Classification of mammographic masses using support vector machines and Bayesian networks

Maurice Samulski[a], Nico Karssemeijer[a], Peter Lucas[b], and Perry Groot[b]

[a]Department of Radiology, Radboud University Medical Centre, Geert Grooteplein Zuid 18, 6525 GA Nijmegen, The Netherlands;
[b]Radboud University, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

ABSTRACT

In this paper, we compare two state-of-the-art classification techniques characterizing masses as either benign or malignant, using a dataset consisting of 271 cases (131 benign and 140 malignant), containing both a MLO and CC view. For suspect regions in a digitized mammogram, 12 out of 81 calculated image features have been selected for investigating the classification accuracy of support vector machines (SVMs) and Bayesian networks (BNs). Additional techniques for improving their performance were included in their comparison: the Manly transformation for achieving a normal distribution of image features and principal component analysis (PCA) for reducing our high-dimensional data. The performance of the classifiers were evaluated with Receiver Operating Characteristics (ROC) analysis. The classifiers were trained and tested using a k-fold cross-validation test method (k=10). It was found that the area under the ROC curve (A_z) of the BN increased significantly (p=0.0002) using the Manly transformation, from $A_z = 0.767$ to $A_z = 0.795$. The Manly transformation did not result in a significant change for SVMs. Also the difference between SVMs and BNs using the transformed dataset was not statistically significant (p=0.78). Applying PCA resulted in an improvement in classification accuracy of the naive Bayesian classifier, from $A_z = 0.767$ to $A_z = 0.786$. The difference in classification performance between BNs and SVMs after applying PCA was small and not statistically significant (p=0.11).

Keywords: Methods: classification and classifier design, pre-processing, Modalities: mammography, Diagnostic task: diagnosis

1. INTRODUCTION

Machine learning techniques to diagnose breast cancer is a very active research area. Several computer-aided diagnosis (CAD) systems have been developed to aid radiologists in mammographic interpretation. These CAD systems analyze mammographic abnormalities and classify lesions as either benign or malignant in order to assist the radiologist in the diagnostic decision making. Some of them are based on Bayesian networks learned on mammographic descriptions provided by radiologists[1] or on features extracted by image processing.[2] Another classification technique that is widely used for the diagnosis of breast tumors are support vector machines.[3–6] The theoretical advantage of SVMs is that by choosing a specific hyperplane among the many that can separate the data in the feature space, the problem of overfitting the training data is reduced. They are often able to characterize a large training set with a small subset of the training points. Also, SVMs allow us to choose features with arbitrary distributions, and we do not need to make any independence assumptions. The advantage of Bayesian networks is that statistical dependences and independences between features are represented explicitly, which facilitates the incorporation of background knowledge. In this study we compare both classification methods and use two techniques, namely dimension reduction by principal component analysis (PCA) and a transformation for achieving a normal distribution of image features, to further improve the accuracy rate of the classifiers. Recently, the combination of PCA and support vector machines (SVMs) has been used in medical imaging, where principal component analysis is applied to extracted image features and the results are used to train a SVM classifier, but not specifically for mammograms.[7]

Further author information: (Send correspondence to Maurice R.M. Samulski)
M.R.M. Samulski, e-mail: m.samulski@rad.umcn.nl, telephone: +31 24 3619811, fax: +31 24 3540866

