December 14, 2001

To the Graduate School:

This thesis entitled "Automated Diagnosis of Retinal Images Using Evidential Reasoning" and written by Reetal Pai is presented to the Graduate School of Clemson University. We recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Computer Engineering.

Dr. Adam W Hoover, Advisor

We have reviewed this thesis
and recommend its acceptance:

Dr. Robert J Schalkoff

Dr. John N Gowdy

Accepted for the Graduate School:

AUTOMATED DIAGNOSIS OF RETINAL IMAGES USING EVIDENTIAL REASONING

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Computer Engineering

---

by
Reetal Pai
December 20, 2001

Advisor: Dr. Adam W Hoover

ABSTRACT

In this project we research issues involved in the design of a robust automated medical diagnostic system. We take a decision theoretical approach to diagnosis. We construct a system to reason about the diagnosis of evidence obtained from a retinal image. The beliefs used to diagnose the evidence are given by an expert physician, in the form of probability values. In this study, we explore several theoretical questions. We consider the use of three mathematical formulations in the reasoning process. We explore the effects of three methods of evidence selection in the reasoning process. We consider the possible responses of the system to the diagnostic question. We examine the use of frequencies in place of expert given beliefs. Finally, we explore the case difficulty of the images contained in our test set.

This thesis is dedicated to my parents and
my sisters whose love and
encouragement have been my
strength.

VITA

Reetal Pai was born in Mysore, India. She completed her Bachelors in Electrical and Electronic Engineering at the University of Mysore, India in September 1998. She enrolled at Clemson University in August 1999. Her graduate research work focused on the development of an automated medical system to diagnose retinal images using evidential reasoning. She completed her Master of Science in Computer Engineering at Clemson University in December 2001.

ACKNOWLEDGEMENTS

I am very grateful to my advisor Dr. Adam Hoover. His support and motivation have made this work possible. I also want to thank him for continually believing in my capabilities and for all the opportunities he has given me to learn.

I thank our expert Dr. Michael Goldbaum for his guidance and helping us understand our problem domain better.

Several people have contributed to the collection of data and knowledge design used in this work. Images were collected by Stuart Burgess (MD), Howard Chen (MD), Michael Goldbaum (MD), Adam Hoover (Ph.D) and Valentina Kouznetsova (Ph.D). Beliefs were supplied by Michael Goldbaum (MD). Annotations were supplied by Michael Goldbaum (MD) and Howard Chen (MD). The lists of diagnoses and relevant manifestations were compiled by Michael Goldbaum (MD). The causal link network was designed by Michael Goldbaum (MD) and Edd Hunter (Ph.D).

Finally, I would like to thank my family and my friends who have always believed in me.

TABLE OF CONTENTS

Table of Contents (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

Table Page

List of Tables (Continued)

Table Page

CHAPTER 1
INTRODUCTION

In this work we construct an automated diagnostic system. The ultimate goal of automating the diagnostic process is to achieve repeatable performance superior to a human expert. For example, a system built by deDombal and his colleagues to diagnose acute abdominal pain [4, 5] performed with an overall diagnostic accuracy of 91.8%. This percentage was significantly higher than that of the most senior member of the clinical team to see each case (79.6%). However, modern automated systems only achieve superior performance when the diagnostic question is relatively simple, such as a single present/absent question. In the more general case, where the diagnosis may be one of tens or hundreds of possibilities, automated methods have yet to achieve performance comparable to human experts.

In recent times the term decision support system (DSS) has replaced the term expert system in the automated diagnostic literature [8]. The emphasis in the terms denotes a subtle change in the goal of the work. Instead of replacing the human expert, the system is intended to assist the expert. For example, a DSS could help a human expert save time in some cases, by eliminating some possibilities. A DSS could also review decisions and help eliminate potential errors. In the system built by deDombal [4, 5] it was observed that the clinician's own diagnostic accuracy improved while using the system. However, a recent study by Elstein et al.[2] provides contrasting evidence. They experimented with a system called ILIAD, developed at the University of Utah, to determine if it affected the diagnosis reached by users with varying levels of expertise. They found that in 70% of the cases the DSS did not change the diagnoses concluded by the human experts.

In this work we pursue complete automation in a complex domain, recognizing that we are as yet unable to achieve performance comparable to a human expert. Our short term goal is to improve upon the potential for complete automation.

Automated diagnostic systems may be classified into three categories: case based expert systems, model based expert systems, and evidential based expert systems. A case based expert system or case based reasoning system is one where the system solves a problem by trying to find previous cases matching the present one. This solution can be seen as one driven by analogy. A model based reasoning system summarizes the underlying mechanisms in the system in a model and uses the model to solve a problem. An evidential reasoning system reasons from the evidence provided to it. In this study we will look at the working of an evidential reasoning system.

Most of the early evidential diagnostic systems were heuristic in nature. The process of diagnosis was guided by a set of rules formulated to model the problem domain. One of the earliest heuristic diagnostic systems built was MYCIN [20], a probabilistic expert system. Built in the 1970's at Stanford, MYCIN diagnoses and recommends treatment for bacterial infections of the blood. It approximates the physician's diagnosis in cases with partial information. MYCIN is coded in LISP. The knowledge about the problem domain is represented in IF-THEN form. A certainty factor is associated with each of the rules. It is a goal directed system that uses the backward chaining procedure. MYCIN is unable to separate the knowledge contained in the database from the results of the rule based chaining process. This leads to inconsistencies in the database. There are other expert systems, like EMYCIN, PUFF and NEOMYCIN [1], that were developed along the lines of MYCIN.

In recent times researchers have been exploring the use of decision theory as a framework for knowledge representation and inference. Decision theory allows us to describe evidence and reason about a decision. Probabilities are used to describe a person's beliefs. An expert system built using decision theory is also referred to as normative expert system. An example of a normative expert system is Pathfinder [11]. It assists surgical pathologists with the diagnosis of lymph-node disease. Pathfinder has knowledge of over 60 diseases and 100 disease findings. It uses similarity networks to build the belief networks, and partitions to assess the probabilities associated with the belief networks.

In this work we take a decision theoretical approach to diagnosis. We construct

a system to reason about the diagnosis of evidence obtained from a retinal image. The beliefs used to diagnose the evidence are given by an expert physician, in the form of probability values. In this study, we explore several theoretical questions. We consider the use of three mathematical formulations in the reasoning process. We explore the effects of three methods of evidence selection in the reasoning process. We consider the possible responses of the system to the diagnostic question. We examine the use of frequencies in place of expert given beliefs. Finally, we explore the case difficulty of the images contained in our test set.

STARE

This thesis is part of the STARE project. The acronym STARE refers to STructured Analysis of the REtina. The ultimate goal of the STARE project is to create a system that can automatically diagnose a retinal image. Additional goals for the project include the capabilities to measure key features, to annotate the image contents, and to compare manifestations in images of a subject taken at different times.

The STARE project was conceived and initiated in 1975 by Michael Goldbaum, M.D., at the University of California, San Diego. Since then, over thirty people with backgrounds in medicine, science and engineering have contributed to the project. Images and clinical data have been provided by the Shiley Eye Center at the University of California, San Diego, and by the Veterans Administration Medical Center in San Diego. The STARE project, including the work presented in this thesis, is currently funded by the National Institutes of Health (U.S.A.).

An opthalmologist is a medical doctor who specializes in the structure, function, and diseases of the human eye. During a clinical examination, an opthalmologist notes findings that are visible in the eyes of the subject. The ophthalmologist uses these findings to reason about the health of the subject, and to conclude a diagnosis in the case of ill health.

The process of retinal imaging provides a permanent record of the patient's eye that can be consulted at any time for diagnostic reasoning. A retinal image is acquired using an optical camera to see through the pupil of the eye to the rear inner surface

of the eyeball. A retinal image generally shows the optic nerve, fovea, surrounding vessels, and the retinal layer.

Figure 1.1 shows an example retinal image. The optic nerve is visible on the far left side of the image. In full color, the optic nerve appears bright red or white. The blood vessels (arteries and veins) in the eye emanate from this spot. The optic nerve supplies the retina with blood, and conducts visual stimuli to the brain. The image in Figure 1.1 shows two findings that are important for diagnostic reasoning. First, there are several lesions. These are the bright spots (yellowish in full color) in the center of the image. Second, the blood vessels are very tortuous. This refers to their twisted and winding appearance. The lesions and the tortuous blood vessels are evidence of some abnormality. An ophthalmologist uses these findings from the observation of the eye (or the retinal image) to reason about the health of the patient. The physician then reaches a conclusion regarding the reason for the anomalies. In our example, the findings of lesions and tortuous blood vessels might lead to a diagnosis of *background diabetic retinopathy*.

Figure 1.2 illustrates the complete process of diagnosis as undertaken in the STARE project. The three key steps are retinal imaging, image annotation, and evidential-based reasoning. Other parts of the STARE project have concentrated on the image annotation process. A few examples are [9], [12], [13] and [14]. This thesis concentrates on the evidential-based reasoning process. We assume that the information observable in the retinal images is available as input. This information is identical to the findings that an ophthalmologist would note during a clinical examination of the eye. The main output of our system is a diagnosis that is formulated to mimic the conclusion that an ophthalmologist would reach about the health of the patient.

System Description

Our system is capable of making thirteen diagnoses. The thirteen diseases and their abbreviations used in this thesis are listed in Table 1.1.

The presence of each of these diseases causes abnormalities to appear in the eye.

Figure 1.1 An example of a retinal image used for diagnosis.

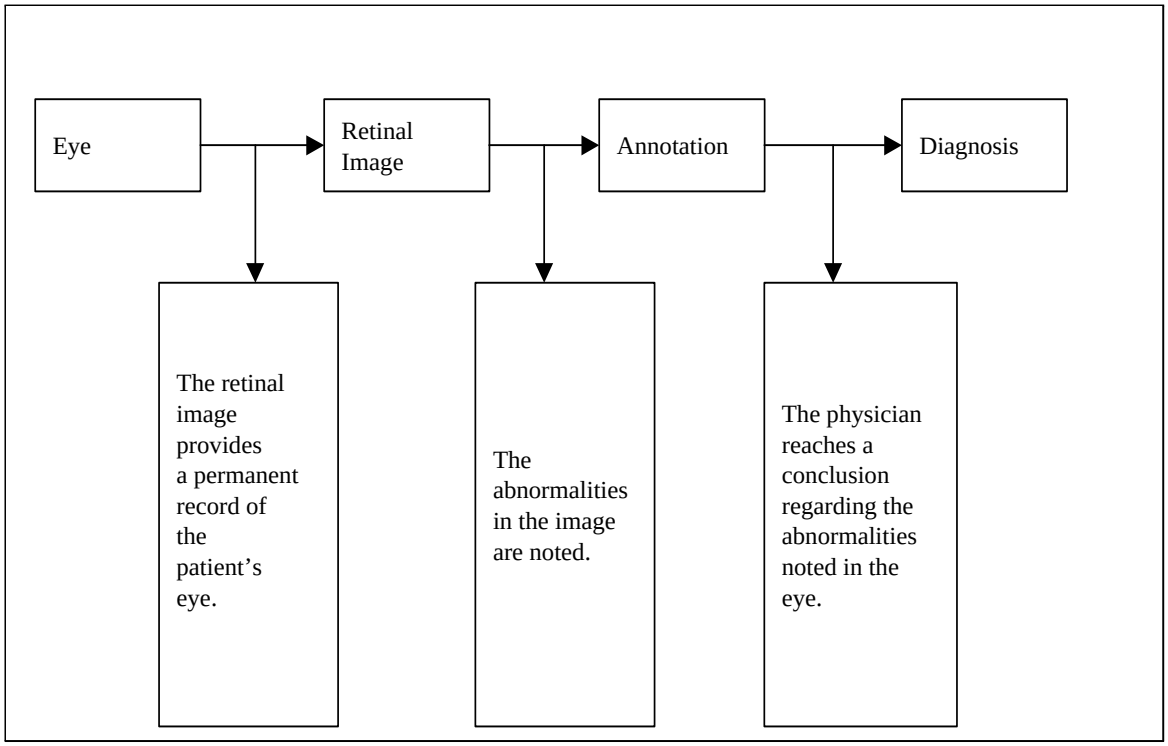

Figure 1.2 The process of retinal diagnosis.

These abnormalities are referred to as manifestations. In our system we identified thirty nine manifestations relevant to the thirteen diagnoses that are being considered. Each of the manifestations is characterized by a set of discrete states. The number of states can vary from two to seven depending on the manifestation. The first state signifies the normal or absent state of the manifestation. The remaining one to six states signify abnormal or present states of the manifestation in varying degrees of severity. Table 1.2 lists two example manifestations, *RPED* and *artery narrowing*, to illustrate how the states for a manifestation can vary. All thirty nine manifestations in all possible states are listed in Appendix A.

The reasoning process of the physician is replicated by mathematical formulations. In a clinical setup, the physician reasons about the abnormalities seen in the image based on his past experiences. In order to mimic this reasoning process, the physician's knowledge about the relationships between the various diseases and manifestations is modeled. This knowledge is tabulated in the form of a belief table. The entries in the belief table are probability values, of the form $P(M = S/D)$. Each entry specifies the probability of the presence of a manifestation M in a state S given the presence of a diagnosis D. The interpretation of a probability as a degree of belief is referred to as the subjective or Bayesian interpretation. The mathematical formulations we explore are therefore based upon the Bayesian approach to probability.

The process of observing the abnormalities in the retinal image is actually the process of observing each manifestation's state. This identification of each manifestation's state is referred to as the process of annotation. The ultimate goal of the STARE project includes the capability to produce image annotations automatically. In this thesis we use annotations that have been hand drafted by an expert physician. These annotations allow us to focus on the diagnostic portion of the problem. They also allow us to assume that we have an annotation process that works as well as can be expected, so that any diagnostic error can be attributed entirely to the reasoning process.

