



University of
Salford
MANCHESTER

Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach

Liu, Shuo, Zeng, Jinshu, Gong, Huizhou, Yang, Hongqin, Zhai, J, Cao, Yi, Liu, Junxiu, Luo, Yuling, Li, Yuhua, Maguire, Liam and Ding, Xuemei

<http://dx.doi.org/10.1016/j.combiomed.2017.11.014>

Title	Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach
Authors	Liu, Shuo, Zeng, Jinshu, Gong, Huizhou, Yang, Hongqin, Zhai, J, Cao, Yi, Liu, Junxiu, Luo, Yuling, Li, Yuhua, Maguire, Liam and Ding, Xuemei
Type	Article
URL	This version is available at: http://usir.salford.ac.uk/44659/
Published Date	2018

USIR is a digital collection of the research output of the University of Salford. Where copyright permits, full text material held in the repository is made freely available online and can be read, downloaded and copied for non-commercial private study or research purposes. Please check the manuscript for any further copyright restrictions.

For more information, including our policy and submission procedure, please contact the Repository Team at: usir@salford.ac.uk.

Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach

Shuo Liu^{1,*}, Jinshu Zeng^{2,*}, Huizhou Gong³, Hongqin Yang^{4,†}, Jia Zhai⁵, Yi Cao⁶, Junxiu Liu⁷,
Yuling Luo⁸, Yuhua Li⁹, Liam Maguire⁷, Xuemei Ding^{1,7,†}

¹ Faculty of Mathematics and Informatics, Fujian Normal University, Qishan Fuzhou, 350108, China

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

³ College of Foreign Languages, Fujian Normal University, Cangshan Fuzhou, 350007, China

⁴ Fujian Provincial Key Laboratory for Photonics Technology, Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Cangshan Fuzhou, 350007, China

⁵ Business School, University of Salford, Manchester, M5 4WT, UK

⁶ Department of Business Transformation and Sustainable Enterprise, Surrey Business School, University of Surrey, Surrey, GU2 7XH, UK

⁷ Faculty of Computing, Engineering, and Built Environment, Ulster University, Londonderry, BT48 7JL, UK

⁸ Faculty of Electronic and Engineering, Guangxi Normal University, Guilin, 541004, China

⁹ School of Computing, Science and Engineering, University of Salford, Manchester, M5 4WT, UK

Abstract

Background: Breast cancer is the most prevalent cancer in women in most countries of the world. Many computer-aided diagnostic methods have been proposed, but there are few studies on quantitative discovery of probabilistic dependencies among breast cancer data features and identification of the contribution of each feature to breast cancer diagnosis.

Methods: This study aims to fill this void by utilizing a Bayesian network (BN) modelling approach. A K2 learning algorithm and statistical computation methods are used to construct BN structure and assess the obtained BN model. The data used in this study were collected from a clinical ultrasound dataset derived from a Chinese local hospital and a fine-needle aspiration cytology (FNAC) dataset from UCI machine learning repository.

Results: Our study suggested that, in terms of ultrasound data, cell shape is the most significant feature for breast cancer diagnosis, and the resistance index presents a strong probabilistic dependency on blood signals. With respect to FNAC data, bare nuclei are the most important discriminating feature of malignant and benign breast tumours, and uniformity of both cell size and cell shape are tightly interdependent.

Contributions: The BN modelling approach can support clinicians in making diagnostic decisions based on the significant features identified by the model, especially when some other features are missing for specific patients. The approach is also applicable to other healthcare data analytics and data modelling for disease diagnosis.

* These authors contributed equally.

† Corresponding authors: xuemeid@fjnu.edu.cn, x.ding@ulster.ac.uk (Xuemei Ding); hqyang@fjnu.edu.cn (Hongqin Yang).

Keywords: Clinical decision support, Data modelling, Bayesian network, Quantitative analysis, Diagnostic contribution, Breast cancer diagnosis

1. Introduction

Breast cancer is the most prevalent cancer in women around the world. It has been reported that approximately 1.3 million women worldwide have been diagnosed with breast cancer since 2011, and approximately 465,000 women die from breast cancer each year [1]. In China, 214,360 women had died from breast cancer by 2008. It has been estimated that the number of Chinese women with breast cancer will reach 2.5 million by 2021 [2]. According to a report published by the Chinese National Cancer Centre in 2017, breast cancer is the most common cancer in Chinese women. Following lung, stomach, liver, oesophageal and colorectal cancers, breast cancer is the sixth largest killer in small- and medium-sized cities, with a mortality rate of 8.44% and 9.59%, respectively, while the mortality rate from breast cancer in large-sized cities is 12.78%, making it the fifth most common cause of death among all cancer types in Chinese women [3]. Due to the rapid increase in the number of breast cancer patients, early identification of women at risk of developing breast cancer is currently an international priority [4].

In order to improve diagnostic accuracy and help domain experts to make more effective decisions, many computer-aided diagnosis (CAD) systems have been developed [5]-[7]. They provide new computational algorithms combined with domain knowledge to support clinical diagnosis. Zeng et al. [8] proposed different nonlinear state-space models for lateral flow immunoassay, which have been commonly used in clinical diagnosis. In clinical medicine, breast cancer could also be diagnosed via several different techniques, such as ultrasound, fine-needle aspiration cytology (FNAC) and magnetic resonance imaging (MRI) scanning.

Other CAD algorithms in [12] were also proposed to detect breast cancer. For example, Eltoukhy et al. [12] proposed a feature extraction method based on a statistical t -test for breast cancer diagnosis from a digital mammogram. They used wavelet and curvelet methods to transform digital

mammography data into vector coefficients and then employed a support vector machine (SVM) algorithm for breast cancer diagnosis. As a result, the highest diagnostic accuracy based on wavelet and curvelet coefficients was 96.56% with 1238 features and 97.30% with 5663 features [12].

As a well-established probabilistic classifier, Bayesian network (BN) analysis has been used widely for data analytics and data modelling in many healthcare area, such as psychotic depression [15], Alzheimer's disease [16], heart disease [17] and social anxiety [18] as well as breast cancer. Wang et al. [19] proposed a three-layer BN for earlier diagnosis of breast cancer. They assessed the performance of the BNs constructed based on non-imaging features, imaging features and both. They found that the BN built on both imaging and non-imaging features performed well and that imaging features dominated BN performance. In 2007, Nicandro et al. [20] evaluated the performances of seven BN classifiers (i.e. Naïve Bayes classifier, Bayes-N, MP-Bayes, Greedy, MP-Bayes + Greedy, PC which is a procedure contained the Tetrad, and a CBL2 algorithm in Power Constructor, which is a software package containing CBL1 and CBL2 algorithms) for breast cancer diagnosis based on fine-needle aspiration from a breast lesion collected by a single observer and multiple observers. They found that the classifiers learnt from different data performed differently, which indicated that the observations would impact the breast cancer diagnostic result. Furthermore, in 2009 [21], Nicandro and his team made use of two decision trees and four different BNs for breast cancer diagnosis. Their study discovered interobserver variability in breast cancer cytodiagnosis, indicating that different observers would focus on different perspectives while making a diagnostic decision. Kalet et al. [22] designed a Bayesian model to detect a misdiagnosis made at the initial diagnostic stage of a disease such as lung, brain or female breast cancer. The BN model they designed produced a better AUC (0.98) than a decision made by clinical experts (0.90).

