A decision support system for breast cancer detection in screening programs

Marina Velikova[1] and **Peter J.F. Lucas**[2] and **Nivea Ferreira**[2] and **Maurice Samulski**[1] and **Nico Karssemeijer**[1]

Abstract. The goal of breast cancer screening programs is to detect cancers at an early (preclinical) stage, by using periodic mammographic examinations in asymptomatic women. In evaluating cases, mammographers insist on reading multiple images (at least two) of each breast as a cancerous lesion tends to be observed in different breast projections (views). Most computer-aided detection (CAD) systems, on the other hand, only analyze single views independently, and thus fail to account for the interaction between the views. In this paper, we propose a Bayesian framework for exploiting multi-view dependencies between the suspected regions detected by a single-view CAD system. The results from experiments with real-life data show that our approach outperforms the single-view CAD system in distinguishing between normal and abnormal cases. Such a system can support screening radiologists to improve the evaluation of breast cancer cases.

1 INTRODUCTION

Breast cancer is the most common form of cancer among women world-wide and its early detection can improve the chances of successful treatment and recovery ([1]). Therefore, many countries have introduced screening programs for the early diagnosis of breast cancer in asymptomatic women.

A screening mammographic examination usually consists of four images, corresponding to each breast scanned in two views: mediolateral-oblique (MLO) and craniocaudal (CC) (Figure 1). The MLO projection is taken under 45° angle and shows part of the pectoral muscle. The CC projection is a top-down view of the breast. In reading mammograms, radiologists judge for the presence of a lesion by comparing both views and breasts. The general rule is that a lesion is to be observed in both views.

Most computer-aided detection (CAD) systems, on the other hand, are only able to analyze each view independently. Hence, the correlations in the lesion characteristics are ignored and the breast cancer detection can be obscured due to the lack of consistency in lesion marking. This limits the usability and the trust in the performance of such systems.

In this paper, we explore multi-view dependencies to improve the breast cancer detection rate at a patient level. We develop a Bayesian network model that combines the information from all the regions detected by a single-view CAD system in MLO and CC to obtain a single measure for suspiciousness of a case.

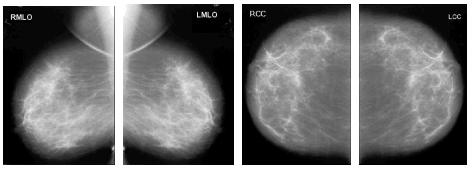

Figure 1. MLO and CC projections of a right and left breast

To get the reader acquainted with the terminology used in the domain of breast cancer and throughout this paper, we next introduce a number of definitions of terms. By *lesion* we refer to a physical cancerous object detected in a patient. We call a contoured area on a mammogram a *region*. A region can be true positive (for example, a lesion marked manually by a radiologist or detected automatically by a CAD system as being suspicious) or false positive. A region detected by a CAD system is described by a number of continuous (real-valued) *features* (e.g., size, location, contrast). By *link* we denote matching (established correspondence) between two regions in MLO and CC views, respectively. The term *case* refers to a patient who has undergone a mammographic exam.

The remainder of the paper is organized as follows. In the next section we briefly review previous research in multi-view breast cancer detection. In Section 3 we introduce basic definitions related to Bayesian networks and then we describe a general Bayesian network framework for multi-view detection. The proposed approach is evaluated on an application of breast cancer detection using actual screening data. The evaluation procedure and the results are presented in Section 4. Section 5 concludes the paper.

2 PREVIOUS RESEARCH

A number of works have already been developed to deal with multi-view breast cancer detection. Van Engeland et al. develop a linking method in [2] based on Linear Discriminant Analysis (LDA) classifier and a set of view-link features to compute a correspondence score for every possible region combination. The proposed approach demonstrated an ability to

[1] Dept. of Radiology, Radboud University Nijmegen Medical Centre, The Netherlands, email: {m.velikova, m.samulski, n.karssemeijer}@rad.umcn.nl

[2] Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands, email: {peterl, nivea}@cs.ru.nl

discriminate between true and false links. In [3], Van Engeland and Karssemeijer extend this matching approach by building a cascaded multiple-classifier system for reclassifying the initially detected region based on the linked candidate region in the other view. Experiments show that the lesion-based detection performance of the two-view detection system is significantly better than that of the single-view detection method. Paquerault et al. also consider established correspondence between suspected regions in both views to improve lesion detection based on LDA ([4]). By combining the resulting correspondence score with its one-view detection score the lesion detection improves and the number of false positives reduces. Only in this study, however, the authors report improvement in the case-based performance based on multi-view information. Therefore more research is required to build CAD systems that discriminate well between normal and suspicious cases–the ultimate goal of breast cancer screening programs.

