A PROBABILISTIC INFERENCE ALGORITHM FOR EARLY DETECTION OF AGE RELATED MACULAR DEGENERATION

Harshini Manoharan
Department of Computer Science,
CHRIST (Deemed to be University), Bengaluru, Karnataka, India
harshini.m@mca.christuniversity.in

Dr. Rajesh R
Associate Professor, Department of Computer Science,
CHRIST (Deemed to be University), Bengaluru, Karnataka, India
r.rajesh@christuniversity.in

Abstract – Age Related Macular Degeneration or ARMD is a retinal disorder that causes blindness over people of older age group. ARMD is associated with age and is a leading cause of blindness around the world. There is no specific medicine to fully cure ARMD but its development can be controlled by regular exercises and a healthy lifestyle if it is detected early. With a rising population of old age group of people, it becomes important to detect ARMD as early as possible in order to contain its development further. This research attempts to develop an algorithm based on probabilistic inference through Bayesian Network by analyzing large datasets collected from previous cases where datasets include elements of risk factors that could cause ARMD along with eye images. Unlike most of the approaches in detecting ARMD this work not only analyses eye images but also includes analysis of various factors causing the disorder. To include the study and analysis of the presence of factors causing ARMD is sensible because those factors are good indicators when the need is an early detection.

Keywords: *Age Related Macular Degeneration, Probabilistic Inference, Bayesian Network*

I. Introduction

Macular degeneration also referred to as Age Related Macular Degeneration or simply ARMD or AMD, is an eye condition that results in blurred or no vision in the center of the visible field known as the macula. Macula which is a pigmented region located near the posterior central portion of the retina is responsible for the central vision field. ARMD is caused by the formation and presence of drusen, cellular polymorphous debris in the macula. There are often no early symptoms. In some people one or the both eyes may be affected by experiencing a gradual worsening of the vision over time. Sometimes it causes visual hallucinations, but these do not involve mental illness. Problems in central vision can make people to feel hard to recognize faces, read, or perform other day to day activities of life although it does not result in complete vision loss [1].

ARMD has affected almost 6.2 millions of people around the world as of 2015 [3]. In 2013 it was the fourth most common cause of blindness after cataracts, preterm birth, and glaucoma [2]. ARMD is found in people over the age of fifty and in the United States it is the most common cause of visual impairment of this age group [1][4]. About 0.4% of people aged between 50 and 60 have this condition, 0.7% of those people aged 60 to 70, 2.3% of those aged 70 to 80, and nearly 12% of people older than 80 years [4].

Genetics and smoking are considered to be the major risk factors of ARMD while Age is also considered to be the major factor because ARMD typically occurs in people of older age groups due to the damage caused in the macula of the retina [1]. ARMD is broadly classified into early, intermediate, and late types [1]. The late type is further sub- divided into "dry" and "wet" category with the dry condition contributing up to 90% of the cases [1] [4].Usually people may not notice ARMD until they have distractions in the visual field. Such slow development of ARMD makes it often difficult for detection. There is only a limited option of treatment available and when the disease is in the lateral stages, there are only limited options of treatment available so it is very much important to diagnose the disease as early as possible [5]. Once the disease is detected in its early stage it can be prevented from further development by leading a healthy lifestyle with proper exercises.

This research attempts to develop an algorithm based on probabilistic inference in Bayesian Network to detect ARMD at early stages. The algorithm infers a probability in Bayesian Network based on analysis of risk factors that could cause ARMD and fundus images from a variety of datasets of the previous cases. We use fundus images because they are taken from the back of the eyes where it is comfortable to spot macula or retinal region. There are notable works in early detection of ARMD [6][7][8] and very few of these works were able to detect ARMD with good accuracy. But these works in most cases do not address the problem of early detection and also there are very few works which give attention to risk factors that could cause ARMD. Our work follows a different way of approaching the problem of early detection. We included the analysis of presence of risk factors in the subject along with eye images whereas most of the existing literatures use only eye images with different techniques for the detection.

