a particular variable's value is not known. We then use the EM algorithm to handle this situation.

3.5 Database Creation

In order to create a data set from which the Bayesian belief networks training and test data could be created, a number of operations were necessary. First, the three separate data sources were merged on the basis of "study id"(SID), a unique identifier the Sleep Disorders Clinic uses to identify subjects. This merged database contained the responses to all questions from the subjective questionnaire and values filled in by the sleep clinic technicians and doctors after the in clinic overnight PSG, as well as the pupillometry variables calculated from the pupil size data for those patients for which this data was available. For certain SID numbers, not all data sources were available. In this situation, all values for the missing data source were marked as missing values for that data record by setting them to "*". Then, in order to maintain anonymity, SID was stripped from this merged database.

Then, from this database, numerous operations were performed to decode the fields in the existing database and convert them into the variables we wish to use in our Bayesian networks. Details of the operations performed for each variable can be found in Appendix 3.

3.6 Experimental Evaluation

In order to evaluate the performance of our networks, sets of training and test data are created from the complete final set of data. There are a number of ways of doing this. The total data is partitioned into n sets of equal size. We train on n-1 sets and make the remaining set a test set, giving us n experiments. If n is too small, we run the risk that a given run is not representative. If n is too large, this may lead to a large correlation among the training sets for each run. We used n=10 for a 10 fold cross validation [65],

[66]. However, we encountered problems with getting Netica™ to properly process our eighth test set. We were unable to resolve these problems. We therefore use nine training and test sets for computing our performance measures.

3.6.1 Experimental Evaluation with Pupillometry Data

The training and test sets for networks including the pupillometry variable PUI are slightly different than the ones used for other networks. In order to reduce variance, we modified the original test and training sets in order to create 10 stratified samples. This way each test set contains a unique set of 3 data records that include PUI data and each training set of data contained the remaining 25 records with PUI data.

3.6.2 Network Performance

In our investigation, the ultimate goal of our Bayesian Belief network is the prediction of AHI level in patients, as a stand-in for a positive diagnosis of OSAS. As a result, we refer to the AHI node as our query node. While AHI is a continuous variable, it is converted to an ordinal variable with values of Low or None and Moderate to Severe. The way in which the AHI variable is discretized is the same in all network configurations and is based on literature review.

No or Low OSA: 0 to 15

Moderate to Severe OSA: 15 and over

A given network's performance is judged in the following way. After having learned the network parameters from the training data set using EM learning, Netica™ reads through each test case, except for any findings for unobserved nodes (the nodes for which we want to predict the value). Netica™ does belief updating based on the

observed values it has been provided with to generate beliefs for the unobserved nodes. It then compares these beliefs to the true values supplied for the unobserved nodes. The performance of the networks in predicting the value of the AHI query node given the findings at other nodes in the network is evaluated for each of the ten (10) test sets and then averaged in order to provide a performance value that accurately reflects the overall predictive ability of a given network. The following statistics are calculated from a confusion matrix which Netica™ reports:

3.6.2.1 Sensitivity

Sensitivity is defined as the fraction of those with the disease that are correctly identified has having moderate to severe OSAS.

3.6.2.2 Specificity

Specificity is defined as the fraction of those without the disease that are correctly identified as have no or low OSAS.

3.6.2.3 Positive predictive value

Positive predictive value is defined as the fraction of people identified as having moderate to severe OSAS that actually have moderate to severe OSAS.

3.7 Investigation of the Effect of Network Topology on a Network's Predictive Ability

A Bayesian Belief Network (BBN) captures relationships that are believed to exist but may be uncertain. Once the set of key variables has been determined, the next

step in building a BBN is to attempt to map out the causal relationships between them. Belief network learning is usually divided into two parts: structure learning and parameter learning. Structure learning determines the causal relationships between variables, that is, the placement and direction of links in the network from a set of data provided. Parameter learning takes an existing structure of nodes and links between them and determines the conditional probability relationship at each node given the data. Netica™ currently only supports parameter learning. As such, a network topology needs to be provided. As part of our research exploration, several different topologies are investigated to see how they perform compared to one another.

3.7.1 Star Topology

As a preliminary investigation, a star topology of the Bayesian network is created. This consists of our query node, AHI, being the root of the graph with links pointing out of it to all other nodes, making it the parent node to all other nodes. In this configuration, other nodes are assumed to be topologically independent and conditionally independent given AHI. When trained in this configuration, Netica™ uses the naïve Bayes model to learn the parameters of the network. With these assumptions, the model is trained using maximum likelihood parameter learning. This network configuration is quick to learn from the training sets because for the AHI node, only prior probabilities are learned, while for the remaining nodes, their conditional probability tables are small, as they are assumed to be dependent only on the values taken on by the AHI query node. Another advantage of using a naïve Bayes classifier is that it works fairly well and deals quite easily with noisy data [46].

This network topology can be seen in Figure 3.2.

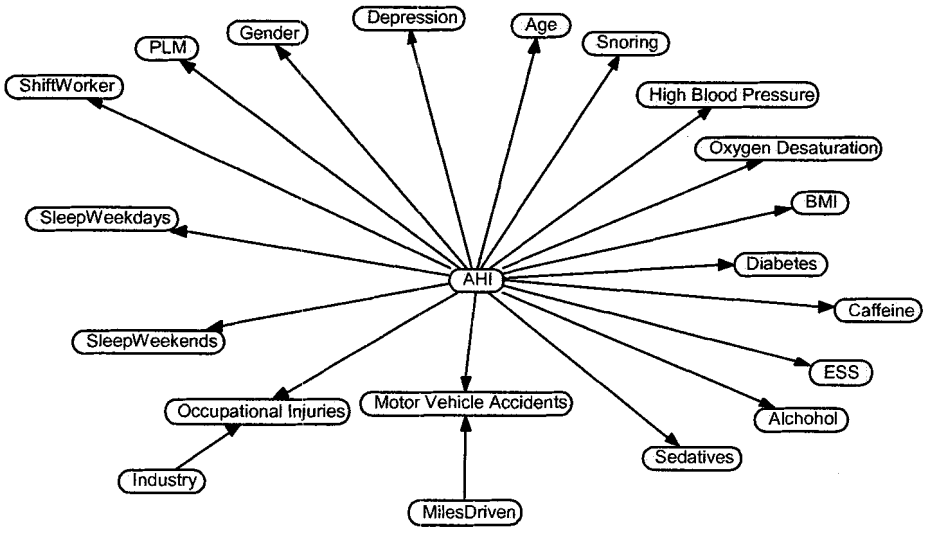

Figure 3.2 Star Topology Bayesian Network

3.7.2 3_In Network Topology

Netica™ provides a function which, on fully trained networks, allows the user to see how much the beliefs or expected values of a given query node are influenced by a single finding at another node in the network. Several sensitivity measures are provided for each of these other nodes (termed "findings node" in the Netica ™ manual) including the following:

1. For each possible state of the query node, the minimum, maximum and current posterior probabilities that each state of the query node can take on due to a finding at a given findings node F.

2. For each possible state of the query node, the square root of the expected change squared of the posterior probability of the query node being in state q, due to a finding at F.

3. Mean of Real Value: for query nodes representing continuous variables, the average expected value the query node would take on due to a finding at F.

4. Variance reduction: the expected reduction in variance of the expected real value of the query node due to a finding at F.

It is this last sensitivity measure on which we focus our attention, as it allows us rank which of the findings nodes will provide the most information about the predicted state of our query node.

For our purposes, the query node is the root node AHI. We use the "Sensitivity to Findings" feature that Netica^{TM} provides on our Star topology networks. Because the joint probability distribution of a trained Bayesian network is dependent on the data used to learn the parameters of the network, the list of findings nodes that most influence our AHI node changes with each training set of data. That being said, the majority of trained networks agree on the top three such finding nodes. This information is then used to create a new type of topology, which we refer to as the 3_In network topology.

In this topology, the AHI node is still the parent of most other network nodes, however, it is no longer the root of the network. It now has three parents, those nodes that ranked as the top three findings nodes in terms of their influence on the AHI query node when the network topology was a star formation. We investigate this network topology in order to explore the cumulative effect of linking the three nodes that most influenced our AHI node on the predictive ability of the networks. This allows us to study the interaction of these three nodes as a group with the AHI node, even though the

links are not representative of the causal relationships at play between the variables in the network. This network topology can be seen in Figure 3.3.

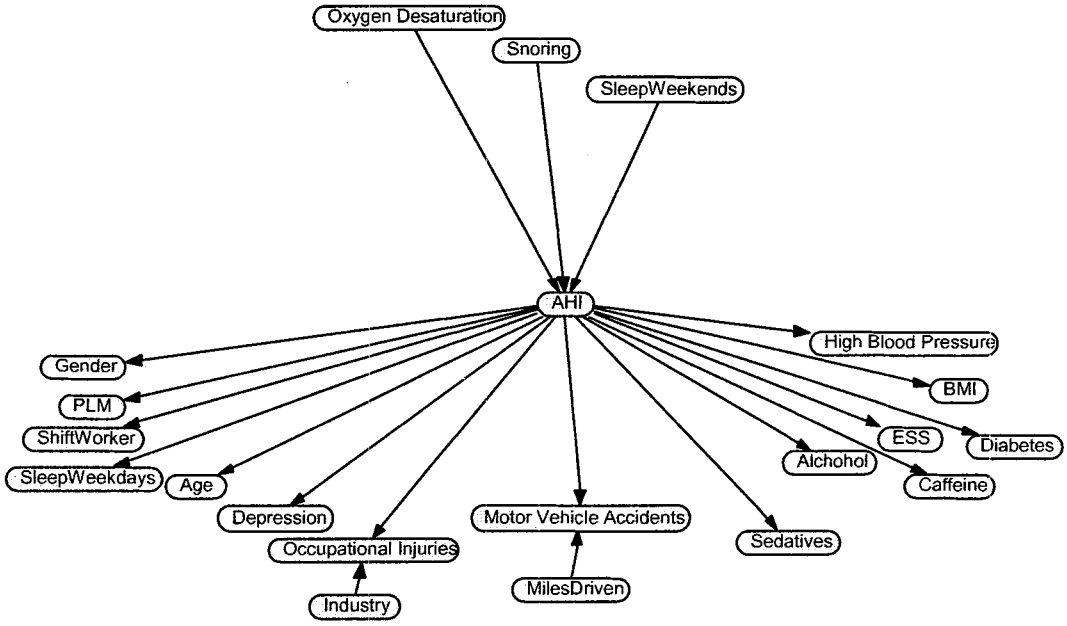

Figure 3.3 3_In Network Topology Bayesian Network

3.7.3 Knowledge Engineered Network Topology

The next network topology that was explored for its predictive abilities was one designed with the help of our medical expert. From the set of variables selected, he mapped out, using his expert opinion, a subset of the causal links between our network nodes. Too many links between nodes results in very long learning times, and large conditional probability tables at certain nodes, thus only the most important causal links were included in this network.

This network topology can be seen in Figure 3.4.

