UiT THE ARCTIC UNIVERSITY OF NORWAY

Faculty of Science and Technology, Department of Mathematics and Statistics

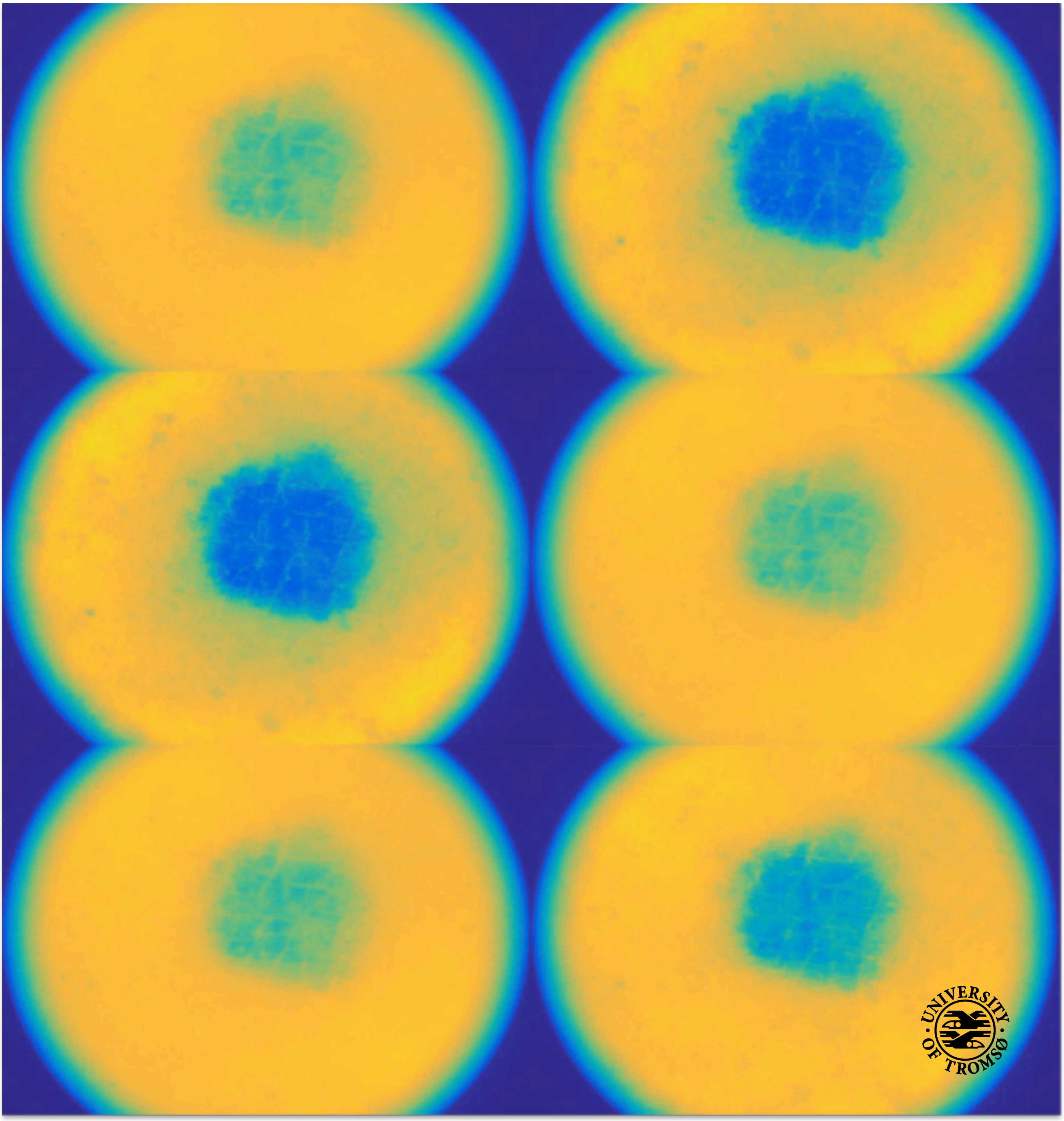
Melanoma detection

Colour, clustering and classification

—

Kajsa Møllersen

A dissertation for the degree of Philosophiae Doctor – November 2015



Abstract

Malignant melanoma is the deadliest form of skin cancer, and successful treatment relies on early detection. Undiagnosed skin lesions can be photographed and the images fed into a computer system that potentially differentiates malignant from benign lesions. To develop the melanoma detection system, various methods from statistics, machine learning and image analysis are applied.

An image consists of millions of pixels, so reducing the enormous amount of data is an important part of image analysis. This can be done by probability density estimation and clustering. In hierarchical clustering, the dissimilarity measure has great influence on the final clustering, but there has been little focus on how to choose an adequate dissimilarity measure for density-based techniques. In this thesis, six properties for density-based dissimilarity measures are therefore proposed as a guide for the user, based on assumptions and previous knowledge about the data set.

An image cannot be fed directly into a classifier because of the amount of data contained in each image; therefore a set of features is extracted from one. In melanoma detection, the colour of the lesion is of special interest. This thesis presents several approaches to colour feature extraction. (i) By clustering the pixel values and then comparing the cluster centres to pre-defined colour values, melanoma-indicative colours are detected. (ii) By estimating the probability density, and then measuring the goodness-of-fit, the variation in colours is accounted for. (iii) By the use of a dissimilarity measure, an unclassified lesion's similarity to the melanoma class or the class of benign lesions can be calculated. Different methods for feature evaluation are discussed.

A thorough presentation of computer systems for melanoma detection is provided, and some of the key elements are discussed. The challenge of feature selection and classifier selection is given special attention. A computer system for melanoma detection, *Nevus Doctor*, is presented. It is based on semi-automatic feature selection of both new and previously developed features, and a new hybrid classifier.

The performance of the system in terms of sensitivity and specificity scores is presented and compared to that of a commercially available system for the same set of lesions. This methodology has previously been used once only, and then in a smaller study. Obstacles associated with small data sets are discussed, including cross-validation and clinical relevance.

Nevus Doctor performed better than the commercially available system. The new colour features add value in computer-aided melanoma detection, both by improving the existing system and by introducing a new class of features. The properties for dissimilarity measures offer a new perspective on clustering and other fields where dissimilarity measures are a core element.

Acknowledgements

First I would like to thank my main supervisor Fred Godtliebsen who has followed (or rather lead) me from beginning to end. His strong belief in my abilities as a researcher combined with realistic probability estimates for manuscript rejection has pushed me higher than I could have achieved by myself but still not over the edge. During the course of my Ph.D. I have walked many paths that lead to dead ends, but since 'you haven't seen where you haven't been', every dead end adds value to the journey, even though there was nothing to be found. Stronger guidance could have resulted in fewer of those, but would have come with the expense of less independence and creativity.

I also owe a great thank you to Thomas R. Schopf, my co-supervisor. His knowledge about skin cancer has been invaluable, but would have been worthless without the eager to share and teach. During my work, I have come across many of the challenges that appear in applied statistics; data collection, communication with future users, incomprehensible medical terms (reading pathology reports in German), challenges that I would not have overcome without help. I have also learned to appreciate multi-disciplinary research, where you walk more dead ends, but at least you have companionship.

Some of the people I have walked the dead ends with, but also crossed some finish lines, are my colleagues and co-authors Jörn, Marc, Kevin, Maciel, Kristian, Stein Olav, Herbert, Jon Yngve and Subhra.

This thesis would not have been finished without my dear friends Heidi and Rebekka, who literally provided me food and shelter in times of need, and whose importance in my life expands far beyond that of a thesis. Now I share food and shelter with Øystein, who has endured long lectures on my recent findings at the dinner table.

Last, but not least, I thank my family in Kirkenes, Oslo and Fortaleza, who has made me who I am, for better or for worse. And my brother in Trondheim who died before he could see me cross the finish line, but who will be with me in my thoughts across this line and all other lines to come.

Just because it's black in the dark doesn't mean there's no colour.
- Laleh

List of Publications

I. **Kajsa Møllersen**, Maciel Zortea, Kristian Hindberg, Thomas R. Schopf, Stein Olav Skrøvseth, Fred Godtliebsen, "Improved skin lesion diagnostics for general practice by computer-aided diagnostics", In *Dermoscopy Image Analysis (M. E. Celebi, T. Mendonca, and J. S. Marques, eds.), pp. 247–292, CRC Press*, October 2015.

II. **Kajsa Møllersen**, Herbert Kirchesch, Maciel Zortea, Thomas R. Schopf, Kristian Hindberg, and Fred Godtliebsen, "Computer-aided decision support for melanoma detection applied on melanocytic and non-melanocytic skin lesions. A comparison of two systems based on automatic analysis of dermoscopic images", *BioMed Research International*, accepted, November 2015.

III. **Kajsa Møllersen**, Jon Y. Hardeberg, Fred Godtliebsen, "Divergence-based colour features for melanoma detection", In *Colour and Visual Computing Symposium (CVCS), pp. 1-6*, August 2015

IV. **Kajsa Møllersen**, Subhra S. Dhar, Fred Godtliebsen, "On data-independent properties for density-based dissimilarity measures in hybrid clustering", Submitted to *Journal of Machine Learning Research*, September 2015.