Our aim is to detect the disease at its early stage while ARMD has very little symptoms and it has a very slow progression in most cases. Presence of risk factors causing the disorder in the subject's lifestyle may very well be a good indicator for early detection. Considering these facts about ARMD it is useful and sensible to include data about risk factors along with fundus images. In general, diagnosis like this requires and is conducted in clinical environments. Our work attempts to develop an algorithm for a system, say a piece of software, that does not require any special environment, is able to be equipped to run on standalone computers. Such systems can be deployed in eye care centers, health care center's etc., where a clinical environment is difficult to access.

II. Methodology

This method extensively uses Bayesian Networks or BN. A Bayesian Network shows the causal probabilistic influence between a set of random variables, their conditional dependences and it is a compact specification of a joint probability distribution [9].

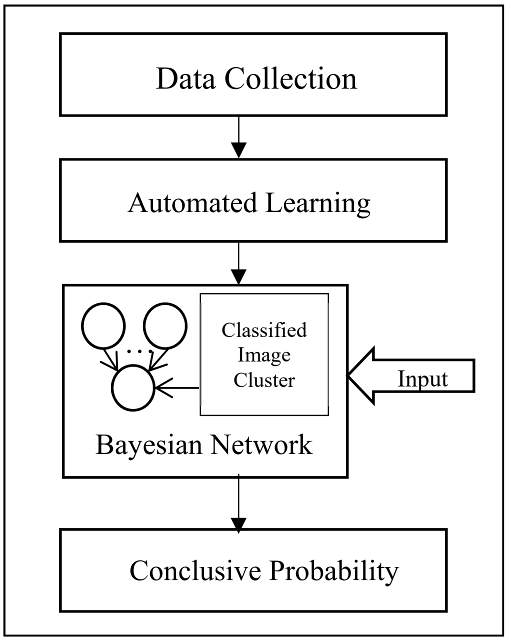


Figure 1: An Overview of the Algorithm

The primary components of a BN are a directed acyclic graph and a set of conditional probability distributions [11]. The graph is a set of nodes where each node represents a random variable in the problem domain. Every node has a conditional probability distribution, that is, the probability (possibility) to precede one or more nodes [11]. In BN if a node A causes node B then there is a probabilistic dependence between the nodes and while making the graphs such nodes are connected by directed edges [11]. We chose Bayesian Network because of its ability to establish a causal relationship between random variables under a complex uncertainty.

The goal of the algorithm is to find the probability of presence of the symptoms, that is, the presence of factors causing ARMD including eye images. We may want to find the probability of ARMD given the factors. This is often complicated to calculate the probability of a condition or disease based on any causes. Because this involves much more variables such as intensity, severity and presence of any other influential causes. But it is practical to calculate the reverse of this conditional probability, that is probability of the symptoms (cause) if the subject has the disease or condition (evidence). Bayesian Network helps to find the probability of every possible cause (by posterior probability distribution) for the given observed evidence.

Firstly, we need data from which BN can be constructed. Data are large, random and varied sets. Datasets include information about normal people and variety of information about ARMD affected people with different stages (see the data considerations chapter). These datasets can serve our purpose to construct the BN.