Additionally, BN was also used in other studies of breast cancer, such as risk factor estimation [23], and causal interaction detection [24]. Nicandro et al. [25] employed a score-based BN approach to estimate the power of thermography for breast cancer diagnosis. The BNs were learned using

Naïve Bayes, hill-climber and repeated hill-climber algorithms with a minimum description length (MDL) metric. The BN learned by a repeated hill-climber algorithm provided the best accuracy for both cancer and non-cancer diagnosis ($75.50 \pm 6.99\%$) and sensitivity of cancer diagnosis (94%). Furthermore, their obtained BN identified five important features for breast cancer diagnosis: 1C (hottest point in only one breast), f unique (total number of hottest points), thermovascular network (number of veins with the highest temperature), curve pattern and asymmetry (temperature difference between the right and left breasts).

Although a BN modelling approach has been used for breast cancer diagnosis, a report on quantitative analysis among different breast cancer features, which is critically important for clinical decision making, is lacking. As a result, some researchers might ignore the relationships between different features, which may lead to a high misdiagnosis rate [19][26]. A clearly explained BN in a medical area can increase the understanding of disease pathology and provide valuable decision-making assistance to domain experts. This paper employed a BN modelling approach to discover the probabilistic relationships between different data features of breast cancer. We also analysed the contribution of each feature to breast cancer diagnosis. The data were focused on ultrasound and FNAC examinations obtained from The First Affiliated Hospital of Fujian Medical University, China and the Breast Cancer Wisconsin Dataset (BCWD) of the UCI machine learning repository [27].

BN modelling can be deconstructed into two sub-processes: structure learning and parameter learning. In this study, a K2 learning algorithm [28] with an MDL score metric was used to learn the BN structure. Our reasons for using a K2 algorithm were the following: 1) K2 is the most commonly used algorithm for BN structure learning [29], 2) K2 is relatively easy to implement [29], 3) K2 only needs to consider a subset of a directed acyclic graph (DAG) and can quickly find the variable with the local maximal score [30] and 4) a K2 algorithm makes good use of experts' knowledge to learn the BN structure.

The contributions of this study are 1) we discovered the most important features which can provide uninitiated observers and doctors objective and quantitative guidance to focus on specific features for early breast cancer diagnosis. 2) We analysed the probabilistic dependencies among different data features and identified the strength of the dependency, which can assist the domain experts in making a quantitatively accurate diagnosis, even using fewer available features. A focus on different features by different observers [21] may cause them to miss some important features, which can significantly influence diagnostic results. The above two contributions are helpful in decreasing the misdiagnosis rate. 3) Our study showed a potential translational application of the BN modelling approach to the breast cancer care pathway.

The remainder of this paper is organized as follows. Section 2 provides the basic theory of BN in detail, as well as a brief introduction about the technique of BN visualization. Section 3 presents the experimental results based on two real-world datasets. Section 4 discusses the results and evaluates the BN modelling approach in comparison with other methods. Finally, Section 5 concludes this paper and discusses potential extensions of the method in future work.

2. Methods and materials

Numerous approaches have been developed to support breast cancer diagnosis. BN analysis has been used widely to improve diagnostic accuracy and to discover probabilistic relationships among features and the influence of joint probability distribution inference.

2.1 Bayesian network

A BN represents a domain which explicitly provides a set of variables belonging to a specific domain and visualizes the relationships between the variables [31]. It can successfully represent uncertain knowledge in various fields [30]. A BN is usually represented using a DAG, $DAG = (V, E)$, where V denotes a set of nodes made up by a set of variables, and E denotes a set of edges between the nodes in V . No cycles are present in the DAG [32]. Each edge is directly linked from one node to

another, and it indicates that the corresponding two nodes are mutually dependent. Otherwise, nodes are independent if there is no link between them.

Consider a given dataset D containing a set of variables $X = \{X_1, X_2, \dots, X_n\}$, the joint probability distribution on X , $p(X_1, \dots, X_n)$, is defined as

$$p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | \pi_i) \quad (1)$$

where π_i is the set of parents of X_i .

2.1.1 Structure learning

BN modelling process contains structure learning and parameter learning. Structure learning aims at identifying the topology of the network in order to display the relationships among the nodes. Parameter learning quantitatively finds how a node relates to its parent nodes [33]. BN structure learning methods are normally classified into search-and-score-, constraint- and dynamic programming-based categories [30]. Compared with the other two categories, the search-and-score-based category is suitable for large data sets in the whole feature space to assist in finding an exact structural topology [30]. As the name implies, it needs a search strategy and a score metric [33].

The K2 learning algorithm [28] is one of the most commonly used search strategies. It starts with a set of ordering nodes, and each node initially has no parents. It then, according to a certain order, iteratively adds the parent nodes for the node of interest, i.e. if X_i is preceded by X_j ($i \neq j$), the edge from X_j to X_i is not allowed. Assume there is a set of ordering nodes, $X = \{X_1, X_2, \dots, X_n\}$, and each node has r_i states. π_i is initialized with empty at the beginning. The function $pred(X_i)$ represents a set of nodes preceding X_i in the X . The K2 learning algorithm will add the node X_m of $pred(X_i)$ into π_i when X_m makes the score f_{new} larger than the old score f_{old} . The formula of the scoring function f is given by equation (2) [28].

$$f(i, \pi_i) = \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} N_{ijk} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \quad (1 \leq i \leq n) \quad (2)$$

where q_i is the number of all possible values of π_i ; N_{ijk} is the number of cases in the given dataset in which the node X_i is in the k th state, and its parent π_i is in the j th state; and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

2.1.2 Parameter learning

The parameters of BN are denoted as θ_{ijk} , which is the conditional probability distribution of node X_i when it takes the k th value and its π_i takes the j th value, i.e.

$$\theta_{ijk} = p(X_i = x_k | \pi_i = j) \quad (3)$$

This learning process is implemented by expectation-maximization (EM) algorithm [34].

2.1.3 Strength of influence

BN structures the probabilistic dependency between different nodes, which can be explained in different ways [35]. Koiter [36] proposed a comprehensive technique to visualize the inference in a BN in order to clearly understand the constructed BN. The technique can show the strength of the influence between two directly linked nodes by automatically adjusting the thickness of the corresponding arcs. Koiter made use of a dynamic model to show this influence without considering the direction of the arc in a BN. The proposed dynamic model is illustrated in Fig.1, where each node represents a variable. The red crossed sign denotes the node as a target node, and the green question mark means that the BN has not yet been updated. Four possible cases in Fig.1 are as follows: (1) an arc from a target node A to a non-target node B, e.g. the influence B-to-A (Fig.1 (a)); (2) an arc from a non-target node A to a target node B, e.g. the influence A-to-B (Fig.1 (b)); (3) an arc between two target nodes (Fig.1 (c)) and (4) an arc between two non-target nodes (Fig.1 (d)). Both cases (3) and (4) illustrate the influences in both direction, e.g. the influence A-to-B and B-to-A.

The strength of the influence (SI) between nodes is determined by the mean distance between different posterior probability distributions [36], i.e.

$$SI = \frac{1}{n} \sum_{i=0}^n D(P(A), P(B | A = a_i)) \quad (4)$$

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

2.2 Materials

We applied a BN modelling approach combined with expert knowledge to analyse the probabilistic relationships among data features with respect to ultrasound and FNAC examinations for breast cancer diagnosis.

