



A probabilistic framework for image information fusion with an application to mammographic analysis

Marina Velikova^{a,*}, Peter J.F. Lucas^a, Maurice Samulski^b, Nico Karssemeijer^b

^a Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands

^b Department of Radiology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

ARTICLE INFO

Article history:

Received 19 May 2011

Received in revised form 20 November 2011

Accepted 16 January 2012

Available online 24 January 2012

Keywords:

Bayesian networks

Information fusion

Multi-view image analysis

Computer-aided detection

Mammography

ABSTRACT

The recent increased interest in information fusion methods for solving complex problem, such as in image analysis, is motivated by the wish to better exploit the multitude of information, available from different sources, to enhance decision-making. In this paper, we propose a novel method, that advances the state of the art of fusing image information from different views, based on a special class of probabilistic graphical models, called causal independence models. The strength of this method is its ability to systematically and naturally capture uncertain domain knowledge, while performing information fusion in a computationally efficient way. We examine the value of the method for mammographic analysis and demonstrate its advantages in terms of explicit knowledge representation and accuracy (increase of at least 6.3% and 5.2% of true positive detection rates at 5% and 10% false positive rates) in comparison with previous single-view and multi-view systems, and benchmark fusion methods such as naïve Bayes and logistic regression.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The increasing number and heterogeneity of information sources and techniques for data acquisition, especially in complex domains such as image analysis, has produced vast amounts of information, giving rise to an increasing demand for combined processing of information to extract valuable knowledge and make better decisions. The term *information fusion* is used when such merging of information from different sources is done automatically.

There have been given numerous definitions of the notion of information fusion in the past 20 years (Boström et al., 2007); its key point is that the synergy of multiple sources should lead to more effective support, i.e., better decisions or actions, e.g. in terms of accuracy, than when these sources were used separately. Yet, this should be done without sacrificing computational efficiency, as otherwise the fusion of information might be intractable. Often, in fusing image information one also has to deal with the inherent *uncertainty* of the information, e.g. due to measurement or interpretation errors. The fusion of such uncertain image information is the subject of this paper.

Given the rich expertise of human interpreters and the computing power of intelligent computer-based systems, we believe that their integration is the main road for building smarter computer-

aided detection (CAD) systems. In this work, we adopted this synergistic principle to develop a novel CAD system for automated multi-view image analysis by combining knowledge derived from the analysis of the way humans interpret images, on the one hand, and image information automatically extracted by a CAD system for image interpretation, on the other hand. In particular, we built a multi-stage system using Bayesian networks—one type of probabilistic graphical model—which are especially promising in bridging the gap between the capabilities of humans and computer-aided interpretation, as they can support the explicit representation of expert knowledge, handle uncertainty and missing information, and allow combining multiple sources of knowledge.

The application domain of this paper is the analysis of X-ray images, or *mammograms*, for breast cancer detection in screening programs, which will be referred to as *mammographic analysis* in the remainder of the paper. In mammographic analysis, the majority of current CAD systems are mainly used to analyze single images only and often act as prompt systems, meant to focus the radiologist's attention to particular image regions, rather than providing overall classification of the patient's condition. They have no or limited capability to capture the working principles employed by radiologists. An example of one of such principles is multi-view mammographic analysis, where the radiologists judge for the presence of cancer on the basis of two projections, or *views*, of the same breast: mediolateral oblique (MLO), taken under 45° angle and showing part of the pectoral muscles, and craniocaudal (CC), taken head to toe. Human readers also normally compare image parts and different images of the breasts to each other, i.e., they interpret

* Corresponding author.

E-mail address: marinav@cs.ru.nl (M. Velikova).

potentially suspicious regions in the context of all other available image information.

The system and the results presented here extend and improve upon our multi-view methods described in (van Engeland et al., 2006; Samulski and Karssemeijer, 2008; Velikova et al., 2009a; Velikova et al., 2009b). The basic idea of all these methods is to use links between the regions automatically detected by a single-view CAD system in two breast views, or projections, in order to improve the region and patient-based classification. In (van Engeland et al., 2006; Samulski and Karssemeijer, 2008), the main focus was on the cancer detection at a region level rather than at a patient level, although the latter is the ultimate aim of breast-cancer screening programs. Linking of regions was performed twice, using each of the views as a starting point, which resulted in asymmetrical links between any two regions characterized by a correspondence score.

In existing methods, supervised classification methods are used to combine information. With increasing complexity, however, this traditional approach becomes less attractive. To model reading of mammograms, many sources of information have to be combined. For example, multiple views may include different projections as well as different exams in a temporal screening sequence. The presence of multiple features such as microcalcifications and masses may have to be combined, where it is not only the detection of these features, but also the question whether they are more indicative for malign or benign processes that plays a role in decision making processes. Most of these image interpretation tasks have been addressed as separate problems in the literature. Combining them in a consistent manner is far from trivial. One of the problems one has to deal with is missing data, as not all information sources are available for all cases in the training database. To tackle this problem we propose to use a hybrid method, where sources of information provided by systems previously developed for specific subtasks are combined in a probabilistic manner. It is noted that this method also allows inclusion of non-imaging information like age and breast cancer risk to be accounted for in a proper manner.

In (Velikova et al., 2009b) we already proposed a multi-view CAD system based on probabilistic modeling. The linking of breast projections and view classification was modeled using Bayesian networks and single-view region features. Some improvement in the discrimination between cancerous and non-cancerous patients was obtained in comparison to the single-view CAD system, but the modeling scheme was rather limited by the requirement for a fixed, equal number of regions detected per breast projection as by the simplified definition of a link as being only true or false. Furthermore, the modeling at a breast and patient level was using a relatively simple logistic regression method. This motivated us to develop the more advanced multi-view fusion method presented in this paper. Major differences are the following: (1) the modeling has unifying and systematic nature based on a special type of Bayesian network, called causal independence models, at all levels of analysis—link, region, breast and patient—following the human reader's practice, (2) the model is redesigned to allow handling of a varying number of suspicious regions detected per projection, and (3) more complex links are introduced, making use of correspondence scores and multi-view features generated by a recently developed two-view analysis method (Samulski and Karssemeijer, 2011). The proposed generic modeling principles allow straightforward application to potentially any problem of information fusion. To evaluate the method we compare our method against standard probabilistic methods, such as naïve Bayes and logistic regression, and our previous single- and multi-view systems (Samulski and Karssemeijer, 2011; Velikova et al., 2009b).

In the next section we formalize the problem of information fusion and discuss the theoretical background related to Bayesian

networks and causal independence models. The main contribution of this work is the causal probabilistic modeling scheme, described in Section 3, followed by its application on real mammographic data and discussion of the results in Section 4. Conclusions are given in Section 5.

2. Uncertain information fusion

2.1. Basic principles

The general problem of uncertain information fusion can be formulated as follows. Suppose that a number of *sources of evidence* are given, indicated by e_1, e_2, \dots, e_n , which provide uncertain information about a *hypothesis* h . Our goal then is to merge the information from the variables e_i to draw a conclusion about the hypothesis h , i.e.,

$$h = f(e_1, e_2, \dots, e_n), \quad (1)$$

where f is a general fusing scheme with the characteristic that the more e_1, e_2, \dots, e_n tell about h , the higher the certainty about h . The problem of uncertainty information fusion inevitably becomes relevant as soon as one wishes to solve problems of sufficient complexity.

The crucial bottleneck of uncertain information fusion, expressed by Eq. (1), is the modeling of the interaction between the uncertain sources of evidence e_1, \dots, e_n . Although this interaction can be quite naturally expressed by the family of probability distributions $P(H|E_1, \dots, E_n)$, where each instantiation of the variables $E_1 = e_1, \dots, E_n = e_n$ leads to a joint conditional probability distribution $P(H|E_1 = e_1, \dots, E_n = e_n)$, this will in general lead to a specification that is exponential in the size of the domains of the variables E_1, \dots, E_n .