These publications are referred to by their roman letters in the following chapters.

Other Contributions

During the course of the PhD research, the author has contributed to other relevant publications. These may serve as background material, but are not regarded as part of the thesis.

1. **Kajsa Møllersen**, Herbert M. Kirchesch, Thomas G. Schopf, Fred Godtliebsen, "Unsupervised segmentation for digital dermoscopic images", *Skin Research and Technology, Vol. 16, No. 4., pp. 401-407*, November 2010.

2. Maciel Zortea, Thomas R. Schopf, Kevin Thon, Marc Geilhufe, Kristian Hindberg, Herbert Kirchesch, **Kajsa Møllersen**, Jörn Schulz, Stein O. Skrøvseth, Fred Godtliebsen, "Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists", *Artificial Intelligence in Medicine, Vol. 60, No. 1., pp. 13-26*, January 2014.

Contents

Chapter 1

Introduction

1.1 Image analysis: Feature extraction, feature selection and classification

Image analysis covers a wide range of objectives, fields, techniques and applications. The objective of the work presented here has been to classify skin lesions based on their colour image representation. To achieve this, different techniques from statistical analysis have been used, some of them closely related to machine learning. New techniques have been developed when existing ones did not meet the specific needs. The final application is a system for use in clinical practice that makes recommendations based on the classification. In the context of statistical analysis, images and their pixel values are treated as observations.

A common digital image consists of pixels spatially organised in a rectangle, where each pixel has one or more numbers associated to it. A colour image typically has three or four numbers for each pixel, e.g. one number for each of the colours red, green and blue (RGB), but many other representations exist. In an RGB image with 8 bit depth there will be more than 16 million possible unique colours. Fig. 1.1 shows the Kodak colour plate "red" photographed through a *dermoscope*, a magnifying lens with surrounding lights. There are more than $20,000$ unique shades of red in this 8 bit image, due to noise and non-homogeneous lightening. The pixel values can be spatially organised in a number of ways, depending on the number of pixels. In practice, all images are unique, even those of the same object.

In many applications it is desirable to group images, either for the purpose of grouping the images themselves or the objects in the images. If the images are labelled according to class, the groups are defined by the classes. Labelling a new image based on its similarity to the pre-defined classes is called classification, and is a form of supervised learning. Grouping independently of class label is referred to as clustering, and is a form of unsupervised learning. Both images and pixel values can be classified or clustered. In medical applications, images often have corresponding class labels, e.g. *benign* and *malignant*, and the objective is to develop a classifier that accurately predicts the class label of a new

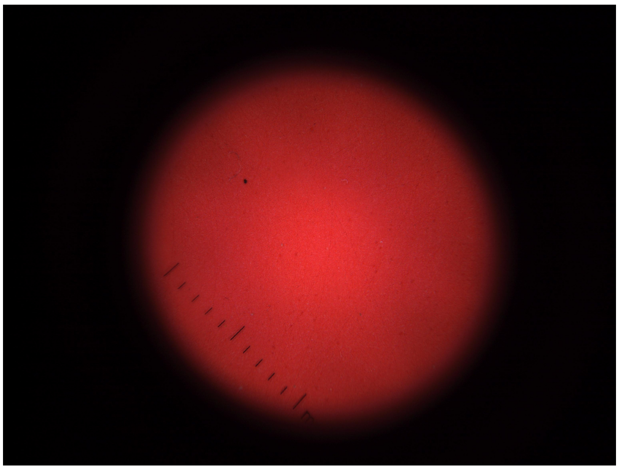


Figure 1.1: Kodak colour plate "red" photographed through a dermoscope.

observation. To make further analysis easier, it is sometimes desirable to group the pixels within an image, e.g. according to a pre-defined number of colours.

In order to be able to group images either by classification or clustering, the enormous amount of data involved, consisting of the pixel values and their spatial organisation, must be reduced. This is done by calculating characteristics of the image that are relevant for the grouping, and is called feature extraction. Deciding on which features to extract depends on the desired grouping. If Fig. 1.1 were one of many images of colour plates, a relevant feature might be the mean pixel value, for the purpose of grouping the plates according to colour. If Fig. 1.1 were one of many images of the same colour plate, taken with different dermoscopes, a relevant feature might be the light intensity decay towards the edges, for grouping the images according to the characteristics of the dermoscopes. Note that these two examples can be both classification problems and clustering problems.

The relevance of a feature is closely connected to its ability, in combination with other features, to separate groups. In clustering, where the desired grouping is defined by certain characteristics of the observations, the relevance of a feature is closely connected to its ability to quantify the specific characteristics. Translating characteristics to a feature extraction algorithm is not necessarily straightforward. Several feature algorithms can describe the light intensity decay towards the edges in Fig. 1.1, but not all are equally relevant. In classification, the relevance of a feature is more directly connected to class separability, since it can be measured against a ground truth. For both clustering and classification, a feature's relevance is dependent on the clustering or classification technique. A feature that increases group separability for one technique does not necessarily increase it for another technique. There is no strict definition of relevance: for a discussion see e.g. Kohavi and John (1997).

An unlimited number of features can be extracted from an image, but for the purpose of classification, large numbers of features should be avoided. If the number of features is high compared to the number of observations in each class, the classifier can be overfitted and become unstable. From a set of proposed features, a subset can be selected according to its ability for class separation. If the original feature set is very large, the feature selection encounters serious obstacles. The fraction of possible subsets that can be evaluated will be small, and possibly relevant feature combinations are excluded from the search. Consequently, the selected subset might not be the best subset. The problem of finding the best subset is known to be NP-hard (Guyon and Elisseeff, 2003; Amaldi and Kann, 1998), which implies that there are no known algorithms that can solve the problem in reasonable time for a large amount of features. Strong correlations between feature values and class labels are more likely to have appeared by chance if the number of features is very large. Features included in the classifier due to spurious correlations with the class labels can make the classifier unstable. Although it can be tempting to develop a series of feature extraction algorithms, often differing only by a parameter value, e.g. the pth percentile of the colours in an image for $p = 5, 10, \ldots, 95$, this should be avoided. If the data set is not adequately sized, spurious correlations are likely to occur, but even with a large enough dataset only a smaller proportion of possible feature subsets can be tested.

As a starting point for feature extraction algorithms, it is natural to turn to the experience of human experts. When human beings classify images the process is complex but the language describing the process is simple. Quantifying features such as “asymmetry” or “colourful” requires identification of the object, segmenting the object from the background, clustering pixels according to colour, etc. Due to the complexity of human processing, single features like “asymmetry” often result in several feature extraction algorithms, and thus, even if the initial set of features for human classification is small, the set of feature extraction algorithms can be large.

Even if the data set is large enough for the classifier to be stable with a large number of features, feature selection is needed to detect irrelevant and redundant features, which will lower the performance of the classifier. Feature selection can be done in a number of ways. Unsupervised feature selection aims at reducing the number of features independently of the class labels. Very highly correlated features and degenerate features can be detected, although this is only relevant if the number of features is very large. If two features are strongly correlated, one of them may be redundant, but to know which one better contributes to the classification, the class labels must be taken into account. It is also not possible to determine, without the class labels, just how strong the correlation must be before one of the features is redundant. For examples and discussion on correlation and redundancy, see e.g. Guyon and Elisseeff (2003). Feature selection based on correlation to class label, but independent of the specific classifier, is called filtering, which is useful to detect features with weak correspondence to the class labels and redundant features. The subset acquired from filtering should be investigated further with a feature selector specific for the classifier, called a wrapper. The chosen classifier might make assumptions about the features, e.g. that their values are normally distributed, that are not taken into account by the filter. Feature selection and classifier choice should be an iterative process. If the features do not meet the assumptions of the classifier, either the features can be transformed, or a different type of classifier can be chosen, and the feature selection repeated.

There is a wide range of classifiers available, from the simple linear discriminant analysis (LDA) and the intuitive k-nearest neighbours (k-NN), to more complex ones like support vector machines (SVM) and artificial neural networks (ANN). A classifier is first trained on a set of labelled data, and then a new observation can be classified. Many classifiers have user-set parameters, e.g. the number of neighbours and distance function in k-NN. There is a great risk of overfitting the classifier if the number of observations in each class is small compared to the number of features and the complexity of the classifier. Choice of classifier can be crucial, since the potential of class separability of the features is not fully exploited if the assumptions of the classifier are not met. Combining several classifiers might improve the performance compared to a system that relies on only one classifier (Ho et al., 1994). Other important aspects can be interpretability for the user, stability, etc.

To sum up, grouping images or other data objects containing lots of information is a complex task, made even more so by the interactions between the different parts: one part cannot be seen independently of the others. As well as those discussed here, other multiple aspects are also relevant for image analysis, e.g. noise reduction, choice of colour space, etc. These aspects are also integrated parts of the process. Then, in addition, there is the relevance of the actual application. In the work presented here, the application is melanoma detection, where images of skin lesions are classified as benign or suspicious. The cost of misclassifying a malignant lesion as benign can be huge, and this must be reflected in all parts of the analysis.

