# Bayesian network modelling on data from fine needle aspiration cytology examination for breast cancer diagnosis

Shuo Liu[1, a], Jinshu Zeng[2], Yuhua Wang[3], Hongqin Yang[3, c], Yuhua Li[4], Liam Maguire[5], Jia Zhai[6], Yi Cao[7], Xuemei Ding[1, 5, b]

[1] Software Faculty, Fujian Normal University, Fuzhou, 350108, China

[2] Department of Ultrasonic Medical, The First Affiliated Hospital of Fujian Medical University, Fuzhou, 350005, China

[3] Fujian Provincial Key Laboratory for Photonics Technology, Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Fuzhou, 350007, China

[4] School of Computing, Science and Engineering, University of Salford, Manchester, M5 4WT, UK

[5] Faculty of Computing and Engineering, Ulster University, Londonderry, BT48 7JL, UK

[6] Business School, University of Salford, Manchester, M5 4WT, UK

[7] Department of Business Transformation and Sustainable Enterprise, Surrey Business School, University of Surrey, Surry, GU2 7XH, UK

[a]shuoliv@163.com [b]xuemeid@fjnu.edu.cn [c]hqyang@fjnu.edu.cn

**Keywords**: Bayesian networks, Data modelling, Quantitative analysis, Breast cancer diagnosis

**Abstract:** The paper employed Bayesian network (BN) modelling approach to discover causal dependencies among different data features of Breast Cancer Wisconsin Dataset (BCWD) derived from openly sourced UCI repository. K2 learning algorithm and k-fold cross validation were used to construct and optimize BN structure. Compared to Naïve Bayes (NB), the obtained BN presented better performance for breast cancer diagnosis based on fine needle aspiration cytology (FNAC) examination. It also showed that, among the available features, bare nuclei most strongly influences diagnosis due to the highest strength of the influence (0.806), followed by uniformity of cell size, then normal nucleoli. The discovered causal dependencies among data features could provide clinicians to make an accurate decision for breast cancer diagnosis, especially when some features might be missing for specific patients. The approach can be potentially applied to other disease diagnosis.

## Introduction

Breast cancer has become one of the most terrible cancer types in China. By 2008, there are more than 210 thounsand people died from breast cancer which has become the sixth killer in all kinds of cancer for Chinese women [1,2]. It is important for patients to be early diagnosed and treated using appropriate methods [3]. Bayesian network (BN) modelling approach has been extensively used for diagnosis of breast cancer. Kalet et al. [4] used a Bayesian model to detect misdiagnoses made at the initial stage of diseases, such as lung, brain and female breast cancer. Wang et al. [5] proposed a three-layer BN for the earlier diagnosis of breast cancer. Hassen et al. [6] used Bayesian network to estimate the risk of metastasis for breast cancer patients. However, there is a lack of report on quantitative analysis among different breast cancer features, neither any research founded about the causal dependency between any pair of different features. A clearly explanation of BN in a specific domain, such as healthcare, can well understand the disease pathology and provide valuable diagnostic decision for domain experts. In this paper, breast cancer features based on fine needle aspiration cytology (FNAC) derived from Breast Cancer Wisconsin Dataset (BCWD) of UCI repository [7] were analyzed using BN modelling approach to dicover the causal relationships

between diagnostic results and different FNAC features, as well as the causal relationships among the features.

## Methods and Materials

We employed widely used BN modelling approach to discover causal dependencies among breast cancer FNAC features, and further quantitate the dependencies.

### Bayesian network

Bayesian network (BN) contains BN structure which was represented by a directed acyclic graph (DAG) and conditional probability tables (CPTs). A BN construction consists of two key steps: BN structure learning and parameters learning. In order to obtain the most reasonable BN structure which fits better the dataset, many algorithms have been developed [8]. K2 learning algorithm [9] is one of search-and-score methods. It starts with a set of ordering variables, and each node has no parents initially. Scoring process iterates based on the certain ordered variables. The node with the highest score will then be added incrementally into the resulting structure as parent node. K2 only needs to consider a subset of DAG and can quickly find the variable with local maximal score. We used K2 learning algorithm to construct a BN structure.

BN structure can visualize the relationship between different nodes, and the strength of the influence can be visualized using a static technique proposed by Koiter [10]. The static technique can show the influence while ignoring direction of an arc in BN. Fig. 1 shows two possible situations as follows: (1) arc from a target node to a non-target node, e.g., Fig. 1 (a) shows the influence that node B has on node A; (2) arc between two non-target nodes, e.g., Fig. 1 (b) shows the influence in both directions, but the influence is determined by average of those they have on each other.