In the early days of information fusion, researchers, therefore, discarded this intuitively appealing probabilistic approach. Instead, typical early solutions tried to take advantage of particular ways of decomposing the function f of Eq. (1), for example by defining f in terms of a binary function g with

$$f(e_1, \dots, e_n) = g(e_1, g(e_2, \dots, g(e_{n-1}, e_n) \dots)).$$

If the order in which the function g is applied does not matter for the result, then the function f is called *symmetric* and the function g is associative, i.e., $g(e_i, g(e_j, e_k)) = g(g(e_i, e_j), e_k)$, and commutative, i.e., $g(e_i, e_j) = g(e_j, e_i)$. These properties make it much easier to fuse uncertain information: the merge can be done pairwise in any possible order.

Many of the early attempts can be characterized in this fashion, such as the *certainty-factor (CF) calculus*, introduced by Shortliffe and Buchanan (Shortliffe and Buchanan, 1975), quantifying the uncertainty in the degree to which available evidence supports a hypothesis. The CFs of rules with the same hypothesis and different pieces of evidence are combined in any order without affecting the result using binary, associative and commutative functions (Lucas and van der Gaag, 1991). Although, CF calculus has been justifiably criticized for its lack of a principled foundation, it has also been proved that a significant fragment of CF calculus can be mapped to Bayesian networks, which perhaps explains the good performance of the systems that employed the CF calculus (Lucas, 2001). To further relax the restrictions of probability theory for mutually exclusiveness of hypotheses, alternative approaches have been developed to handle uncertainty in a general sense. One example is *belief theory* with its representative: Dempster–Shafer theory of evidence (Dempster, 1967; Shafer, 1976). The theory has been popular for a long time as it includes a function for fusion of uncertain information, known as *Dempster's combination rule* (Dempster, 1967), which is again associative and commutative.

Alternative information fusion methods, which preserve the associative and commutative properties, are based on simple aggregation operators such as (weighted) average, minimum, maximum, or a combination of them. Such methods have been investigated in a recent paper for fusion of multiple images at an exam level to improve the retinal image screening for diabetic retinopathy (Niemeijer et al., 2009). Other current studies for decision- and feature-based fusion in medical imaging, using statistical methods or traditional classifiers (such as neural networks), have been proposed in (Jesneck et al., 2006; Calhoun and Adali, 2009).

2.2. Probabilistic methods

2.2.1. Bayes theorem and Bayesian networks

Although, initially probability theory was mostly discarded as a method for information fusion, for the reasons mentioned at the beginning of this section, there is early work that explored probabilistic theory by employing Bayes' theorem:

$$P(H|E_1, \dots, E_n) = \alpha P(E_1, \dots, E_n|H)P(H), \quad (2)$$

where α is a normalizing constant. Assuming that the pieces of evidence E_i are all conditionally independent given the hypothesis H , i.e., $P(E_i|H, E_j) = P(E_i|H)$ for $i \neq j$, yields:

$$P(E_1, \dots, E_n|H) = \prod_{k=1}^n P(E_k|H).$$

This approach is often called *naïve Bayes* or *idiot's Bayes*. Note that this independence assumption renders the use of Bayes' theorem again associative and commutative in evidence E_i . In order to use the method, one needs a probability distribution $P(E_i|H)$ for each information source, and the prior probability distribution $P(H)$. Hence, the naïve Bayes (NB) classifier combines the naïve Bayesian network model with a decision rule ((Domingos and Paz-zani, 1997)). A common choice is the most probable hypothesis – a maximization criterion – and then the classifier is defined by:

$$\text{NB}(e_1, \dots, e_n) = \arg \max_h P(H = h) \prod_{k=1}^n P(E_k = e_k|H = h). \quad (3)$$

Despite its naïve design and apparently oversimplified assumptions, the naïve Bayes classifier often demonstrates good classification performance in many complex real-world situations; for example, in robotics it is used for information fusion of maps obtained by different sensors of a robot exploring the same area (Moravec, 1988). Another advantage is that it requires a small amount of training data to estimate the network parameters (the conditional probability distributions). Therefore, it is also considered in this study as one of the benchmark probabilistic methods to compare against the proposed causal model.

More recently, researchers have explored more general independence constraints on a joint probability distribution $P(H, E_1, \dots, E_n)$, in the form of *Bayesian networks*. A Bayesian network (BN) is defined as a pair $B = (G, P)$, where G is an acyclic directed graph (ADG) $G = (V, E)$ and P is a joint probability distribution of a set of random variables. There exists a 1–1 correspondence between the nodes in V and the random variables; the (directed) edges, or arcs, $E \subseteq (V \times V)$ correspond to direct causal relationships between the variables. A Bayesian network offers a compact representation of the joint probability distribution P in terms of local *conditional probability distributions* (CPDs) or *tables* (CPTs), if the variables are discrete, by taking into account the conditional independences represented by the ADG. In contrast to discriminative (classification) tools such as neural networks, a Bayesian network is a generative model (no input and output variables are explicitly defined), which allows easy integration of domain knowledge and work with

missing data, while the network can still be used for discriminative tasks.

Bayesian networks have already been exploited for information fusion in a number of studies. Mnatsakanyan et al. propose in (Mnatsakanyan et al., 2009) a manually constructed Bayesian network using the knowledge of experienced epidemiologists to fuse data from various sources to improve the identification of influenza-like epidemiologically relevant events. To account for time changes, Zhang and Ji propose a dynamic Bayesian network (DBN) for merging image sequences to perform facial expression understanding (Zhang and Ji, 2005).

2.2.2. Causal independence models

To provide an efficient way to specify interactions among random variables in a compact fashion, the notion of *causal independence* has been introduced (Heckerman and Breese, 1996). Causal independence arises when multiple causes (parent nodes) lead to a common effect (child node) through interaction of independent uncertain processes. This type of models allows in a systematic way to decompose a probability distribution in terms of Boolean interactions among local parameters. Such a decomposition makes it easier and tractable to deal with problems involving a large number of causes, which is often the case in many domains such as medicine.

The general structure of a causal-independence model, as defined in (Visscher et al., 2009), expresses the idea that causes C_1, \dots, C_n influence a given common effect E through intermediate variables I_1, \dots, I_n ; the intermediate variable I_k is considered to be a contribution of the cause variable C_k to the common effect E . The way in which the intermediate effects I_k , and indirectly also the causes C_k , interact is represented by an *interaction function* f . A natural class of interaction functions f are the *symmetric* Boolean functions. There are 2^{n+1} of such functions, with n the number of arguments; typical examples are the binary logical functions OR, AND and XOR, which are all associative and commutative. In mammography, for example, the order of regions does not play a role in determining whether or not there is suspicion for breast cancer.

A useful feature of symmetric Boolean functions is their decomposability in terms of *exact* Boolean functions. The exact function ε_q checks whether there are *exactly* q trues among its arguments, i.e., $\varepsilon_q(I_1, \dots, I_n) = 1$, if $\sum_{k=1}^n I_k = q$ with I_k equal to 0 or 1. In decision making under uncertainty there is a natural tendency to aggregate available uncertain information until a threshold is passed. The *threshold function* τ_q is a symmetric Boolean function that allows us to model this principle; it checks whether there are *at least* q trues among its arguments, i.e., $\tau_q(I_1, \dots, I_n) = 1$, if $\sum_{k=1}^n I_k \geq q$. Note that the logical OR function is a threshold function τ_q with $q = 1$ and the AND function is a threshold function τ_q with $q = n$. The conditional probability of the effect variable E given the causes C_1, \dots, C_n in a noisy threshold model with interaction function τ_q is given by:

$$P_{\tau_q}(e|C_1, \dots, C_n) = \sum_{q \leq l \leq n} \sum_{\varepsilon_l(I_1, \dots, I_n)} \prod_{k=1}^n P(I_k|C_k). \quad (4)$$

Hence, the computation of $P_{\tau_q}(e|C_1, \dots, C_n)$ in a noisy threshold model involves only summation and product of a linear number of parameters and this can again be done in pairs by counting the number of truths, making its application easy and attractive for complex realistic problems. An example illustrating causal independence models with exact and threshold functions is given in [Appendix A](#). More details on this topic can be found in (Visscher et al., 2009).