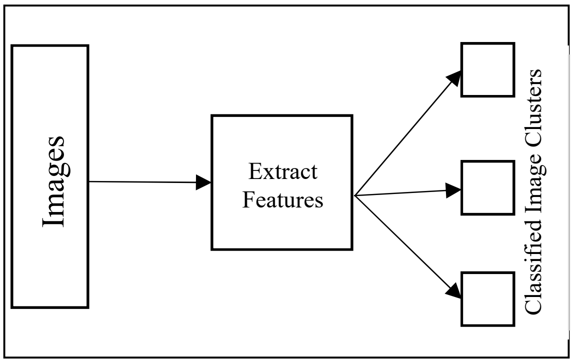


Figure 2: Schematic representation of how Image Clusters are classified

Now that we have prepared the data, BN can be constructed out of it. The process of creating BN graphs, that is, set of nodes from given data is called learning [11]. Learning can be done manually by creating node structures after analyzing the data, select random variables and find the probability distributions. Alternatively use tools that could create BN structures. The latter is called automatic learning. For this work, we tend to use automatic learning. We chose to invoke an automatic learning because 1) Automatic learning doesn't require underlying domain knowledge 2) Wide variety of readily available tools [11]. These tools are generally software (for example, GeNIe by BayesFusion) or one can be written with any of the high level programming languages (for example, Python. Automatic learning is capable of creating node structures and finding the relevant conditional probability distributions with software advantages.

Next, we focus on establishing probability distributions for eye images. It is significant to learn the features of normal eye images and ARMD affected eye images of various stages to confirm the presence of drusen but finding probabilities for every image may be meaningless, considering images are not random variables. To address this issue in constructing BN nodes for images, this work, through its automated learning adds one more tier to the node structure. We adapt a slightly different version of J. Luoa et. al.'s [13] work of understanding semantic features of images in Bayesian Network. We extract common features between multiple images that are of same category and put them into the same cluster. We call it as Classified Image Cluster. For example, by extracting common features of normal eye images and put them into a cluster. Similarly, more clusters can be created for eyes images that are early, late or intermediate ARMD affected. Formation of more clusters (that is we try to create more categories) can yield better accuracy. Now these clusters can be treated as nodes for which probability distributions are derived. Hence these clusters become part of the BN node structure.

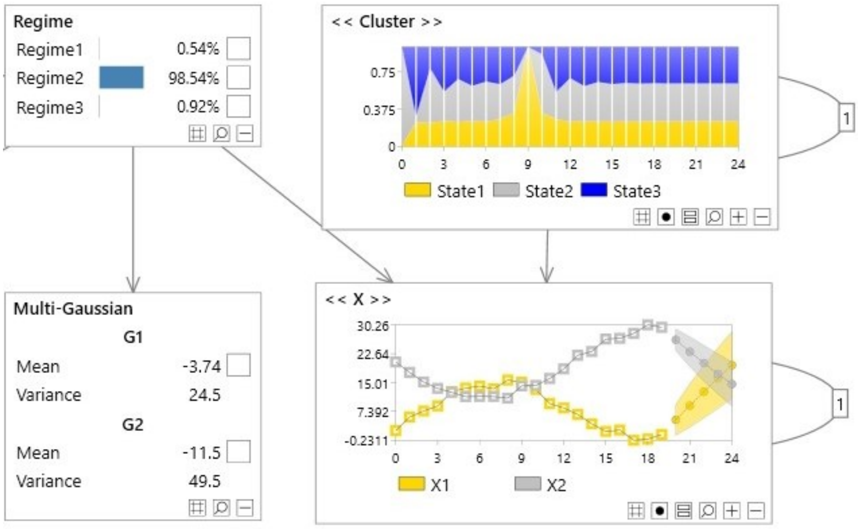
Upon construction of the BN, because of the influences and probabilistic relations among variables can be described readily in a BN[12], we assume that the subject has the condition, that is, we are setting evidence. According to Bayes theorem, let the graph be **G** (over one or more variables), with our new evidence **e**,

$$\mathrm{P(G|e)} = \frac{\mathrm{P(e|G)\,.\,P(G)}}{\mathrm{P(e)}} \qquad \textbf{(1)}$$

the term P(G|e) is the posterior probability distribution of G and the term P(G) is the prior probability or marginal probability distribution.