2.2.1 Clinical ultrasound dataset

We collected a total of 1993 complete data samples from clinical ultrasonic examination in The First Affiliated Hospital of Fujian Medical University, China. The data are composed of 1494 benign (non-cancerous) and 499 malignant (cancerous) samples. Each sample contains five ultrasonic features of the breast tumour: shape (SH), resistances index (RI), calcification (CA), blood signal (BS) and the diameter-to-width ratio (DW).

The feature SH is either regular or irregular and was labelled as either 0 or 1, respectively. RI has three categories: none, less than 0.7 and greater than or equal to 0.7, labelled as 0, 1 and 2, respectively. Therefore 0.7 is an important cut-off point to distinguish between benign and malignant tumours in clinics. The feature CA contains no calcification, micro-calcification and macro-calcification. In our experiment, each CA category was labelled with 0, 1 and 2, respectively. Feature BS has two categories, i.e. no and yes, labelled with 0 and 1, which respectively indicates that a blood signal in the breast tumour is absent or present. In clinics, the BS in a malignant tumour is richer than that in a benign tumour. The two categories of the feature DW were labelled as 0 (the tumour's diameter is greater than its width) and 1 (the tumour's width is greater than or equal to its

diameter). However, our study excluded the feature DW because 1305 out of 1993 DW values were missing. The ultrasound data are described in Table 1.

2.2.2 FNAC examination dataset

The BCWD of the UCI machine learning repository contains a total of 683 complete FNAC data samples (444 benign and 239 malignant). The nine scored cytological features are bare nuclei (BANU), uniformity of cell size (UCSI), single epithelial cell size (SECS), uniformity of cell shape (UCSH), normal nucleoli (NN), marginal adhesion (MA), bland chromatin (BC), clump thickness (CT) and mitosis (MITO). All these features were computed from a digital image of a breast tumour FNAC examination, and they could be used to describe the characteristics of the cell nuclei shown in the image. Each feature was scored using an integer value ranging from 1 to 10, where 1 represented the most benign characteristic, and 10 represented the most malignant characteristic. According to [27], each sample was classified into either the benign or the malignant diagnostic category, labelled as 0 or 1, respectively. The FNAC data description are detailed in Table 1.

2.3 Data pre-processing

As BN structure learning is based on discretized data, we discretized the ultrasound data features according to their corresponding definitions, while each FNAC feature was discretized into two categories according to [37], [38] (see Table 1). We then applied an information gain algorithm [39] to rank each feature's relevance to the diagnostic result (i.e. benign or malignant). The obtained information gain score (IGs) of each feature is listed in descending order in the last column of Table 1. Finally, a K2 learning algorithm was applied to learn the BN structure of the given datasets with IGs-ordered features. All experiments were carried out on a PC using Weka [40] and GeNIe [41] software.

3. Results

We randomly split both datasets into 10 folds for cross validation. Each fold of the ultrasound dataset contained 1000 training data points (750 benign samples and 250 malignant samples) and 993 testing data points (744 benign samples and 249 malignant samples). For FNAC data, each fold contained 355 training data points (230 benign samples and 125 malignant samples) and 328 testing data points (214 benign samples and 114 malignant samples). After 10-fold cross validation, the obtained BNs were visualized and are shown in Fig.2, where the rectangular boxes stand for feature nodes. In each box, the feature's name is labelled in the upper cell, the categories of each discretized feature and the corresponding probability distributions are shown in the lower left cell and the lower right cell intuitively shows the probabilities using different coloured bars. The arrows indicate the probabilistic influence between two features, and the thickness of the arrow represents the strength of the corresponding influence. The thicker the arrow, the stronger the influence. Since diagnosis in the obtained BN models (Fig.2) was set as a target while other features were non-target, the influences between diagnosis and other features should be in line with Fig.1 (a) case, while the influences among non-targets is in line with Fig.1 (d).

The BN model in Fig.2 (a) depicts that all four ultrasound features have influences on diagnosis. The links between SH, RI and BS show their relatedness, while CA is independent of the other three. The BN model in Fig.2 (b) shows that all nine FNAC features have influences on diagnosis. Additionally, UCSI has a relationship with MA, UCSH, NN, CT, SECS and BANU, where NN is also related to MITO and BC. The strength of the influences of both ultrasound and FNAC features on diagnosis are shown in Fig.3. These could reflect each feature's contribution to diagnosis. For example, Fig.3 (a) shows that SH makes the highest contribution to diagnosis (SI: 0.3201), followed by RI, CA and BS. Hence, SH would be the most important diagnostic marker. Fig.3 (b) shows the SI of BANU to be 0.379, which is the highest among those of FNAC features. This indicates that BANU would be the most valuable diagnostic feature. On the other hand, SECS presented the

weakest influence on diagnosis (SI: 0.165). In addition, we also investigated the extent of dependence between different features, the results of which are shown in Fig.4. In order to clearly illustrate the dependency, we used *1 (* denotes the feature's name, such as SH1) to represent the impact producer. Fig.4 (a) shows that SH influences RI and BS with a strength of 0.1452 and 0.1465, respectively (see the diamond markers). The strengths of influence of RI and BS on SH are 0.2269 and 0.1415, respectively (the square and triangle markers in the left column), while RI and BS are strongly dependent each other (SI: 0.4398 and 0.4246, see the highest triangle and square markers). Fig.4 (b) clearly shows the strength of the dependencies among FNAC features. The highest triangle and square markers indicate that UCSI and UCSH are strongly dependent each other.

4. Discussion

In order to verify and validate the obtained BNs, statistical analysis based on pure data was also carried out. Fig.5 shows the true-positive rate (TPr, malignant tumour is correctly classified as malignant), the true-negative rate (TNr, benign tumour is correctly classified as benign) and the overall correct diagnostic accuracy, using diamond, square and triangle markers, respectively, in terms of individual ultrasound features and their combinations.

The result suggested that, considering individual features, SH presented the highest diagnostic accuracy and TNr and a higher TPr. Although BS had the highest TPr, its accuracy and TNr were very low. Therefore, the feature SH presented higher diagnostic performance than other features, while BS showed the lowest diagnostic accuracy. With respect to the combinations of any two features, SH combined with RI presented the highest TNr of 0.995, a high TPr of 0.886 and the highest accuracy of 0.967. In terms of combinations of any three features, the combination of SH, RI and CA contributed the highest TNr of 0.999, a high TPr of 0.9 and the highest malignant tumour diagnosis accuracy of 0.978. Furthermore, the combination of all features (i.e. SH + RI + BS + CA)

showed the highest performance in terms of TPr (0.934), TNr (0.999) and malignant tumour diagnosis accuracy (0.982). This is very consistent with the obtained BN shown in Fig.2 (a).

Many CAD methods have been proposed to support breast cancer diagnosis. Some studies [19][26] assumed that the features of breast cancer data should be independent, i.e. the Naïve Bayes (NB) modelling method would be appropriate. In addition, the decision tree methods, such as ID3 and J48, are also commonly used [21], while the NBTree is a classifier of the combination of the Naïve Bayes and decision tree methods. In this paper, we evaluated the performance of these representative and frequently-used methods; we then compared them with that of a BN modelling approach based on ultrasound and FNAC datasets using 10-fold cross-validation in order to obtain a fair comparison of classification accuracy for breast cancer diagnosis.