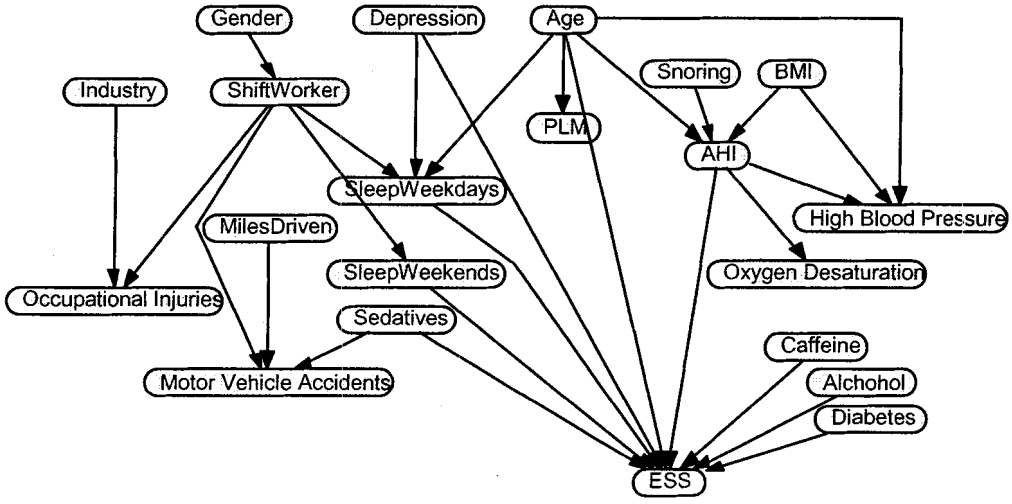

Figure 3.4 Knowledge Engineered Bayesian Network

3.8 Investigation of Discretization of Network Variables on Predictive Ability of Networks

Many data variables used in the study of OSAS are measurements of physiological states or events, or quantities/ frequencies and as such are considered continuous variables. In our data set these include the AHI, the Periodic Limb Movement Arousal Index (PLM), Percentage of time spent working the night shift (ShiftWorker), alcohol consumed per month, caffeine consumed per month, ESS Score, Body Mass Index, Oxygen Desaturation, Age and Miles Driven. We hypothesize that the way in which these variables are discretized (broken up into categories or ranges) for the purposes of creating conditional probability tables for their child nodes may have an effect on the predictive ability of a given Bayesian network. Therefore, the next stage of exploration involves using different strategies for discretizing these variables and observing the results on the performance of the various networks created in section (3) above.

3.8.1 Binning as a Strategy For Discretization of Continuous or Ordinal Variables

Netica™ provides a feature which allows for the easy discretization of such variables. This allows the user to specify how many bins into which the total range of a continuous variable should be divided. When a variable is binned, Netica™ divides the continuous variable into ranges or bins such that each bin will have a roughly equal number of records with a value falling in that particular bin's range.

We investigate breaking up these continuous variables into two, three and five bins of equal size and what effect this has on network performance. We hypothesize that three bins will provide the best performance as it theoretically provides enough separation between extreme cases and a middle ("average") zone for a given variable and yet will not result in a conditional probability table that is too large for that node's children, as is the case when a variable is divided into five bins. The ranges for each continuous variable, once discretized into the various bins can be found in Appendix 4 and examples of the resulting networks can be seen in Figure 3.5 for the Star topology and Figure 3.6 for the 3_In topology.

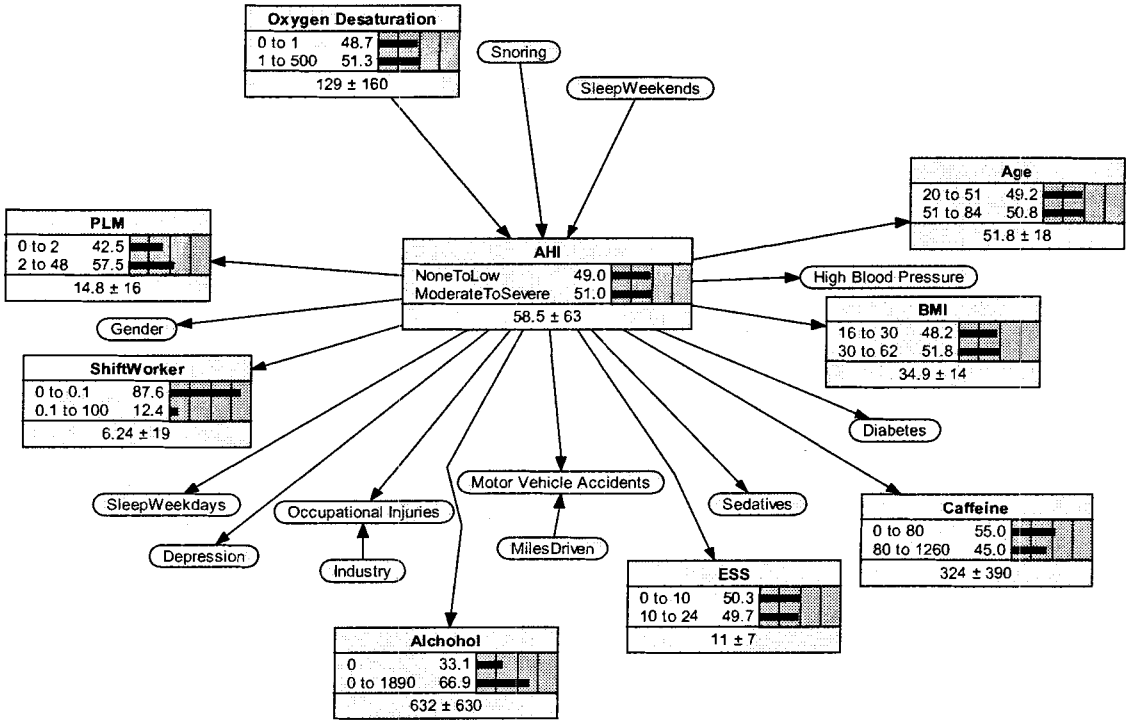

Figure 3.5 Star Topology Network with Continuous Variables Discretized into Two Equal Sized Bins

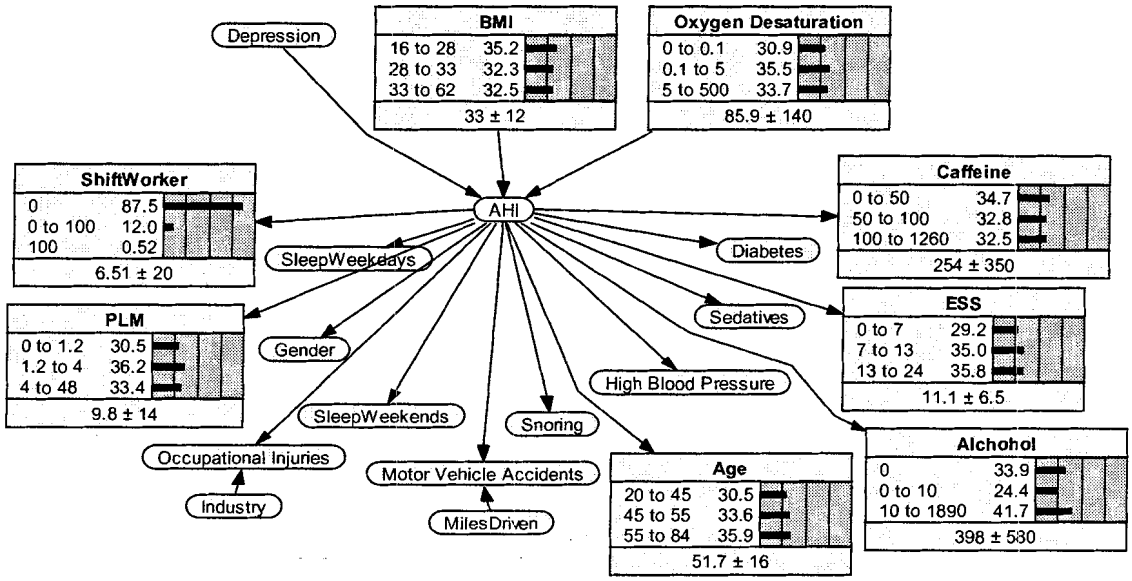

Figure 3.6 3_In Network Topology with Continuous Variables Discretized into Three Equal Sized Bins

3.8.2 Expert Discretization of Continuous Variables

A different strategy for breaking up continuous variables into ranges or bins involves looking at statistics for these variables and how they relate to AHI. In this way, continuous variables can be broken up into ranges that favour the differentiation between cases that have low, medium and high values of AHI. In order to accomplish this, box plots of each continuous variable against AHI are created using SPSS 13.0 for Windows, a software package by SPSS Inc. used for generating statistics. The continuous variables are first re-categorized as ordinal variables and recoded using SPSS in a variety of ways. The breakdown that seems to provide the optimal separation between low and high AHI is chosen as the 'expert' way of discretizing that particular continuous variable. This is done by examining a box plot of each categorization plotted against the AHI as a continuous variable. A box plot is a visual display that summarizes data using a ``box and whiskers" format to show the minimum and maximum values (ends of the whiskers), interquartile range (length of the box), and median (line through the box). Each variable

is broken up into two ranges or categories to avoid the AHI query node having too many entries in its conditional probability table. This is done for the following variables: BMI, ESS, Age, Snoring and Oxygen Desaturation as these are the continuous variables that most directly affect AHI in the knowledge engineered network topology. Age, Snoring and BMI are parents of the AHI node while High Blood Pressure, ESS and Oxygen Desaturation are children of the AHI node. Therefore, these are the nodes that should have the greatest causal links to the AHI node.

3.9 Adding a Node to an Existing Bayesian Network

When working with Bayesian networks, it is interesting to add variables into a network in order to see the effect they have on the predictive ability of the network. In our network two additional variables of interest are investigated as nodes to insert into the network.

When using large data sets, libraries of network components can be trained separately to decrease the time it takes to train a network. Then, these network components can be re-assembled to do 'belief updating' or inference on the combined network. We could have created a network component for each new node and its parents that we were inserting, trained these network components separately and then added them to the existing networks however, it was decided this would take more time than simply retraining networks with the new node added in. In order to achieve this, the training and test data sets were simply augmented by another variable, and a new set of networks created with the added node. These are then trained and subsequently tested on the augmented data sets.

3.9.1 Adding a Node with Small Amounts of Data Relative to Existing Data Set

As mentioned in Section 2.3.2, pupillometry has been investigated as a possible source of autonomic measures of sleepiness. Our hypothesis is that such data will enhance the predictive ability of our Bayesian networks either in combination with the ESS node (a standard though controversial subjective measure of sleepiness) or alone. As mentioned above in section 3.2.2.3, we gathered pupillometry data from 44 but were only able to use the data from 28 of these patients. As a result of this, data records containing valid values for the pupillometry variable PUI constitute only 4.29% of the total number of valid data records. Unfortunately, PUI has no significant correlations with any other OSAS variables (AHI, ESS, BMI, Oxygen Desaturation). A scatter plot of PUI against AHI can be seen in Figure 3.7. Other variables that were investigated for correlations with PUI included total sleep time, sleep efficiency and sleep latency the night of the overnight PSG. None of these correlates significantly with PUI either. For the remainder of the data records in the database created, PUI has missing values.

