Radboud Universiteit Nijmegen

Computing Science Department

Learning Bayesian models using mammographic features.

Master Thesis

Author:
Niels Radstake

Supervisor:
dr. Peter J.F. Lucas

Second supervisor:
dr. Elena Marchiori

Thesis number:
625

Date:
January 13, 2010

Abstract

The most frequent type of cancer among women worldwide is breast cancer. When breast cancer is detected in an early stage, chances of successful treatment are high. Screening programmes have shown to reduce the mortality rate of breast cancer. Studies have shown that radiologists fail to identify a significant number of cases with breast cancer due to misinterpretation.

To address the problem of interpretation failure by radiologists, the B-SCREEN project investigates the use of Bayesian networks and Bayesian classifiers. This study focuses on the learning of Bayesian networks from data that is available from the Dutch breast cancer screening programme using different structure and parameter learning techniques. The possibility to use these techniques is verified using experiments.

This study concludes that it is possible to use structure and parameter learning techniques to learn Bayesian classifiers that perform reasonable using data from breast cancer screening programmes. The network structures give insights in the correlation of certain variables in the breast cancer domain. However, the performance of these classifiers is still less than when other classification methods are used.

Acknowledgements

I would like to thank a number of people who have supported me during the writing of this thesis.

First, I would like to thank dr. Peter J.F. Lucas for initiating this research project and giving advice on the direction of my research. His knowledge on Bayesian networks helped me a lot. I would also like to thank him for the useful feedback an co-operation during the writing of this thesis.

I would like to thank dr. Nivea Carvalho Ferreira for introducing me to the field of breast cancer, the explanation of features and models, providing MATLAB code from previous experiments, giving very useful feedback on draft versions of this thesis and many more things. Without her help, this thesis would not have been possible.

I would also want to thank dr. Elena Marchiori for reviewing my thesis, and the people of the Department of Radiology at the UMCN, especially dr. Marina V. Velikova, for introducing me to the field of breast cancer.

Last, but certainly not least, I could not have finished my thesis without the support of my girlfriend Marloes Voltman and friends and family.

Contents

1 Introduction

Worldwide, the most frequent type of cancer (in order of the number of global deaths) among women is breast cancer [36]. In the Western world, 10 percent of all women are confronted with breast cancer in their lives [36]. The success of treatment of breast cancer largely depends on the stage of the tumor at the time of detection. When breast cancer is detected in an early stage, chances of successful treatment are high.

To detect breast cancer in an early stage, many nations have set up screening programmes. In these programmes, female breasts are examined using X-rays, resulting in screening mammograms. These mammograms are read by radiologists for abnormalities. Studies have shown that radiologists fail to identify a significant number of cases with breast cancer due to interpretation failure. Audits have shown that in the Dutch screening programme more than 25% of the detected cancers already show relatively clear signs of abnormality on mammograms made during a previous screening, while another 25% show minimal signs [3]. If these abnormalities had been detected in previous screenings, a more effective treatment would have been possible.

Starting in 2006, all mammograms of the Dutch breast cancer screening programme are being digitized and stored in one central national archive. This large database provides an opportunity for the development of decision-support systems, which can assist radiologists in reading mammograms.

At the end of 2006, the *B-SCREEN: Bayesian Decision Support in Medical Screening project* started as a collaboration of the Institute for Computer and Information Science of the Radboud University Nijmegen and the Department of Radiology, UMC St. Radboud Nijmegen[1]. The aim of this project is to "use Bayesian networks and Bayesian classifiers to further address the problem of interpretation failures by radiologists" [3]. To develop these new improved Bayesian classifiers, advanced image analysis and domain knowledge from the breast cancer screening domain are being used. The breast cancer models that are being developed can be used in decision support systems for radiologists and help to improve the detection of breast cancer in an early stage.

In previous research in the B-SCREEN project [12], a Bayesian network model was constructed using medical expert knowledge. This Bayesian model was built from a conceptual model, which is based on a causal picture of the domain. The feature-based model, shown in Figure 10, is based on elements of the domain that can be observed and measured. A problem with the network model that was constructed

[1] The UMC St. Radboud Nijmegen plays an active role in the Dutch breast cancer screening programme. It houses the LRCB (Dutch: Landelijk Referentiecentrum voor Bevolkingsonderzoek) which is responsible for quality assurance and training of screening radiologists.

using medical expert knowledge is that the classification performance of the network was lower than expected. Construction of models using expert knowledge is also very time consuming and with the availability of large datasets it becomes possible to learn Bayesian network structures using data.

The study reported in this document focuses on the learning of Bayesian models in the breast cancer domain using structure and parameter learning techniques. Learning the optimal network structure is not possible, because the number of possible network structures is super exponential in the number of variables used. Using heuristic methods, it is possible to search a part of the space of network structures and find structures that fit well to the data.

Results from research in the B-SCREEN project [33, 34, 35, 28, 29] have been used as a starting point for this study.

The purpose of this study is to investigate:

1. to what extent structure and parameter learning techniques can be used in breast cancer research;
2. whether the correlation of certain variables in the dataset can be observed in the learned models;
3. the possibility of improving classification performance by combining data from different mammographic views.

This can help the development of better performing Bayesian classifiers and possibly improve detection of breast cancer in an early stage. It is not the aim of this project to find a perfect performing classifier or to include all available mammographic features in the learning process.

In section 2, an overview of the breast cancer domain is given. In section 3 the theory of Bayesian network is explained briefly. To learn the Bayesian classifiers, a number of existing learning algorithms are being used. Which algorithms are used for learning Bayesian networks from data is explained in section 4. The experiments and results of this research are in section 5. This thesis concludes with the conclusions, discussion and suggestions for future research in section 6.

2 Breast cancer domain

2.1 Breast cancer

Breast cancer is a form of cancer that starts in the cells of the breasts of women and men. It is the result of uncontrolled division and growth of breast cells. Most breast cancers have their origins in the cells of the ducts and some in cells of the lobules, which are milk producing glands. See Figure 1 for the anatomy of a female breast.

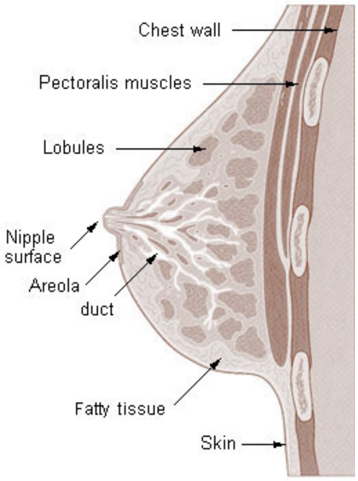


Figure 1: Breast anatomy. Image taken from http://www.wikimedia.org

A malignant cancer, or *carcinoma*, has various stages starting at *carcinoma in-situ*, or CIS, which is a pre-malignant condition in which the tumor is in-situ (Latin for 'in its place') and there is no invasion of surrounding tissue. A CIS is considered a precursor form of cancer that can develop into a malignant cancer, which can spread to surrounding tissue or metastasize to other parts of the body through the lymphatic system. It is generally assumed that all invasive cancers develop from a CIS condition, but not every CIS develops eventually to an invasive carcinoma.

Most breast carcinomas (75%) are invasive ductal carcinomas, or IDCs. About 10-15% are invasive lobular carcinomas, or ILCs. Other breast carcinomas include invasive medullary carcinomas (5-7%), invasive tubular carcinomas (2-6%), invasive mucinuous carcinomas (3%) and invasive papillary carcinomas (2%) [13].

Breast cancer is an unilateral disease, which means that it develops, usually, in one breast. However, in 1-2% of the cases breast cancer is bilateral at the moment of detection. The incidence of breast cancer in men is about a hundred times less common than in women [36], because women have more breast tissue and their breasts cells are exposed to the female hormones estrogen and progesterone. In this research, the focus in only on breast cancer in women.

2.2 Risk factors

There are several risk factors that increase the probability of developing breast cancer [13].

Age Breast cancer occurs more frequent in older women. Only 12-15% of all invasive breast cancers are found in women younger than 45 [36].

Genes Presence of BRCA1/2 genes leads to an increased risk of developing breast cancer irrespective of other risk factors.

Age at menarche A lower age at the menarche increases the risk of developing breast cancer.

Pregnancy Nulliparous women have a 25% higher chance of developing breast cancer compared to women who have had a full-term pregnancy.

Hereditary factors Women with a mother or sister who has had breast cancer have about twice as much chance of developing breast cancer.

Hormonal stimulation The use of the combined oral contraceptive pill (COCP) or hormone replacement therapy (HRT) increases the risk of developing breast cancer.

Life style factors Factors related to diets, alcohol consumption, physical activity and other life style factors do have influence on the risk of developing breast cancer.

Ionizing radiation Exposure to ionizing radiation increases the risk of developing breast cancer.

History of breast cancer Women with previous breast cancer diagnosis have an increased risk of developing breast cancer.

2.3 Breast cancer screening

The success of treatment of breast cancer largely depends on the stage of the tumor at the time of detection. When a tumor is smaller than 20 mm and it has not metastasized to other parts of the body, chances of a successful curative treatment are high. Screening programmes have shown to contribute to the detection of breast cancer in an early stage [31, 24].

Many nations have set up a screening programme for breast cancer. The aim of these programmes is to detect breast cancer in an early stage and reduce the mortality rate of breast cancer. In previous research by Nyström et al., screening programmes have shown to reduce the mortality rate of breast cancer by 25% to 30% [24]. Using recent data it is estimated that the breast cancer mortality in The Netherlands has decreased by 800 cases per year [6]. If the screening methods can be improved, the breast cancer mortality can be reduced even further.

This research focuses on the Dutch breast cancer screening programme. In The Netherlands, asymptomatic women — women showing no indication of the presence of breast cancer — between the ages of 50 and 75 can voluntarily participate in the breast cancer screening programme. There is a high participation level: 82% of all invited women participated in 2006 [22].

The women that participate are invited biennially (every two years) to have an examination of both breasts. For the initial screening, two mammographic views — which are described in the next section — of each breast are taken: the *mediolateral oblique* (MLO) view and the *craniocaudal* or (CC) view. For subsequent screenings only a MLO view is taken, unless there is an indication that a CC view might be beneficial. Acquisition of a CC view occurs in about 30% of subsequent screening rounds [9].

2.4 Mammography

Mammography is the diagnostic procedure to detect breast cancer in the female breast using low-dose X-rays. In the process, mammographic images, called mammograms, are made from different angles (*views* or *projections*). The most common projections of the breast are the *mediolateral oblique* (MLO) view and the *craniocaudal* (CC) view, which are shown in Figure 2. The MLO view is a 45 degree angled side view, usually showing a part of the pectoral muscles (see Figure 1). The craniocaudal (literally: from head to tail) view is an projection of the breast from above with the nipple centered in the image.

Because a mammogram is a projection of the breast, the parts of which the breast is consisted (see Figure 1) are superimposed. The X-ray attenuation (due to absorption and scattering of photons) describes the density of a region. In the resulting mammogram, this can be seen as the contrast, or whiteness, of a region. The darker areas of the breast are non-dense tissue and consist mainly of fatty tissue. The lighter areas are denser tissue which contain lobules, ducts and possibly masses.

Diagnosis of breast cancer through mammography is more accurate in non-dense breasts, because dense breasts make mammograms difficult to read. Dense breast tissue can hide masses that are potentially malignant. This is especially true for young women, since breasts gradually becomes less dense over time. Because the

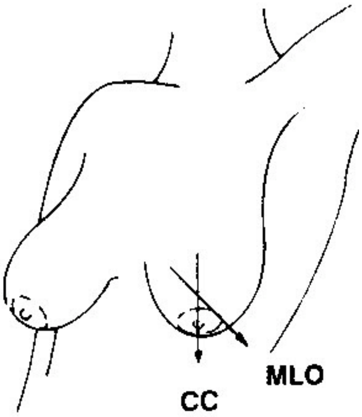


Figure 2: CC and MLO projections. Image ©http://www.imaginis.com.

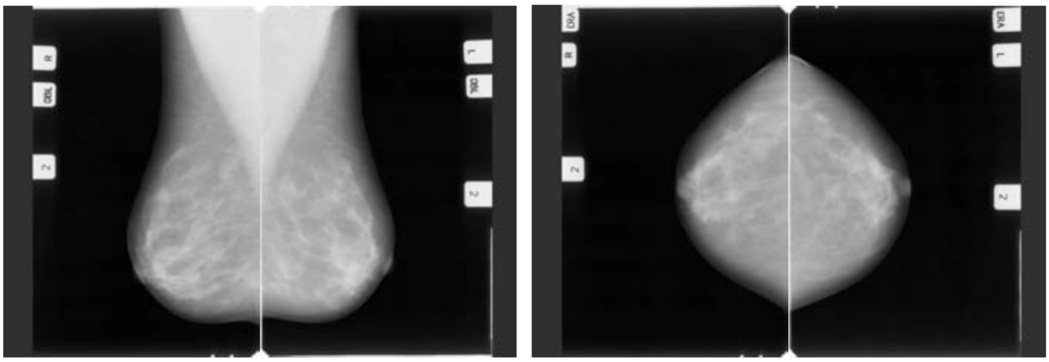


Figure 3: Mammograms. From left to right: MLO (right), MLO (left), CC (right), CC (left). Image taken from [9]

incidence of breast cancer in younger woman is very low and the mammograms of younger woman are in general difficult to read, only woman over the age of 50 are invited to participate the Dutch screening programme.

In Figure 4, four categories of breast density are shown. A mass would be easily detectable in the leftmost non-dense breast, whereas it might not be detected in the rightmost dense breast.