One challenge for applied problems is verification of the proposed solution. The goal for the system is to be able to correctly classify objects that were not used in developing the system. The only way to guarantee this is to test the system on a data set that was not available during development. The performance of the system improves, up to a certain point, as the number of observations available for training increases. For real data applications, the data set is most often limited, and there is a trade-off between optimal training and reliable verification. Cross-validation is a well-established method for using the whole data set, both for training and testing, which is especially useful if the data set is small compared to the complexity of the problem (see e.g. Hastie et al. (2009, pp. 241-9)). The data set is divided into K folds and for each repetition, $K-1$ folds are used for training and one fold is used for testing. Correct use of cross-validation requires that all parts of the training that depend on the class labels are repeated for each fold, not only the training of the classifier. The training involves

- Parameter adjustment of the feature extraction algorithms: many feature extraction algorithms include user-set parameters, which are often adjusted according to class label.

- Feature selection: if feature selection is done on the whole data set, the classification result can be *severely* biased.

- Classifier model selection: if the type of classifier (e.g. ANN, SVM, k-NN) and the user-set parameters (e.g. number of hidden layers) are chosen on the basis of the whole data set, the result can be *severely* biased.

There are numerous examples of wrong use of cross-validation in the literature. The problem has been addressed by authors in fields like bioinformatics (Smialowski et al., 2010) and medicine (Babyak, 2004). Doing feature selection on the whole data set is fairly common and it seems as though many authors are unaware of the bias that they may introduce. Classifier model selection is often not discussed, and it is therefore not possible for the reader to know whether the selection of the classifier was based on class labels. If the assumptions of cross-validation, independently and identically sampled observations, are not met, the result can be biased. This can be a problem, especially if samples are taken over a long time period, where the underlying distribution of the population is slowly changing.

In applied problems with real data that is not publicly available, how the performance of the system is reported requires special attention. When simulated or publicly available data is used, interested parties have the opportunity to replicate the experiment and measure different aspects of the system's performance, for example to compare it to their own method. However, in many medical applications the data is not publicly available, due to patient privacy regulations. Then the performance of the system should be reported more thoroughly: for example is the entire receiver operating characteristic (ROC) curve more informative than the area under the curve (AUC) or a single sensitivity/specificity pair. In addition, the clinical relevance must be taken into consideration when the performance is reported, so that the clinical relevant aspects are clearly shown.

1.2 Clustering and dissimilarity measures

The aim of clustering is to group observations so that the observations in one group (cluster) are more closely related to each other than to the observations in another group. In clustering, there is no ground truth, so the evaluation of a technique depends both on the specific data set and the problem-specific cluster concept. This has led to an enormous variety in clustering techniques, each one introduced to cover a specific aspect of clustering. Frequently, labelled data is used to evaluate clustering techniques since objective criteria are lacking. In an attempt to provide an overview, it may be useful to categorise the different techniques. Some of the more widely used categories are as follows - hard versus fuzzy, where fuzzy assigns a probability of cluster membership to each observation; deterministic versus stochastic, where stochastic techniques can provide different outcomes for the same data set; hierarchical versus partitional, where hierarchical provides a nested series of clusters. For more categorisations, see e.g. Jain et al. (1999). A rarer categorisation is distance-based versus model-based (Zhong and Ghosh, 2003). It can be argued that categorising clustering techniques is a form of clustering in itself, and that the variety of categorisations is necessarily as rich as the variety of techniques.

Deciding on which clustering technique is more adequate depends on the problem at hand in terms of the cluster concept. Choosing a clustering technique on the basis of the observed data is counter-intuitive; since clustering is a tool for data exploration, it would be redundant if the data had already been explored to the extent that the most adequate clustering technique could be selected. A small proportion of the data set could be used to choose a clustering technique, but because of the enormous amount of techniques available, the choice must be narrowed. The cluster concept can be used to choose a clustering technique, for example by restricting the clusters' ability to intersect, or penalisation of small clusters. Clustering techniques rely on dissimilarity measures, and for this reason, knowledge about the measures' properties is needed.

Dissimilarity measures can be categorised as distance-based versus density-based, which is similar to the clustering technique categorisation of Zhong and Ghosh (2003). Consider clustering the colours of an image. An option is to start with every pixel constituting a subcluster on its own, and then merge the two subclusters that are most similar. This is called hierarchical agglomerative clustering (Ward, 1963). At the first stage, a distance function can be used directly, since all subclusters are points in the space. At the next stages, if a distance function is used, a representation of the points in each subcluster or a function of the pairwise distances is needed. At later stages, when the number of subclusters is low compared to the number of observations, a density estimate can represent the points, and the distance function can replaced by a density-based dissimilarity measure.

The term *divergence* does not have a strict mathematical definition, and is often used about functions that aim at measuring dissimilarities, but that are not proper distance functions, or that belong to a class which includes such functions. Bregman divergences (Bregman, 1967) and f-divergences are used in clustering, and their properties can therefore be relevant when choosing the most adequate technique. The class of Bregman divergences includes the squared Euclidean distance and the Mahalanobis distance, and can in addition measure the distance between discrete probability distributions. Bregman divergences were investigated by Banerjee et al. (2005) in the clustering context. The class of f-divergences measures the distance between two probability distributions, and is therefore suited for density-based dissimilarity measures. The f-divergences possess information monotonicity, explained by Ali and Silvey (1966) as *We should not be able to increase our ability to distinguish between two different distributions by "grouping observations together".* Amari (2009) proved that any decomposable divergence that fulfils information monotonicity is an f-divergence, and stated the more general conjecture that any divergence that fulfils information monotonicity is a function of an f-divergence. The Kullback-Leibler information (often referred to as Kullback-Leibler divergence) is both an f-divergence and a Bregman divergence, therefore possesses both classes' properties (Amari, 2009), and is a much-used divergence for clustering.

Dissimilarity measures are fundamental in clustering since they define the *relation* between observations or groups of observations. Their properties are relevant in other contexts as well, e.g. when comparing reference models to proposed models. A dissimilarity measure will then indicate which of the proposed models best approximates the reference model. An example of this is content-based image retrieval, where a new image is compared to images with known class labels by the use of a dissimilarity measure. In filter feature selection, a proxy measure is used to evaluate each feature. This can, for example, be class separability measured by divergence; see e.g. Guyon and Elisseeff (2003).

An image consists of millions of pixels, which correspond to observations. The length of the vector associated to each pixel (three or four, depending on the colour space) corresponds to the dimension of the sample space. Any observed sparsity can therefore be assumed to reflect the underlying distribution, and it is therefore possible to make good density estimates. Assuming that the observations are independently drawn from an underlying distribution, a probability distribution can be estimated, and the density can be summarised in very few parameters from which features can be calculated. The challenge then remains as to how to estimate the density and how specifically to use the density estimates as features. The mixture of Gaussian distributions can be used for density estimation since it approximates any continuous distribution with arbitrary accuracy (Maz'ya and Schmidt, 1996; Huber et al., 2008). A possibility for feature extraction is to compare the estimated density to other densities, and for that a dissimilarity measure is needed.

Chapter 2

Computer systems for melanoma detection

2.1 Melanoma detection and dermoscopy

Skin cancer is the most common human cancer, with high incidence rates especially in regions with predominantly fair-skinned populations, such as Europe, North-America, Australia and New-Zealand (Ferlay et al., 2013). The most common skin cancer type is basal cell carcinoma (BCC), which is not considered deadly and therefore excluded from most cancer registries and studies related to skin cancer. The remaining skin cancers are usually divided into two groups; malignant melanoma (melanoma) and non-melanoma skin cancer (NMSC). Both melanoma and NMSC can metastasise and can therefore be deadly. NMSC is approximately twice as common as melanoma, and BCC is approximately five times as common as the other two combined (Socialstyrelsen, 2014). It is believed that BCC and NMSC are under-reported, and that this is because of their low mortality rates. Hence, incidence rates must be handled with care. Melanoma accounts for nearly 90% of skin cancer deaths in Norway (Cancer Registry of Norway, 2015), and is by far the most deadly type of skin cancer. It is the second most common cancer type in the age group 25-49 years in Norway, both for men and women, and the incidence rate is increasing (Cancer Registry of Norway, 2015). Increasing incidence rates are found both in Europe (Forsea et al., 2012), the USA (American Cancer Society, 2014) and Australia (Australian Institute of Health and Welfare, 2015).

Melanomas originate from melanocytes; see Fig. 2.1. These are the same cells that produce melanin, the main component in skin colour. Benign melanocytic lesions, commonly referred to as moles, also originate from melanocytes. BCC originates from basal cells, and the most common NMSC, squamous cell carcinoma (SCC), originates from squamous cells - both examples of non-melanocytic lesions. Benign lesions can also be non-melanocytic. Although melanocytic and non-melanocytic lesions originate from different cells, and thus have different physiology, their visual appearance can be similar.