Figure 1.3 overviews our approach to the reasoning process. The first step is to decide on the evidence to be used. This refers to the choice in manifestations to be

	Diagnosis	Abbreviation
1	Emboli	Emboli
2	Branch retinal artery occlusion	BRAO
3	Cilio-retinal artery occlusion	CRAO
4	Branch retinal vein occlusion	BRVO
5	Central retinal vein occlusion	CRVO
6	Hemi-central retinal vein occlusion	Hemi-CRVO
7	Background diabetic retinopathy	BDR
8	Proliferative diabetic retinopathy	PDR
9	Arteriosclerotic retinopathy	ASR
10	Hypertensive retinopathy	HTR
11	Coats' disease	Coats'
12	Macroaneurism	Macroaneurism
13	Choroidal neovascularization	CNV

Table 1.1 List of the diagnoses that are recognized by our system and their abbreviations used in this thesis.

Manifestation	States
RPED	Absent
	Present
Artery narrow	Normal
	Focal,one or more segments
	Moderate,branch or single
	Moderate,global
	Extreme,global

Table 1.2 Two manifestations and their possible states.

considered during reasoning. For example, we can use only those manifestations that have been noted as present in some state in the annotation, or we could use both the presence and absence of manifestations as diagnostic indicators. The topic of evidence selection is discussed in Chapter 2.

Given a set of evidence, we explore three mathematical formulations to reach a diagnostic conclusion. These three formulations are based upon the Bayesian probabilistic approach. We discuss two known formulations, their drawbacks, and present a third novel formulation. Chapter 3 discusses these three formulations.

The simplest output of our system is a single diagnosis from the set of thirteen diagnoses considered. However, in some cases there may be multiple diagnoses present, or an unknown disease, or no disease at all. The topic of system response is discussed

in Chapter 4.

One way to obtain knowledge about a problem domain is to gather beliefs from a physician. Obtaining these beliefs is time consuming, and potentially prone to errors. Chapter 5 explores the alternative use of frequencies in place of expert beliefs.

Throughout this work we have assumed that the test images used for evaluation are "good" examples of the thirteen diagnoses in our set. In Chapter 6 we explore the classification of the "goodness" of our examples, and examine the performance of our system based upon case goodness.

Related Work

Bayesian belief networks are being increasingly used as a knowledge representation for reasoning under uncertainty. In 1990 Heckerman et al.[10], designed a normative expert system, Pathfinder, which assists surgical pathologists with the diagnosis of lymph node disease. The issues explored are the graphical representations of conditional independence and the use of similarity networks and partition representations. In 1995 Pradhan et al.[19], examined the effect of imprecision in the probabilities on diagnostic performance in a Bayesian belief network. They examined the effect of varying the mapping from qualitative frequency weights into numerical probabilities and the effect of simplifying quaternary domains for manifestations to binary domains on diagnostic performance. The noisy MAX and noisy OR formulations were used in diagnosis.

This thesis applies a Bayesian belief network to the domain of retinal images. To our knowledge, this is the first time that this medical domain has been explored in automated diagnosis. We describe a novel formulation that outperforms the noisy MAX formulation in our system. In addition to examining the diagnostic performance based on these formulations, we also examine the effect of evidence selection, the use of frequencies instead of beliefs and the system performance on diagnoses unknown to the system.

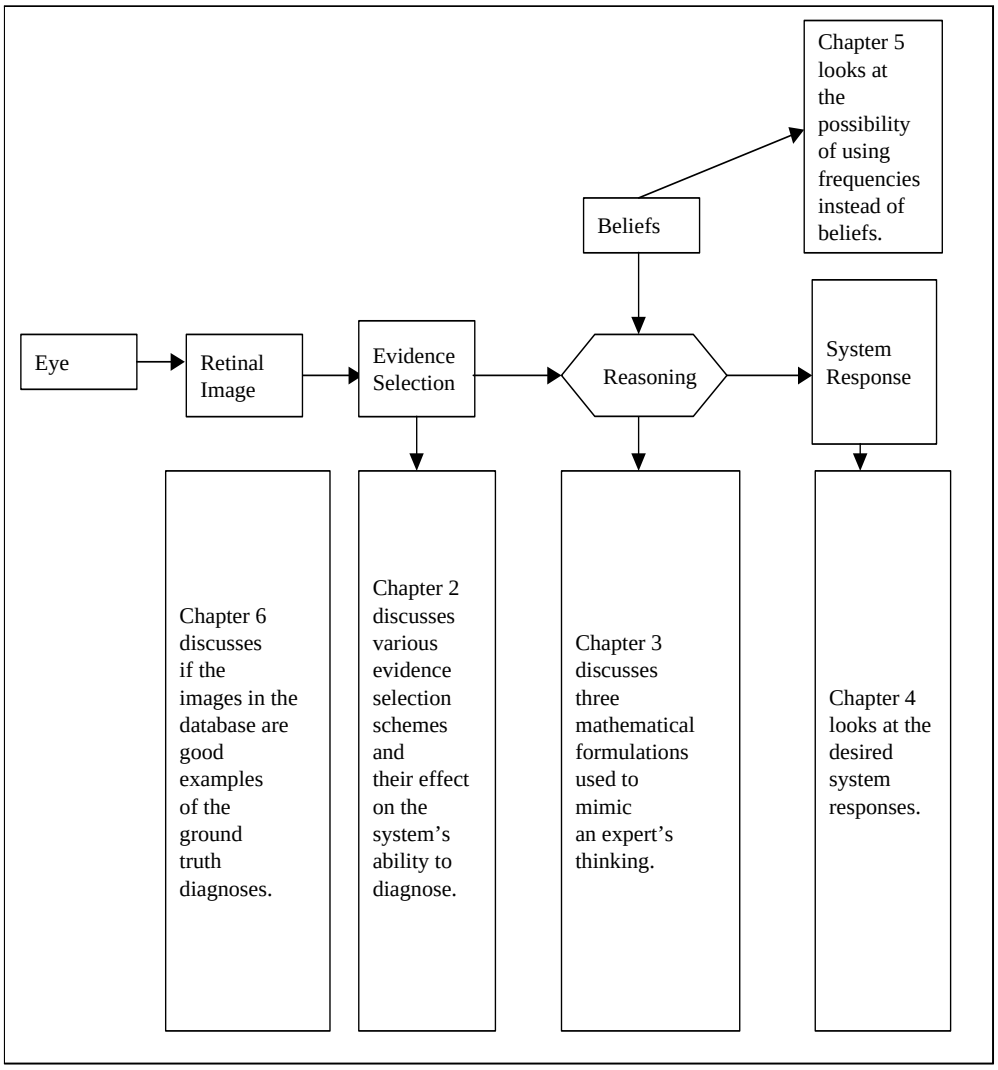

Figure 1.3 Organization of this thesis.

CHAPTER 2

EVIDENCE SELECTION

Each disease can cause a particular set of manifestations to appear. The evidence provided by all the manifestations observed, taken together, should indicate a unique diagnosis. The problem of evidence selection is to decide which set of manifestations, including those that are absent, should be used in the reasoning process.

Each manifestation is associated with a set of diagnoses that can cause that manifestation. These associations can be represented using links. In our system the number of links for a manifestation varies from one to eleven. Table 2.1 shows two manifestations and their linked diagnoses. The links for all thirty nine manifestations are listed in Appendix A.

Manifestation	Links
RPED	CNV
CME	CNV, NPDR, PDR, BRVO, CRVO,
	Hemi-CRVO, Coats, Macroaneurism

Table 2.1 Two manifestations and their causal diseases.

The links between the manifestations and the diagnoses can be represented graphically. Each of the thirty nine manifestations and the thirteen diagnoses is represented by a node. Each link is represented using an arrow. The direction of the arrow shows the direction of the dependence. This representation is called a directed graph. In the domain of reasoning it is also called an influence diagram, or a causal network.

Figure 2.1 shows the influence diagram for one of the manifestations in our system, *ghost blood vessels*. The ellipse represents the manifestation, and each of the rectangles represents a diagnosis. In Figure 2.1, the two diagnoses *emboli* and *CNV* are not connected to the manifestation. This implies that the presence of either of these two diagnoses does not cause the manifestation *ghost blood vessels*.

Figure 2.2 shows the influence diagram for all thirteen diagnoses and all thirty nine manifestations in our system. The complexity of retinal diagnosis is apparent

in the density of the links[1]. This complexity is largely due to the fact that a single manifestation may be caused by multiple different diagnoses.

In Figure 2.1 and in Figure 2.2, the states of the manifestations have not been represented. Each of the manifestations is assumed to be either absent (normal), or present in one of its abnormal states.

We constructed our influence diagram using the assumption of causal independence. This assumption implies that no manifestation can cause another manifestation; only a diagnosis can cause a manifestation. This assumption simplifies the design of the automated system.

Additional diagnoses and manifestations can be added to our system so long as the assumption of causal independence is held. The methods we describe herein should scale nicely to hundreds of diagnoses, although we have not tested this hypothesis.

Choice of Evidence

In any image, each manifestation may be said to have a specific state, whether absent or present. The entire set of specific states of manifestations is referred to as the evidence. The evidence comprises the input to the system. The evidence presented to the system is specific to each individual image.

We consider the evidence in three different contexts: *all*, *linked* and *annotated*. Each context defines a different subset of the total evidence. In the context of *all* evidence, a specific state for every manifestation is considered during the reasoning process. In the context of *linked* evidence, only those manifestations which might be caused by each disease contribute to the consideration of the presence of the disease. The fact that an unrelated manifestation is absent is not considered. In the context of *annotated* evidence, only those manifestations which might be caused by each disease, and are annotated as present, contribute to the consideration of the presence of the disease. The fact that an unrelated or related manifestation is absent is not considered.

[1]There are hundreds of unique diagnoses of the retina, so that even our influence diagram for thirteen diagnoses does little justice to the complexity of the problem.

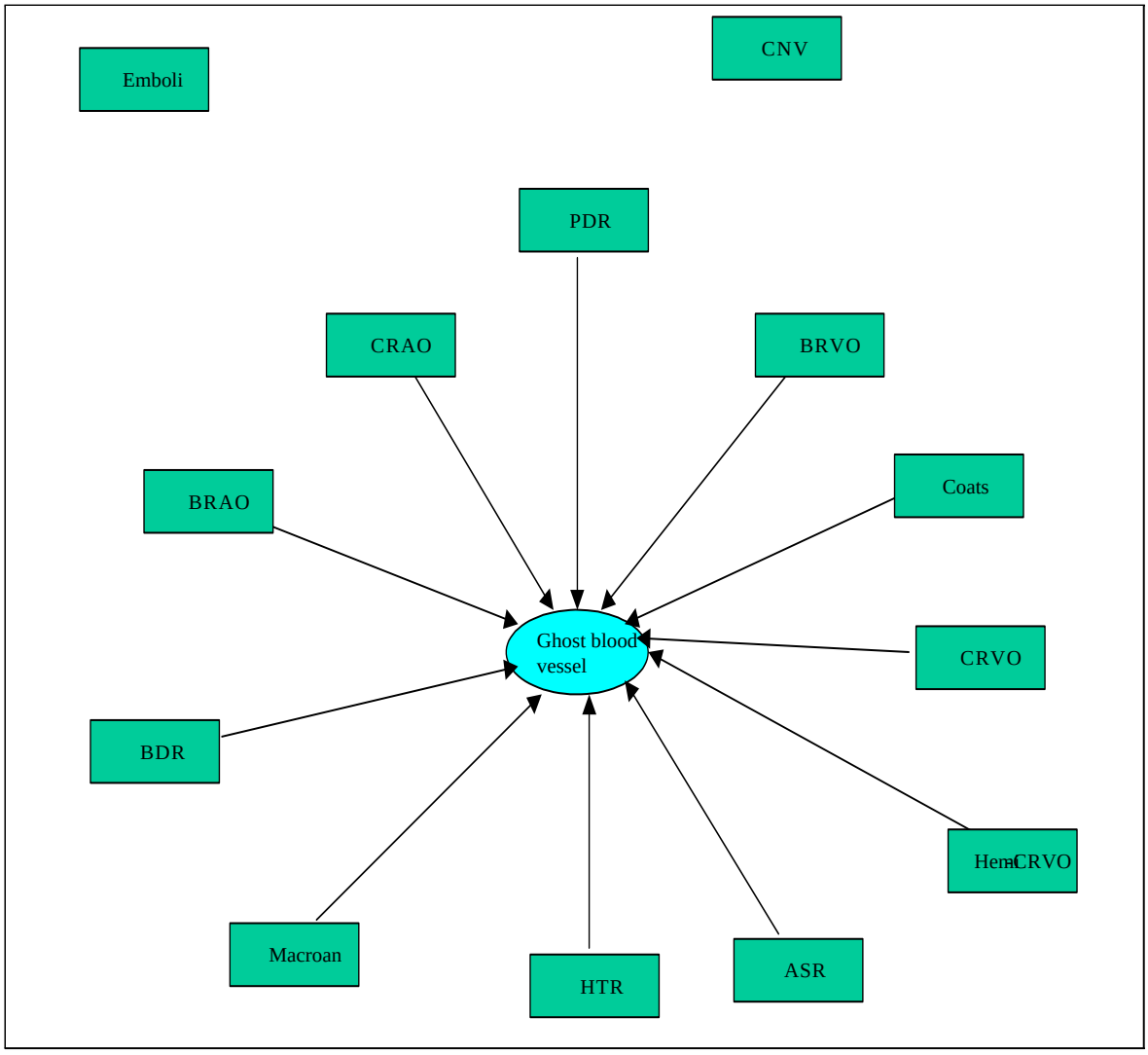

Figure 2.1 Influence diagram for a single manifestation (*ghost blood vessels*).