Table 2 lists the sensitivity (TPr) and specificity (TNr) of malignant tumour diagnosis, the overall correct classification accuracy for both benign and malignant classes and the AUC (area under the ROC curve) performance in terms of the five classifiers applied to the two given datasets. For the ultrasound dataset, J48 showed the highest performance in terms of sensitivity (0.9110), specificity (0.9100), accuracy (0.909) and AUC (0.941), followed by ID3 (sensitivity 0.8847, specificity 0.9083, accuracy 0.899 and AUC 0.9258) and lastly, the BN (sensitivity 0.8762, specificity 0.8978, accuracy 0.898 and AUC 0.9298). The BN approach performed competitively with ID3, with an even higher AUC than that of ID3, while NB and NBTree presented relatively worse performance. Although J48 performed the best, it could not successfully represent the probabilistic relationships among different data features (see the J48 tree constructed on ultrasound data in Fig. 6). Moreover, the diagnostic result in terms of the RI feature is unacceptable. According to the definition of RI, if the RI is less than 0.7, the tumour will normally be considered as a benign sign; otherwise, it will be considered as a malignant sign. However, the obtained J48 tree classified the tumour into the malignant category regardless of whether its RI was less than 0.7. We then reviewed our obtained BN model, and it could correctly diagnose a tumour with an RI less than 0.7 as benign. Apparently, this verified that

the BN model can provide more accurate and reasonable diagnosis than the J48 tree. For the FNAC dataset, the BN performed best in terms of sensitivity (0.9659) and specificity (0.9617) of malignant tumour diagnosis, correct classification accuracy of 0.965 and an AUC of 0.9887.

Hence, the experimental results verified that the BN model is more reliable than other models with which it was compared to discover probabilistic dependencies among data features for breast cancer diagnosis and for diagnostic classification accuracy improvement. More importantly, with the BN modelling approach, FNAC examination showed more accurate performance in terms of malignant tumour diagnosis and the correct classification for both malignant and benign tumours than ultrasound examination. However, FNAC is more complex than ultrasound examination.

To enhance the above analysis, we also calculated three classification errors, i.e. mean absolute error (MAE) [42], root mean square error (RMSE) [43] and relative absolute error (RAE) [43], of the five classifiers applied to the two datasets. Fig.7 (a) shows that BN constructed on the ultrasound dataset obtained the lowest MAE (0.143) and RAE (0.381). ID3 had the lowest RMSE (0.269), and BN had the second lowest RMSE (0.271). For the FNAC dataset, BN presented the lowest MAE (0.044), RMSE (0.179) and RAE (0.098) (see Fig.7 (b)). This further strengthens the competitiveness of the BN model in comparison with other representative models.

Furthermore, the kappa coefficients (κ) of the five classification models based on ultrasound and FNAC datasets were calculated. The κ is an essential metric to evaluate the agreement between different classifiers [44]. The assessment of κ was based on Landis and Koch [45]. They considered $0.21 \leq \kappa \leq 0.40$ to indicate fair agreement, $0.41 \leq \kappa \leq 0.60$ to indicate moderate agreement, $0.61 \leq \kappa \leq 0.80$ to indicate tentative agreement and $0.81 \leq \kappa \leq 1$ to indicate definite agreement. Fig.8 showed that all five models constructed from ultrasound data were in stronger agreement with a random classifier, which indicated that they all performed well in breast cancer diagnosis. Especially, ID3 showed the highest κ of 0.7434, followed by BN ($\kappa = 0.7378$). For FNAC, the five models worked perfectly, as the rankings (in descending order) of κ of BN, NB, NBTree, ID3 and J48 were 0.9222,

0.9114, 0.8941, 0.8881 and 0.8681, respectively. This is further evidence that FNAC would be a more effective examination than ultrasound for breast cancer diagnosis, which is also consistent with our above-mentioned experimental results shown in Table 2.

5. Conclusions

This study employed a BN modelling approach to support decision making on whether a breast tumour is diagnosed as benign or malignant. A benign breast tumour can be cured by treatment, but a malignant tumour, which is one of the most terrible diseases, cannot be completely cured. Early diagnosis and treatment will improve patient survival. A BN modelling approach has been used widely for breast cancer diagnosis, but a report quantitatively analysing the contributions of different breast cancer features to diagnosis is lacking. Moreover, studies about probabilistic relationships between different breast tumour features based on ultrasound examination and FNAC examination, respectively, are few. In this paper, we attempted to meet these two challenges using a BN modelling approach. The BN model we obtained can support clinical decisions in an automated manner by using a set of machine learning algorithms. We used a clinical ultrasound dataset from a local hospital and an FNAC dataset from openly sourced UCI machine learning repository. Numeric data features were discretized in line with their descriptions. The BNs were structured according to a K2 learning algorithm and validated by 10-fold cross validation.

We carried out extensive experiments to evaluate the performance of the employed BN modelling approach and four commonly used methods, i.e. NB, ID3, J48 and NBTree, in terms of sensitivity and specificity of malignant tumour diagnosis, the overall correct classification accuracy, and the AUC metrics. It turned out that the BN model is competitive and promising for breast cancer diagnosis. The BN model can explicitly present the probabilistic relationships between directly linked data features. SI mirrors the importance of different features to diagnosis and the intensity of dependence between features. The most important ultrasound marker with respect to breast cancer

diagnosis would be SH, followed by RI and CA. CA is independent of other features, but it cannot be ignored when making a diagnostic decision. RI is tightly associated with BS. We can select RI as a measure for diagnosis, but cannot use BS alone due to its weak influence. In addition to the ultrasound features we investigated, other features from ultrasound scanning were accessible, such as echo halo and posterior shadowing; however, they fall into the area of medical image processing, which is outside the scope of this study. In terms of the FNAC dataset, the best-performing marker for breast cancer diagnosis is BANU, followed by UCSI, UCSH, NN, MITO and MA. Clinicians can refer to the importance ranking of these features when making diagnostic decisions. Due to the strong dependencies between UCSI and UCSH, a breast tumour will be considered benign if both the size and shape of the cells are uniform. Malignancy should be associated with a large number of BANU.

This study can provide a valuable guide and assist observers in focusing on the most valuable features when collecting data; it can also indirectly aid breast cancer oncologists in making more accurate diagnoses when few features are known. Moreover, the BN-observed probabilistic relationships between different clinical features of breast tumours, identified from both ultrasound and FNAC examinations, can help oncologists to make more exact inferences, especially for specific patients with some missing feature values. In the meantime, the BN modelling approach can be extended to the diagnosis of other diseases in the healthcare area, which is our potential application.

Although the BN modelling approach performed well in our study, it requires that the data used be discrete [46]. Therefore, we need to discretize continuous data prior to data modelling. However, it is difficult to determine a stable cut-point value for discretization of specific features, which may cause inflated performance estimates. Therefore, a regression method that suits continuous data will be investigated in our future work.

Breast cancer contains many subtypes at the molecular level [47], and the treatments may differ between subtypes [49]. Our study so far focuses on two datasets collected from different countries,

and the data only reflect binary diagnostic categories, which might limit our research in its current stage. However, this study is attempting to fill a requirement using machine learning and data modelling approaches to discover holistic quantitative relationships among breast cancer data, which are few in the area. We are not searching for a new aetiology, and the experimental results from our study based on available data are consistent with the known aetiology of breast cancer. This can be used to validate our approach for use with potential new data types. Therefore, in next research stage, we will collect more data types and time-evolved data from our collaborative hospital to extend our study not only to breast cancer diagnosis but also to prognosis.