Recent developments in melanoma treatment are promising, but the only effective treat-

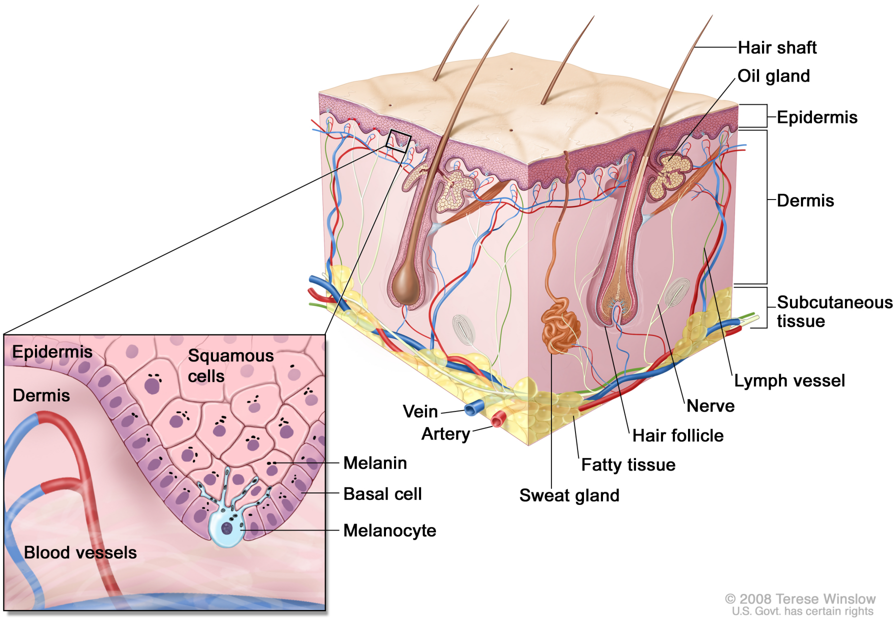


Figure 2.1: Anatomy of the skin.

ment for melanoma is still excision of the tumour before metastasis (Garbe et al., 2011; Niezgoda et al., 2015). The five-year survival rate for patients diagnosed with melanoma is close to 90% if the tumour has not metastasised, but drops to about 20% for tumours with distant spread metastasis (Cancer Registry of Norway, 2015). Early detection is therefore crucial. In its early stages, a melanoma resembles a common benign skin lesion, and detection is therefore challenging. A final melanoma diagnosis can only be made by excising the lesion for histopathological examination of the tissue, as seen in Fig. 2.2.

The decision to excise a lesion is made by a dermatologist or a general practitioner (GP) based on visual inspection and patient history. Due to the high risk for the patient if a melanoma is left untreated, the decision to excise is based on a low grade of suspicion. Excising a lesion is fairly easy, and is often done by the GP. Due to the similarity of melanoma and benign lesions, and since low grade suspicion triggers excision of the lesion, the number of benign lesions per melanoma being excised is high. For the dermatologist, the excision ratio of benign lesions per melanoma varies from 10-50:1 in reported studies, whereas the excision ratio for GPs is significantly higher and typically lies around 50-100:1 (Lindelöf et al., 2008; Marks et al., 1997). Histopathological examination of a suspicious lesion is time-consuming and requires expert knowledge. It is considered a bottleneck in melanoma diagnosis.

Most of the light that impinges the skin is reflected by the uppermost skin layer, and only the information contained there is available for naked-eye examination. A dermoscope (dermatoscope) is a device consisting of a magnifying lens with surrounding lights and a glass plate, used to examine skin conditions. The glass plate has approximately the same refraction index as the uppermost skin layer, and is used in combination with fluid (water,

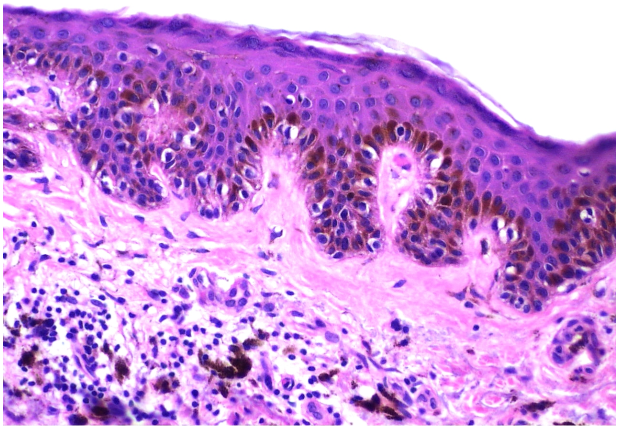


Figure 2.2: Tissue sample prepared for histopathological examination. L. Wozniak & K. W. Zielinski /CC BY-SA 3.0

alcohol) to avoid air bubbles between the glass plate and the skin. The light can now penetrate the uppermost skin layer, and more information about the lesion is revealed (Stolz et al., 2002). An alternative to the glass plate is polarised light, which penetrates the skin even deeper. However, the information in the superficial layer is lost (Pan et al., 2008). The structures and colours (features) revealed by the use of a dermoscope can help in differentiating between melanomas and benign skin lesions, but the interpretation of dermoscopic features requires training and experience (Kittler et al., 2002). Several rules have been drawn up to help the inexperienced dermoscopy user, e.g. the ABCD-rule of dermoscopy, based on (A) asymmetry, (B) border, (C) colour and (D) differential structure of the lesion (Nachbar et al., 1994). A value is assigned to each dermoscopic feature, and a total score is calculated based on the weighted sum of the feature values. The calculation of the total score is easy, but assigning values to the dermoscopic features is not straightforward; even expert dermoscopy users have low inter-observer agreement (Argenziano et al., 2003). The dermoscope itself is easy to use, but, with the exception of Australia (Rosendahl et al., 2012), dermoscopy is not widely used among GPs, probably because of the need for training and experience.

The increasing incidence rate for melanoma, public awareness campaigns (Sneyd and Cox, 2013) and screening programs (Katalinic et al., 2012) all present a challenge to health care systems, namley to increase the detection rate for early stage melanomas without increasing the excision ratio of benign lesions per melanoma, since pathology resources are limited. A potential solution is to rely more on computer systems to avoid unnecessary excision of benign lesions. Possibly the first description of a computer system for skin lesions was published in 1984 (Vanker and Stoecker, 1984), and since then, both computer and image technology have made a giant leap forward. A dermoscope can easily be coupled with a digital camera, as seen in Fig. 2.3(a), and the result is a dermoscopic image, as seen in Fig. 2.3(b). This equipment is off-the-shelf and often referred to as digital dermoscopy.

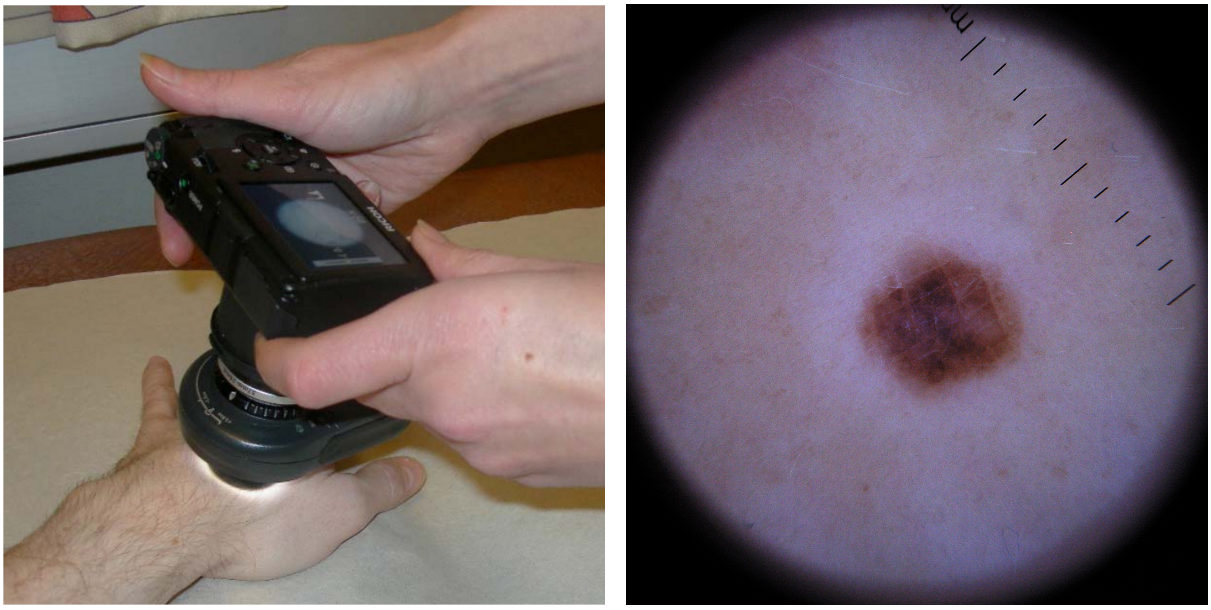


Figure 2.3: (a) A dermoscope attached to a digital camera. (b) Dermoscopic image of benign lesion.

