

Medical images modality classification using discrete Bayesian Networks

Jacinto Arias^a, Jesus Martínez-Gómez^{a,b}, Jose A. Gámez^a, Alba G. Seco de Herrera^c, Henning Müller^c

^a*University of Castilla-La Mancha, Spain*

^b*University of Alicante, Spain*

^c*University of Applied Sciences Western Switzerland, Switzerland*

Abstract

In this paper we propose a complete pipeline for medical image modality classification focused on the application of discrete Bayesian network classifiers. Modality refers to the categorization of biomedical images from the literature according to a previously defined set of image types, such as X-ray, graph or gene sequence. We describe an extensive pipeline starting with feature extraction from images, data combination, pre-processing and a range of different classification techniques and models. We study the expressive power of several image descriptors along with supervised discretization and feature selection to show the performance of discrete Bayesian networks compared to the usual deterministic classifier used in image classification. We perform an exhaustive experimentation by using the ImageCLEFmed 2013 collection. This problem presents a high number of classes so we propose several hierarchical approaches. In a first set of experiments we evaluate a wide range of parameters for our pipeline along with several classification models. Finally, we perform a comparison by setting up the competition environment between our selected approaches and the best ones of the original competition. Results show that the Bayesian Network classifiers obtain very competitive results. Furthermore, the proposed approach is stable and it can be applied to other problems that present inherent hierarchical structures of classes.

Keywords: Medical Image Analysis, Visual Features Extraction, Bayesian Networks, Hierarchical Classification

1. Introduction

Medical images are essential for diagnosis and treatment planning. These types of images are produced in ever-increasing quantities and varieties [1]. A

Email addresses: Jacinto.Arias@uclm.es (Jacinto Arias), Jesus.Martinez@uclm.es (Jesus Martínez-Gómez), Jose.Gamez@uclm.es (Jose A. Gámez), alba.garcia@hevs.ch (Alba G. Seco de Herrera), henning.mueller@hevs.ch (Henning Müller)

recent European report estimates medical images of all kind occupied 30% of the global digital storage in 2010 [2].

Clinicians use images of past cases in comparison with current images to determine the diagnosis and potential treatment options of new patients. Images are also used in teaching and research [3, 4]. Thus, the goal of a clinician is often to solve a new problem by making use of previous similar cases/images together with contextual information, by reusing information and knowledge [5].

Systematic and quantitative evaluation activities using shared tasks on shared resources have been instrumental in contributing to the success of information retrieval as a research field and as an application area in the past few decades. Evaluation campaigns have enabled the reproducible and comparative evaluation of new approaches, algorithms, theories and models through the use of standardized resources and common evaluation methodologies within regular and systematized evaluation cycles. The tasks organized over the years by ImageCLEF¹ [6] have provided an evaluation forum and framework for evaluating the state of the art in biomedical image retrieval.

In this article, we evaluate a probabilistic approach for medical image classification by using the ImageCLEF 2013 medical image modality classification task [7] as a benchmarking environment. We propose a reproducible methodology based on the use of discrete Bayesian Network Classifiers (BNCs) [8]. Our approach first explores details of the modality classification and medical image processing. This is done by defining a basic pipeline in which the application of BNCs is straightforward. The defined pipeline includes a problem transformation from its usual continuous domain to a discrete one that is more natural for probabilistic approaches. In the classification stage we dive into the problem of modality classification to evaluate several proposals in order to deal with the relatively large number of classes (31) that the task presents; for this we focus on hierarchical classification and multi-class approaches. Fig. 1 shows a scheme with all the stages taking part in the described pipeline for the classifier generation process.

Our proposal is based on an extensive experimental evaluation of the proposed pipeline using a collection of probabilistic instead of deterministic classifiers that are the common approaches to tackle this kind of problem [9, 10]. The BN models selected for the experimental analysis show an advantage in their trade-off between efficiency and quality. They also demonstrate a promising performance in a wide range of domains such as document classification [11], object detection [12] or semantic localization [13].

Moreover, probabilistic graphical models are suitable for the integration of contextual categorical variables in conjunction with descriptors extracted from computer vision techniques. Contextual information, such as human annotations or semantic attributes, are becoming more frequent and they can be obtained automatically by means of external tools like Amazon mechanical turk [14]. This kind of information cannot be directly incorporated in other

¹<http://imageclef.org/>

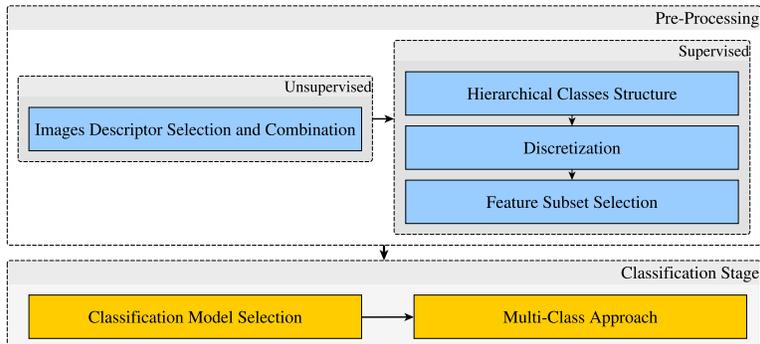


Figure 1: Classifier Generation Pipeline.

traditionally used methods while maintaining its descriptive properties; black box designed algorithms such as SVMs or Neural Networks are not suitable to properly represent dependency relation or other qualitative information in the data.

Regarding the evaluation of our proposed technique, we conducted experiments by following a competition scheme using the ImageCLEF 2013 training and test sets of images separately, for model selection and evaluation respectively. The purpose of this decision is twofold: first of all, to compare the proposed approaches with the best systems participating in the competition by using only the training subset of images for model selection, and to evaluate the most promising approaches on the test data; second, we use the original ImageCLEFmed training and test sets from the competition to keep the original challenge difficulty, allowing our model selection to remain unbiased and to avoid overfitting. To this extent, we did not perform any parameter tuning, thus leaving our basic approach using only baseline models and strategies as the goal of this paper is to show the competitiveness of BNC over the standard approaches used for this problem and not to adjust the model to this exact benchmark. The final results show that our approach would have ranked 3rd in the competition (where 7 groups sent their results of a variety of techniques) being a promising result as it is a non-specific approach that leaves room to several improvements as it will be discussed at the end of the paper. The better results also used extended training sets, which was not used in the work described here.

The rest of the article is organized as follows. We first review the related work in the field of medical image classification in Section 2. Sections 3 and 4 describe the background on computing visual features and Bayesian Network classifiers. The pre-processing of the data is then discussed in Section 5 and the different strategies for hierarchical and multi-class classification are introduced in Section 6. Finally, an extensive experimentation is conducted and analyzed in Section 7 and in Section 8 we summarize the obtained results and obtain a

brief perspective of future work and extensions.

2. Background and Related work

The interest of the visual retrieval community in the automatic analysis of medical information was motivated in part thanks to the medical ImageCLEF challenge. The medical task has run at ImageCLEF since 2004, with many changes between different editions [6]. The underlying objective of this challenge is the retrieval of similar images to fulfill a precise information need and image classification. The 2009 edition of the task [15] focused on the retrieval of articles from the biomedical literature that might best suit a provided medical case description (including images). In 2010 [16], the organizers introduced the modality classification task.