The strength of the influence (SI) can be measured by calculating the distance of different posterior probability distributions, i.e.,

$$\mathrm{SI}=\frac{1}{\mathrm{n}}\sum_{i=0}^{n} D\left(P\left(A\right),P\left(B\middle|A=a_i\right)\right). \tag{1}$$

Where n is the number of states about a node. *A* and *B* are the directly linked nodes. *P(A)* is the priori probability of variable *A*, and $P(B|A=a_i)$ is the posterior probability of B, given a certain state of *A*. *D* is the Euclidean distance function between these two probability distributions.

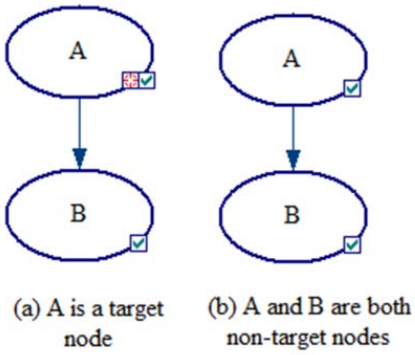


Fig. 1 Two kinds of relationship between two nodes.

### Materials

The BCWD data contains totally 699 records (458 benign and 241 malignant records). Each records was described by ten features: clump thickness (CT), uniformity of cell size (UCSI), uniformity of cell shape (UCSH), marginal adhesion (MA), single epithelial cell size (SECZ), bare nuclei (BN), bland chromatin (BC), normal nucleoli (NN), mitoses (MITO) and corresponding diagnostic categories. The diagnosis feature labelled benign and malignant for each record with 1and 2, respectively. Other features were scored using an integer value ranging from 1 to 10, where 1 represents the most benign characteristic and 10 represents the most malignant characteristic.

According to the data description, BCWD were discretized into 2 categories due to binary diagnostic categories. We considered 1 as benign and the other descriptions (2-10) as malignant for each feature. K2 learning algorithm requires ordering features, therefore, we applied information gain algorithm to rank each feature. The obtained information gain score (IGs) of each feature was listed (from the highest ranking) in Table 1. All experiments were carried out in Weka [11] and GeNIe [12].

Table 1 BCWD dataset description and information gain score of each feature

| Features | Values | IGs |
|---|---|---|
| Bare Nuclei | 1-10 | 0.571 |
| Uniformity of Cell Size | 1-10 | 0.525 |
| Single Epithelial Cell Size | 1-10 | 0.494 |
| Uniformity of Cell Shape | 1-10 | 0.475 |
| Normal Nucleoli | 1-10 | 0.416 |
| Marginal Adhesion | 1-10 | 0.347 |
| Bland Chromatin | 1-10 | 0.323 |
| Clump thickness | 1-10 | 0.291 |
| Mitoses | 1-10 | 0.199 |

**Results and discussion**

The BN was estimated by 5-fold cross validation. Each fold contains 355 training data (230 benign samples and 125 malignant samples) and 344 testing data (228 benign samples and 116 malignant samples). The obtained BN is shown in Fig. 2 where each rectangle box stands for a feature node. The arrows indicate the causal influences between two features, and the thickness of the arrow represents the strength of the corresponding influence. The thicker the arrow, the stronger the influence.

The strength of influences between diagnosis and features were shown in Fig. 3, which reflects different features‘ importance to diagnosis. BANU would be the most important feature, followed by UCSI, then NN, while SECS has the weakest influence to diagnosis. According to the thickness of arcs between breast cancer features, we could know, UCSI has the strongest dependency on UCSI, followed by MA, and then NN, in terms of the corresponding SI being 0.537, 0.336 and 0.326, respectively.

Some literatures, such as [5, 13], supposed features of breast tumor were independent. According to this assumption, we trained and tested Naïve Bayes (NB) model also based on 5-fold cross validation. Diagnostic performances of both NB and our proposed BN are list in Table 2, with respect to classification accuracy, sensitivity, specificity and the AUC (Area under the ROC curve) metrics. Due to a higher accuracy of 0.931, a higher sensitivity of 0.930, a higher specificity of 0.941 and a higher AUC of 0.971, our BN showed a more competitive performance than the NB.