It can be used for a second opinion or consensus diagnosis without having to refer the patient (Kittler et al., 2002). Digital dermoscopy is used with high-risk patients to follow up suspicious lesions over time, since it is not possible to excise all such lesions (Argenziano et al., 2008). In addition, digital dermoscopy can be used for automatic image analysis as part of computer-aided decision support.

2.2 Computer-aided clinical decision support systems for melanoma detection

Clinical decision support aims at providing a recommendation based on patient-specific information to aid the clinician in making the right decision. A simple example is check lists for surgeons. In computer-aided clinical support systems (CDSS), also referred to as computer-aided diagnostic (CAD) systems, the recommendation is the output of some computer algorithm. Whether such systems actually improve clinical practice is a question that does not yet have a definite answer. A successful CDSS must be able to give the correct recommendation, but also have impact on the actual decision. There are numerous examples of successful CDSSs, but also of the opposite (Kawamoto et al., 2005). For a CDSS to have the potential of clinical impact, both the clinician and the patient must accept its use, and the CDSS must have the potential of improving patient care. For a CDSS to have the potential of improving patient care, the sensitivity and specificity of the CDSS must be higher than that of the clinician.

Whether CDSSs for melanoma detection can improve clinical practice is an open question. The results from the study of Dreiseitl and Binder (2005) indicated that clinicians are willing to follow a CDSS's recommendation if they are uncertain of their own judgement. Frühauf et al. (2012) concluded that most patients accept the use of an image-based CDSS for melanoma detection. The main focus of studies on CDSSs for melanoma detection has been performance in terms of sensitivity and specificity compared to that of dermatologists, and the overall conclusion is that the CDSSs do not perform better (Rosado et al., 2003; Vestergaard and Menzies, 2008; Korotkov and Garcia, 2012). However, GPs in general have lower diagnostic accuracy than dermatologists, so there might be a potential for use in general practice. According to a recent review (Koelink et al., 2014), only one study (Walter et al., 2012) has investigated the value of adding a CDSS in general practice. The results did not show any improvement when the CDSS was added.

The computer system described in Vanker and Stoecker (1984) required that the clinician detected the lesion features and fed them into a classifier, but automatic and semi-automatic image analysis followed shortly. The early computer systems did not use dermoscopic images, but *clinical* images of skin lesions. In this context, clinical images refer to images captured without a dermoscope attached in front. In the late 1990's there was a shift towards dermoscopic images, and by early 2000 almost all computer systems were based on dermoscopic images (Rosado et al., 2003). Other image modalities such as multispectral imaging then followed, but dermoscopic images seem the most popular so far (Korotkov and Garcia, 2012). More recently, both electrical impedance spectroscopy (Malvehy et al., 2014) and Raman spectroscopy (Lui et al., 2012) have been used in CDSSs for melanoma detection. There is little evidence that one technology is better than the others, and research and development on CDSSs for melanoma detection continue in parallel for the different modalities, with the exception of clinical images.

Review papers on the performance of CDSSs for melanoma detection give little reason to believe that a computer system can make a sufficiently accurate diagnosis to be trusted completely by the clinician (Rosado et al., 2003; Vestergaard and Menzies, 2008; Korotkov and Garcia, 2012). This is to be expected, since the CDSS aims at reproducing the histopathological classification of a skin lesion based on a tissue sample (Fig. 2.2), based on the information contained in an in vivo measurement (Fig. 2.3(b)). Since the CDSS classification cannot be trusted completely, the advantage of image-based technologies is that the clinician is given the opportunity to verify the CDSS's findings by the visual appearance of the skin lesion. From this point of view, the term 'computer-aided clinical *decision support* system' is more accurate than 'computer-aided *diagnostic* system' because the systems are not able to give a diagnosis. They can help the clinicians in their decision by either classifying the lesion as suspicious (excise), or classify the lesion as benign (not excise).

Most image-based CDSSs for melanoma detection follow the same main steps; the image is pre-processed, the lesion is segmented from the background skin, feature values are extracted, and the lesion is classified, as illustrated in Fig. 2.4. A few systems for content-based image retrieval also exist (Baldi et al., 2014).

Many CDSSs are already commercially available and many are described in the litera-

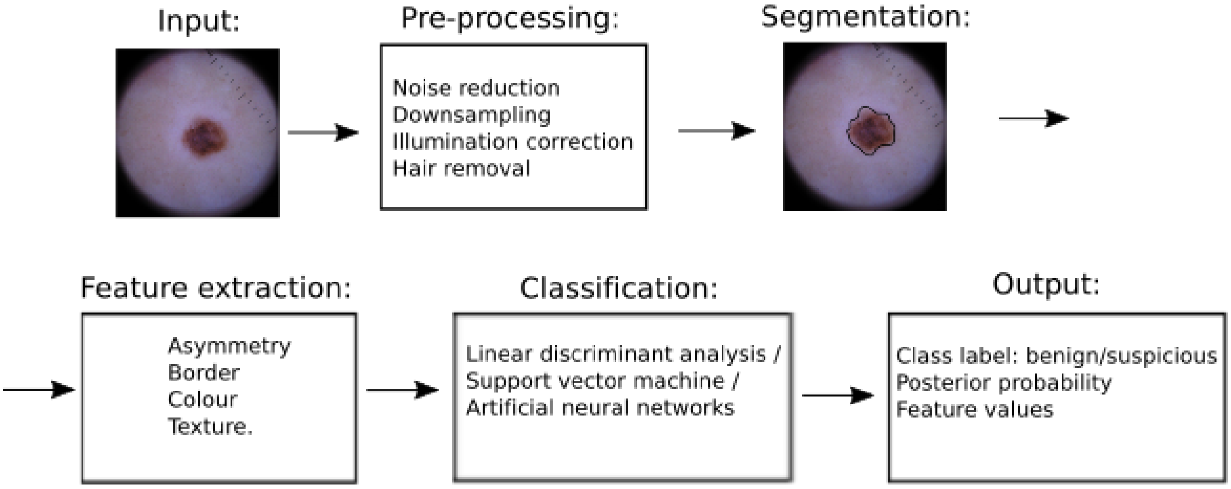


Figure 2.4: Common structure for image-based CDSSs for melanoma detection.

ture. Sensitivity and specificity scores close to 100% have been reported; see e.g. the review paper of Blum et al. (2008). These studies have major shortcomings such as post-hoc exclusion criteria based on histopathological diagnosis, and wrong use of cross-validation. Distinguishing between melanocytic and non-melanocytic lesions is not straightforward, not even for expert dermatologists (Argenziano et al., 2003). There will therefore be a certain number of non-melanocytic lesions among the lesions that are examined due to suspicion of melanoma. Excluding these lesions post-hoc based on the histopathology report biases the sensitivity and specificity scores.

In practice it is not feasible to do parameter adjustment, feature selection and classifier model selection for each cross-validation repetition, and therefore an independent test set is needed. An illustration of just how biased non-independent test sets can be are the two MelaFind studies: Elbaum et al. (2001) reported sensitivity of 100% and specificity of 84% using cross-validation. Ten years later, Monheit et al. (2011) reported sensitivity of 95% and specificity of 10% on an independent test set.

There are very few studies with independent test sets, consecutively collected data, well-defined exclusion criteria and moderate to large number of melanomas. Even when these studies are well-designed, they have different inclusion and exclusion criteria resulting in different diagnostic difficulty of the respective data sets. It is therefore not possible to rank them based on the reported sensitivity and specificity scores.

Colour features

Colour is an important feature in melanoma detection, both for dermoscopy users and computer-aided systems. The ABCD-rule of dermoscopy defines six possible colours in a lesion; white, red, light and dark brown, blue-grey, and black. Brown and black are frequently found in benign lesions, whereas white and blue-grey can be regarded as melanoma-indicative colours. Red can be found in benign lesions, but can also indicate malignancy, and special attention is therefore needed if the lesion contains red colour. In addition to detection of specific colours, large variation in colour and asymmetric colour distribution within the lesion are indicative of melanoma. Recognising colours is a challenge, both for people and computers.

In the CDSS for melanoma detection literature, the colour features can be divided into two main classes: features that *describe* the colours and features that *divide* colours into classes or *detect* certain colours (Korotkov and Garcia, 2012). When dividing colours into classes, the thousands of colours in the lesion image are categorised into fewer classes, typically five or six. These classes can be pre-defined by the user, or be data driven. Detection of blue-grey colour, often referred to as blue-white veil, is a common feature in the divide-detect class. Colour descriptors are characterisations of the pixel value distribution, typically mean, variation, skewness, range, percentiles, entropy, energy, etc. When building a CDSS, a large number of candidate colour features are often calculated, and then reduced through automatic feature selection, but the number of candidate colour features should be kept to a minimum.