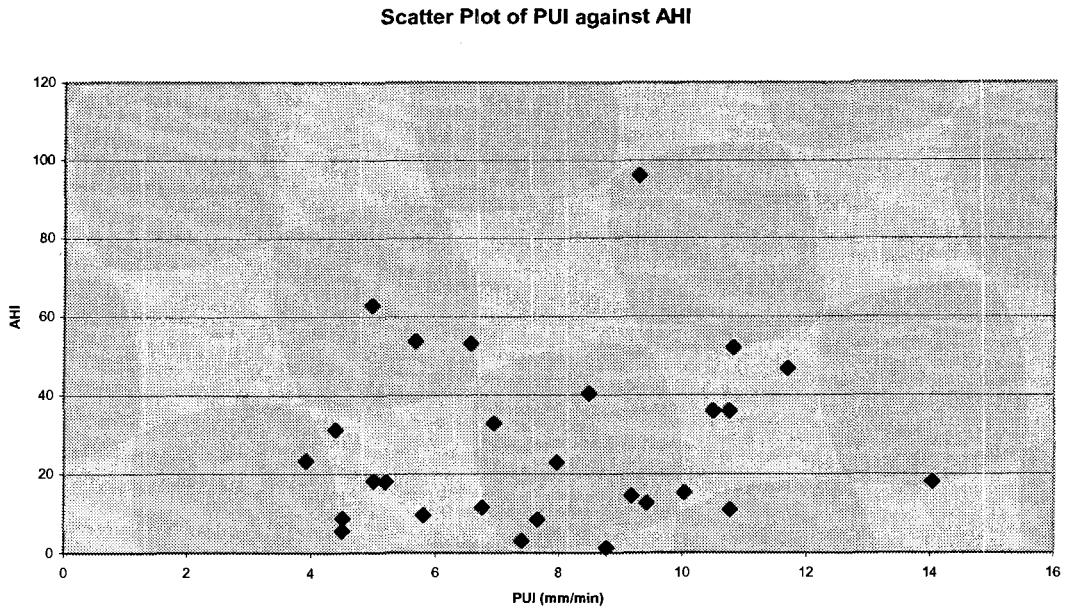

Figure 3.7 Scatter Plot of PUI against AHI

We investigate what effect, if any, the addition of a node into the network with so few training examples may have on the ability of the network to predict the value of our query node AHI. We begin with the PUI node, adding it to the best performing network, 3_In_3Levels. PUI was added as a child node of AHI, as sleepiness is a consequence of OSAS, not a cause of it. This resulted in a network structure as seen in Figure 3.8.

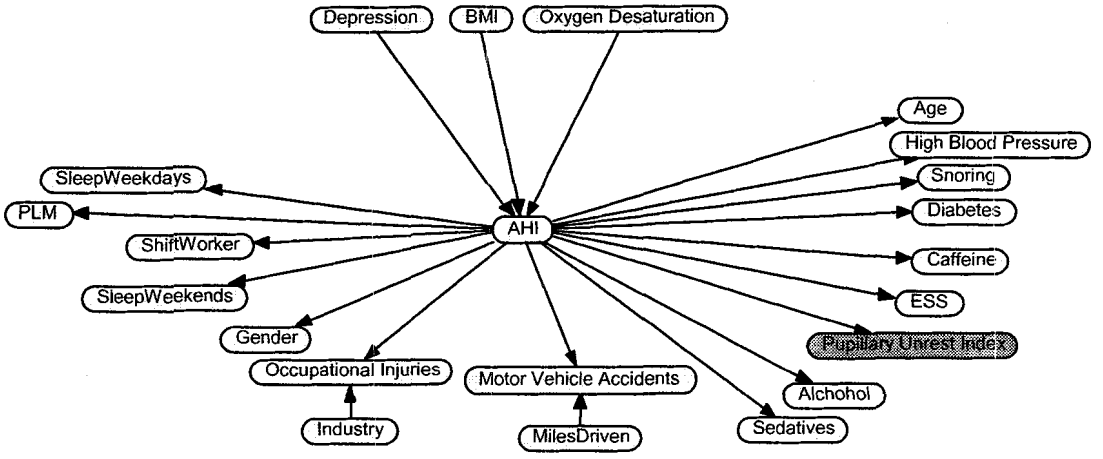

Figure 3.8 3_In Bayesian Network with PUI node added

3.9.2 Adding a Node with Moderate Amounts of Data Relative to Existing Data Set

The next area of investigation became adding a variable for which there were significantly fewer missing values. We chose neck circumference because it is a simple measurement reflecting obesity in the upper airway. While it alone is still not sufficient to predict obstructive sleep apnea severity without an overnight sleep study, it has been found to be more useful, especially when corrected for height, in predicting the severity of OSAS than measures of general obesity such as BMI [67]. Other studies show that when apnoeic and non-apnoeic patients are matched one-for-one for BMI and age, neck circumference was significantly higher in apnoeic patients [68]. Unfortunately, the neck circumference value in our data set is not a clinical measure. Instead, it is a response to one of the questions in the subjective questionnaire (Data Source one). As a result, results will be less accurate than clinical measurements. In addition, patients are able to respond that they are unsure of their neck circumference. Such responses were recoded as missing values. After this recoding, of the total 652 data records, 303 data records contained valid subjective values for the neck circumference attribute. This constitutes

46.5% of the total records, over ten times more valid records than the number of records containing valid PUI data as discussed above.

We initially chose to discretize neck circumference into three ranges (Low[0, 13], Medium [14, 17] and High [18, over 22] inches). However, when plotted against AHI it was found that the medium and high ranges had significant overlap as seen below in Figure 3.9.

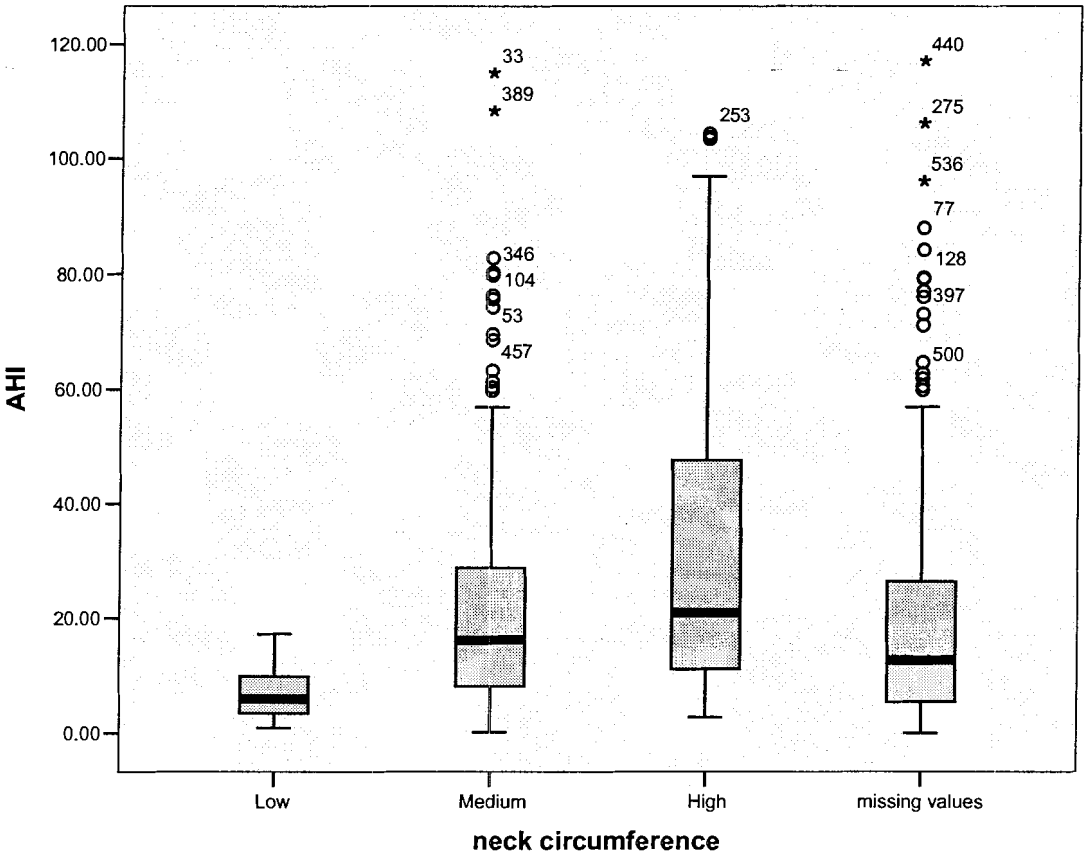

Figure 3.9 Box Plot of Neck Circumference (3 categories) against AHI

As a result, we recategorized the neck circumference into two ranges ([0, 15], [16, over 22] inches. This seems to provide better separation between ranges when plotted against AHI as seen in Figure 3.10.

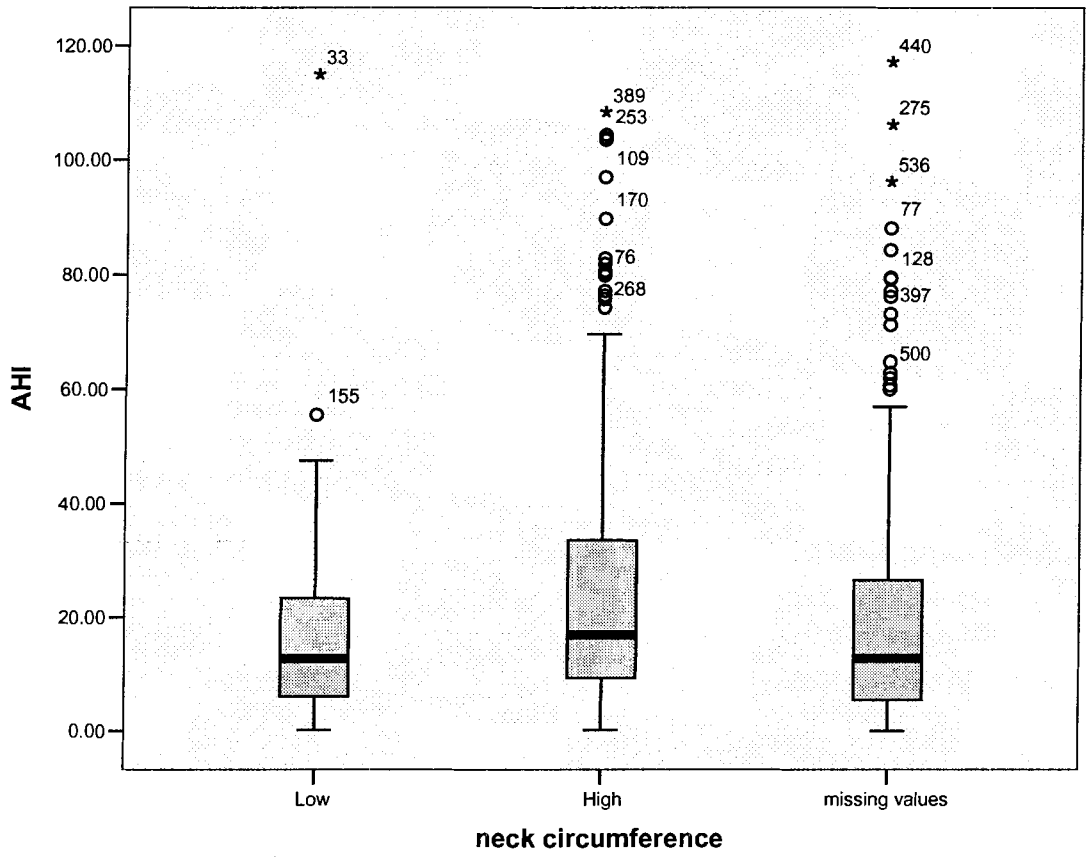

Figure 3.10 Box Plot of Neck Circumference (2 categories) plotted against AHI

We chose the same network to add the neck circumference node to as the one to which we add the PUI node, that is, 3_In_3Levels, in order to be able to properly compare results.