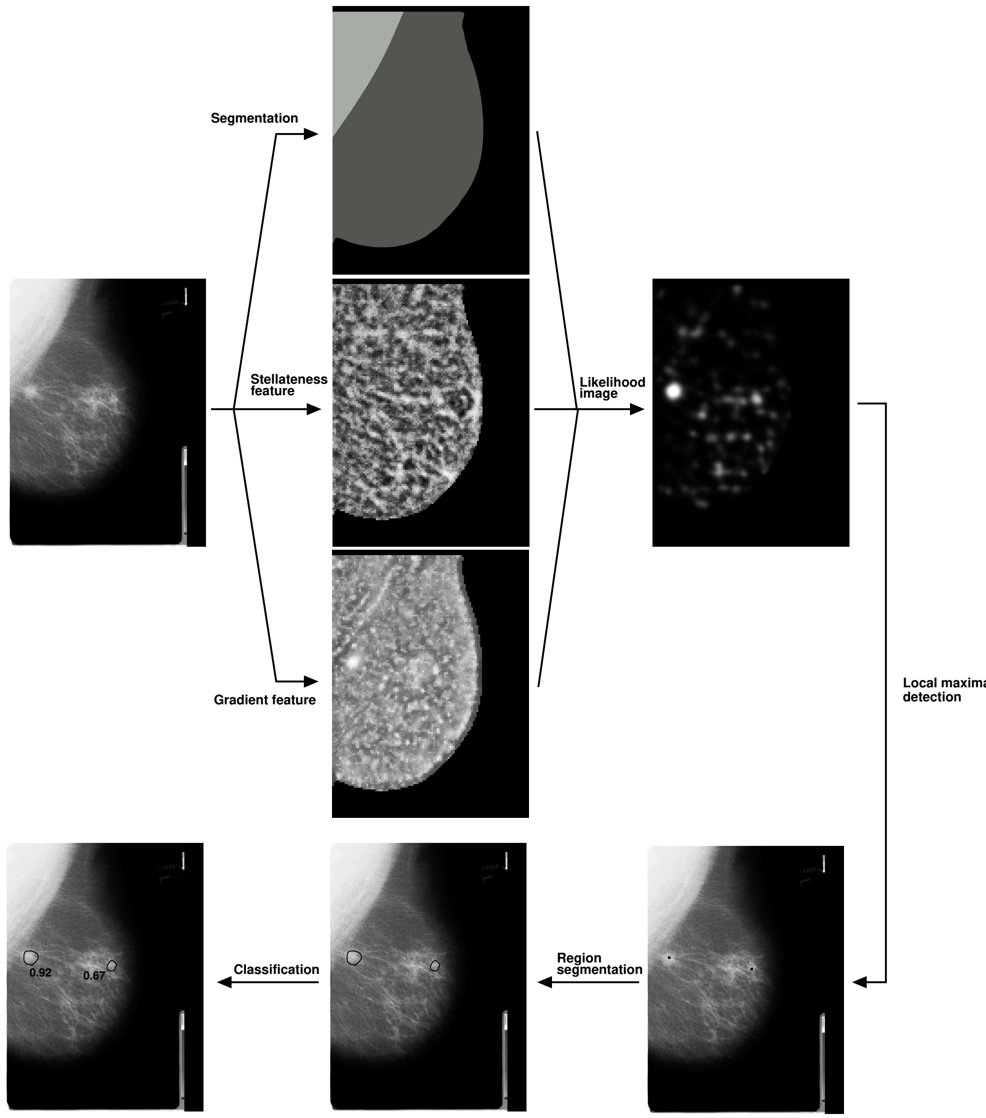

Figure 1: Schematic overview of the CAD scheme employed in this paper. First the mammogram is segmented into breast tissue, background tissue and the pectoral muscle. We then calculate at each location two stellateness features for the detection of spiculation and two gradient features for the detection of a focal mass. A neural network classifier combines these features into a likelihood of a mass at that location, resulting in a likelihood image. The most suspicious locations on the likelihood image (bright spots) are selected and used as seed points for the region segmentation. After that, features are calculated for each segmented region. Finally a second classifier combines these features into a malignancy score that represents the likelihood that the region is malignant.

2. MATERIALS AND METHODS

The digitized mammograms that were used in this study have been obtained from the Dutch Breast Cancer Screening Program. In this program two mammographic views of each breast were obtained in the initial screening: the medio-lateral oblique (MLO) view and a cranio caudal (CC) view. In this study 271 cases were used. Of these cases, 131 were benign and 140 were malignant. All cases had four-view mammograms.

To each image in the dataset a CAD scheme was applied that was previously developed in our group.[8] The CAD scheme consists of the following steps (Figure 1):

- Segmentation of the mammogram into breast tissue, pectoral muscle (if image is a MLO view), and background area
- Initial detection step resulting in a likelihood image and a number of suspect image locations (local maxima)
- Region segmentation, by dynamic programming, using the suspicious locations as seed points
- Final classification step to classify regions as true abnormalities and false positives.

These steps will be described in more detail in the following subsection.

2.1. Likelihood detection

Segmentation of the mammogram The first step of our CAD scheme is the segmentation of an image into breast tissue and background, using a skin line detection algorithm. Additionally, it finds the edge of the pectoralis muscle if the image is a MLO view.[9] After these steps, a thickness equalization algorithm is applied to enhance the periphery of the breast.[10] A similar algorithm is used to equalize background intensity in the pectoralis muscle, to avoid problems with detection of masses located on or near the pectoral boundary.

Initial mass detection step In this step we use a pixel-level method: for each pixel inside the breast area there are a small number of features calculated that represent presence of a central mass and the presence of spiculation.[11] A neural network classifies each pixel using these features and assigns a level of suspiciousness to it. The neural network is trained using pixels sampled inside and outside of a representative series of malignant masses. The result is an image in which pixel values represents the likelihood that a malignant mass or architectural distortion is present. This likelihood image is then slightly smoothed and a local maxima detection is performed. A local maximum is detected when the likelihood is above a certain threshold and no other nearby locations have a higher likelihood value. This results in a number of suspicious locations. Finally an algorithm searches for local maxima that are located closer than 8 mm together and remove multiple candidate locations to avoid multiple suspicious locations on the same lesion.

Region segmentation Each of the detected local maxima in the previous step are used as seed points for region segmentation, based on dynamic programming.[12]

Final classification For each segmented region, 81 features are calculated related to lesion size, roughness of the boundary, linear texture, location of the region, contour smoothness, contrast, and other image characteristics.

In the conducted experiments we used a subset of 12 features out of 81 features. They were selected using a k-nearest neighbor (KNN) algorithm and sequential forward procedure to find the most useful features for classifying lesions as benign or malignant. The procedure is described in detail in previous research.[4] We will give a short description of the used features in the following subsection.

