USER-TAILORED
CLINICAL DECISION SUPPORT SYSTEMS

Andrea Sboner

Advisor:
Dott. Ing. Paolo Traverso
ITC-irst – Center for Scientific and Technological Research, Trento, I
Abstract

Computational tools for building clinical decision support systems (CDSS) have been developed for decades. Several aspects must be taken into account when designing and building CDSSs. First, data uncertainty is a common aspect in medicine. Moreover, different skill levels and expertise of health care providers lead to a high grade of inter-observer variability in the assessment of patient’s conditions. Inter-observer variability, i.e. different observations or interpretations by different human experts of the same clinical parameters, plays a great role when no gold standard measurements are available. Rather surprisingly, despite this problem is well known, it has been largely ignored when developing CDSSs. To our knowledge, no solution is available in the literature to soundly cope with this problem.

Moreover, the output of a support system, besides accurate, should also be valuable for the physician. In other words, a system should provide correct, new and sufficient information to the user and that information should be interpretable and comprehensible. This aspect is fundamental as the final aim of a CDSS is to improve the performances of unaided physician.

In this work, we focused on the development of a CDSS, addressing in particular the inter-observer variability problem, by means of a combination of artificial intelligence techniques: a critiquing system, based on a novel machine learning approach and a consulting system based on case-based reasoning methodology.

Concerning the critiquing module, the novelty of this machine learning based approach regards the development of user tailored learning systems that exploit knowledge and skills specific of each physician. Concerning the consultation module, the CDSS applies concepts from the case-based reasoning (CBR) approach for sharing implicit knowledge of expert physicians.

The medical domain of this research is the early diagnosis of melanoma, the most dangerous form of skin tumor. The goal of the CDSS is to provide physicians with a tool supporting the early recognition and treatment of malignant lesions on the basis of
subjective evaluations of clinical parameters and the sharing of implicit, subjective knowledge of experts.

**Keywords**
Clinical decision support systems, inter-user variability, artificial intelligence, machine learning, case-based reasoning, multi-modal reasoning.
Contents

CHAPTER 1 .......................................................................................... 1
  1. INTRODUCTION ................................................................................ 1
      1.1. The Context ............................................................................. 1
      1.2. The Problem ........................................................................... 3
      1.3. The Solution ............................................................................ 4
      1.4. Innovative Aspects .................................................................. 6
      1.5. Structure of the Thesis ............................................................ 7

CHAPTER 2 .......................................................................................... 9
  2. STATE OF THE ART .......................................................................... 9
      2.1. Clinical decision support systems ......................................... 10
      2.2. Machine learning .................................................................. 13
      2.3. Case-based reasoning in medicine ....................................... 15
      2.4. CDSSs for the early diagnosis of melanoma ........................ 16

CHAPTER 3 ........................................................................................ 17
  3. THE PROBLEM ............................................................................... 17
      3.1. General model of a CDSS ..................................................... 19
      3.2. Early melanoma diagnosis ................................................... 22
      3.2.1 Advantages and drawbacks of current CDSSs systems for early
           melanoma diagnosis .................................................................23

CHAPTER 4 ........................................................................................ 27
  4. THE PROPOSED APPROACH ........................................................... 27
      4.1. User-tailored CDSS: components......................................... 28
      4.1.1 Critiquing module ......................................................................28
      4.1.2 Consulting module .....................................................................28

CHAPTER 5 ........................................................................................ 28
  5. EXPERIMENTAL RESULTS .............................................................. 28
      5.1. Dataset description ............................................................... 28
      5.2. Preliminary results ............................................................... 28
      5.3. Evaluation ............................................................................ 28
      5.3.1 Subjectivity Assessment ............................................................28
      5.3.2 Physician Performances ............................................................28
Acknowledgements

Numerous people deserve my special thanks for their help in so many ways, without which I never have completed this research study.

Starting from my closest colleagues and friends of ITC-irst, I’d like to warmly thank: Antonella Graiff, Paolo Traverso, my advisor, Francesca Demichelis, Rossana Dell’Anna, Federica Ciocchetta, and Luigia Carlucci Aiello, former director of ITC-irst.

Next, the dermatologists who participated in this study, as well as in other studies, The Trento group: Paolo Bauer, Mario Cristofolini, Antonella Bergamo, Giuseppe Zumiani, Trento, I – The Florence group: Gianfranco Gensini, Paolo Carli, Alessandra Chiarugi, Vincenzo De Giorgi, Francesca Mannone, Paolo Nardini, Camilla Salvini, Marcello Stante – University of Florence, I.

All the people at the Laboratory for Medical Informatics, University of Pavia; I and in particular Mario Stefanelli (wishing him all the best).

The wonderful people at Vanderbilt University, Nashville, TN, where I had the opportunity to spend a couple of months during my thesis: Nanci M. Lorency, Constantin F. Aliferis, Ioannis Tsamardinos, Randolph A. Miller, Alexander R. Statnikov, Lewis J. Frey, Rischelle Jenkins and all the other people I met there.

I would like to acknowledge the financial support of the Pezcoller Foundation and the Fondazione Trentina per la Ricerca sui Tumori.
Last, but not least, a special thank to Marzia, my wife, for her support, understanding and tolerance throughout these years. This thesis is dedicated to my beloved wife.
Chapter 1

1. Introduction

Diagnostic, prognostic or treatment models have been studying for long time in several medical disciplines to improve quality of patient care, and as long as computational tools came available, they were employed in medicine to assist physicians in their decision making. In particular, a great deal of research activity has been devoted to the development of Clinical Decision Support Systems (CDSSs).

This is the context of Medical Informatics or Health Informatics, an inter-disciplinary field, applying concepts, methodologies, and tools of information technology to medicine and health care [1].

This thesis regards the problem of the development of a CDSS for the early recognition of melanoma, the most dangerous form of skin cancer, accounting for the inter-user variability in pigmented lesion assessment. It provides the user with a specific critiquing system tailored on user’s expertise, and a consulting module supporting the sharing of implicit knowledge of expert dermatologists.

1.1. The Context

The UK Health Informatics Society defined Medical Informatics as “the understanding, skills and tools that enable the sharing and use of information to deliver healthcare and promote health” [2]. The U.S. Institute of Medicine’s report “To Err Is Human: Building a Safer Health System” [3] started the debate on medical errors. As a matter of fact, preventable errors results in 44,000-98,000 deaths a year in the USA. These figures exceed the deaths
CHAPTER 1. INTRODUCTION

due to motor vehicle accident, breast cancer, or AIDS. The report highlighted that information technology can be a means to reduce those errors and costs. In particular, it suggested that decision support systems are important components to achieve this goal.

“Clinical decision support systems (CDSSs) form a significant part of the field of clinical knowledge management technologies through their capacity to support the clinical process and use of knowledge, from diagnosis and investigation through treatment and long-term care.”[4].

The pioneering studies of the sixties and seventies report the application of statistics and artificial intelligence techniques for helping physicians in their everyday tasks. The seminal work of those days has been the basics for subsequent development of Medical Informatics field and in particular for the development of Clinical Decision Support Systems.

Those studies showed two main different approaches: statistical approaches, such as in the Leeds Abdominal Pain System [5] which provides physicians with explanations for acute abdominal pain through Bayesian probability theory, and rule-based approaches addressing diagnostic and patient’s management problems, such as in MYCIN, and in Internist-1/Quick Medical Reference (QMR) [6,7,8]. In the latter systems logic-based approach were employed together with heuristics which were able to cope with uncertainty that is always present in medicine.

In general, the rationale of CDSSs is that proper management of clinical knowledge is required to support physician’s daily decisions, as their personal knowledge might be outdated, or time-constraints and limited information may impair their reasoning procedure.

The current pressure on healthcare organizations to ensure both quality of care and cost containment is driving them towards a more effective management of medical knowledge. CDSSs are important tools to help to accomplish this goal.
1.2. The Problem

Uncertainty in data is a common problem in medicine: besides the intrinsic variability of biologic conditions, patients cannot describe exactly how they feel, physicians cannot define exactly what they observe, and measurements from medical devices always present some degree of error. Moreover, different skill levels and expertise of health care providers lead to a high grade of inter-user variability in the assessment of patient’s conditions. This variability is even more evident when no clear and definite biomedical knowledge is available and therefore the interpretation of symptoms is deeply dependent on physician’s expertise and skill.

From a clinical viewpoint, inter-user variability has been extensively investigated in pathology and radiology, where interpretation of images is highly dependent on individual expertise [9,10]. As a matter of fact, almost every field of medicine presents this problem. From a more general viewpoint inter-user variability is due to the amount of subjective, tacit or implicit knowledge that is always present in every decision making process. Rather surprisingly, despite this problem is well known, it has been largely ignored when developing CDSSs [11]. Searching for scientific articles in PubMed1 with the Medical Subject Heading (MeSH) terms: “Decision Support Systems, Clinical” and “Observer-variation”, resulted in only a dozen papers. Among those, Cross et al. [12] highlighted inter-user variability as the main cause for the failure of an even promising CDSS (see section 2.1). Clinical Decision Support Systems built so far deal mainly with the explicit knowledge. Computerized guidelines, for instance, use Evidence Based Medicine (EBM) paradigm to provide help to physicians, fostering the dissemination of “best practices” and improving the quality of care.

1 PubMed, National Library of Medicine, National Institute of Health
CHAPTER 1. INTRODUCTION

Another problem to consider regards the output of a support system: besides accurate, it should also be valuable for the physician. In other words, a system should provide correct, new and enough information to the user and that information should be interpretable and comprehensible. This aspect is fundamental as the final aim of a CDSS is to improve the performances of unaided physician: a CDSS is useless if it provides users with information they already know or not pertained to the problem [13].

This thesis addresses the development of a clinical decision support system for the early diagnosis of melanoma. Among the tools dermatologists can use for the diagnosis of malignant melanoma, dermoscopy has been proven to be valuable, increasing the diagnostic accuracy for pigmented skin lesions, especially for early melanoma [14,15]. However, to really benefit from this technique, a well experienced dermatologist is needed [16]. One of the main reasons for that is that the explicit definition of the dermoscopic criteria is rather difficult, since it involves a description of visual parameters. This leads to a high degree of subjectivity in feature assessment [17].

Up to now, most CDSSs for assisting physicians in the early recognition of melanoma rely on automated image analysis, in an attempt to overcome this problem. However, some drawbacks limited the development and use of those image based systems in the clinical practice.