Despite, the wide spectrum of information fusion methods there is still a need for further research to allow effective, computationally feasible and systematic merge of information sources.

Given the sound theoretical basis and computational advantages of causal independence models, we use them as basic elements in the novel model for multi-view image analysis, presented in the next section, which possesses the properties for practically useful application.

3. Causal model for image information fusion

3.1. Multi-view image analysis

The objective of multi-view image analysis of a physical object (e.g., the breast) is to determine whether or not the object has certain characteristics (e.g., being cancerous) based on the characteristics of regions (subparts) in multiple object views. Fig. 1 depicts the general multi-view detection scheme.

We have a physical object, shown in Fig. 1a, referring to an organ, which is projected in two views, *View-A* and *View-B* shown in Fig. 1b. Suppose we have an abnormal physical subpart (e.g., cancer) of the object (the dark gray ball in the middle), which is also projected in both views (the dark gray ovals); hence, the whole object is considered to be abnormal. An automatic single-view system attempts to establish whether there are regions that are abnormal in each view separately based on a number of single-view real-valued extracted features, e.g., size, location, contrast. In the figure, regions A_1 and B_1 are correct detections of the abnormal physical subpart, i.e., these are true positive (TP) regions, whereas regions A_2 and B_2 are false positive (FP) regions. Since we deal with projections of the same physical object, correspondence between subparts is represented by a *link* ($LINK_{ij}$) between the detected regions in each view, A_i and B_j . To every link, a value $LINK_{ij} = \ell_{ij}$ is assigned, where $\ell_{ij} \in \{TPTP, TFP, FPTP, FFP\}$; for example in the figure $\ell_{21} = FPTP$. Every link is described by a set of *multi-view features*, e.g., contrast difference and location difference. For every region and object, a class with values of *true* (abnormal) or *false* is assumed to be provided apriori, e.g., by pathology or an expert, which supervizes the training process and provides the ground-truth data for evaluation during testing.

Multiple views corresponding to the same abnormal subpart contain correlated characteristics whereas views corresponding to normal parts tend to be less correlated. For example, in mammography an artifactual density might appear in one view, whereas it disappears in the other view due to the superposition of normal tissue. To account for the interaction between the object projections, we next present a novel Bayesian network framework for combining image information in a probabilistic manner, where view dependencies are modeled explicitly.

3.2. Causal probabilistic model

Our novel method is presented in Fig. 2, modeling the problem presented in Fig. 1. We emphasize that the graphical structure of the model is manually built using the domain knowledge whereas the parameters (probability distributions) are learned from data.

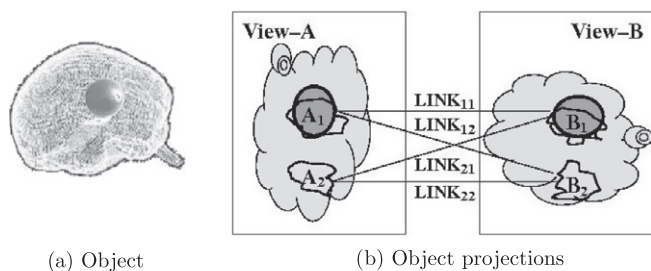


Fig. 1. Schematic representation of multi-view image analysis with automatically detected regions in (b) of a physical object with abnormal subpart (the dark gray ball) in (a).

We start by modeling the multi-view dependencies between the regions in View-A and View-B. For each of the four link values ℓ_{ij} we consider the links $LINK_{ij}$ with a respective set of descriptors \mathbf{D} such as multi-view features **MVFeat** and correspondence scores $CorrSc(LINK_{ij})$, $CorrSc(LINK_{ji})$ obtained from another linkage systems (e.g., as the one presented in (Samulski and Karssemeijer, 2011)). We then apply logistic regression to reliably compute the probability $P(LINK_{ij} = \ell_{ij} | \mathbf{D}) = 1/(1 + \exp(-z))$, where z is a linear combination of the descriptors in \mathbf{D} . Thus, for every $LINK_{ij}$ we obtain four probabilities corresponding to each link value.

At the second stage, we compute the probability of a region being abnormal given the link information about the regions in the complementary view. This is done by combining link probabilities obtained from the first stage using a causal independence model, where the link probabilities are the cause variables and the region probability is the effect variable. In computing the probability of a region in View-A having certain abnormal characteristics (e.g. cancer), we combine only the link probabilities for the classes TPTP and TFP as they correspond to a TP region in View-A, whereas for a region in View-B, the link classes considered are TPTP and FPTP. These link probabilities interact through the XOR function, as only one of them can be true. Here I_k denotes the intermediate variable corresponding to the link probability index k and the region index j in the other view. Next the logical OR is used to represent the knowledge that the probability of a region having certain characteristics is true if at least one of the link probabilities is true.

At the third stage, we focus at the object level where the region probabilities for the respective views computed at the previous stage are combined using a causal independence model with a threshold function τ_q , where q is the threshold. In the combining scheme, we also integrate other features, which may contribute to the description of the regions; for example in the current mammography application we use the *normality scores* (NormSc), computed by a single-view CAD system as good indicators for suspiciousness, i.e., the smaller the normality score the higher the likelihood for cancer. By varying the threshold q from 1 to the maximum number of regions detected in both views, one can get insight into the causal interactions between the regions and the object. Having a value of $q = 1$ implies that at least one region is required to be detected as suspicious, and thus visible, in one view to classify the object as suspicious; hence, the model can tackle situations where the abnormality is not visible on one of the views. The best performing threshold values will depend on the domain characteristics. For example, in mammography one can expect that models with small threshold values (i.e., smaller number of regions are required to be cancerous to define a breast as cancerous) would be able to distinguish well between cancerous and normal breasts whereas models with larger values of q might not be able to make the distinction. This expectation follows from the fact that breast cancer in its early stages is mostly unifocal, i.e., located in a single region, and not observed on multiple locations in the breast (view).

The model just described can easily be extended with next stages if, for example, more objects are simultaneously analyzed. Such extension in the domain of mammographic analysis is presented in Section 4.1.

3.3. Objective evaluation scheme

Given the multi-stage nature of the proposed model, we next propose an objective cross-validation scheme to evaluate the classification performance of the model. We split the data in a number of subsets (folds) and for every data split i and every stage S , one subset is used only for testing ($Test^S$) and never for training at different stages of the model. In such a way, the test and training sets are kept independent over the stages, as done in practice.