2.2. Region features

Spiculation features Malignant mammographic densities are often surrounded by a radiating pattern of linear spicules. For the detection of these stellate patterns of straight lines directed toward the center pixel of a lesion, two features have been designed by Karssemeijer and te Brake.[11] The idea is that if an increase of pixels pointing to a given region is found then this region may be suspicious, especially if, viewed from that region, such an increase is found in many directions. The first feature *Stellateness 1* is a normalized measure for the fraction of pixels with a line orientation directed towards the center pixel. We call this set of pixels F. For calculating the second feature *Stellateness 2*, the circular neighborhood is divided into 24 angular sections. This feature measures to what extent the pixels in set F are uniformly distributed among all angular sections. Also the mean values of *Stellateness 1* and *Stellateness 2* inside the region are included in the subset.

Region Size Some features depend on the size of the lesion, like the contrast feature. Bigger lesions have a higher contrast than smaller lesions. This morphological feature captures this difference.

Compactness Compactness represents the roughness of an object's boundary relative to its area. This feature is included because benign masses often have a round or oval shape compared to a more irregular shape of malignant masses. Compactness (C) is defined as the ratio of the squared perimeter (P) to the area (A), i.e.,

$$C = \frac{P^2}{A}$$

The smallest value of compactness is $C = \frac{(2\pi r)^2}{\pi r^2} = 4\pi$ which is for a circle. For more complex shapes, the compactness becomes larger. In our dataset this feature is normalized by dividing the compactness by 4π.

Linear Texture Normal breast tissue often has different texture characteristics than tumor tissue. Therefore Karssemeijer and te Brake[11] have developed a texture feature that represents presence of linear structures inside the segmented region. Malignant lesions tend to have less linear structures than normal tissue or benign lesions.

Relative Location The relative location of a lesion is important since more malignancies develop in the upper outer quadrant[13] of the breast toward the armpit. Therefore, some features have been constructed that represent the relative location of a lesion using a new coordinate system.[14] This internal coordinate system is different for MLO and CC views. In MLO views the pectoral edge is used as the y-axis. The x-axis is determined by drawing a line perpendicular to the y-axis where the distance between the y-axis and the breast boundary is maximum. We assume that the end of this line is close to the nipple. In CC views the chest wall is used as y-axis. In this internal coordinate system we calculate the x- and y-location of the centre of the segmented region and normalize with the effective radius of the breast $r = \sqrt{\frac{A}{\pi}}$, where A is the size of the segmented breast area. In this way, positions of cancers in different mammograms can be compared.

Maximum Second Order Derivative Correlation This border feature indicates the smoothness of the contour and is especially useful to discriminate between benign and malignant lesions. Most benign lesions have a well-defined contour and the margins of these lesions are sharply confined with a sharp transition between the lesion and the surrounding tissue which indicates that there is no infiltration.[14]

Contrast Regions with high contrast or a higher intensity than other similar structures in the image are more likely to be a mass since tumor tissue on average absorbs more X-rays than fat and also slightly more than glandular tissue. The distance measure we used to indicate differences in contrast is the squared difference in intensity between the segmented region and its surround, divided by both standard deviations,

$$\frac{(\overline{Y}(R) - \overline{Y}(S))^2}{\sigma_Y(R) + \sigma_Y(S)}$$

where R is the set of pixels in the segmented region, S is the set of pixels in the surroundings of the segmented region. $\overline{Y}(X)$ is the mean grey level of the pixels in set X, and $\sigma_Y(X)$ is the grey level standard deviation of the pixels in set X.

Number of Calcifications The presence of clustered microcalcifications is one of the most important signs of cancer on a mammogram. They occur in about 90% of the non-invasive cancers. Therefore we include a feature representing the number of calcifications found in the segmented region.

2.3. Statistical analysis

For every feature the first four moments of the distribution of feature values in the dataset have been computed. These are shown in Table 1. The third moment, skewness, is a measure of the lack of symmetry. The skewness for a normal distribution is zero, and any near-symmetric data should have a skewness near zero. The fourth moment, also called kurtosis, is a measure of whether the data are peaked or flat relative to a normal distribution. The kurtosis for a standard normal distribution is three.

Combining the 12 features of the MLO views with the 12 features of the corresponding CC views gives a total of 24 features per case. The continuous output of the classifier is analyzed using ROC methodology, using the LABROC program[15] of Metz et al. The statistical significance of the difference between ROC curves was tested using the CLABROC program[16] of Metz et al. The classifiers were trained and tested using a k-fold