1.3. The Solution

In this work, we focused on the development of CDSSs, addressing in particular the inter-observer variability problem, by means of a combination of artificial intelligence techniques that exploit user’s skill and expertise as well as support the sharing of experts’ tacit knowledge. The medical domain of this research is the early diagnosis of melanoma, the most dangerous skin tumor.

The approach here proposed follows current trends of CDSSs research, which aim at the combination of different problem-
CHAPTER 1. INTRODUCTION

solving approaches [11,18,19]. The CDSS goal is to provide physicians with a tool supporting the early recognition and treatment of malignant lesions on the basis of subjective evaluations of clinical parameters. Inter-user variability in the clinical assessment of such pigmented lesions is extremely critical, as noted by Kittler et al. [16].

The CDSS is based on two modules: a critiquing module and a consulting module, both of them relying on a reference dataset. Such dataset is composed by a number of known cases. Objective information (such as patient’s age) as well as the “gold standard” data (in the melanoma scenario, the histological diagnosis) are present in the dataset.

Subjective information is given by a group of expert dermatologists, who evaluated those cases independently from each other. They provide the information that completes the description of a case.

The reference dataset together with dermatologists’ evaluations is the reference case base.

Each new user of the CDSS is first required to assess the reference dataset by providing clinical parameter evaluations and diagnostic and treatment decisions. This procedure allows:

- Building a critiquing system by applying a novel machine learning-based approach, able to use the evaluations of the single physicians and to exploit his/her diagnostic capability. The critiquing module applies a user-tailored classification model, learned by focusing on the most “difficult” cases of the reference data set, i.e. the cases in which user’s decision was wrong.

- Building a consulting system by comparing the user subjective evaluations with the ones contained in the reference case based. This allows: i) assessing similarities between the user and the experts in evaluating the lesions; ii) selecting the expert “more close” to the user in terms of the subjective evaluations; iii) assessing the diagnostic and treatment capabilities of the user with respect to the experts ones. The con-
sulting module aims to share the implicit knowledge of experts.

When a new case is provided by the user to the CDSS, the critiquing module would suggest the user to re-evaluate his/her decisions if deemed necessary, i.e., when the clinical decisions and model outputs are different [20], while the consulting system shows experts feature evaluations and decisions on similar cases in the case base upon user request: the retrieval takes into account the inter-user similarity estimated on the reference case base.

To investigate this model, a prototype web-based application was developed and used by a group of physicians with different expertise in the diagnosis of melanoma. The distributed nature of the system allowed sharing information and knowledge across multiple institutions.

1.4. Innovative Aspects

To our knowledge, this thesis is the first study to investigate a user-tailored CDSS, which exploits physician’s expertise and support the sharing of experts’ implicit knowledge.

Concerning the critiquing module, the originality of this machine learning based approach regards the development of personalized learning systems that exploit knowledge and skills specific of each physician. The classical machine learning approach for building decision support system is not suitable in this case. In fact, machine learning based CDSSs build models from a set of known data. Typically, in medicine an expert, or a group of experts, provides this set of data and the learning algorithm uses these data to create a model, in an attempt to elicit from them the expert knowledge. This resulting model is then supposed to be used by novices or less experienced physicians.

This approach of creating CDSS is successful if the input data (features) are objective ones (for instance, physiological measurements, demographic data, etc.), or if the input data are subjective, but highly reproducible among physicians because well defined. Whenever the features are subjective, i.e. depending on
CHAPTER 1. INTRODUCTION

physician’s expertise, the model created so far, with experts data, may fail when used by novices [12]. The novel approach proposed in this thesis creates a tailored model on each physician instead of a general one, properly combining the human expertise with the computer analysis. The obtained results outperform standard machine learning approaches on the melanoma domain [20].

Concerning the consultation module, the CDSS applies concepts from the case-based reasoning (CBR) approach to deal with sharing of implicit knowledge of expert physicians. CBR methodology can be appropriate to support medical decisions because it resembles the analogical reasoning of physicians, and may enable the retrieval of other physician’s expertise and knowledge [21,22]. The retrieval of the most similar case cannot be performed on the reference case base, without dealing with inter-user variability in feature assessment. In fact, case retrieval should account for the different expertise of the physicians and their personal, subjective way of assessing patients.

The consulting system provides novices with the implicit knowledge of expert dermatologists included into their evaluations on the reference case base. The novice can compare the retrieved case with that s/he has at hand in terms of feature evaluations and clinical diagnosis. The retrieved cases also show expert’s treatment choices that can further help the proper management of patients, and support the process of “socialization” that allows sharing tacit knowledge.

1.5. Structure of the Thesis

In the next chapter there is a review of the state-of-the-art of clinical decision support systems, as well as the description of CDSS in the field of the early recognition of melanoma. Chapter 3 and 4 describes the problem we faced and the proposed solution, respectively. Chapter 5 describes the experimental evaluation and results. Finally, in chapter 6 there is a discussion of the work and the future steps to pursue.
CHAPTER 1. INTRODUCTION

In the appendix, the description of the web based system and the clinical background and are reported.
Chapter 2

2. State of the Art

A great deal of research in Bio-Medical Informatics has been devoted to the creation of computerized systems helping physicians taking proper decisions. Computational tools for building clinical decision support systems (CDSS) have been developed for decades.

The pioneer studies of early seventies showed basically two different approaches: statistical approaches, such as in the Leeds Abdominal Pain System [5] which provides physicians with explanations for acute abdominal pain through Bayesian probability theory, rule-based approaches addressing diagnostic and patient’s management problems, such as in MYCIN, and in Internist-1/Quick Medical Reference (QMR) [6,7,8].

These works still represent milestones for the subsequent investigations on clinical decision support systems, in particular on knowledge-based systems.

In the next section a discussion of the evolution of Clinical Decision Support Systems is reported. For a more detailed description, see R.A. Miller [23,24].

Since the content of my thesis regards more specifically the use of machine learning tools and case-based reasoning methodology for creating a CDSS for the early diagnosis of melanoma, next sections reports the application of these methodologies in the medical field.
2.1. Clinical decision support systems

Clinical Decision Support Systems can be defined as "active knowledge systems which use two or more items of patient data to generate case-specific advice" [25]. Given this definition, CDSSs include, and exploit, three elements: a knowledge base, patient data, and an inference engine to provide case-specific advices.

Different approaches were applied for the development of clinical decision support systems, since the early sixties. Two main dichotomous approaches were pursued: the probabilistic approach and the logical approach. The latter approach requires the CDSS a very detailed knowledge of pathophysiology or extensive epidemiologic data about the domain of application, in order to divide the possible output decisions into non-overlapping sets. This approach is suitable in medicine only into narrow domains where detailed medical knowledge is present. Unfortunately, medical decision making often presents a certain level of uncertainty that makes the application of pure logic systems unsuitable.

On the other hand, Bayesian probabilistic reasoning was successfully applied to several medical domains. The independence assumption among diagnosis and findings, and a way to overcome this limitation, has led to the development of Bayesian Network that is currently a very active research and application topic.

An alternative between the two approaches combines characteristics of both through heuristics reasoning, based on empirical rule-of-thumb. Most of the work on expert systems in the seventies and eighties is based on heuristics systems employing symbolic reasoning. MYCIN [6], Internist-1 [7,8], Iliad [26], DXPlain [27], to name some of the well-known systems.

The commonest form of CDSS is the knowledge-based systems. Clinical knowledge, often in a limited medical discipline, is embedded into system. An inference engine is then able to use this knowledge in order to provide patient specific advices. Typically, medical knowledge in the knowledge base is represented through
a set of rules, but other approaches can be applied, such as Bayesian Networks.

Despite the long time research in this field, inter-user variability problem has not been taken into account in the development of the system [11]. For instance, searching for scientific articles in PubMed\(^2\) with the Medical Subject Heading (MeSH) terms: “Decision Support Systems, Clinical” and “Observer-variation”, resulted in only a dozen papers [12,28,29,30,31,32,33,34,35,36,37]. Most of them state that computerized aid may help in reducing or directly overcome the inter-user variability problem and propose methods to achieve this goal.

Ambrosiadou et al. [35] employ a DELPHI approach to help a number of diabetologists to arrive at a consensus about insulin administration regimes. They point out that the consensus facilitates performance evaluation and further knowledge acquisition to be used by DIABETES, a computerized decision support system for insulin administration.

Some of the systems deal with medical images, as it is difficult to formally express the content of an image, posing the basis for subjectivity and therefore inter-user variability. In many case, automated image processing is advocated to overcome this problem. Van der Laak and colleagues [33] applied image processing, a decision tree and linear discriminant analysis to automatically identify diploid reference cells in the field of cytology. Their system avoids manual selection which is poorly reproducible. Coutts et al. [31] advocate the use of image based computerized support using in the assessment of perfusion volume in case of ischemic stroke, due to the poor inter-user agreement even among experts. Wenzel [34], on the other hand, evaluated the performances of an automated system for caries detection and showed that its performances were even lower than human observer. In addition, inter-observer agreement between dentists did not improve by using

\(^2\) PubMed, National Library of Medicine, National Institute of Health
the system. Hu et al. [29] propose a framework for identifying salient visual features through eye tracking. They aim at improving the quality of decision support system by discovering those factors that consciously or subconsciously are applied by radiologists during visual assessment.

The evaluation carried out by Tsai and colleagues [30] pointed out that the advices provided by CDSS should be careful assessed as they can even decrease physician’s diagnostic performances. They evaluated the diagnostic performances of physicians (non cardiologist) using an expert system for electrocardiograms (EKGs) interpretation. The results showed that physicians agree with the incorrect system output twice as often when using the system than physicians without aid.

Smeets et al. [36,37] showed that a rule based CDSS created on the basis of knowledge elicited from one expert and medical literature agrees with experts at the same levels they do in the field of drug treatment for epilepsy. Provided the low agreement between neurologists, their system is an attempt towards the application of medical evidence medicine in drug treatment of epilepsy. Bindels et al. [32] studied the high level of inter-user variability in the assessment of appropriateness of diagnostic test ordering. A reminder system based on practice guidelines for primary care was developed. They argue that such system can effectively help physicians since practice guidelines are clearly formulated and therefore easily to put in a formal language. Van Ast et al. [28] proposed a method for expert knowledge eliciting with the aim of knowledge base construction, identifying expert with convergent and divergent opinion.