The goal of modality classification is to classify the images of the literature into medical modalities and other image types, such as Computer Tomography (CT), X-ray or general graphs. Medical image classification by modality can improve the retrieval step by filtering or re-ranking the results lists [17]. Moreover, it can reduce the search space to a set of relevant categories improving the speed and precision of the retrieval [18]. Some example images used in the modality classification task are shown in Fig. 2.

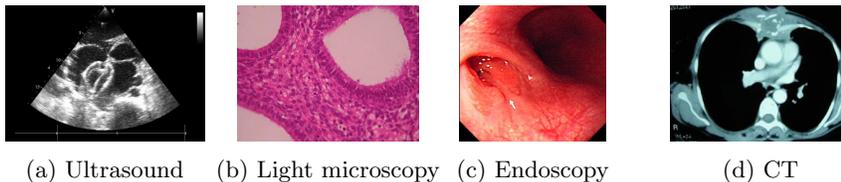


Figure 2: Examples of medical images from various modalities.

The modality classification task was maintained until 2013 [6]. The image collection provided in 2013 was used for the evaluation in this article. This collection includes 2896 annotated training images and 2582 test images to be classified. Each image belongs to one of the 31 classes that present an intrinsic hierarchy shown in Fig. 3.

The techniques presented, rely mainly on two key stages: a.- extraction of visual features from the images, and b.- generation of classification models. Regarding the feature extraction, several studies have shown that the modality can be extracted from the visual content using only visual features [19, 20]. The extracted visual features must then be transformed to obtain meaningful descriptors. For instance, Kitanovski et al. [9] use a spatial pyramid pooling in combination with dense sampling using an opponentSIFT descriptor for each image patch.

Support Vector Machines (SVMs), in combination with the χ^2 kernel, are the most common classification model used in the competition [21, 9, 10]. However, other deterministic classifiers such as k -Nearest Neighbor [22] were also

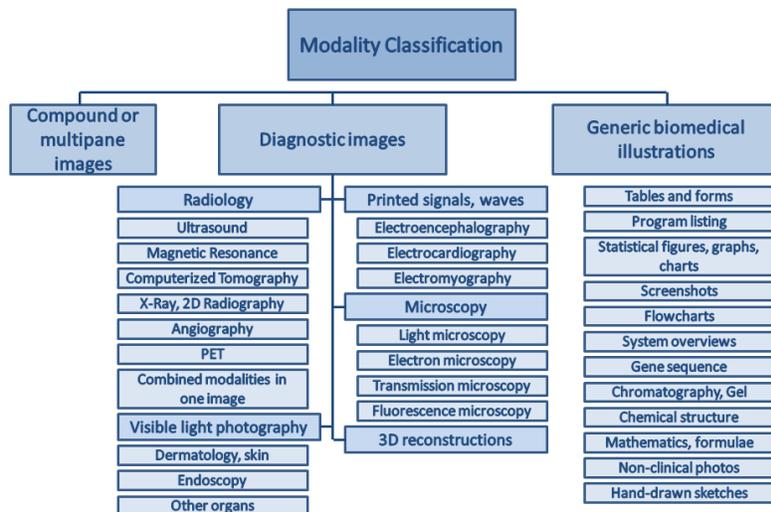


Figure 3: The image class hierarchy provided by ImageCLEFmed for document images occurring in the biomedical open access literature.

used. Despite their lack of descriptive capabilities, these classification models present properties that have encouraged their use in image classification problems. They can work with numeric input data, which is the most common output from feature extraction techniques. They can properly cope with the high dimensionality of the image descriptors.

The hierarchical relationship between the categories in the medical task can also be found in other problems, such as object categorization [23]. In this problem, we can find solutions where SVM classifiers are applied over the combination of different 3D descriptors [24], as well as hierarchical decomposition of the descriptors [25] but the explicit management of the object hierarchy is very seldom adopted.

Though different Bayesian and non-Bayesian probabilistic methods have been applied for medical image analysis (see e.g. [26]), the use of BNCs has been scarce to the best of our knowledge. None of the participants of the ImageCLEFmed modality classification task presented approaches based on BNCs. One of the main reasons could be the continuous domain of the features used in medical image classification, while the developments for BNCs have been mainly devoted to the discrete case. In fact, some BN models are available to deal with numerical variables, but they have two major shortcomings: the Gaussian assumption and structural constraints, e.g. a discrete variable cannot be conditioned on a numerical one. Nevertheless, some approaches to medical image analysis problems have been carried out by using discrete BN [27], al-

though they reduce to the use of Naive Bayes and Tree Augmented Naive Bayes (TAN) algorithms.

3. Feature extraction and descriptor generation

The transformation of an input image into a set of features that describe it is a key stage for a subsequent classification task. This process is known as feature extraction and can be accomplished in different ways, as it is discussed in [28]. In this paper, we follow the scheme proposed in [29], where a combination of multiple low-level visual features is explored. In this paper, the same combination of visual descriptors is applied. Therefore, the descriptors used are the following:

- *Bag of Visual Words (BoVW)* using Scale Invariant Feature Transform (SIFT) (BoVW-SIFT) [30] – Each image is represented by an histogram symbolizing a set of local descriptors represented in visual words from a vocabulary previously learned with 238 visual words. This leads to a 238 bin histogram;
- *Bag of Colors (BoC)* [31] – Each image is represented by a 100 bin histogram symbolizing the colors from a vocabulary previously learned;
- *Color and Edge Directivity Descriptor (CEDD)* [32] – Color and texture information is produced by a 144 bin histogram. Only little computational power is required for its extraction;
- *Fuzzy Color and Texture Histogram (FCTH)* [33] – This descriptor contains results from the combination of 3 fuzzy systems including color and texture information in 192 bin histogram;
- *Fuzzy Color Histogram (FCH)* [34] – The color similarity of each pixel's color associated with a 192 histogram bin through a fuzzy-set membership function is used;

These descriptors are extracted using the ParaDISE (Parallel Distributed Image Search Engine) [35]. The combination of all these features generates descriptors with a dimensionality of 866.

4. Bayesian Network classifiers

Bayesian Networks (BNs) [36] are one of the most frequently used knowledge representation techniques when dealing with uncertainty, mostly owing to their predictive/descriptive capabilities. They are based on sound mathematical principles and, as a probabilistic graphical model, they output a graphical structure that provides an interpretative representation of the relationships between the variables of the problem.

Learning general BNs is known to be a complex problem [37] involving the task of structural learning as well as parameter estimation [38]. As learning general BNs is usually problematic, this has led to the definition and wide usage of specific models that are explicitly designed to tackle the standard classification problem, these are commonly known as Bayesian Network Classifiers (BNCs) [39, 8].

The simplest BNC model is the Naive Bayes classifier (NB) that avoids structural learning by assuming that all attributes are conditionally independent given the value of the class. Although this independence assumption can be considered too strong for some domains, the NB classifier has shown very good results in many real applications such as computing, marketing and medicine [40]. Its results can be improved by slightly alleviating its independence assumption and using a more complex graphical structure. The techniques using this principle are known as semi-naive Bayesian network classifiers, some of them are among the most competitive classification techniques.

We evaluated the performance of different semi-naive BNC models to solve the proposed classification problem, specifically: NB, TAN, K-Dependence Bayesian classifiers (KDB) and Average One Dependence Estimators (AODE).