Computer-aided clinical decision support systems in Northern Norway

Northern Norway is a sparsely populated region. Although its northernmost county, Finnmark, is bigger than Denmark, it has only 75, 000 inhabitants; a visit to the nearest specialist doctor requires more often than not hours of travelling. There are only two dermatologists in Finnmark, so the local GPs need to diagnose and treat more advanced conditions than their colleagues in more densely populated areas. They also have to do better triage (deciding which patients must be referred to a dermatologist), since the expense in term of travel cost and lost work hours is high. Dermatology is not the only speciality with few practitioners in Finnmark, and one cannot expect the GPs to have superior knowledge in all fields of medicine. Therefore, CDSSs that can improve the GP's decision are especially welcomed in sparsely populated regions. Even in more densely populated areas, the limited availability of dermatologists results in long waiting lists. The short-term solution is to excise the lesion at the GP's office for histopathological examination, but in the long term this increases the work load for the pathologist, and is therefore not a desirable outcome.

Chapter 3

Results and Discussion

This chapter presents the four papers which make up the thesis. The findings in each paper are discussed, and suggestions for future work are put forward.

3.1 Paper I - Improved skin lesion diagnostics for general practice by computer-aided diagnostics

A thorough presentation of the various elements in computer-aided systems for melanoma detection is provided. A computer-aided system and its performance from an earlier study are presented, since it forms the basis of a new system named *Nevus Doctor*. The performance of Nevus Doctor is then reported in terms of sensitivity and specificity scores on an independent test set.

The paper also discusses feature selection and classifier selection, two issues that have not received much attention in the melanoma detection literature. A semi-automatic feature selection is performed, based on a combination of a filter method and a wrapper method. The selected feature set depends on the chosen classifier for the wrapper and the proxy measure for the filter. Additional criteria for feature evaluation, such as interpretability and computational burden are discussed.

There is no evidence in the literature that any one specific classifier is preferred for melanoma detection. Studies that compare different classifiers are difficult to implement because of the complexity of the systems. The results cannot be generalised, because the performance of the classifiers depends on the data set, segmentation algorithm, feature set, feature selector, etc. As long as the sensitivity and specificity scores of a system are too low for the user to trust the system completely, the interpretability of the classifier is

The final version of the paper is in black and white and contains references to other chapters in the book where it appears. The version presented here is in colour and without references to other chapters, but otherwise identical to the final version.

The dermoscope in this study was DermLite FOTO, not Dermlite Pro II HR as referred to in the paper.

an important aspect. Interpretability includes both additional outcome, such as posterior probability, and classification processes with restricted interaction between the feature values.

Two new colour features are presented, one that describes the colours in an image, and one that detects certain colours. The colour detector measures the presence of the melanoma-indicative colours - blue, red and whitish. Clustering is performed on the lesion pixels with a pre-defined number of clusters. The cluster centres represent the colours of a lesion, and are compared to the pre-defined melanoma-indicative colours. The descriptor measures how well a Gaussian mixture model with a pre-defined number of components fits the lesion pixel values. A lesion with low variability in the colours can be fitted better than a lesion with high variability in the colours for the same restriction on distribution and number of components. The two features both use unsupervised learning, first, to reduce the number of colours in a lesion from the number of unique pixel values to a pre-defined number of clusters, and second, to describe the variability in a sample through clustering and density estimation.

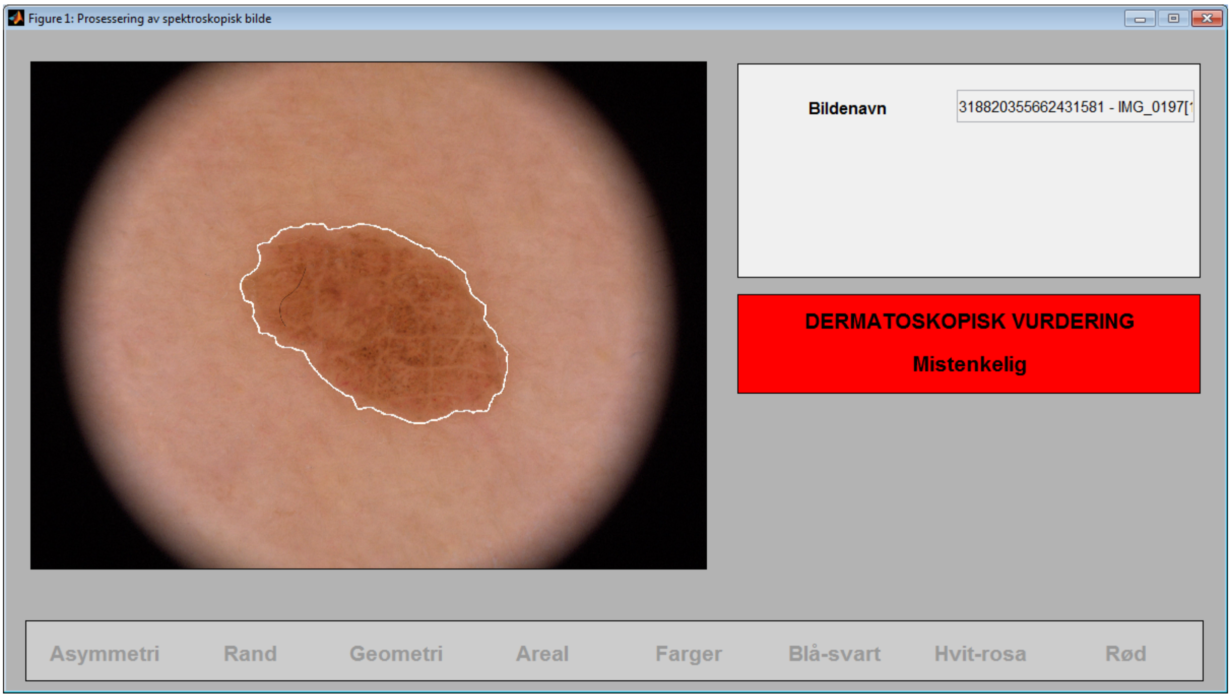
A hybrid classifier is proposed, consisting of linear discriminant analysis and cut-off values for single features. The proposed classifier is an attempt to provide more interpretable feedback to the user. In addition to the class label, an indication is given as to whether the threshold for single feature values is exceeded.

Nevus Doctor is tested on an independent test set, and its performance compared to that of an independent dermatologist in terms of benign to malignant excision ratio, where the gold standard is the pathology reports. In addition, sensitivity and specificity scores are reported, where the gold standard is the dermatologist's excision recommendation combined with the pathology reports. The independent test data are not consecutively collected, and one of the main quality criteria for performance studies is therefore violated. Another quality criterion is that the classification of the system should be compared to human diagnosis. Combining these two criteria with a reasonable number of melanomas requires that independent dermatologists classify thousands of lesion images. This heavy work load makes recruitment for such studies difficult.

Future work

Nevus Doctor is currently being used in a pilot study for a clinical trial being conducted at a rural GP's office in Finnmark, Norway. The objectives of the pilot study are to evaluate the user friendliness and interpretability of the system. The current outcome presented to the user is shown in Fig. 3.1. The preliminary results indicate that more detailed feedback is needed and that the computation time is too high. In addition, the possibility of sending the images to a dermatologist for an expert opinion should be incorporated.

The current hybrid classifier consists of a complex (LDA) and an interpretable (the single feature values) part. The result is a classifier that can only occasionally give the user an interpretation of the class label. Interpretable classifiers such as LDA and decision trees become complex when the number of features is high, so reducing the number of features is a necessity. Today, several image features are needed for each dermoscopic feature. If



Bildenavn 318820355662431581 - IMG_0197[

DERMATOSKOPISK VURDERING

Mistenkelig

Figure 3.1: The outcome of Nevus Doctor. None of the eight single feature value thresholds were exceeded, shown in grey in the bottom row. The lesion was labelled ‘suspicious’, but no explanation is offered.

these features are grouped, and a single value can be retrieved, then the complexity of the classifier can be reduced. A possibility is to have one classifier for each dermoscopic feature, whose outputs are the inputs of a final classifier. The final classifier can then become simple enough for visualisation and interpretation.

3.2 Paper II - Computer-aided decision support for melanoma detection applied on melanocytic and non-melanocytic skin lesions. A comparison of two systems based on automatic analysis of dermoscopic images

The CDSS for melanoma detection described in Paper I, Nevus Doctor, is tested here on an independent test set of consecutively collected images from clinical practice. A thorough description of the data set is provided, including diagnosis for both malignant and benign lesions and Breslow depth for the invasive melanomas. As opposed to Paper I, the test set here was consecutively collected, which means that the proportion of melanoma, NMSC and the different benign lesions is representative of the clinical practice where the data set was collected. One of the main obstacles when comparing CDSSs for melanoma detection is that the data sets on which they have been tested can vary significantly in diagnostic difficulty. Indications of the level of diagnostic difficulty are: proportions of melanomas in situ, Breslow depth, proportion of non-melanocytic lesions, and clinical diagnosis, but only the data set itself can give a full description of the diagnostic difficulty. A commercially available CDSS for melanoma detection is therefore applied to the same set of lesions, and compared to the performance of Nevus Doctor. The results indicate that Nevus Doctor performs equally well for melanocytic lesions, and better for non-melanocytic lesions. The two CDSSs misclassify different melanomas, and a simple classifier combining the two CDSSs' class labels with an OR operation is presented. For melanocytic lesions only, where the two CDSSs performed approximately equally, the OR classifier had better performance.