	Mean	Std dev	Min	Max	Skewness	Kurtosis
Benign (cases: 258)						
Stellateness 1	1.1256	0.1710	0.7800	2.1400	2.3002	13.4307
Stellateness 2	1.0241	0.1160	0.8300	2.1900	4.7815	44.8670
Stellateness 1 Mean	1.1189	0.1316	0.8600	1.5630	0.8565	3.6986
Stellateness 2 Mean	1.0215	0.0713	0.8380	1.2990	0.5482	3.6256
Region Size	0.4070	0.3915	0.0200	3.4510	3.0272	17.9799
Contrast	0.5502	0.2558	0.1260	2.0110	1.9986	9.8575
Compactness	1.2141	0.0906	1.0470	1.5600	0.9308	3.8448
Linear Texture	0.1750	0.1444	0.0130	1.0240	2.2365	10.1391
Relative Location X	0.6705	0.3024	-0.0670	1.5470	0.0470	2.7819
Relative Location Y	0.2160	0.4262	-0.9680	1.2990	-0.2289	2.4769
Max. 2nd order Drv Corr.	0.6800	0.1008	0.4520	0.9060	0.0436	2.3011
Number of Calcifications	0.7871	2.6723	0.0000	19.0000	3.8831	19.2635
Malignant (cases: 274)						
Stellateness 1	1.2273	0.1730	0.8200	1.7300	0.5060	3.0005
Stellateness 2	1.0827	0.0965	0.7900	1.3500	0.1468	2.8634
Stellateness 1 Mean	1.2357	0.1736	0.8290	1.7740	0.6844	3.1281
Stellateness 2 Mean	1.0868	0.0946	0.8530	1.4140	0.4533	3.0175
Region Size	0.4471	0.3272	0.0160	1.8040	1.2728	4.4259
Contrast	0.6272	0.2777	0.0110	1.5090	0.7688	3.2074
Compactness	1.2111	0.0983	1.0410	1.7080	1.5022	6.3482
Linear Texture	0.1578	0.1161	0.0040	0.9490	2.2258	11.5829
Relative Location X	0.6130	0.3046	-0.0710	1.3080	0.0140	2.3298
Relative Location Y	0.2080	0.4449	-0.9770	1.2180	-0.2483	2.7594
Max. 2nd order Drv Corr.	0.6354	0.0951	0.4040	0.9320	0.1608	2.9336
Number of Calcifications	2.0645	6.7471	0.0000	50.0000	4.4524	25.7707

Table 1: Statistics of benign and malignant cases in the used dataset

cross-validation test method (k=10), in which each of 10 different combinations of training and test data sets included 244 and 27 cases, respectively. For each test partition, the classification accuracy was evaluated as the area A_z under the ROC curve.

2.4. Classifiers

2.4.1. Naive Bayesian classifier

The naive Bayesian classifier (Figure 2) is a Bayesian network with a limited topology[17] applicable to learning tasks where each instance is described by a conjunction of feature values and a class value. To learn the Bayesian network a set of training examples has to be provided. Classification using this Bayes' probability model is done by picking the most probable hypothesis which is also known as the *maximum a posteriori*. The corresponding classifier function can be defined as follows:

$$C_{MAP} = \operatorname*{arg\,max}_{c_j \in C} \; P(c_j | f_1, f_2, \ldots, f_n) \tag{1}$$

where $\{f_1, f_2, \ldots, f_n\}$ is the set of feature values that decribe the new instance, and C_{MAP} is the most probable hypothesis. Using Bayes theorem, Equation 1 can be rewritten as follows:

$$\begin{aligned} C_{MAP} &= \operatorname*{arg\,max}_{c_j \in C} \; \frac{P(c_j)P(f_1, f_2, \ldots, f_n | c_j)}{P(f_1, f_2, \ldots, f_n)} \\ &= \operatorname*{arg\,max}_{c_j \in C} \; P(c_j)P(f_1, f_2, \ldots, f_n | c_j) \end{aligned} \tag{2}$$

Using training data the two terms $P(c_j)$ and $P(f_1, f_2, \ldots, f_n|c_j)$ have to be calculated. The *class prior probability* $P(c_j)$ can be easily estimated by counting the frequency of occurence of the class value c_j in the training data. However, estimating the different $P(f_1, f_2, \ldots, f_n|c_j)$ terms is difficult and is only possible if a huge set of training data is available. To dramatically simplify the classification task we can use the following simplifying assumption: each feature f_i is conditionally independent of every other feature f_j for $i \neq j$. This fairly strong assumption of independence leads to the name naive Bayes, with the assumption often being naive in that, by making this

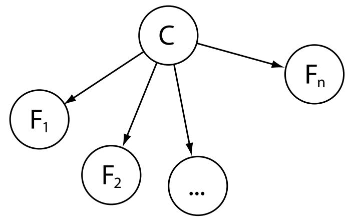

Figure 2: A graphical representation of a naive Bayesian classifier

assumption, the algorithm does not take into account dependencies that may exist. By using the conditionally independence assumptions we can express Equation 2 as:

$$C_{MAP} = \underset{c_j \in C}{\arg\max}\ P(c_j) \prod_{i=1}^{n} P(f_i|c_j) \tag{3}$$

The model in this form is much more manageable, since it factors into a so-called *class prior probability* $P(c_j)$ and independent probability distributions $P(f_i|c_j)$. These class conditional probabilities $P(f_i|c_j)$ can be calculated separately for each variable which reduces complexity enormously. Even with such strong simplifying assumptions, it does not seem to greatly affect the posterior probabilities, especially in regions near the decision boundaries which leaves the classification task unaffected. Some papers show that such naive Bayesian classifiers yield surprisingly powerful classifiers.[18]