3.10 Investigation of the Effect of Node Removal on Network Performance

Once having observed the effects that adding a node to a BBN can have on a network's predictive performance, the next investigation is the removal of a node from an existing network. We hypothesize that nodes representing variables that do not correlate well with AHI could possibly hinder accurate prediction of the value of AHI as opposed

to facilitate it. The Pearson correlation coefficients of five nodes representing continuous variables most likely to affect AHI, especially in the expert designed network layout, were computed and can be seen in Table 3.2. The Pearson correlation coefficient is a measure of how well a linear equation describes the relation between two variables measured on the same object or organism. From these correlations, it would seem likely that eliminating Age, ESS, and Snoring nodes from networks should decrease the network prediction errors. Eliminating the BMI may decrease network prediction errors, although, in our data set it is better correlated with AHI than the other three variables.

Table 3.2 Pearson Correlation Coefficients of Different network Variables with AHI

Variable	Pearson Correlation Coefficient r with AHI
Age	0.033
ESS	0.132 (significant at the 0.01 level)
Snoring	0.141 (significant at the 0.01 level)
BMI	0.286 (significant at the 0.01 level)
Oxygen Desaturation	0.515 (significant at the 0.01 level)

We begin with the base network, (both in terms of network topology and in the way that continuous variables are discretized into ranges) that we had used in the node addition experiments, the 3_In_3Levels network. We proceed in three phases. The first phase removes only one node from the network at a time. The second phase removes two nodes from the network at a time, and the third phase removes three nodes from the network at one time. As with the topology and node addition experiments, for each resulting network we use nine training and test data sets, and we average the prediction errors together in order to obtain an average network prediction error for the AHI query node. These training and test sets are the same as those used in the other experiments in order to maintain consistency.

CHAPTER 4:RESULTS AND DISCUSSION

4.1 Topology Exploration

4.1.1 Star Topology

The Star topology networks result in average sensitivities, specificities and positive predictive values as seen in column1 of Table 4.1, Table 4.2 and Table 4.3. The Star topology represents a naïve Bayesian network. In this network, no arcs are allowed between variables, just one from the variable being classified (AHI) to all other variables in the network. While this topology does not model the causal links between all the variables in the network, researchers have shown that when used for classification problems, naïve Bayes networks tend to outperform more complicated networks [69]. Performance depends on how the continuous variables in a given network are discretized, but in terms of sensitivity, the Star topology performed best in half of all cases, and second best in the other half. In terms of specificity, the star topology network faired slightly less well, coming in second best 75% of the time and first in the remaining network. However, in terms of positive predictive value, the star topology fared the best, coming in first place in all different discretized network configurations.

Table 4.1 Sensitivity of different network topologies

	Star topology	3_In topology	Knowledge Engineered topology
Discretization Strategy			
2 bins	71.5% +/- 7%	67.6% +/- 10%	67.3% +/- 11%
3 bins	67.5% +/- 9%	72.4% +/- 6%	59.9% +/- 8%
5 bins	69.6% +/- 8%	66.0% +/-12%	61.7% +/-10%
Expert	66.2% +/- 7%	65.8% +/- 8%	71.4% +/-11%

Table 4.2 Specificity of different network topologies

	Star topology	3_In topology	Knowledge Engineered topology
Discretization Strategy			
2 bins	50.0% +/- 6.6%	49.2% +/- 6.2%	51.2% +/- 4.9%
3 bins	53.7% +/- 6.6%	49.5% +/- 8.5%	53.7% +/- 8.8%
5 bins	53.1% +/- 5.3%	51.0% +/- 9.8%	54.0% +/- 8.8%
Expert	45.3% +/- 3.8%	42.1% +/- 3.2%	36.2% +/- 10.9%

Table 4.3 Positive Predictive Value of different network topologies

	Star topology	3_In topology	Knowledge Engineered topology
Discretization Strategy			
2 bins	58.6% +/-5.1%	56.8% +/- 6.6%	57.5% +/- 5.2%
3 bins	59.0% +/- 3.2%	58.9% +/- 4.3%	56.5% +/- 8.0%
5 bins	59.4% +/- 6.5%	57.1% +/-7.4%	57.0% +/- 7.6%
Expert	54.4% +/- 5.8%	52.8% +/- 5.4%	52.7% +/- 6.2%

As discussed in Section 3.7.1, we also investigated which nodes in these star topology networks most affected the predicted value of our query node AHI, by using the "Sensitivity to Findings" feature that Netica™ provides. These results can be seen in Table 4.4. It is not surprising that Oxygen Desaturation influenced AHI's predicted value more than other variables in all network configurations. Of all the variables in the network, it is the most strongly correlated with AHI. However, other variables in the "top three" list were a bit of a surprise. If one looks at the features as ranked by

correlation coefficients in Table 4.5, one expects BMI and PLM to be the other two variables that influence the predictive value of AHI the most. This is not the case. Instead, it is Sleep Weekends, Snoring and Depression that appear most often in the top 3 Findings nodes.

Table 4.4 Top 3 Findings Nodes in terms of influencing predicted AHI value

Discretization Strategy	Top 3 Finding Nodes for that network
Binning – 2 Levels	Oxygen Desaturation, Sleep Weekends, Snoring
Binning – 3 Levels	Oxygen Desaturation, Depression, Body Mass Index (BMI)
Binning – 5 Levels	Oxygen Desaturation, Snoring, Depression
Expert Discretization – 2 Levels	Oxygen Desaturation, Sleep Weekends, Snoring

Table 4.5 Correlation Coefficients of different Variables with AHI

	Pearson Correlation Coefficient r
Oxygen Desaturation	.515(**)
BMI	.286(**)
PLM	.149(*)
Snoring	.141(**)
ESS	.132(**)
Sleep Weekends	.110(*)
Sleep Weekdays	.110(*)
Shift Worker	.105(*)
Caffeine	-.049
Age	.033
Alcohol	-.015

*. Correlation is significant at the 0.05 level (2-tailed)
**. Correlation is significant at the 0.01 level (2-tailed)

Discussion of discretization of continuous variables will be detailed further in Section 4.2.

4.1.2 3_In Topology

The 3_In network topology was created by reversing the links between the top three Findings nodes and the AHI node such that the links point in towards the AHI node

instead of out of it. These networks result in average sensitivities, specificities and positive predictive values as seen in column 2 of Table 4.1, Table 4.2 and Table 4.3.

The 3_In topology does not result in significantly better sensitivity or specificity. Instead, in comparison to the Naïve Bayes Star network topology, it performs consistently worse, with one exception. The sensitivity of the 3_In_3Levels network is the best of all the networks (72.4%). This is the network which has the Oxygen Desaturation, Depression and Body Mass Index(BMI) nodes as parents of the AHI node and its continuous variables are discretized into three levels. It would seem that the interaction between these three variables and AHI, as well as the discretization level favour the identification of those individuals with moderate to severe OSAS. That said however, the same network perform significantly worse than the other two topologies in terms of specificity, indicating there are many false positives. This is a trend that is seen in all networks: the higher the sensitivity, the lower the specificity of a given network. The 3_In networks fare slightly better in terms of positive predictive value, where it is second best in all but one network configuration. One possible explanation for the relatively poor performance of this topology overall is that, of the nodes that most influence the AHI node, only two of them, Snoring and BMI are causally related to AHI in the direction indicated. Snoring can be indicative of a blockage in the throat or pharynx which in turn can cause a decrease in the oxygen entering or exiting the body, affecting AHI. BMI indicates overall obesity. A fatty neck can also favour a blockage of the pharynx while sleeping. However, Oxygen Desaturation is affected by, but does not affect the AHI of a given patient. SleepWeekends, while it may affect sleepiness or be an indication of the level of sleepiness experienced during the weekdays, does not itself

influence AHI. The same goes for Depression. It does not affect AHI directly. The fact that these 3_In network topologies do not reflect the causal relationships between the network variables, especially with the AHI query node, may explain why such network topologies when trained on real data give poor predictive results. However, this network topology was studied to investigate the interactions between groups of variables and AHI to see how this affected the predictive ability of a network.

4.1.3 Knowledge Engineered Topology

The knowledge engineered topology was created with the assistance of our medical sleep expert. It attempts to map out the important causal links between the variables in our network. These network topologies result in average sensitivities, specificities and positive predictive values as seen in the last column of Table 4.1, Table 4.2 and Table 4.3.

This network topology is sensitive to how continuous variables are discretized. In this topology, the AHI query node has three parents; Age, Snoring and BMI and it in turn is the parent of the three nodes; High Blood Pressure, Oxygen Desaturation and ESS (a measure of sleepiness). Together these six nodes can directly affect the predicted AHI value. In terms of sensitivity, with expert discretization, it performs within 1% of the best of all networks, the 3_In_3Level network. However, when continuous variables are discretized into three or five bins using Netica^{TM}, the network topology performs worse than the other two networks. In terms of specificity, with the exception of the network with expert discretization (which performs very poorly), this topology yields good results.

However, all these results are not very different from those of the naïve Bayes (star) topology or the 3_In topology, even though its structure should reflect the actual causal links between the variables in the network. There are several possible reasons for this. One is that in each network, there are six or less nodes which are very important in terms of influencing values at our query node AHI and that in each network topology, these nodes are within the Markov blanket of the AHI node. This consists of the AHI node's children, those children's parent nodes, and the AHI's parent nodes. As a result, all the nodes of importance are present and influencing the AHI node in all the topologies. Another possible reason for this poor performance might be overfitting. In order to see if this was the case, we trained all three network topologies on the full data set and then tested each on the full data set that was used to train them. This is often viewed as an internal consistency test to see how well the independence assumptions encoded in the network topology are reflected in the actual data set used to train and test. The resulting sensitivities can be seen in Table 4.6, specificities in Table 4.7 and positive predictive values in Table 4.8.