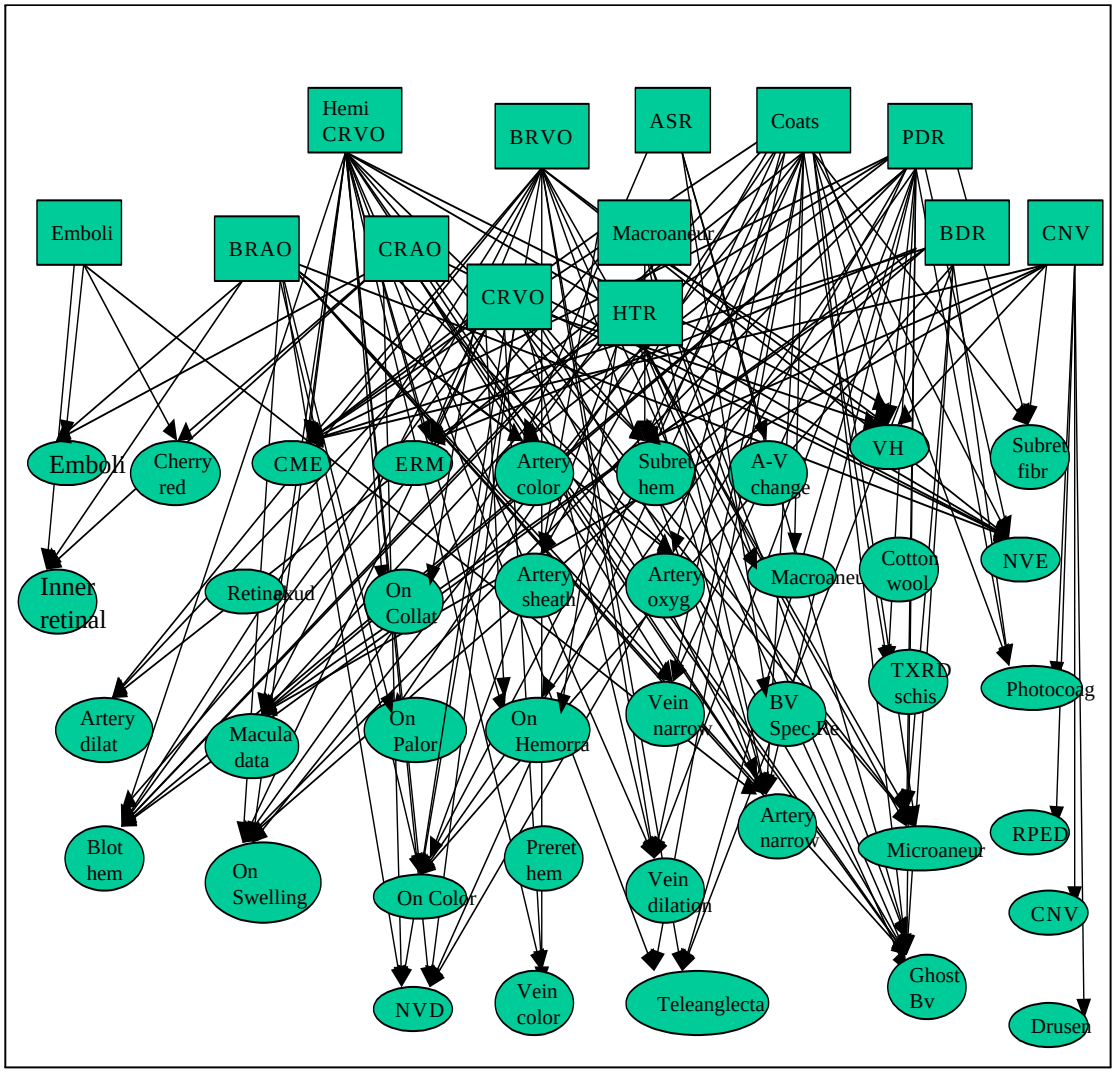

Figure 2.2 Influence diagram for the entire network. The ovals represent the manifestations and the squares represent the diagnoses. The states of the manifestations are not pictured in this diagram.

We explore the contexts of linked and annotated evidence to determine if a narrower view of the available evidence improves diagnosis. We hypothesize that in some cases, considering all the evidence "overloads" the system. In some cases a narrower focus may improve the overall diagnostic ability.

Figure 2.3 demonstrates the three contexts of evidence. In this example, the manifestations *inner retinal infract, cherry red spot*, *artery color* and *cotton wool spots* are observed to be present. Two of the relevant diagnoses (*emboli and ASR*) and their links are pictured. Using the context of all evidence, all thirty nine manifestations are considered while diagnosing the image as having either *emboli* or *ASR*. Notice that thirty two of the manifestations are not linked to either diagnosis, so that their usefulness in distinguishing between the two given diagnoses is questionable. Using the context of linked evidence, only the manifestations *cotton wool spot*, *inner retinal infract*, *cherry red spot* and *emboli* are considered to find the belief in the presence of *emboli*. Similarly only the manifestations *artery color*, *bv specular reflex*, *av change* and *cotton wool spot* are used to find the belief in the presence of the diagnosis *ASR*. We observe that although the manifestations *cotton wool*, *emboli*, *bv specular reflex* and *av change* are not present in the given image they will still contribute to the reasoning process.

Using the context of annotated evidence, only the manifestations linked to each diagnosis, and observed to be present, are considered to determine the belief in the presence of the diagnosis. In our example, only the manifestations *cherry red spot* and *inner retinal infract* are used to diagnose the presence of *emboli*. This is because these are the only manifestations that are present and are also linked to the diagnosis *emboli*. Similarly only the manifestation *artery color* is used to diagnose the presence of *ASR*. This is because it is the only manifestation that is present and is also linked to the diagnosis *ASR*.

Throughout this work we consider all three contexts of evidence. Chapter 4 presents the specific results of how well our system performs using each context.

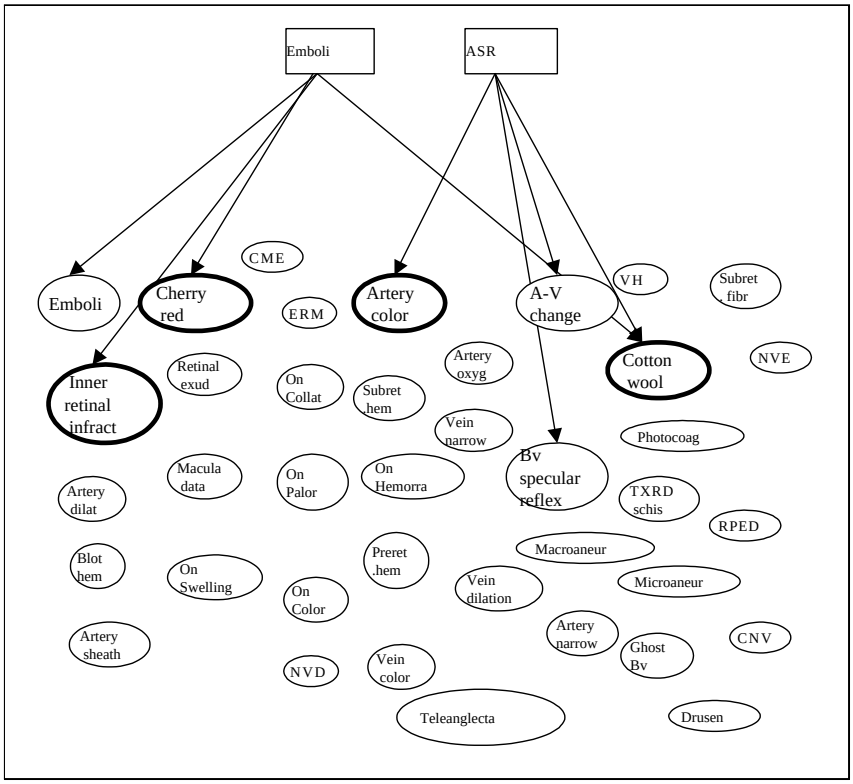

Figure 2.3 An example influence diagram. The manifestations *inner retinal infract*, *cherry red*, *artery color* and *cotton wool spots* are observed as present in some state. The evidence can be considered in three contexts: all manifestations, only those linked, or only those present.

CHAPTER 3
FORMULATIONS FOR DIAGNOSIS

This chapter discusses three mathematical formulations to mimic reasoning. The reasoning problem involves examining the evidence provided in order to conclude a diagnosis. The process of diagnosis can be formalized as

$$D_{out} = F(P(M = S/D), M_{input}) \tag{3.1}$$

where D_{out} refers to the diagnostic output of the system, or system diagnosis. The knowledge used during reasoning is given by probability values $P(M = S/D)$. Each $P(M = S/D)$ is the probability of a manifestation M being present in state S given a diagnosis D. The evidence used during reasoning is given by an annotation file M_{input}, which describes the state of each manifestation in a given image. The function F denotes the mathematical formulation used to process the evidence to conclude the output diagnosis.

Using Equation 3.1, the diagnosis in part depends on the probability values given in a belief table[2]. The use of probabilities to express beliefs is a Bayesian technique. Bayesian techniques are popular in the field of automated medical diagnosis (see for example [6, 10, 11, 18]). In the next section we look at the Bayesian approach and its applicability to our system.

Baye's Rule

The Bayesian treatment of beliefs is based upon conditional probability. Whenever a statement of the probability of an event is given, denoted $P(A)$, it is conditioned by other known factors. For example, the belief that "Given the event B, the probability of the event A is x" is written as $P(A/B) = x$.

In the formulation of our diagnostic system the values in the belief table are given by $P(M = S/D)$. They represent the probability of manifestation M being in state

[2]For now, it is assumed that the belief table is supplied by a human expert. Chapter 5 discusses this issue in detail.

S given the presence of diagnosis D. Since the beliefs in our system are conditional probabilities, we explore the application of Bayes' Rule to the reasoning process.

Bayes' Rule is based on conditional probability and is derived from

$$P(A/B) * P(B) = P(A, B) \tag{3.2}$$

where P(A,B) is the probability of the joint event A and B. By symmetrically substituting B and A in Equation 3.2 we may write

$$P(A, B) = P(A/B) * P(B) = P(B/A) * P(A) \tag{3.3}$$

Solving for one of the conditional probabilities in terms of the other yields Bayes' Rule

$$P(B/A) = \frac{P(A/B) * P(B)}{P(A)} \tag{3.4}$$

By letting B represent a diagnosis D and A represent a manifestation M, Bayes' Rule can be applied to our system as follows:

$$P(D/M) = \frac{P(M/D) * P(D)}{P(M)} \tag{3.5}$$

The term $P(M/D)$ represents the belief about the presence of the manifestation M given the disease D, and is known a priori. The term $P(D)$ represents the belief about the presence of the disease in the population, and is known a priori. The term $P(M)$ represents the belief about the presence of the manifestation in the population, and is known a priori. Equation 3.5 concludes the presence of the disease given the manifestation, represented by the term $P(D/M)$.

Equation 3.5 considers only one diagnosis. The formulation can be expanded to consider multiple diagnoses as follows:

$$P(D_i/M) = \frac{P(M/D_i) * P(D_i)}{\sum_{x=1}^{I} P(M/D_x)} \tag{3.6}$$

where $i = 1...I$ refers to each one of the total I diagnoses being considered. Note that there is a different and unique output belief $P(D_i/M)$ in the presence of each diagnosis. Since multiple diagnoses are being considered the total probability of the presence of manifestation M is given by the sum of the probabilities of its presence

given all the I different diagnoses. This sum of probabilities is denoted in Equation 3.6 by the denominator term.

Equation 3.6 considers only one manifestation as evidence. The formulation can be expanded to include multiple manifestations as evidence.

$$P(D_i/M_j) = \frac{\left[\prod_{j=1}^{J} P(M_j/D_i)\right] * P(D_i)}{\prod_{j=1}^{J} \sum_{x=1}^{I} P(M_j/D_x)} \tag{3.7}$$

where $j = 1...J$ refers to each of the total J manifestations being considered. Since multiple manifestations are being considered the total probability of the presence of the set of manifestations M_j is given by the product of the probabilities of each of the individual manifestations M_j. Both the numerator and denominator terms in Equation 3.6 contain products.

Finally, Equation 3.7 assumes that each manifestation can be present in only two states, absent or present. In our system each manifestation can take on an arbitrary number of discrete states. For each manifestation j let k represent the state of the manifestation. Then the formulation can be written as follows:

$$P(D_i/M_j = S_k) = \frac{\left[\prod_{j=1}^{J} P(M_j = S_k/D_i)\right] * P(D_i)}{\prod_{j=1}^{J} \sum_{x=1}^{I} P(M_j = S_k/D_x)} \tag{3.8}$$

where $k = 1...K_j$ refers to each of the total K states of the j^{th} manifestation. This formulation can be interpreted to mean that the probability of a diagnosis being present given a manifestation in a particular state is equal to the probability of the manifestation in that particular state given the presence of the diagnosis multiplied by the apriori probability of the diagnosis and scaled by the overall probability of the manifestation being in that particular state given all of the I diagnoses.

Equation 3.8, based upon Bayes' Rule, provides a method to compute a belief in each of the I diagnoses. The I values represent the system's beliefs in each of the I diagnoses, where the largest belief could be chosen to indicate the system diagnosis. Strict application of Bayes' Rule as derived above is problematic. We can identify three problems:

1. The belief table typically contains many zeros. This is because the physician believes that certain manifestations would never appear given a particular disease.

Equation 3.8 is a multiplicative combination of all the evidence. If the evidence presented to the system has even one manifestation with a zero conditional probability, it can invalidate the evidence provided by the other manifestations. For example, suppose that it is believed that *emboli* never causes manifestation M_3, so that $P(M_3/emboli) = 0$. If the evidence considered includes manifestation M_3, then $P(emboli/M_{input}) = 0$. This problem has also been referred to in the work on Pathfinder. Heckerman states that "... the expert was too cavalier when assigning 0 to the probability of many events. In preliminary evaluations of Pathfinder, we found that over 10 percent of the cases were diagnosed incorrectly, because the correct disease was ruled out by a feature that was unlikely (but not impossible) to be seen in that disease...."[10, pg. 106]. To overcome the presence of the zeros in the belief table we substitute all zeros with a very small non-zero number.

2. A patient may have multiple diseases or conditions. Roughly one-third of the images in our test set have more than one ground truth diagnosis. This implies that multiple diagnoses could influence the manifestations present in a retinal image. Equation 3.8 does not account for the influence of multiple diagnoses on the presence of a manifestation. We explore using the noisy MAX formulation [19] to determine the probability values for a manifestation given multiple diagnoses. The noisy MAX formulation allows the propagation of the influence of multiple diagnoses on a particular manifestation. The noisy MAX formulation also allows us to find the probability values for the presence of two or more diseases. We will look at this formulation in detail in the next section.