2.4.2. Support vector machines

The SVM algorithm has been introduced by Cortes and Vapnik[3] for solving classification tasks and have been successfully applied in various areas of research. The basic idea of SVM is that it projects datapoints from a given two-class training set in a higher dimensional space and finds an optimal hyperplane. The optimal one is the one that separates the data with the maximal margin. SVMs identify the datapoints near the optimal separating hyperplane which are called support vectors. The distance between the separating hyperplane and the nearest of the positive and negative datapoints is called the margin of the SVM classifier. The separating hyperplane is defined as

$$D(x) = (w \cdot x) + b \tag{4}$$

where x is a vector of the dataset mapped to a high dimensional space, and w and b are parameters of the hyperplane that the SVM will estimate. The nearest datapoints to the maximum margin hyperplane lie on the planes

$$\begin{aligned} (w \cdot x) + b &= +1 \quad \textit{for } y = +1 \\ (w \cdot x) + b &= -1 \quad \textit{for } y = -1 \end{aligned} \tag{5}$$

where $y = +1$ for class ω_1 and $y = -1$ for class ω_2. The width of the margin is given by $m = \frac{2}{||w||}$. Computing w and x is then the problem of finding the minimum of a function with the following constraints:

$$\begin{aligned} \textit{minimize} \quad & m(w) = \frac{1}{2}(w \cdot w) \\ \textit{subject to constraints} \quad & y_i[w \cdot x_i + b] \geq 1 \end{aligned} \tag{6}$$

In its simplest form, a SVM attempts to find a linear separator, as shown in Figure 3. In practice however, there may be no good linear separator of the data. In that case, SVMs can project the dataset to a significant higher dimensional feature space to make the separation easier, using a kernel function to produce separators that are non-linear. Unfortunately there is no theory about deciding which kernel is the best.[19]

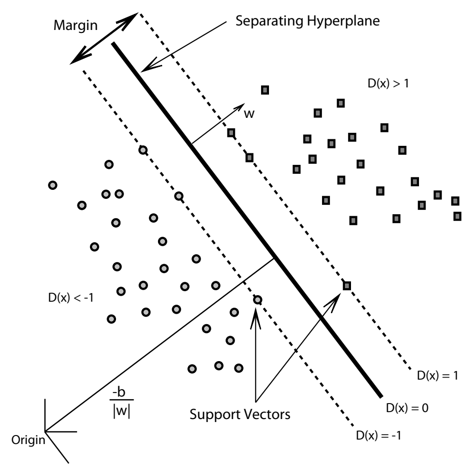

Figure 3: Linear separating hyperplanes for the separable case.

2.5. Preprocessing

2.5.1. Manly transformation

Many Bayesian learning algorithms that deal with continuous nodes, including the learning algorithms in Kevin Murphy's Bayesian Networks Toolbox,[20] are based on the assumption that the features are normally distributed. Unfortunately, most of the image features we use do not follow a normal distribution. We used Manly's exponential transformation to make the non-normal data resemble normal data by reducing skewness, which is a transformation from y to $y^{(\lambda)}$ with parameter λ. This transform is most effective if the probability distribution of a feature can be described as a function which contains powers, logarithms, or exponentials. The transform is given by:

$$y^{(\lambda)} = \begin{cases} \frac{e^{\lambda y}-1}{\lambda} & \text{if } \lambda \neq 0 \\ y & \text{if } \lambda = 0 \end{cases} \tag{7}$$

The assumption made by this transformation is that $y^{(\lambda)}$ follows a normal linear model with parameters β and σ^2 for some value of λ. Given a value of λ, we can estimate the linear model parameters β and σ^2 as usual, except that we work with the transformed variable $y^{(\lambda)}$ instead of y. To select an appropriate transformation we need to find the optimal value of λ using an optimization criteria. We used a technique based on the normal probability plot. The data is plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line if the data is normal distributed. Deviations of this straight line mean that the data is less normally distributed. We can use that property to plot the correlation coefficient of the normality plot against a range of λ's. The lambda resulting in the largest correlation coefficient is chosen.

2.5.2. Principal component analysis

One might think that the use of more features will automatically improve the classification power of the classifier. However the number of samples needed to train a classifier with a certain level of accuracy increases exponentially with the number of features. Therefore, we used principal component analysis[21] as a preprocessing technique to reduce the dimensionality of our dataset. The assumption made in PCA is that most of the information is carried in the variance of the features: the higher the variance in one dimension (feature), the more information is carried by that feature. The general idea is therefore to preserve the most variance in the data using the least number of dimensions. One of the major drawbacks of PCA is that it is an unsupervised algorithm, i.e., it does