In contrast to the clinical situation, in the screening setting the detected lesions are usually small and due to breast compression they are sometimes difficult to observe in both views. However, there is a strong correlation between the characteristics of the breast projections, which can assist the decision process of classifying a case as normal or suspicious.

3 BAYESIAN MULTI-VIEW DETECTION

3.1 Basic Definitions

A Bayesian network is defined as a pair BN $= (G, P)$ where G is an acyclic directed graph (ADG) $G = (\mathbf{V}, \mathbf{A})$ with a set of nodes $\mathbf{V}$ corresponding $1-1$ to a finite set of random variables X and a set of arcs $\mathbf{A} \subseteq (\mathbf{V} \times \mathbf{V})$ corresponding to direct causal relationships between the variables. Here P denotes a joint probability distribution of X. We say that G is an *I–map* of P if any independence represented in G, denoted by $U \perp\!\!\!\perp_G V \mid W$ with $U, V, W \subseteq \mathbf{V}$ mutually disjoint sets of nodes, is satisfied by P, i.e.,

$$U \perp\!\!\!\perp_G V \mid W \quad \Longrightarrow \quad X_U \perp\!\!\!\perp_P X_V \mid X_W,$$

where U, V and W are sets of nodes of the ADG G and X_U, X_V and X_W are the sets of random variables corresponding to the sets of nodes U, V and W, respectively. A Bayesian network BN allows a compact representation of independence information about the joint probability distribution P by specifying a *conditional probability table* (CPT) for each random variable. This table describes the conditional distribution of the node given each possible combination of values of its parents. The joint probability can be computed by simply multiplying the CPTs. For more detailed recent description of Bayesian networks, the reader is referred to [5].

One way to specify interactions among statistical variables in a compact fashion is offered by the notion of *causal independence* [6]. The general structure of a causal-independence model is shown in Figure 2; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through intermediate variables $I_1, \ldots, I_n$. A value of a variable is denoted by a lower-case letter, e.g., i_k stands for $I_k = \top$ (true) and $\bar{i_k}$ otherwise. The *interaction function* f represents in which way the intermediate effects I_k, and indirectly also the causes

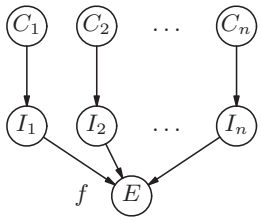

Figure 2. Causal-independence model.

C_k, interact. This function f is defined in such a way that when a relationship between the I_k's and $E = \top$ is satisfied, then it holds that $f(I_1, \ldots, I_n) = e$; otherwise, it holds that $f(I_1, \ldots, I_n) = \bar{e}$. Furthermore, it is assumed that if $f(I_1, \ldots, I_n) = e$ then $P(e \mid I_1, \ldots, I_n) = 1$; otherwise, if $f(I_1, \ldots, I_n) = \bar{e}$, then $P(e \mid I_1, \ldots, I_n) = 0$. Using information from the topology of the network, the notion of causal independence can be formalised for the occurrence of effect E, i.e. $E = \top$, in terms of probability theory as follows:

$$P(e \mid C_1, \ldots, C_n) = \sum_{f(I_1,\ldots,I_n)=e} \prod_{k=1}^{n} P(I_k \mid C_k)$$

Finally, it is assumed that $P(i_k \mid \bar{c}_k) = 0$ (absent causes do not contribute to the effect); otherwise, $P(I_k \mid C_k) > 0$.

An important subclass of causal-independence models is obtained if the deterministic function f is defined in terms of separate binary functions g_k; it is then called a *decomposable* causal-independence model [6]. Usually, all functions $g_k(I_k, I_{k+1})$ are identical for each k. Typical examples of decomposable causal-independence models are the noisy-OR [7] models, where the function g represents a logical OR. This function is used in the general theoretical model presented in the next section.