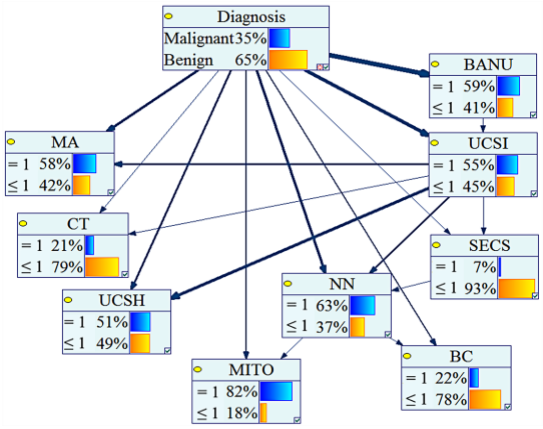


Fig. 2 Bayesian Network constructed on the BCWD

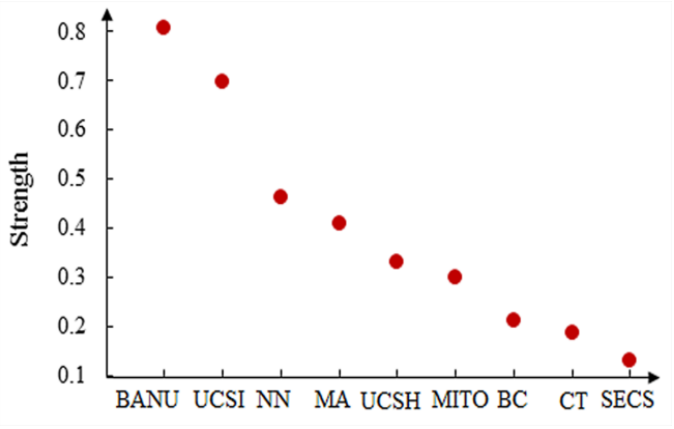


Fig. 3 The strength of the influence of each feature in terms of diagnosis

Table 2 The performance comparison between the BN and the NB

| Model | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| BN | *0.931* | *0.930* | *0.941* | *0.971* |
| NB | 0.918 | 0.918 | 0.932 | 0.964 |

## Conclusions

The paper mainly made two contributions: first, it discovered the most valuable FNAC features for diagnosis of breast cancer. Second, it found the causal relationships among different features. The former proves that BN modelling approach can be used to quantitatively discover which feature is important for disease diagnosis. The latter illustrates that BN can quantitatively analyze the dependencies among different disease features. The findings can help domain experts make more accurate and objective decisions, especially given less number of features, for breast cancer diagnosis. Our potential work will test our method on different data types collected from local hospital.

## Acknowledgements

This work was partly supported by the National Key Basic Research Program of China (2015CB352006), the National Natural Science Foundation of China (61335011), Scientific Research Founds for the Returned Overseas Chinese Scholars, State Education Ministry, Young Key Program of Education Department, Fujian Province, China (JZ160425), Program of Education Department of Fujian Province, China (I201501005) and the Program for Changjiang Scholars, Innovative Research Team in University (IRT_15R10).

## References

[1] W.Q. Chen, R. Zheng and S.W. Zhang. Ca Cancer J Clin 66 (2016) p.115-132

[2] L. Fan, K.S. Weippl, J.J. Li, J. St. Louis, D.M. Finkelstein, K.D. Yu, W.Q. Chen, Z.M. Shao and P.E. Goss, Lancet Oncol. 15 (2014) p. 279-89

[3] D.R. Chen, Y.L. Huang and S.H. Lin. Computerized Medical Imaging and Graphics 35 (2011) p. 220-226

[4] A.M. Kalet, J.H. Gennari and E.C. Ford. Phys. Med. Biol. 60 (2015) p. 2735-2749

[5] X.H. Wang, B. Zheng and W. F. Good. International Journal of Medical Information 54 (1999) p. 115-126

[6] H.B. Hassen, I. Kallen and L. Bouchaala, International Journal of Biomathematics 6 (2013)

[7] Information on http://archive.ics.uci.edu/ml/

[8] R. Daly, Q. Shen and S. Aitken, The Knowledge Engineering Review 26 (2011) p. 99-157

[9] G. E. Cooper and E. Herskovits, Machine Learning 9 (1992) p. 309-347

[10] J.R.Koiter, Nature Biotechnology 24 (2006) p. 39-61

[11] Information on http://weka.wikispaces.com/

[12] Information on https://www.bayesfusion.com/

[13] M. Karabatak, Measurement 72 (2015) p. 32–36