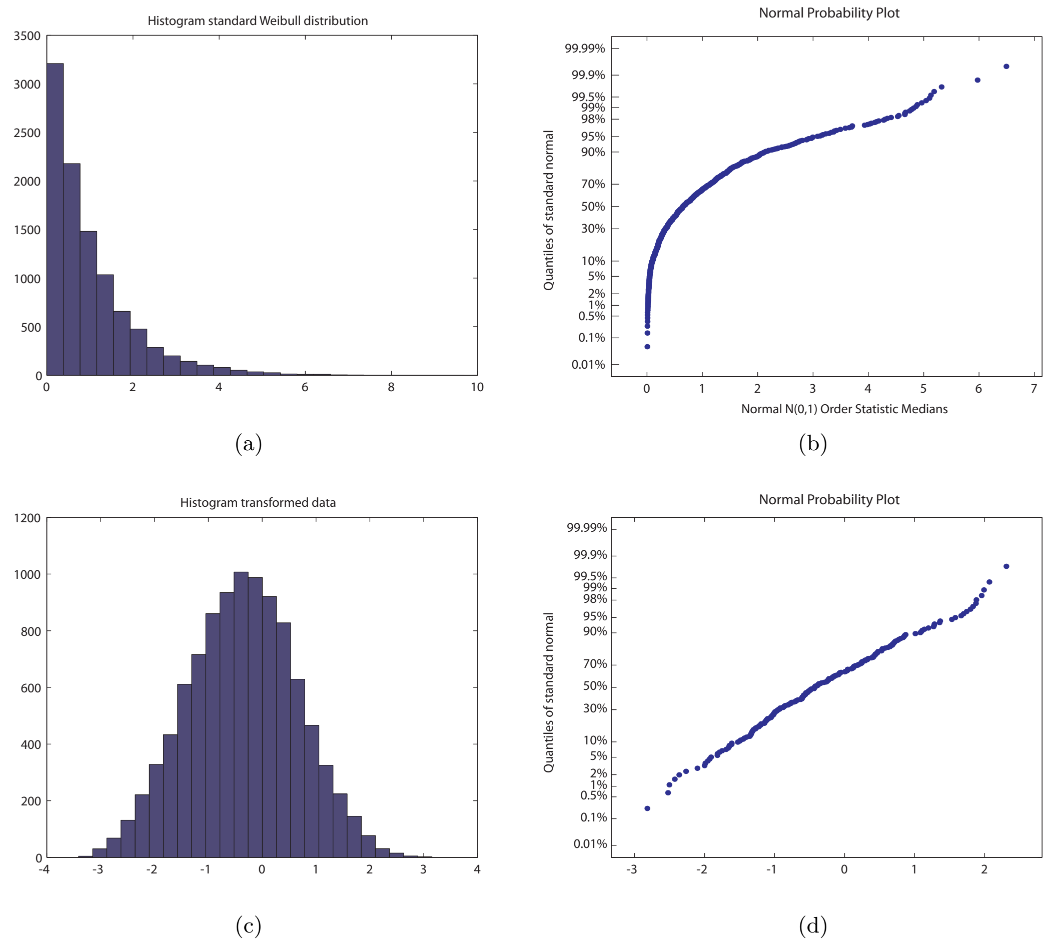

Figure 4: An example Manly transformation: (a) histogram of a feature that is Weibull distributed, (b) normality plot of the feature, (c) histogram of the transformed feature, and (d) normality plot of the transformed feature

not take the class label in account. It can therefore eliminate a dimension that is best for discriminating positive from negative cases.

3. RESULTS

The dataset we used contained a lot of features that were highly skewed and therefore did not follow a normal distribution. The learning algorithms in Murphy's BNT toolbox[20] for Bayesian networks with continuous nodes, assume that within each state of the class the observed continuous features follow a normal distribution. These continuous nodes have therefore two parameters per class, mean and variance, to represent the characteristics of the training data. We evaluated the classification performance of the naive Bayes classifier after applying the Manly transformation on the dataset. The *Stellateness Mean* and the *Maximum Second Order Derivative Correlation* features are approximately normal distributed in their original form and did not perform well when transformed. We chose therefore to not transform these features. Also the *Number of Calcifications* feature was not a useful candidate to transform, because of its discrete nature. Statistical information about the transformed dataset can be found in Table 2

The calculated area under the ROC curve (A_z value) of the Bayesian classifier without transforming the dataset was 0.767. After applying the Manly transformation it increased to 0.795, which is statistically significant

	Mean	Std dev	Min	Max	Skewness	Kurtosis
All cases (cases: 542)						
Stellateness 1	0.5102	0.0229	0.4351	0.5799	0.0000	3.1743
Stellateness 2	0.4077	0.0095	0.3739	0.4495	0.0000	3.7745
Stellateness 1 Mean	1.1790	0.1653	0.8290	1.7740	0.8638	3.5949
Stellateness 2 Mean	1.0551	0.0902	0.8380	1.4140	0.6548	3.4349
Region Size	0.2148	0.0878	0.0157	0.3706	0.0000	1.9075
Contrast	0.3599	0.0924	0.0109	0.5939	0.0000	2.7383
Compactness	0.2070	0.0002	0.2063	0.2076	0.0000	2.4822
Linear Texture	0.0931	0.0381	0.0040	0.1725	0.0000	2.3284
Relative Location X	0.6312	0.2974	-0.0711	1.5016	0.0000	2.5448
Relative Location Y	0.2404	0.4554	-0.8736	1.5173	0.0000	2.5943
Max. 2nd order Drv Corr.	0.6571	0.1004	0.4040	0.9320	0.1290	2.5924
Number of Calcifications	1.4446	5.2255	0.0000	50.0000	5.4429	39.8079
Benign (cases: 263)						
Stellateness 1	0.5029	0.0219	0.4351	0.5799	0.2717	4.1962
Stellateness 2	0.4047	0.0091	0.3806	0.4495	0.5072	5.8195
Stellateness 1 Mean	1.1189	0.1316	0.8600	1.5630	0.8565	3.6986
Stellateness 2 Mean	1.0215	0.0713	0.8380	1.2990	0.5482	3.6256
Region Size	0.2048	0.0882	0.0195	0.3706	0.1791	1.9413
Contrast	0.3463	0.0865	0.1140	0.5939	0.2547	2.8499
Compactness	0.2070	0.0002	0.2063	0.2075	-0.1018	2.4716
Linear Texture	0.0946	0.0396	0.0125	0.1725	-0.0221	2.2135
Relative Location X	0.6601	0.2946	-0.0671	1.5016	0.0159	2.7614
Relative Location Y	0.2437	0.4457	-0.8665	1.5173	-0.0159	2.4540
Max. 2nd order Drv Corr.	0.6800	0.1008	0.4520	0.9060	0.0436	2.3011
Number of Calcifications	0.7871	2.6723	0.0000	19.0000	3.8831	19.2635
Malignant (cases: 279)						
Stellateness 1	0.5171	0.0217	0.4456	0.5636	-0.2453	2.9714
Stellateness 2	0.4106	0.0090	0.3739	0.4301	-0.4677	3.4270
Stellateness 1 Mean	1.2357	0.1736	0.8290	1.7740	0.6844	3.1281
Stellateness 2 Mean	1.0868	0.0946	0.8530	1.4140	0.4533	3.0175
Region Size	0.2242	0.0864	0.0157	0.3678	-0.1658	1.9722
Contrast	0.3728	0.0960	0.0109	0.5640	-0.2511	2.8418
Compactness	0.2070	0.0002	0.2063	0.2076	0.0943	2.5107
Linear Texture	0.0917	0.0365	0.0040	0.1722	0.0067	2.4444
Relative Location X	0.6040	0.2975	-0.0711	1.2754	-0.0094	2.3235
Relative Location Y	0.2372	0.4643	-0.8736	1.4087	0.0149	2.6997
Max. 2nd order Drv Corr.	0.6354	0.0951	0.4040	0.9320	0.1608	2.9336
Number of Calcifications	2.0645	6.7471	0.0000	50.0000	4.4524	25.7707