3. The belief table consists of a large quantity of values (1,469 in our system) whose relative ratios are assumed to be globally uniform. For example, if $P(M_{13}/D_{12}) = 0.3$ and $P(M_{36}/D_4) = 0.3$, then these beliefs are assumed to be equal. However, as an expert fills such a table, we believe that the relativity varies. In the work on Pathfinder they found that " .... the expert was not comfortable with many of the probability assessments that he provided ... Specifically, he assessed the probability matrix required by the simple Baye's model, p (feature/disease, x) for all diseases and features, by fixing a disease and assessing probabilities across all features.

In an analysis that followed the completion of the knowledge base, we found that he strongly preferred making assessments by fixing a feature and assessing probabilities of that feature across all diseases...." [10, pg. 106] . We explore a novel formulation, which we call normalized sums, to disregard the assumption of global relativity in the belief table. This formulation is explained after the description of the noisy MAX formulation.

Noisy MAX

The presence of the manifestations in an image need not be a result of just one disease. The manifestations could be the result of multiple diseases. Each of the values in the belief table is the probability of a manifestation given a single disease. In accounting for the presence of multiple diagnoses we need to find the combined probability values for all the manifestations in all the possible states, given combinations of diseases.

There is a disjunctive interaction between the manifestations and the multiple diagnoses causing the manifestations. Disjunctive interaction occurs when any member of a set of conditions is likely to cause a certain event and this likelihood does not diminish when several of these conditions prevail simultaneously. For example if a retinal image has diseases *emboli* and *BRAO*, then the likelihood of observing the manifestation *emboli* is increased. This is because both the diseases are causes for the presence of the manifestation *emboli*. If the same retinal image also has diagnosis *CNV*, a diagnosis that does not cause the manifestation *emboli*, then the probability of observing the manifestation *emboli* is not altered. In [3] and [18] a mathematical model called noisy OR is described to compute disjunctive interactions.

The noisy OR is a model of probabilistic causal influence between a binary effect variable and a set of binary variables that represent its causes. In our problem the effect variable, the manifestation, is not a binary variable. Therefore we use the noisy MAX formulation that is a generalized version of the noisy OR. The noisy MAX formulation accounts for the multiple states of the manifestations. This formulation has been looked at in detail in [19]. This formulation is also used by the commercial

software package Hugin [16] to propagate beliefs in a causal network.

In this work the noisy MAX formulation is used to evaluate the probability of a manifestation in a particular state under the influence of multiple diagnoses. These combined probabilities are denoted as the noisy MAX probabilities. The noisy MAX formulation for two diagnoses D_a and D_b is

$$P(M = S_k/D_a\&D_b) = \left[\sum_{n=0}^{k} P(M = S_n/D_a) * \sum_{n=0}^{k} P(M = S_n/D_b)\right] - A \qquad (3.9)$$

$$A = P(M = S_{k-1}/D_a\&D_b)$$

where k represents the state of the manifestation for which the combined probability is being found. The terms $\sum_{n=0}^{k} P(M = S_n/D_b)$ and $\sum_{n=0}^{k} P(M = S_n/D_a)$ represent the sums of the probabilities for a range of states. These terms require that the states be ordered in increasing severity, from normal (absent) to the worst case of abnormal (present).

The noisy MAX formulation is recursive. The term A in Equation 3.9 represents $P(M = S_{k-1}/D_a\&D_b)$, the joint probability of the manifestation in its next less severe state. For the least severe case, when $k = 0$, the term $A = 0$.

In order to calculate the noisy MAX probabilities we must consider all the possible combinations of diseases. In our system we consider thirteen diagnoses, so that in theory there are $\sum_{x=1}^{13} x!$ disease combinations. However, not all of the diagnoses in our system can occur with each other. The diagnoses that are mutually exclusive are shown in Table 3.1

By eliminating combinations containing mutually exclusive diagnoses, we obtain 867 possible diagnosis combinations. The maximum number of diagnoses in any combination under these conditions was found to be seven for our system. Equation 3.9 can be expanded to compute the joint probability of a manifestation in a particular state given an arbitrary number C of diagnoses as follows:

$$P(M = S_k/\{D\}) = \left[\prod_{c=1}^{C} \sum_{n=0}^{k} P(M = S_n/D_c)\right] - A \qquad (3.10)$$

	Diagnosis Pairs
1	BRAO and CRAO
2	BRVO and CRVO
3	BRVO and Hemi-CRVO
4	CRVO and Hemi-CRVO
5	BDR and PDR
6	Coats and Macroaneurism

Table 3.1 List of incompatible diagnosis pairs.

$$A = P(M = S_{k-1}/\{D\})$$

where $\{D\}$ represents the set of C diagnoses in the combination, and D_c represents each member of the set $\{D\}$.

Equation 3.10 yields 867 probability values indicating the beliefs in the 867 diagnosis combinations. These output beliefs could be interpreted individually, but the large quantity of numbers clouds distinction. Instead, we combine these 867 values into thirteen values, where each value represents the belief in a single diagnosis. This combination is done by summation as follows

$$P(D_i/M = S_k) \sim \sum_{x=1}^{867} \begin{array}{ll} P(\{D\}_x/M = S_k) & \text{if } D_i \in \{D\}_x \\ 0 & \text{if } D_i \notin \{D\}_x \end{array} \tag{3.11}$$

where D_i is each of the thirteen singular diagnoses and D_x is each of the 867 sets of combinative diagnoses. The resulting sum is no longer a probability which we emphasize by using the similarity operator in Equation 3.11. However, the concept is similar, so we maintain the notation $P(M = S_k/D_i)$ to represent the system output values.

The noisy MAX formulation is computationally expensive. Since it involves computing all the possible diagnosis combinations, the computational complexity grows exponentially as the number of diagnoses is increased. It is also unclear to us how to improve upon our output combination scheme. In this scheme the final belief in the presence of a diagnosis is increased if it occurs in many diagnosis combinations.

Normalized Sums

The values in the belief table are assumed to be globally relative. As an expert fills a belief table, it is assumed that the relative ratios between all values are consistent. Given the quantity of values in a belief table, we question this assumption. It can take hours for an expert to fill a belief table of the size used in our system. In addition, an expert is typically not used to thinking about all possible manifestations and causes in a single global scale. We develop the following approach, which we call normalized sums, to disregard the assumption of global relativity in the belief table.

There are two steps involved in calculating the normalized sums values. In the first step the probability values given in the belief table are normalized. The normalization is done independently for each state of each manifestation across the set of diagnoses:

$$\overline{P(M = S/D_i)} = ||P(M = S/D_i)|| \tag{3.12}$$

This normalization preserves the relativity across all diagnoses for each manifestation state. We believe this most closely resembles how the expert perceives the importance of relativity when filling the belief table. The belief values are normalized on a scale from zero to one. The normalization is done as follows:

$$||P(M = S/D_i)|| = \frac{P(M = S/D_i) - \min_i (P(M = S/D_i))}{\max_i (P(M = S/D_i)) - \min_i (P(M = S/D_i))} \tag{3.13}$$

where $\max_i (P(M = S/D_i))$ is equal to the maximum of the belief values for manifestation M in state S across all i diagnoses. Similarly $\min_i (P(M = S/D_i))$ is equal to the minimum of the belief values for manifestation M in state S across all i diagnoses.

The following example demonstrates the normalization. In this example there is a manifestation M which has two states S_1 and S_2. The original belief table for this manifestation is as follows:

Manif	Diagnoses												
M	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
S_1	0.3	1	1	1	0.75	0.8	1	1	1	1	1	1	1
S_2	0.7	0	0	0	0.25	0.2	0	0	0	0	0	0	0

Consider the manifestation M in state S_1. Using Equations 3.12 and 3.13 the normalized probabilities $\overline{P(M = S_1/D_i)}$ are computed. For $M = S_1$, the maximum

$P(M = S_1/D_i) = 1$ and the minimum $P(M = S_1/D_i) = 0.3$. Using Equations 3.12 and 3.13 the normalized values $||\ P(M = S_1/D_i)||$ become

Manif	Diagnoses												
M	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
S_1	0	1	1	1	0.64	0.71	1	1	1	1	1	1	1

This procedure is repeated for the manifestation in state S_2. The new belief table for manifestation M now looks as follows:

Manif	Diagnoses												
M	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
S_1	0	1	1	1	0.64	0.71	1	1	1	1	1	1	1
S_2	1	0	0	0	0.36	0.29	0	0	0	0	0	0	0

The normalized values are no longer strictly considered as probabilities. Instead, we view them as indicators or contributors towards a causal diagnosis. In this sense, the more contributors found in a given image, the higher the belief in a particular diagnosis. This concept can be formalized as follows:

$$P(D_i/M_j = S) \backsim \sum_{j=1}^{J} \overline{P(M_j = S/D_i)} \tag{3.14}$$

where $P(D_i/M_j = S)$ is the system output value for each diagnosis. This term is no longer a probability, which we emphasize by using the similarity operator in Equation 3.14. However, we maintain the notation $P(D_i/M_j = S)$ to indicate that the result is similar to the belief in diagnosis given a set of evidence.

The following example demonstrates the formulation. In this example there are two manifestations M_1 and M_2, both of which are in state S_2. The belief table for the two manifestations is as follows:

Manifs	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
$M_1 = S_2$	0.4	0	0	0	0.1	0.5	0	0	0	0	0	0	0
$M_2 = S_2$	0.8	0	0	0	0.3	0.2	0	0	0	0	0	0	0

Since both the manifestations are present in state S_2, only the belief values for the two manifestations in that particular state are considered. The first step is to find the normalized probability values. These are calculated similar to the example shown above. For this example the normalized probability values are as follows:

Manifs	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
M_1= S_2	0.8	0	0	0	0.2	1	0	0	0	0	0	0	0
M_2= S_2	1	0	0	0	0.38	0.25	0	0	0	0	0	0	0

These values are then used to calculate the normalized sums values. In this case the normalized sum for each diagnosis is found by adding the normalized probability values for manifestations M_1 and M_2. The normalized sums values are as follows:

Normalized sums for the thirteen diagnoses												
D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
1.8	0	0	0	0.58	1.25	0	0	0	0	0	0	0

In this case diagnosis D_1 is the most likely diagnosis. Again note that these numbers are no longer probability values. They do not range between zero and one.

Conclusions

In this chapter we have looked at three formulations to reason about the given evidence. The formulations based upon Bayes' Rule and the noisy MAX operator have been implemented based upon known works. We have reviewed these two formulations and examined their relevance to our system. The normalized sums formulation is to our knowledge novel. We have been unable to find this approach in the existing literature. In the next chapter we report the performance of these three formulations in our system.

CHAPTER 4
SYSTEM RESPONSE

The goal of this work is to be able to identify the thirteen diseases under consideration. However, the response of the system will not be a unique selection of the thirteen possible diseases. It is possible for more than one disease to appear simultaneously in the retina. Therefore the response of the system will be a set of diagnoses, naming one or more diseases.

In addition, we explore cases where the retinal image exhibits a disease outside the thirteen diseases known to the system. In some cases, the image may exhibit one disease that is known to the system, and one disease that is unknown to the system. We explore the potential for the system to be able to recognize these unfamiliar or partially familiar cases.

We also examine the normal case, where the retinal image exhibits no disease. Although this case is not modeled in our belief table, we discuss possible methods to diagnose an image as normal using heuristics based upon our formulation.

Finally, there is one more system response of interest. The system may be unable to make a decision, in this case responding "I don't know". We did not pursue this response in this thesis, but we note it here for the sake of completeness.

Figure 4.1 summarizes these possible system responses. This chapter details methods to compute these responses, and provides results on a set of test images. The complete test set consists of 354 images, and is fully described in Appendix A. Various experiments reported in this chapter make use of different subsets of the available images. The importance of each subset and the criteria for selecting the relevant images is discussed in each section.

Familiar Diagnoses

The primary response of the system is a set of one to three diagnoses from the set of thirteen diagnoses that are familiar to the system. The range of one to three

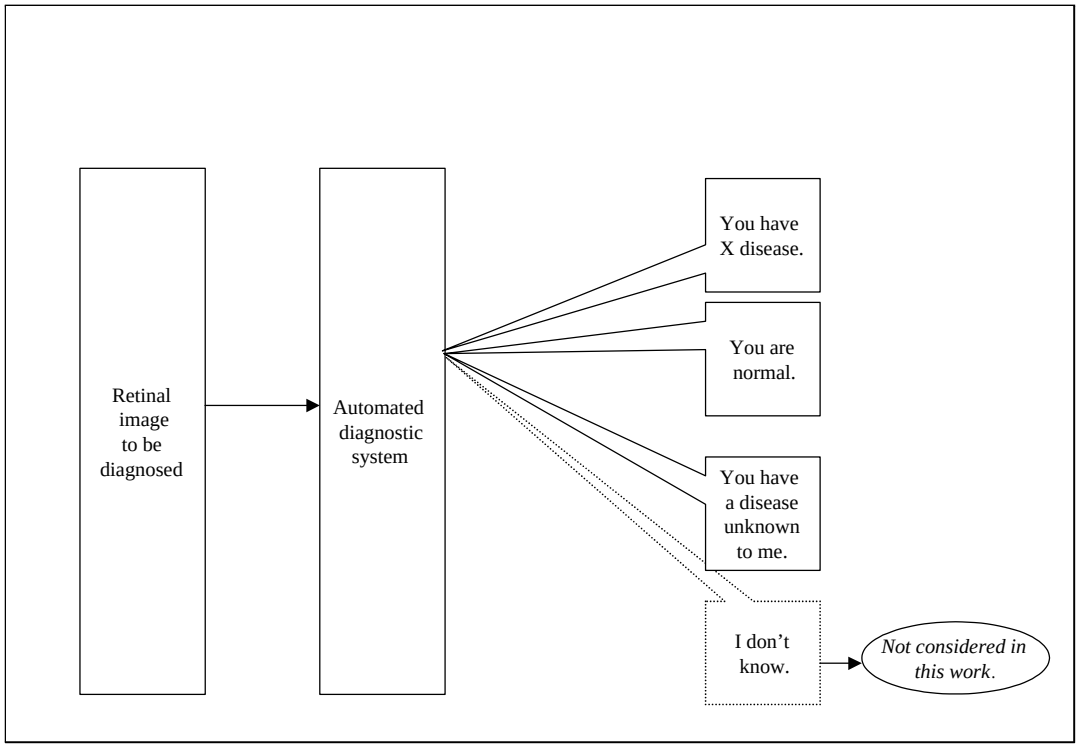

Figure 4.1 Desired system responses.

diagnoses was chosen based on the expert's diagnoses of our test images. Out of 354 images, 250 had one diagnosis, 91 had two diagnoses, 13 had three diagnoses and one had four diagnoses, which was the highest number of diagnoses per image.