Other imaging techniques like *ultrasound* and *magnetic resonance imaging* (MRI) are available, but these techniques are not useful for screening programmes because of their high costs and respectively low *specificity* (see section 3.8) and low sensitivity. They can however by used as complementary to mammography. For instance, ultrasound is used to investigate detected lesions which can not be classified only

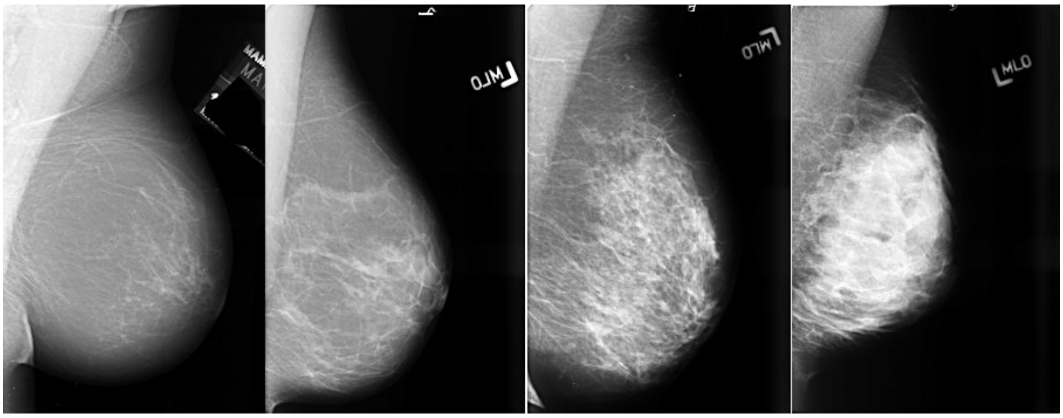


Figure 4: Four different categories of breast density as defined by the American College of Radiology [2]. From left to right: 'almost entirely fat', 'scattered fibroglandular densities', 'heterogeneously dense' and 'extremely dense'. Image ©http://sprojects.mmi.mcgill.ca/Mammography

using mammography images.

2.5 Interpretation of mammograms

The interpretation of mammograms by radiologists includes the identification of a number of *regions of interest*, or *regions* for short. A region can also be referred to as *lesion* or *abnormality.*

Each region is characterized by a number of properties called *features.* Examples of these features are, for instance, the area of the region, its location, its shape and its density. The features of a region of interest together might suggest a certain level of suspiciousness for the presence of breast cancer.

A radiologist reads the mammograms in search of abnormalities. When both MLO and CC views are present, regions that appear in both views are compared. When previous mammograms are present, the current mammograms are compared to previous ones to see how the breast has changed over time. See Figure 5 for an example.

In the Dutch screening programme the mammograms are independently interpreted by two radiologists. If both recommend to have a further investigation, the patient is recalled. If one of the radiologists recommends to recall the patient and the other does not, the decision for recall is reached by discussion (consensus double reading).

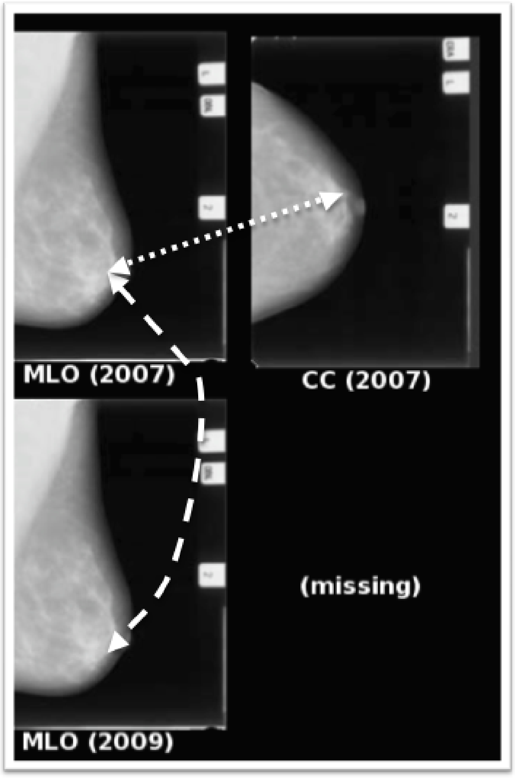


Figure 5: Comparing regions in different views (dotted line) and over time (dashed line)

2.6 Malign and benign tumors

The majority of breast tumors that are detected through mammography are benign. These tumors, such as cysts, are non-cancerous and cannot metastasize to other parts of the body. Using mammography it is in some cases difficult to distinguish between malign and benign tumors. However, there are some features of regions that can indicate whether a tumor is benign or malign. This section describes the most important features that radiologists use to distinguish between benign and malign tumors.

2.6.1 Margins

In general, benign tumors have clear, well-defined, circumscribed borders whereas malign tumors have less distinguishable borders or are surrounded by a radiating pattern of spicules. However, when the cancer is in its initial stage it can appear to have a well-defined (circumscribed) margin. A margin can be obscured by other superimposed or adjacent tissue. In mammography, five different classes of margins are distinguished (see Figure 6).

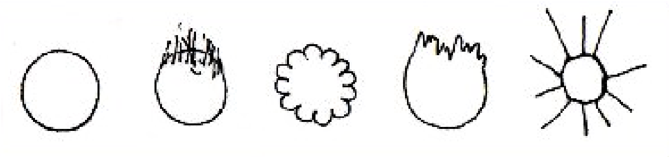


Figure 6: Mass margins as defined by BI-RADS [2]. From left to right: 'circumscribed', 'obscured', 'micro-lobulated', 'ill-defined' and 'spiculated'. Image ©http://sprojects.mmi.mcgill.ca/Mammography

2.6.2 Shape

The shape of a tumor is an important feature to indicate the presence of breast cancer. The five shapes as defined by BI-RADS [2] are shown in Figure 7. When there is no mass visible, but the normal architecture is distorted, a tumor is classified as 'architectural distortion'. This includes a spiculated margin.

2.6.3 Size

The size of a tumor combined with other features, such as its shape and location is an indication whether a tumor is malignant or not. Malignant masses are on average larger than benign masses [32].

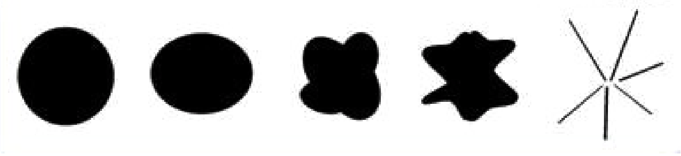


Figure 7: Mass shapes as defined by BI-RADS [2]. From left to right: 'round', 'oval', 'lobular', 'irregular' and 'architectural distortion'. Image ©http://sprojects.mmi.mcgill.ca/Mammography

2.6.4 Location

The location of a mass is an indication for the presence of breast cancer because most malignancies (45%) develop in the upper outer quadrant of the breast toward the armpit [5].

2.6.5 Density

Malignancies have in general a greater density than benign masses. In a mammogram, this shows as the contrast (whiteness) of a region. Lesions with the same density as the surrounding tissue can remain invisible on mammograms.

2.6.6 Calcifications

Calcifications are small bits of calcium that can appear in breast tissue. They can give further information about the presence of breast cancer and appear as white dots on mammograms. There are two kind of breast calcifications:

Macrocalcifications are calcifications that are larger than 2 mm and are usually not an indication for breast cancer.

Microcalcifications are calcifications smaller than 1 mm that are associated with breast caner. They can appear in different patterns. The number of microcalcifications, their grouping of the calcifications and its pattern give an indication for breast cancer.

2.7 Computer-aided detection

Studies have shown that radiologists fail to identify a significant number of cases with breast cancer (false negatives) due to misinterpretation. The causes for these misses are not clear [9]. Audits have shown that abnormalities that are clearly visible in retrospect must have been overlooked or its signs were misinterpreted [3].

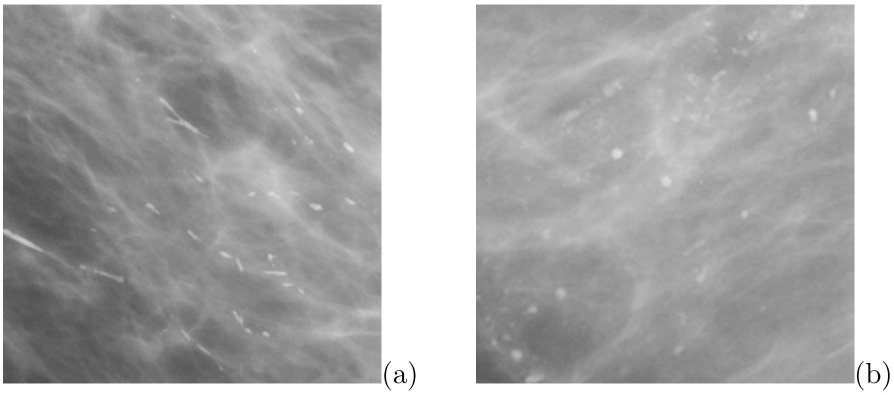


Figure 8: Examples of calcifications. Normal calcifications in (a) and calcifications with a high probability of malignancy in (b).

To increase detection rate, *computer-aided detection* (CAD) systems are being developed. These systems use pattern recognition techniques to identify features in a mammogram that can give an indication for the presence of breast cancer. Using these features, a CAD system is able to identify regions in a mammogram that are possibly suspicious. With this information, the CAD system can assist the radiologist while reviewing mammograms with the identification and interpretation of breast abnormalities.

CAD systems are not intended to replace the radiologist, but to assist the radiologist during reviewing by indicating suspicious regions. Advantages of a CAD system are that these systems are objective en consistent, and are able to identify suspicious regions that might be overlooked or misinterpreted. The use of CAD systems can (potentially) decrease the need of multiple readings.

The CAD system at the UMC St. Radboud Nijmegen uses four steps to classify regions [9]. These steps are presented in Figure 9.

1. The mammographic image is segmented into breast tissue, background and pectoral muscle and some corrections are being made;
2. Initial detection of suspicious pixel-based locations is performed;
3. Regions and region-based features are extracted;
4. Regions are classified using a classifier.

The suspiciousness for each region is calculated by the CAD system, based on other features. Per mammogram the number of regions is limited to the five most suspicious regions.

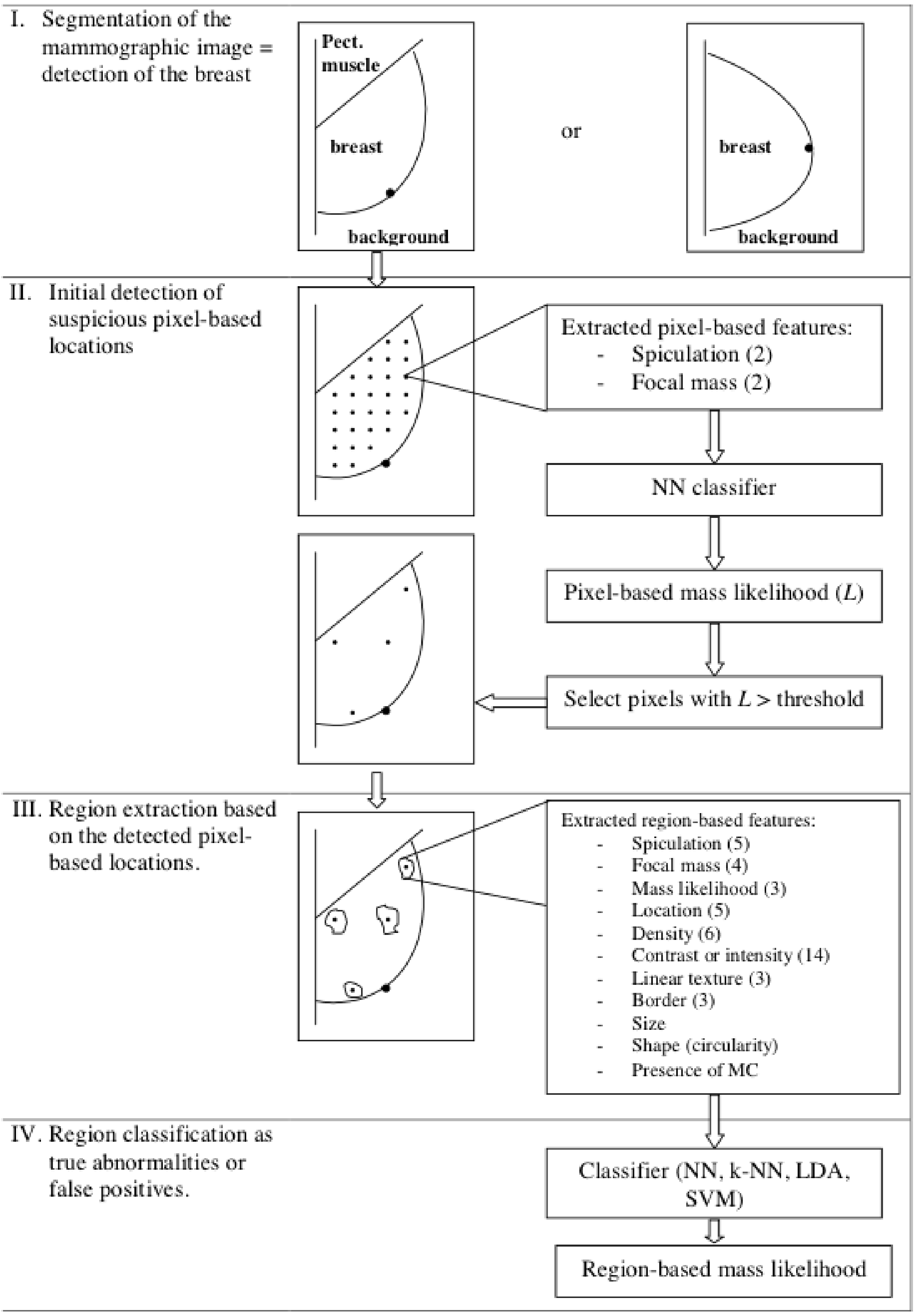

I. Segmentation of the mammographic image = detection of the breast
Pect. muscle
breast
background
or
breast
background
II. Initial detection of suspicious pixel-based locations
Extracted pixel-based features:
- Spiculation (2)
- Focal mass (2)
NN classifier
Pixel-based mass likelihood (L)
Select pixels with L > threshold
III. Region extraction based on the detected pixel-based locations.
Extracted region-based features:
- Spiculation (5)
- Focal mass (4)
- Mass likelihood (3)
- Location (5)
- Density (6)
- Contrast or intensity (14)
- Linear texture (3)
- Border (3)
- Size
- Shape (circularity)
- Presence of MC
IV. Region classification as true abnormalities or false positives.
Classifier (NN, k-NN, LDA, SVM)
Region-based mass likelihood

Figure 9: General scheme of the UMCN CAD system

2.8 Features

For each region, specific *features*, are calculated by the CAD system after the preprocessing step. These features are properties of the region such as its area, the distance to skin or its contrast. The features are represented by continuous values. This section gives a description of the features that are used throughout this thesis.

In total there are 81 features calculated for each region [28]. Some of them represent in essence the same feature, but are calculated using different algorithms (see the `Contrast` feature below).