4.1. Naive Bayes

NB classifier [41] is the simplest BNC, due to its independence assumption which avoids any needs for structural learning. This classifier uses a fixed graphical structure in which all predictive attributes are considered independent given the class, as it is depicted in Fig. 4a. This implies the following factorization: $\forall c \in \Omega_C p(\vec{e}|c) = \prod_{i=1}^n p(a_i|c)$. Here, the maximum a posteriori (MAP) hypothesis is used to classify:

$$c_{MAP} = \underset{c \in \Omega_C}{\operatorname{argmax}} p(c|\vec{e}) = \underset{c \in \Omega_C}{\operatorname{argmax}} \left(p(c) \prod_{i=1}^n p(a_i|c) \right). \quad (1)$$

4.2. Tree-Augmented Naive Bayes

The TAN classifier [39] can be considered a structural augmentation of the NB classifier in which the conditional independence assumption is relaxed by allowing a restricted number of relationships between the predictive attributes. This strategy implies that a structural learning process must be performed. However, TAN can still obtain competitive learning times with moderate datasets establishing a good trade-off between model complexity and model accuracy. In particular, every predictive attribute is allowed to have an extra parent in the model in addition to the class. In order to learn these dependencies a Maximum Weighted Spanning tree (MWST) is learned by using the Chow-Liu algorithm [42] with the conditional mutual information between each pair of attributes and the class as metric to measure each arc weight:

$$MI(A_i, A_l | C) = \sum_{r=1} p(c_r) \sum_{i=1} \sum_{j=1} p(a_i, a_j | c_r) \log \frac{p(a_i, a_j | c_r)}{p(a_i | c_r)p(a_j | c_r)} \quad (2)$$

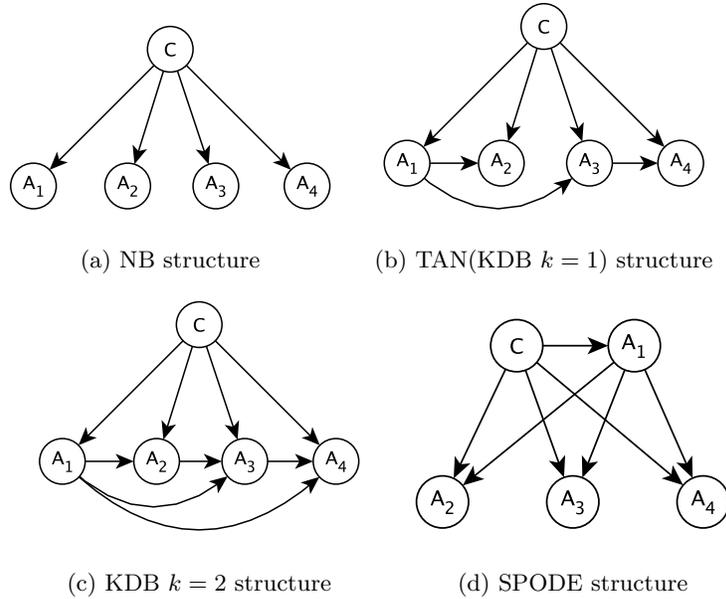


Figure 4: Graphical structure of semi-naive Bayesian Network Classifiers.

This process guarantees that the tree learned from the training data is optimal, i.e. it is the best possible probabilistic representation from the available data as a tree as it maximizes the log likelihood. Once the tree is obtained, an arbitrary node is selected as the root of the tree and the edges are oriented to create a directed acyclic graph in which all attributes are conditioned to the class, creating the final BNC structure of a TAN classifier. An example is shown in Fig. 4b.

4.3. K -Dependence Bayesian classifiers

The basic KDB classifier is based on the notion of a k -dependence estimator introduced by Sahami [43] in which the structure of a basic NB classifier can be augmented by allowing an attribute to be conditioned to a maximum of k parent attributes in addition to the class, thus covering the full spectrum from the NB classifier to a general full BN structure by varying the parameter k . An example for a given value of k is shown in Fig. 4c.

The KDB classifier itself performs a three-stage learning process:

1. A ranking is established between the predictive attributes by means of their mutual information with the class variable.
2. For each attribute A_i , i being its position on the previous ranking, the k attributes taken from $\{A_1, \dots, A_{i-1}\}$ with the highest conditional mutual information $MI(\cdot, A_i | C)$ are set as the parents of A_i .

3. The class variable C is added as a parent for all the predictive attributes.

This classifier presents a more flexible approach to the TAN classifier (in fact, the TAN classifier is a particular case of KBD with $K = 1$), as it is capable of adjusting the mentioned trade-off between model complexity and model quality.

4.4. Averaged One Dependence Estimators

AODE [44] are an alternative to other semi-naive BNC approaches. They present a fixed structure model that avoids structural learning, improving its efficiency when compared to other BNCs that require the step. Moreover, AODE maintains very competitive model quality.

Therefore, AODE is restricted exclusively to 1-dependence estimators. This classifier can be seen as an ensemble of models, concretely, it considers each model belonging to a specific family of classifiers (known as SPODEs (Superparent One-Dependence Estimators)) in which every attribute is dependent on the class and on another shared attribute, designated as superparent. The structure of a specific SPODE is depicted in Fig. 4d.

For classification, AODE computes the average of the n possible SPODE classifiers:

$$C_{MAP} = \arg \max_{c \in \Omega_C} \left(\sum_{j=1, N(a_j > q)}^n p(c) \cdot p(a_j | c) \prod_{i=1, i \neq j}^n p(a_i | c, a_j) \right) \quad (3)$$

5. Data pre-processing

Input data can be pre-processed to meet the requirements of the classification models, to reduce their complexity but also to increase their performance. Here, we enumerate three data pre-processing techniques. First, we propose a combination of the different descriptors extracted from the image. Then, we discretize the input data to obtain nominal variables suitable for their use in the classification models. Finally, we select a subset of variables from the whole set.

5.1. Descriptor combination

In this article, we opted to use five descriptors generated from visual features extracted from the input images: BoVW-SIFT, CEDD, FCTH, FCH, BoC (see Section 3). Every descriptor consists of numeric variables with dimensionality between 100 and 238, which represents visual feature frequencies (see Sec. 3 for more details on the visual features). They can be concatenated to create a single descriptor, as it is commonly done when working with SVMs. However, we can also follow an aggregation approach to merge descriptors in a recursive way, where we can include pruning strategies. From an initial set of n descriptors, we can generate the following number of combinations,

$$\sum_{i=1}^n \frac{n!}{i!(n-i)!}$$

which would result in 31 different combinations from an initial set of 5 descriptors (5+10+10+5+1).

5.2. Discretization

Some of the internal variables of the image descriptors contain discriminant information but others can be useless for the problem we are facing. While Bayesian classifiers exist tackling numeric variables [45] (e.g. Gaussian Naive Bayes), we opted for the exclusive use of discrete classification models. The main reason for this decision is that some numeric versions of the classifiers assume that input data can be modeled with uni-modal distributions [46], which is not always true. Here, we propose the use of the Fayyad Irani discretization method [47], which can be considered a standard approach. This discretization method takes into account the class information when selecting the number of bins and breaking points by using mutual information. Moreover, it selects the optimal number of bins separately for each input variable. As seen in the experimentation (see Section 7), this discretization step produces a number of binary partitions. It also discretizes input variables into a single bin, which means that the variable has no discriminant power with respect to the class, thus resulting in useless variables that are removed in subsequent steps.