Conflict of interest statement

The authors declare no conflicts of interest.

Acknowledgements

This work was partly supported by the National Key Basic Research Program of China [grant number 2015CB352006] in data collection, the National Natural Science Foundation of China [grant number 61335011] in the decision to submit the manuscript for publication, the Scientific Research Funds for the Returned Overseas Chinese Scholars, State Education Ministry in the study design, the Young Key Program of Education Department, Fujian Province, China [grant number JZ160425] in the analysis and interpretation of data, the Program of Education Department of Fujian Province, China [grant number I201501005] and the Program for Changjiang Scholars, Innovative Research Team in University [grant number IRT_15R10] in the writing of the manuscript.

References

- [1] E.L. Davies, Breast cancer, *Medicine* (2015) 42-46.
- [2] L. Fan, K. Strasser-Weippl, J.J. Li, J. St. Louis, D. M. Finkelstein, K.D. Yu, W.Q. Chen, Z.M. Shao, P.E. Goss, Breast cancer in China, *Lancet Oncology* 15 (2014) 279-89. [http://dx.doi.org/10.1016/S1470-2045\(13\)70567-9](http://dx.doi.org/10.1016/S1470-2045(13)70567-9)
- [3] Chinese National Cancer Center, The latest Chinese cities' cancer data report of 2017, Chinese National Cancer Center (2017).

- [4] D.R. Chen, Y.L. Huang, S.H. Lin, Computer-aided diagnosis with textural features for breast lesions in sonograms, *Computerized Medical Imaging and Graphics* 35 (2011) 220-226.
- [5] R. Fusco, M. Sansone, Pattern recognition approaches for breast cancer DCE-MRI classification: a systematic review, *Journal of Medical & Biological Engineering* 36 (2016) 449-459.
- [6] J. Shan, S.K. Alam, B. Garra, Y. Zhang, T. Ahmed, Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods, *Ultrasound in Medicine and Biology* (2016) 980-988.
- [7] W.K. Moon, C.M. Lo, N. Cho, J.M. Chang, C.S. Huang, J.H. Chen, R.F. Chang, Computer-aided diagnosis of breast masses using quantified BI-RADS findings, *Computer Methods and Programs in Biomedicine* (2013) 84-92. <http://dx.doi.org/10.1016/j.cmpb.2013.03.017>
- [8] N. Zeng, Z. Wang, Y. Li, M. Hu, X. Liu, Inference of nonlinear state-space models for sandwich-type lateral flow immunoassay using extended Kalman filtering. *IEEE Transaction Biomedical Engineering* 58 (2011) 1959-1966.
- [9] N. Zeng, Z. Wang, Y. Li, M. Hu, X. Liu, A hybrid EKF and switching PSO algorithm for joint state and parameter estimation of lateral flow immunoassay models, *IEEE/ACM Transaction on Computational Biology and Bioinformatics* 9 (2012) 321-329.
- [10] N. Zeng, Z. Wang, H. Zhang, Inferring nonlinear lateral flow immunoassay state-space models via an unscented Kalman filter, *Science China Information Sciences* 59 (2016) 1-10.
- [11] W.K. Moon, C.M. Lo, J.M. Chang, C.S. Huang, J.H. Chen, R.F. Chang, Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses, *Journal of Digital Imaging* 26 (2013) 1091-1098. <http://dx.doi.org/10.1007/s10278-013-9593-8>
- [12] M.M. Eltoukhy, I. Faye, B.B. Samir, A statistical based feature extraction methods for breast cancer diagnosis in digital mammogram using multiresolution representation, *Computers in Biology and Medicine* 42 (2012) 123-128.
- [13] E.A. Mohammed, C.T. Naugler, B.H. Far, Breast tumor classification using a new OWA operator, *Expert Systems With Applications* 61 (2016) 302-313.
- [14] A.C. Comstantinou, N. Fenton, W. Marsh, L. Radkubski, From complex questionnaire and interviewing data to intelligent Bayesian Network models for medical decision support, *Artificial Intelligence in Medicine* 67 (2016) 75-93. <http://dx.doi.org/10.1016/j.artmed.2016.01.002/>
- [15] M.L.P. Nueno, A. Hommersom, P.J.F. Lucas, M. Lappenschaar, J.G.E. Janzing, Understanding disease processes by partitioned dynamic Bayesian networks, *Journal of Biomedical Informatics* 61 (2016) 289-297.
- [16] F.L. Seixas, B. Zadrozny, J. Laks, A. Connci, A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment, *Computers in Biology and Medicine* 51 (2014) 140-158.
- [17] Y. Liu, Z. Cheng, Y. Zhao, X. Chang, C. Chan, Y. Bai, N. Cheng, Efficacy and safety of long-term treatment with statins for coronary heart disease: A Bayesian network meta-analysis, *Atherosclerosis* 254 (2016) 215-227.
- [18] Z.S. Estabragh, M.R. Kashani, F.J. Moghaddam, S. Sari, Z. Taherifar, S. M. Moosavy, and K. S. Oskooyee, Bayesian network modeling for diagnosis of social anxiety using some cognitive-behavioral factors, *Network Modeling Analysis in Health Informatics and Bioinformatics* 2 (2013) 257-265.
- [19] X.H. Wang, B. Zheng, W.F. Good, J.L. King, Y.H. Chang, Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network, *International Journal of Medical Information* 54 (1999) 115-126.
- [20] C.R. Nicandro, H.G. Acosta-Mesa, H. Carrillo-Calvet, L.A. Nava-Fernández, R.E. Barrientos-Martínez, Diagnosis of breast cancer using Bayesian networks: A case study, *Computers in Biology and Medicine* 37 (2007) 1553-1564.
- [21] C.R. Nicandro, H.G. Acosta-Mesa, H. Carrillo-Calvet, L. A. Nava-Fernández, R. E. Barrientos-Martínez, Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks, *Applied Soft Computing* 9 (2009) 1331-1342.
- [22] A.M. Kalet, J.H. Gennari, E.C. Ford, M.H. Phillips, Bayesian network models for error detection in radiotherapy plans, *Physics in Medicine and Biology* 60 (2015) 2735-2749.
- [23] H.B. Hassen, I. Kallen, L. Bouchaala, A. Rebai, Analysis of breast cancer profiles using Bayesian Network Modeling, *International Journal of Biomathematics* 6 (2013) 1350014. <http://dx.doi.org/10.1142/S1793524513500149>
- [24] Z.X. Zeng, X. Jiang, R. Neapolitan, Discovering causal interactions using Bayesian network scoring and information gain, *BMC Bioinformatics* (2016).
- [25] C.R. Nicandro, M.M. Efrén, A.A.M. Yaneli, M.D.C.M. Enrique, A.M.H. Gabriel, P.C. Nancy, G.H. Alejandro, H.R. Guillermo de Jesús and B. M. R. Erandi, Evaluation of the diagnostic power of thermography in breast