The features can be categorized into two groups:

1. Features that have been extracted directly from the mammographic image by the CAD system during step three (see previous section). These features can be observed directly in the mammogram. For instance, the distance of a region to the skin or the contrast of a region compared to its surrounding area. These features will be referenced to as *observed features.*

2. Features that are calculated using the CAD system. These features are calculated using region-based or pixel-based features. These features include false-positive level (FPLevel) and mass likelihood (MassLik). These features will be referenced to as *calculated features*

In this research, a subset of 9 observed features that are expected to contribute most to the detection of cancer have been selected to learn Bayesian models from the data. This set of features consists of `{LocX, LocY, d2skin, Contrast, Isodense, Spiculation, FocalMass, LinTexN, RegSize}`. Also, two calculated features (`FPLevel`, `MassLik`) are being used. Each of these 11 features is described below.

The features that are directly observed from the mammogram:

Relative location: The relative location is captured by the position of the region on the x- and y axis in a normalized breast. The location of a region is important because most malignancies (45%) develop in the upper outer quadrant of the breast toward the armpit [28]. The location features will be referenced to as `LocX` and `LocY`.

Distance to skin: The distance to skin is the shortest distance of the region to the skin. This feature will be referenced to as `d2skin`.

Contrast: The contrast of a region is its whiteness relative to other breast tissue. A region with a higher contrast than other parts of the mammographic image is more likely to be a mass, so this feature is important. There are different ways of calculating the contrast. The contrast value used in this research is the difference in the mean pixel value between the pixels inside the region and the pixels outside the region.

Spiculation: The spiculation feature indicates if the margins of a region are *spiculated*, a characteristic of the margin of tumors. A region is spiculated if straight, thin lines radiate from a central point or mass. This feature will be referenced to as `Spiculation`.

Focal mass: Indicates the presence of a circumscribed lesion. This feature will be referenced to as `FocalMass`.

Linear texture: Indicates if the region has linear texture characteristics, which are common to normal breast tissue. This feature is expected to be correlated to the `Spiculation` feature: if linear texture is present, the region can not be spiculated and vice-versa. This feature will be referenced to as `LinTex`.

Size of the region: This feature represents the area of the region in cm^2 relative to a typical breast tumor. A small value indicates that the region is similar in size to a typical breast tumor. Most breast tumors are about 2 cm^2 in size. This feature will be referenced to as `RegSize` and is defined as

$$RegSize =\mid size - typicalsize \mid$$

The features that are not extracted directly from the preprocessed mammographic image, but are calculated by the CAD system based on observed features, including the ones used in this research. These calculated features are:

False positive level: The false-positive level indicates the average number of normal regions in a image with the same or higher likelihood scores. A higher value means that similar regions occur frequently and that the region is most likely benign. This feature will be referenced to as `FPLevel`.

Mass likelihood: The malignancy likelihood calculated by the CAD system, based on a neural network supervised learning (taking into account automatically computed region features). This feature will be referenced to as `MassLik`.

Calcifications

This research concentrates on the detection of masses and no features that relate to the presence of calcifications (see Section 2.6.6) have been used during the research. The reasons for this are:

- Calcifications are usually quite easily detected during the reading, and some automatic systems already show good performance on detecting them [15],

- Masses occur more frequently as an indication of breast cancer development, and,
- Breast masses misinterpretation seems to be a more common cause of cancers being missed during mammogram reading [15].

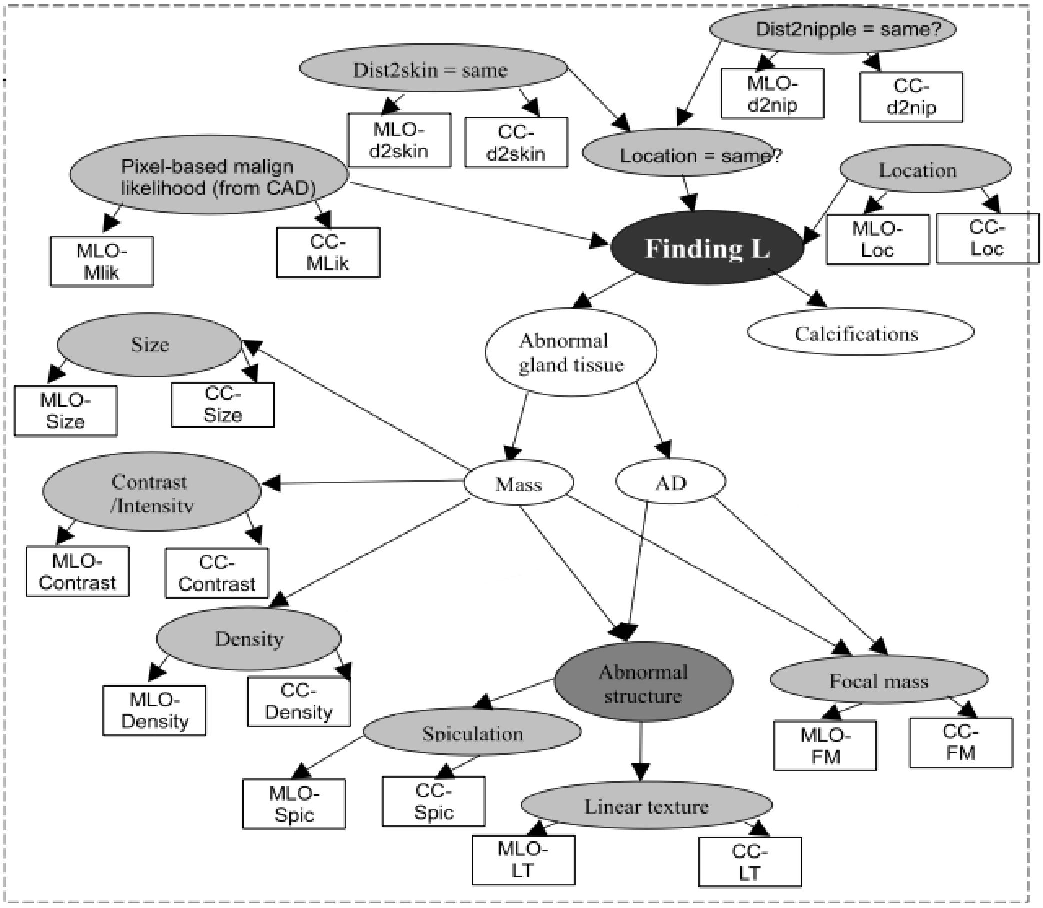


Figure 10: Mammography-feature-based model [12]

3 Bayesian network theory

A Bayesian network (BN) is a probabilistic graphical model that represents a set of random variables and their probabilistic independencies. Formally, it is defined as a pair (G, P) where:

- G is a directed acyclic graph (DAG) $G = (\mathbf{V}, \mathbf{E})$ that encodes a set of conditional independence assertions about variables in the set of variables $\mathbf{X}$: $\{X_1, ..., X_n\}$. The nodes (vertices) $\mathbf{V}$ of G represent the variables $\mathbf{X}$ in a one-to-one correspondence. The *lack* of arcs (edges) $\mathbf{E}$ between nodes represent conditional independencies between the corresponding variables.
- P is a set of local probability distributions associated with the variables.

The local probability distributions $p \in P$ define the joint probability distribution

$$P(X) = \prod_{i=1}^{n} p(X_i|Pa(X_i))$$

on the variables. A Bayesian network structure G is an I-map (independency mapping) of P. In I-maps each independence relationship modelled in the graph G has to be consistent with the joint probability distribution P and each dependence relationship represented in the joint probability distribution P has to be present in the graph representation G [14].

If the value of a node is observed, the node is called an evidence node. If a node has no parents, its local probability distribution is unconditional, otherwise it is conditional.

Under some conditions, the DAG can be interpreted causally. In this case, the nodes that correspond to the random variables from the domain and the arcs are direct causal relations between these variables [25, 18]. This means that the parent variables have a causal influence on the values of their child variables.

A BN can be constructed from domain knowledge. An expert determines the variables in the domain of interest and the relationships among these variables, so a DAG G can be constructed. Then the conditional probabilities given G are determined.

Domain knowledge is often not sufficient to construct a complete BN. If data is available, both the structure and parameters can be learned using this data. Section 4 is dedicated to learning BNs from data.

3.1 Bayesian networks

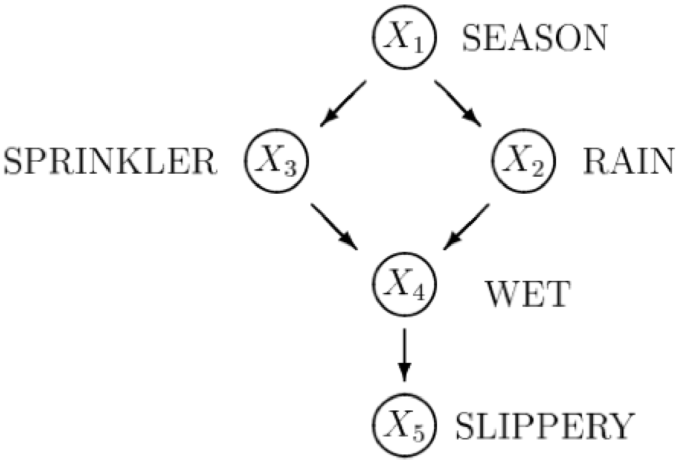


Figure 11: A Bayesian network representing causal influences among five variables [26]

Figure 11 illustrates a simple Bayesian network that describes the relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during that season, whether the pavement would get wet (X_4), and whether the pavement would be slippery (X_5).

At each node, the conditional probability distribution (CPD) is specified. If the variables are discrete, this distribution can be represented as a table (CPT) which contains for each combination of values of the nodes' parents the probability that the node takes on each of its different values.

In the sprinkler example, for example, the CPD's can be specified as the following tables (CPT's).

SEASON: Since X_1 has no parents, the CPT specifies the prior probability that it is the rainy season. A probability of 0.5 is assumed for having the rainy season.

$P(X_1 = F) = P(X_1 = T) = 0.5$

RAIN: When it is the rainy season ($X_1 = T$), the probability of rain is 0.8. When it is not the rainy season ($X_1 = F$), the probability of rain is 0.2.

X_1	$P(X_2 = F)$	$P(X_2 = T)$
F	0.8	0.2
T	0.2	0.8

SPRINKLER: The probability that the sprinkler is on during the rainy season is low (0.1), but 0.5 during the dry season.

X_1	$P(X_3 = F)$	$P(X_3 = T)$
F	0.5	0.5
T	0.9	0.1

WET: In the model we can see that the event 'wet pavement' has two possible causes: either the sprinkler is on, or it is raining. This table specifies the probability of the pavement being wet, conditional on the values of these causes.

X_2	X_3	$P(X_4 = F)$	$P(X_4 = T)$
F	F	1.0	0.0
F	T	0.1	0.9
T	F	0.1	0.9
T	T	0.01	0.99

3.2 Basic probability theory

A probability distribution is function of Boolean expressions to the closed real interval [0,1]. Two basic probabilistic rules are used to compute probabilities of interest from the specification of a BN. The first rule, *marginalization*, is used to sum out irrelevant variables from a joint probability distribution:

$$P(A) = \sum_{b \in B} P(A, b) \tag{1}$$

In additions, often conditional probabilities are computed to determine the effect of observations or evidence on uncertainty. This is done using Bayes' theorem, is a formula used for calculating conditional probabilities. In BNs it is used for inference in which evidence E is used to update the probability that a hypothesis H may be true. Bayes' theorem adjusts probabilities given new evidence in the following way:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{2}$$

Where:

- H is a specific hypothesis;
- $P(H)$ is the prior probability of H.

- $P(E)$ is the marginal probability of E: the a priori probability of observing E under all possible hypothesis.
- $P(E|H)$ is the conditional probability of observing evidence E, given H being true.
- $P(H|E)$ is the posterior probability of H being true, given E is observed.

If we consider the sprinkler example and observe that the pavement is wet ($X_4 = T$). From the model we know that there are two causes for this: either it is raining ($X_2 = T$), or the sprinkler is on ($X_3 = T$). We can use Bayes' rule to infer the posterior probability of both causes (respectively $P(X_2 = T|X_4 = T)$ and $P(X_3 = T|X_4 = T)$) and see which cause is more likely:

The probability that the sprinkler is on, given the wet pavement:

$$P(X_3 = T|X_4 = T) = \frac{P(X_3 = T, X_4 = T)}{P(X_4 = T)} =$$

$$\frac{\sum_{x_1,x_2 \in \{T,F\}} P(X_1 = x_1, X_2 = x_2, X_3 = T, X_4 = T)}{\sum_{x_1,x_2,x_3 \in \{T,F\}} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = T)} =$$

$$\frac{0.2781}{0.6471} = 0.430$$

The probability that it rains, given the wet pavement:

$$P(X_2 = T|X_4 = T) = \frac{P(X_2 = T, X_4 = T)}{P(X_4 = T)} =$$

$$\frac{\sum_{x_1,x_3 \in \{T,F\}} P(X_1 = x_1, X_2 = T, X_3 = x_3, X_4 = T)}{\sum_{x_1,x_2,x_3 \in \{T,F\}} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = T)} =$$

$$\frac{0.4581}{0.6471} = 0.708$$

So we can see that it is more likely that the grass is wet because it is raining.

3.3 d-separation

The DAG G encodes independencies between variables. Conditional independence can be determined by the (graphical) property of d-separation, in which 'd' stands for ‘directional’. If two sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ are d-separated in G by a third set $\mathbf{Z}$, the corresponding sets of variables $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ are independent given the variables in $Z \in \mathbf{Z}$.

The basic idea of d-separation is to associate *dependence* between variables with the existence of a connecting path between the corresponding nodes and *independence* with separation (no connecting path). If we have disjoint sets of nodes $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{Y}$ are d-connected by $\mathbf{Z}$ if and only if:

- there exists an undirected path U between $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$;
- for every *collider* C in U, C or a descendant of C is in $\mathbf{Z}$. A collider in a path is a node with two incoming arcs. In Figure 12, node B is the collider;
- no *non-collider* in U is in $\mathbf{Z}$. Every node that is not a collider node is a non-collider.

$\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in G if and only if they are not d-connected by $\mathbf{Z}$ in G.

This implies that two variables are (unconditionally) independent if the corresponding nodes are d-separated by the empty set $\emptyset$.

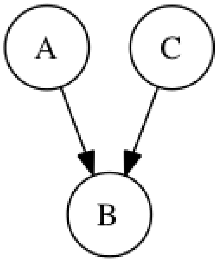
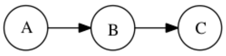


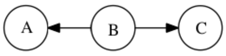
Figure 12: Collider

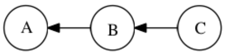
3.4 Markov equivalence

Definition 3.1. *Two DAGs are said to be Markov equivalent (noted $\equiv$) if they imply the same set of conditional dependencies. The Markov equivalent classes set (named $\mathcal{E}$) is defined as $\mathcal{E} = \mathcal{A}/_{\equiv}$ where we named $\mathcal{A}$ the DAGs’ set. [20]*

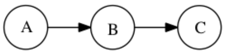
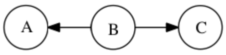
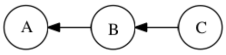
This means that the DAGs



, 

and 

are equivalents and can be denoted as

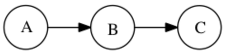
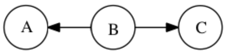
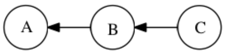
 ≡  ≡ .