A study more related to the problem addressed in this thesis is that of Cross et al. [12]. They developed and evaluated a CDSS for supporting the cytological diagnosis of fine needle aspirates of breast lesions. The purpose of the system was to help less-experienced pathologists in this diagnostic task. They applied a logistic regression model and an artificial neural network on retrospective data collected by an expert pathologist. The data were coded in binary features and correspond to parameter with great
diagnostic importance and typically used in cytology. A group of 19 pathologists evaluated other 322 specimens in the same way as the expert. While the logistic model and the ANN performed reasonably on the expert data, the performances were not acceptable when applied to the multiple pathologists’ dataset. The authors ascribed this poor result to the high level of inter-user variability among pathologists and advocate a better training for the staff.

Another study facing the inter-user variability is that of Price and colleagues [38]. They developed a decision support system for the histological interpretation of pre-invasive cervical squamous cells. Knowledge and uncertainty were represented through a Bayesian Network. Diagnostic clues were assessed by the user and correspond to various histopathological parameters. The system helped to achieve a slightly better diagnostic agreement among experts and junior pathologists. However, as the input information is entered by the user, it could be the case that the system reaches the wrong conclusion. The disagreement they found in the study is ascribed to the different assessment of the features. They argue that using on-screen image templates, additional consistency could be obtained.

2.2. Machine learning

Machine Learning is part of Artificial Intelligence that deals with the creation of algorithms and methods which enable a computer to “learn” [39].

Machine learning methods were successfully applied to problems that cannot be formally defined, but which data are available for, such as artificial vision, hand-written character recognition, speech recognition, etc. as well as in specific medical applications [39,40,41].

The basic paradigm of machine learning is the creation of computer-based algorithms able to learn from examples, as human beings do, and to discover new knowledge “hidden” in the available data. As a matter of fact, the data mining step in the knowledge discovery in database process (KDD) is based on the use of ma-
CHAPTER 2. STATE OF THE ART

Machine learning tools as well as more standard statistical methods [42].

There is an increasing interest in the application of such tools in medicine given the increasingly widespread use of medical information systems and growth of medical databases. These methods are particularly suited for supporting the interpretation of a variety of complex input sources of medical information, such as electrocardiograms, computerized axial tomography scans, ultrasound images, and, in general for signal processing and analysis. Moreover, machine learning approaches are also applied to biomedical data, particularly for the interpretation of gene and protein expression data.

Medical datasets are characterized by some peculiarities [40]:

- Incompleteness: missing values
- Incorrectness: due to systematic or random noise in the data;
- Sparseness: few data available per patient;
- Inexactness: inappropriate selection of parameters for a given task.

Machine learning tools in medical application were advocated to deal with these characteristics.

The creation of diagnostic, prognostic or models from data has been gaining attention recently due to the diffusion of electronic medical records that allow collecting high number of information to derive their models.

Machine learning methods present very different approaches to the learning task: induction of symbolic rules [43], creation of decision trees [44], approaches coming from statistical or pattern recognition [45,46], creation of artificial neural networks [47], to depict the most common ones.

Machine learning methods were applied in almost every medical field [48]: radiology [49,50], pathology [51], oncology [52,53], cardiology [54,55], just to name a few.

In dermatology, machine learning methods were mainly applied for the automated classification of digital images. In section 2.4 a more detailed review is reported.
2.3. Case-based reasoning in medicine

Case-based reasoning (CBR) has become a successful technique for knowledge-based systems in many domains [56,21]. In medicine, it seems particularly useful because it resembles the clinical analogy reasoning. Moreover, CBR is particularly suited for the automatic acquisition of subjective knowledge.

CBR uses the knowledge contained in previously solved cases to reason about new problems. It consists of two main tasks: the retrieval of the most similar cases, and the adaptation of the previous solution to fit the current case. One of the earliest CDSSs employing CBR was CASEY [57], dealing with heart failure diagnosis. It also employs a rule-based domain theory which can be used when no suitable cases are found.

The application of CBR in medicine requires specific attention for the adaptation task. This is a common topic in CBR, which is especially evident in medicine, where case description usually involves a large number of features. A possible solution to avoid the adaptation task is to build retrieval-only systems. The rationale is that physicians are free to assess the information from the retrieved cases. Example of retrieval only system are mainly in the field of medical imaging [58]. In addition, combining CBR with other artificial intelligence methods, i.e. following a multimodal approach, may overcome the adaptation problem [59,60,61].

Concerning an approach similar to that addressed in this thesis there is the work of Le Bozec and colleagues [62,63,64] on histopathology on breast cancer. In their application, the system uses cases derived from written histopathology reports. Each case has an internal tree structure and the similarity is computed in a complex way, accounting for different feature weighting and medical knowledge of patho-physiology. The information describing the case is a subjective evaluation of several histological parameters. The system retrieves the cases relying on this case description. The inter-variability in the subjective evaluation of the features drove them to build a consensus module that enables the experts
to reach an agreement on the subjective features so as to provide case descriptions which are reproducible, recognizable and clinically relevant [64]. Those features which do not have these properties are discarded.

Current trends of CBR in medicine show that CBR is employed in combination with other artificial intelligence methods [65] due to the complexity of medical domain.

2.4. CDSSs for the early diagnosis of melanoma

Since the early nineties, several studies have been investigated the feasibility of computerized support systems for helping dermatologists in the early diagnosis of melanoma. Nearly all of those computerized systems rely on the automated processing of digital images. This can be viewed as the first crucial step for the creation of automated support systems. This first step can be further divided into two sequential parts: the segmentation, which aims to recognize the region of lesion into the whole image, and the feature extraction, which aims to compute several parameters that describe the image in a numeric way. The second step leads to the creation of a classifier that employs those features as input and produces a diagnosis as output. Different classification algorithms were used for this diagnostic purpose: linear discriminant analysis [66], k-nearest neighbor [67], decision trees [68][69], neural networks [70,71], support vector machines [72], a combination of methods [73], etc.

The evaluation of these types of computerized support systems, in literature is typically described in terms of sensitivity and specificity, as usually do physicians. Some of them use the area under the curve (AUC) of the corresponding Receiving Operating Curve (ROC) [74]. One paper evaluates the volume under the curve (VUC) because the authors take into account dysplastic lesions, a third class of skin lesions, besides the malignant and the benign ones [75]. A more detailed description of the advantages and drawback of these systems is described in section 3.2.1.
Chapter 3

3. The Problem

This thesis focuses on the development of a CDSS addressing the problem of inter-user variability in the assessment of clinical parameters, which comes into play in different phases of CDSSs user interaction.

The first section of the chapter describes the methodological problem, and the second section illustrates the clinical problem regarding the early melanoma diagnosis.

Inter-user variability is a well known issue in medicine. The Medical Subject Headings (MeSH) of the National Library of Medicine define inter-user variability under the term “Observer Variation” as:

“The failure by the observer to measure or identify a phenomenon accurately, which results in an error. Sources for this may be due to the observer’s missing an abnormality, or to faulty technique resulting in incorrect test measurement, or to misinterpretation of the data. Two varieties are inter-observer variation (the amount observers vary from one another when reporting on the same material) and intra-observer variation (the amount one observer varies between observations when reporting more than once on the same material).”

It is well accepted that inter-user variability may be caused by subjectivity and by the lack of absolute reference values. These are common factors in medicine, whenever no clear and definite knowledge of a disease is available [76].

The assessment of inter-user variability in medicine has been investigated a lot, mostly in the field of radiology and human pathology. The need for reducing as far as possible this problem es-
especially in those disciplines which, in many cases, have a potential effects on diagnosis and treatment (histo-pathology for instance, provides a gold standard reference for the definition of many kinds of diseases) has led to protocols for quality control and to constant training sessions to ensure comparable levels of performances among physicians (e.g. in cytology, radiology, etc.). Subjectivity is a quite common characteristic in many decision processes. Large part of knowledge is not explicit, but tacit. Tacit knowledge is personal, context specific and thus hard to formalize and communicate. On the other hand, explicit knowledge can be easily communicated through a formal language for representation. Medical textbooks are examples of explicit knowledge; instead apprenticeship deals with tacit knowledge. As a matter of fact, medical training all over the world includes some form of apprenticeship that helps to teach the implicit medical knowledge to novices.

From the clinical viewpoint, Evidence Based Medicine (EBM) [77] is a new paradigm that aims at reducing subjectivity providing a set of scientifically-based advices to physicians. EBM relies typically on randomized controlled trials that provide the ultimate source of medical knowledge. From the medical informatics viewpoint, this new clinical paradigm has led to the development of computerized clinical guidelines exploiting the research in knowledge representation already active in medical informatics. From a Knowledge Management (KM) perspective, EBM and computerized guideline support the dissemination of up-to-date scientifically sound medical knowledge trying to ensure the best quality of care to the patients. The kind of knowledge which EBM deals with is explicit knowledge. This new paradigm is important to speed up the process to transfer results of the research into the clinical practice, but it does not deal with the implicit or tacit knowledge always present in the medical practice, especially when general rules have to be applied in a real situation [78]. The development of computerized guidelines, initiatives for continuous medical education help the dissemination of the medical explicit knowledge.
Generally, CDSSs typically deal with this explicit knowledge; however tacit knowledge, may still play a great role and affect the performances of those systems. To our knowledge, the research on CDSSs has not properly addressed this problem so far.

### 3.1. General model of a CDSS

The general model of clinical decision support system can be described by four main components: input data, a knowledge base, an inference engine and the output data. Figure 1 shows those components and their relations [79].

![Diagram of General Model of a CDSS](image)

**Figure 1 – General model of a clinical decision support system.**

Within this framework, the user supplies input data to the system and the system provides its output by using the inference engine. The inference engine combines the data provided by the user and the data included into its knowledge base to produce its output. The knowledge base encodes the medical knowledge which is relevant for the system to solve the tasks it was built for. Depending on the choice of the inference engine, the knowledge base should encode this knowledge in a proper representation. As an example, if the inference engine is a production rule system which employs predicate logic for combining statements, the knowledge base should include basic relation of symptoms and diseases.
We can abstract this model for decision support systems as a “functional” relation among some input variables (e.g. symptoms, clinical findings, etc.) and the output (e.g. differential diagnosis). In a diagnostic support system, let’s $x$ be the set of input data and $y$ the output class. The CDSS can be described by a “transform function” $h$ that associates $x$ to $y$.

**Figure 2 – Abstract view of a CDSS.**

It is worth noting that this is not a function in a strict mathematical sense. It is only an abstract representation of the general CDSS model.

The history of CDSSs shows a great effort in the definition of proper knowledge bases and inference engines, thus the definition of “function” $h$, to properly support physicians in their everyday tasks.