The paper discusses the inadequacy of comparing CDSSs based on their sensitivity and specificity scores, due to the scores' dependence on the data set. This is demonstrated by replacing one melanoma with another in a set-up consisting of only half of the melanomas of the original test set. The results then show that the specificity score increases for one CDSS and decreases for the other, and by replacing only one image, the ranking of the two systems shifts. The paper also points out that low sensitivity scores are not clinically relevant, and therefore the area under the receiver operating characteristic curve is not an adequate measure for the performance of a system.

Future work

This study is one of the largest ever implemented in terms of the number of melanomas in an independent, consecutively collected test set from clinical practice, but the data set is still too small to generalise the findings. Ongoing data collection can increase the size of the independent test set, but there is also reason to believe that more melanomas in the training set can improve the performance.

The detailed results show that Nevus Doctor misclassifies sebhorreic keratosis, a benign non-melanocytic skin lesion. Developing features that aim at differentiating sebhorreic keratosis from malignant lesions will increase the specificity of the system and potentially save resources in the pathology department.

The results from the OR classifier pose the question as to whether this strategy is applicable within a CDSS. Today, CDSSs for melanoma detection apply only one classifier, but it may be beneficial to combine several different classifiers. The wrapper feature selection should then be repeated for each classifier in the ensemble. The drawback of combining classifiers is that it introduces a new level of complexity to the system, which hampers interpretability and requires a larger data set to avoid overfitting.

3.3 Paper III - Divergence-based colour features for melanoma detection

A new type of colour feature for melanoma detection is presented. The concept is to build one model for benign lesions and one model for malignant lesions based on labelled data, and to measure the divergence between an unlabelled lesion and each of the two models. The divergence is measured between the estimated probability distribution of an unlabelled lesion and the estimated probability distribution of each of the two models, and therefore the problem does not require a symmetric divergence function. If the two models are separable and an appropriate divergence function is chosen, the divergence and the class labels will be correlated. The challenge is to choose the appropriate divergence function. It is assumed that the model distribution will envelop the lesion distribution if the model is the same as the true class of the unlabelled lesion, and only partly overlap otherwise. Based on these assumptions it can be shown that the problem requires a non-symmetric divergence function; the Kullback-Leibler information is suggested.

The probability densities are assumed continuous, and are estimated by Gaussian mixture distributions. There is no closed form for the Kullback-Leibler information, and importance sampling can be used to approximate the integrals. In order to emphasise the region where the two models differ, the observations are sampled from one model but the integrand is weighted by the other model. The result is improper importance sampling.

The performance of the two features is measured on the data set used for training in Paper I. Three methods for evaluating the features are used: (i) sensitivity and specificity scores for the individual features, (ii) result from a classifier-independent feature selection, and (iii) sensitivity and specificity scores for a classifier with and without the new features. The results show that the two new features can add value to melanoma detection, but the data set is too small to enable firm conclusions to be drawn. The method can be used for other problems as well.

The paper discusses the three methods for feature performance evaluation. Individual sensitivity and specificity scores cannot describe a feature's relevance in a more complex system, because it depends on the correlation to the other features, and whether the feature meets the assumptions of the classifier. Classifier-independent feature selection can indicate the feature's value since the correlation to other features are accounted for, but different proxy measures can lead to different results. To truly know if a feature increases the performance of a specific CDSS, the performance can be measured when adding the feature. However, the increase/decrease depends on the whole feature set and the chosen classifier, and the result cannot necessarily be generalised to other feature sets and classifiers.

Future work

The results from the paper should be verified on the independent test set from Paper II, and on the hybrid classifier described in Paper I. If the results are good, the features can be added to Nevus Doctor.

The amount of time needed for each lesion image is too high due to the fitting of several Gaussian mixture distributions. It should be investigated whether it is possible to use a fixed number of components, such as for the colour descriptor in Paper I, without significant decrease in performance. The notation used to describe the assumptions should be changed to (δ, ε)-notation, consistent with the notation in Paper IV. A more theoretical justification of the improper importance sampling and its behaviour as a function of the divergence between the two models can add value.

3.4 Paper IV - On data-independent properties for density-based dissimilarity measures in hybrid clustering

In partitional-hierarchical hybrid clustering, the dissimilarity measure involved in the hierarchical clustering has a large influence on the final outcome. Data-independent properties for distance-based dissimilarity measures have been investigated for decades, but density-based dissimilarity measures have not received the same attention. This paper proposes six data-independent properties for density-based dissimilarity measures. Through properties on symmetry, equality and orthogonality, some of the basic concepts of a dissimilarity measure for hierarchical clustering are covered. Since hierarchical merging involves pairwise divergences, and not divergence from a reference model to proposed models, a symmetry property is proposed. The equality and orthogonality properties follow from the divergences' concept of measuring a distance. Three properties are proposed regarding outliers, noise and light-tailed models for heavy-tailed clusters. These properties are adequate for categorising dissimilarity measures. With appropriate categories, the most adequate dissimilarity measure for the problem at hand can be chosen. The need for the proposed properties is illustrated by examples and references in the literature. The impact of different dissimilarity measures in view of which properties they fulfil is illustrated on real and simulated data.

The paper sheds light on a neglected issue associated with partitional-hierarchical hybrid clustering. Since the properties are not axiomatic but serve to categorise, it will not be possible to provide a complete list. It is hoped that more properties will be added in the future to categorise dissimilarity measures from different viewpoints, or fine-grain already existing categories. The investigation of previously proposed dissimilarity measures revealed that some of them are not adequate for partitional-hierarchical hybrid clustering. This was not known when the dissimilarity measures were proposed, since they were only evaluated on specific data sets.

The use of density-based dissimilarity measures is not limited to hybrid clustering, and some of the proposed properties can be relevant in other settings.

Future work

The property regarding light-tailed models for heavy-tailed clusters is limited to Gaussian distributions with equally shaped covariance matrices. A more general property, e.g. for unimodal, symmetric distributions, should be proposed. An important issue in clustering is estimation of the number of clusters, and dissimilarity measures can be used for this purpose. Properties regarding this should also be proposed.

The properties' relevance for other areas where density-based dissimilarity measures are used should be investigated, e.g. proxy measures for feature selection.

Chapter 4

Conclusions

This thesis presents work on computer-systems for melanoma detection in different aspects and from different viewpoints. It provides an overall introduction to image-based computer-systems for melanoma detection, and discusses some of the main issues involved. Feature extraction algorithms for colour are proposed. The descriptive feature based on Gaussian mixtures and a measure of fit can be viewed as a predecessor to the more elaborate divergence-based features. However, the introduction of divergence-based features has not made the measure-of-fit feature redundant. A hybrid classifier is proposed, based on the need for interpretable computer systems.

The CDSS for melanoma detection, Nevus Doctor, was tested on a consecutively collected, independent test set, which is a necessary verification of the promising results from previous studies. A new methodology for comparing CDSSs is introduced - direct comparison on the same set of lesions. Although this method has been used once before, it was in a very small study. The discussion on evaluation methods may be valuable in a field where statistical knowledge is relatively poor, and methods are often used without any previous discussion as to how or why.

The divergence-based colour features expand the field of descriptive colour features from simple ad-hoc to more elaborate and justified ones. Divergence-based colour features are not limited to the proposed divergence function, and thus a whole new class of colour features is provided.

The properties for density-based dissimilarity measures offer a new perspective on partitional-hierarchical hybrid clustering. In image analysis, the number of observations (pixel values) is large and pure hierarchical clustering techniques suffer from the drawback of being computationally expensive. The number of dimensions (three in an RGB image) is low, so good density estimates are available, and hybrid clustering will be particularly well suited. The properties can offer new perspectives on any field where density-based dissimilarity measures are important.

Bibliography

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1): 131–142, 1966.

E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1-2):237–260, December 1998.

S. I. Amari. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, November 2009.

American Cancer Society. Cancer Facts & Figures 2014. Technical report, American Cancer Society, 2014.

G. Argenziano, I. Mordente, G. Ferrara, A. Sgambato, P. Annese, and I. Zalaudek. Dermoscopic monitoring of melanocytic skin lesions: Clinical outcome and patient compliance vary according to follow-up protocols. *British Journal of Dermatology*, 159(2):331–336, August 2008.