3.5 Essential graphs

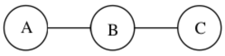
Definition 3.2. *An arc is said to be reversible if its reversion leads to a graph which is equivalent to the first one. [20]*

The space of essential graphs is defined as the set of chain graphs, i.e., acyclic graphs that have directed and undirected edges. The essential graph acts as a class representative for BNs that encodes the same probabilistic independence information [14].

This means that the DAGs

,  and 

can be represented by the essential graph:

.

3.6 Specific Bayesian networks structures

Two network structures that are used in this thesis are described in this section.

3.6.1 Naïve Bayesian network

A naïve Bayesian network, or NB, is a special case of a Bayesian network. It consists of one class variable C which is conditional on a set of feature variables $\mathbf{F}$: $\{F_1, ..., F_n\}$. All variables in $\mathbf{F}$ are assumed to be conditional independent from each other given C, which means that $P(F_i|C, \mathbf{F}\backslash F_i) = P(F_i|C)$ for each i. The arcs are going from the class node C to all feature nodes $F \in \mathbf{F}$.

Despite their naïve design and over-simplified independence assumptions, naïve Bayes *classifiers* often work well in many complex real-world situations.

3.6.2 Tree augmented naïve Bayes network

The major drawback of naïve Bayes is that it assumes all feature variables to be conditionally independent given the class variable. In practice, these variables are often strongly related to each other. The tree augmented naïve Bayes, or TAN,

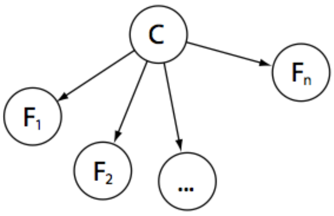


Figure 13: Naïve Bayes network

model retains the basic structure of naïve Bayes, but also permits each feature node to have at most one other feature node as a parent. This allows the model to capture dependencies between the feature nodes.

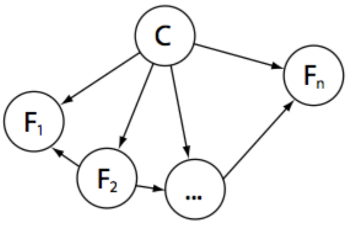


Figure 14: Tree augmented naïve Bayes network

3.7 Bayesian classifiers

Bayesian networks can be used to infer probabilities. For instance, in this research we use Bayesian networks to infer the probability of a region being cancerous.

A *classifier* is a mapping from discrete or continuous values to a set of labeled classes. In this research, we want to map the output of a Bayesian network to the two classes `cancerous` and `non-cancerous`.

By applying different threshold values the probabilities can be mapped into the different labeled classes `cancerous` and `non-cancerous`. For instance, a region with a probability of being cancerous of 0.4 will be classified as `non-cancerous` with a threshold of 0.5. However, if the threshold is lowered to 0.2, the case will be classified as being `cancerous`.

With a binary classifier (two classes, e.g. positive (p) and negative (n)), there are four possible outcomes. If the outcome of the classifier is p and the actual value is also p, this is called a true positive (TP) or 'hit'. However, if the actual value is n, this is called a false positive (FP) or 'false alarm'. When the predicted value is n and the actual value is also n, this is called a true negative (TN) or 'correct rejection'. If the actual value is p, this is called a false negative (FN) or 'miss'.

For example, in this research this means:

True positive (TP): A cancerous case is classified as cancerous;

False positive (FP): A non-cancerous case is classified as cancerous;

True negative (TN): A non-cancerous case is classified as non-cancerous;

False negative (FN): A cancerous case is classified as non-cancerous.

With multiple cases and a classifier we can calculate the *sensitivity* and the *specificity* of a classifier. The sensitivity is the number of correctly classified cancerous cases divided by the total number of cancerous cases.

$$sensitivity = \frac{TP}{TP + FN} \tag{3}$$

The specificity is the number of correctly classified non-cancerous cases divided by the total number of non-cancerous cases.

$$specificity = \frac{TN}{TN + FP} \tag{4}$$

3.8 Evaluation of classifier performance

To evaluate the performance of a classifier, a measure is needed. Using the percentage of correctly classified cases is not a good measure. Since the probability of breast cancer being present in a mammogram is very low, a high accuracy could simply be reached by predicting that no breast cancer is present for each mammogram. Therefore, the performance of a classifier should be described by its sensitivity and its specificity.

The receiver operating characteristic (ROC) curve is in the field of Radiology an widely accepted methodology for evaluating the performance of a classifier. In ROC curves the true positive rate (TPR), or *sensitivity*, versus the false positive rate (FPR) of a classifier for different threshold values is plotted.

With a threshold of 0, every case is classified as cancerous, so the sensitivity and 1 - specificity are both 1. However, all non-cancerous cases are also identified as cancerous. With a threshold of 1, all cases are identified as non-cancerous, so the sensitivity and 1 - specificity are both 0.

To compare classifiers, we bring back the ROC curve to a single scalar value that represents the expected performance. A common used method is to calculate the area

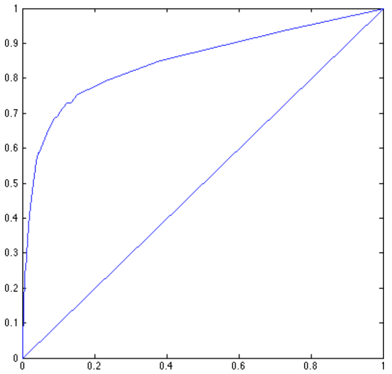


Figure 15: Example of ROC curve

under the ROC curve AUC. The AUC value lies in the range 0.5 - 1. A worthless classifier's curve is a straight line from (0,0) to (1,1) and has an AUC value of 0.5.

When a classifier has a low FPR and a high TPR, it has a high AUC value. This means that the classifier will perform better than a classifier with a lower AUC value.

The AUC value has also an important statistical property: the AUC value of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [10].

In this thesis, the AUC value will be used to represent expected classifier performance.

4 Learning Bayesian Networks from data

As explained in the previous section, a Bayesian network consists of two parts, its *structure* G and the parameters P, which is the conditional probability distribution associated with the network topology. It is possible to learn both of these from data. However, it is much harder to learn the structure of the network than its parameters. Sections 4.1 and 4.2 will respectively give details on structure learning and parameter learning.

4.1 Structure Learning

A naïve idea to find the best network structure is to score all possible DAGs using a *scoring function* and choose the DAG that has the highest score. The problem with this method is that the number of possible DAGs is super exponential in the number of nodes. Robinson has proved in [27] that the number of different network structures $r(n)$ for a network with n nodes, is given by the formula of Equation 5.

$$r(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} r(n-i) \tag{5}$$

As we can see in Table 1 it becomes impossible to search the space of DAGs exhaustively in reasonable time with values of $n \geq 6$, so heuristic methods are needed to find (sub)optimal network structures.

Number of variables n	Number of possible DAGs
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	1,138,779,265
8	78,370,2329,343
9	1,213,442,454,842,881
10	4,175,098,976,430,598,100

Table 1: Number of possible DAGs for different number of variables

In general there are two different approaches to structure learning. Both view the structure learning problem differently.

Constraint-based: This approach tries to measure the dependencies and conditional independencies in the data. These dependency and independency relationships are measured by a statistical tests on the data set. In general, constraint-based methods start with a fully connected graph from which edges are removed when certain conditional independencies are present in the data.

Optimization-based search: This approach searches the space of possible DAGs and returns the structure that best fits the data. Different algorithms exist to limit the size of the space that is searched for possible DAGs.

In this research, we use the more popular optimization-based search approach. This heuristic approach requires:

A scoring function which measures how well a network structure fits the data;

A search algorithm, which tries to find the network structure with the highest score.

Both scoring functions and search algorithms are described in this section.

4.1.1 Scoring functions

To measure how well a network structure fits the data, we need a scoring function which calculates the probability of network graph G given data set D, $P(G|D)$. This scoring function should be *equivalent*, which means that the measure should return the same score for Markov equivalent DAGs. Such a scoring function can be used to rank Bayesian network structures using:

$$q = \frac{P(G|D)}{P(G'|D)} \tag{6}$$

From Bayes' theorem it follows that the model that maximizes the posterior probability is the model that maximizes $P(D,G)$:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \tag{7}$$

Thus:

$$q = \frac{P(D|G)P(G)/P(D)}{P(D|G')P(G')/P(D)}$$

$$q = \frac{P(D|G)P(G)}{P(D|G')P(G')}$$

$$P(G, D) = P(D|G)P(G)$$

So for each Bayesian network, we need to determine:

$$log\ P(G, D) = log\ P(D|G) + log\ P(G)$$

If we assume:

1. No missing values in data set D;
2. Cases $c \in D$ have occurred independently;
3. Discrete network parameters.

This can be calculated by:

$$\log P(D|G) = \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} n_{ijk} \cdot \log_2 \left(\frac{n_{ijk}}{n_{ij}} \right) \tag{8}$$

Where

- N is the number of variables in the network;
- q_i is the number of states over the parents $Pa(X_i)$ in the graph, with $q_i = 1$ if $Pa(X_i) = \emptyset$;
- r_i is the number of states for variable X_i;
- n_{ijk} is the number of cases in D with X_i in state k and $Pa(X_i)$ in state j;
- n_{ij} is the number of cases in D with $Pa(X_i)$ in state j.

This measure estimates the maximum likelihood parameters for the model.

Overfitting

One thing that should be taken into consideration is that the model with the maximum likelihood is a complete (fully connected) graph, because adding an arc never decreases likelihood on the training data. This can lead to severe overfitting.

This can be solved in two ways:

- Using a prior to specify that we prefer sparse models. The effect of using a prior is equivalent to penalizing complex models, since the prior probability of a complex model is lower than from a less complex model.

- Adding a penalty term

We can see the prior probability for the structure $P(G)$. If we assume that all Bayesian networks are equally likely, $P(G)$ is a uniform probability distribution and can be replaced by a constant $c \in \mathbb{R}$.

$$logP(G, D) = logP(D|G) + c \tag{9}$$

We can also incorporate background knowledge and specify that we prefer sparse models.

$P(G, D)$ is usually higher for complex networks. Another solution to prevent overfitting is to add a penalty term r that penalizes complexity:

$$r = -\frac{1}{2}k \cdot \log n$$

where k is the number of parameters needed to specify the joint probability distribution.

Scoring functions

The two most widely used scoring functions are the:

Bayesian score [17], which is the marginal likelihood of the model (parameters are marginalized out).

BIC (Bayesian Information Criterion) score [17], which is the log likelihood score with a term added to penalize model complexity. Since two equivalent graphs have the same likelihood and the same complexity, the BIC score is also equivalent.

The Bayesian score is a more accurate scoring function, but it needs more computation

The BIC score can be derived as a large sample approximation to the marginal likelihood, the second component of Bayesian score. In practice, the sample size does not need to be very large for the approximation to be good. The BIC score is defined as follows:

$$\log \Pr(D|G) \approx \log \Pr(D|G, \hat{\Theta}_G) - \frac{\log N}{2} \#G \tag{10}$$

Where

- N is the number of samples;
- $\hat{\Theta}_G$ is the Maximum Likelihood estimation (MLE) of the parameter;
- $\#G$ is the dimension of the model.

As can be seen in Equation 10, the first term is the likelihood and the second term penalizes complex structures. Also, the BIC score does not depend on a prior.

4.1.2 Search algorithms

The aim of a search algorithm is to find the most probable network structure given a dataset. The algorithms described in this section have different ways of searching for this structure.

Exhaustive search

The exhaustive search method does not minimize the search space of possible DAGs, but generates all possible DAGs and chooses the one with the highest score. The score of the resulting network is the global optimum, so the structure is the best possible structure. As mentioned before, this method does not find the optimal network in reasonable time for networks with more than 6 nodes, since the number of possible DAGs is super exponential in the number of nodes [27].

Greedy search

A simple search algorithm is greedy search. The greedy search algorithm starts with a initial network structure G. For each step, it defines a set of neighborhood graphs and computes the scores for every graph in this neighborhood set. The neighbor graph with the highest score is selected and used for the next iteration. The search is stopped when there is no neighborhood network graph with a higher score than the current structure.

The most common way of defining the set of neighborhood graphs is the set of graphs that differ only one arc operation with the current graph, while taking the acyclicity constraint into consideration. The possible arc operations are:

- insert arc,
- remove arc, and
- reverse arc.

Other definitions of the set of neighborhood graphs are possible (see [7]).

The greedy search algorithm can be initialized with any graph, e.g., the empty graph, a random graph or a network structure determined using another search algorithm. By default, the algorithm is initialized with the empty graph.

The greedy search algoritm does not necessarily reach a global maximum, but can also converge to a local maximum. A solution to this problem is to add a second phase in which the network structure is randomly disturbed when a maximum is reached and repeat the algorithm for a number of times.

Greedy equivalence search

The greedy equivalence search algorithm works similar to the greedy search algorithm described in the previous paragraph. However, it searches the space of essential graphs (see section 3.5). It does not require an initial graph as an input, but the empty graph is used as a starting point for the algorithm.

The algorithm consists of two phases. First, arcs are added or removed until an network structure is found that gives the highest score. After this first phase, arcs are removed in the second phase.

The advantage of using greedy equivalence search is that the space of essential graphs is a subspace of all graphs and the amount of search reduces the search space.

K2 algorithm

The K2 algorithm was introduced by Cooper and Herskovits in [8] and is a special case of a greedy search algorithm. It minimizes the search space by having a initial ordering on the nodes and limiting the number of parents a node can have. It requires:

- A set of n nodes;
- An ordering on the nodes;
- An upper bound u on the number of parents a node may have;
- A dataset D containing cases.

To minimize the search space, the K2 algorithm requires an specific ordering in the nodes. A node can be only the parent of nodes which are appearing after it in the ordering. So, the first node in this ordening can't have a parent. The second node can only have node one as a parent. The n^{th} node can only have (some) of the $n-1$ nodes as parents. The search space becomes the subspace of all the DAGs admitting this order as topological order. As a consequence, the search space under this constraint is much smaller than the entire search space.