The rationale for this framework is that physicians’ knowledge can be incomplete in various ways: personal knowledge can be outdated or imprecise, medical knowledge contains uncertainties that physicians are not able to properly manage (due to the lack of definite biological understanding of disease), the development of biomedical equipment and instrumentation expanded investigative and therapeutic possibility of physicians, but overwhelming them with new information, etc.

Therefore, CDSSs can help them in managing this quantity of information in a better and more proper way to improve quality of their decisions.

Among the problems in the development of CDSS, the management of uncertainties in the medical knowledge has been deeply investigated, giving rise to a various set of methodological tools and solutions.
Those uncertainties were typically included into the knowledge base through various approaches: probabilities in Bayesian systems, certainty factors in rule based system, to name a few.

However, let us suppose we have a perfect knowledge of the patho-physiologic mechanism of a disease, i.e. we perfectly know how to diagnose a certain situation given a set of clinical findings. In this scenario a CDSS would apply its inference engine (activate a rule, compute a function, calculates probabilities, etc.) and provide the correct output given those clinical findings. This is an ideal situation which is far to be common in medicine.

If the clinical findings are subjective, their assessment depends on the specific skills and on a set of personal judgments of the physician. If the degree of inter-user variability of those parameters is very high, the CDSS may fail in providing a diagnosis, even if “perfect” medical knowledge is available.

Obviously, this is an oversimplified example, where the effects of inter-user variability are extremely emphasized. Nevertheless, in a real situation, the lack of definite biomedical knowledge makes the medical diagnostic process even more difficult.

Cross et al. [12] reports that inter-user variability was the main causes of failure of a support system for the cyto-diagnosis of fine needle aspirates of breast lesions. In their CDSS, diagnostic models were created by using logistic regression and artificial neural networks on data collected by an expert physician. After comparing and assessing the performances of those models through hold-out methods on the expert data, they were used by multiple physicians. The inter-user variability in the assessment of the features was identified as the principal cause of the inadequate performances of the system when used by the other physicians.

Another aspect to consider in the development of CDSSs regards the output of a support system: besides accurate, should also be valuable for the physician. In other words, a system should provide new and enough information to the user and that information should be interpretable and comprehensible. This aspect is fun-
CHAPTER 3. THE PROBLEM

damental as the final aim of a CDSS is to improve the performances of unaided physician: a CDSS is useless if it provides users with information they already know [13].
The demise of the “Greek Oracle” model for a CDSS has become clear since the early nineties [80]. Within this model, a CDSS is supposed to provide the correct diagnosis and explain its reasoning, reducing physician’s work to data input.
A critiquing approach emerged to face this problem, trying to couple user’s knowledge with system’s abilities. Within this framework the system generates comments based on the user’s choices. This model has been typically applied for the treatment, drug prescribing and care planning. Once more, the typical critiquing systems rely on explicit knowledge to provide their advices.

A further aspect to consider is the applicability of such kind of systems in real clinical settings. Their seamlessly integration into clinical workflow has been recognized as a key aspect for their successful use. The current pressure on healthcare organizations to ensure both quality of care and cost containment is driving them towards a more effective management of medical knowledge. CDSSs are important tools to help accomplishing this goal.

3.2. Early melanoma diagnosis
In the specific medical field of this thesis, dermatology, the above explained aspects are presents in various degrees. In the early recognition of melanoma, unfortunately, there are no clinical signs or physiological parameters that can diagnose malignant lesions without uncertainty. In other clinical situations, this could not be the case, as for instance in diagnosing Type I diabetes.

3 A fasting plasma glucose test result greater than 125 mg/dl, confirmed in a second time, diagnoses Type I diabetes.
Dermoscopy (see Appendix 8) is a powerful tool that increases the diagnostic accuracy of the dermatologist. Nevertheless, dermoscopy enhances the inter-user variability problem, since many scientific studies claim the need for specific training of the user before the technique would become effective.

To be more explicit, the features that are assessed by dermatologists depend on the visual interpretation of skin lesions. Therefore this is typical pattern recognition situation, where, even for experts, explicitly expressing through formal languages their reasoning is extremely difficult, if not impossible. As a matter of fact, the most effective clinical protocol for the early diagnosis is called pattern analysis. It comprises a thorough evaluation of the dermoscopy parameters considered on the whole. The typical training consists in evaluating skin lesions images on textbooks or atlas illustrating those parameters in exemplary cases.

Recognizing that dermoscopy analysis is an effective procedure for early recognition of malignant lesion and that this analysis requires well-trained personnel, a number of researches focused on the development of automated computerized support through image analysis techniques (see 2.4). The rationale for those systems is to avoid inter-user variability by substituting the clinical inspection of skin lesions by automated analysis of dermoscopy images.

Next section discusses the advantages and the drawbacks of the current decision support systems in the field of melanoma diagnosis.

3.2.1 Advantages and drawbacks of current CDSSs systems for early melanoma diagnosis

The main advantage advocated by proposed image-based support systems is their independence from subjective evaluation of skin lesions. Actually, the extracted features are objective and reproducible over time, i.e. they do not suffer of inter or intra-observer variability problems. Ideally, image-based support systems could be used by non-specialists or even directly by the patient them-
selves. However, these systems are affected by some important drawbacks [81].

As the image-processing phase entirely relies in the “numbers” contained in the digital image, the “stability” of the image represents a very critical aspect. As a matter of fact, image-based systems are strictly dependent on the image acquisition device and on the image acquisition procedure. For instance, a variation in the illumination can lead to a complete change in the colorimetric characteristics of images. Moreover, different image acquisition devices have different colorimetric responses, and thus images coming from different devices are not “numerically” comparable. The standardization of colorimetric characteristics of digital images is still an open problem. This leads to difficult, if not impossible, comparison or integration of different acquisition devices [81]. Moreover, image-based systems are extremely susceptible to failure due to poor quality images. In addition, some acquisition devices cannot acquire images of large or prominent lesions.

In addition, each system performs segmentation and feature extraction in its own way. Segmentation regards the identification of the relevant or interesting objects from the background; whereas feature extraction means the computation or determination of features that describe, typically numerically, the image.

At present, there is no universally accepted segmentation method that has been proven to work on a large representative image database [81]. Furthermore, there are no well-defined standard features which are proven to be effective for diagnostic purposes. Most of the computerized systems attempt to identify diagnostic patterns similar to those assessed by dermatologists, but no standard image-processing algorithm has been described yet. In addition, new research projects are required to collect their own images, and develop their own technique more or less from scratch [81]. In this context, the basic characteristic of research, i.e. building models and applications on previously done work, cannot be performed effectively.

Furthermore, the extracted features, though objective, usually represent low-level concepts (for instance: mean value of red, stan-
standard deviation of hue, etc.). They are difficult to interpret by clinicians. Moreover, a classifier's output is based on some combination of those features; therefore the suggested diagnosis is somehow difficult to understand. In medical diagnosis it is crucial that a computerized system is able to explain and justify its decisions. Physicians do not like black boxes that tell them what to do in critical situations without any justification or explanation [82].

All of these critical problems could have prevented the development and the dissemination of systems used by dermatologists as a support. As noted by Day and Talbot [81], despite the length of time that research has been conducted, no automated diagnosis tool is in standard clinical use. Among the hurdles to be overcome, they underline practical feasibility of such instruments and the acceptance from the medical community and the patients as major concerns. Regarding these issues, some aspects are still a matter of research.

Published works show that such system could outperform general practitioners, at least. However, the evaluation of these support systems is usually focused on the performances of the classifiers in artificial experimental settings. Nothing has been said about their practical feasibility in a clinical setting. None of the systems has been evaluated in order to understand their supporting capability, i.e. testing whether the information provided (a diagnosis) is really “valuable” to the dermatologist. As already mentioned, this is a more general issue that affects decision support systems in medicine. Rousseau et al. [13] recently reported the results of a study aimed at understanding the factors influencing the adoption of clinical decision support system in the general practice. In this study, several barriers were found to the use of clinical decision support systems. Among the key issues, relevance and accuracy of messages were highlighted.

A way to overcome some of these problems could be moving the attention from the automatic processing of digital images to the information provided by dermatologists, as in [83]. The main advantage of such support systems is the avoidance of all the problems related to image acquisition devices, color calibration, re-
prodicibility of digital images, etc. Moreover, clinical information usually represents high-level concepts, thus potentially improving the interpretability of the system outputs, and, in the end, resulting in a greater use by dermatologists.

Nevertheless, the crucial drawback of this kind of systems is due to the dermatologists’ diagnostic variability in the evaluation of skin lesions. In fact, the most important features for early diagnosis are mainly related to ELM analysis. However, as already mentioned, only experienced dermatologists carry out this procedure in an effective way. As a matter of fact, the correct recognition of ELM criteria can be difficult or even misleading, for physicians unaccustomed with this clinical evaluation of the skin [16,84]. Therefore, it is mandatory to deal with this problem in order to build an effective clinical decision support system for the early recognition of melanoma.

In the next chapter, the solution we propose for the development of a clinical decision support system accounting for this problem is described.
Chapter 4

4. The Proposed Approach

As previously described, the inter-user variability in the assessment of patient’s condition can affect the performances of Clinical Decision Support Systems, if they do not properly deal with this problem. Current trends of CDSSs research aim both at the combination of different problem-solving approaches and at the actual integration of these systems into clinical practice. In this thesis a novel architecture for Clinical Decision Support Systems is proposed to properly deal with this problem. The general framework is shown in Figure 3. The scenario of the early recognition of melanoma is particularly suited for this investigation, as described in chapter 2 and 3.

Figure 3 – General architecture of the User-tailored CDSS.
The proposed CDSS is based on two modules: a critiquing system and a consulting module; both of them relying on a reference dataset. The specific composition of the reference datasets for melanoma is described in section 5.1. However, the proposed approach is independent from the specific medical domain. Next section describes in more details the components of this CDSS with reference to the melanoma scenario.

4.1. User-tailored CDSS: components
The principal component is the reference dataset. This dataset is composed by a set of known cases, i.e. cases for which a "gold standard" exists. In the melanoma scenario the gold standard is represented by the histological diagnosis. Each case in the reference dataset, i.e. each pigmented skin lesion, is described by a set of objective features, such as patient's age, lesion location, images, etc. In addition, subjective features, i.e. features that can be differently assessed by different physicians (e.g. dermoscopy parameters, see 5.1) are included. A group of expert physicians with several years of experience in dermoscopy, evaluates each case in the reference dataset independently from each other. They provide the subjective clinical parameters as well as their diagnostic and or treatment decisions. The reference dataset together with expert physicians' evaluations is the reference case base. The reference case base includes the implicit knowledge of experts through their subjective evaluations and clinical decisions.