The reasoning formulations presented in Chapter 3 result in thirteen values, one for each of the thirteen familiar diagnoses. Each value is an indicator of the strength of the belief in the presence of the disease, where a higher value indicates a stronger belief. These values may be considered as belonging to two sets, where one set indicates the diagnoses responsible for the evidence observed in the retina and the other set indicates the diagnoses not related to the evidence observed. In order to compute the separation of values into two sets, we apply Fisher's linear discriminant test. Appendix B reviews Fisher's test and details its application to our system.

Of the 354 images in our test set, there are 198 images with all the ground truth diagnoses within the set of thirteen recognizable diagnoses. These are referred to as the *familiar images*. These images are used to test the ability of our system to correctly diagnose familiar cases. Of these 198 images, 43 images have multiple ground truth diagnoses.

We tested all three formulations to reason about the evidence discussed in Chapter 3. The evidence is considered in the three contexts of all, linked and annotated, as discussed in Chapter 2. Crossing the formulations with the contexts of evidence yields nine total system tests.

The system performance is evaluated in terms of diagnostic accuracy. The accuracy is measured by comparing the system diagnoses with the ground truth diagnoses. We define two measures of accuracy as functions of the overlap between the ground truth and system diagnoses. Formally, let GT=$\{D_{g_1},....\}$ and MS=$\{D_{m_1},...\}$, where GT represents the set of ground truth diagnoses and MS represents the set of system diagnoses. The system diagnosis is considered a match if $GT \cap MS \neq \oslash$. The system diagnosis is considered a perfect match if $GT = MS$.

Results

The results of applying the three formulations in the three contexts of evidence to the set of 198 images are shown in Table 4.1. The entries summarize the system

performance in percentage of images matched.

	Bayes' Rule	Noisy MAX	Normalized sums
All Evidence	1%	63%	70%
Linked Evidence	7%	30%	22%
Annotated Evidence	7%	36%	75%

Table 4.1 Summary of system performance using the match criterion.

	Bayes' Rule	Noisy MAX	Normalized sums
All Evidence	0%	20%	15%
Linked Evidence	0%	0%	5%
Annotated Evidence	0%	10%	23%

Table 4.2 Summary of system performance using the perfect match criterion.

Conclusions

An examination of Table 4.1 and 4.2 shows that the normalized sums formulation out-performed the noisy MAX formulation. Although this result is for only one system in one problem domain, it does offer empirical evidence to support our philosophical reasoning about the nature of the belief table.

The normalized sums formulation showed a higher recognition rate in the context of annotated evidence than in the context of all evidence. The noisy max formulation did not show a similar performance ratio. This shows that the normalized sums formulation is designed to make the most of positive indicators (evidence), while it gains little from considering negative (absent) indicators.

The performance of the noisy MAX formulation in the context of all evidence and the normalized sums formulation in the context of annotated evidence are broken down by diagnosis in Table 4.3. These two systems were selected because of their better performance. An examination of Table 4.3 shows that the performance varies across the diagnoses. It is also interesting to note that the performance of the two systems on a particular diagnosis is similar. For example both systems have high recognition rates for *CNV* and low recognition rates for *Emboli*. The reason for this variation has not been studied in detail. One reason might be the lack of good

examples for certain diagnoses. Some diagnoses in our test set are only found in images with multiple diagnoses. This might make it harder for the system to detect these diagnoses. This possibility could be studied further in future work.

The perfect match criterion shows that our system is far from optimal. The best system performance using the basic match criterion is 75%. The system is able to recognize at least one of the diagnoses indicated in an image three out of four times. However, the best system performance using the perfect match criterion is 23%. The system is able to recognize all the diagnoses indicated in the image, with no extraneous diagnoses, in only one out of four cases.

Diagnosis	Formulations	
	Noisy MAX	Normalized sums
	All Evidence	Annotated Evidence
Emboli	15%	8%
BRAO	86%	71%
CRAO	50%	38%
BRVO	9%	18%
CRVO	64%	88%
Hemi-CRVO	50%	75%
BDR	60%	74%
PDR	67%	71%
ASR	29%	42%
HTR	33%	23%
Coats	57%	79%
Macroaneurism	50%	88%
CNV	91%	100%

Table 4.3 Performance of the formulations by diagnosis.

Unfamiliar Diagnoses

In this section we consider the problem of recognizing retinal images exhibiting diseases outside our set of thirteen diagnoses. Of the 354 images in our test set, there are 56 images with all the ground truth diagnoses outside the set of thirteen recognizable diagnoses. These images are referred to as *unfamiliar images*. There are also 62 images with at least one of the ground truth diagnoses outside the set of thirteen recognizable diagnoses, and at least one of the ground truth diagnoses

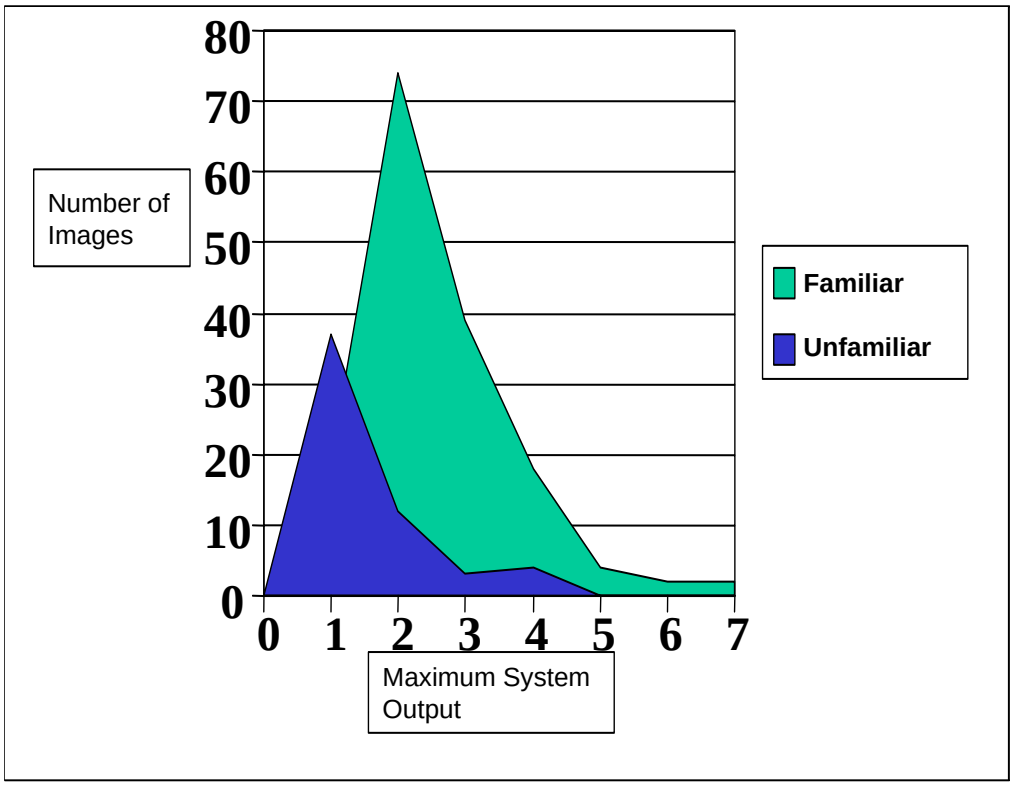

Figure 4.2 Variation of the maximum system output for the familiar and unfamiliar images.

within the set of thirteen recognizable diagnoses. These images are referred to as *partially-familiar images*.

The method of partitioning the system output values into two sets depends upon the values exhibiting a bimodal distribution. We assume that in images with an unknown diagnosis, there is not enough evidence to support any of the thirteen diagnoses within our set. Therefore in these images the system output values for all the thirteen diagnoses within our set should be very low.

To test this hypothesis, we plot the highest system output value for each image. The values are labelled as familiar or unfamiliar according to the ground truth diagnoses of the 254 images (198 familiar + 56 unfamiliar) in our test set. Figure 4.2 shows the variation of the maximum system output for the unfamiliar and familiar images. The normalized sums with annotated evidence system is used to diagnose the images.

Figure 4.2 shows that images with unfamiliar diagnoses have lower system output values. To categorize an image as familiar or unfamiliar, we set a threshold at the valley between the distributions of the familiar and the unfamiliar images. In Figure 4.2, this threshold is at 1.2. Images with a maximum system output less than 1.2 are classified as unfamiliar and images with a maximum system output greater than 1.2 are classified as familiar. Formally, we consider the unfamiliar diagnosis as a fourteenth possible system output. However, this output can only be made in isolation, and not in combination with the thirteen familiar diagnoses.

Results

We tested the method of recognizing the unfamiliar images on the set of 316 images (198 familiar + 56 unfamiliar+ 62 partially familiar). System performance was measured using the match criterion. The results are given in Table 4.4.

Image Class	Images In Class	System Performance
Familiar	198	60%
Unfamiliar	56	71%
Partially Unfamiliar	62	95%
All	316	69%

Table 4.4 System performance based on diagnosis familiarity.

Conclusions

We see from Figure 4.2 that there is considerable overlap between the distributions of the familiar and the unfamiliar images. There are familiar images with a system output less than 1.2. Therefore including the ability to recognize unfamiliar images causes a decrease in performance. Examining Table 4.1 and Table 4.4, we observe that the system's ability to diagnose the familiar images has fallen from 75% to 60%.

It should also be noted that all the unfamiliar diagnoses are lumped together. There is no way to distinguish between two different diagnoses outside our set of thirteen familiar diagnoses.

The system performs with a very high accuracy on the partially familiar images. The reason for this high accuracy has not been studied in this work.

As an alternative to using the maximum system output value to distinguish between familiar and unfamiliar diagnoses, we also considered using the maximum value from Fisher's test. The idea is that if the distribution is not bimodal, the largest value from Fisher's test should be relatively small. However, we found this method did not produce an adequate partitioning.

Normal cases

In this section we consider the problem of recognizing retinal images exhibiting no disease. Of the 354 images in our test set, 38 have been diagnosed as *normal*. Looking at the annotations, we found that 25 of these images did not have any abnormalities. All the manifestations were either absent or normal. The diagnosis of these 25 images as normal is trivial.

The remaining thirteen images were diagnosed as normal despite the presence of one or more abnormalities. We examined these annotations for similarities. Presumably there may be some manifestations that can appear in the retina without the presence of a disease. We might recognize these as patterns. We also examined the system output values for all the normal images. The idea is that we might be able to identify a threshold to classify an image as normal.

Results

We could not discover a pattern in the annotations. No two annotations were exactly alike. The system output values also did not exhibit a noticeable pattern.

Conclusions

We hypothesize two possible reasons for our failure to find a pattern or discriminant for the normal images. First, it is possible that with the thirteen images under consideration, there was a problem in data collection. For example, the annotation may be inaccurate, the diagnosis may be mistaken, or there may have been an error during the recording of the data. Second, it is possible that more normal images are needed to define a pattern. We did not explore this issue, and we leave it for future work.

CHAPTER 5

FREQUENCIES VERSUS BELIEFS

The values in the belief table represent opinions. Each value is an expert's belief in the likelihood of a manifestation given the presence of a disease. Although the opinion is well-informed, the beliefs are not actual measurements of probabilities. In this chapter we study the effects of substituting the beliefs with frequencies. Frequencies are actual measurements of the probabilities.

The difficulty in using frequencies lies in making the actual measurements. In order to be statistically meaningful, the frequencies must be measured using "large" populations. In our case, this could mean making measurements of the retinal health of thousands (or millions) of people. This difficulty is why it is attractive to use an expert's beliefs in place of frequencies. Although it is time consuming and potentially error-prone to obtain an expert's beliefs, it is relatively more difficult to obtain actual frequencies.

In the course of this research we have obtained hand drafted annotations and diagnoses of hundreds of retinal images. The frequency of each manifestation in the presence of a disease can be computed from this data. After computing these frequencies, we can substitute them for the beliefs and rerun our system on the image database. We are interested in discovering the performance differences, if any, between the use of frequencies and beliefs.

The frequencies to be used to diagnose the retinal images have to be computed. To compute a frequency, we count the number of times manifestation M in state S has occurred in the presence of diagnosis D in all the test images. This count is then scaled by the number of images in the set with diagnosis D.

In order to obtain unbiased system diagnoses, we separate the entire set of images into training and test sets. We use the well known method called "leave one out". One image is chosen to be the test image. The rest of the images are then used to calculate the frequencies. Using these frequencies the test image is diagnosed. This

method is repeated so that each image is the test image exactly once. Any image that is to be diagnosed is excluded from the training set.

Results

We tested this methodology on the set of 198 familiar images (see Chapter 4). We tested two systems, one using the noisy max formulation in the context of all evidence, and one using the normalized sums formulation in the context of annotated evidence. The results are tabulated by diagnosis in Table 5.1.