The algorithm assumes for a node that it has no parents. Then it incrementally adds the parent whose addition increases the probability of the resulting structure most. The algorithm stops adding parents to the node if the number of parents equals u or when the addition of no single parent can increase the probability.

The major disadvantage of K2 is that the input node ordering is of great influence on the performance of the algorithm. Using an improper order will result in poor results.

Naive Bayes

The algorithm for learning a naive Bayes network is trivial. The search space is empty, since the search space is minimized by allowing only NB structures (exactly one network with a fixed class node C).

TAN

A very simple structure learning algorithm is used for finding the optimal TAN structure. The search space of DAGs is minimized by allowing only TAN structures. The best tree is obtained using the *Maximum Weight Spanning Tree algorithm* [16].

4.2 Parameter learning

To fully specify the joint probability distribution, it is necessary to specify the local conditional probability distribution for each node N (the probability distribution for N conditional upon its parents). This distribution can have any form, such as discrete or Gaussian. If the distribution is discrete, a conditional probability table (CPT) can be used.

The goal is to find the parameters of each CPD which maximizes the likelihood of the training data D, where all N cases in D are assumed to be independent.

5 Experiments

This section describes the experiments that have been performed in this research and their results. The experiments have been divided into three parts, each with a different goal. These parts, including their goals are:

1. Learn restricted models. In this experiment, only 4 feature variables are used. The goal of this experiment is to compare the models obtained using structure learning with a model that is constructed using expert knowledge and a model found using exhaustive search. The result of this experiment is described in Section 5.2.
2. See what the influence of certain features variables is on the resulting model. In this experiment, a difference is made between observed feature variables and feature variables that have been calculated by the CAD system. Models are learned using:
    - 9 observed feature variables and 2 calculated feature variables, and
    - with the same 9 observed feature variables (without calculated feature variables).

   The results of this experiment is described in Section 5.3.
3. Learn models with feature variables from both the CC and MLO view. Since a MLO and a CC view of the same breast are related, it is interesting to take features of the regions of one view into account when learning the classification of the regions of the other view. In these experiments, variables of regions in the CC view are used when learning the classification of regions from the MLO view. The result of this experiment is described in Section 5.4.

The set up of the experiments is described first in this section. It covers:

- The software and toolboxes that have been used are described in Section 5.1.1;
- The dataset that has been used, the features that have been selected for the experiments and the data discretization techniques are described in Section 5.1.2;
- The initialisation of the structure learning algorithms that have been used is described in Section 5.1.4.

5.1 Set up

For calculating the scores of networks, the Bayesian score is used, unless stated otherwise. The Bayesian score is a more accurate scoring function than the BIC score, but it needs more computation.

5.1.1 Software and toolboxes

The experiments described in this section are performed using Matlab 7.5.0.338 (R2007b) for Mac OS X. Additional toolboxes which provide functions for the use with Bayesian networks have also been used. These toolboxes are:

Bayes Net Toolbox by Kevin Murphy [21]. This toolbox provides the basic functionality for using Bayesian networks. The version used in this research is version 1.0.4, which was last updated February 11, 2007.

BNT Structure Learning Package by Philippe Leray and Olivier Francois. This toolbox provides additional structure learning functions for the use with Murphy's Bayes Net Toolbox. It is proposed and documented in [20] and distributed on the GNU Library General Public License.

5.1.2 Dataset

The dataset with mammograms that are used for the experiments have been obtained from the Dutch breast cancer screening programme 'Bevolkingsonderzoek Borstkanker Nederland'. The dataset contains data of 1063 patients or *cases*. Screening was performed in each of the patient's breasts, giving a total of 2126 screened breasts. Both the CC- and MLO-views of are present for all breasts.

For each mammogram the number of regions is limited to the 5 most suspicious regions (among the — at most — 10 regions detected and classified by the CAD system). In total there are 10478 MLO regions and 10343 CC regions in the dataset. For each case it is known if the patient has breast cancer. In total, 385 regions are known to be cancerous by pathological report. This means that a cancer has been identified in 36.2% of the cases and that it is known which regions represent the cancer. On region level, this means that 1.8% of the regions is cancerous.

The dataset contains features from both the CC and MLO views. In the first two experiments, this data is kept separated. This means that structures are learned using either CC data or MLO data. In the third experiment, the data of both views are combined.

The dataset is divided into a trainingset, used for learning the models, and a testset that is used for scoring the models afterwards. These sets have an equal distribution of cancerous regions.

Identification of region

Each mammogram is identified by a code. The first 8 positions are used for case identification. This is followed by 'm' or 'c' (respectively MLO or CC view) and 'l' or 'r' (respectively left or right breast). For example: `l0100025ml` means "MLO view of left breast of case l0100025". For each mammogram there are multiple regions available which are identified by a number $\{0..10\}$, which is the number the CAD system has given to the region. The features that are calculated by the CAD system are described in Section 2.8.

5.1.3 Discretization of the data

Most structure learning implementations work with discrete values for the variables. The values of the features in the dataset are real-valued, so to perform structure learning, the data needs to be discretized.

For the discretization of the data, the `hist_ic` implementation present in the BNT Structure Learning Package was used. This function discretizes the data into an optimal number of bins according to a cost function based on Akaike's Criterion [20]. The function takes the continuous data and a penalty term as input. For the penalty term, the default setting for the algorithm is used.

After applying this function to the dataset, the resulting number of bins per variable varied between 2 and 33. For some variables, this number of bins was too high to obtain useful results. This was verified by experiments. The number of bins for the resulting discrete dataset is reduced by merging smaller bins into bigger bins until a maximum number m of bins remains.

To see the influence of different discretization, the value of the maximum number of bins m has been varied from 2 (binary data) to 20. The class variable `Finding` has always 2 bins, that correspond to the two output classes `cancerous` and `non-cancerous`.

Using the resulting discrete datasets, TAN and GS structures and parameters have been learned. For each resulting structure, the Bayesian score has been calculated to see how well the model fits the data. Also the area under the ROC curve (AUC) has been calculated as an indication of how well the trained model performs as a classifier. The results are presented in appendix A in Table 11.

We can see that classification performance decreases rapidly for the TAN structure when the number of bins becomes greater than 10. The classification performance of the structures learned with GS varies a lot. Taking results from both TAN and GS into account, maximizing the number of bins to 7 seemed a reasonable choice.

This means that the number of bins for each variable in the discretized dataset is between 2 and 7.

Since discretization is not the main aim of this research and finding the optimal number of bins for each variable is computationally expensive, the maximum number of bins of 7 is used for all features as a good estimate. Using further statistical analysis, the optimal number of bins for each feature can be determined and it is expected that this will increase the quality of the models.

5.1.4 Initialisation of the algorithms

Order of the variables for K2

The K2 algorithm needs an ordering of the nodes to limit the search space. With a small number of variables experiment, it is possible to perform structure learning using K2 for all permutations of the variables. In the first experiment this is possible since with 5 variables there are only $(5! =)$ 120 permutations of the variables possible. For models with more variables this is not feasible. For example: if we have 10 variables, there are over 10^6 permutations; with 20 variables this increases to over 10^{18} permutations. So, for the second and third experiment we need to determine some pre-ordering on the nodes. The ordering is obtained using the expert model and experiments with different orderings to fine-tune the order. The K2 algorithm used the BIC scoring function for evaluation of the intermediate results while performing the algorithm. This setting is chosen instead of the Bayesian score, because it needs less computation and the results are good. It is also the default scoring function for the K2 algorithm in BNT [21].

Initial network structure for GS

The greedy search algorithm needs an initial network structure as a starting point for the algorithm. During the research, different variations have been used for the initial network structures. The models did not differ much in structure or score and were in some cases equivalent to other models learned using greedy search with another initialization. In this thesis only the results of the greedy search algorithm with an empty network structure as initial structure are presented for clarity reasons. Using the empty structure is the default choice for the greedy search implementation in BNT [21].

Neighborhood definition for GS

There are different ways of defining the set of neighborhood structures. In this research, the implementation from BNT for Matlab [21] is used, which defines the set of neighborhood structures as described in section 4.1.2.

5.2 Restricted model

A small sub-model containing 5 variables is selected from the expert model (shown in Figure 16). The choice to use 5 variables is made so it is possible to perform exhaustive search. Exhaustive search finds the network structure that fits best to the data and is feasible in reasonable time for 5 variables. With more than five variables, this is not feasible as the number of possible possible DAGs is super exponential in the number of nodes.

The goal of using a part of the expert model is that we can compare the part of the expert model with the results of structure learning and the model found using exhaustive search.

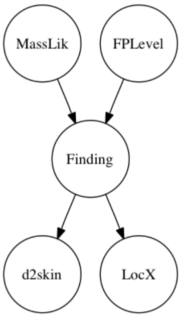


Figure 16: Sub model extracted from the expert model

To compare the fitness of the expert model, two models which do not incorporate any of the knowledge incorporated in the expert model are made. These models are kept very simple and consist of a class variable C `Finding` and the same set of feature variables $\mathbf{F}$ `{MassLik, FPLevel,d2skin, LocX}` as the expert model.

The two reference models are:

Independent variables. All variables are considered independent, which means there are no arcs in the model. This model is described in Figure 17(a).

Naive Bayes. This model follows from the description given in Section 3.6.1. The class variable `Finding` is conditionally dependent from the four feature variables. All feature variables are considered to be independent. This model is described in Figure 17(b).

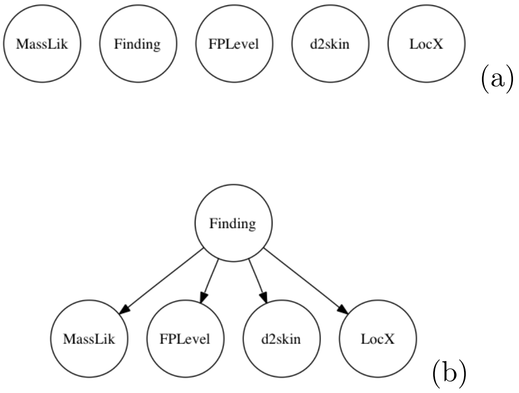


Figure 17: Models with independent variables and conditional independence (naïve Bayes).

5.2.1 Learned models

TAN

When learning the optimal TAN structure, it is needed to provide the class node C and the root node of the tree F_{root}. The class node is fixed (Finding) and there are only four possible variables to use as the root node. All possible TAN structures were learned. The resulting structures are Markov equivalent and are shown in Figure 28. One of the models (using FPLevel as root node) is shown in Figure 18.

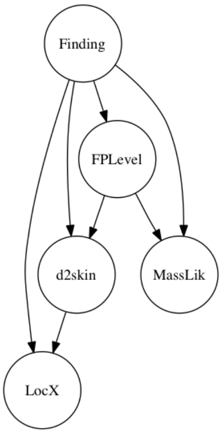


Figure 18: Learned TAN structure

K2

For the K2 algorithm, it is needed to provide an ordering on the input nodes. Since our model consists of only five nodes, it is possible to learn the structure with all (5! = 120) permutations of the required ordering on the nodes. For all resulting models, the scores are calculated and the model with the highest score is chosen. The model with the highest score is shown in Figure 19.

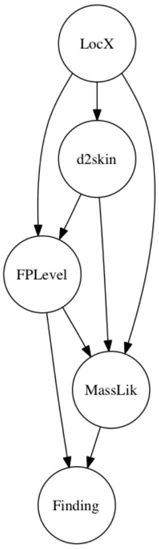


Figure 19: Models found with K2, GS, GES and exhaustive search

Greedy search

The greedy search algorithm needs a initial network structure as a starting point for the algorithm. For the initial structure, different network structures have been used:

1. Empty structure
2. Naive Bayes structure
3. Structure learned using K2

The results of the greedy search algorithm are 3 equivalent dags, which are also equivalent to the structure learned using K2 (see Figure 19).

Exhaustive search

Finally, we learn the optimal network structure using exhaustive search. Since the number of variables is 5, this is feasible. All possible network structures with 5 variables (29281) are scored.

The best performing model found with exhaustive search is equivalent to the model found using K2 and greedy search (see Figure 19. This means that these search algorithms have found the structure that fits the best to the data. Since the search space of all possible DAGs is relatively small and the model is very restricted, this is not surprising.

5.2.2 Scores

The results of the reference models (independent variables and naïve Bayes) and the learned models (TAN, K2, GS, GES) are shown in Table 2. Since the K2, GS, GES and exhaustive search models are Markov equivalent, the scores of these models are equal. The scores of these models are also higher than the reference models (independent variables and naïve Bayes) and slightly higher than the TAN model. This means that these models describe the data more accurately.

	Indep.	NB	TAN	K2	GS	GES	Exhaustive
Bayesian	-30772	-30508	-28506	-28220	-28220	-28220	-28220
BIC	-30771	-30509	-28506	-28213	-28213	-28213	-28213

Table 2: Results of the simple structures

When looking at the learned structures, we can see that in these models Finding becomes conditioned on MassLik and FPLevel, just like in the expert model. The variables LocX and d2skin are not conditioned on Finding as in the expert model.

The values of the Bayesian score and the BIC score are almost the same. This confirms that the BIC score is a good estimate, even with a relatively small dataset.

5.3 Influence of calculated features

This section describes the results of experiments that were performed to investigate the influence of calculated features on both the performance and network structures when learning models from data. By using the calculated features MassLik and FPLevel, the classifier becomes in fact a second-order classifier, since it uses features that are the result of a first classifier (the CAD system).

In the experiment, models are learned using all variables (observed and calculated) and only using observed variables. The structure learning algorithms introduced in Section 4 are used to learn the models. The NB structure follows from the definition given in Section 3.6.1.

5.3.1 All variables

In section 2.8 we have made a distinction between observed features and calculated features. The models learned in this experiment include all features as variables. First, all selected features (`{LocX, LocY, FPLevel, MassLik, d2skin, Contrast, Isodense, Spiculation, FocalMass, LinTexN, RegSize}`), are being used to learn the models. These models are scored using the Bayesian score and the AUC value of the classifier is calculated. The results are shown in Table 3.

	CC		MLO	
	Bayesian score ($\times 10^4$)	AUC	Bayesian score ($\times 10^4$)	AUC
NB	-9.9787	0.8606	-10.1710	0.8319
TAN	-9.5442	0.8092	-9.7866	0.8033
K2	-9.4059	0.8490	-9.6489	0.8299
GS	-9.3385	0.8489	-9.6023	0.8299
GES	-9.3385	0.8489	-9.5944	0.8299

Table 3: Results using all variables

The models that were learned are included in Appendix C.