3.2 Model Description

The objective of multi-view detection of a physical object is to determine whether or not the object has certain characteristics (e.g., being suspicious) based on the characteristics of regions (subparts) in multiple object views (projections). Figure 3 depicts a schematic representation of multi-view detection.

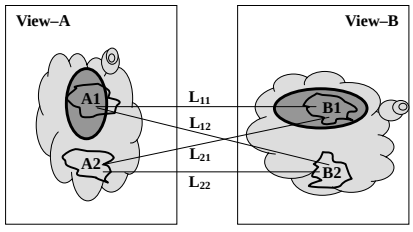

Figure 3. Schematic representation of multi-view analysis of a physical object with automatically detected regions

We have a physical object (displayed as a gray cloud), which is projected in two views, *View-A* and *View-B*. The ovals represent the projections of a suspicious physical subpart of the

object; thus, the whole object is suspicious. An automatic single-view system detects regions in both views and a number of real-valued features are extracted to describe every region. In the figure regions $A1$ and $B1$ are correct detection of the suspicious physical subpart, i.e., these are true positive (TP) regions whereas $A2$ and $B2$ are false positive (FP) regions. Since we deal with projections of the same physical object we introduce links (L_{ij}) between the detected regions in both views, A_i and B_j. Every link has a class (label) $L_{ij} = \ell_{ij}$ defined as follows

$$\ell_{ij} = \begin{cases} false & \text{if } A_i \text{ and } B_j \text{ are FP,} \\ true & \text{if } A_i \text{ or } B_j \text{ are TP.} \end{cases} \quad (1)$$

A region, view and the whole object has also a binary class with a value of $false$ if all the corresponding links $\ell_{ij} = false$; otherwise it is $true$. This definition allows us to maintain information about the suspiciousness of the physical object even if there is no detected TP region in one of the views. In any case, multiple views corresponding to the same TP subpart contain correlated characteristics whereas views of FPs tend to be less correlated.

To account for view interaction, we propose a two-step Bayesian network framework where all the regions from corresponding views are considered simultaneously to compute a single measure for suspiciousness for the physical object as whole. Figure 4 represents the framework.

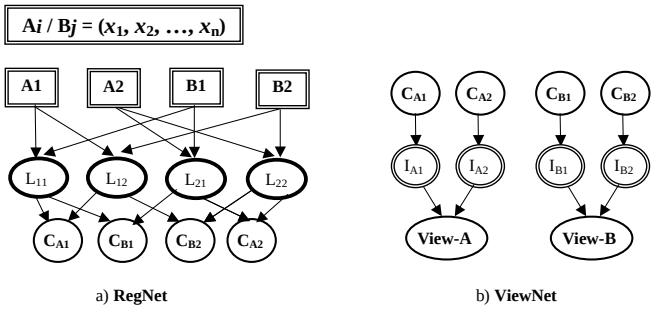

Figure 4. Bayesian network framework for representing the dependencies between multiple views of an object

At first we compute the probability that a region in one view is classified as $true$ given its links to the regions in the other view. A straightforward way to model a link L_{ij} is to use the corresponding regions A_i and B_j as causes for the link class, i.e., $A_i \longrightarrow L_{ij} \longleftarrow B_j$. Since the link variable is discrete and the regions are represented by a vector of real-valued features $(x_1, x_2, \ldots, x_n)$ extracted from an automatic detection system, we apply logistic regression to compute $P(L_{ij} = \ell_{ij} | A_i, B_j)$:

$$P(L_{ij} = \ell_{ij} | A_i, B_j) = \frac{\exp\left(\beta_0^{\ell_{ij}} + \beta_1^{\ell_{ij}} x_1 + \cdots + \beta_k^{\ell_{ij}} x_k\right)}{1 + \exp\left(\beta_0^{\ell_{ij}} + \beta_1^{\ell_{ij}} x_1 + \cdots + \beta_k^{\ell_{ij}} x_k\right)}$$

where β's are the model parameters we optimize. Logistic regression ensures that the outputs $P(L_{ij} = \ell_{ij} | A_i, B_j)$ lie in the range $[0, 1]$ and they sum up to 1.