Each new user of the CDSS is first required to evaluate the reference dataset by providing the subjective evaluation of the cases as well as the clinical decisions, in the same way as experts did. The reference dataset with the user's subjective evaluation and clinical decisions is the user's case base4.

4 We will refer to the user’s case base also with physician’s case base.
Upon completion of the reference dataset evaluation, the two modules can be implemented. In fact, such evaluations allow:

- Building a critiquing module by applying a novel machine learning-based approach, able to use the evaluations of the single physicians and to exploit his/her diagnostic skill. The critiquing system applies a classification model, learned by focusing on the most “difficult” cases of the reference data set, i.e. the cases in which user’s decision was wrong.
- Building a consulting system by comparing the user subjective evaluations with the ones contained in the reference case based. This allows: i) assessing similarities between the user and the experts in evaluating the lesions; ii) selecting the expert “more close” to the user in terms of lesion subjective evaluations; iii) assessing the diagnostic and treatment capabilities of the user with respect to the experts ones. The consulting system enables the retrieval of previously solved cases.

Scenario
The CDSS is intended to be used by different kinds of users: dermatology students, dermatologists with little experience in dermoscopy, general practitioners, etc.
The main advantage with respect to other CDSSs for the early recognition of melanoma is that it can be used without having digital image acquisition devices. This can be an important aspect, for instance, when used by a dermatologist who “moves” through different outpatient clinics, or by general practitioners who cannot afford expensive image acquisition devices.
Furthermore, the consulting module can also be employed as an “intelligent” tutor for dermatology students. It is different from standard dermatology textbooks since it can show how experts deal with similar challenging cases not only in diagnostic terms, but also regarding the treatment choice as well as the clinical parameter evaluations.
CHAPTER 4. THE PROPOSED APPROACH

Once the single physician completes the evaluation of the reference dataset, the CDSS can be used through two main modalities: the critiquing and the consulting.

When a new case is evaluated by the user and provided to the system, the critiquing module would suggest the user to re-evaluate his/her decisions if deemed necessary, i.e., when the clinical decisions and model outputs are different. The focus here is on the diagnosis. The critiquing module can be viewed on top of an electronic medical record, where the user already collects patients’ information. This interaction modality for the CDSS is therefore that of unsolicited advices: while the user is inserting a case, the system activated itself autonomously.

On the other hand, the consulting module requires the user to explicitly request the advice of the system. The system then shows previously solved cases by the experts, by computing a similarity measures between the case at hand and those in the reference case base (see section 4.1.2 for a more detailed description). Despite the retrieval approach the system applies, the user is first required to assess the actual similarity of the retrieved cases by comparing the images and the objective data. In case the retrieved cases are judged comparable, the user can assess how experts evaluated the subjective features, which their decisions, i.e. diagnosis and treatment choice, were, on real similar case. If this is not the case, the user can also evaluate why the similarity in the feature description of that case resulted in rather different cases.

The purpose of this module is to provide the user with the experts’ tacit knowledge: in other words, it provides some sort of “computerized” apprenticeship.

Up to now, the CDSS has been implemented in a prototypical web-based tele-dermatology application, which aims at sharing information among dermatologists and general practitioners on a wider area of dermatological problems.
4.1.1 Critiquing module

The critiquing module employs a machine learning algorithm to provide alerts to the user of the CDSS. The alerts are tailored on the specific skill of that user. The machine learning algorithm provides the advice only when deemed necessary, i.e. when the output of the system is different from the clinical decision. In the following paragraphs this modality is described in more details. Machine learning is particularly suited in medicine whenever no clear definite biomedical knowledge is available and when it is difficult, or impossible, to formally define a problem.

Typically, machine learning methods are applied within the field of supervised learning. In this context, the task is to learn general models from a set of specific examples, i.e. to generalize information from known data to unseen data. For example, given a set of patient with cancer and a set without cancer, each case can be described by a set of patho-physiological measurements. A Machine Learning algorithm can discriminate the ill patients from the healthy by recognizing some complex relation among the measurements.

For sake of completion, another topic of research in machine learning is the unsupervised learning. In this context, the algorithm finds natural associations in the data without knowing a target class. For the purpose of this thesis, supervised learning is used.

The next section briefly discusses machine learning methods in the context of clinical decision support systems.

Machine learning

Machine learning methods were successfully applied to problems that cannot be formally defined, but which data are available for, such as artificial vision, hand-written character recognition, speech recognition, etc. as well as in specific medical applications [39,40].

The basic paradigm of machine learning is the creation of computer-based algorithms able to learn from examples, as human be-
CHAPTER 4. THE PROPOSED APPROACH

ings do, and to discover new knowledge “hidden” in the available data.

In a classification problem, such as the diagnostic one, the learning task can be the identification of a model able to discriminate two categories of patients (ill and healthy ones) from a proper combination of clinical measurements. More specifically, machine learning methods attempt to find a function $h$ able to map a given input vector $x$ to a class label (or target) $y$. The vector $x$ includes the parameters, or features, describing the case. The corresponding class label $y$ describes the class the case belongs to. In our example, the vector $x$ represents the patho-physiological measurements; whereas the class label $y$ describes patients with cancer and healthy patients.

During the training phase, the learning algorithm creates the model $h$ which is then used to make prediction on new cases. Those cases are described only by the vector of features $x$. The learned model $h$ is able to make a prediction on the class $y^*$ the new case belongs to. Figure 4 depicts the typical workflow for building predictive models.

![Figure 4 – Usual machine learning-based approach.](image)

The main advantage of the application of such methods for clinical decision support systems is that no formal knowledge is required to create the model. In medicine, whenever no detailed or
definite biomedical knowledge of a disease is available, those methods present some advantages over knowledge based approaches.

Figure 4 shows a schematic process diagram of what typically happens by employing machine learning algorithms to create computerized support systems.

In this scenario two aspects may have a potential effect on the application of machine learning algorithms for building support systems.

The first one is the nature of input data. Whenever input data are subjective evaluations of medical parameters, and therefore may undergo the inter-user variability problem, the model \( h \) created using a certain training data (e.g., expert data) may fail to provide correct predictions if used by other physician (e.g., novices).

A second aspect regards the output of the decision support system. Even a good model \( h \), i.e., a model with good performances itself, may not provide new or enough information to its user. If this is the case, user’s performances would not improve with system aid.

We argue that a “local” model, tailored on a each physician and used by him/her, supports better his/her future decisions, by exploiting his/her specific skill and expertise. Instead a standard model, based, for instance, on expert data may fail when used by other physicians with different, personal interpretation of clinical findings.

The rationale is that physician’s subjectivity evaluation of parameters is stable over time, i.e., intra-user variability is lower with respect to inter-user variability.

**PhysicianBoost**

We propose a novel approach to deal with inter-user variability and supporting capability in developing decision support systems through the application of machine learning methods.

The proposed method for the critiquing module, called PhysicianBoost, builds a model for each physician (a local model), starting from his/her feature evaluations, and then combines its
CHAPTER 4. THE PROPOSED APPROACH

output with clinical decision through a suitable combination rule. The final aim is improving overall performances, i.e. the performances of physicians with the aid of the support system. The system accounts for both the physician’s capability of correctly classify cases and his/her assessment of feature values, which may turn out to be correlated.

An abstract view of the system is illustrated in Figure 5.

Figure 5 – Abstract view of PhysicianBoost

Training data provided by the single user must include the subjective evaluations (e.g. dermoscopy features), clinical decisions (e.g. clinical diagnosis) and the “gold standard” (e.g. histological diagnosis).

One could argue that if the learning algorithm is based on the data provided by the single physician, the learned model would be no better than the single physician. A CDSS of this kind would be therefore useless for the user.

Keeping in mind that the CDSS should support the user providing them with new or enough information to take proper decision, PhysicianBoost first assess user’s performances by comparing the clinical decisions to the “gold standard”. Then the learning phase of the algorithm is focused on the most difficult case for the physician, i.e. the cases for which physician’s conclusions were different from the gold standard. Finally system’s output and clinical
decisions are properly combined to exploit their respective abilities. The combination is problem dependent and may include cost-sensitive issues, rather common in medicine.

The combination rule constitutes the basis for the critiquing module. As a matter of fact, the choice of the combination rule affects the way the system would provide the advices to the user.

In a preliminary study we analyzed this model in an artificial setting [85]. The results, briefly presented in 5.2, were promising to lead to the development of a more complete system.

A more detailed description of the critiquing approach for the CDSS is illustrated in the following paragraphs.

In Box 1 the pseudo code of our approach for binary classification problem is outlined.

In step 2, physician’s error is computed accounting for the cost-sensitive nature of many medical decision tasks, i.e. through sensitivity and specificity (see 5.3.2). Moreover, this choice also account for unbalanced dataset, another typical aspect of medical dataset. In an unbalanced datasets the distribution of the examples in the classes is not uniform.

Before starting the learning phase the method weights more physician’s misclassified instances. These newly weighted instances, are then used as input of the learning algorithm.

This weighting procedure is one of the methods to force model $h$ in focusing to the “most difficult” cases for the physician. Next, the output of the model, i.e. the hypothesis $h(x)$, is combined with the physician diagnosis $\text{diag}(x)$.

We propose to use a simple voting scheme that is able to account for cost-sensitive issues. For example, we can tune the combination to enhance sensitivity or specificity in a simple way: to enhance sensitivity (specificity) it is sufficient to output negative (positive) only if both “models” output negative (positive).
CHAPTER 4. THE PROPOSED APPROACH

Box 1: PhysicianBoost: a critiquing user tailored decision support system

<table>
<thead>
<tr>
<th>Given</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Training Set (TS) = {w_i(x_i, y_i), i = 1, \ldots, N}, x_i \in \mathbf{X}; y_i \in {0,1}</td>
</tr>
<tr>
<td>- \text{features } x_i \text{ assessed by a single physician; } y_i \text{ is the target class; } w_i = 1 \text{ is the weight of } i\text{-th instance; } w_i \in (0,1].</td>
</tr>
<tr>
<td>- Learning algorithm: \text{LA.}</td>
</tr>
</tbody>
</table>

Do:
1. Compute physician’s sensitivity (Se) and specificity (Sp).
2. Compute error \( \varepsilon = (1 – Se) + (1 – Sp), \) \( \varepsilon < 1/2. \)
3. Compute \( \beta = \varepsilon / (1 – \varepsilon), \) \( \beta \in (0,1]. \)
4. For all instances correctly classified by the physician set \( w_i = \beta. \)
5. Train the learner \( \text{LA} \) using the weighted TS* = \{w_i(x_i, y_i)\}
6. Get hypothesis \( h(x) : \mathbf{X} \rightarrow \{0,1\} \)
7. Combine physician’s diagnosis \( \text{diag}(x) \) with \( h(x): \) \( y^* = \text{comb(diag}(x), h(x)) \)

PhysicianBoost approach reminds the boosting approach [86], where weak learners are trained subsequently on weighed training samples and eventually combined together to give an overall prediction.