G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, R. Hofmann-Wellenhof, M. Landthaler, S. W. Menzies, H. Pehamberger, D. Piccolo, H. S. Rabinovitz, R. Schiffner, S. Staibano, W. Stolz, I. Bartenjev, A. Blum, R. Braun, H. Cabo, P. Carli, V. De Giorgi, M. G. Fleming, J. M. Grichnik, C. M. Grin, A. C. Halpern, R. Johr, B. Katz, R. O. Kenet, H. Kittler, J. Kreusch, J. Malvehy, G. Mazzocchetti, M. Oliviero, F. Zdemir, K. Peris, R. Perotti, A. Perusquia, M. A. Pizzichetta, S. Puig, B. Rao, P. Rubegni, T. Saida, M. Scalvenzi, S. Seidenari, I. Stanganelli, M. Tanaka, K. Westerhoff, I. H. Wolf, O. Braun-Falco, H. Kerl, T. Nishikawa, K. Wolff, and A. W. Kopf. Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *Journal of the American Academy of Dermatology*, 48(5):679–693, May 2003.

Australian Institute of Health and Welfare. Australian cancer incidence and mortality (ACIM) books: Melanoma of the skin, 2015. URL `http://www.aihw.gov.au/acim-books`.

M. A. Babyak. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3):411–421, May 2004.

A. Baldi, R. Murace, E. Dragonetti, M. Manganaro, and S. Bizzi. Automated content-based image retrieval: Application on dermoscopic images of pigmented skin lesions. In A. Baldi, P. Pasquali, and E. P. Spugnini, editors, *Skin Cancer*, Current Clinical Pathology, pages 523–528. Springer New York, 2014.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, December 2005.

A. Blum, I. Zalaudek, and G. Argenziano. Digital image analysis for diagnosis of skin tumors. *Seminars in Cutaneous Medicine and Surgery*, 27:11–15, 2008.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, January 1967.

Cancer Registry of Norway. Cancer in Norway 2013 - cancer incidence, mortality, survival and prevalence in Norway. Technical report, Cancer Registry of Norway, 2015.

S. Dreiseitl and M. Binder. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artificial Intelligence in Medicine*, 33 (1):25–30, January 2005.

M. Elbaum, A. W. Kopf, H. S. Rabinovitz, R. G. B. Langley, H. Kamino, M. C. Mihm, A. J. Sober, G. L. Peck, A. Bogdan, D. Gutkowicz-Krusin, M. Greenebaum, S. Keem, M. Oliviero, and S. Wang. Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: A feasibility study. *Journal of the American Academy of Dermatology*, 44(2):207–218, February 2001.

J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC cancerbase no. 11 [internet]. Technical report, International Agency for Research on Cancer, 2013. URL `http://globocan.iarc.fr`.

A. M. Forsea, V. del Marmol, E. de Vries, E. E. Bailey, and A. C. Geller. Melanoma incidence and mortality in Europe: New estimates, persistent disparities. *British Journal of Dermatology*, 167(5):1124–1130, November 2012.

J. Frühauf, B. Leinweber, R. Fink-Puches, V. Ahlgrimm-Siess, E. Richtig, I. H. Wolf, A. Niederkorn, F. Quehenberger, and R. Hofmann-Wellenhof. Patient acceptance and diagnostic utility of automated digital image analysis of pigmented skin lesions. *Journal of the European Academy of Dermatology and Venereology*, 26(3):368–372, March 2012.

C. Garbe, T. K. Eigentler, U. Keilholz, A. Hauschild, and J. M. Kirkwood. Systematic review of medical treatment in melanoma: Current status and future prospects. *The Oncologist*, 16(1):5–24, January 2011.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, 2nd edition, 2009.

T. K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(1):66–75, January 1994.

M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pages 181–188. IEEE, August 2008. ISBN 978-1-4244-2143-5.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

A. Katalinic, A. Waldmann, M. A. Weinstock, A. C. Geller, N. Eisemann, R. Greinert, B. Volkmer, and E. Breitbart. Does skin cancer screening save lives? *Cancer*, 118(21): 5395–5402, November 2012.

K. Kawamoto, C. A. Houlihan, A. A. Balas, and D. F. Lobach. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ (Clinical research ed.)*, 330(7494):765+, April 2005.

H. Kittler, H. Pehamberger, K. Wolff, and M. Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165, March 2002.

C. J. Koelink, M. F. Jonkman, K. Van Der Meer, and W. K. Van Der Heide. Examination of skin lesions for cancer: Which clinical decision aids and tools are available in general practice? *European journal of dermatology*, 24(3):297–304, 2014.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, December 1997.

K. Korotkov and R. Garcia. Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine*, 56(2):69–90, October 2012.

B. Lindelöf, M.-A. Hedblad, and U. Ringborg. Nevus eller malignt melanom? Rätt kompetens vid diagnostik ger lägre kostnader. *Läkartidningen*, 105(39):2666–2669, 2008.

H. Lui, J. Zhao, D. McLean, and H. Zeng. Real-time Raman spectroscopy for *in vivo* skin cancer diagnosis. *Cancer Research*, 72(10):2491–2500, May 2012.

J. Malvehy, A. Hauschild, C. Curiel-Lewandrowski, P. Mohr, R. Hofmann-Wellenhof, R. Motley, C. Berking, D. Grossman, J. Paoli, C. Loquai, J. Olah, U. Reinhold, H. Wenger, T. Dirschka, S. Davis, C. Henderson, H. Rabinovitz, J. Welzel, D. Schadendorf, and U. Birgersson. Clinical performance of the Nevisense system in cutaneous melanoma detection: An international, multicentre, prospective and blinded clinical trial on efficacy and safety. *The British journal of dermatology*, 171(5):1099–1107, November 2014.

R. Marks, D. Jolley, C. McCormack, and A. P. Dorevitch. Who removes pigmented skin lesions?: A study of the ratio of melanoma to other benign pigmented tumors removed by different categories of physicians in Australia in 1989 and 1994. *Journal of the American Academy of Dermatology*, 36(5):721–726, May 1997.

V. Maz'ya and G. Schmidt. On approximate approximations using Gaussian kernels. In *IMA Journal of Numerical Analysis*, pages 13–29, 1996.

G. Monheit, A. B. Cognetta, L. Ferris, H. Rabinovitz, K. Gross, M. Martini, J. M. Grichnik, M. Mihm, V. G. Prieto, P. Googe, R. King, A. Toledano, N. Kabelev, M. Wojton, and D. Gutkowicz-Krusin. The performance of MelaFind: A prospective multicenter study. *Archives of dermatology*, 147(2):188–194, February 2011.

F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, April 1994.

A. Niezgoda, P. Niezgoda, and R. Czajkowski. Novel approaches to treatment of advanced melanoma: A review on targeted therapy and immunotherapy. *BioMed Research International*, 2015:1–16, 2015.

Y. Pan, D. S. Gareau, A. Scope, M. Rajadhyaksha, N. A. Mullani, and A. A. Marghoob. Polarized and nonpolarized dermoscopy. *Archives of Dermatology*, 144(6):828–829, June 2008.

B. Rosado, S. Menzies, A. Harbauer, H. Pehamberger, K. Wolff, M. Binder, and H. Kittler. Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis. *Archives of Dermatology*, 139(3):361–367, March 2003.

C. Rosendahl, G. Williams, D. Eley, T. Wilson, G. Canning, J. Keir, I. McColl, and D. Wilkinson. The impact of subspecialization and dermatoscopy use on accuracy of melanoma diagnosis among primary care doctors in Australia. *Journal of the American Academy of Dermatology*, 67(5):846–852, November 2012.

P. Smialowski, D. Frishman, and S. Kramer. Pitfalls of supervised feature selection. *Bioinformatics*, 26(3):440–443, February 2010.

M. J. Sneyd and B. Cox. A comparison of trends in melanoma mortality in New Zealand and Australia: The two countries with the highest melanoma incidence and mortality in the world. *BMC Cancer*, 13(1):372+, August 2013.

Socialstyrelsen. Cancerincidens i Sverige 2013. Technical report, Socialstyrelsen, 2014.

W. Stolz, O. Braun-Falco, P. Bilek, M. Landthaler, W. H. C. Burgdorf, and A. B. Cognetta. *Color Atlas of Dermatoscopy*. Blackwell Wissenschafts-Verlag, Berlin, 2nd edition, 2002.

A. D. Vanker and W. V. Stoecker. AI/DERM: Diagnosis of skin tumors. In D. A. B. Lindberg and M. F. Collen, editors, *AAMSI Congress 1984*, pages 213–217. American Association for Medical Systems and Informatics, 1984.

M. E. Vestergaard and S. W. Menzies. Automated diagnostic instruments for cutaneous melanoma. *Seminars in cutaneous medicine and surgery*, 27(1):32–36, March 2008.

F. M. Walter, H. C. Morris, E. Humphrys, P. N. Hall, A. T. Prevost, N. Burrows, L. Bradshaw, E. C. F. Wilson, P. Norris, J. Walls, M. Johnson, A. L. Kinmonth, and J. D. Emery. Effect of adding a diagnostic aid to best practice to manage suspicious pigmented lesions in primary care: Randomised controlled trial. *BMJ (Clinical research ed.)*, 345, July 2012.

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.

S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, December 2003.

Chapter 5

Papers I - IV