The next step is to compute the probabilities $P(C_{Ai} = 1 | L_{ij} = \ell_{ij})$ and $P(C_{Bj} = 1 | L_{ij} = \ell_{ij})$ where C_{Ai} and C_{Bj} are the classes of regions A_i and B_j, respectively. Given our class definition in (1), we can easily model these relations through a causal independence model using the logical OR. The Bayesian network `RegNet` models this scheme.

At the second step of our Bayesian network framework we simply combine the computed region probabilities from `RegNet` by using again a causal independence model with the logical OR to obtain the probability of the respective view being $true$. We call this Bayesian network `ViewNet`.

Finally, we combine the view probabilities obtained from `ViewNet` into a single probabilistic measure for the object as a whole by using different schemes. The first simplest scheme is taking the average of both view probabilities. In another more advanced scheme, we take into account the class of the object ($false$ or $true$) by using a logistic regression model with the view probabilities as input variables. We refer to the whole multi-view detection scheme thus described as `MultiView` model.

4 APPLICATION TO BREAST CANCER

As mentioned earlier, multi-view analysis plays a crucial role in the breast cancer detection on mammograms. Here, we describe the application of the proposed Bayesian network framework in this domain.

4.1 Data Description

As input for our multi-view detection scheme we use the regions detected by a single-view CAD system that consists of the following main steps: 1) Segmentation of the mammogram into background area, breast, and for MLO, pectoral muscle; 2) Initial detection of pixel-based locations of interest; 3) Region extraction with dynamic programming using the detected locations as seed points. For each region a number of real-valued features are computed based on breast and local area information; 4) Region classification as "false" and "abnormal" based on the region features. A measure for suspiciousness is computed based on supervised learning with a neural network (NN) and converted into *normality score* (NormSc): the average number of normal regions in a view (image) with the same or higher suspiciousness measure.

The proposed model was evaluated using a data set containing 1063 screening exams from which 383 are cancerous. All exams contained both MLO and CC views. The total number of breasts were 2126. All cancerous breasts had one visible lesion in at least one view, which was verified by pathology reports to be malignant. Lesion contours were marked by, or under supervision of, an experienced screening radiologist.

For each image (mammogram) we selected the first 5 regions with the lowest NormSc computed from the CAD system. In total there were 10478 MLO regions and 10343 CC regions. Every region from MLO view was linked with every region in CC view, thus obtaining 51088 links in total.

We constructed the data such that every row contains the features of all regions belonging to the MLO and CC view for one breast, i.e., first 5 MLO and then 5 CC regions. The regions per image were sorted according to their NormSc. Every region is described by 11 continuous features automatically

extracted by the system, which tend to be relatively invariant across the views. These features include the neural network's output from the single-view CAD and lesion characteristics such as spiculation (star-like shape), focal mass, size, contrast, linear texture and location coordinates. Finally we add the class variable with binary values *false* ("normal") and *true* ("suspicious") for each link following the definition in (1). Hence a region, view, breast and case has class values of "normal" and "suspicious". Given the data processing procedure, we obtained a dataset with 2126 rows and 135 columns (10 regions × 11 features + 25 link classes).

4.2 Evaluation

We applied our `MultiView` model to the data thus described. Both `RegNet` and `ViewNet` networks have been built, trained and tested using the Bayesian Network Toolbox in Matlab ([8]). The learning has been done using the EM algorithm as the OR-nodes are hidden variables. The evaluation of each network performance is done using two-fold cross validation: the dataset is split into two subsets with approximately equal number of observations and proportion of cancerous breasts. Each half is used as a training set and as a test set.

The view probabilities for MLO and CC obtained from `ViewNet` are combined by the averaging scheme `Avg(MLO,CC)` for computing the probability of breast being suspicious. For the logistic regression combining scheme we use as input variables not only the view probabilities for MLO and CC but also the minimum NormScs for each view. The breast data is split in two halves and each half is used once as a train and a test set. This splitting is repeated 10 times. As a result for each case we obtain 20 probabilities in total–2 × 10 probabilities corresponding to each breast. Out of these 20 probabilities we first choose the maximum probability and assign it to the respective breast. Then consider only the 10 probabilities for the other breast and take the minimum as a final measure for suspiciousness. This scheme is referred to as `10-fold LR`.