III. Implementation & Results

For this work, we incorporate the AREDS (Age Related Eye Disease Study) dataset of the NEI (National Eye Institute). AREDS is one of the comprehensive studies available with over 4000 participants in our desired age group for this work. AREDS is a study which lasted over 5 years and results to a rich, varied and large dataset. On implementation of our early detection algorithm we also make use of BayesServer, a software API that supports multiple platforms and processing dataset efficiently in Bayesian Network models. This work adapts Java API of BayesServer. BayesServer API is capable of learning a Bayesian network from given data that can be used effectively to predict, diagnose or to automate our decision making. Fig. 3 shows a sample of how results in BayesServer API appear (Image credit: bayesserver.com).



Regime
Regime1 0.54%
Regime2 98.54%
Regime3 0.92%

Multi-Gaussian
G1
Mean -3.74
Variance 24.5
G2
Mean -11.5
Variance 49.5

Figure 3: Results in BayesServer API

The main aim of our work is to develop a standalone piece of software that uses the proposed algorithm with AREDS dataset in BayesServer API environment. The following figure is the screenshot taken out of BayesServer environment upon constructing the network out of AREDS dataset. Nodes are represented as boxes and the edges show the relationship that is conditional probability of that event to happen from the previous node. These edges are derived through automated learning of BayesServer environment. This is the outcome of the clustered learning.

Through this work we are able to observe some satisfying results. We follow the following steps:

Step 1: Training and Query

We take our data into a training set and a query set. Training sets are used to train models and evaluate it against the query set to arrive at a decisive probability.

Step 2: Training, Query & Validation

Since we work with different clusters during the construction of the network in our work, the usage of mere training and query sets alone is not going to be enough, so we add a validation set, to address the dependency on the query sets. These validation sets could also help the system to get better results over time.

Step3: Cross validation

Cross validation is necessary, especially to the kind of algorithm we are proposing here that segments data into multiple clusters. For each cluster c, a model is evaluated on c to have been trained on all the data excluding c. Thereby we evaluate the network through all of the data available, still it remains hidden to each model.

We estimate the accuracy of our model using confusion matrix in Bayes networks. A confusion matrix is a square matrix in which rows and columns represent actual and predictions respectively. Given below is the resultant confusion matrix obtained from our training dataset,

Table 1: Confusion Matrix in Bayes Network

	PREDICTED	PREDICTED
ACTUAL	59	3
ACTUAL	8	66

Then accuracy is given by:

$$\text{Accuracy}=\frac{(\text{No.of True Positives}+\text{No.of True Negatives})}{(\text{No.of True Positives}+\text{No.of False Positives}+\text{No.of True Negatives}+\text{No.of False Negatives})}$$

$$\text{Accuracy}=\frac{(59+66)}{(59+5+8+66)}=90.57 \qquad \textbf{(2)}$$

Our initial query and validation sets with 10 training sets in a BN constructed out of AREDS yielded over 90% (based on outcomes of confusion matrix for each training set) accuracy through this work.

IV. Data considerations and compliances

For the development of this algorithm, we need ARMD affected patient's data. Datasets consist of various personal information such as age, gender, race etc., biological information such as height, weight etc., habitual information such as diet, smoker, drinker etc. and eye images. We add similar information of normal people or people that may be tested negative for ARMD to the datasets to increase the reliability of Bayesian Network. While predicting the probability the sample space needs to be comprehensive and random. Fortunately, over the time hospitals have generated enough data and in clinical practices, it is common to collect the information mentioned above. As an alternate, the organizations that conduct clinical studies or support clinical studies by providing Statistical Analyzing Services (SAS) possess large and comprehensive datasets. Since our work involves an automated learning, it can utilize multiple formats of data (like csv, SAS etc.) [11].