We developed PhysicianBoost in Java using Weka package libraries [87].

Learning algorithm: Naïve Bayes

The method we proposed does not rely on a specific choice of the learning algorithm. We chose Naïve Bayes as learning algorithm due to its stability property given the small dataset we have at our disposal. Naïve Bayes error has a low variance component with respect to other methods [88].

Moreover, Naïve Bayes output, a probability, is supposed to be also quite interpretable for physicians. Moreover, the decision of a Bayesian classifier can be interpreted as a sum of information gains of the features for or against the given class [89].
Interpretability is one of the peculiarities of the application of machine learning methods in the medical field [40,90]. Regarding the latter aspect other methods are certainly more suitable for interpretability, such induction rule algorithms, decision trees, etc. However, their instability is also well known. Naïve Bayes applies the Bayes rule of conditional probability to compute the posterior probabilities of a class \( c_i \) given the values \( x_j \) of all the \( n \) attributes for a given instance. Bayes rule states that:

\[
p(c_i | x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n | c_i) p(c_i)}{p(x_1, \ldots, x_n)}
\]

Assuming the conditional independence of the attributes, given the class then:

\[
p(x_1, \ldots, x_n | c_i) = \prod_{j=1,n} p(x_j | c_i)
\]

and therefore:

\[
p(c_i | x_1, \ldots, x_n) = \frac{\prod_{j=1,n} p(x_j | c_i) p(c_i)}{p(x_1, \ldots, x_n)}
\]

Since \( p(x_1, \ldots, x_n) \) is independent from the class, it can be ignored when comparing values of \( p(c_i | x_1, \ldots, x_n) \) for the different classes \( c_i \) to choose the class that maximizes this quantity. In the end, we have that:

\[
p(c_i | x_1, \ldots, x_n) \propto \prod_{j=1,n} p(x_j | c_i) p(c_i)
\]

### 4.1.2 Consulting module

The consulting module is based on the physician’s and experts’ evaluations included into the reference case base.
CHAPTER 4. THE PROPOSED APPROACH

The purpose of this module is to share expert’s implicit knowledge contained into the subjective evaluations of cases in the reference case base. The consulting module adopts case based reasoning paradigm to provide the user with expert knowledge. Case based reasoning is particularly suited for medical decision making when subjective knowledge is involved [21,22]. Moreover, reasoning with cases corresponds with the typical decision making process of physicians; incorporating new cases in the reference case base updates the implicit knowledge. The consulting module is supposed to be activated by a specific request of the current user of the system. The scenario is therefore the following:

- the physician examines a new patient;
- he/she collects the objective data, if available and requested by the system;
- he/she provides the system with the subjective information as well as clinical decisions: diagnosis, treatment, etc.
- the physician then asks the system to seek for previously solved cases similar to that at hand;
- the system compute a similarity measure with the cases into the reference case base, ranking the cases accordingly;
- the user can browse the ordered cases.

The retrieved cases entail both the subjective feature evaluations as well as the clinical decisions of the experts. By comparing the retrieved cases in terms of subjective feature evaluations and clinical decisions, we argue that the module support the dissemination of expert tacit knowledge. The consulting module benefits from the multiple independent evaluations of the same cases given by the experts. The most important point here is the retrieval of the similar cases. Again the inter-user variability prevents the naive application of search queries into the reference case-base.
Therefore, to account for this problem, two approaches were investigated to retrieve the similar cases:

1. we assess the similarity between the new case with those previously evaluated by the single physician included into his/her own reference dataset. The similarity provides a ranking of the cases in the reference dataset. The user can see his/her “closer” examples, as well as directly compare experts’ evaluations on those cases in the common reference case base;

2. we first assess who is the expert most similar to the single physician through the common reference case base. Once the “most similar” expert is identified, the system searches for the most similar cases in that expert’s reference case base.

Despite which reference case base we search for the retrieval of cases, both approaches rely on a notion of similarity among cases. This similarity has to deal with numeric and nominal features as well as missing values. Next section describes the approach we adopted regarding this topic.

**Similarity**

Assume that a case $u$ is defined by a set of $m$ attributes $(u_1, \ldots, u_m)$. The similarity among two cases $u$ and $v$ is defined as:

$$sim(u, v) = 1 - \frac{1}{m} \sum_{i=1}^{m} d(u_i, v_i)$$

Where $d(u_i, v_i)$ is a distance measure between the values $u_i$ and $v_i$ for the features $i$. We employed two types of distance measure to account for the numeric and nominal features:

$$d(u_i, v_i) = \begin{cases} 
1 & \text{if } u_i \neq v_i \\
0 & \text{if } u_i = v_i 
\end{cases}$$

for nominal features.
CHAPTER 4. THE PROPOSED APPROACH

\[ d(u, v) = \frac{u_i - v_i}{\max(i) - \min(i)} \]

for numeric features

where \(\max(i)\) and \(\min(i)\) are the maximum and minimum of attribute \(i\), respectively. The normalization factor for numeric features ensures that \(d(u, v) \in [0,1]\).

To deal with missing values, we chose to assign \(d(u, v) = 0.5\), if one of the parameter is missing. The rationale of this choice is to consider a missing data somehow between the complete agreement (\(d(u, v) = 0\)) and the complete disagreement (\(d(u, v) = 1\)).

The distance measure divided by the number of the features such that \(\text{sim}(u, v) \in [0,1]\).

Concordance

In the second approach, we first assessed the similarity of the user to the group of expert, concerning subjective feature evaluations. We estimated this similarity through the Cohen’s Kappa, which is a robust measure of concordance between categorical data [91]. It has been developed in the context of psychological studies to measure the concordance between two subjects in classifying objects into mutual exclusive categories. It is defined in the following way:

\[ \kappa = \frac{p_{obs} - p_{exp}}{1 - p_{exp}} \] (1)

where \(p_{obs}\) is the proportion of observed agreements and \(p_{exp}\) is the proportions of expected agreements. The possible values of \(\kappa\) range from -1 to 1. A \(\kappa\) equal to zero means that agreement is entirely attributable to chance; \(\kappa\) greater (less) than zero means that agreement is greater (less) than that expected only by chance.

Given a contingency 2x2 table, \(p_{obs}\) is the sum of the diagonal cells divided by the total number of elements in the table; whereas \(p_{exp}\) is the sum of the diagonal.

In case of ordered categories, we used the weighted form of Cohen’s kappa. The weighted method takes into account the rela-
tive concordances, besides the absolute concordances. The rationale for this choice is that two close ordinal values should be considered differently with respect to a greater difference. In other words, the difference between category 1 and 2 should be considered less than the difference between category 1 and 3. The weighted kappa includes this aspect into the computation. We chose a linear weighting:

$$w = 1 - \frac{|distance|}{maxpossibledistance}$$  \hspace{1cm} (2)

To decide which expert’s case base were to use, we first computed the Cohen’s kappa for each subjective features $f$ between the physician $i$ and the expert $e$. We then averaged the results over all features, obtaining the average agreement between the physician and the expert. The expert $\hat{e}$ with the maximum agreement in terms of kappa was selected. In the end, the search phase was performed into the reference case base of expert $\hat{e}$, in the same way as in the previous approach.

**Retrieved information**

With both approaches, the user can view the retrieved cases in terms of the complete information: objective data (patient’s age, lesion location, etc.), digital images and the own or expert’s subjective evaluations and decisions. Moreover, the module can show the overall similarity value, as well as the partial similarities computed using the subjective features or only the objective ones. They are straightforwardly defined as:

$$sim_{subj}(u,v) = 1 - \frac{1}{m_{subj}} \sum_{i \in Subj} d(u_i,v_i)$$

$$sim_{obj}(u,v) = 1 - \frac{1}{m_{obj}} \sum_{i \in Obj} d(u_i,v_i)$$
where $\text{Subj} = \{ i \mid \text{feature } f_i \text{ is subjective}\}$ and $\text{Obj} = \{ i \mid \text{feature } f_i \text{ is objective}\}$.

These measures helps physician to evaluate the reason why the system proposes those cases and are useful to assess the relevancy of the retrieved cases with respect to that at hand.
Chapter 5

5. Experimental Results

In this section we focus on the analyses that have been performed for the supporting modules of the CDSS system, i.e. for the critiquing and consulting modules. All the descriptive and statistical computation were performed with R statistical package [92].

5.1. Dataset description

The dataset is made up by 177 pigmented skin lesions, diagnosed as 76 malignant melanomas and 101 benign lesions by histological examination. The histological diagnosis represents the “true class” and it is accepted as the ultimate diagnosis. All pigmented lesions were consecutively clinically and histologically examined at S. Chiara Hospital in Trento (Italy) during the usual activity from June 1999 to September 2002. The features describing a case correspond to the actual information collected in the Department of Dermatology of S. Chiara Hospital during usual clinical practice, which are recorded into the Integrated Electronic Medical Record of the Department [93].

For each case a dermatologist recorded objective information, such as patient’s age, lesion evolution, etc., during face-to-face examination, and acquired digital images of the lesions. Different image acquisition devices were used to acquire pictures of the lesions, due to technological innovation. From an automated image processing perspective, this fact prevents the possibility to develop an automated support system, because up to now, no image processing based system is robust enough to adapt to various kind of digital images.
The lesions in the dataset were judged suspected by the dermatologist and were surgically excised to undergo the histological analysis. After the histological examination, which provides the gold standard diagnosis, the case was inserted into the reference dataset. If no images or no objective data were taken, the lesions were not included into the dataset.

Table 1 reports the distribution of the malignant lesions according to the Breslow thickness [102]. The upper-bound values of 5 year survival according to Breslow thickness are also reported. If other criteria such as ulceration, number of lymph nodes, metastases, etc. are present, the 5-year survival is even poorer.

Among the lesions, there were 33 dysplastic nevi. Those lesions were included into the dataset as benign lesions, because they are no longer considered direct precursor of malignant melanoma. Nevertheless, the presence of several dysplastic nevi for a patient should be considered as indicator of an increased risk for developing melanoma, instead of a precursor of the malignant cancer [94].