5.3. Feature Subset Selection

The number of input variables, which comes from the descriptors combination, can be successfully reduced by following an appropriate procedure. In addition to data reduction, this step can also increase the accuracy of the classification model by finding redundant or irrelevant input variables. Moreover, using fewer variables also provides non-overfitted and more interpretable classification models, which requires shorter training times. In order to avoid the bias introduced by the classification method while using wrapper approaches, we selected a filter strategy. We opted for the Correlation Feature Selection (CFS [48]), which has shown its value in several scenarios.

6. Strategies

In addition to the classical data pre-processing (discretization and feature subset selection, which only affects the predictive attributes), we can also take advantage from strategies that cope with multi-class problems. The first alternative consists of partitioning the original problem into recursive sub-problems using a hierarchy of classes. The second strategy splits the multi-class problem into n binary problems with decisions being merged to select the final decision. This second strategy was discarded in a first round of preliminary experiments where no significant improvements were obtained when using several multi-class approaches, such as One-versus-All (OvA) [49]. The improvements of OvA were limited to the absence of hierarchical approaches due to the large number of classes (31).

6.1. Hierarchical classification

The intrinsic relationships between class values in a multi-class problem provides us with relevant information that is commonly discarded. However, this information can be used to create a multi-layer classification scheme [50] where each level corresponds to a degree in the hierarchy of the classes. This involves the generation of more classification models but the training stage in second (and subsequent) layers is performed from a set with a smaller number of instances.

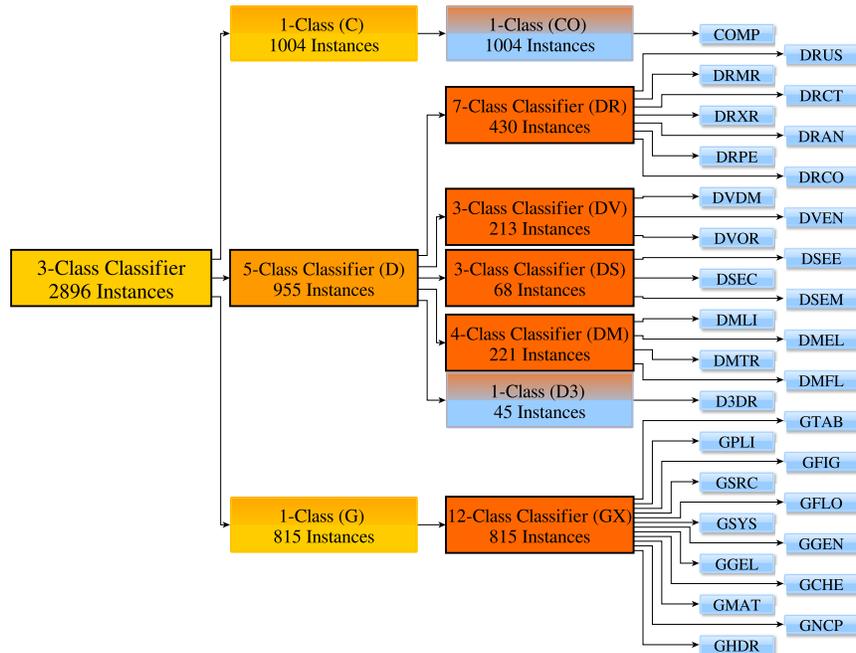


Figure 5: Overall class hierarchy and instance distribution using a 3-level approach

Fig. 5 shows a 3-level hierarchy generated from the inherent relationships between the classes of the ImageCLEF 2013 medical classification task. This figure also includes the number of training instances that affect the generation for each one of the seven classifiers (1-class classifiers are not considered) when applying a three layer hierarchy. As it can be observed, classifiers from levels 2 and 3 are generated from a lower number of instances than those from level 1. In addition to the 3-level hierarchy shown in Fig. 5, we can adopt either a 1-level approach (no hierarchy) or a 2-level one, which generates one single or eight different classifiers respectively.

7. Results

7.1. Parameter Selection

The first round of experiments was conducted with two main objectives: validating the proposed methodology, and performing an initial selection from the set of internal parameters. The experiments were carried out using the training set from the ImageCLEF 2013 medical classification task, which includes 2896 images and 31 nominal classes by performing a 5 fold cross-validation.

7.1.1. Descriptor Combination

There are 31 descriptor combinations as result of merging the 5 initial sets of descriptors: SIFT(D_1), BoC(D_2), CEDD (D_3), FCTH(D_4), and FCH (D_5). Each combination was evaluated using the following options:

- Hierarchical Classification (see Fig. 5)
 - 1-level hierarchy (no hierarchy)
 - 2-level hierarchy
 - 3-level hierarchy
- Discretization
 - Fayyad-Irani
- Feature Subset Selection
 - Correlation Feature Selection (CFS)
- Classification Model / Algorithm
 - Naive Bayes (NB)
 - Tree Augmented Naive Bayes (TAN)
 - Average One Dependence Estimators (AODE)
 - K-Dependence Bayesian Classifier (KDB)

By combining all these options for the described parameters, we generated and tested a total of 465 classifiers. Results are graphically presented in Fig. 6 and detailed in Table A.4. Results show a clear difference in performance between the combinations of descriptors where higher accuracy is obtained from larger descriptor combination. This result implies that integrating descriptors increases the expressive power of the classifier and that the proposed pre-processing pipeline is effective in practice, as the feature subset selection (FSS) process copes properly with an increasing number of input attributes, retaining useful information for each one of the combined descriptors. Fig. 7 shows the evolution of the number of attributes in the final combination when we increase the size of the combined descriptors by adding new ones; the figure

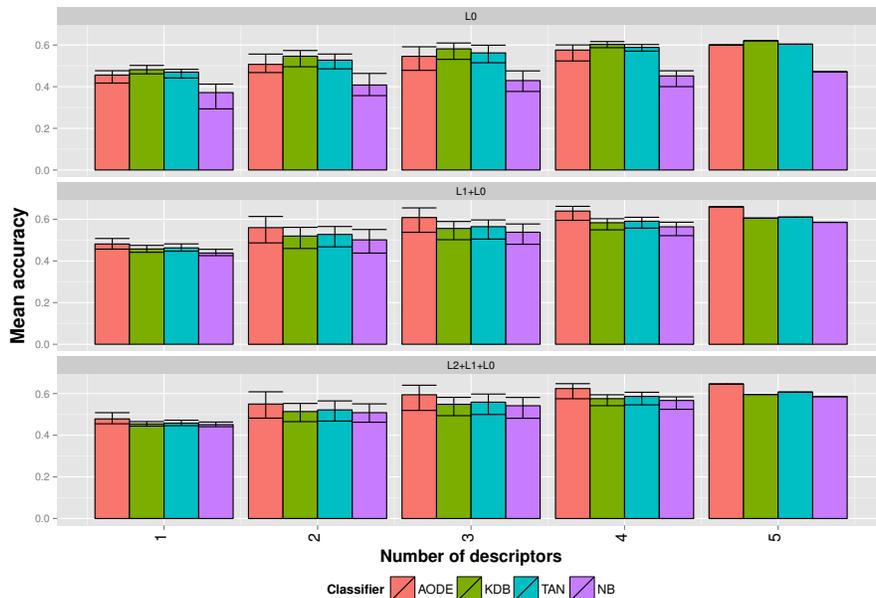


Figure 6: Preliminary results for each hierarchical approach, classifier and descriptor combination using a 5-fold cross validation over the full training set. The results are summarised as the mean accuracy for all the different combination of the same number of descriptors. The error bars correspond to the maximum and minimum values for such an experiment.

exposes how a combination of a larger number of descriptors does not necessarily involves higher dimensionality once FSS is applied, in fact we can observe an asymptotic behavior when the number of descriptors in the combination is increased. Given these results, we decided to continue the experimentation by selecting the entire combination of descriptors $D_{1,2,3,4,5}$ as the candidate.