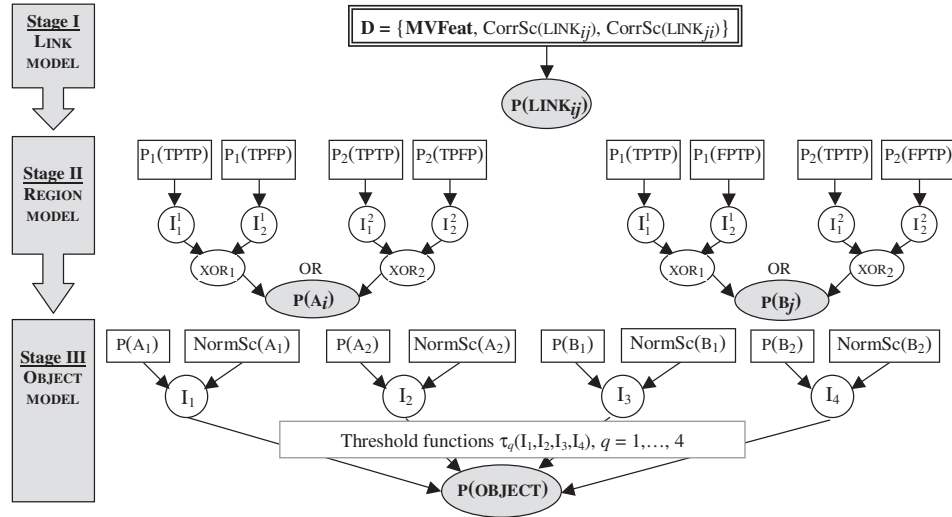


Fig. 2. Multi-stage causal probabilistic model for the problem of multi-view image analysis presented in Fig. 1.

Furthermore, we split the training data ($Train^S$) into two equally-sized subsets $Set-j$, which are used to train two classifiers $Classifier-j$, $j = 1, 2$, at each stage. These classifiers are applied to the alternative training subset to create the training data ($Train^{S+1}$) for the subsequent stage in the model, which can be considered as validation data. This approach helps prevent overtraining the model in the later stages by reducing the tendency of the model to get fitter to the same training data only. The two classifiers are also applied to the test set, resulting into two sets of test results at every stage. These two sets are further combined by taking their *mean*, which produces again one test set for the next stage of the model ($Test^{S+1}$). The evaluation scheme is given in Fig. 3.

We note that the proposed scheme can be extended or modified depending on the available data and task at hand. For example, depending on the number of cases, which in screening mammography is becoming easily available, one may also consider more splits of the training data and thus training more “experts”. Also, other combination functions, such as *max*, may also be applied, but for the application at hand it is sufficient to consider the mean.

4. Application to mammographic analysis

In this section we explore the use of the proposed causal independence model for practical image analysis tasks and we consider mammographic analysis as an example, selected because we have

much experience with it. In contrast to the previous research in this area (Good et al., 1999; van Engeland and Karssemeijer, 2007; Paquerault et al., 2002; Qian et al., 2007; Zheng et al., 2006; Wei et al., 2009), which mostly explores neural networks or linear discriminant analysis for multi-view mammographic analysis, the probabilistic methodology proposed in the current study has the advantages of fusing uncertain information typically inherent in the domain (e.g., obscure lesions) in a systematic fashion at all levels of analysis, being easily extended to more information sources (such as multimodality or patient data), and providing not only strong predictive power but also an insight in the results obtained—properties desired especially by medical experts.

4.1. Causal model for mammographic analysis

We applied the causal probabilistic model described in Section 3.2 to multi-view mammographic analysis. The model is built upon the results from our previous CAD systems. Fig. 4 depicts the automatic interpretation of the mammographic examination of one patient, consisting of four mammograms (two MLO and two CC images). The original images are input to the single-view CAD (SV-CAD) system, which for every image independently segments the breast, detects regions of interest, extract region features and use them to compute the region’s likelihood for cancer ($NormSc(A_i)$ and $NormSc(B_j)$ in Fig. 2) using a neural network classifier.

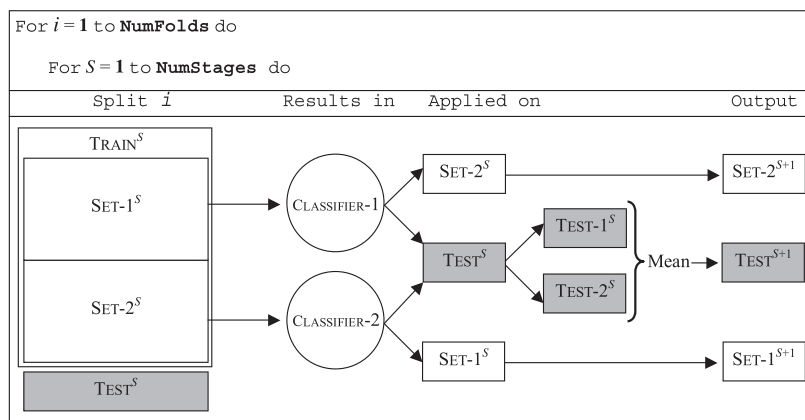


Fig. 3. Evaluation scheme for a multi-stage model with cross-validation.

The resulting regions from the MLO and CC views for the same breast are then analyzed in parallel to establish links between them. This is done by a multi-view linking (*Linker*) system using a kNN classifier, where multi-view features are extracted and used as input to compute a correspondence score for two regions representing the same lesion; for the full description of *SV-CAD* and *Linker*, their methodological principles, scheme and corresponding features the reader is referred to (Samulski and Karssemeijer, 2011).

Finally, our multi-view causal model (*MV-CAD-Causal*) uses the multi-view features and the correspondence scores as input to compute the probability for a link, region, breast and the whole exam being cancerous considering the available information from all the images simultaneously. This modeling process is inspired by the way the radiologists perform their analysis work in practice and is an example of information fusion. In comparison to the *Linker* system, the advantage of our causal method is making the links symmetric, i.e., for every two regions in MLO and CC views there is only one probability per link class, in contrast to the former system where two independent link probabilities for every two regions exist.

The multi-view causal model summarized in Fig. 2 was subsequently employed for mammographic analysis. A breast corresponds to the physical object in Fig. 1a and the MLO and CC views correspond to View-A and View-B in Fig. 1b, with respective detected regions. The object characteristic we are interested in is the detection of cancer, so we compute probabilities for cancer of the regions at the second stage of the model and of both left and right breast at the third stage. For this specific application, we extend the multi-view causal model with a fourth stage, modeling the exam level considered eventually in the screening practice; see Fig. 5. The term *exam* refers to a patient who has undergone a mammographic examination, which usually comprises the set

of four mammographic images from the two views of both breasts. At this stage, we combine the probabilities for the left and right breast, obtained from the third stage, and their respective single-view normality scores ($\text{NormSc}(\text{BREAST})$) using a causal independence model to compute the probability for an exam being cancerous. Two combination functions are used and compared: the logical OR and the MAX function. As two alternatives to this information fusion approach at an exam level, we use naïve Bayes (*MV-CAD-NB*) (see Eq. (3)) and logistic regression (*MV-CAD-LR*) as combination methods for computing the exam probability for cancer using the same input information as for the causal model. The naïve Bayes model contains one parent node ($P(\text{EXAM})$) and 4 children nodes ($P(\text{LBr})$, $\text{NormSc}(\text{LBr})$, $P(\text{RBr})$ and $\text{NormSc}(\text{RBr})$) whereas logistic regression makes the prediction by fitting the data to a logistic curve: $P(\text{EXAM} = \text{cancer}) = 1/(1 + \exp(-z))$, where z is a linear combination of the probabilities for both breasts being cancerous and their respective normality scores. Naïve Bayes and logistic regression are two common methods used for probabilistic classification tasks due to their simple nature, easy training and in many cases superior performance in comparison to more complex models. We also note that the naïve Bayes classifier and logistic regression are the only two other efficient methods that respect the principles of information fusion as described in Section 2. Therefore we choose them as benchmark methods for comparison with the causal independence model in this study.