<table>
<thead>
<tr>
<th>Breslow [mm]</th>
<th>Number</th>
<th>%</th>
<th>5-year survival [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>In situ, 0-1.0</td>
<td>55</td>
<td>72.4</td>
<td>95.3</td>
</tr>
<tr>
<td>1.01-2.0</td>
<td>14</td>
<td>18.4</td>
<td>89.0</td>
</tr>
<tr>
<td>2.01-4.0</td>
<td>4</td>
<td>5.3</td>
<td>78.7</td>
</tr>
<tr>
<td>&gt;4.0</td>
<td>3</td>
<td>3.9</td>
<td>67.4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>76</strong></td>
<td><strong>100</strong></td>
<td></td>
</tr>
</tbody>
</table>

Each case in the reference dataset was then submitted through a web system to six dermatologists: objective information and a digital image set of the lesion were available to physicians so as to provide them with as much information as possible to perform subjective feature evaluations. In our experimental protocol we used the same clinical parameters that are currently collected in the clinical practice in Trento. In addition to subjective feature
evaluation, physicians also provide their clinical decisions in terms of clinical diagnosis (benign or malignant lesion) as well as the choice for excisional biopsy (yes or no).

Three expert dermatologists, with several years of experience in dermoscopy, evaluated the reference dataset. The reference dataset with the subjective evaluations as well as the clinical diagnosis and treatment choice, constitutes the reference case base. Subjective features were evaluated by each dermatologist independently from others.

Three less-experienced dermatologists were then involved in the evaluation phase of the CDSS.

Table 2 shows the objective and the subjective features describing the cases. The subjective features represent usual parameters that are assessed by dermatologists [17].
Table 2. List of the features in the dataset. Objective features were collected during the face-to-face visit.

<table>
<thead>
<tr>
<th>Objective Features</th>
<th>Type/values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesion location</td>
<td>{head and neck, trunk, arm, sacral, leg, foot}</td>
</tr>
<tr>
<td>Max-diameter</td>
<td>numeric (mm)</td>
</tr>
<tr>
<td>Min-diameter</td>
<td>numeric (mm)</td>
</tr>
<tr>
<td>Evolution</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>Age</td>
<td>numeric (year)</td>
</tr>
<tr>
<td>Gender</td>
<td>{male, female}</td>
</tr>
<tr>
<td>Family history of melanoma</td>
<td>{yes, no, not determined}</td>
</tr>
<tr>
<td>Fitzpatrick’s Photo-type</td>
<td>{1,2,3,4, not determined}</td>
</tr>
<tr>
<td>Sunburn</td>
<td>{yes, no, not determined}</td>
</tr>
<tr>
<td>Ephelis</td>
<td>{yes, no, not determined}</td>
</tr>
<tr>
<td>Lentigos</td>
<td>{yes, no, not determined}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>Type/values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>Irregular Border</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>Number of colors</td>
<td>ordinal [1-6]</td>
</tr>
<tr>
<td>Atypical pigmented network</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Streaks (radial streaming, pseudopods)</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Slate-blue veil</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Globular elements</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Comedo-like openings, milia-like cysts</td>
<td>{present, absent}</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>{present, absent}</td>
</tr>
</tbody>
</table>

5.2. Preliminary results

A preliminary investigation was carried out in a more artificial setting [85]. The purpose of that investigation was to assess the feasibility of the critiquing approach, by taking into account the single physician in the learning procedure. In that experiment, a less experienced dermatologist and an expert dermatologist were required to evaluate only one digital image of a pigmented skin lesion, providing the dermoscopy features and the clinical diagnosis.
Various learning algorithms were tested, namely Naïve Bayes, decision tree and k-nearest neighbor, and a rule induction algorithm. Differently from the present approach, a manual weighting procedure was applied to physicians’ misclassified cases, with weights ranging from 1 to 80 times greater than correctly diagnosed cases. The combination rule we used (step 7, Box 1) promotes sensitivity, as in the present case (see section 5.4.1).

The principal result showed that a user tailored system allows improving physician performances to a greater degree than standard approach. For example, the performances of the less experienced dermatologists regarding the recognition of melanoma improved from 0.66 to 0.93, while specificity did not decrease below 0.70.

In addition, even the expert improved the performances from 0.86 to 0.90, given that the classifier (Naive Bayes) was able to correctly recognize those few malignant lesions misclassified by the expert.
Figure 6 - Preliminary results. Sensitivity and specificity curves for the weighted NB and its combination with expert dermatologist. The vertical line represents a reasonable choice for the weight [85].

Those promising results drove us to the development of a more complete CDSS, trying to reproduce a more actual clinical set-
CHAPTER 5. EXPERIMENTAL RESULTS

ting. Therefore, we moved ahead to our present prototypical system where a more complete description of skin lesions is provided to the user. Moreover, an automated computation of the weight for the physician’s misclassified case is reported.

5.3. Evaluation

This section describes the experimental evaluation we performed. The first paragraph reports the results of the subjectivity assessments. Subsequent paragraphs describe the experimental evaluations of the critiquing module and the consulting module.

5.3.1 Subjectivity Assessment

To assess whether subjectivity plays a role, at least in the dataset we have at our disposal, we firstly measured physicians’ agreement in terms of Cohen’s kappa and overall disagreement (see section 4.1.2 for a description of Cohen’s kappa).

We compute $\kappa$ between each pair of physicians ($\Phi_i$ and $\Phi_j$) and for each feature $f$: $\kappa_{ij}^f$.

Afterwards, we averaged the results for all pairs of physicians:

$$ K' = \frac{1}{T \cdot (T - 1)} \sum_{i} \sum_{j \neq i} \kappa_{ij}^f $$

where $T$ is the number of physicians.

We also estimated inter-user variability through the overall disagreement. We defined, for each feature $f$:

$$ dis^f(\Phi_i, \Phi_j) = dis^f_j = \frac{1}{N} \sum_{r=1}^{N} d^f(\Phi_i(r), \Phi_j(r)) $$

where:

$$ d^f(\Phi_i(r), \Phi_j(r)) = \begin{cases} 1 & \text{if } \Phi_i(r) \neq \Phi_j(r) \\ 0 & \text{otherwise.} \end{cases} $$

$dis^f_j$ represents the mean number of disagreements between physician $\Phi_i$ and physician $\Phi_j$ with respect to all the $N$ cases in evaluating one feature $f$. This formula applies for nominal features. For
CHAPTER 5. EXPERIMENTAL RESULTS

ordinal features we used a weighted scheme as for weighted Cohen’s kappa:

\[ d^f \left( \Phi_i(r), \Phi_j(r) \right) = w^f \cdot \left\| \Phi_i(r) - \Phi_j(r) \right\| \]

where:

\[ w^f = \frac{\text{distance}}{\text{max possible distance}} \] (see section 4.1.2)

We then computed the overall average disagreement of features \( f \) with respect to all the pairs of physicians:

\[ D^f = \frac{1}{T \cdot (T - 1)} \sum_{i=1}^{T} \sum_{j=1, j \neq i}^{T} \text{dis}^f_{ij} \] \hspace{1cm} (6)

The overall disagreement provides a more intuitive measure of agreement, but does not account for unbalanced feature distribution. For example (see Table 3), the *comedo-like openings, and milia-like cysts* feature shows a disagreement value of only 4%, but the Cohen’s kappa is as low as 0.10. This is due to the very low presence of this feature in the dataset. As a matter of fact, in only 4 cases, on average, this feature was present.

Table 3 reports the results of this evaluation. Generally, values of \( \kappa < 0.40 \) are considered *poor agreement*, \( \kappa \in [0.40, 0.75] \) *fair to good agreement*, and \( \kappa > 0.75 \) *excellent agreement*. 

50
Table 3 – Subjective assessment. The column represents the average (with its standard deviation) of inter-user variability in subjective feature evaluation among all six physicians computed in terms of Cohen’s kappa and overall disagreement (see 5.3.1).

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>Type</th>
<th>Inter-user variability</th>
<th>(\kappa)</th>
<th>((Df)) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>binary</td>
<td>0.34 ± 0.11</td>
<td>31 ± 8</td>
<td></td>
</tr>
<tr>
<td>Border</td>
<td>binary</td>
<td>0.29 ± 0.12</td>
<td>36 ± 8</td>
<td></td>
</tr>
<tr>
<td>Number of colors</td>
<td>ordinal</td>
<td>0.38 ± 0.11</td>
<td>15 ± 3</td>
<td></td>
</tr>
<tr>
<td>Atypical pigm. network</td>
<td>binary</td>
<td>0.36 ± 0.11</td>
<td>31 ± 7</td>
<td></td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>binary</td>
<td>0.27 ± 0.13</td>
<td>35 ± 10</td>
<td></td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>binary</td>
<td>0.47 ± 0.11</td>
<td>22 ± 5</td>
<td></td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>binary</td>
<td>0.16 ± 0.12</td>
<td>33 ± 7</td>
<td></td>
</tr>
<tr>
<td>Streaks</td>
<td>binary</td>
<td>0.48 ± 0.10</td>
<td>21 ± 5</td>
<td></td>
</tr>
<tr>
<td>Slate blue veil</td>
<td>binary</td>
<td>0.48 ± 0.10</td>
<td>23 ± 5</td>
<td></td>
</tr>
<tr>
<td>Whitish veil</td>
<td>binary</td>
<td>0.17 ± 0.16</td>
<td>17 ± 4</td>
<td></td>
</tr>
<tr>
<td>Globular elements</td>
<td>binary</td>
<td>0.35 ± 0.14</td>
<td>31 ± 7</td>
<td></td>
</tr>
<tr>
<td>Comedo-like openings</td>
<td>binary</td>
<td>0.10 ± 0.19</td>
<td>4 ± 2</td>
<td></td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>binary</td>
<td>0.23 ± 0.17</td>
<td>11 ± 5</td>
<td></td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>binary</td>
<td>0.38 ± 0.14</td>
<td>26 ± 9</td>
<td></td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>binary</td>
<td>0.63 ± 0.08</td>
<td>19 ± 4</td>
<td></td>
</tr>
</tbody>
</table>

To assess whether expert physicians were more reliable in their evaluation, we estimated inter-variability on the expert subgroup in the same way. The results are shown in Table 4.
The results show that there is a high degree of inter-user variability in feature assessment, even if almost every feature is binary. The average $\kappa$ across all the subjective features is $0.32 \pm 0.12$; which is considered a poor agreement.