Regarding the classification models and the proposed hierarchies, results from the Table A.4 show that the usage of a hierarchy improves upon the results when compared to dealing with the 31 original classes using a single classifier. If we compare the models we can observe a better performance of AODE when performing hierarchical classification, however KDB is superior when dealing with the flat class structure (31 classes). To analyze these results we performed a statistical evaluation on the performance of each model for each one of the hierarchy approaches. The Friedman test [51] with a 0.05 confidence level rejects the three hypotheses of all classifier being equivalent when using any of the hierarchy levels. A further post-hoc statistical analysis using the Holm procedure is included in Table 1, in which we compare the performance of all classifiers with the best one for each one of the hierarchies. This confirms how KDB clearly outperforms the other classification models within a 1-level hierarchy. However, AODE stood out when taking advantage of the 2 or 3 levels of the hierarchy. This exposes KDB as the optimal classifier when coping with a large number of

classes, while AODE performs notoriously better in the remaining scenarios.

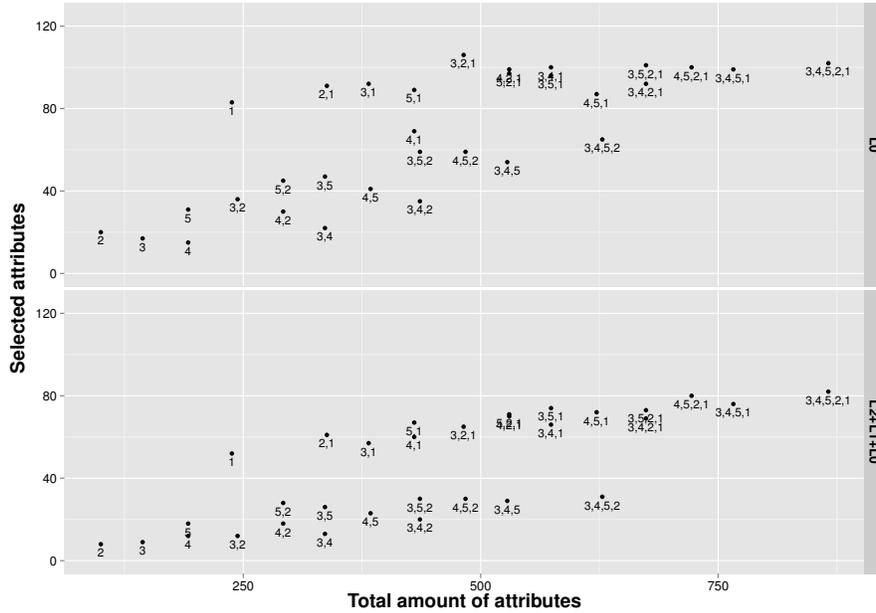


Figure 7: Effect on the FSS procedure for two different hierarchical approaches when using different combinations of descriptors.

Table 1: Classification model ranking. w/t/l denote the number of scenarios the first ranking method wins(w) ties(t) and loses(l) the evaluated method

method	rank	pvalue	w/t/l	method	rank	pvalue	w/t/l	method	rank	pvalue	w/t/l
KDB	1.03			AODE	1.00			AODE	1.00		
TAN	2.26	<0.01	31/0/0	TAN	2.03	<0.01	31/0/0	TAN	2.13	<0.01	31/0/0
AODE	2.71	<0.01	30/0/1	KDB	2.97	<0.01	31/0/0	KDB	3.16	<0.01	31/0/0
NB	4.00	<0.01	31/0/0	NB	4.00	<0.01	31/0/0	NB	3.71	<0.01	31/0/0
1-level hierarchy				2-level hierarchy				3-level hierarchy			

7.2. Final Results

The last experimental stage was carried out following the procedure proposed in the ImageCLEF 2013 medical classification task. Namely, we learned classification systems using the 2896 training images and evaluated such systems against the 2582 test images. In order to perform a fair evaluation we followed the original rules of the competition. We selected the best 10 classification systems that achieved highest results in previous experiments using only the training data, i.e., we evaluated the selected models over the provided the test set and computing the accuracy of our systems. We used the ten best results because this was the limit in the number of task submissions. The internal

parameters for the systems, as well as the accuracy obtained both against the training set that was used to rank the models (with 5-folds cross-validation) and test sequences are shown in Table 2.

Table 2: Final results obtained against the test set with the 10 configurations that ranked first in previous experiments.

Ranking (training)	Parameters			Accuracy	
	Classifier	Hierarchy	Descriptor	Training (5-cv)	Test
1st	AODE	L1+L0	$D_{1,2,4,5}$	0.6622	0.6840
2nd	AODE	L1+L0	$D_{1,2,3,4,5}$	0.6600	0.6774
3rd	AODE	L1+L0	$D_{1,2,3,5}$	0.6569	0.6739
4th	AODE	L1+L0	$D_{1,2,5}$	0.6547	0.6700
5th	AODE	L2+L1+L0	$D_{1,2,4,5}$	0.6474	0.6921
6th	AODE	L2+L1+L0	$D_{1,2,3,4,5}$	0.6455	0.6905
7th	AODE	L1+L0	$D_{1,3,4,5}$	0.6433	0.6642
8th	AODE	L2+L1+L0	$D_{1,2,3,5}$	0.6419	0.6863
9th	AODE	L1+L0	$D_{1,2,3,4}$	0.6401	0.6607
10th	AODE	L2+L1+L0	$D_{1,2,5}$	0.6400	0.6789

If we compare the training and test columns in Table 2, we can observe that we are obtaining higher results over the test set. This trend supports the stability of our methodology, allowing the generation of non-overfitted classification systems. All the evaluated systems share the following internal parameters: AODE as classification model, 2 or 3 levels of hierarchy and large combinations of initial descriptors. As previously pointed out, the multi-class approach did not obtain significant improvements on the accuracy once the hierarchy is applied. An important detail is that the best result for the training set is obtained by using the 2 level hierarchy whereas the best results for the test set corresponds to those approaches using the 3 level hierarchy. This is probably due to the small size of the training set compared to the size of the test set, which supports the generalization capabilities of the proposed framework.

7.3. ImageCLEF 2013 medical task results

Our maximum accuracy obtained over the test set was 69.21%. This result ranked 3rd in the modality classification task of ImageCLEF 2013, as shown in Table 3. We can observe how most of the teams used Support Vector Machines (SVMs) as classifier, while none of them used a probabilistic model. Apart from SVMs, only the K-Nearest Neighbor (K-NN) and the Stochastic gradient descent (SGD) were used. With respect to the image descriptors used by the participants, most were based on histogram representations or bag of words approaches managing continuous information, but other more complex techniques (such as the spatial pyramid [52]) were also used. The best results both used a training set expansion, which can explain part of the gains. On the contrary, our proposal relies on just the training provided by the task organizers, which

is not biased nor optimized to artificially increase the accuracy at the expense of lack of generalization.