- cancer using Bayesian Network classifiers, *Computational and Mathematical Methods in Medicine* (2013). <http://dx.doi.org/10.1155/2013/264246>
- [26] M. Karabatak, A new classifier for breast cancer detection based on Naïve Bayesian, *Measurement* 72 (2015) 32–36. <http://dx.doi.org/10.1016/j.measurement.2015.04.028>
- [27] [dataset] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, Breast Cancer Wisconsin Dataset, UCI Repository of Machine Learning Databases. (1998). <http://archive.ics.uci.edu/ml/>
- [28] D. Heckerman, D. Geiger, and D.M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, *Machine Learning* 20 (1995) 197-243.
- [29] S.L. Yang, and K.C. Chang, Comparison of score metrics for Bayesian network learning, *IEEE Transactions on Systems, man and cybernetics-part A: Systems and Humans* (2002).
- [30] R. Daly, Q. Shen, S. Aitken, Learning Bayesian networks: approaches and issues, *The Knowledge Engineering Review* 26 (2011) 99-157. <http://dx.doi.org/10.1017/S0269888910000251/>
- [31] G.E. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* 9 (1992) 309-347.
- [32] F.J. Diez, J. Mira, DIAVAL, a Bayesian expert system for echocardiography, *Artificial Intelligence in Medicine* 10 (1997) 59-73.
- [33] L. Bouchaala, A. Masmoudi, F. Gargouri, A. Rebai, Improving algorithms for structure learning in Bayesian networks using a new implicit score, *Expert Systems with Application* 37 (2010) 5470-5475.
- [34] S.L. Lauritzen, The EM algorithm for graphical association models with missing data, *Computational Statistics & Data Analysis* 19 (1995) 191-201.
- [35] G.E. Yap, A.H. Tan, H.H. Pang, Explaining inferences in Bayesian networks, *Applied Intelligence* 29 (2008) 263-278. <http://dx.doi.org/10.1007/s10489-007-0093-8>
- [36] J.R. Koiter, Visualizing Inference in Bayesian Networks, Master Thesis, Faculty of Electrical Engineering, Mathematics, and Computer Science, Department of Man-Machine Interaction (2006)
- [37] R. Ohashi, M. Matsubara, Y. Watarai, K. Yanagihara, K. Yamashita, S. Tsuchiya, H. Takei, Z. Naito, Diagnostic value of fine needle aspiration and core needle biopsy in special types of breast cancer, *Breast Cancer* 23 (2016) 675-683. <http://dx.doi.org/10.1007/s12282-015-0624-9>
- [38] P.W. Hamilton, N. Anderson, P.H. Bartels, D. Thompson, Expert system support using Bayesian belief networks in the diagnosis of fine needle aspiration biopsy specimens of the breast, *Journal of Clinical Pathology* 47 (1994) 329-336.
- [39] C.E. Shannon, A mathematical theory of communication. *Bell Labs Technical Journal* 27 (1948) 379-423. <http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [40] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second ed., Morgan Kaufmann, San Francisco (2005).
- [41] P.C. Kraaijeveld, M.J. Druzdzal, GeNIeRate: An Interactive Generator of Diagnostic Bayesian Network Models, *International Workshop on Principles of Diagnosis* (2008) 175-180.
- [42] E.J. Coyle, J.H. Lin, Stack filters and the mean absolute error criterion, *IEEE Transactions on acoustics speech and signal processing* 36 (1988) 1244-1254.
- [43] T. Chai, R.R. Draxler, Root mean square error (RMSE) or means absolute error (MAE)? Arguments against avoiding RMSE in the literature, *Geoscientific Model Development* 7 (2014) 1245-1250.
- [44] B.D. Eugenio, M. Glass, The Kappa Statistic: A Second Look, *Computational Linguistic* 30 (2004) 95-101.
- [45] J.R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* 33 (1977) 159-174. <http://dx.doi.org/10.2307/2529310>
- [46] R.R. Bounckaert, Bayesian network classifier s in Weka, *Artificial Intelligence Tools* 29 (2004) 131-163.
- [47] E.A. Rakha, J.S. Reis-Filho, F. Baehner, D.J. Dabbs, T. Deckers, V. Eusebi, S.B. Fox, S. Ichihara, J. Jacquemier, S.R. Lakhani, J. Palacios, A.L. Richardson, S.J. Schnitt, F.C. Schmitt, P.Tan, G.M. Tse, S. Badve, I.O. Ellis, Breast cancer prognostic classification in the molecular era: the role of histological grade, *Breast Cancer Research* 12 (2010): 207. <http://breast-cancer-research.com/content/12/4/207>
- [48] M.T. Weigel, M. Dowsett, Current and emerging biomarkers in breast cancer: prognosis and prediction, *Endocrine-Related Cancer* 17 (2010): 245-262.
- [49] G.W. Sledge, E.P. Mamounas, G.N. Hortobagyi, H.J. Burstein, P.J. Goodwin, A.C. Wolff, Past, present and future challenges in breast cancer treatment, *Journal of Clinical Oncology* 32 (2014)

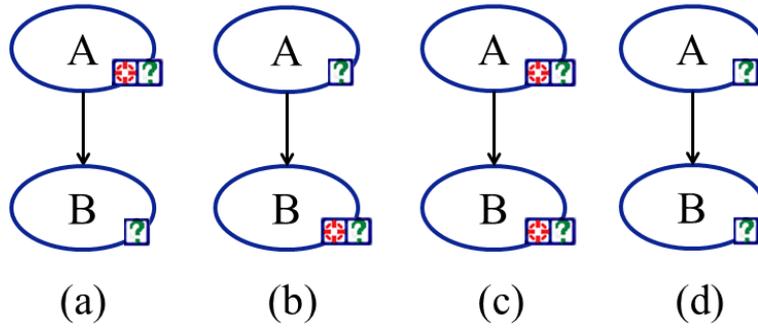


Fig.1. Four cases of a dynamic BN model. Each elliptic node stands for a variable. The red crossed sign denotes a target node, and the green question mark means that the BN has not been updated. (a) Influence from a target node A to a non-target node B. (b) Influence from a non-target node A to a target node B. (c) Influence between two target nodes A and B. (d) Influence between two non-target nodes A and B.