5.3.2 Only observed variables

Secondly, only features that are observed in the mammographic image are used. This means that the features `MassLik` and `FPLevel` are not used as variables during the learning process. For the other features, the dataset remained the same. The scores and AUC value are shown in Table 4.

	CC		MLO	
	Bayesian score ($\times 10^4$)	AUC	Bayesian score ($\times 10^4$)	AUC
NB	-8.0157	0.8269	-8.1824	0.7786
TAN	-7.7521	0.7700	-7.9304	0.7494
K2	-7.5682	0.7019	-7.7859	0.6950
GS	-7.5612	0.7019	-7.7728	0.6950
GES	-7.5591	0.7425	-7.7634	0.6706

Table 4: Results using only observed variables

The models that were learned are included in Appendix C.

5.3.3 Results

When we compare the models, we notice that the AUC value for the models learned using MLO data and the GS, K2 and GES algorithms are the same (0.8299) when FPLevel and MassLik are included. When comparing these models, it appears that in the structures, Finding is conditioned on FPLevel and not on other variables. This means that when using the model as a classifier, only the FPLevel feature is taken into account and that the entire structure could be replaced by the very simple model without that is shown in Figure 20.

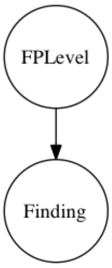


Figure 20: Finding is conditioned on FPLevel

This can be explained because the FPLevel feature is the result of another classifier. When not taking the calculated features FPLevel and MassLik into consideration, a number of observations can be made.

One thing that can be noticed is that the features LocX, LocY and d2skin are related in all learned models. These features describe the location of the region in the breast. In some models, especially those learned using CC data, these variables are independent of the other variables.

It was expected that the Spiculation and LinTex features are related and that models would show relations as shown in Figure 22. The features are more or less complementary: if linear texture is present, the region is not spiculated and vice-versa. In some models this can be observed. However, in only 25% of the learned models this relation is present.

5.4 Combining CC and MLO data

An interesting feature of learning relational models is the ability to consider the features of objects and links that are related to the object that is classified. Since a

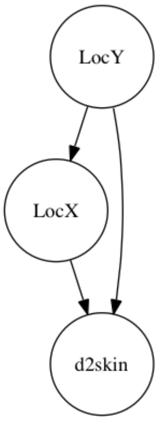


Figure 21: Location features

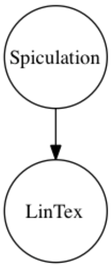


Figure 22: Spiculation and linear texture

MLO and a CC view of the same breast are related, it would, for example, be interesting to take features of the CC regions into account when learning the classification of MLO regions.

This section describes the experiments and results of taking features of CC regions into account while learning the classification of MLO regions. There are two common techniques to combine the datasets of MLO and CC views. In this research, experiments have been performed using both these techniques and the results have been compared.

In this section, each feature variable is prefixed with ‘MLO-’ or ‘CC-’ to indicate which dataset was used.

5.4.1 Aggregate function

The datasets can be combined using an aggregate function. This means that instead of considering the value of certain features, the values of these features are aggregated into a new feature using a function, such as an average, a max, a proportion, or a count of a specific value. To make this more clear, an example is given:

We want to classify an object O and also have a dataset of a related object O_R. The dataset of the object O and the dataset of the related object O_R are shown in Table 5.

F1	F2	F3
1	1	3
1	2	2
2	1	3

F1'	F2'	F3'
2	1	2
1	1	3
2	2	3

Table 5: Dataset of object O and related object O_R

Suppose that feature F2' of the related object O_R contains useful information for the classification of object O. We can use an aggregate function, such as the average or max value, to add another feature to the dataset O. For this example we use the `max` function on feature F2'. The dataset that is used in the learning process becomes:

F1	F2	F3	F2'
1	1	3	2
1	2	2	2
2	1	3	2

Table 6: Aggregated dataset of object O and related object O_R.

In this research, object O translates to a MLO view. The related object O_R translates to the CC view of the same breast. The rows in the datasets translate to the regions in this view and the columns to the features of these regions. The MLO dataset contains 10478 regions from 2126 screened breasts. The aggregated features from the CC dataset are added for each MLO region. Half of the dataset is used for learning a model using greedy search, the other half is used for validation and scoring.

Based on expert knowledge, three features, CC-Spiculation, CC-Contrast and CC-RegSize, have been selected. These features from the CC dataset are aggregated and added to the MLO dataset. The three features have been aggregated using the maximum value and the average value for all linked CC regions. The calculated features MLO-FPLevel and MLO-MassLik are not used in this dataset, because in the previous experiment we have shown that these features are not of interest for learning models from data.

Six models are learned from this dataset using greedy search. These six models are the models for each of the three features with both aggregate functions. The results are included in Table 7.

	Max		Average	
	Bayesian score ($\times 10^4$)	AUC	Bayesian score ($\times 10^4$)	AUC
CC-Spiculation	-8.7686	0.6950	-8.7492	0.6950
CC-Contrast	-8.6639	0.6952	-8.7393	0.6950
CC-RegSize	-8.7552	0.6950	-8.7459	0.6950

Table 7: Results using aggregated features.

When looking at the learned network structures for these models, it appears that the aggregated CC-Spiculation and CC-RegSize variables are independent of the other variables. The aggregated CC-Contrast variable becomes conditioned on the MLO-Contrast feature for both aggregate functions.

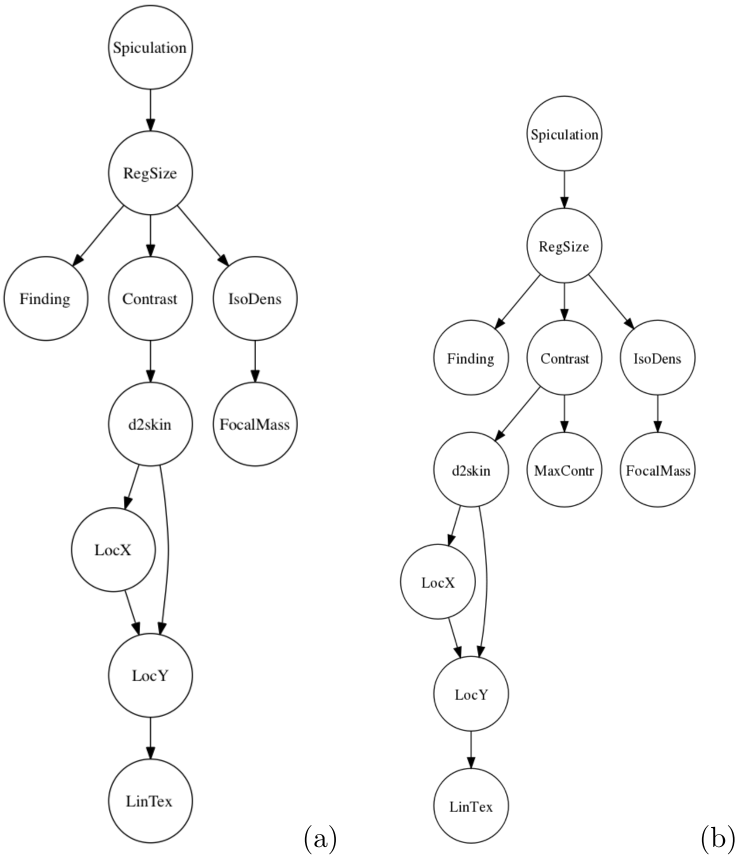


Figure 23: Learned structures using greedy search.

5.4.2 Cartesian product

The Cartesian product of set A and B, denoted as $A \times B$, is the set of all ordered pairs (a, b) with $a \in A$ and $b \in B$. Using the Cartesian product, each MLO region of a patient's breast is combined with each CC region from the same breast. For most views are 5 regions available and for a small number of views 4 regions. When these views are combined using the Cartesian product, there are at most 25 combinations per screened breast. The combined datasets contains data for 2126 screened breasts and has 51088 records.

In this experiment, the same features that are used for the previous experiment have been selected. Four models are learned from this dataset using greedy search:

1. Using MLO data and CC-Spiculation,
2. Using MLO data and CC-Contrast,
3. Using MLO data and CC-RegSize,
4. Using MLO data and CC-Spiculation, CC-Contrast and CC-RegSize.

When CC-Spiculation and CC-RegSize are included, these variables are independent of the rest of the (MLO) variables in the model. When CC-Contrast is included, this feature is conditioned on the MLO-Contrast feature. When CC-Spiculation, CC-Contrast and CC-RegSize are included, CC-Contrast is also conditioned on MLO-Contrast.

The models that are learned using greedy search with MLO data and CC-Contrast, and CC-Spiculation, CC-Contrast and CC-RegSize are included in Figure 24.

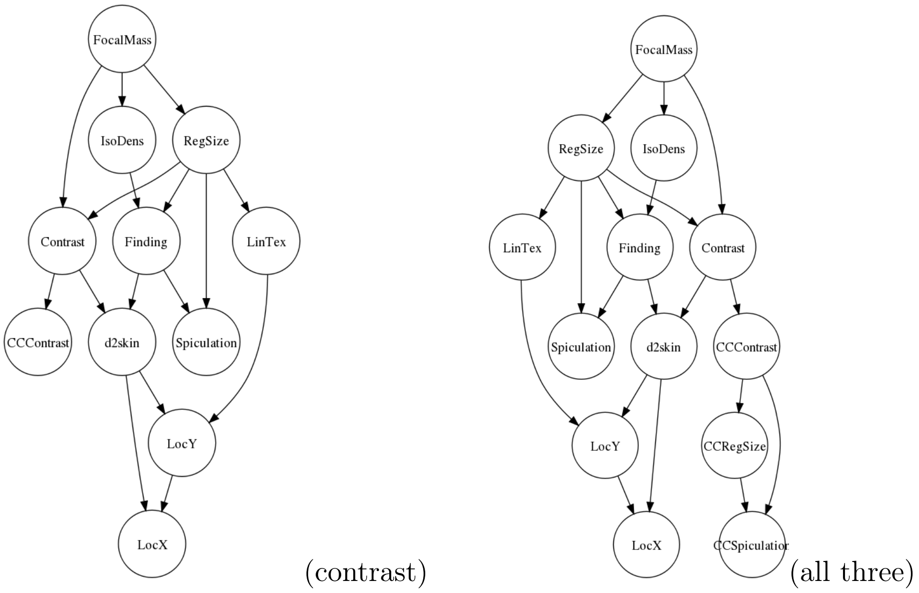
	Bayesian score ($\times 10^5$)	AUC
Spiculation	-4.6290	0.8356
Contrast	-4.6256	0.8356
RegSize	-4.6326	0.8356
Spiculation, Contrast and RegSize	-5.5569	0.8356

Table 8: Results using Cartesian product dataset.

5.5 Modifying greedy search

In Section 5.3, we have seen that in all cases where MassLik and FPLevel variables are present, `Finding` becomes conditioned on `FPLevel` (see Figure 20). This can be explained because FPLevel is already the result of a classifier.

To see if classification results can be improved by learning models from data without incorporating FPLevel and MassLik in the learning process, but enforcing this



(contrast) (all three)

Figure 24: Learned structures using greedy search.

structure in the resulting structures, the greedy search algorithm is modified. This section describes the results of these learning using this updated algorithm.

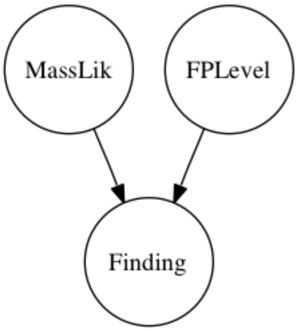


Figure 25: Finding is conditioned on FPLevel and MassLik

5.5.1 Adding features afterwards

The greedy search algorithm is modified, so that the learning process does not use MassLik and FPLevel. Instead, these variables are added as parent of Finding to the resulting structure. These resulting structures are shown in Figure 26 and the scores are presented in Table 9.

When the AUC scores are compared with the scores of the models that were learned using GS with MassLik and FPLevel included, we see a minor decrease of classification performance of 0.0175 for CC and 0,0322 for MLO.

	Bayesian score ($\times 10^4$)	AUC
CC	-9.5819	0.8314
MLO	-9.9670	0.7977

Table 9: Results with FPLevel and MassLik being added afterwards

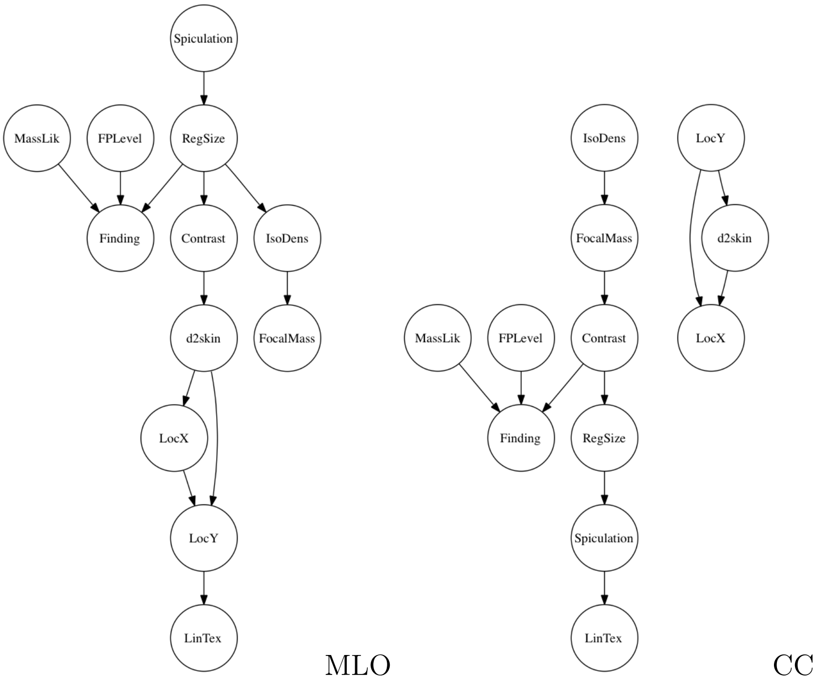


Figure 26: DAG structures learned without FPLevel and MassLik and added afterwards

5.5.2 Adding during scoring step

An other way to modify the greedy search algorithm is by including FPLevel and MassLik only in the scoring step of the algorithm.

The greedy search algorithm starts with a initial network structure G, which consist of the nodes `{Finding, LocX, LocY, d2skin, Contrast, Isodense, Spiculation, FocalMass, LinTexN, RegSize}` without arcs. For each step, it defines a set **NG** of neighborhood graphs. A copy of this set **NG**′ is made and each DAG in **NG**′ is modified: FPLevel and MassLik are added to the variables and used to condition on Finding (see Figure 25). For each modified DAG, the score is computed. The (modified) graph with the highest score is selected and its (unmodified) original version is used for the next iteration. The search is stopped when there is no neighborhood network graph with a higher score than the current structure. The resulting network structure is modified in the same way as before.