The Star topology did not perform very well, with sensitivities of 72.6%, 71.8%, 72.2% and 62.0% as can be seen in column one of Table 4.6, even though essentially, all these networks had to do is "remember" the data they had learned. These results imply that this structure is too far from the actual relationships present in the data to get accurate results. The Knowledge engineered topology however, performs extremely well when testing on the full training data set, in terms of sensitivity, specificity and positive predictive value. This suggests that this network topology does in fact model the relationships between the various variables. If this is the case, why did the Knowledge

Engineered topology perform so badly when doing 10 fold cross validation? Our hypothesis is that the poor performance is a result of overfitting. This is a phenomenon which occurs when a learning algorithm adapts so well to a training set, that random disturbances in training set (noise) are included as being meaningful. This is evidenced by predictive performance on a test set being much lower than when testing on the training set. Our Knowledge Engineered topology performed fairly well with Expert discretization, however, it performed poorly with other discretization levels when we look at the ten-fold cross validation results. However, when testing on the training set of data, the Knowledge Engineered topology performs very well, with over 90% sensitivity and specificity for the 3 and 5 bin networks, as seen in rows 2 and 3 of Table 4.6 and Table 4.7 respectively. This is indicative that the structure of this network topology does indeed match the relationships the variables display in the data set. It is most likely that learning was performed too long and that some form of early stoppage will result in a network which will generalize better than the current trained Knowledge Engineered network topology does. Having additional data for training will also alleviate this problem.

Table 4.6 Sensitivity of Network Topologies When Testing on Full Training Data

	Star topology	**3_In topology**	**Knowledge Engineered topology**
Discretization Strategy			
2 bins	72.55%	73.68%	74.4%
3 bins	71.80%	76.69%	90.97%
5 bins	72.18%	75.56%	90.60%
Expert	62.03%	63.16%	86.09%

Table 4.7 Specificity of Network Topologies When Testing on Full Training Data

	Star topology	3_In topology	Knowledge Engineered topology
Discretization Strategy			
2 bins	71.37%	69.41%	78.03%
3 bins	70.61%	63.52%	91.76%
5 bins	71.37%	73.73%	92.55%
Expert	66.67%	64.31%	78.01%

Table 4.8 Positive Predictive Values of Network Topologies When Testing on Full Training Data

	Star topology	3_In topology	Knowledge Engineered topology
Discretization Strategy			
2 bins	72.56%	71.53%	78.00%
3 bins	69.96%	68.69%	92.02%
5 bins	72.45%	75.0%	92.69%
Expert	66.0%	64.86%	80.35%

4.2 Discretization of Continuous Variables

4.2.1 Binning Strategy

The average performance of the various network topologies when continuous variables are discretized into 2, 3 and 5 approximately equal sized bins are summarized in the first three rows of Table 4.1, Table 4.2 and Table 4.3. There appears to be no trend in terms of which level of discretization provides the best performance, except for the fact that in terms of sensitivity, the 5 level network tends to perform the worst, and in terms of

specificity it performs the best. As before, networks that perform well in terms of sensitivity tend to perform poorly in terms of specificity. When ranked in terms of Positive predictive value, each level of discretization performs in first place in one topology, in second place in another and in third place with a third topology, and as such perform equally well. Instead, how the networks perform varies with network topology. It seems to be more the interaction between a topology and a discretization level that results in a network that performs well. Discretization using binning does not seem to provide a tangible change that affects performance regardless of network topology.

4.2.2 Expert Discretization

The resulting ranges for each variable when discretized using statistical knowledge of the variables and how they relate to AHI can be seen in Table 4.9 and results in the sensitivities, specificities and positive predictive values as found in the last row of Table 4.1, Table 4.2 and Table 4.3. This strategy yielded little results. With the single exception of the network that combines Expert discretization with the knowledge engineered topology, all networks with expert discretization performed in last place in terms of sensitivity, specificity and positive predictive value. This goes against what we hypothesized. It is not known why the networks performed so poorly when discretized in this manner. However, the combination of knowledge engineered topology and knowledge engineered discretization does result in a network that performs very well in terms of sensitivity. This same network (knowledge engineered topology and expert discretization) performs very badly in terms of specificity. Most likely this mix of high sensitivity and low specificity is also reflective of the fact that in general, medical experts are biased towards erring on the side of false positives as opposed to false negatives.

Table 4.9 Expert Discretization Ranges of Continuous Variables

Variable Name	Ranges used for discretizing the variable
BMI	[16, 28], [28.1, 62]
ESS Score	[0, 8], [9, 24]
Snoring	[0, 4] time per week, [5, 7] times per week
Age	[20, 45], [46, 84]
Oxygen Desaturation	[0,50], [51, 200]

4.3 Adding Nodes to Existing Networks

4.3.1 Adding a PUI Node

Contrary to our hypothesis, PUI did not statistically significantly correlate with AHI (Pearson correlation coefficient r=0.043). That being said, ESS, a standard for assessing subjective sleepiness, also correlated very little to AHI (Pearson correlation coefficient r= .132, significant at the 0.01 level). After the pupillometry data had been gathered, when discussing the poor results with the equipment manufacturer, they mentioned that they had performed a study when pupil data had been gathered first thing in the morning. They found that the data gathered in the first 90 minutes of being awake did not correlate with any other sleepiness measures. It seems that individuals vary greatly in terms of sleep inertia (the feeling of grogginess after awakening). This is not something we had taken into account in our study. We recorded the pupil sizes of up to 3 subjects per morning, from immediately after their awakening to 45 minutes after their awakening. This was done in order to facilitate getting volunteers for our study as the Sleep Clinic patients were required to stay at the clinic in the morning until a sleep doctor arrived to consult with them on their overnight PSG results. Afterwards, most patients rushed off to work, and most patients did not want to stay after their visit with the sleep doctor. Sleep inertia temporarily reduces a person's ability to perform even simple tasks

and can last from 1 minute to 4 hours, but typically lasts 15-30 minutes in normal individuals. However, effects can be severe if a person is very sleep deprived or has been woken from a deep sleep stage. It may be that our pupillometry data should have been gathered slightly later in the day, after at least a 90 minute period had elapsed, allowing for all subjects to have recovered from their sleep inertia. Nonetheless, we still proceeded with the experiment of adding a node with very sparse data into our networks to observe the results.

Inserting a PUI node, when PUI is discretized into three categories ([3, 7], [7, 11], [11, 15]) very slightly decreases the sensitivity while it very slightly increases the specificity and positive predictive value of a network, averaged over the training and test sets, as can be seen in row 2 of Table 4.10. This is a small change, most likely attributeable to the slight differences between these testing and training sets and the ones used to train and test the network in the topology and discretization experiments. This change was necessary to make the training and test sets stratified so that they included the same number of pupillometry records in each test set and covered all of them.

Table 4.10 Sensitivity, Specificity and Positive Predictive Value with Node Added Networks

Network Topology	Sensitivity	Specificity	Positive Predictive Value
3_In_3Levels	72.4% +/- 6.0%	49.5% +/- 8.5%	58.9% +/- 4.3%
3_In_3Levels with PUI inserted (3 Levels)	72.1% +/- 9.3%	53.4% +/- 9.8%	60.7% +/- 7.0%
3_In_3Levels with neck_circumference inserted	70.5% +/- 7.6%	52.0% +/- 7.6%	59.3% +/- 5.4%

4.3.2 Adding a Neck Circumference Node

Our second experiment involves adding a node representing a variable which, while it still has a great deal of missing values, still has 46.5% of the total records which contain valid neck circumference data. We hypothesize that the addition this node to the network will increase the sensitivity of the resulting network. Firstly, in our data set, neck circumference correlates mildly with AHI severity (Pearson correlation coefficient r = 0.224, significant at the 0.01 level). Secondly, it has ten times more data records containing valid data than that which exists for the PUI node.

Once trained, this augmented network is able to predict the presence of moderate to severe OSAS with a sensitivity 2% lower than our base network, a specificity 2.5% higher than our base network and positive predictive value 0.4% higher than the base network, as can be seen in row 3 of Table 4.10. This is not what was expected. We speculate that there is some adverse interaction between this neck circumference node and another in the network.

4.4 Removing Nodes from Exiting Networks

It was hypothesized that removing nodes with low correlation to AHI could increase network performance in terms of its ability to accurately predict moderate to severe OSA (AHI node). We felt that most likely removing a combination of the two lowest-correlating variables (Age and ESS) would result in a network with the best sensitivity. Removing three nodes at one time might remove too much information from the network and instead increase errors. The sensitivities, specificities and positive predictive values for the resulting networks can be seen in Table 4.11. Out of the eleven new networks created, the top 3 performers in terms of sensitivities are the following:

with Age and BMI removed, with Age and Snoring removed and with ESS and BMI removed. What is surprising is that the network with Age and ESS removed, which was predicted to perform best, has an average sensitivity even lower than those of the base network (in which no node was removed). The reason for this is unknown. Another surprise is that every network in which the BMI node is removed performs better than the base network, in spite of BMI being relatively well correlated with AHI. This being said, the changes in sensitivity resulting from node removal are within 1% to 4% of those of the base network, and as such are not very large. Therefore, it seems that in this data set, removing these nodes that represent variables that are not highly correlated with AHI does not have a large effect on sensitivities and specificities. Even removing the three nodes Age, ESS and Snoring results in a sensitivity only 1% higher than the base network.

Table 4.11 Sensitivity, Specificity and Positive Predictive Value for Moderate to Severe OSAS when Removing Nodes from Network

Network Topology	Sensitivity	Specificity	Positive Predictive Value
3_In_3Levels (base network)	72.4% +/- 6.0%	49.5% +/- 8.5%	58.9% +/- 4.3%
3_In_3Levels without Age	72.5% +/- 7.2%	51.5% +/- 5.6%	60.3% +/- 4.9%
3_In_3Levels without ESS	71.4% +/- 8.1%	53.1% +/- 9.0%	60.2% +/- 5.9%
3_In_3Levels without Snoring	72.1% +/- 6.3%	49.7% +/- 5.5%	58.7% +/- 3.5%
3_In_3Levels without BMI	72.8% +/- 3.9%	53.1% +/- 17.9%	57.6% +/- 5.5%
3_In_3Levels without Age, ESS	72.1% +/- 7.9%	52.6% +/- 8.5%	60.2% +/- 6.0%
3_In_3Levels without Age, Snoring	75.0% +/- 3.7%	51.1% +/- 12.0%	61.0% +/- 6.0%
3_In_3Levels without Age, BMI	75.5% +/- 4.3%	47.2% +/- 6.0%	58.6% +/- 5.7%
3_In_3Levels without ESS, Snoring	71.7% +/- 7.7%	49.7% +/- 7.0%	58.6% +/- 4.5%
3_In_3Levels without ESS, BMI	73.5% +/- 2.9%	47.5% +/- 6.1%	58.2% +/- 6.2%
3_In_3Levels without Snoring, BMI	72.6% +/- 6.0%	45.2% +/- 5.4%	56.8% +/- 4.6%
3_In_3Levels without Age, ESS, Snoring	73.4% +/- 7.9%	50.0% +/- 5.0%	59.1% +/- 4.3%

CHAPTER 5:CONCLUSIONS AND FUTURE WORK

Contributions of this thesis include development of a set of Bayesian network models of Obstructive Sleep Apnea. The networks include the naïve Bayesian network, one created by experimenting with the naïve Bayesian network and a knowledge engineered network. We have investigated varying topology and varying the way in which continuous variables are discretized. We have also studied the effects of simplifying existing network models by removing nodes representing variables that do not correlate well with our query variable, observing the impact this has on the predictive abilities of the networks. Lastly, we have studied the effects of adding new nodes to networks when there are both small and large amounts of missing data.