Table 3: Highest accuracies (by group) in the ImageCLEF 2013 modality classification task including our best result.

Ranking	Group Name	Classifier	Accuracy
1st	IBM [53]	SVMs	80.79
2nd	FCSE [9]	SVMs	77.14
3rd	Our proposal	BNs(AODE)	69.21
4th	MiiLab [10]	SVMs,K-NN	66.46
5th	medGIFT [29]	K-NN	63.78
6th	ITI [54]	SVMs	61.50
7th	CITI [55]	SGD	56.62
8th	IPL [56]	SVMs	52.05

7.4. Analysis of the results

We carried a posterior analysis of the results to identify the points suitable for improvements of our methodology. The first study examined the accuracy of the internal classifiers involved in the hierarchical classification. We evaluated the system that ranked 5th in Table 2 which is best in the test set. It uses a 3-level hierarchy with 7 internal classifiers (see Fig. 5). For each of these seven classifiers, we computed two accuracy measures: Acc(a).- the overall accuracy, and Acc(b).- the accuracy obtained only with instances that were correctly classified in upper levels. Acc(b) reflects the isolated behavior of each classifier, while Acc(a) is affected by the rest of the system.

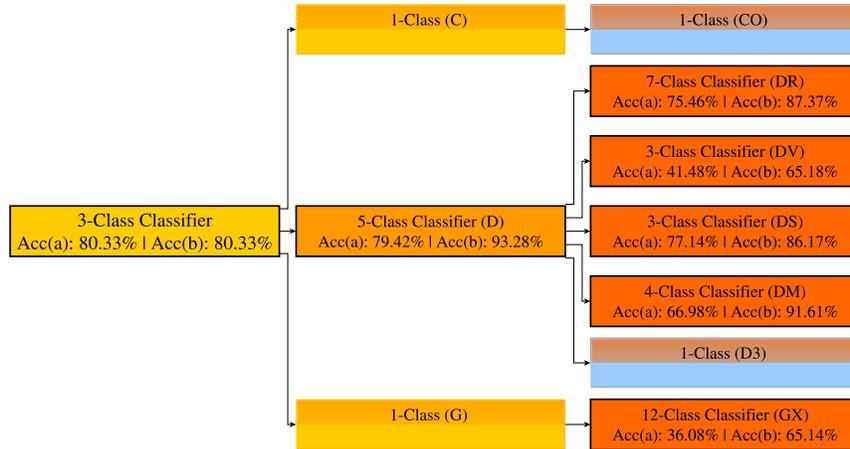


Figure 8: Acc(a) and Acc(b) computed for a classification system including $D_{1,2,3,4,5}$ descriptors, 3-hierarchy levels and AODE as classification model without multi-class approach.

The accuracy from the first level classifier (80.33% Fig. 8) shows how nearly 20% of the test instances were wrongly classified at this level. The classifier is crucial because it processes the full test sequence and its errors are therefore propagated through the hierarchy structure. Concretely, 62.46% of the classification error came from the first classifier, while the remaining 37.54% came from the other 6 classifiers.

8. Conclusions

We propose a pipeline for modality classification of medical images by using probabilistic classifiers, namely BNCs. We have identified an extensive descriptor set, a combination of descriptors, a detailed pre-processing scheme and several approaches for hierarchical and multi-class classification. We evaluated a large number of parameter combinations by using a selected range of the most popular BNCs.

Evaluation was carried out on the ImageCLEFmed 2013 collection. The AODE classifier shows superior results when combined with hierarchical classification and a large number of combined descriptors over the training set. For model selection we replicated the competition conditions by using the test set of images to compare our proposal with the best results in the competition. We obtain results that rank 3rd. From this analysis we can draw useful conclusions:

- Descriptor combinations have proven to be an expressive tool showing also the robustness of discrete supervised preprocessing techniques such as MDL discretization and feature selection. The hierarchical approach proved to be an excellent pairing with these methods, as they can be replicated easily for the hierarchy levels.
- Among all the BNCs evaluated, ensemble methods such as AODE prove to obtain highest discrimination power and thus overall best classification results. The best results are obtained when a deeper hierarchy and a larger number are combined.
- The above points suggests that, AODE, being a low-bias learner can be a suitable candidate to tackle these problems. To contrast this we performed an additional experiment in which we evaluate the models when the training set is increased incrementally. We used the best parameter combination mentioned above and the training/test split modifying the size of the training set by means of 10% random sample partitions. The results are shown in Fig. A.9 where one can clearly observe the superior behaviour of AODE as well as a tendency to improve its results at the presence of additional data.

We believe that this is a positive result for probabilistic classifiers, as this methodology is not the most popular to be applied for solving these kind of problems. Furthermore, our selected models were not tuned or adjusted to optimize the results for the competition dataset. This means that usual techniques for

learning can still be applied to improve upon the results for the ImageCLEFmed 2013 collection such as model averaging, ensemble learning or training data expansion.

Finally, we have conducted a brief analysis of the behavior of our approach by using a probabilistic view, trying to detect the main weak points in the discrimination capabilities of the different levels of the hierarchy. The result shows that the classification error is high for many instances, either correctly or wrongly classified. In future work, we will explore the possibilities of the proposed pipeline in other classification problems where classes present an intrinsic hierarchy. We plan to apply our system to indoor scene classification problems as well.

Acknowledgments

This work is partially funded by the FEDER funds and the Spanish Government (MICINN) through projects TIN2013-46638-C3-3-P, PPII-2014-015-P, and TIN2015-65686-C5-3-R. Jacinto Arias is funded by the MECD grant FPU13/00202. Jesus Martínez-Gómez is funded by the JCCM grant POST2014/8171.

References

- [1] C. Akgül, D. Rubin, S. Napel, C. Beaulieu, H. Greenspan, B. Acar, Content-based image retrieval in radiology: Current status and future directions, *Journal of Digital Imaging* 24 (2) (2011) 208–222.
- [2] Unknown, Riding the wave: How europe can gain from the rising tide of scientific data, Submission to the European Commission, available online at <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf> (October 2010).
- [3] S. Montani, R. Bellazzi, Supporting decisions in medical applications: the knowledge management perspective, *International Journal of Medical Informatics* 68 (2002) 79–90.
- [4] P. Welter, T. Deserno, B. Fischer, R. Günther, C. Spreckelsen, Towards case-based medical learning in radiological decision making using content-based image retrieval, *BMC Medical Informatics and decision Making* 11 (68).
- [5] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI communications* 7 (1) (1994) 39–59.
- [6] J. Kalpathy-Cramer, A. García Seco de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, H. Müller, Evaluating performance of biomedical image retrieval systems— an overview of the medical image retrieval task at ImageCLEF 2004–2014, *Computerized Medical Imaging and Graphics*.