4.2. Data description, training and evaluation

The data set contained 392 (332 diagnostic +60 missed by radiologists) exams from which 218 (185 diagnostic +33 missed by radiologists) were cancerous. All exams contained both MLO and CC views. All cancerous breasts had one visible lesion in at

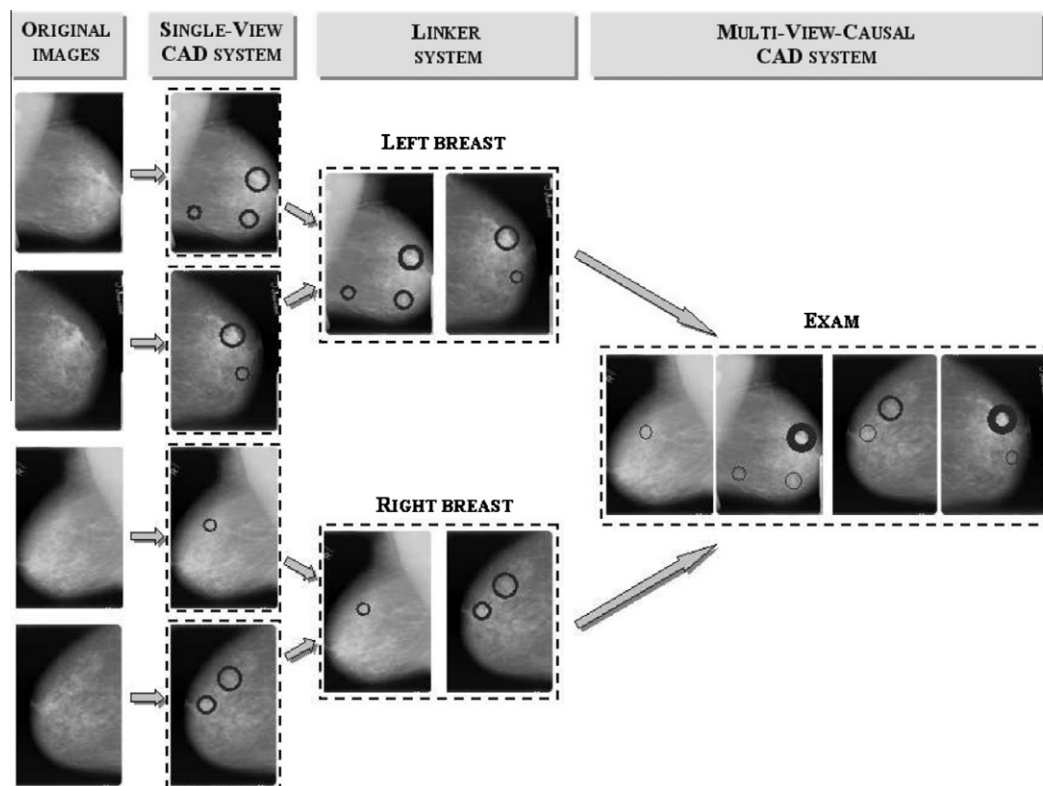


Fig. 4. Schematic representation of the multi-view CAD system development based on the images of one patient. The dashed lines represent the final level of modeling—image, breast and exam—for each of the three systems.

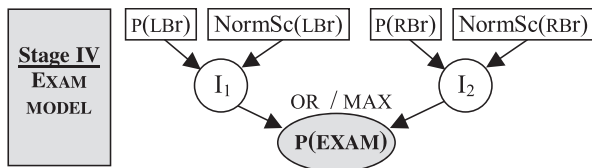


Fig. 5. Extension of the causal model in Fig. 2 for modeling exam level.

least one view, which was verified by pathology reports to be cancerous. Lesion contours were marked by a mammogram reader.

For each image (mammogram) we have a number of regions (between 1 and 5) detected as suspicious by the single-view CAD system (Samulski and Karssemeijer, 2011). The variation in the number of regions per image was explicitly tested in our previous work (Velikova et al., 2009b) and it was not the main focus of this work. Based on the ground-truth data, for each region we assigned a class value of *true* (TP) if the region is the true detection of cancer and *false* (FP) otherwise. Every region from MLO was linked with every region in CC and every link was described by a set of multi-view features (e.g., compactness difference, distance-to-nipple difference, pixelwise correlation, histogram correlation, linear texture difference) and the correspondence scores computed by the *Linker* system; the full description of the features and scores is given (Samulski and Karssemeijer, 2011). To every link we assigned one of the four link class values depending on both region classes. To every exam we assigned a binary class with values of *true* (cancerous) and *false* (normal) based on the ground-truth information.

The proposed model has been built, trained and tested using the Matlab-based Bayesian Network Toolbox ((Murphy, 2007)). The evaluation of the model is done using ten-fold cross validation with the same data split as the one used in (Samulski and Karssemeijer, 2011). The split of the data is stratified such that the percentage of the cancerous cases per fold-on average $55\% \pm 6\%$ – reflects the proportion of cancerous cases in the whole data. At a link level the performance of our multi-view model is compared with the *Linker* system. At a region and exam level, the benchmark for comparison is the *SV-CAD* system. For the latter, the likelihood for an exam being cancerous is computed by taking the likelihood of the most suspicious region. In addition, at an exam level the multi-view causal model is compared with *MV-CAD-NB*, *MV-CAD-LR* and the *MV-CAD-LR-LR* model we proposed in (Velikova et al., 2009b), using the current (splits of) data and evaluation scheme. The classification performance is evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), a standard performance measure in the medical imaging research. The significance of the differences obtained in the AUC measures is tested using the ROCKIT software for fully paired data ((Metz et al., 1984)).

4.3. Results

We next present the test results obtained from the proposed multi-view causal model and the benchmark methods for various tasks. In addition, in Section 4.3.3, we also report results obtained from the training data to obtain insight in the proposed training/testing evaluation scheme.

4.3.1. Classification performance

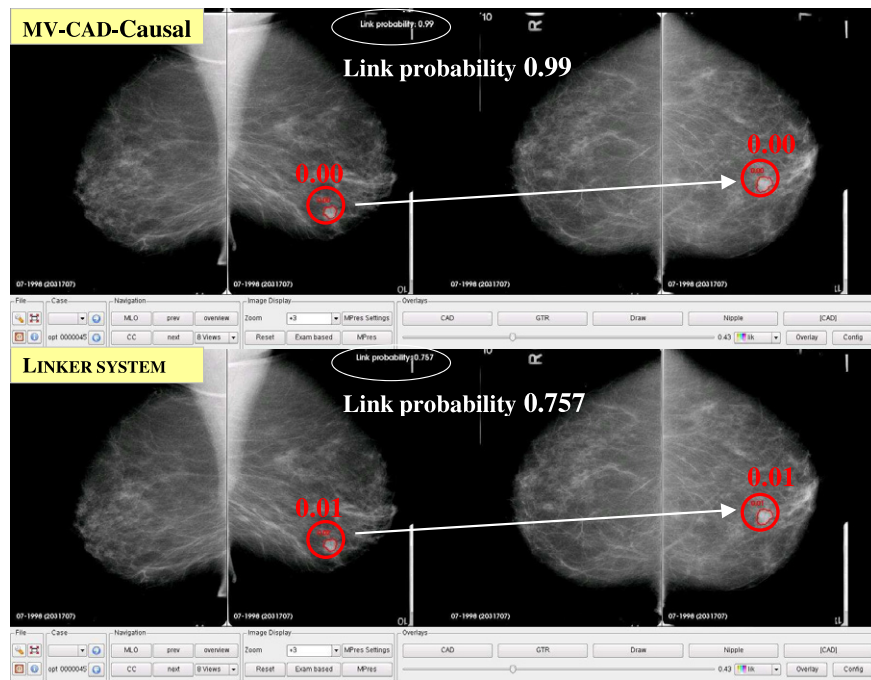
Previous results presented in (Velikova et al., 2009a) showed that building upon the results from the *Linker* system for re-computing the link probabilities in our link model helps improving the classification for the four link types. To demonstrate the advantage of our system making the links symmetric, in Fig. 6 we display