Using Bayesian network modelling software such as Netica^{TM}, these models provide a graphical user interface that is simple to use and adapt. In addition, for those networks knowledge engineered with the help of a medical expert, the networks show a causal map of OSAS risk.

We have found that the 3_In_3Levels network topology, with the Age and BMI nodes removed performed best overall, with a sensitivity of 75.5%, a specificity of 47.2% and a positive predictive value of 58.6%. Overall, neither network topology nor discretization strategy alone influences network performance in terms of being able to accurately predict if a subject has moderate or severe OSAS. Instead, it seems that the interaction between the topology and how continuous variables are discretized is influencial in determining the predictive accuracy of the network. This being said, all

network topologies perform similarly when doing 10 fold cross validation. However, it is suspected that there may be overfitting occuring and that the Knowledge Engineered network topology can actually result in much higher sensitivities and specificities than those shown in these results. The reason for this belief is that when testing on the complete training data set, sensitivities and specificities in the Knowledge Engineered network topology, especially the 3 and 5 bin networks increased to over 90%. It is hoped in the future, a combination of additional data and some form of early stoppage will allow the creation of Knowledge Engineered network which will generalize better than the one we currently have. That being said, the current Knowledge Engineered network is a good base from which to continue future experiments.

In addition, it seems that networks with fewer nodes perform better overall than networks with a large number of nodes. Adding the PUI node influenced the performance of the network very little, which is not surprising given the small amount of data records with actual values for the PUI variable. But even adding neck circumference, a variable that is correlated in our data set to AHI index and has a large number of valid values, does not increase performance. Instead, it actually decreases sensitivity by 2%. In comparison, removing nodes from the 3_In_Level network resulted in higher sensitivity for predicting moderate to severe OSAS in seven out of eleven cases, even when the variables being removed were correlated significantly with AHI (BMI, Snoring). This is thought to be in part because of the small amount of training data. If there were a great deal more data available for training, then it is likely that removing nodes would not increase networks' predictive abilities.

Performance in terms of specificity was lower across all network configurations and discretization strategies when compared to sensitivity. There were significantly more false positives than false negatives in the results overall. In general, changes made to networks that resulted in higher sensitivities also resulted in lower specificities indicating that with an increasing number of true positives, there was an accompanying increase in false positives. This was expected to a certain extent because the cases used to train the network were all clinical cases referred to the sleep clinic for suspicion of OSAS, heavily biasing the data towards a positive diagnosis. Also, physicians are biased towards making false positives rather than false negatives.

There remains much territory to be explored.

It would be interesting to collect more pupillometry data, but this time when subjects are not under the influence of varying amounts of sleep inertia – either later in the morning, or perhaps in the evening before they go to sleep. Acquiring more data from all three data sources would also be useful in making the training data sets more robust. This combined with early stoppage would hopefully alleviate the problems of overfitting which seem to be present in the current networks.

Another experiment would be to adapt the data set, by removing missing values, such that we could run a structure learning algorithm and compare the topologies that the program would suggest with the ones we have created.

Another direction of future exploration would be the treatment of missing data. Currently, we simply treat missing values as having an unknown value. Other strategies that have been explored by others include removing data records with missing values, modelling missing values as a separate category, namely, "missing" or replacing missing

values with the mean values for their corresponding variables. Our data sets could be altered to try these different strategies and see whether this influences predictive ability in a positive or negative way.

In addition, as it seems that simpler models perform better in terms of their predictive ability, another set of future experiments would include node removal experiments on the knowledge engineered network topology to see resulting performance.

If it were possible to obtain matched normal controls and access to a sleep clinic for overnight polysomnography, it would be very interesting to see how the networks we have developed perform as a screening system. An OSAS screening technology could be used as a way of ensuring that individuals at high risk of having OSAS but who would not otherwise present themselves to a sleep clinic, either due to their location (far from major urban centre) or denial of the problem, could be flagged as needing to be examined by a sleep expert. In theory, if these networks are able to separate clinical cases that have moderate to severe OSAS from those who do not, they should be able to predict with even greater specificity those out of a general population most likely have moderate to severe OSAS.

Finally, whatever the set of networks would result from this future work, it will be necessary to validate the predictive results with a doctor on different data and analyze whether the results would be better or worse than current diagnostic procedures in correctly identifying patients with moderate to severe OSAS.

APPENDICES

Appendix 1. Eyecheck™ Pupillometer Specifications

A pupillometer is an instrument used to measure the changes in the size of the pupil of the eye over time. The EyeCheck™ pupillometer relies on scattering techniques. An infrared beam is shone into the eye and the pupillometer collects of light leaving the pupil and uses this to calculate pupil size. This type of process entitles EyeCheck™ to measure pupil diameter and it's reaction to light in real time, for ABSOLUTE pupil dynamic measurements. It allows both eyes to be tested, albeit one at a time.

The subject is seated in a comfortable chair in a quiet room. The pupillometer is held up to the subject's eyes, and the subject simply peers into the unit's viewing area. The visor around the viewing area blocks out outside light. The eyes are illuminated with infrared light and the scattering of this light from the eyes is collected in order to calculate the size of the pupil. These measurements are taken 100 times per second (100Hz) and the data stored on a computer.

Dark Adaptation Test:

The subject's pupil sizes are recorded for 10 minutes, in a dark, quiet room. Following a brief period of adjustment after the lights are turned off, the pupil of an alert person maintains a constant size while that of a sleepy person on fluctuates slowly, a number of times during the 10 minutes. The amount of instability in pupil size can be calculated with the Pupillary Unrest Index (PUI), a quantitative measure of the average amount of change in size that occurs during the pupillometry test. Other measurements include the total cumulative miosis of the pupil over the 11 minutes, the starting size of

the pupil during the dark adaptation period, and the power of fluctuation frequencies below 0.8 Hz.

Pupil Light Response Test:

During the Pupil Light Response test (PLR) data is collected on changing pupil size in response to three light flashes. The light source is a yellowish green. Parameters that can be measured during this test include initial pupil diameter, final pupil diameter at maximum constriction, latency to onset of pupil constriction, total time to minimum pupil size, and reflect amplitude percentage. All of these are based on the pupil size and how it change over time in response to the administration of light.

The specifics of the EyeCheck^{TM} pupillometer are as follows:

Handheld Pupillometer Unit
Weight 40 ounces
Battery Pack w/Charger
7.2V 1500 mAh rechargeable battery
CD with Systems Display Software
Operator's Manual
DB9 cable or USB adaptor

Appendix 2. C++ Code for Algorithm Used to Calculate PUI from Raw 10 Hz Pupillometry Data

```
#include "stdafx.h"
#include "math.h"
#include <fstream>
#include <vector>
#include <stdlib.h>

using namespace std;

int main(int argc, char* argv[])
{
        // open input file
        ifstream inFile;
        char inFileName[100];
        strcpy(inFileName,argv[1]);
        inFile.open(inFileName);
        if ( inFile.fail())
                return 0;

        // open output file
        ofstream outFile;
        char outFileName[100];
        strcat(outFileName,

        "C:\\Sleep\\PupillometryData\\DarkData\\GoodFirst4Min\\PUIPerMinFirst4Min.txt");
        outFile.open(outFileName, ios::app);
        if (outFile.fail())
                return 0;

        // output the name of the input file for a given line of PUIs
        std::string inName(inFileName);
        std::string inSubstring;
        inSubstring = inName.substr((inName.find_last_of("\\"))+1);
        outFile << inSubstring.c_str() << "\t\t";

        // Begin processing the input file.
        // discard first 6 lines of file b/c they contain no useful data
        char introLine[200];
        for (int i=0; i<2; i++)
                inFile.getline(introLine, 200, 13);
        inFile.getline(introLine, 200);
        inFile.getline(introLine,200,13);
        inFile.getline(introLine,200,13);
        inFile.getline(introLine,200, 13);

        vector<int> minute1;
        vector<int> minute2;
```

```
vector<int> minute3;
vector<int> minute4;
vector<vector<int> > allSegments;

allSegments.push_back(minute1);
allSegments.push_back(minute2);
allSegments.push_back(minute3);
allSegments.push_back(minute4);

for (int k=0; k<4; k++)
{
        // We want PUI per minute:
        // read in all data points for a given minute, average each 6 values
        // and store them in one of the segment vectors
        if (!inFile.eof())
        {
                for (i=1; i<600; ) //there are 10 data points per second, 60 seconds
                {
                        if (! inFile.eof())
                        {
                                // read in 6 values and average them together
                                // each value is the first double on a given line
                                // store this in the corresponding vector
                                int avgValue = 0;
                                int newValue = 0;
                                char newValueString[4];

                                // read in newvalue from input stream
                                for (int j=0; j<6; j++)
                                {
                                        if (! inFile.eof())
                                        {
                                                inFile.get(newValueString,9);
                                                newValue = atoi(newValueString);
                                                avgValue += newValue;
                        // trash the rest of the line so the next time we read from the beginning
                                                inFile.getline(introLine, 100, 13);
                                                i++;
                                        }
                                        else
                                                break;

                                }

                                avgValue = avgValue/(j+1);          // take the average
                                allSegments.at(k).push_back(avgValue);
                                avgValue =0; // reset to 0

                        }
                        else
                                break;
                }
        }
        else
                break;
}
```

```
        // For each value in each vector, calculate the absolute difference
        // with its neighbour,
        vector<double> PUIPerMin;
        k=0;
        for (k=0; k<4; k++)
        {
                double sumDiffs = 0.00;
                for (unsigned int j =0; j+1 < allSegments.at(k).size(); j++)
                {
                        sumDiffs+= abs(allSegments.at(k).at(j) - allSegments.at(k).at(j+1) );
                }
                // Also, PUI in the Eyecheck file in recorded as 10 exp -5 m, so
                // to convert to mm we need to divide by 100.
                PUIPerMin.push_back(sumDiffs/100);
                outFile << sumDiffs/100 << "\t";    // no normalize since each vector is one minute
                //reset SumDiffs
                sumDiffs = 0.00;
        }

        outFile << PUIPerMin.at(0) << "\t";
        outFile << (PUIPerMin.at(0) + PUIPerMin.at(1))/2 << "\t";
        outFile << (PUIPerMin.at(0) + PUIPerMin.at(1) + PUIPerMin.at(2))/3 << "\t";
        outFile << (PUIPerMin.at(0) + PUIPerMin.at(1) + PUIPerMin.at(2) + PUIPerMin.at(3))/4 << "\t";

        outFile << endl;
        inFile.close();
        outFile.close();
        return 0;
}
```