Diagnosis	Normalized sums		Noisy MAX	
	annotated evidence		all evidence	
	beliefs	frequencies	beliefs	frequencies
Emboli	8%	15%	15%	23%
BRAO	71%	71%	86%	14%
CRAO	38%	50%	50%	13%
BRVO	18%	82%	9%	9%
CRVO	88%	84%	64%	32%
Hemi-CRVO	75%	83%	50%	42%
BDR	74%	88%	60%	54%
PDR	71%	76%	67%	62%
ASR	42%	38%	29%	4%
HTR	23%	33%	33%	24%
Coats	79%	93%	57%	57%
Macroaneurism	88%	88%	50%	38%
CNV	100%	91%	91%	55%
Total	75%	86%	63%	51%

Table 5.1 System performance on 198 familiar images by diagnosis, comparing frequencies with beliefs.

Also, we compared the frequencies with the expert given beliefs to find the distribution of variance. This distribution is plotted in Figure 5.1.

Conclusions

Figure 5.1 shows that the difference between each frequency and belief is less than 0.2 in most cases. The instances where the difference is larger can be studied in future work.

The normalized sums system improved from 75% to 86% in diagnostic efficiency using frequencies in place of beliefs. The performance of the noisy MAX system

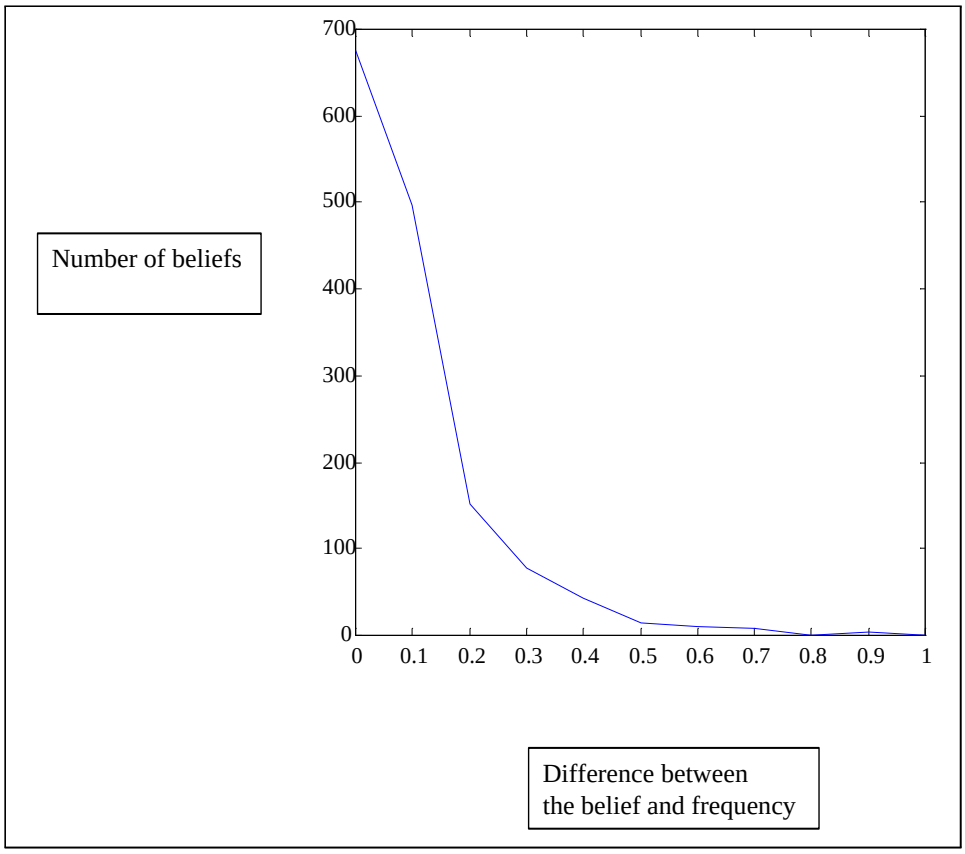

Figure 5.1 Differences between the expert given beliefs and the calculated frequencies.

decreased from 63% to 51% in diagnostic efficiency using frequencies in place of beliefs. These results appear contradictory. It is not clear to us why the performance of the noisy MAX system decreased.

For the normalized sums system, the performance difference for most diagnoses was less than ±15%. The one exception was BRVO, which improved by 64%. This dramatic change might indicate a biased belief. In future work we would like to reexamine this particular diagnosis.

We note that our sample population is still relatively small. A sample of hundreds does not necessarily provide confidence that the improvement in performance of the normalized sums system witnessed in this experiment is statistically significant.

The problem of frequency acquisition remains more difficult than belief acquisition. For this project, the effort required to gather the data to compute the frequencies took months, while the effort required to gather the beliefs took only days. None-the-less, there may be an alternative to strict frequency acquisition. It may be possible to gather frequencies based on smaller samples and provide them to an expert during the acquisition of beliefs. In future work we plan to pursue the acquisition and evaluation of such "frequency anchored beliefs".

CHAPTER 6
CASE DIFFICULTY

It is possible to assume that all diagnostic cases are of equal difficulty. For our system, this would imply that it is equally difficult to diagnose each image in our test set. In this chapter we test this assumption. We hypothesize that the test set can be divided into easy, moderate, and difficult cases. We define the ideal annotation and the separable annotation in order to classify the diagnostic difficulty of each image. The difficulty of each image is measured according to the difference of its annotation from the separable annotation of its corresponding ground truth diagnosis. We examine the performance of the system on these classes to verify the hypothesis.

Ideal Annotation

Each diagnosis is assumed to have an ideal annotation. The ideal annotation consists of manifestations which strongly identify the given diagnosis. We call these manifestations key indicators. Manifestations that only weakly identify the given diagnosis are considered support evidence, and are excluded from the ideal annotation.

There are 39 possible manifestations in our system, having from two to seven states each. Ignoring the absent or normal state of each manifestation, there are 75 total possible manifestation states. We ignore the absent or normal state of each manifestation, for this experiment, because we are only interested in positive indicators of disease.

We assume that the manifestations of a particular diagnosis may be considered as belonging to two sets, key indicators and support evidence. The beliefs associated with these manifestations should have larger values for the key indicators, and smaller values for the support evidence. We use Fisher's linear discriminant (see Appendix B) to separate the 75 total possible manifestation states for each diagnosis into two sets. The set of larger values identifies the key indicators, which form the ideal annotation.

It is possible for two or more states of a single manifestation to have larger belief values, so that multiple states of a single manifestation can be key indicators. The result is that there may be more than one ideal annotation for a diagnosis.

The ideal annotations computed for each diagnosis are shown in Table 6.1. Numbers separated by a slash in a single box indicate multiple states of that manifestation are key indicators for the given diagnosis. Blank entries indicate support evidence, or manifestations that are not causally linked to the given diagnosis.

According to our definition, it is possible for more than one diagnosis to have the same ideal annotation. In Table 6.1, diagnoses D5 (*CRVO*) and D6 (*hemi-CRVO*) have the same ideal annotation. In this case it is necessary to use the support evidence to distinguish between the two diagnoses. To account for this possibility, we define the separable annotation.

Separable Annotation

Each diagnosis is assumed to have a separable annotation. The separable annotation consists of manifestations which identify more strongly with one unique diagnosis than with any of the other possible diagnoses. We call these manifestations strong separators. Manifestations that equally identify with multiple diagnoses, whether strongly or weakly, are considered weak separators, and are excluded from the separable annotation.

Each state of each manifestation is considered independently during the computation of the separable annotations. There are thirteen possible diagnoses in our system, so that there are thirteen possible belief values for each manifestation state. We assume that the belief values of a manifestation state across the set of diagnoses may be considered as belonging to two sets, strong separators and weak separators. The beliefs should have larger values for the strong separators, and smaller values for the weak separators. We use Fisher's linear discriminant (see Appendix B) to separate the thirteen belief values for each manifestation state into two sets. The set of larger values identifies the strong separators.

This separation is repeated for the 75 abnormal manifestation states. For this

experiment, we ignore the normal (absent) state of each manifestation because we are only interested in positive indicators of disease. Since each manifestation state is separated independently, it is possible for two or more states of a single manifestation to be strong separators. The result is that there may be more than one separable annotation for a diagnosis.

The separable annotations computed for each diagnosis are shown in Table 6.2. Numbers separated by a slash in a single box indicate multiple states of that manifestation are strong separators for the given diagnosis. Blank entries indicate weak separators, or manifestations that are not causally linked to the given diagnosis.

According to our definition, it is not possible for more than one diagnosis to have the same separable annotation. Although the separable annotation is less definitive than the ideal annotation for a given diagnosis, its uniqueness provides for a more precise measure of case difficulty.

In order to verify that the ideal and separable annotations are good (ideal) cases of the diagnoses, we process them through our system. We expect that the system should obtain near 100% diagnostic accuracy on these annotations. We test two systems, one using the noisy MAX formulation in the context of all evidence, and one using the normalized sums formulation in the context of annotated evidence. There are 16 ideal annotations (three diagnoses have two ideal annotations each), and 25 separable annotations (6 diagnoses have multiple separable annotations). Table 6.3 shows the results.

Image classes

We hypothesize that it is possible to partition the 198 familiar images in our test set into three classes of diagnostic difficulty: easy, moderate and difficult. We measure the diagnostic difficulty of an image according to the difference between its annotation and the separable annotation for its ground truth diagnosis. The manifestations present in the separable annotation are referred to as strong separators. The image annotation can have some or all of the strong separators missing. It can also have manifestations other than the strong separators present. We assume these extra

manifestations detract attention from any strong separators that are present.

We define c_k as the ratio of the number of strong separators in the image annotation to the number of strong separators present in the separable annotation. The value of c_k can range from 0 to 1. When c_k =0, it implies that all the strong separators are absent from the image annotation, making the image hard to diagnose. When c_k =1, all the strong separators are present in the image annotation, making the image easy to diagnose.

We define c_d as the ratio of the number of detractor manifestations in the image annotation to the number of strong separators in the separable annotation. The range of c_d is determined by the number of manifestations being considered. When $c_d > 1$, the detractor manifestations outnumber the strong separators, presumably making the image more difficult to diagnose.

Figure 6.1 shows a plot of the values of c_k and c_d for the 198 familiar images in our test set. This plot shows that the distribution varies largely with c_k, but very little with c_d. Figure 6.2 shows a plot of the values of c_k only for the 198 familiar images in our test set. Since we hypothesized three classes of case difficulty, we identify three ranges of values of c_k suggested by this plot. Images with $c_k <= 0.6$ are classified as easy. Images with $0.6 < c_k <= 0.95$ are classified as moderate. Images with $0.95 < c_k$ are classified as difficult. Based on these boundary ranges, the 198 familiar images are distributed as listed in Table 6.4. System performance on these images, distributed according to class, is listed in Table 6.5. For this experiment the system used is the normalized sums formulation in the context of annotated evidence.

Conclusions

Table 6.5 suggests that there are only two classes of images: easy and difficult. Images with $c_k <= 0.95$ can be classified as easy or moderate and those with $c_k > 0.95$ can be classified as difficult. This suggests that for an image to be easily diagnosed it must have at least one of the strong separators of its ground truth diagnoses. Images which have no strong separators and rely only upon support evidence are harder to diagnose.

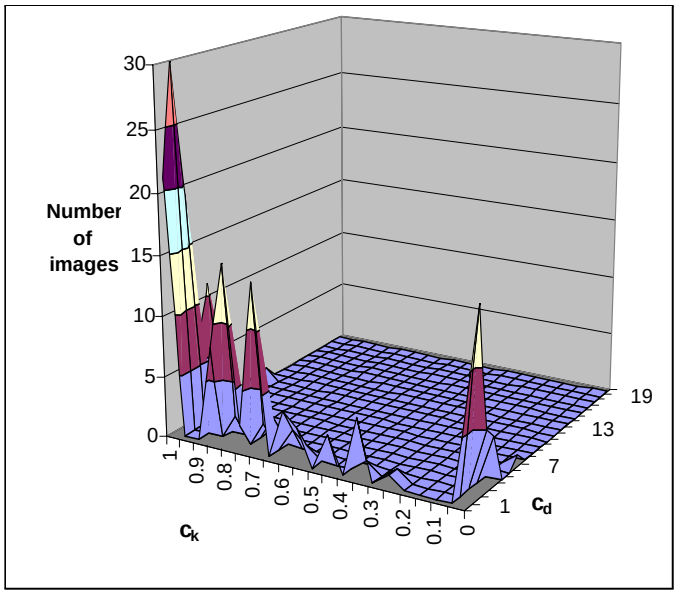

Figure 6.1 Variation of the images with c_k and c_d.

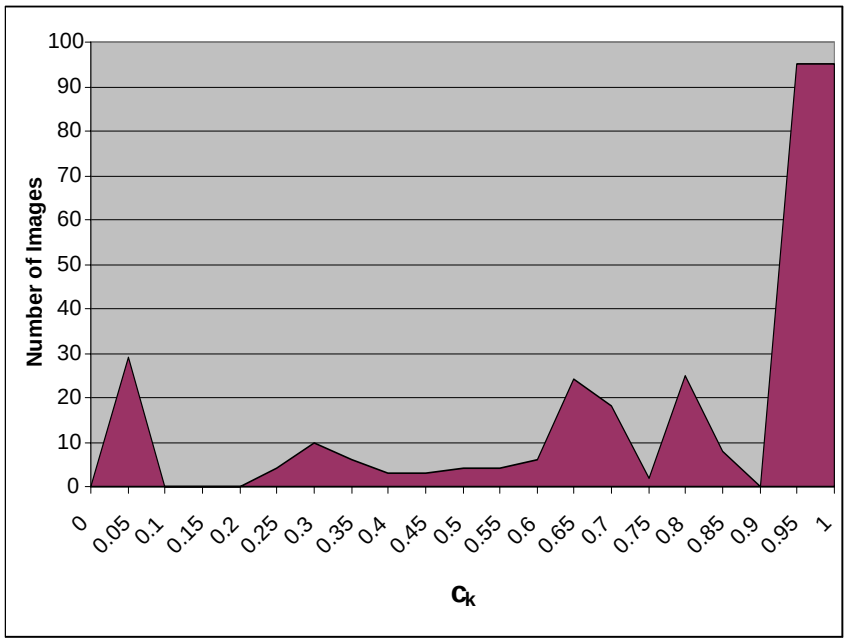

Figure 6.2 Variation of images with c_k.