To compute the probability of a case being "suspicious" we apply the most straightforward scheme of taking the maximum of the two breast probabilities. However for the `10-fold LR` model, which accounts for the breast classes, it is expected that the absolute difference between the left and right breast probabilities should be larger for the suspicious cases than that of the normal cases and thus allowing for a better case distinction. We use this difference as a third measure for suspiciousness at the case level (`10-fold LR-diff`).

We compare the performance of our model with the performance of the single-view CAD system (`SV-CAD`). For the latter the breast (case) probability is computed by taking the minimum NormSc out of all the regions in both views (breasts). The comparison analysis is done using Receiver Operating Characteristic (ROC) curve ([9]) and the Area Under the Curve (AUC) as a performance measure. The significance of the AUC differences between our multi-view model and the benchmark `SV-CAD` system is tested by using the software package LABROC4 ([10]).

4.3 Results

Based on the results from `ViewNet`, Figure 5 presents the classification outcome with the respective AUC measures per MLO and CC view, respectively. To check the significance of the difference between the AUC measures we test the hypothesis that the AUC measures are equal against the one-sided alternative hypothesis that the multi-view system yields higher AUC for MLO and CC views. The p-values obtained are: 0.000 for MLO view and 0.035 for CC view. The results clearly indicate an overall improvement in the discrimination between suspicious and normal views for both MLO and CC projections. Such an improvement is expected as the classification of each view in our multi-view system takes into account region information not only from the view itself but also from the regions in the other view.

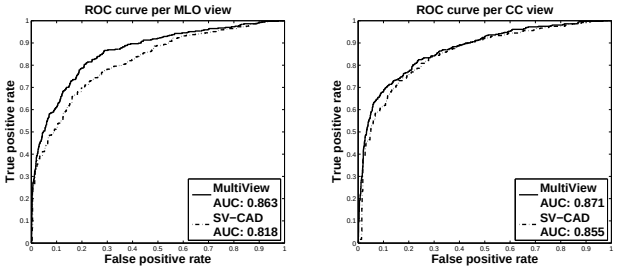

Figure 5. ROC analysis per MLO and CC view

While the view results are very promising, from a radiologists' point of view it is more important to look at the breast and case level performance; in Table 1 the AUCs from our `MultiView` and `SV-CAD` are presented. Overall we see that `MultiView` outperforms `SV-CAD` in terms of an increased true detection rate at both breast and case level. Although the simple averaging method `Avg(MLO,CC)` tends to show better distinction between normal and suspicious breasts (cases) than `SV-CAD`, the differences in the AUC measures is statistically insignificant at breast and case level. However, taking into account the breast classes and performing new training as done in `10-fold LR` leads to a significant improvement in the classification outcome. The best performance is achieved for `10-fold LR-diff`, confirming our expectation that the probability difference between the breasts for suspicious cases must be larger than that for the normal cases.

Table 1. AUCs obtained from the single- and multi-view system

Method	Breast	p-value	Case	p-value
SV-CAD	0.850	–	0.797	–
Avg(MLO,CC)	0.864	0.123	0.827	0.135
10-fold LR-max	0.875	0.001	0.832	0.014
10-fold LR-diff	0.875	0.001	0.838	0.006

To get more insight into the areas of improvement we plotted ROC curves for each of our models against the single-view CAD system. For all the plots we observed the same tendency of an increased true positive rate at (very) low false positive rates (< 0.5)–a result ultimately desired at the screening practice where the number of normal cases is considerably larger than the suspicious ones; Figure 6 presents the ROC plot for the best performing `10-fold LR-diff` model.

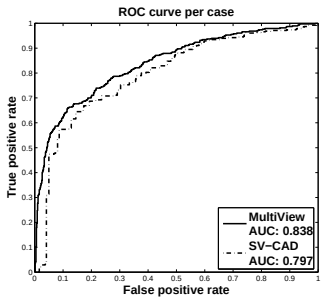

Figure 6. ROC analysis per case

To have a closer look at the quality of classification for the models that produce a real probability measure for suspiciousness (`Avg(MLO,CC)` and `10-fold LR-max`), we compute the average log-likelihood (ALL) of the probabilities for different *units*–link, region, MLO/CC view, breast and case–by:

$$ALL(C) = \frac{1}{N}\sum_{i=1}^{N} -\ln P(C_i|\mathcal{E}_i), \tag{2}$$

where N is the number of the unit, C_i and $\mathcal{E}_i$ is the class value and the feature vector of the i-th observation, respectively. Thus, the value of $ALL(C)$ indicates how close the posterior probability distribution is to reality: when $P(C_i|\mathcal{E}_i) = 1$ then $\ln P(C_i|\mathcal{E}_i) = 0$ (no extra information); otherwise $-\ln P(C_i|\mathcal{E}_i) > 0$.