Table 2: Statistics of benign and malignant cases after transformation.

(p=0.0002). For the SVM classifier, the Manly transformation had no noticeable effect on the performance. Comparing the performance between BNs and SVMs using the transformed dataset showed that the difference was not statistically significant (p=0.78).

Additionally, we evaluated the classification performance of the naive Bayesian and SVM classifier after applying dimensionality reduction on our dataset. Figure 5 shows the classification performance of the naive Bayesian classifier, where horizontally the number of principal components is plotted and vertically the area under the ROC curve. The principal component vectors were calculated using the training set only. These principal component vectors are then used to transform both the training and test set. The best result was obtained with 14 principal components. The performance remained almost constant when adding more dimensions. With SVMs the best result was obtained with only 6 principal components and decreased gradually if more components were added which is shown in Figure 6. The difference in classification performance between BNs and SVMs was statistically insignificant (p=0.11) when we used the optimal number of principal components for the classifier. In an additional experiment we trained a SVM on all the available features (81 per view). This led to the classification results shown in Figure 7. The maximum performance was reached in 10 components ($A_z = 0.811$) but this was not significantly higher than the maximum performance obtained in the experiment with the subset of the 12 most important features ($A_z = 0.793$).

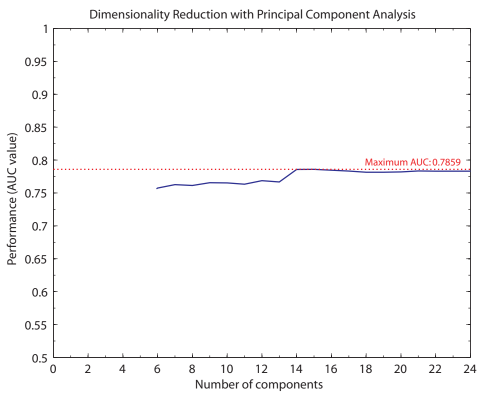

Figure 5: Case based performance naive Bayes classifier after dimensionality reduction with PCA, averaged over 5 runs.

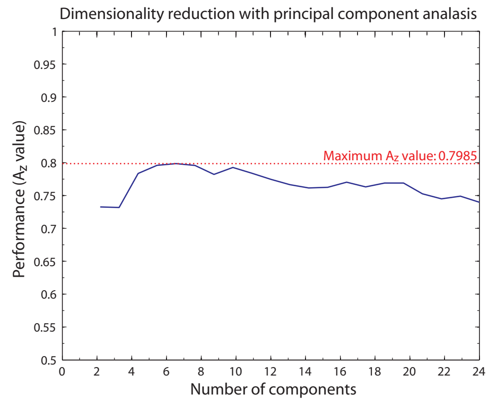

Figure 6: Case based performance SVM classifier with radial kernel function after dimensionality reduction with PCA, averaged over 5 runs.

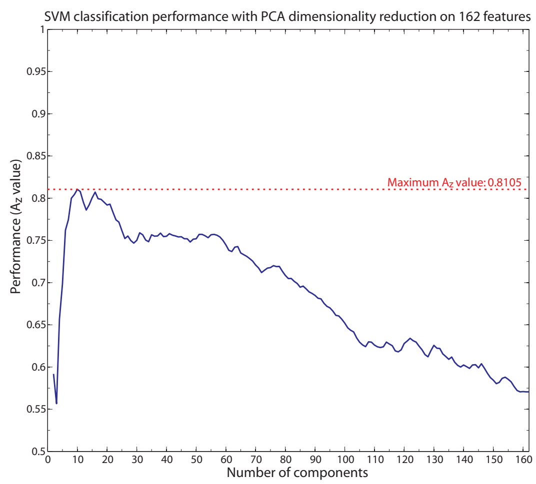

Figure 7: Case based performance SVM classifier with radial kernel function after dimensionality reduction of all features (81 per view) with PCA, averaged over 5 runs.