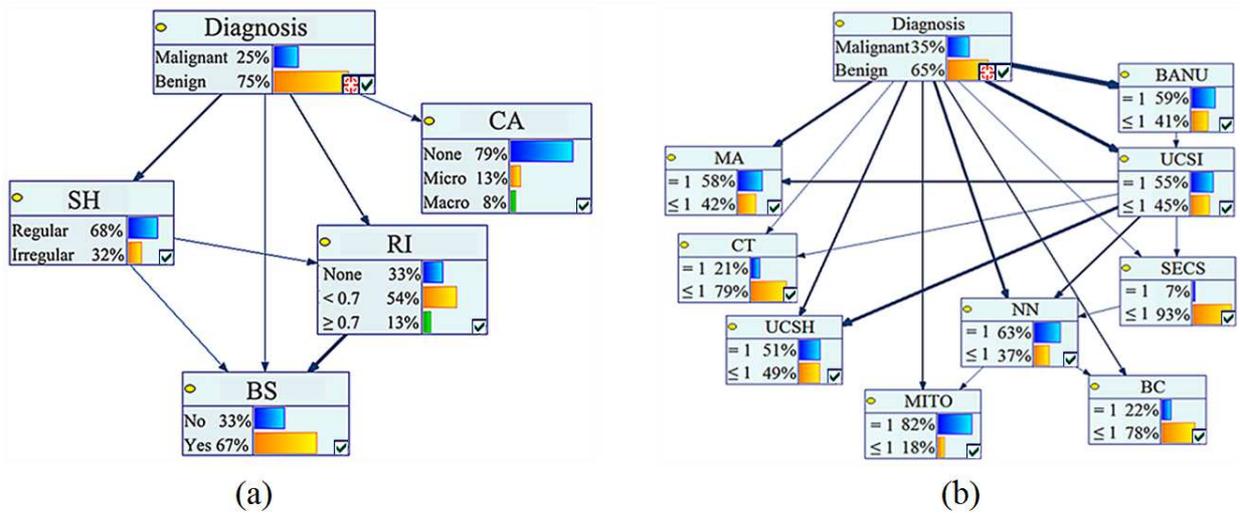


Fig.2. Bayesian network models constructed based on ultrasound (a) and FNAC (b) datasets. Rectangular boxes stand for feature nodes. In each box, the feature's name is labelled in the upper cell, the categories of each discretized feature and the corresponding probability distributions are shown in the lower left cell and the lower right cell intuitively shows the probabilities using different coloured bars. The red crossed sign marks the target node, and the green check mark means the BN has been updated. The arrows indicate the probabilistic dependencies between two features, and the thickness of the arrow represents the strength of the corresponding dependency.

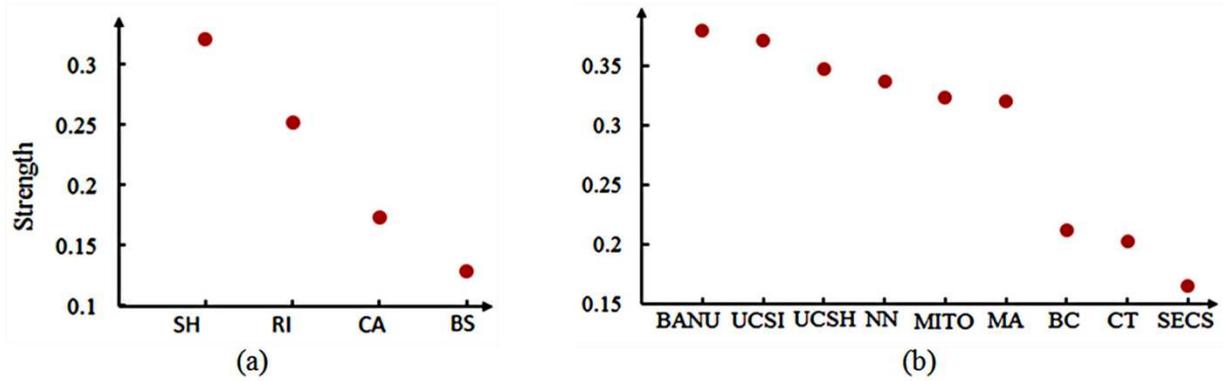


Fig.3. Strength of the influences between diagnosis and each single feature in terms of ultrasound (a) and FNAC (b) datasets.

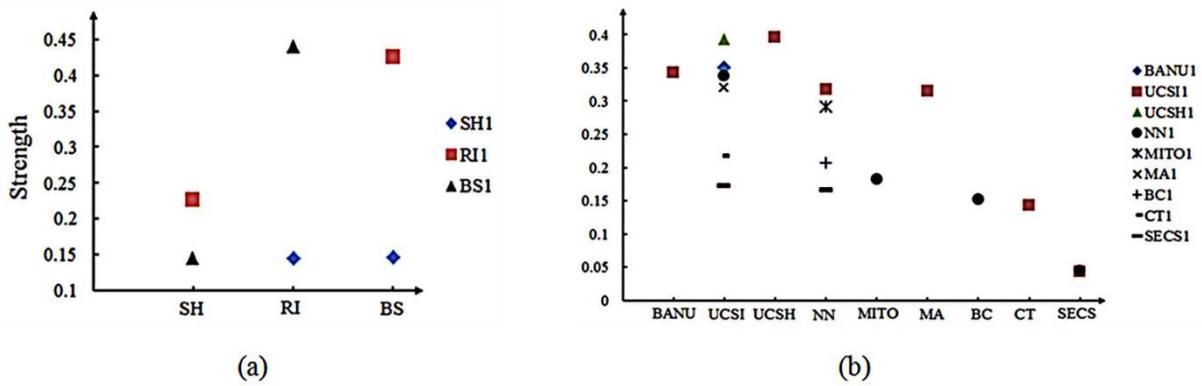


Fig.4. Strength of the dependencies among different features in terms of ultrasound (a) and FNAC (b) datasets. *1 is used to represent the impact producer, e.g. SH1 in (a) where the diamond markers stand for the influences of shape (SH) on the resistance index (RI) and blood signal (BS). The square markers stand for the influences of RI on SH and BS. The triangle markers represent the influence of BS on SH and RI. Similar presentation in (b).

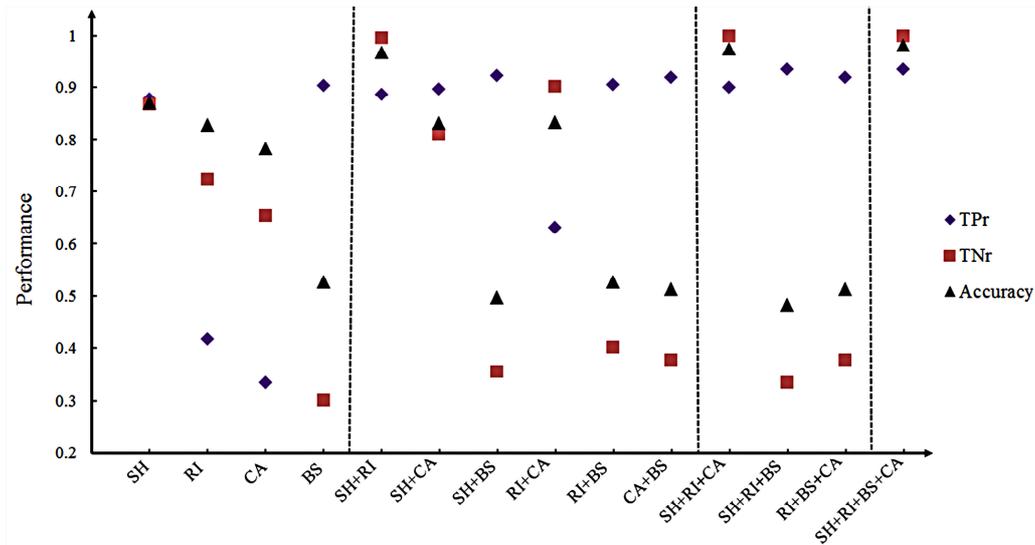


Fig.5. Performance of each individual feature and their combination in diagnosis based on the ultrasound dataset. The diamond, square and triangular markers denote the true-positive rate (TPr), true-negative rate (TNr) and the overall correct diagnostic accuracy.