Manifs	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
1													
2					2	2	2						
3				2									
4													
5													2
6													
7													
8													2
9							2						
10													
11				4 / 5			6			6			
12							2			2			
13										2 / 3			
14													
15										3			
16													
17													
18										3			
19								3					
20													
21													
22													
23													
24		3								2			
25				2									
26													
27				2							2		
28											3		
29									2				
30												2	
31													
32													
33													
34		2	3										
35													
36													
37													
38								4					
39	2 / 3												

Table 6.1 Ideal annotations for the thirteen diagnoses.

Manif	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
1													
2													
3													
4													2
5													2
6													3/4/5
7													
8													2
9							2						
10													
11					2 / 3								
12											3	5	6/ 7
13										2 / 3			
14					2								
15					3								
16													
17		2											
18					3	2							
19								2 / 3					
20		2											
21									2				
22											2		
23											3		
24		3	4 / 5									2	
25				2	3								
26			3										
27													
28											3		
29									2				
30												2	
31									2				
32													
33											2 / 3		
34		2	3										
35			2										
36													
37								2 / 3					
38								4					3
39	3												

Table 6.2 Separable annotations for the thirteen diagnoses.

	Normalized sums	Noisy MAX
	annotated evidence	all evidence
match against ideal annotation	100%	94%
perfect match against ideal annotation	69%	88%
match against separable annotation	100%	96%
perfect match against separable annotation	100%	96%

Table 6.3 System performance on ideal and separable annotations.

Image Class	Number of Imges
Easy	58
Moderate	45
Difficult	95

Table 6.4 Number of images in each hypothesized class of diagnostic difficulty.

Image Class	Images In Class	Images	System
		Correctly Diagnosed	Performance
Easy	58	55	95%
Moderate	45	42	93%
Difficult	95	52	55%

Table 6.5 System performance based on case difficulty.

APPENDICES

Appendix A

Domain Materials

This appendix details the data used in this thesis. These materials are specific to the domain considered in this work, retinal image diagnosis.

Diagnoses

	Diagnoses	Abbreviation
1	Emboli	Emboli
2	Branch retinal artery occlusion	BRAO
3	Cilio-retinal artery occlusion	CRAO
4	Branch retinal vein occlusion	BRVO
5	Central retinal vein occlusion	CRVO
6	Hemi-central retinal vein occlusion	Hemi-CRVO
7	Background diabetic retinopathy	BDR
8	Proliferative diabetic retinopathy	PDR
9	Arteriosclerotic retinopathy	ASR
10	Hypertensive retinopathy	HTR
11	Coats' disease	Coats
12	Macroaneurism	Macroaneurism
13	Choroidal neovascularization	CNV

Table A.1 List of the diagnoses.

Manifestations

	Manifestation	States
1	RPED	Absent
		Present
2	CME	Not visible
		Visible
3	ERM	Absent or not visible
		Present
4	Subretinal fibrosis	Absent
		Present
5	CNV	Not observable
		Observable
6	Drusen	Absent
		Fine, few
		Fine, many
		Large, soft, few
		Large, soft,many
7	Preretinal	Absent
	hemorrahge	Present anywhere
8	Subretinal	Absent
	hemorrahge	Present anywhere
9	Microaneurism	Absent
	or dot	Few anywhere
	hemorrahge	Many anywhere
10	VH	Absent
		Present anywhere

Table A.2 List of manifestations and their states.

	Manifestation	States
11	Small or	Absent
	medium blot or	Low density, not regional
	flame	High density, not regional
	hemorrahge	Low density,regional not crossing horiz. meridian
		High density,regional not crossing horiz. meridian
		Low density,regional crossing horiz. meridian
		High density,regional crossing horiz. meridian
12	Retinal or	Absent
	subretinal	Low severity, no circinate
	exudate	High severity, no circinate
		Low severity, circinate not surrounding fovea
		High severity, circinate not surrounding fovea
		Low severity, circinate surrounding fovea
		High severity, circinate surrounding fovea
13	Macula data	Absent
		Present incomplete
		Present 360 deg
14	ON collateral	Absent
		Present
15	ON swelling	Absent
		Low severity
		High severity
16	ON hemmorahge	Absent
		Splinter
		Blob
17	ON palor	Normal
		Sector palor
		Pale or white, whole nerve

Table A.3 List of manifestations and their states (continued).

	Manifestation	States
18	ON color	Normal
		Sector erythema
		Rosy or red, whole nerve
19	NVD	Absent
		Less than one disk area
		Greater than one disk area
20	Artery oxygen	Normal
		Dark (deoxygenated)
21	Artery color	Normal
		Copper wire
		Silver wire
22	Artery sheath	Absent
		Present
23	Vein color	Normal
		Gray or white (ghost vessel)
		Yellow (sheathed)
24	Artery narrow	Normal
		Focal, one or more segments
		Moderate, branch or single
		Moderate, global
		Extreme, global
25	Artery dilation	Normal
		Tortuosity, branch or single
		Tortuosity, global
26	Vein narrow	Normal
		Tributary vein or single
		Entire venous tree
27	Vein dilation	Normal
		Tortuosity, tributary vein or single
		Tortuosity, global
28	Teleanglectasis	Absent or not visible
		Present, not crossing horizontal median
		Present, crossing horizantal meridian
29	Bv specular reflex	Normal
		Wide and bright

Table A.4 List of manifestations and their states (continued).

	Manifestation	States
30	Macroaneurism	Absent
		Single
		Multiple
31	AV change	Absent or not in picture
		One or more examples
32	TXRD schisis	Absent
		Present
33	Cotton-wool spot	Absent
		Few
		Many
34	Inner retinal infarct	Absent
		Not involving the macula
		Involving the macula (cherry red spot)
35	Cherry red spot	Absent
		Present
36	Ghost bv	Absent
		Present
37	NVE	Absent
		Few or small
		Many or large
		Absent
38	Photocoagular scar	Fine (grid)
		Large, in or near macula
		Around five hundred †m,
		round, arcade or peripheral, many
39	Emboli manifestation	Absent
		One
		More than one

Table A.5 List of manifestations and their states (continued).

Causal Links

	Manifestation	Linked Diagnoses
1	RPED	CNV
2	CME	CNV, BDR,PDR, BRVO,CRVO,
		Hemi-CRVO, Coats, Macroaneurism
3	ERM	BDR, PDR,BRVO,CRVO, Hemi-CRVO
4	Subretinal fibrosis	CNV, PDR, Coats
5	CNV	CNV
6	Drusen	CNV
7	Preretinal hemorrhage	CNV, HTR, PDR,
		BRAO,CRAO, BRVO,CRVO, Hemi -CRVO
		Macroaneurism, Coats
8	Subretinal hemorrhage	CNV, BRVO,CRVO, Hemi-CRVO ,
		Macroaneurism, Coats
9	Microaneurism	HTR, BDR,PDR,BRVO,
		CRVO, Hemi-CRVO, Coats
10	VH	CNV,PDR, BRVO,CRVO, Hemi-CRVO,
		Macroaneurism, Coats
11	Small or medium	HTR,PDR,BDR,BRVO,CRVO,
	blot hemorrhage	Hemi-CRVO,Macroaneurism, Coats
12	Retinal or subretinal	CNV, HTR, PDR,BDR, BRVO,
	exudates	Hemi-CRVO,Macroaneurism, Coats
13	Macula data	CNV, HTR, PDR,BDR, BRVO,
		Hemi-CRVO,Macroaneurism, Coats
14	ON collateral	CRVO, Hemi-CRVO
15	ON swelling	BDR, PDR, BRAO, CRVO, Hemi-CRVO,
		HTR, Coats
16	ON hemorrhage	HTR, CRVO, Hemi-CRVO,Coats
17	ON palor	BRAO, Coats
18	ON color	PDR, BDR, BRAO, CRVO, Hemi-CRVO
		HTR, Coats

Table A.6 List of manifestations and their causal links.

	Manifestation	Linked Diagnoses
19	NVD	PDR, BRAO, CRAO, BRVO, CRVO,
		Hemi-CRVO, Coats
20	Artery oxygen	BRAO, CRAO, Coats
21	Artery color	PDR, BRAO, ASR
22	Artery sheathing	Macroaneurism, Coats
23	Vein color	BRVO, CRVO, Hemi CRVO, Coats
24	Artery narrowing	HTR, PDR, BRAO, CRAO, Emboli,
		BRVO, CRVO, Hemi-CRVO,
		Macroaneurism, Coats
25	Artery dilation	BRVO, CRVO, Coats
26	Vein narrowing	PDR, BRAO,CRAO, Coats
27	Vein dilation	PDR, BRVO, CRVO, Hemi-CRVO, Coats
28	Teleanglectasis	BDR, PDR, BRVO, Hemi-CRVO, Coats
29	Bv specular reflex	ASR
30	Macroaneurism	HTR, Coats, Macroaneurism
31	AV change	ASR
32	TXRD schisis	PDR, Coats
33	Cotton wool spot	PDR, BDR, BRAO, Emboli, BRVO,
		CRVO, Hemi-CRVO, ASR, Coats, HTR
34	Inner retinal infract	BRAO, CRAO, Emboli
35	Cherry red spot	CRAO, Emboli
36	Ghost bv	HTR, PDR, BDR, BRAO,
		CRAO, BRVO, CRVO, Hemi -CRVO
		Macroaneurism, Coats, ASR
37	NVE	PDR, Coats, BRAO,CRAO,BRVO,
		CRVO, Hemi CRVO
38	Photocoagular scar	CNV, PDR,BDR, Coats
39	Emboli manifestation	Emboli, BRAO,CRAO

Table A.7 List of manifestations and their causal links (continued).

Image Database

The images in the STARE database are designated imXXXX where XXXX is a four digit number. The available images are numbered from 0001 to 0402. Five numbers in this range are missing. Of the 397 available images, 43 images have missing or incomplete annotations or ground truth diagnoses, as listed in Table A.8. In these cases the expert physician felt that the image was impossible to diagnose alone (a physician might use verbally acquired evidence from the patient, as well as multiple images of multiple portions of the patient's eye in order to conclude a diagnosis). The remaining 354 images with annotations were used for experiments in this thesis.

14, 22, 35, 42, 55, 56, 71, 75, 79, 99, 102,
107, 111, 146, 152, 153, 180, 181, 182,
183, 184, 261, 264, 272, 292, 294, 295,
303, 304, 307, 346, 374, 377, 386, 387,
388, 389, 390, 391, 392, 393, 394, 395

Table A.8 Images with missing or incomplete annotations or ground truth diagnoses.

We define these 354 images according to four categories. A familiar image has all its ground truth diagnoses within the set of thirteen recognizable diagnoses. An unfamiliar image has all its ground truth diagnoses outside the set of thirteen recognizable diagnoses. A partially familiar image has at least one of its ground truth diagnoses outside the set of thirteen recognizable diagnoses and at least one of the ground truth diagnoses within the set of thirteen recognizable diagnoses. A normal image exhibit no disease. Tables A.9-A.12 list the images according to these four categories.

1, 2, 5, 9, 13, 16, 17, 18, 19, 20, 21, 25, 26, 27, 28,
31, 33, 38, 40, 48, 49, 50, 51, 52, 57, 58, 60, 64, 65,
69, 70, 73, 74, 85, 87, 89, 90, 93, 94, 95, 96, 101,
103, 113, 114, 115, 116, 117, 118, 121, 122, 123,
124, 125, 126, 127, 133, 134, 135, 136, 137, 138,
139, 140, 141, 148, 150, 155, 156, 157, 158, 160,
161, 165, 166, 168, 169, 177, 178, 179, 186, 187,
188, 189, 191, 194, 195, 196, 197, 200, 201, 202,
203, 204, 205, 206, 207, 208, 209, 210, 211, 214,
215, 217, 218, 220, 222, 223, 224, 225, 226, 227,
228, 230, 232, 233, 246, 247, 248, 251, 257, 258,
259, 271, 274, 275, 276, 277, 279, 280, 281, 284,
293, 297, 299, 306, 308, 309, 311, 312, 313, 314,
315, 316, 317, 318, 319, 320, 321, 322, 323, 324,
325, 326, 327, 328, 329, 330, 331, 332, 333, 334,
335, 336, 337, 338, 340, 341, 342, 343, 345, 347,
348, 349, 350, 351, 352, 353, 354, 356, 357, 358,
359, 361, 362, 363, 364, 365, 366, 367, 368, 369,
371, 396, 397, 398, 399, 400

Table A.9 Familiar images.

3, 6, 10, 12, 15, 23, 24, 29, 39, 41, 43, 44,
46, 63, 66, 67, 68, 72, 78, 86, 88, 97, 112,
128, 129, 130, 131, 132, 159, 173, 174, 175,
176, 185, 221, 250, 273, 282, 283, 285, 286,
287, 288, 289, 290, 291, 296, 298, 300, 301,
302, 305, 344, 360, 401, 402

Table A.10 Unfamiliar images.

4, 7, 8, 11, 34, 36, 37, 45, 53, 54, 61, 62, 83, 84,
91, 92, 98, 100, 104, 105, 106, 110, 142, 143,
145, 147, 149, 151, 154, 171, 192, 193, 212,
229, 256, 260, 262, 263, 265, 266, 267, 268,
269, 270, 310, 339, 355, 370, 372, 373, 375,
376, 378, 379, 380, 381, 382, 383, 384, 385

Table A.11 Partially familiar images.

30, 32, 76, 77, 80, 81, 82, 119, 120, 162, 163, 164,
170, 190, 198, 199, 213, 216, 219, 231, 234, 235,
236, 237, 238, 239, 240, 241, 241, 243, 244, 245,
249, 252, 253, 254, 255, 278

Table A.12 Normal images.

Annotations

A hand drafted annotation of the manifestations in each image was obtained from the expert. When the annotations were obtained, the sets of diagnoses and manifestations under consideration were slightly different from the final sets used in this thesis. Each annotation contains 44 manifestations, compared to the 39 manifestations used in this thesis. Figure A.1 lists the two formats and highlights the differences.