The parameters are learned for the resulting structure.

	Bayesian score ($\times 10^4$)	AUC
CC	-9.5461	0.8495
MLO	-9.7830	0.8179

Table 10: Results using modified GS

When we compare the results, we see a slightly higher score for CC (0.0006) and a slightly lower score for MLO (0.002).

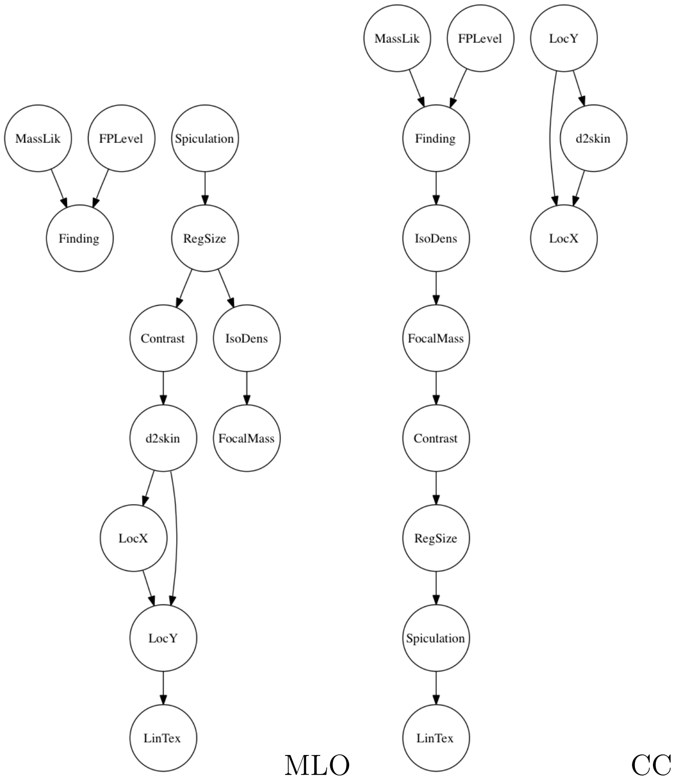


Figure 27: DAG structures learned with modified GS algorithm

6 Conclusions and discussion

In the introduction of this thesis, the goals of this research have been set. This section evaluates the research.

The purpose of this study was to investigate:

1. to what extent structure and parameter learning techniques can be used in breast cancer research;
2. if the correlation of certain variables in the dataset can be observed in the learned models;
3. the possibility of improving classification performance by combining data from different mammographic views;

Looking at these goals, a number of observations can be made.

It is possible to use structure and parameter learning techniques to learn Bayesian networks that perform reasonable using data from breast cancer screening programmes. Different algorithms can be used and classification performance of the resulting networks does not differ much. The simple Naive Bayes classifier still outperforms all learned classifiers. A remarkable observation is that the Naive Bayes classifier performs better than the Tree Augmented Naive Bayes classifier.

Including the variables FPLevel and MassLik greatly improves the performance of a classifier. In all cases where `FPLevel` and `MassLik` are used as variables, `Finding` becomes conditioned on `FPLevel`. This was expected, since the FPLevel feature is the result of another classifier, making the resulting classifier a sort of a second-order classifier and not very useful.

When `FPLevel` and `MassLik` are absent, `Finding` is conditioned on `Contrast` or `RegSize`. Without these variables, classifiers — in particular the Naive Bayes classifier — still perform reasonable. This means that the other variables do contain useful information about the presence of breast cancer. The choice for the features in the dataset, and how they are calculated, is affected by the fact that they are used in the classifier that calculates FPLevel and MassLik. This makes it harder to learn Bayesian classifiers that show a good performance, and certain features, and the way these features are calculated, could have been determined in another way which would have affected the results.

The results for using data from the CC-view are slightly better than for the MLO-view. This was expected, because the MLO-view is harder to interpret.

Two modifications to the greedy search algorithm have been proposed to see if classification results can be improved by learning models from data without incorporating FPLevel and MassLik in the learning process, but enforcing them in the resulting network structures. The performance of these modified versions of the greedy search algorithm is not significantly different than using FPLevel as a classifier.

Future research

In this section, some suggestions for future research on this topic are presented.

When looking at the goal of improving classifier performance, it is useful to include other features in the dataset. The dataset used in this research concentrated on the detection of masses. Features on microcalcifications (see Section 2.6.6) are used by radiologists when screening mammograms. From previous research it is known that microcalcifications are usually quite easily detected during the reading, and some automatic systems already show good performance on detecting them. Freer and Ulissey concluded in [15] that "the ability to detect clustered microcalcifications with a CAD system produced the most profound effect on the performance".

In Section 2.2, a number of risk factors that increase the probability of developing breast cancer are mentioned. These risk factors can not be observed from mammograms, but can be translated into features that provide relevant information and which can be useful to use for classification purposes. These risk factors include age, presence of certain genes, previous incidence of breast cancer and dietary and other lifestyle factors such as alcohol consumption. The inclusion of these 'risk factor features' could possibly lead to a better classification.

In the dataset are 81 features available, which are chosen and calculated for optimum performance of the classifier that calculates FPLevel. In this research, a subset of 11 features that are expected to contribute most to the detection of cancer have been selected to learn network structures from the data. The influence of other features on resulting structures and classification performance is an interesting topic for future research.

In this research, I have made use of the discretization methods that existed in the software libraries that were used for this research (see section 5.1.1). Other software libraries or tools, such as, for example, the Waikato Environment for Knowledge Analysis (WEKA) machine learning suite, provide more comprehensive libraries for discretization and other preprocessing steps of the data. An in-depth study exploring the possibilities and their influences is an interesting subject for future work.

References

[1] S. Acid, L.M. de Campos, and J.G. Castellano. *Learning Bayesian Network Classifiers: Searching in a Space of Partially Directed Acyclic Graphs.* Machine Learning Volume 59, Issue 3 (June 2005), Pages: 213-235, 2005.

[2] American College of Radiology (ACR). *Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas).* Reston, American College of Radiology, 2003.

[3] B-SCREEN project information on Bricks (Basic Research in Informatics for Creating the Knowledge Society) website. (2009). `http://www.bsik-bricks.nl/projects/is8.shtml` Retrieved 22 februari 2008.

[4] Casscells, W., Schoenberger, A., and Graboys, T. (1978). *Interpretation by physicians of clinical laboratory results.* New England Journal of Medicine, 299, 999-1001.

[5] S. Caulkin, S. Astley, J. Asquith, and C. Boggis. *Sites of occurrence of malignancies in mammograms.* In N. Karssemeijer, M. Thijssen, J. Hendriks, and L. Erning, editors, Digital Mammography Nijmegen, pages 279âĂŞ282. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[6] CBO Quality institute for healthcare. `http://www.cbo.nl/`. Retrieved April 8, 2008.

[7] D. M. Chickering. (2003) *Optimal structure identification with greedy search.* The Journal of Machine Learning Research, Volume 3, Pages: 507 - 554, 2003.

[8] Cooper, Gregory F., Herskovits, Edward. (1992). *A Bayesian Method for the Induction of Probabilistic Networks from Data.* Machine Learning, 9, Pages 309-347 (1992).

[9] van Engeland, S. (2006). *Detection of mass lesions in mammograms by using multiple views.* PhD thesis, Radboud University Nijmegen.

[10] Fawcett, T., *ROC graphs : Notes and practical considerations for researchers*, Technical report, HP Laboratories, April 2004.

[11] Ferreira, Nivea., Velikova, Marina., Lucas, Peter. *Bayesian Network Variations on Multi-View Breast Cancer Modelling.* Radboud University Nijmegen.

[12] N. de Carvalho Ferreira, M. Velikova. *Modelling Breast Cancer*. Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008.

[13] N. de Carvalho Ferreira, M. Velikova. *Overview of the breast cancer domain (internal document).* Radboud University Nijmegen, August 2007.

[14] Ildikó Flesch and Peter Lucas. *Markov Equivalence in Bayesian Networks* Institute for Computing and Information Sciences, University of Nijmegen

[15] T. Freer and M. Ulissey. (2001) *Screening mammography with computer-aided detection: Prospective study of 12860 patients in a community breast center.* Radiology 220:781-786, 2001.

[16] Geiger, D. (1992). *An entropy-based learning algorithm of Bayesian conditional trees.* Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference (UAI-1992), p. 92-97, San Mateo, CA: Morgan Kaufmann Publishers.

[17] Heckerman, David (1995). *A Tutorial on Learning With Bayesian Networks.* MSR-TR-95-06. Microsoft Research.

[18] Finn V. Jensen (2001). *Bayesian Networks and Decision Graphs*, Springer, 2001, ISBN 978-0387952598.

[19] Kevin B. Korb., Ann E. Nicholson. *Bayesian Artificial Intelligence.* ISBN 1-58488-387-1. Chapman & Hall/CRC press.

[20] Philippe Leray, Olivier Francois (2004). *BNT Structure Learning Package: Documentation and Experiments.* Technical Report, Laboratoire PSI - INSA Rouen.

[21] Murphy, K. (1997-2002). *Bayesian Network Toolbox for Matlab.*

[22] Nationaal Kompas Volksgezondheid (2008). *Borstkankerpreventie.* Version 3.16. `http://www.rivm.nl/vtv/object\class/kom\prevborstkanker.html` Retrieved 18 december 2008.

[23] Neapolitan, Richard E. (2003). *Learning Bayesian Networks.* ISBN 0130125342. Prentice Hall.

[24] Nyström L., Andersson I, Bjurstam N, Frisell J, NordenskjÃűld B, Rutqvist LE. (2002) *Long-term effects of mammography screening: updated overview of the Swedish randomised trails.* Lancet 2002; 359: 909-919.

[25] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo: Morgan Kaufmann, 1988.

[26] Pearl, J.*Bayesian Networks.* Cognitive Systems Laboratory, Computer Science Department, University of California. `ftp://ftp.cs.ucla.edu/pub/stat_ser/R246.ps` Retrieved 1 december 2009.

[27] Robinson, R. W. (1977). *Counting unlabeled acyclic digraphs.* C. H. C. Little, Ed., Combinatorial Mathematics V, volume 622 of Lecture Notes in Mathematics, p. 28-43, Berlin: Springer.

[28] Samulski, M. R. M. (2006). *Classification of Breast Lesions in Digital Mammograms.* Master thesis, Radboud University Nijmegen.

[29] M. Samulski, N. Karssemeijer, P. Lucas M.D., P. Groot, *Classification of mammographic masses using support vector machines and Bayesian networks*, Proceedings of SPIE, Vol. 6514, Medical Imaging 2007: Computer-Aided Diagnosis, 2007.

[30] M. Singh and M. Valtorta (1995). *Construction of Bayesian Network Structures From Data: A Brief Survey and an Efficient Algorithm* International Journal of Approximate Reasoning 1995; 12:111-131, 1995.

[31] Tabár L, Fagerberg CJ, Gad A, Baldetorp L, Holmberg LH, Gröntoft O, Ljungquist U, Lundström B, Mánson JC, Eklund G, et al. (1985). *Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare.* Lancet. 1985 Apr 13;1(8433):829-32.

[32] S. Timp, S. van Engeland, and N. Karssemeijer. *A regional registration method to find corresponding mass lesions in temporal mammograms.* Medical Physics, 32(8):2629âĂŞ2638, 2005.

[33] M. Velikova, M. Samulski, N.Karssemeijer, P. Lucas, *Toward expert knowledge representation for automatic breast cancer detection*, In Proceedings of the 13th biennial International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA), LNAI 5253, pp. 333-344, 2008.

[34] M. Velikova, P. Lucas, N. de Carvalho Ferreira, M. Samulski, N.Karssemeijer, *A decision support system for breast cancer detection in screening programs*, In Proceedings of the 18th biennial European Conference on Artificial Intelligence (ECAI), Vol. 178, 2008.

[35] M. Velikova, N. de Carvalho Ferreira, P. Lucas, *Bayesian network decomposition for modeling breast cancer detection*, In Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME) 2007, LNA 4594, pp. 346-350, 2007.

[36] World Health Organization (2009), *World Health Organization Cancer Fact Sheet.* `http://www.who.int/mediacentre/factsheets/fs297/en/index.html` Retrieved at January 17, 2009.