Appendix 3. Variables and their Derivation

Variable Name	Definition	How Derived From Source Data	Possible States
Apnea Hypopnea Index (AHI)	the average number of apneas and hypopneas per hour recorded during sleep as measured by polysomnography	Data Source: overnight PSG study, entry c 25	Continuous variable
Periodic Limb Movement Arousal Index (PLM)	number of PLMS per hour of sleep leading to arousal, as defined by (American Sleep Disorders Association and Sleep Research Society Task Force criteria[70])	Data Source: overnight PSG study, entry c 15a	Continuous variable
Shift Worker	The estimated average percentage(over the last 2 years) of work time spent working night shift (11 pm to 7 am)	Data source: subjective questionnaire, question 24c	Continuous variable
SleepWeekdays	The estimated average number of hours of sleep obtained per night on weekdays	Data Source: subjective questionnaire, question 7a	<=5,6,7,8,9,>=10
SleepWeekends	The estimated average number of hours of sleep obtained per night on weekends	Data Source: subjective questionnaire question 7b	<=5,6,7,8,9,>=10
Occupational Injuries	The number of occupational injuries suffered over the last 24 months	Data Source: subjective questionnaire, question 26a	0,1,2,3,>=4
Motor Vehicle Accidents	The number of motor vehicle accidents over the last 24 months in which the patient was driving	Data Source: subjective questionnaire, question 45a	0,1,2,3,4,>=5
Sedatives	The frequency with which prescription medication (sleeping aids) is taken to get or stay asleep by the patient	Data Source: subjective questionnaire, question 36	Never, very rarely (less than once per month), rarely (less than one night per week but more than once per month), sometimes (1-2 nights per week), frequently (3-4 nights per week), almost always (5-7 nights per week)

Variable Name	Definition	How Derived From Source Data	Possible States
Alcohol	The total number of alcoholic beverages (beer, wine or liquor) consumed, on average (during the past year), per month	Data Source: subjective questionnaire, question 18 d,e and f converted to number of drinks per month and summed	Continuous variable
Epworth Sleepiness Scale Score (ESS)	The score on the Epworth Sleepiness Scale Questionnaire [17]	Data Source: subjective questionnaire, question 16 a,b,c,d,e,f,g,h, scored as per [17]	Possible scores are whole numbers ranging from 0 to 24
Caffeine	The total number of caffeinated beverages (carbonated beverages with caffeine, tea or coffee) consumed on average (during the past year), per month	Data Source: subjective questionnaire, question 18 a, b and c converted to number of drinks per month and summed	Continuous variable
Diabetes	Whether the patient has been diagnosed by a physician with Diabetes	Data Source: subjective questionnaire, question 33n	Absent, Present
Body Mass Index (BMI)	The body mass index of the patient, as calculated from their height and weight. BMI is defined as being equal to: Weight(kg)/(Height(cm) * Height(cm)) * 1000	Data Source: overnight PSG study, entries c2 and c3, converted to BMI, as per definition	Continuous variable
Gender	The gender of the patient	Data Source: subjective questionnaire, question 2	Male, Female
Depression	Whether the patient has been diagnosed by a physician with a major mood disorder (depression)	Data Source: subjective questionnaire, question 33a	Yes, No
Age	The age of the patient	Calculated from Data Source: subjective questionnaire, question 1.	Continuous variable
Snoring	The average (over the last month) the patient has been told they snored or noticed they were snoring	Data Source: subjective questionnaire, question 11	Never, rarely(less than once per week), sometimes (1-2 times per week), frequently (3-4 times per week), almost always (5 -7 per week), not sure

Variable Name	Definition	How Derived From Source Data	Possible States
High Blood Pressure	Whether the patient has been diagnosed by a physician as having hypertension	Data Source: subjective questionnaire, question 33i	Yes, No
Industry	The industry in which the subject has been occupied over the last 24 months	Data Source: subjective questionnaire, question 23	Agriculture, Fishing, Forestry, Oil and Gas/Mineral Resources, Food and Beverage Manufacturing, Metal/Non-Metallic Mineral Product Manufacturing, Petroleum or Coal or Rubber or Plastic or Chemical product manufacturing, wood and paper product manufacturing, other manufacturing, general construction, heavy construction, road construction or maintenance, warehousing, transportation and related services, Retail trade, wholesale trade, working for the federal government, Military service, Public administration other than the federal government, accommodation or food or leisure services, business services, healthcare and social assistance, professional or scientific or technical services, other services, other
Miles Driven	The average(over the last 24 months) number of kilometers driven per week	Data Source: subjective questionnaire, question 44	Continuous variable
Oxygen Desaturation	The total fraction (out of 200) of time spent during sleep with a oxygen desaturation below 90%	Data Source: overnight PSG report question, sum of questions 32 b,c, 33 b,c, 34 b,c, 35 b,c and 36 b,c	Continuous variable (max 200)
PUI	Pupillary Unrest Index, the average amount of pupil oscillation over time (mm/min) during a period of dark adaptation	Data Source: pupil size recording during dark adaptation the morning after the in-clinic overnight PSG recording, run through the PUI algorithm and averaged over the first four minutes of dark adaptation	Continuous variable

Variable Name	Definition	How Derived From Source Data	Possible States
RA%, Reflex Amplitude in percentage	The difference, expressed as a percentage between the pupil size before a flash of light is shone in the eye, and the pupil size 3 seconds after a light has been shone into the eye (thus triggering the pupillary light reflex)	DataSource: pupillometry data on PLR (pupillary light reflex)	Continuous variable
TTM	The time, in milliseconds, between a light flash in the eye and the point of maximum pupillary constriction as a result of the pupillary light reflex)	DataSource: pupillometry data on PLR (pupillary light reflex)	Continuous variable
TTI	The time, in milliseconds, between a light flash in the eye and the initiation of pupillary constriction	DataSource: pupillometry data on PLR (pupillary light reflex)	Continuous variable

Appendix 4. Continuous Variable Ranges when Discretized by Netica™

Variable Name	Ranges for 2 bins	Ranges for 3 bins	Ranges for 5 bins
Age	[0,51],[52,84]	[0,45], [46, 55], [56, 84]	[0,40], [41, 48], [49, 53], [54, 59], [59, 84]
Snoring	[0,4],[5,7] times per week	[0,3], [4,5], [5,7] times per week	[0, less than once per week], [1, 2], [3, 4], [5, 7] time per week
Oxygen Desaturation	[0,1],[1.1,200] (sum of percentage time spent below 90% saturation during REM and non-REM sleep	[0,0.1], [0.2, 5], [5.1, 200]	[0,0], [0, 0.3], [0.31, 3], [3.1, 14], [14, 200]
BMI	[16,302]],[30.1,62]	[16,28], [28.1, 33], [33.1, 62]	[16, 26], [26.1, 29], [29.1, 32], [32.1, 36], [36.1, 6
Caffeine	[0,80],[81,1260] beverages per month	[0, 50], [51, 100], [101, 1260] beverages per month	[0, 30], [31, 60], [61, 90], [91, 150], [150, 1260] beverages per month
Alcohol	[0,2],[3,1890] beverages per month	[0, 0], [0, 10], [11, 1890] beverages per month	[0, 0], [0, 2], [3, 12], [13, 30], [31, 1890] beverages per month
ESS Score	[0,19],[11,24]	[0, 7] ,[8, 13], [14, 24]	[0, 5],[6, 8],[9, 11], [12, 15], [16, 24]
PLM Index	[0, 2], [2.1, 48]	[0, 1.2], [1.3, 4], [4.1, 48]	[0, 0.7], [0.8, 1.6], [1.7, 3.2], [3.3, 6.0], [6.1, 48]

REFERENCE LIST

[1] M. Henrion, J. S. Breese, and E. J. Horvitz, "Decision Analysis and Expert Systems," *AI Magazine*, vol. 12, pp. 64-91, 1991.

[2] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell, "Bayesian Analysis in Expert Systems," *Statistical Science*, vol. 8, pp. 219 - 283, 1993.

[3] T. M. Mitchell, "Bayesian Learning," in *Machine learning*. New York ; London: McGraw-Hill, 1997, pp. 154-200.

[4] T. Young, L. Evans, L. Finn, and M. Palta, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women.," *Sleep*, vol. 20, pp. 705-706, 1997.

[5] S. Coren, *Sleep Thieves: An Eye-opening Exploration into the Science and Mysteries of Sleep*. New York: The Free Press, 1996.

[6] H. Hakkanen, H. Summala, M. Partinen, M. Tihonen, and J. Silvo, "Blink duration as an indicator of driver sleepiness in professional bus drivers.," *Sleep*, vol. 22, pp. 1-5, 1999.

[7] J. L. Krichmar, M. L. Thomas, L. Rowland, H. S. Sing, and D. Thorne, "Pupil dynamics and eye movements as indicators of fatigue and sleepiness," *Investigative Ophthalmology & Visual Science*, vol. 38, pp. S71, 1997.

[8] H. Danker-Hopfe, S. Kraemer, H. Dorn, A. Schmidt, I. Ehlert, and W. M. Herrmann, "Time-of-day variations in different measures of sleepiness(MSLT, pupillography, and SSS) and their interrelations," *Psychophysiology*, vol. 38, pp. 828-835, 2001.

[9] S. L. Merritt, H. C. Schnyders, P. Mercer, and X. Zhou, "The sensitivity of pupillography to circadian aspects of sleepiness.," *J. Sleep Res.*, vol. 7, pp. 176, 1998.

[10] S. L. Meritt, H. C. Schnyders, and P. Mercer, "Circadian aspects of pupillary unrest reflecting daytime sleepiness.," *Sleep Research Online*, vol. 2, pp. 773, 1999.

[11] B. Wilhelm, H. Wilhelm, H. Ludtke, P. Streicher, and M. Adler, "Pupillographic assessment of sleepiness in sleep deprived healthy subjects.," *Sleep*, vol. 21, pp. 258-265, 1998.

[12] W. C. Dement and C. Vaughan, *The Promise of Sleep*. New York: Dell Publishing, 1999.

[13] N. J. Douglas, *Clinicians' Guide to Sleep Medicine*. London: Arnold, 2002.

[14] K. Kinnari, J. H. Peter, A. Pietarinen, L. Groete, T. Penzel, A. Vam, A. Varri, P. Laippala, A. Saastamoinen, W. Cassel, and J. Hasan, "Vigilance stages and performance in OSAS patients in a monotonous reaction time task," *Clinical Neurophysiology*, vol. 111, pp. 1130-1136, 2000.

[15] G. Curcio, M. Casagrande, and M. Bertini, "Sleepiness: evaluating and quantifying methods.," *International Journal of psychophysiology*, vol. 41, pp. 251-263, 2001.

[16] T. Akerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *International Journal of Neuroscience*, vol. 52, pp. 29-37, 1990.

[17] M. W. Johns, "A new method of measuring daytime sleepiness: the Epworth Sleepiness Scale," *Sleep*, vol. 14, pp. 540-545, 1991.

[18] M. W. Johns, "Sleepiness in different situations measured by the Epworth Sleepiness Scale.," *Sleep*, vol. 17, pp. 703-710, 1994.

[19] M. W. Johns, "Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth sleepiness scale: failure of the MSLT as a gold standard.," *J. Sleep Res.*, vol. 9, pp. 5-11, 2000.