- [7] A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, H. Müller, Overview of the ImageCLEF 2013 medical tasks, in: Working Notes of CLEF 2013, 2013.
- [8] C. Bielza, P. Larranaga, Discrete bayesian network classifiers: a survey, *ACM Computing Surveys (CSUR)* 47 (1) (2014) 5.
- [9] I. Kitanovski, I. Dimitrovski, S. Loskovska, FCSE at medical tasks of ImageCLEF 2013, in: Working Notes of CLEF 2013, 2013.
- [10] X. Zhou, M. Han, Y. Song, Q. Li, Fast filtering techniques in medical image classification and retrieval, in: Working Notes of CLEF 2013, 2013.
- [11] L. Denoyer, P. Gallinari, Bayesian network model for semi-structured document classification, *Information processing & management* 40 (5) (2004) 807–827.
- [12] H. Schneiderman, Learning a restricted bayesian network for object detection, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition., Vol. 2, IEEE, 2004, pp. 631–639.
- [13] F. Rubio, M. J. Flores, J. Martínez-Gómez, A. Nicholson, Dynamic semantic network for semantic localization, in: 15th Workshop of Physical Agents., 2014, pp. 144–155.
- [14] P. G. Ipeirotis, Analyzing the amazon mechanical turk marketplace, *XRDS: Crossroads, The ACM Magazine for Students* 17 (2) (2010) 16–21.
- [15] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, J. Kahn, C. E., W. Hersh, Overview of the clef 2009 medical image retrieval track, in: Multilingual Information Access Evaluation II. Multimedia Experiments, Vol. 6242 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 72–84.
- [16] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, J. Reisetter, C. E. Kahn Jr., W. Hersh, Overview of the CLEF 2010 medical image retrieval track, in: Working Notes of CLEF 2010, 2010.
- [17] P. Tirilly, K. Lu, X. Mu, T. Zhao, Y. Cao, On modality classification and its use in text-based image retrieval in medical databases, in: 9th International Workshop on Content-Based Multimedia Indexing, 2011.
- [18] M. Rahman, D. You, M. S. Simpson, S. K. Antani, D. Demner-Fushman, G. R. Thoma, Multimodal biomedical image retrieval using hierarchical classification and modality fusion, *International Journal of Multimedia Information Retrieval* 2 (3) (2013) 159–173.
- [19] A. P. Pentland, R. W. Picard, S. Scarloff, Photobook: Tools for content-based manipulation of image databases, *International Journal of Computer Vision* 18 (3) (1996) 233–254.

- [20] A. K. Jain, A. Vailaya, Image retrieval using color and shape, *Pattern Recognition* 29 (8) (1996) 1233–1244.
- [21] M. Simpson, M. Rahman, S. Phadnis, E. Apostolova, D. Demmer-Fushman, S. Antani, G. Thoma, Text- and content-based approaches to image modality classification and retrieval for the ImageCLEF 2011 medical retrieval track, in: *Working Notes of CLEF 2011*, 2011.
- [22] D. Markonis, I. Eggel, A. García Seco de Herrera, H. Müller, The medGIFT group in ImageCLEFmed 2011, in: *Working Notes of CLEF 2011*, 2011.
- [23] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view rgb-d object dataset, in: *IEEE International Conference on Robotics and Automation.*, IEEE, 2011, pp. 1817–1824.
- [24] H. Ali, Z.-C. Marton, Evaluation of feature selection and model training strategies for object category recognition, in: *IEEE/RSJ International Conference on Intelligent Robots and Systems.*, IEEE, 2014, pp. 5036–5042.
- [25] V. Kramarev, S. Zurek, J. L. Wyatt, A. Leonardis, Object categorization from range images using a hierarchical compositional representation, in: *22nd International Conference on Pattern Recognition.*, IEEE, 2014, pp. 586–591.
- [26] P. A. Bromiley, N. A. Thacker, M. L. Scott, M. Pokrić, A. Lacey, T. F. Cootes, Bayesian and non-bayesian probabilistic models for medical image analysis, *Image and Vision Computing* 21 (10) (2003) 851–864.
- [27] M. Velikova, P. J. Lucas, M. Samulski, N. Karssemeijer, On the interplay of machine learning and background knowledge in image interpretation by bayesian networks, *Artificial Intelligence in Medicine* 57 (1) (2013) 73–86.
- [28] J. Martínez-Gómez, A. Fernández-Caballero, I. García-Varea, L. Rodríguez, C. Romero-González, A taxonomy of vision systems for ground mobile robots, *International Journal of Advanced Robotic Systems* 11 (2014) 1–26.
- [29] A. García Seco de Herrera, D. Markonis, R. Schaer, I. Eggel, H. Müller, The medGIFT group in ImageCLEFmed 2013, in: *Working Notes of CLEF 2013*, 2013.
- [30] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60 (2) (2004) 91–110.
- [31] A. García Seco de Herrera, D. Markonis, H. Müller, Bag of colors for biomedical document image classification, in: H. Greenspan, H. Müller (Eds.), *Medical Content-based Retrieval for Clinical Decision Support, MCBR-CDS 2012, Lecture Notes in Computer Sciences (LNCS)*, 2013, pp. 110–121.

- [32] S. A. Chatzichristofis, Y. S. Boutalis, CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval, in: Lecture notes in Computer Sciences, Vol. 5008, 2008, pp. 312–322.
- [33] S. A. Chatzichristofis, Y. S. Boutalis, FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval, in: Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Service, 2008, pp. 191–196.
- [34] J. Han, K. Ma, Fuzzy color histogram and its use in color image retrieval, *IEEE Transactions on Image Processing* 11 (8) (2002) 944–952.
- [35] R. Schaer, D. Markonis, H. Müller, Architecture and applications of the parallel distributed image search engine (ParaDISE), in: FoRESEE 2014, 1st International Workshop on Future Search Engines, 2014.
- [36] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 2014.
- [37] D. M. Chickering, Learning bayesian networks is np-complete, in: Learning from Data: Artificial Intelligence and Statistics V, Springer-Verlag, 1996, pp. 121–130.
- [38] F. V. Jensen, T. D. Nielsen, Bayesian Networks and Decision Graphs, 2nd Edition, Springer Verlag, New York, 2007.
- [39] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, *Machine Learning* 29 (2-3) (1997) 131–163.
- [40] M. J. Flores, J. A. Gámez, A. M. Martínez, Supervised classification with bayesian networks, in: R. M. et al. (Ed.), *Intelligent Data Analysis for Real-Life Applications: Theory and Practice*, IGI Global, 2012, pp. 72–102.
- [41] M. Minsky, Steps toward artificial intelligence, *Proceedings of the IRE* 49 (1) (1961) 8–30.
- [42] C. I. Chow, C. N. Liu, Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information Theory* 14 (1968) 462–467.
- [43] M. Sahami, Learning limited dependence bayesian classifiers, in: *KDD*, Vol. 96, 1996, pp. 335–338.
- [44] G. Webb, J. Boughton, Z. Wang, Not so naive bayes: Aggregating one-dependence estimators, *Machine Learning* 58 (1) (2005) 5–24.
- [45] M. J. Flores, J. A. Gámez, A. M. Martínez, J. M. Puerta, Gaode and haode: two proposals based on aode to deal with continuous variables, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009, pp. 313–320.