Datasets are large, rich and varied. We emphasize large datasets because machine learning algorithms make use of the rich, varied datasets and relate them by finding high dimensional interactions among datasets [10]. Sample sizes are large enough to include variety of data consisting various information about the AMD affected and the unaffected. Automated learning in the algorithm requires a bit more data to overcome missing data problem in datasets while constructing Bayesian Network [11]. When it comes to applications that require more data, security considerations are high. As a measure the following steps are taken. 1) Data are anonymous 2) No prospect data collection from patients 3) Only necessary data are collected 4) Encrypted and secure Storage 5) Destroy data when no longer needed 6) Compliance with clinical terminologies 7) Compliance with security, data transfer standards and laws that vary country to country.

V. Conclusion

Through this work we expressed an algorithm for early detection of ARMD that makes the best use of Bayesian Networks with an accuracy of over 90% through our initial training sets. There are very less literature available in the area of machine learning when it comes to clinical and biological solutions. On that interest, we still have a very long way to go. It is our aim to study, analyze and provide more exploratory solutions especially for clinical problems in our future works. We continue to refine the proposed algorithm and the system in our future works. The outcomes of the system are to be carefully studied. Usually machine learning systems get better as time progresses and so the outcomes are more accurate with time. Also, we continue to study ARMD and how machine learning could be equipped to provide better solutions for biological and clinical problems.

Acknowledgement

The Authors thank the Management and the Head of the Department of CHRIST (Deemed to be University), Bangalore for permitting and supporting us to carry out the research. We would also like to thank Mr. Suriya Prakash B, Software Engineer and Mr. Ullas, Little Flower Hospital, for helping us in our research.

References

[1] Bressler NM, "Age-related macular degeneration is the leading cause of blindness" JAMA 2004; 291(15):1900-1901 DOI:10.1001/jama.291.15.1900.

[2] Global Burden of disease Study 2013 Collaborators "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013", The Lancet. 386 (9995): 743–800. August 2015, DOI: 10.1016/S0140-6736(15)60692-4.

[3] GBD 2015 Disease and Injury Incidence and Prevalence Collaborators (October 2016), "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015", The Lancet. 388 (10053):1545–1602, DOI:https://doi.org/10.1016/S0140-6736(16)31678-6, PMID 27733282.

[4] Mehta S, "Age-Related Macular Degeneration", Prim Care, 2015 September;42(3):377-91, DOI: 10.1016/j.pop.2015.05.009, PMID 26319344.

[5] Ratnapriya, R, and E Y Chew, "Age-Related Macular Degeneration– Clinical Review and Genetics Update", Clinical Genetics 2013 August;84(2):160-6, DOI: 10.1111/cge.12206 PMC. Web. 22 Apr. 2016.

[6] P. Burlina, D.E.Freund, N.Joshi, Y.Wolfson, N.M.Bressler, "Detection of Age-related Macular Degeneration via Deep Learning", IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2016 - Proceedings (Vol. 2016-June, pp. 184-188). [7493240] IEEE Computer Society, DOI: https://doi.org/10.1109/ISBI.2016.7493240.

[7] Liang Z, Wong DW, Liu J, Chan KL, Wong TY, "Towards automatic detection of age-related macular degeneration in retinal fundus image", 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, August 31 - September4,2010,**DOI:** 10.1109/IEMBS.2010.5627289.

[8] S Albert Jerome, Vyshakh Asokan, "Computer Aided Approach for Detection of Age Related Macular Degeneration from Retinal Fundus Images", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT] March 2016, DOI:10.1109/ICCPCT.2016.7530348.

[9] Murphy K. (1998), "A Brief Introduction to Graphical Models and Bayesian Networks".

[10] Kevin P Murthy, "Machine Learning: A Probabilistic Perspective", Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts, MIT Press; 2012.

[11] Michal Horný, "Bayesian Networks", Boston University, April 18, 2014.

[12] Lucas P, "Bayesian Networks in Medicine: a Model-based Approach to Medical Decision Making", University of Aberdeen, December 2001.

[13] Jiebo Luoa, Andreas E. Savakisb, Amit Singhal, "A Bayesian network based framework for semantic image understanding", Pattern Recognition 38 (2005), pp:919 – 934.