snapshots from the *MV-CAD-Causal* and *Linker* systems with the mammographic examination of one cancerous patient. Fig. 6a shows the results when the user asks the CAD system for the probability of a region in the MLO view and the system's response includes not only the requested probability but also the probability for the most likely link with a region in CC view. The regions are true detections of the cancer in both MLO and CC views, showing the good linking capabilities of both multi-view systems. However, based on the results from the link model (Stage I in Fig. 2) *MV-CAD-Causal* demonstrates a better capability to classify a TPTP link, as the link probability is nearly 1. The improved link probabilities are used in the region model (Stage II) to obtain better region probabilities in the multi-view causal model in comparison to *SV-CAD*. Note that the reported scores in Fig. 6a are the normality scores of the regions. Fig. 6b shows the results for the CC view analysis performed by the user with the help of the CAD systems, where the region probability in the CC view is simultaneously displayed with the link probability and the corresponding region in the MLO view. *MV-CAD-Causal* establishes the same true link in the CC-MLO direction as in the MLO-CC direction, demonstrating the symmetric matching naturally modeled by the multi-view causal method. For the *Linker* system, however, this is not the case – the established CC-MLO link with the highest probability (close to 1) is with a FP region with a very low normality score. This result is, of course, undesirable in practice where the radiologists would expect that the linking is unique.

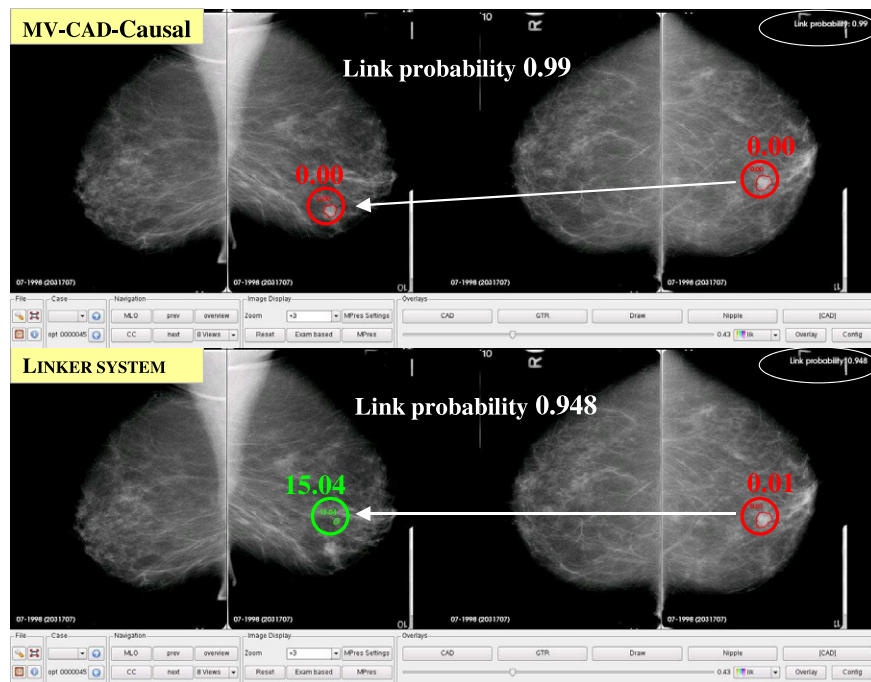
In screening, the most important question eventually is whether or not a patient is suspicious for cancer and needs to be referred for further examination. To answer this question, in this study, we focus on the results obtained from the last stage of our model (*ExamModel*) based on the best three *ObjectModel* (here 'Object' is 'Breast') results with threshold $q = 1, 2, 3$, which is minimal number of cancerous regions required to define a breast as cancerous. For threshold functions with $q \geq 4$, the performance of *ObjectModel* decreases linearly starting from AUC of 0.794 and reaching AUCs of 0.506 for $q = 10$ (this is the maximal number of detected regions as suspicious by *SV-CAD*). This supports our hypothesis that the models with smaller values of q in the screening mammography will perform better, as breast cancer is usually unifocal in the initial stages of detection.

The test exam results are reported in Table 1, including the two-sided p -values and confidence intervals for the statistical tests against the single-view system. We observe overall improvement in the breast cancer detection rate achieved by our multi-view system. Furthermore, we notice that using MAX as a combination function for the breast probabilities leads to a better distinction between cancerous and normal exams than using the logical OR. A possible explanation might be that the latter tends to overestimate the probability of normal exams by considering both breasts, whereas the MAX function seems to be more appropriate given that in screening mammography mostly one of the breasts is cancerous. We also notice the improvement in the exam detection rate achieved by *MV-CAD-LR-LR* is statistically insignificant with respect to *SV-CAD* in this study. This might be explained by the logistic regression modeling at a breast and exam level of *MV-CAD-LR-LR*, applied on a smaller dataset than the one used in (Velikova et al., 2009b), and the use of single-view region features, which are less descriptive for the link modeling.

Next to get a better insight in the improvement of the breast cancer detection rates, we plot the ROC curves for the best *MV-CAD-Causal* (MAX), *MV-CAD-LR-LR* and *SV-CAD* at an exam level; see Fig. 7. It is interesting to observe that *MV-CAD-LR-LR* performs worst in the very lower FP range (<15%), the range that is most relevant in screening, and the best in the upper FP range (>50%). In Table 2, we report the true positive detection rates at



(a) MLO-CC linking



(b) CC-MLO linking

Fig. 6. Linking results for an exam based on MV-CAD-Causal and Linker.

false positive rates of 5% and 10% for the three systems. The results show that the highest increase in the exam true positive detection rate is achieved for MV-CAD-Causal, especially for $q = 1$.

4.3.2. Case selection in screening

An interesting application in the screening setting is to use a CAD system for the selection of the most suspicious cases and present them to the human expert for further analysis. Thus, the selected cases would get special attention and if cancer is present it is more likely that it would be detected. To test the capability of

the proposed causal fusion method, MV-CAD-LR-LR and SV-CAD for this application, we compared the percentage of the most suspicious cancerous cases against various levels at which normal exams are misclassified as cancerous. The results are presented in Table 3.

The results show that the fusion method outperforms the single-view CAD system by providing a higher percentage of true cancer detected exams as most suspicious cases. It is also interesting that this improvement is achieved for the low false positive levels: 8%, 5% and 6% more cancerous detected exams at 2%, 5% and

Table 1

AUCs \pm std. dev. obtained from the multi- and single-view systems at an *exam* level with the corresponding two-sided *p*-values and confidence intervals for the statistical tests against SV-CAD.

Model		<i>q</i>		
		1	2	3
MV-CAD-causal	MAX	0.889 \pm 0.016 <i>p</i> = 0.034 (.002,.048)	0.884 \pm 0.017 <i>p</i> = 0.045 (.000,.041)	0.883 \pm 0.017 <i>p</i> = 0.019 (.004,.043)
	OR	0.879 \pm 0.017 <i>p</i> = 0.249 (-.010,.038)	0.882 \pm 0.017 <i>p</i> = 0.110 (-.004,.041)	0.877 \pm 0.017 <i>p</i> = 0.195 (-.007,.036)
MV-CAD-NB		0.872 \pm 0.017 <i>p</i> = 0.673 (-.024,.037)	0.861 \pm 0.018 <i>p</i> = 0.766 (-.038,.028)	0.797 \pm 0.022 <i>p</i> = 0.001 (-.109,-.029)
MV-CAD-LR		0.861 \pm 0.018 <i>p</i> = 0.745 (-.039,.028)	0.873 \pm 0.017 <i>p</i> = 0.596 (-.021,.037)	0.822 \pm 0.021 <i>p</i> = 0.022 (-.080,-.006)
MV-CAD-LR-LR		0.875 \pm 0.017 <i>p</i> = 0.526 (-.014,.028)		
SV-CAD		0.865 \pm 0.018		