The log-likelihood results are given in Table 2. The lowest $ALL(C)$ is achieved for the links meaning that the estimated probabilities best fit the link probability distribution. A possible explanation is that in our Bayesian network framework the links are directly dependent on the original region features and thus they are better fitted. On the other hand, the rest of the units are based on combining estimated probabilities from previous levels where noise could play a role. Overall, however, our log-likelihood results show that `MultiView` fits closely the probability distributions for different units.

Table 2. Average log-likelihood of the class based on `MultiView`

Method	Average log-likelihood of the class				
	Case	Breast	MLO/CC	Link	Region
Avg(MLO,CC)	0.50	0.32	0.34/0.31	0.19	0.38
10-fold LR-max	0.47	0.30			

5 CONCLUSIONS

Using the proposed Bayesian network framework and expert knowledge on multi-view analysis of mammograms we showed that the detection rate of breast cancer is larger at low false positive rates than that of a single-view CAD system. This improvement is achieved at view, breast and case level and it is due to a number of factors. First, we built upon a single-view CAD system that already demonstrates relatively good detection performance. By applying a probabilistic causal model we linked the original features extracted by CAD for all the regions in MLO and CC views and we combined all the links for one breast to obtain a single measure for suspiciousness of a view, breast and case. Another factor for the improved classification is that our approach incorporated domain knowledge. Following radiologists' practice, we applied a straightforward scheme to account for multi-view dependencies such that (i) correlations between the regions in MLO and CC views are considered per breast as whole and (ii) the classification of breast/case as "suspicious" is employed through the logical OR. Thus the proposed methodology can be applied to any domain (e.g., fault detection in manufacturing processes) where similar definitions and objectives hold.

Although we demonstrated that the proposed framework has the potential to assist screening radiologists to improve the evaluation of breast cancer cases, we consider a number of directions for extension. First, the features used in the current model are independently computed per region. We expect that the inclusion of multi-view region features such as the distance to the nipple or correlation features would further improve the system's performance by considering explicitly multi-view dependences. Another possible extension is based on the model structure. Following our Bayesian network framework with using logistic regression and logical OR at a link and view level, we can also apply similar combining schemes at a breast and case level. Thus we can allow for better handling of missing or noisy information in the estimation of the breast/case probabilities for suspiciousness.

ACKNOWLEDGEMENTS

Work funded by the Netherlands Organisation for Scientific Research under BRICKS/FOCUS grant number 642.066.605.

REFERENCES

[1] Breast cancer and screening. Technical report, World Health Organization, `http://www.emro.who.int/ncd/publications/breastcancerscreening.pdf`, accessed on 25-02-2008.

[2] S. van Engeland, S. Timp, and N. Karssemeijer. Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views. *Medical Physics*, 33(9):3203–3212, 2006.

[3] S. van Engeland and N. Karssemeijer. Combining two mammographic projections in a computer aided mass detection method. *Medical Physics*, 34(3):898–905, 2007.

[4] S. Paquerault, N. Petrick, H. Chan, B. Sahiner, and M. A. Helvie. Improvement of computerized mass detection on mammograms: Fusion of two-view information. *Medical Physics*, 29(2):238–247, 2002.

[5] F.V. Jensen and T.D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2007.

[6] D. Heckerman and J. S. Breese. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 26(6):826–831, 1996.

[7] F. Diez. Parameter adjustment in Bayes networks: The generalized noisy or-gate. In *Proceedings of the Ninth Conference on UAI, San Francisco, CA*. Morgan Kaufmann, 1993.

[8] K. Murphy. Bayesian Network Toolbox (BNT), `http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html`.

[9] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[10] C.E. Metz. Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, 24:234–245, 1988.