4. CONCLUSION

We performed a study to compare two state-of-the-art classification techniques characterizing masses as either benign or malignant. We evaluated the effectiveness of dimension reduction and normal distribution transformation in improving the classification accuracy. The Manly transformation method significantly improved classification accuracy of the naive Bayesian classifier. We believe that this is due the fact that, by transforming the distribution of the non-normal data to a distribution closer to normal, the assumptions of the naive Bayesian classifier are violated less. We also found that this transformation does not work for all data, i.e., transforming features that were already approximately normal within their class. We believe that by selecting one gamma for Manly's transformation, without looking to the class label, can negatively effect the binormal distribution (i.e., two normal distributions: one for benign and another for malignant cases) of the *Stellateness Mean* features. For the SVM classifier, the data does not need to be normally distributed which explains why this transformation did not have effect on the performance of the SVM classifier. After transformation, the difference in performance of the SVM classifier and the naive Bayesian classifier was not statistically significant. Bayesian networks allow incorporating background knowledge, which may be exploited to improve their performance in the future. Despite the major drawback of principal component analysis, i.e., it can eliminate a dimension that is good for discriminating positive cases from negative cases, this unsupervised dimension reduction algorithm improved the classification accuracy of both classifiers. The performance of the two classifiers after applying PCA was very similar, with no statistical differences in the area under the ROC curve.

REFERENCES

1. E. Burnside, D. Rubin, and R. Schachter, "A Bayesian network for mammography," in *Proceedings of AMIA Annual Symposium*, pp. 106–110, 2000.
2. X. Wang, B. Zheng, W. Good, J. King, and Y. Chang, "Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network," *International Journal of Medical Informatics* **54**, pp. 115–126, May 1999.
3. C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning* **20**(3), pp. 273–297, 1995.
4. S. Timp, *Analysis of Temporal Mammogram Pairs to Detect and Characterise Mass Lesions.* PhD in medical sciences, Radboud University Nijmegen, 2006. ISBN 9090205500.
5. T. Nattkemper, B. Arnrich, O. Lichte, W. Timm, A. Degenhard, L. Pointon, C. Hayes, and M. Leach, "Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods," *Artificial Intelligence Medical* **34**, pp. 129–139, June 2004.
6. M. Mavroforakis, H. Georgiou, N. Dimitropoulos, D. Cavouras, and S. Theodoridis, "Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines," *European Journal of Radiology* **54**, pp. 80–89, April 2005.
7. S. Li, T. Fevens, and A. Krzyzak, "Automatic clinical image segmentation using pathological modeling, PCA and SVM," *Engineering Applications of Artificial Intelligence* **19**, pp. 403–410, June 2006.
8. G. te Brake, N. Karssemeijer, and J. Hendriks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms[1]," *Physics in Medicine and Biology* **45**(10), pp. 2843–2857, 2000.
9. N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Physics in Medicine and Biology* **43**(2), pp. 365–378, 1998.
10. P. Snoeren and N. Karssemeijer, "Thickness correction of mammographic images by anisotropic filtering and interpolation of dense tissue," in *Medical Imaging 2005: Image Processing. Edited by Fitzpatrick, J. Michael; Reinhardt, Joseph M. Proceedings of the SPIE, Volume 5747, pp. 1521-1527 (2005).*, J. M. Fitzpatrick and J. M. Reinhardt, eds., pp. 1521–1527, April.
11. N. Karssemeijer and G. te Brake, "Detection of stellate distortions in mammograms," *IEEE Transactions on Medical Imaging* **15**(5), pp. 611–619, 1996.
12. S. Timp and N. Karssemeijer, "A new 2d segmentation method based on dynamic programming applied to computer aided detection in mammography," *Medical Physics* (5), pp. 958–971, 2004.
13. S. Caulkin, S. Astley, J. Asquith, and C. Boggis, *Sites of occurrence of malignancies in mammograms*, vol. 13 of *Digital Mammography Nijmegen*, pp. 279–282. Kluwer Academic Publishers, Dordrecht, The Netherlands, first ed., December 1998. Editors: N. Karssemeijer and M. Thijssen and J. Hendriks and L. Erning.
14. C. Varela, S. Timp, and N. Karssemeijer, "Use of border information in the classification of mammographic masses," *Physics in Medicine and Biology* , January 2006.
15. C. Metz, B. Herman, and J. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statistics in Medicine* **17**, pp. 1033–1053, 1998.
16. C. Metz, P. Wang, and H. Kronman, "A new approach for testing the significance for differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging*, F. Deconinck, ed., pp. 432–445, The Hague: Nijhoff, 1984.
17. R. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, NJ, 2003.
18. P. Domingos and M. J. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning* **29**(2-3), pp. 103–130, 1997.
19. A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing* **14**(3), pp. 199–222, 2004.
20. K. Murphy, "The bayes net toolbox for Matlab," 2001.
21. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley, second ed., 2001.