The original annotation includes states for *retinitis* (16), *optic nerve in picture* (17), *retinal angioma* (38), *chorioretinal scar* (40), *grape clusters* (42), *nevus* (43) and *geographic angioma* (44), which are not used in this thesis. The manifestation *on color* (21) was split into *on palor* (17) and *on color* (18) for this thesis. This lets the states for this abnormality be listed monotonically increasing in severity, which is a requirement for the noisy max formulation. Similarly, the manifestation *artery color* (23) was split into *artery oxygen* (20) and *artery color* (21), the manifestation *artery diameter* (26) was split into *artery narrow* (24) and *artery dilation* (25) and the manifestation *vein diameter* (27) was split into *vein narrow* (26) and *vein dilation* (27). The manifestations *small or medium blot hemorrhage* (11) and *retinal hemorrhage* (12) were combined to form the manifestation *small or medium blot hemorrhage* (11). Similarly the manifestations *retinal or subretinal exudate* (13) and *circinate pattern* (14) were combined to form the manifestation *retinal or subretinal exudate* (12).

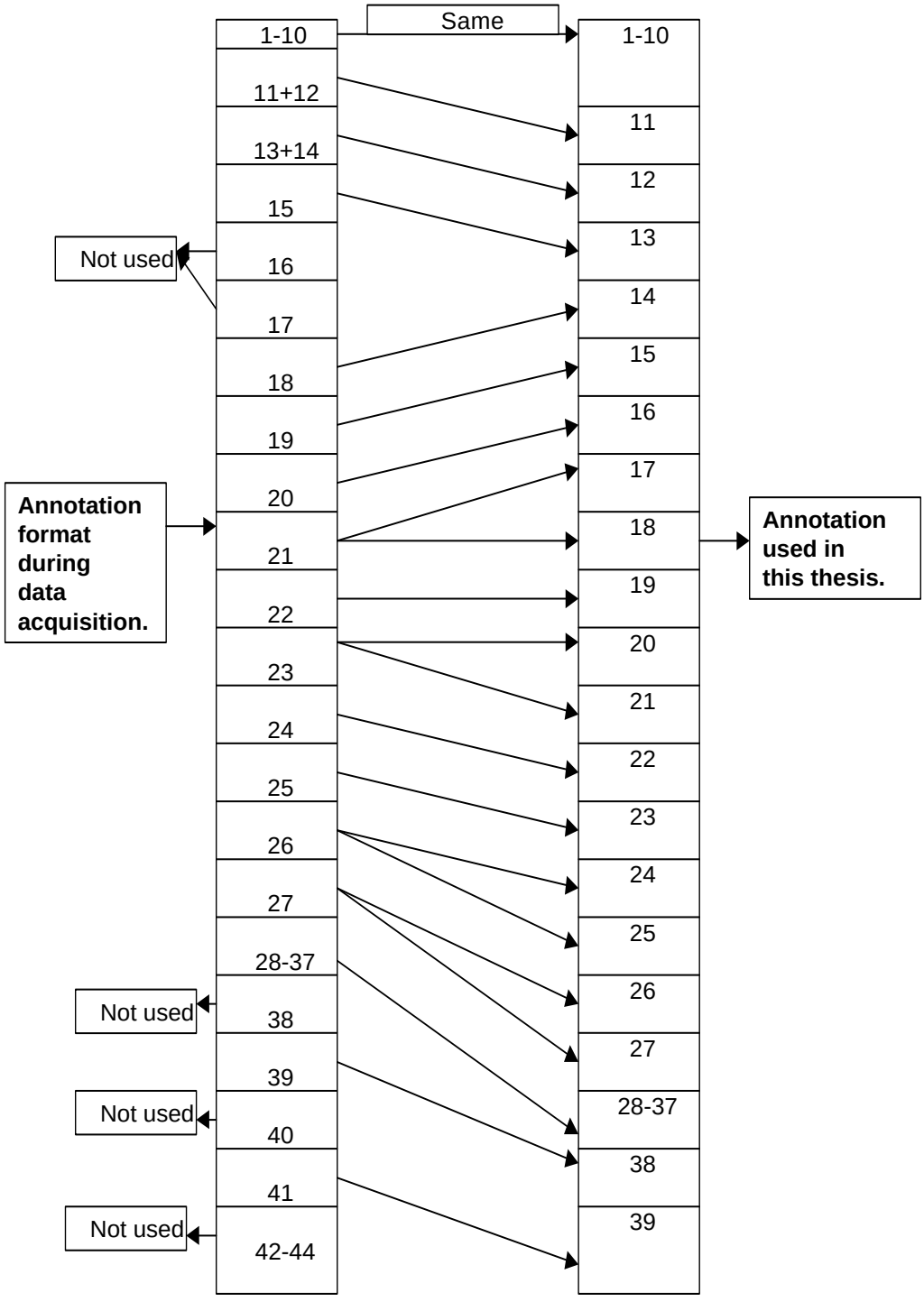

Figure A.1 A comparison of the annotation format and the format used in this thesis.

Appendix B

Fisher's Linear Discriminant Test

Linear discriminant analysis is a well known statistical tool used to classify data belonging to different classes or groups. This test finds the linear projection that maximizes the squared distances between the group means. The distances are standardized by the sample variance. A new observation can be classified based on its projection on the line. The linear discriminant tends to have good separation properties even in the case when the data is not normally distributed.

In this study Fisher's linear discriminant test is used to find the system diagnoses. Before applying Fisher's test to the diagnostic probabilities, they are arranged in descending order.

$$P(D_{i_1}),\ P(D_{i_2}), ... \ P(D_{i_I}) \qquad \text{such that } P(D_{i_x}) > P(D_{i_{x+1}})$$

where $i = 1...I$ refers to each of the total I diagnoses and $P(D_i)$ refers to the final diagnostic probability output by the system. After these probabilities have been sorted in descending order they are partitioned into two sets A and B. The partitioning is done $I-1$ times as follows:

$$A = \{P(D_{i_1})...P(D_{i_f})\},\ B = \{P(D_{i_{f+1}})...P(D_{i_I})\} \qquad \text{for all } f = 1...I-1 \tag{B.1}$$

Fisher's linear discriminant is calculated for each partitioning as follows:

$$F_f = \frac{(\mu_A - \mu_B)^2}{\sigma_A^2 + \sigma_B^2} \tag{B.2}$$

where μ_A and μ_B are the means of the two groups or classes, and σ_A and σ_B are the variances of the two groups. The value of F_f gives the "separateness" between the two groups. We use the linear discriminant test to classify the final system output values associated with each of the I diagnoses into two sets.

In order to diagnose an image we find F_f for the different partitions of the system output values associated with the diagnoses. The partition with the highest F_f is then chosen. The diagnoses in set A for the partition with the highest F_f form the set of

system diagnoses. The test gives us only those diagnoses in set A whose probability values are higher and well separated from the other probabilities.

The following example demonstrates Fisher's test. In this example the system has output the following values:

Probability values for diagnoses												
D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}
0.1	0	0.05	0	0	0	0.15	0.7	0	0.74	0	0	0

First, the $P(D_i)$ values are sorted in descending order.

Probability values for diagnoses												
D_{10}	D_8	D_7	D_1	D_3	D_2	D_4	D_5	D_6	D_9	D_{11}	D_{12}	D_{13}
0.74	0.7	0.15	0.1	0.05	0	0	0	0	0	0	0	0

The values are partitioned twelve times into 2 sets A and B using Equation B.1. In the first partition set A contains D_{10} and set B contains all the other diagnoses. The value of F_f for this partition is then found using Equation B.2. This step is repeated for all twelve partitions. The various F_f values for the twelve partitions in this example are as follows:

F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
318.3	14275.7	50.3	15.1	9.2	6.5	5.3	4.7	4.3	4.1	3.9	3.7

From this table we see that the second partitioning has the largest value from Fisher's test. In this partition set A will contain the first two elements of the sorted list of diagnoses. Therefore, the system diagnoses are D_{10} and D_8.

In our system there are many images with multiple ground truth diagnoses. The number of ground truth diagnoses for all the images in our test set range from one to three. The number of system diagnoses is therefore also be restricted to a maximum value of three.

Fisher's test poses no restrictions on the number of elements in set A. The number of diagnoses in set A can range from one to $I - 1$. Therefore we have imposed a restriction on the number of elements in set A by limiting the value of f to be less than or equal to three.

Appendix C

Glossary

Diagnosis	Disease of the human retina.
Manifestation	Abnormality observed in the retina or retinal image.
Annotation	List of each manifestation's state in each image.
Belief Table	A probability table summarizing the physician's beliefs in the presence of each manifestation given the presence of each disease.
Ground truth diagnosis	The physician's diagnosis of a retinal image.
System diagnosis	The system's diagnosis of a retinal image.
Link	An indicator of a cause between a diagnosis and a manifestation.
Influence diagram	Graphical representation of the causal relationships between diagnoses and manifestations.
Formulation	Mathematical formula used to mimic the expert's reasoning process.
Context of evidence	The evidence observed in the retinal image can be used in three different contexts, all, linked or annotated.
System	A specific formulation and context of evidence used for an experiment.
Familiar image	Image with all the ground truth diagnoses within the set of thirteen recognizable diagnoses.
Unfamiliar image	Image with all the ground truth diagnoses outside the set of thirteen recognizable diagnoses.
Partially familiar image	Image with at least one of the ground truth diagnoses outside the set of thirteen recognizable diagnoses, and at least one of the ground truth diagnoses within the set of thirteen recognizable diagnoses.
Normal Image	An image that exhibits no disease.
Frequency	Number of times an event occurs.
Key indicator	Manifestation that strongly contributes to the belief in a diagnosis.
Strong separator	Manifestation that contributes to the belief in only one diagnosis.
Weak separator	Manifestation that contributes to the belief in two or more diagnoses.
Support evidence	Manifestation that does not contribute very strongly to the belief in a diagnosis, but does contribute a little to the belief ina diagnosis.
Ideal Annotation	An annotation with only key indicators.
Separable Annotation	An annotation with only strong separators.
Image Classes	Classification of images based on their distance from the separable annotation.

REFERENCES

[1] JS Bennett and RS Engelmore, B Buchanan and E Shortliffe, "Experience using EMYCIN, In Rule based Expert Systems", *Addison-Wesley press*, Reading, MA, 1984.

[2] AS Elstein, CP Friedman, FM Wolf, G Murphy, J Miller, P Fine, P Heckerling, T Miller, J Sisson, S Barlas, K Biolsi, M Ng, X Mei, T Franz and A Capitano, "Effects of a Decision Support System on the Diagnostic Accuracy of Users: A Preliminary Report", *Journal of the American Medical Informatics Association*, 1996, Vol 3, pp 422-428.

[3] FJ Diez, "Parameter Adjustment in Bayes networks. The generalized noisy OR-gate", *Uncertainty in Artificial Intelligence,* 1993, pp 99-105.

[4] FT de Dombal, DJ Leaper, JC Horrocks, JR Staniland, and AP McCain. "Computer Aided Diagnosis of Acute Abdominal Pain", *British Medical Journal*, 1972, Vol 2, pp 9-13.

[5] FT de Dombal, DJ Leaper, JC Horrocks, JR Staniland, and AP McCain. "Computer Aided Diagnosis: Description of an Adaptable System, and Operational Experience with 2034 Cases", *British Medical Journal*, 1972, Vol 2, pp 5-9.

[6] FV Jensen, An Introduction to Bayesian Networks, *Springer-Verlag New York, Inc.*, 1996.

[7] M Goldbaum, S Moezzi, A Taylor, S Chatterjee, J Boyd, E Hunter and R Jain, "Automated Diagnosis and Image Understanding with Object Extraction, Object Classification and Inferencing in Retinal Images", in the proceedings of *International Conference on Image Processing*, 1996, Vol 3, pp 695-698.

[8] R Greiner, C Darken and NI Santoso, "Efficient Reasoning", *ACM Computing Surveys*, March 2001, Vol 33, pp 1-30.

[9] WE Hart, M Goldbaum, B Cote, P Kube, MR Nelson, "Measurement and classification of retinal vascular tortuosity", *Intl J Medical Informatics*, 1999, 53:239+.

[10] D Heckerman. Probabilistic Similarity Networks, *The MIT Press*, 1990.

[11] D Heckerman. "A Tutorialon Learning With Bayesian Networks", Technical Report, *Microsoft Research*, March 1995.

[12] A Hoover and M Goldbaum, "Fuzzy convergence", in the proceedings of IEEE Computer Vision and Pattern Recognition, 1998, pp. 716-721.

[13] A Hoover, V Kouznetsova and M Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response", in the proceedings of *AMIA Annual Symposium*, 1998.

[14] A Hoover, V.Kouznetsova and MH Goldbaum, "Locating Blood Vessels in Retinal Images by Piece-wise Threhsold Probing of a Matched Filter Response", *IEEE Transactions on Medical Imaging*, Vol. 19 no. 3, March 2000, pp. 203-210.

[15] EJ Horovitz , JS. Breese and M.Henrion. "Decision Theory in Expert Systems and Artificial Intelligence", *International Journal of Approximate Reasoning*, 1988, Vol 2, pp 247-302.

[16] Hugin Expert A/S. Hugin GUI Version 4.5, 1996.

[17] JT McClave, F.H.Dietrich II and T.Sincich. Statistics, *Prentice Hall*, 1997.

[18] J Pearl. Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference. *Morgan Kaufmann Publishers*, 1988.

[19] M Pradhan, M Henrion, G Provan, BD Favero and K Huang, "The sensitivity of belief networks to imprecise probabilities: an experimental investigation", *Artificial Intelligence*, 1996, Vol 85, pp 363-397.

[20] EH Shortliffe, "Computer Based Medical Consultations: MYCIN", *Elsevier North Holland Inc.*, 1976.

[21] WW Daniel. Biostatistics- A Foundation for Analysis in the Health Sciences, *John Wiley & Sons,Inc*,1999.

[22] NL Zhang and D Poole. "Exploiting Causal Independence in Bayesian Network Inference", *Journal of Artificial Intelligence Research*, 1996, Vol 5, pp 301-328.