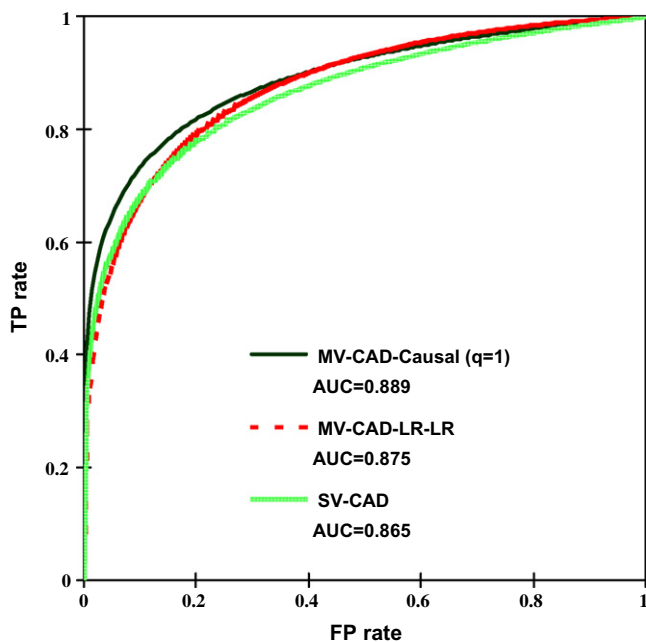


Fig. 7. ROC curves at an *exam* level obtained from MV-CAD-Causal (MAX), MV-CAD-LR-LR and SV-CAD.

Table 2

TP rates corresponding to FP rates of 5% and 10% obtained from MV-CAD-Causal (MAX), MV-CAD-LR-LR and SV-CAD at an *exam* level.

Model		TP rate	
		FPr = 5%	FPr = 10%
MV-CAD- Causal (MAX)	<i>q</i> = 1	0.648	0.731
	<i>q</i> = 2	0.641	0.724
	<i>q</i> = 3	0.621	0.712
		0.563	0.674
MV-CAD-LR-LR		0.585	0.679
SV-CAD			

10% of normal cases, respectively. This is a desired result in screening as the number of normal cases is considerably larger than the

Table 3

Percentage of selected cancerous cases vs. percentage of normal cases classified as highly suspicious by the multi- and single-view systems.

Model	Normal cases				
	2%	5%	10%	15%	50%
MV-CAD-Causal	59%	65%	73%	76%	93%
MV-CAD-LR-LR	39%	64%	73%	76%	92%
SV-CAD	51%	60%	67%	75%	89%

cancerous ones, and thus the specificity needs to be high. Interestingly, MV-CAD-LR-LR tends to have comparable percentages of selected cancerous exams as MV-CAD-Causal except at 2% of normal cases where the former selects 39% of the cancerous cases against 59% selected by the latter method.

4.3.3. Evaluation of the training/testing procedure

Finally, we evaluated the proposed objective training and testing procedure presented in Section 3.3. In particular, we compared for every of the 10 folds, the results obtained from the classifiers built on two subsets of the training data, which are subsequently applied to the other training subset and the test set. The concatenation of the results from the testing on the training subsets results in the validation set for each fold. Table 4 presents the results obtained at an *exam* level from MV-CAD-Causal for *q* = 1 and MV-CAD-LR-LR. For the remaining values of *q* and the other fusion models we observed analogous outcomes. To facilitate the comparison we also report the AUCs from the single-view CAD system per fold although they are not based on our evaluation procedure.

Next to the advantage of being objective and producing a reliable estimate of generalization performance, our evaluation procedure seems to gain in classification performance as well. In majority of cases combining the outcome of two classifiers applied on the test set at every stage improves the prediction accuracy with respect to the single classifiers applied on the validation set. This is a straightforward way for fusing the “opinions” of two experts to get better predictions. In addition we observe a slight tendency of the multi-view causal model to produce the smallest test standard deviation across the folds as reported in the last row of Table 4 indicating robustness. Note that these standard deviations are across the folds and differ from the one reported in Table 1

Table 4

AUCs for the validation and the respective test sets obtained from multi- and single-view systems at an exam level.

Fold	MV-CAD-Causal ($q = 1$)				MV-CAD-LR-LR		SV-CAD	
	MAX		OR		Val	Test	Val	Test
	Val	Test	Val	Test				
1	0.779	0.898	0.768	0.904	0.718	0.900	0.865	0.843
2	0.884	0.924	0.876	0.930	0.801	0.703	0.866	0.737
3	0.783	0.875	0.774	0.869	0.655	0.865	0.865	0.858
4	0.881	0.930	0.873	0.933	0.728	0.948	0.861	0.913
5	0.890	0.897	0.883	0.885	0.761	0.911	0.872	0.864
6	0.889	0.860	0.887	0.847	0.720	0.852	0.867	0.820
7	0.826	0.907	0.820	0.890	0.683	0.939	0.863	0.939
8	0.874	0.859	0.870	0.864	0.697	0.904	0.867	0.887
9	0.870	0.786	0.867	0.764	0.769	0.890	0.878	0.890
10	0.871	0.976	0.866	0.979	0.717	0.892	0.868	0.901
Std. dev	0.043	0.051	0.045	0.058	0.043	0.069	0.005	0.057

where the test results for all exams have first been pulled in one set and then the AUC and standard deviation are computed.

5. Conclusion

In this paper, we presented a unified probabilistic framework for information fusion of complimentary image views. This systematic methodology has all the characteristics one usually finds in information fusion approaches. It is based on the theory of causal independence, providing: (i) a systematic combination of causal information via symmetric Boolean interaction functions; (ii) a natural and easy to understand representation of expert knowledge; and (iii) a way to handle uncertain information in a probabilistic manner. This makes the proposed framework suitable as a general information fusion method that can be extended easily. We demonstrated the application of our method to the domain of automated mammographic analysis, where the information from multiple breast images has been merged to provide an overall probability for a patient having cancer. The results with real mammographic data showed that the causal fusion method outperforms the previous single- and multi-view systems, and the

benchmark probabilistic methods. The results should encourage future development of information fusion methods in the areas of image analysis, certainly in the face of the many different techniques already available (e.g. magnetic resonance imaging, 3D mammography, ultrasound) with undoubtedly more to come. As pointed out in (Collins et al., 2006), the fusion of these modalities in a single CAD system can considerably facilitate physicians in the image interpretation and reduce misdiagnosis. This is one of our current research areas.

Acknowledgment

Funded by the Netherlands Organization for Scientific Research (NWO) under BRICKS/FOCUS Grant No. 642.066.605.

Appendix A. An example of causal independence model with threshold functions

Fig. A.8a depicts an example of a Bayesian network for breast cancer prediction. We have two binary cause variables *mass* (MASS) and *microcalcifications* (MCAL) – the two main mammographic indicators for *breast cancer* (BC). In the terms of BN theory, the former are the *causes*, indicated by the outgoing directed arcs, and the latter is the *effect* variable. The CPT table for BC is also displayed, illustrating how the size of the CPT depends exponentially on the number of causal variables; for example, adding another causal binary variable to this model would lead to 8 (2^3) entries for the CPT of BC.

By introducing the two intermediate variables I_1 and I_2 , as done in Fig. A.8b, such computational complexity can be tackled efficiently. It is known that masses are more frequent occurring sign, which is reflected in the conditional probability distributions $P(I_1|\text{MASS})$ and $P(I_2|\text{MCAL})$. The interaction function $f(I_1, I_2)$ for the effect BC can be defined as an exact (ϵ_q) or a threshold (τ_q) function with a specific threshold q (in our example – 1 or 2).