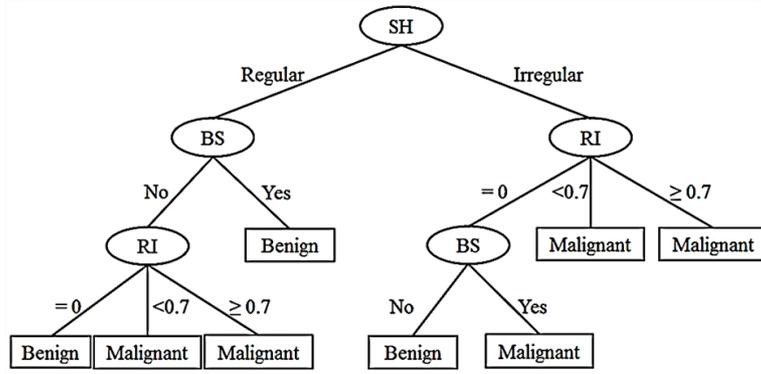


Fig.6. J48 tree structure constructed on ultrasound data. The elliptical nodes stand for ultrasonic features, the rectangle boxes represent the diagnostic results and the labels next to the links between upper and lower features are the states of each upper feature.

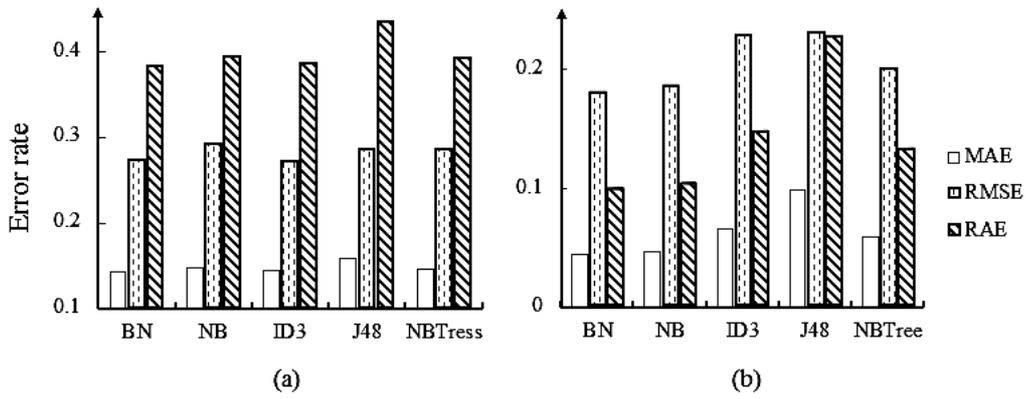


Fig.7. 10-fold cross validation based classification error rates of the five classifiers applied to both ultrasound (a) and FNAC (b) datasets. The blank bar is the mean absolute error (MAE), the bar filled with points is the root mean square error (RMSE) and the bar filled with oblique lines stands for the relative absolute error (RAE).

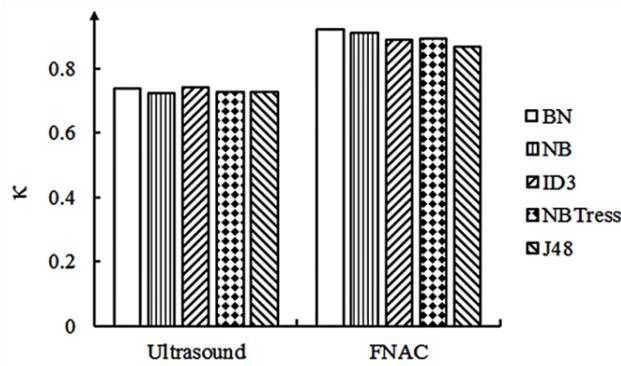


Fig.8. Kappa coefficients (κ) about BN, NB, ID3, NBTress and J48 classifiers applied to both ultrasound and FNAC datasets. The blank bar represents the Bayesian network (BN) model, the bar with vertical lines represents the Naïve Bayes (NB) model, the bar with left slash lines represents ID3 algorithm model, the bar with diamond marks represents the NBTress algorithm and the bar with right slash lines represents the J48 algorithm. The height of the bars suggests the κ value.

Table 1. Data description for both ultrasound and fine needle aspiration cytology (FNAC) datasets.

IGs: information gain scores.

Datasets	Features	Values	Labels	IGs
Ultrasound	Shape (SH)	Regular	0	0.343
		Irregular	1	
	Resistance index (RI)	None	0	0.182
		RI <0.7	1	
		RI ≥0.7	2	
	Calcification (CA)	None	0	0.078
Micro		1		
Macro		2		
Blood signal (BS)	No	0	0.065	
	Yes	1		
Diameter/width ratio (DW)	DW <1	0	---	
	DW ≥1	1		
FNAC	Bare Nuclei (BANU)	1	0	0.571
		2–10	1	
	Uniformity of Cell Size (UCSI)	1	0	0.525
		2–10	1	
	Single Epithelial Cell Size (SECS)	1	0	0.494
		2–10	1	
	Uniformity of Cell Shape (UCSH)	1	0	0.475
		2–10	1	
	Normal Nucleoli (NN)	1	0	0.416
		2–10	1	
	Marginal Adhesion (MA)	1	0	0.347
		2–10	1	
Bland Chromatin (BC)	1	0	0.323	
	2–10	1		
Clump Thickness (CT)	1	0	0.291	
	2–10	1		
Mitosis (MITO)	1	0	0.199	
	2–10	1		

Table 2. Classification performance of Naïve Bayes (NB), Bayesian network (BN), ID3, J48 and NBTree classifiers applied to both ultrasound and FNAC datasets.

Dataset	Model	Sensitivity	Specificity	Accuracy	AUC
Ultrasound	NB	0.8620	0.8927	0.891	0.8958
	BN	0.8762	0.8978	0.898	0.9298
	ID3	0.8847	0.9083	0.899	0.9258
	J48	0.9110	0.9100	0.909	0.9410
	NBTree	0.8768	0.8941	0.893	0.9076
FNAC	NB	0.9603	0.9497	0.960	0.9880
	BN	0.9649	0.9617	0.965	0.9887
	ID3	0.9450	0.9277	0.945	0.9484
	J48	0.9414	0.9393	0.941	0.9447
	NBTree	0.9519	0.8533	0.952	0.9822

Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach

Shuo Liu^{1,*}, Jinshu Zeng^{2,*}, Huizhou Gong³, Hongqin Yang^{4,†}, Jia Zhai⁵, Yi Cao⁶, Junxiu Liu⁷, Yuling Luo⁸, Yuhua Li⁹, Liam Maguire⁷, Xuemei Ding^{1,7,†}

¹ Faculty of Mathematics and Informatics, Fujian Normal University, Qishan Fuzhou, 350108, China

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

³ College of Foreign Languages, Fujian Normal University, Cangshan Fuzhou, 350007, China

⁴ Fujian Provincial Key Laboratory for Photonics Technology, Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Cangshan Fuzhou, 350007, China

⁵ Business School, University of Salford, Manchester, M5 4WT, UK

⁶ Department of Business Transformation and Sustainable Enterprise, Surrey Business School, University of Surrey, Surrey, GU2 7XH, UK

⁷ Faculty of Computing, Engineering, and Built Environment, Ulster University, Londonderry, BT48 7JL, UK

⁸ Faculty of Electronic and Engineering, Guangxi Normal University, Guilin, 541004, China

⁹ School of Computing, Science and Engineering, University of Salford, Manchester, M5 4WT, UK

Highlights

- Using Bayesian network modelling approach, probabilistic dependencies among different data types were discovered for decision support of breast cancer diagnosis.
- Diagnostic contributions of ultrasound and fine-needle aspiration cytology examinations were quantitatively analyzed.
- Bayesian network modelling approach is competitive and promising for breast cancer early diagnosis due to its high classification accuracy, high sensitivity, and high specificity, in comparison with other representative used methods.