- [46] G. H. John, P. Langley, Estimating continuous distributions in bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.
- [47] U. Fayyad, K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning.
- [48] M. A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, 2000, pp. 359–366.
- [49] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, *Pattern Recognition* 44 (8) (2011) 1761–1776.
- [50] A. D. Gordon, A review of hierarchical classification, *Journal of the Royal Statistical Society.* (1987) 119–137.
- [51] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, *The Annals of Mathematical Statistics* (1940) 86–92.
- [52] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, Vol. 2, IEEE, 2006, pp. 2169–2178.
- [53] M. Abedini, L. Cao, N. Codella, J. H. Connell, R. Garnavi, A. Geva, M. Merler, Q.-B. Nguyen, S. U. Pankanti, J. R. Smith, et al., IBM research at ImageCLEF 2013 medical tasks.
- [54] M. S. Simpson, D. You, M. Rahman, D. Demner-Fushman, S. Antani, G. Thoma, ITI’s participation in the 2013 medical track of ImageCLEF, in: *Working Notes of CLEF 2013*, 2013.
- [55] A. Mourao, F. Martins, J. Magalhaes, Novasearch on medical imageclef 2013, in: *Working Notes of CLEF 2013*, Vol. 2013, 2013, pp. 1–10.
- [56] S. Stathopoulos, I. Lourentzou, A. Kyriakopoulou, T. Kalamboukis, IPL at CLEF 2013 medical retrieval task, in: *Working Notes of CLEF 2013*, 2013.

Appendix A. Detailed Results

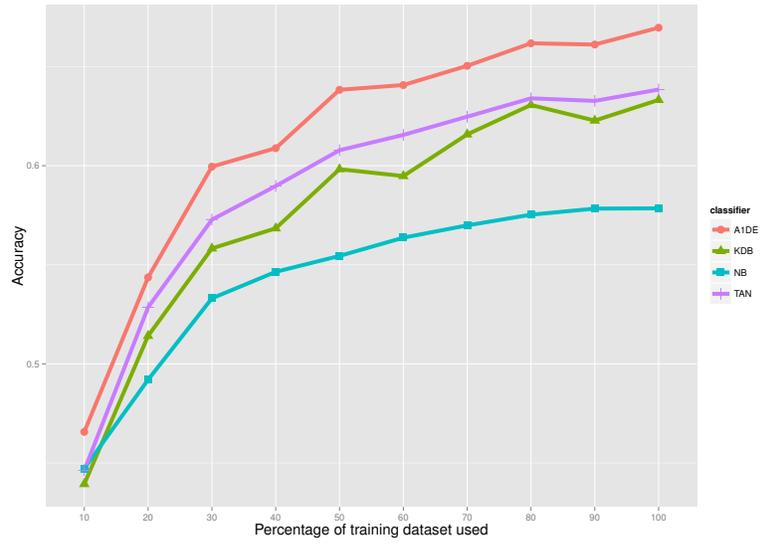


Figure A.9: Accuracy obtained when incrementally increasing the training set. Results are averaged by classifier.

Table A.4: Results for the descriptor combination in conjunction with the 3 hierarchies and 4 classification models. Highest accuracy (for each hierarchy) is shown in bold.

Descriptor Combination	1-level hierarchy				2-levels hierarchy				3-levels hierarchy			
	AODE	KDB	TAN	NB	AODE	KDB	TAN	NB	AODE	KDB	TAN	NB
D_1	46.38	46.15	38.17	44.19	50.78	45.75	45.59	47.04	50.81	45.83	46.22	47.20
D_2	44.93	48.72	39.70	48.34	48.88	46.37	45.08	46.33	48.32	46.53	46.32	45.99
D_3	47.19	47.41	37.36	47.15	46.30	45.01	42.89	45.05	46.16	44.57	44.13	44.88
D_4	47.67	48.29	41.25	47.60	45.68	44.16	42.56	44.67	45.48	44.39	44.67	44.57
D_5	41.73	50.21	29.34	47.61	49.05	47.47	42.89	48.19	48.36	45.66	44.09	46.34
$D_{1,2}$	55.65	57.37	46.36	55.67	61.37	56.15	55.10	56.54	60.84	55.26	55.05	56.47
$D_{1,3}$	54.42	55.11	44.23	52.47	59.18	53.92	52.85	54.96	58.18	53.06	53.38	54.60
$D_{1,4}$	54.19	54.54	45.73	52.75	59.12	54.30	53.16	55.30	58.72	53.67	53.82	55.46
$D_{1,5}$	54.61	57.17	42.00	55.46	60.73	54.38	53.18	55.57	60.01	53.76	53.90	55.52
$D_{2,3}$	46.93	52.66	38.99	51.04	51.67	48.43	46.90	49.56	50.32	48.30	48.02	48.48
$D_{2,4}$	48.46	51.84	40.99	50.79	51.76	48.18	47.73	48.78	50.91	48.75	47.82	48.56
$D_{2,5}$	49.61	56.87	38.14	54.23	56.75	53.56	49.63	53.91	55.01	52.16	50.44	52.40
$D_{3,4}$	46.78	49.57	37.18	48.56	48.67	45.99	43.76	46.80	48.20	46.52	46.24	46.76
$D_{3,5}$	47.57	55.33	35.71	52.95	55.28	52.35	48.84	52.91	53.77	50.87	49.53	51.28
$D_{4,5}$	48.99	55.80	38.57	53.39	55.43	52.17	50.12	52.96	53.69	51.25	50.08	51.93
$D_{1,2,3}$	57.55	59.39	46.48	57.05	62.87	56.91	56.10	58.11	62.09	56.17	56.33	57.73
$D_{1,2,4}$	57.51	58.96	47.58	57.34	63.70	57.47	56.75	58.67	62.38	56.94	56.71	57.75
$D_{1,2,5}$	59.16	60.97	46.51	59.92	65.47	58.95	57.75	59.70	64.00	58.19	58.14	59.76
$D_{1,3,4}$	55.50	55.95	44.67	54.12	60.35	55.44	53.52	56.58	59.86	54.27	54.29	56.20
$D_{1,3,5}$	57.09	59.82	43.69	57.96	63.38	57.01	56.13	58.20	62.20	56.83	56.51	58.15
$D_{1,4,5}$	57.51	59.28	45.15	57.96	63.94	57.61	56.35	58.74	62.81	57.57	57.10	58.63
$D_{2,3,4}$	47.85	53.12	37.98	51.51	53.76	50.20	48.01	50.48	51.91	49.38	48.13	49.95
$D_{2,3,5}$	51.46	58.93	38.91	55.93	58.67	54.88	50.88	55.31	57.03	53.43	51.75	53.83
$D_{2,4,5}$	52.80	58.43	40.70	56.05	59.04	54.55	52.00	55.28	56.91	53.42	51.72	53.92
$D_{3,4,5}$	49.27	56.69	37.71	54.29	57.22	53.07	50.41	53.62	55.16	52.20	50.83	52.68
$D_{1,2,3,4}$	58.02	59.25	46.68	58.20	64.01	57.76	56.85	58.99	62.77	57.15	56.95	58.33
$D_{1,2,3,5}$	59.61	61.67	46.62	60.25	65.69	60.05	57.85	60.39	64.19	58.70	58.38	60.19
$D_{1,2,4,5}$	60.04	61.52	47.64	60.17	66.22	60.28	58.60	60.97	64.74	59.42	58.42	60.61
$D_{1,3,4,5}$	57.84	60.34	44.80	58.61	64.33	58.74	56.57	59.10	62.78	58.32	57.23	59.41
$D_{2,3,4,5}$	52.35	58.72	40.03	57.03	59.50	54.94	52.13	55.75	57.51	54.17	52.44	54.57
$D_{1,2,3,4,5}$	60.03	62.07	47.15	60.38	66.00	60.54	58.55	61.01	64.55	59.59	58.56	60.68