Suppose that we want to compute the probability of having breast cancer given the states of MASS and MCAL using the threshold function (τ_2), implying that at least 2 trues must occur among its arguments in order for the function to be true, so:

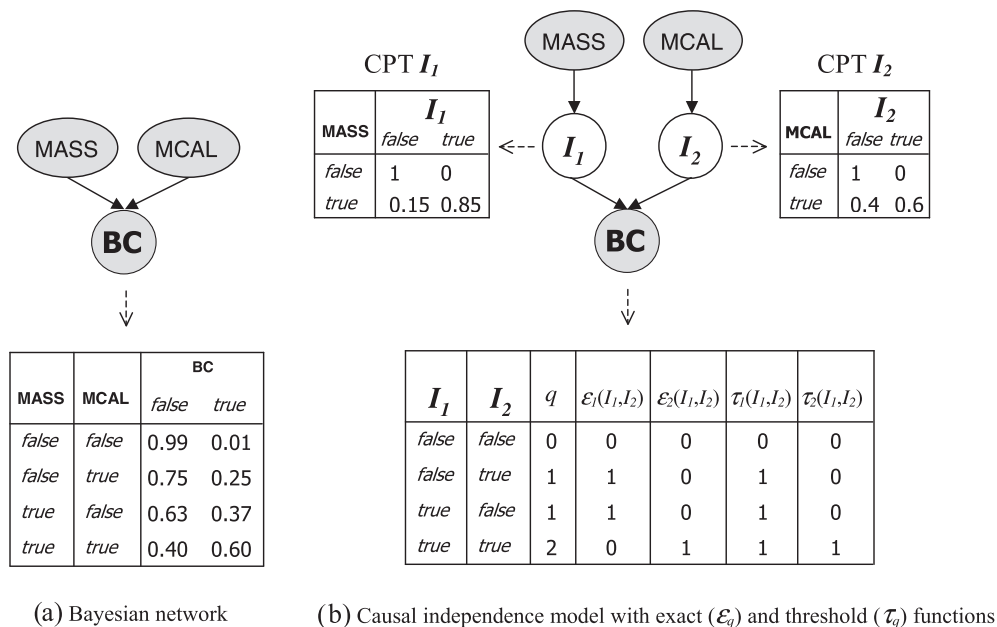


Fig. A.8. Example of two probabilistic models for breast cancer prediction.

$$P_{t_2}(BC = true|MASS, MCAL) = \sum_{2 \leq i \leq 2} \sum_{a_i(I_1, I_2)} P(I_1|MASS)P(I_2|MCAL) \\ = P(I_1 = true|MASS)P(I_2 = true|MCAL).$$

Now suppose that $MASS = true$ and $MCAL = true$. Then,

$$P_{t_2}(BC = true|MASS, MCAL) = 0.85 \cdot 0.6 = 0.51.$$

References

- Boström, H., Andler, S.F., Brohede, M., Johansson, R., Karlsson, A., van Laere, J., Niklasson, L., Nilsson, M., Persson, A., Ziemke, T., 2007. On the definition of information fusion as a field of research. Tech. Rep. HS-IKI-TR-07-006, School of Humanities and Informatics, Univ. of Skövde, Sweden.
- Calhoun, V.D., Adali, T., 2009. Feature-based fusion of medical imaging data. *IEEE Trans. Inf. Technol. Biomed.* 13 (5), 711–720.
- Collins, M., Hoffmeister, J., Worrell, S., 2006. Computer-aided detection and diagnosis of breast cancer. *Semin. Ultrasound CT MRI* 27 (4), 351–355.
- Dempster, A.P., 1967. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* 38 (2), 325–339.
- Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* 29, 103–130.
- Good, W., Zheng, B., Chang, Y., Wang, X., Maitz, G., Gur, D., 1999. Multi-image cad employing features derived from ipsilateral mammographic views. In: *Proc. of SPIE, Med. Imag.* vol. 3661.
- Heckerman, D., Breese, J.S., 1996. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Trans. SMC – A* 26 (6), 826–831.
- Jesneck, J.L., Nolte, L.W., Baker, J.A., Floyd, C.E., Lo, J.Y., 2006. Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis. *Med. Phys.* 33 (8), 2945–2954.
- Lucas, P., van der Gaag, L., 1991. *Principles of Expert Systems*. Addison-Wesley, Wokingham.
- Lucas, P.J.F., 2001. Certainty-factor-like structures in Bayesian belief networks. *Knowl.-Based Syst.* 14 (7), 327–335.
- Metz, C., Wang, P., Kronman, H., 1984. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: *Information Processing in Medical Imaging*. Nijhoff.
- Mnatsakanyan, Z.R., Burkom, H.S., Coberly, J.S., Lombardo, J.S., 2009. Bayesian information fusion networks for biosurveillance applications. *J. Am. Med. Inform. Assoc.* 16 (6), 855–863.
- Moravec, H., 1988. Sensor fusion in certainty grids for mobile robots. *AI Mag.* 9 (2).
- Murphy, K., 2007. Bayesian Network Toolbox (BNT). <<http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>>.
- Niemeijer, M., Abramoff, M.D., van Ginneken, B., 2009. Information fusion for diabetic retinopathy cad in digital color fundus photographs. *IEEE Trans. Med. Imag.* 28, 775–785.
- Paquerault, S., Petrick, N., Chan, H., Sahiner, B., Helvie, M.A., 2002. Improvement of computerized mass detection on mammograms: fusion of two-view information. *Med. Phys.* 29 (2), 238–247.
- Qian, W., Song, D., Lei, M., Sankar, R., Eikman, E., 2007. Computer-aided mass detection based on ipsilateral multiview mammograms. *Acad. Rad.* 14 (5), 530–538.
- Samulski, M., Karssemeijer, N., 2008. Matching mammographic regions in mediolateral oblique and cranio caudal views: a probabilistic approach. In: *Proc. of SPIE, Med. Imag.*, vol. 6915.
- Samulski, M., Karssemeijer, N., 2011. Optimizing case-based detection performance in a multiview CAD system for mammography. *IEEE Trans. Med. Imag.* 30 (4), 1001–1009.
- Shafer, G., 1976. *A Mathematical Theory of Evidence*. Princeton University Press.
- Shortliffe, E., Buchanan, B., 1975. A model of inexact reasoning in medicine. *Math. Biosci.* 23, 351–379.
- van Engeland, S., Karssemeijer, N., 2007. Combining two mammographic projections in a computer aided mass detection method. *Med. Phys.* 34 (3), 898–905.
- van Engeland, S., Timp, S., Karssemeijer, N., 2006. Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views. *Med. Phys.* 33 (9), 3203–3212.
- Velikova, M., Samulski, M., Lucas, P.J.F., Karssemeijer, N., 2009a. Causal probabilistic modelling for two-view mammographic analysis. *Lect. Notes AI* 5651, 395–404.
- Velikova, M., Samulski, M., Lucas, P.J.F., Karssemeijer, N., 2009b. Improved mammographic CAD performance using multi-view information: a Bayesian network framework. *Phys. Med. Biol.* 54, 1131–1147.
- Visscher, S., Lucas, P.J.F., Schurink, C.A.M., Bonten, M.J.M., 2009. Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions. *Artif. Intel. Med.* 46, 251–266.
- Wei, J., Chan, H., Sahiner, B., Zhou, C., Hadjiiski, L.M., Roubidoux, M.A., Helvie, M.A., 2009. Computer-aided detection of breast masses on mammograms: dual system approach with two-view analysis. *Med. Phys.* 36 (10), 4451–4460.
- Zhang, Y., Ji, Q., 2005. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. PAMI* 27 (5), 699–714.
- Zheng, B., Leader, J.K., Abrams, G.S., Lu, A.H., Wallace, L.P., Maitz, G.S., Gur, D., 2006. Multiview-based computer-aided detection scheme for breast masses. *Med. Phys.* 33 (9), 3135–3143.