[20] C. Violani, F. Lucidi, E. Robusto, A. Devoto, M. Zucconi, and L. F. Strambi, "The assessment of daytime sleep propensity: A comparison between the Epworth sleepiness scale and a newly developed resistance to sleepiness scale.," *Clinical Neurophysiology*, vol. 114, pp. 1027-1033, 2003.

[21] H. Furuta, R. Kaneda, K. Kosaka, H. Arai, J. Sano, and Y. Koshino, "Epworth sleepiness scale and sleep studies in patients with obstructive sleep apnea syndrome.," *Psychiatry and Clinical Neurosciences*, vol. 53, pp. 301-302, 1999.

[22] H. Firat, S. Ardic, E. Gelir, S. Ozsahin, and B. Izci, "The relationship between arousal index and Epworth Sleepiness Scale among patients with suspected Obstructive Sleep Apnea Syndrome.," *Journal of Sleep Research*, vol. 9, pp. 66, 2000.

[23] L. G. Olson, M. F. Cole, and A. Ambrogetti, "Correlations among Epwroth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms.," *Journal of Sleep Research*, vol. 7, pp. 248-253, 1998.

[24] C. Stampi, A. Aguirre, M. Macchi, and S. Hashimoto, "Relationship between pupillary/oculomotor measures, and physiological (EEG) and behavioral indices of sleepiness.," *Sleep Research*, vol. 24, pp. 454, 1995.

[25] T. L. Morris and J. C. Miller, "Electrooculographic and performance indices of fatigue during simulated flight," *Biological Psychology*, vol. 42, pp. 343-360, 1996.

[26] M. L. Lobb and J. A. Stern, "Pattern of eyelid motion predictive of decision errors during drowsiness: oculomotor indices of altered states.," *International Journal of Neuroscience*, vol. 30, pp. 17-22, 1986.

[27] G. Barbato, G. Ficca, M. Beatrice, M. Casiello, G. Muscettola, and F. Rinaldi, "Effects of sleep deprivation on spontaneous eye blink rate and Alpha EEG power.," *Biological Psychiatry*, vol. 38, pp. 340-341, 1995.

[28] S. Porcu, M. Ferrara, L. Urbani, A. Bellatreccia, and C. M., "Smooth Pursuit and Saccadic Eye Movements as Possible indicators of nighttime sleepiness," *Physiology & Behavior*, vol. 65, pp. 437-443, 1998.

[29] M. Ferrara, L. De Gennaro, and M. Bertini, "Voluntary oculomotor performance upon awakening after total sleep deprivation.," *Sleep*, vol. 23, pp. 801-811, 2000.

[30] L. De Gennaro, M. Ferrara, G. Curcio, and M. Bertini, "Visual search performance across 40 h of continuous wakefulness: Measures of speed and accuracy and relation with oculomotor performance," *Physiology & Behavior*, vol. 74, pp. 197-204, 2001.

[31] P. Lavie, "Ultradian rhythm in alertness. A pupillometric study," *Biological Psychology*, vol. 9, pp. 49-62, 1979.

[32] M. Russo, M. Thomas, D. Thorne, H. Sing, D. Redmond, L. Rowland, D. Johnson, S. Hall, J. Krichmanr, and T. Balkin, "Oculomotor impairment during chronic partial sleep deprivation," *Clinical Neurophysiology*, vol. 114, pp. 723-736, 2003.

[33] J. Newman and R. Broughton, "Pupillometric assessment of excessive daytime sleepiness in narcolepsy-cataplexy.," *Sleep*, vol. 14, pp. 121-129, 1991.

[34] R. Ranzjin and L. Lack, "The pupillary light reflex cannot be used to measure sleepiness.," *Psychophysiology*, vol. 34, pp. 17-22, 1997.

[35] O. Lowenstein, R. Feinberg, and I. E. Loewenfeld, " Pupillary movements during acute and chronic fatigue.," *Investigative Ophthalmology & Visual Science*, vol. 2, pp. 138-157, 1963.

[36] R. E. Yoss, N. J. Moyer, and K. N. Ogle, "The pupillogram and narcolepsy: A method to measure decreased levels of wakefulness," *Neurology*, vol. 19, pp. 921-928, 1969.

[37] H. Ludtke, B. Wilhelm, M. Adler, F. Schaeffel, and H. Wilhelm, "Mathematical procedures in data recording and processing of pupillary fatigue waves.," *Vision Research*, vol. 38, pp. 2889-2896., 1998.

[38] B. Wilhelm, H. Wilhelm, D. Widmaier, H. Ludtke, and K.-H. Ruhle, "Therapy control in sleep apnea patients by means of a pupillographic sleepiness test.," *Sleep*, vol. 21, pp. 266, 1998.

[39] S. L. Merritt, H. C. Schnyders, M. Patel, R. C. Basner, and W. O'Neill, "Pupil staging and EEG measurement of sleepiness.," *International Journal of psychophysiology*, vol. 52, pp. 97-112, 2004.

[40] H. Wilhelm, H. Lüdtke, and B. Wilhelm, "Pupillographic sleepiness test applied in hypersomniacs and normals.," *Graefe's Arch Clin Exp Opthalmol*, vol. 236, pp. 725-729, 1998.

[41] S. J. Russell and P. Norvig, "Uncertainty," in *Artificial intelligence : a modern approach*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 2003, pp. 462-487.

[42] J. Pearl, "Introduction to Probabilities, Graphs and Causal Models," in *Causality : models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000, pp. 1-40.

[43] P. J. Lucas, L. C. van der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and health-care.," *Artificial Intelligence in medicine*, vol. 30, pp. 201-214, 2004.

[44] S. J. Russell and P. Norvig, "Probabilistic Reasoning," in *Artificial intelligence : a modern approach*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 2003, pp. 492-?

[45] G. Cooper, "The Computational Complexity of Probabilistic Inference Using Belief Networks.," *Artificial Intelligence*, vol. 42, pp. 393-405, 1990.

[46] S. J. Russell and P. Norvig, "Statistical Learning Methods," in *Artificial intelligence : a modern approach*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 2003, pp. 712-?

[47] S. Andreassen, M. Woldbye, B. Falck, and S. K. Andersen, "MUNIN - a causal probabilistic network for interpretation of electromyographic findings.," presented at 10th International Joint Conference on Artificial Intelligence, Los Altos, CA, 1987.

[48] D. E. Heckerman, E. Horvitz, and B. Nathwani, "Toward normative expert systems: Part I. The Pathfinder project.," *Methods of Information in Medicine*, vol. 31, pp. 90-105, 1992.

[49] D. E. Heckerman and B. Nathwani, "Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference.," *Methods of information in medicine*, vol. 31, pp. 106 - 116, 1992.

[50] S. Andreassen, C. Riekehr, B. Kristensen, H. C. Schonheyder, and L. Leibovici, "Using probabilistic and decision-theoretic methods in treatment and prognosis modeling.," *Artificial Intelligence in medicine*, vol. 15, pp. 121-134, 1999.

[51] P. J. Lucas, N. C. De Bruijn, K. Schurink, and I. M. Hoepelman, "A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU.," *Artificial Intelligence in medicine*, vol. 19, pp. 251-279, 2000.

[52] L. C. van der Gaag, S. Renooiji, C. L. M. Witteman, B. M. P. Aleman, and B. G. Taal, "Probabilities for a probabilistic network: a case study in oesophageal cancer.," *Artificial Intelligence in medicine*, vol. 25, pp. 123-148, 2002.

[53] F. de Dombal, D. Leaper, J. Horrocks, and A. McCann, "Computer-Aided Diagnosis of Acute Abdominal Pain," *British Medical Journal*, vol. 2, pp. 9-13, 1972.

[54] F. de Dombal, D. Leaper, J. Horrocks, J. Staniland, and A. McCann, "Human and Computer-Aided Diagnosis of Abdominal Pain: Further Report with Emphasis on Performance.," *British Medical Journal*, vol. 1, pp. 376-380, 1974.

[55] Z. Li and C. Chan, "Inferring pathways and networks with a Bayesian framework.," *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, vol. 18, pp. 746-748, 2004.

[56] N. I. R. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian network to analyze expression data.," *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[57] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor, "Using literature and data to learn Bayesian networks as clinical models of ovarian cancer.," *Artificial Intelligence in medicine*, vol. 30, pp. 257-281, 2004.

[58] I. Milho, A. Fred, J. Albano, N. Baptista, and P. Sena, "A user-friendly development tool for medical diagnosis based on Bayesian networks.," presented at Second International Conference on Enterprise Information Systems, 2000.

[59] A. Schloegl, M. Woertz, E. Trenker, P. Rappelsberger, and G. Pfurtscheller, "Automatic detection of sleep spindles using a Bayesian approach.," *Journal of Sleep Research*, vol. 7, pp. 242, 1998.

[60] A. Rechtschaffen, M. A. Gilliland, B. M. Bergmann, and J. B. Winter, "Physiological Correlates of Prolonged Sleep Deprivation in Rats," *Science*, vol. 221, pp. 182-184, 1983.

[61] P. Sykacek, S. Roberts, I. Rezek, A. Flexer, and G. Dorffner, "A probabilistic approach to high-resolution sleep analysis.," presented at ICANN (International Conference on Artificial Neural Networks), Vienna, Austria, 2001.

[62] R. B. Wittebrood, L. P. J. Veelenturf, and O. E. Herrmann, "Automatic classification of polysomnographic respiration signals," presented at ProRISC Workshop on Circuits, Systems and Signal Processing, Mierlo, the Netherlands, 1997.

[63] K. D. Nielsen, A. Kjaer, W. Jensen, T. Dyrby, L. Andreasen, J. Andersen, and S. Andreassen, "Causal probabilistic network and power spectral estimation used in sleep stage classification," *Methods of Information in Medicine*, vol. 36, pp. 345-348, 1997.

[64] I. Milho, A. Fred, J. Albano, N. Baptista, and P. Sena, "An Auxiliary System for Medical Diagnosis based on Bayesian Belief Networks," presented at Portuguese Conference on Pattern Recognition - RECPAD, Porto, Portugal, 2000.

[65] C. R. Twardy, A. E. Nicholson, K. B. Korb, and J. McNeil, "Data mining cardiovascular Bayesian networks.," Monash University, School of Computer Science & Software Engineering, Melbourne 165, 2004.

[66] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms.," *Neural Computation*, vol. 7, pp. 1895 - 1924, 1998.

[67] R. J. Davies, N. J. Ali, and J. R. Stradling, "Neck circumference and other clinical features in the diagnosis of the obstructive sleep apnoea syndrome.," *Thorax*, vol. 47, pp. 101-105, 1992.

[68] V. Hoffstein and S. Mateika, "Differences in abdominal and neck circumferences in patients with and without obstructive sleep apnoea.," *European Respiratory Journal*, vol. 5, pp. 377-381, 1992.

[69] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss.," *Machine Learning*, vol. 29, pp. 103-130, 1997.

[70] "American Sleep Disorders Association and Sleep Research Society EEG arousals: scoring rules and examples, a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association.," *Sleep*, vol. 15, pp. 173-184, 1992.