A Discretization

	TAN		GS	
Bins	AUC	Score ($\times 10^5$)	AUC	Score ($\times 10^5$)
2	0.7259	-0.3687	0.7769	-0.3586
3	0.7618	-0.5801	0.7955	-0.5669
4	0.7678	-0.7133	0.7921	-0.7002
5	0.7857	-0.8252	0.8037	-0.8112
6	0.7777	-0.9121	0.8122	-0.8954
7	0.8033	-0.9787	0.8299	-0.9602
8	0.7574	-1.0561	0.8151	-1.0424
9	0.7519	-1.1056	0.8287	-1.0886
10	0.7469	-1.1612	0.8365	-1.1411
11	0.7224	-1.1990	0.8289	-1.1756
12	0.7115	-1.2513	0.8196	-1.2213
13	0.6602	-1.2984	0.8168	-1.2604
14	0.6416	-1.3312	0.8198	-1.2886
15	0.6403	-1.3896	0.8146	-1.3369
16	0.6370	-1.4262	0.8333	-1.3665
17	0.5939	-1.4752	0.8280	-1.3985
18	0.6000	-1.5036	0.8278	-1.4177
19	0.5934	-1.5252	0.8217	-1.4301
20	0.5844	-1.5682	0.8218	-1.4561

Table 11: Results for different discretization parameters

B TAN structures for 5 variables

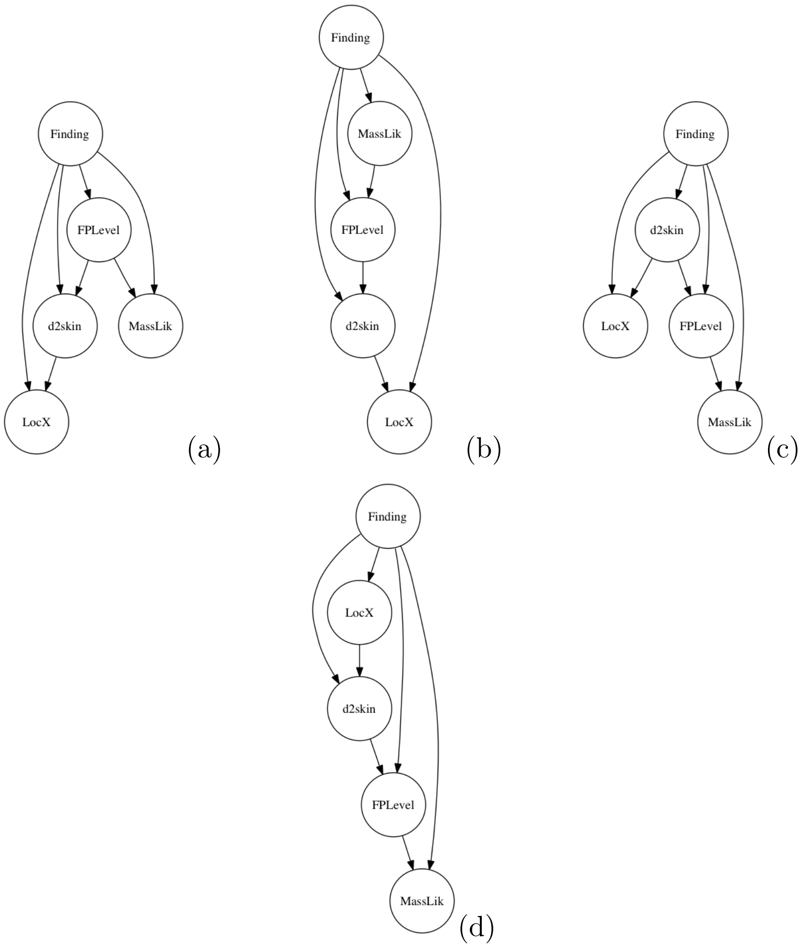


Figure 28: Equivalent tree augmented networks with different root nodes for the tree part of the DAG: (a) FPLevel; (b) MassLik; (c) d2skin; (d) LocX

C Network structures

C.1 Naive Bayes

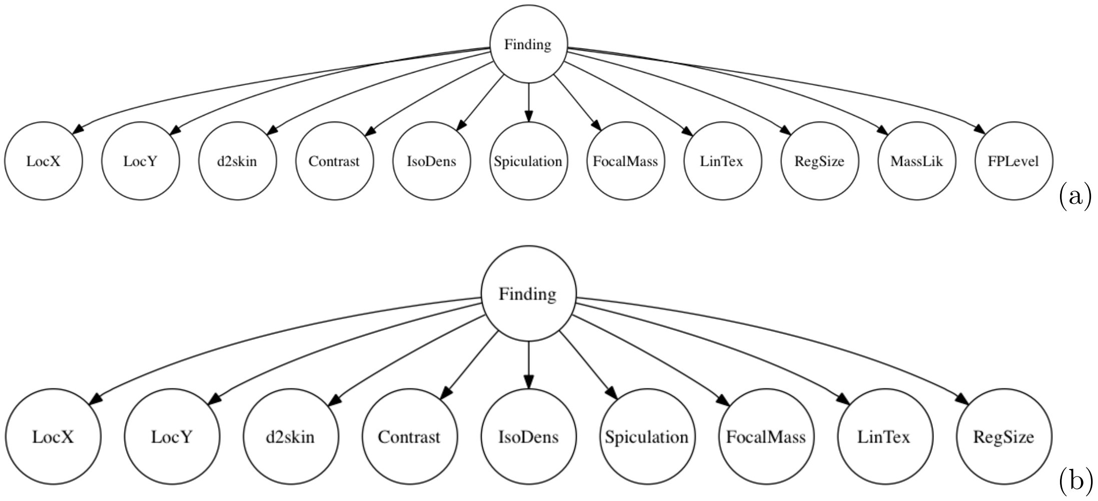
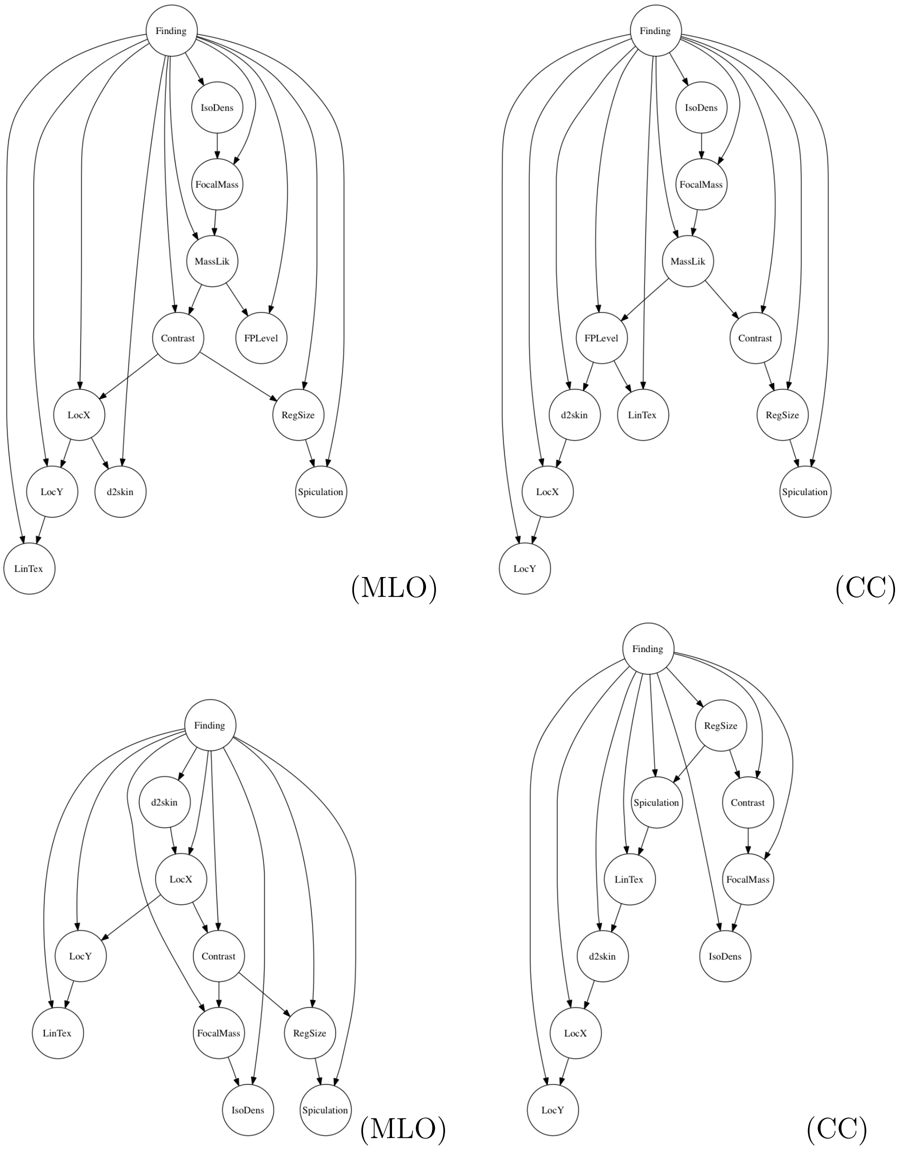


Figure 29: Naive Bayes structures.

C.2 Tree augmented networks



(MLO) (CC)

(MLO) (CC)

Figure 30: Tree augmented networks.

C.3 Structures learned using K2

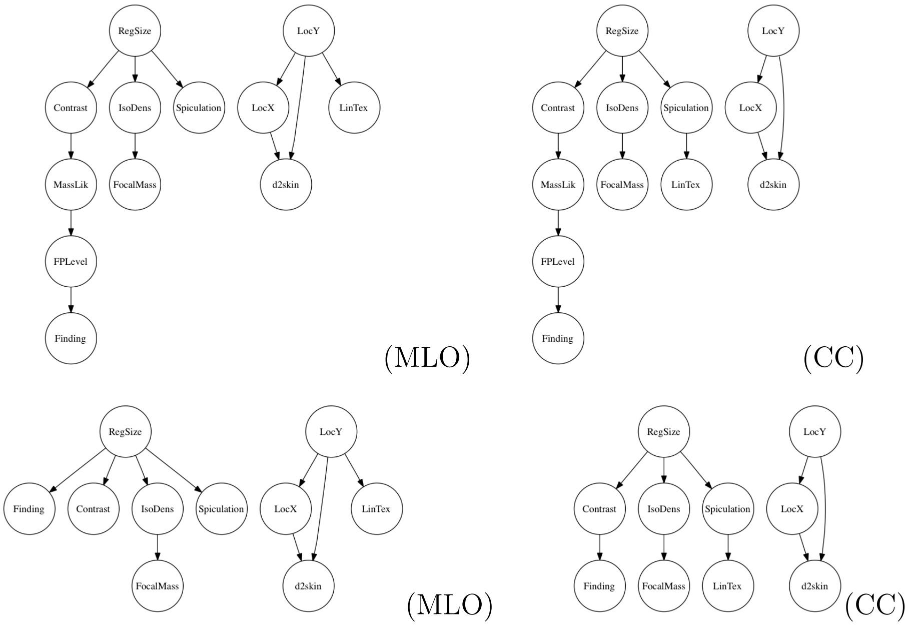
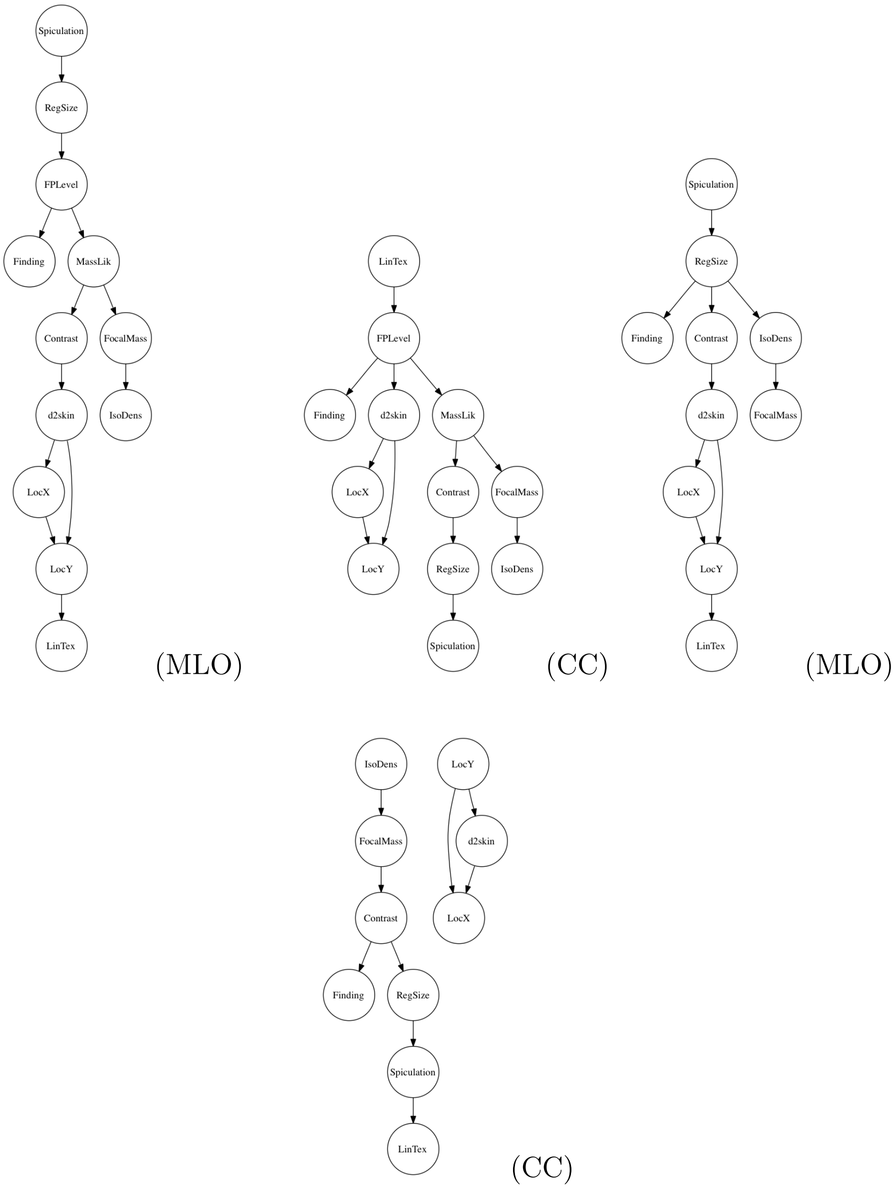


Figure 31: Structures learned using K2.

C.4 Structures learned using greedy search

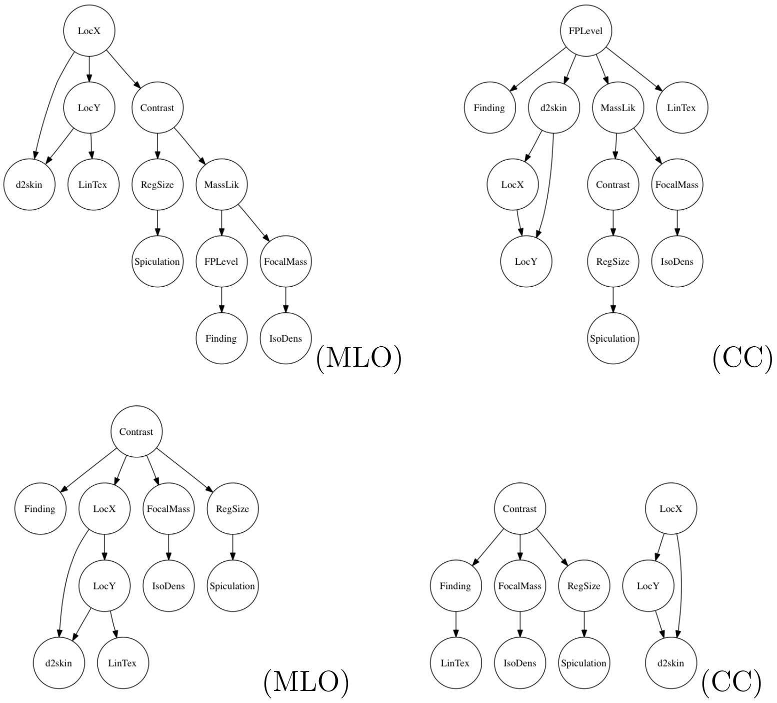


(MLO) (CC) (MLO)

(CC)

Figure 32: Structures learned using greedy search.

C.5 Structures learned using greedy equivalence search



(MLO) (CC)

(MLO) (CC)

Figure 33: Structures learned using greedy equivalence search.