Considering the group of experts only, there is slightly better agreement (mean kappa $0.39 \pm 0.15$), but it is still low.

For sake of completion, the mean kappa for the non-expert group across all the subjective features is $0.27 \pm 0.13$.

Despite the low level of agreement there is a substantial concordance in the clinical diagnosis, both if we consider the whole group of dermatologists ($0.63 \pm 0.08$) or the expert group only ($0.72 \pm 0.11$).

These results suggest that even if there is a great variability in the feature assessments, the diagnostic conclusion are rather concordant. In addition, the treatment decision (excisional biopsy) show a moderate agreement among experts, while considering the whole group there is a poor level of agreement.

### Table 4 - Experts subjective assessment (see Table 3 caption)

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>Type</th>
<th>Inter-user variability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa$</td>
<td>$%\left(\overline{D}^2\right)$</td>
</tr>
<tr>
<td>Asymmetry</td>
<td>binary</td>
<td>$0.40 \pm 0.12$</td>
</tr>
<tr>
<td>Border</td>
<td>binary</td>
<td>$0.35 \pm 0.13$</td>
</tr>
<tr>
<td>Number of colors</td>
<td>ordinal</td>
<td>$0.52 \pm 0.13$</td>
</tr>
<tr>
<td>Atypical pigm. network</td>
<td>binary</td>
<td>$0.45 \pm 0.17$</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>binary</td>
<td>$0.36 \pm 0.22$</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>binary</td>
<td>$0.61 \pm 0.16$</td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>binary</td>
<td>$0.26 \pm 0.20$</td>
</tr>
<tr>
<td>Streaks</td>
<td>binary</td>
<td>$0.60 \pm 0.13$</td>
</tr>
<tr>
<td>Slate blue veil</td>
<td>binary</td>
<td>$0.52 \pm 0.17$</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>binary</td>
<td>$0.12 \pm 0.26$</td>
</tr>
<tr>
<td>Globular elements</td>
<td>binary</td>
<td>$0.37 \pm 0.26$</td>
</tr>
<tr>
<td>Comedo-like openings</td>
<td>binary</td>
<td>$0.20 \pm 0.35$</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>binary</td>
<td>$0.28 \pm 0.38$</td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>binary</td>
<td>$0.52 \pm 0.18$</td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>binary</td>
<td>$0.72 \pm 0.11$</td>
</tr>
</tbody>
</table>
CHAPTER 5. EXPERIMENTAL RESULTS

Apparently subjective judgment is highly present in feature evaluation, however individual diagnostic reasoning lead to similar conclusions. Two main reasons may explain this result:
- the differences in features evaluations are differently accounted by each physicians in his/her diagnostic reasoning;
- the diagnostic reasoning is based on other considerations outside the feature set.

The latter is surely true, because the dermoscopy features do not completely capture the complexity of a visual inspection of lesion images. However, it is likely that a combination of these two aspects is present, and this confirms that tacit, implicit knowledge is an important aspect in the early melanoma diagnosis.

In addition, biopsy decision agreement is lower than diagnostic agreement even considering the expert group only. As a matter of fact, dermatologists may decide to remove some lesions they diagnosed as benign, and this is the reason why we notice this difference. This fact confirms that implicit knowledge plays a role in experts’ decision making, because explicit knowledge generally deals with the diagnosis aspect only.

5.3.2 Physician Performances

We estimated physician performances by computing accuracy, sensitivity and specificity for all physicians. Sensitivity and specificity are performance measures typically used in medicine, whenever binary classification problem is under study. Given a confusion matrix ConfMatrix, defined in Table 5:

<table>
<thead>
<tr>
<th>ConfMatrix</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>TP</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

TP and TN are the number of positive and negative cases correctly identified, respectively; FP is the number of negative cases
CHAPTER 5. EXPERIMENTAL RESULTS

misclassified as positive, and FN is the number of positive cases misclassified as negative. Accordingly to ConfMatrix, Sensitivity and Specificity are defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

In our case, sensitivity measures the ability of the “classifier”, either human or computerized, to identify positive, i.e. malignant, cases; whereas the specificity represents the ability to diagnose negative, i.e. benign, cases. Accuracy, which is one of the common parameters used in machine learning community to assess performances of learning algorithm, is not properly suited in this domain, where cost-sensitive consideration are important. Moreover, it does not account for unbalanced dataset, i.e. datasets where the instances are unevenly distributed among the classes. In any case, for sake of completion, we reported also this parameter.

The same parameters were used to estimate the performances of PhysicianBoost (see 5.4.1).

Table 6 reports the results for the group of dermatologists. To discriminate the group of expert dermatologist, i.e. dermatologists who are experienced with the dermoscopy technique, from the non-expert dermatologists, hereinafter we call the former experts and the latter physicians.
Table 6 – Physicians’ performances. Accuracy, sensitivity and specificity with standard deviation (in parentheses) are shown.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician 1</td>
<td>0.76</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>Physician 2</td>
<td>0.75</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>Physician 3</td>
<td>0.83</td>
<td>0.76</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>Average Phys.</strong></td>
<td><strong>0.78 (0.04)</strong></td>
<td><strong>0.74 (0.02)</strong></td>
<td><strong>0.81 (0.06)</strong></td>
</tr>
<tr>
<td>Expert 1</td>
<td>0.81</td>
<td>0.93</td>
<td>0.71</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.82</td>
<td>0.90</td>
<td>0.76</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.79</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Average Exp.</strong></td>
<td><strong>0.81 (0.02)</strong></td>
<td><strong>0.88 (0.07)</strong></td>
<td><strong>0.75 (0.03)</strong></td>
</tr>
</tbody>
</table>

The group of experts shows a greater sensitivity with respect to physicians, whereas experts’ average specificity is lower than physicians’ specificity in this dataset at least.

5.4. Results

5.4.1 Critiquing system

One of the steps of PhysicianBoost regards the choice of the combination rule (step 7, Box 1). We combined learning algorithm outputs and physician’s diagnoses in order to improve sensitivity, which is the more critical parameter in this application domain, especially for non-expert dermatologists. Therefore the combination rule is the following:
CHAPTER 5. EXPERIMENTAL RESULTS

\[
\begin{align*}
&\text{IF } (\text{diag}(x) = \text{benign AND } h(x) = \text{benign}) \\
&\text{THEN } \\
&\quad \text{comb(diag}(x), h(x)) = \text{benign} \\
&\text{ELSE } \\
&\quad \text{comb(diag}(x), h(x)) = \text{melanoma}
\end{align*}
\]

In other words, only if both “models” predict “benign lesion” the combined system output is benign. This choice resembles the clinical setting where the system advises physicians only for potentially malignant lesions which were judged as benign by the physician. It represents somehow a “safety” condition. In other words the CDSS does not suggest the physician that a lesion is benign, potentially exposing the patient to the risk of not being properly treated. This choice set strict bounds on overall sensitivity and specificity, i.e. the sensitivity and specificity of physician’s helped by the learning algorithm (LA):

\[
\begin{align*}
\text{Sensitivity}_{\text{overall}} & \geq \max(\text{Sensitivity}_{\text{physician}}, \text{Sensitivity}_{\text{LA}}) \\
\text{Specificity}_{\text{overall}} & \leq \min(\text{Specificity}_{\text{physician}}, \text{Specificity}_{\text{LA}})
\end{align*}
\]

Nevertheless, there is no \textit{a priori} guarantee that the learning algorithm would improve physician’s sensitivity or decrease physician’s specificity.

We compared PhysicianBoost, with the unaided physicians and with the standard machine learning based approach. Regarding the latter aspect, we assumed that a “good” dataset is provided by the experts. The learning algorithm then learns the model from that dataset and, finally, the model is used by each physician. This is the standard approach, already described in section 4.1.2.

In more details, we assumed that the dataset is made up by a “consensus” among the three experts dermatologist on each feature. We constructed a “clean” dataset by combining feature evaluations of three experts (using a majority vote on binary features and the rounded mean for ordinal one). This procedure
eliminates variability in feature assessments and should allow for the creation of the “best” diagnostic model. A Naïve Bayes classifier learned the model $M_{\text{Expert}}$ using this dataset. We then compared performances of PhysicianBoost method with:

i) diagnostic performances of physicians,
ii) classification performances of $M_{\text{Expert}}$ with physician’s data as input.

Performances are expressed as accuracy, sensitivity, and specificity and were estimated by using 10-fold cross validation. The 10-fold cross-validation results of $M_{\text{Expert}}$ on the clean dataset were $0.79 \pm 0.08$, $0.76 \pm 0.19$, $0.84 \pm 0.05$ for accuracy, sensitivity, and specificity, respectively.

Table 7 reports PhysicianBoost’s results, whereas Table 8 reports those of the $M_{\text{Expert}}$ model created on the “clean” dataset and then tested with the features evaluated by physician $\Phi_i$. The notation $M_{\text{Expert}}(\Phi_i)$ stands for the model created on the “clean” dataset and then tested with the features evaluated by physician $\Phi_i$.

We used a paired Wilcoxon test to assess statistical differences.

---

Concerning physicians’ performances, we compute accuracy, sensitivity and specificity on the same testing partitions as the classifier. In this way, we can estimate the average values as well as compute the paired Wilcoxon test to statistically estimate the differences.
Table 7 - Performances of Physician-Boost for the three physicians. The corresponding physician’s performances are also reported (see footnote at page 28) (in bold: paired Wilcoxon test with a p value <0.05).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician-Boost</td>
<td>0.77 ± 0.08</td>
<td>0.83 ± 0.19</td>
<td>0.73 ± 0.12</td>
</tr>
<tr>
<td>Physician-Boost</td>
<td>0.75 ± 0.07</td>
<td>0.77 ± 0.19</td>
<td>0.74 ± 0.14</td>
</tr>
<tr>
<td>Physician-Boost</td>
<td>0.78 ± 0.11</td>
<td>0.84 ± 0.13</td>
<td>0.73 ± 0.18</td>
</tr>
<tr>
<td>Physician_1</td>
<td>0.76 ± 0.07</td>
<td>0.73 ± 0.17</td>
<td>0.78 ± 0.11</td>
</tr>
<tr>
<td>Physician_2</td>
<td>0.75 ± 0.11</td>
<td>0.71 ± 0.23</td>
<td>0.77 ± 0.13</td>
</tr>
<tr>
<td>Physician_3</td>
<td>0.82 ± 0.11</td>
<td>0.76 ± 0.15</td>
<td>0.87 ± 0.14</td>
</tr>
</tbody>
</table>

The results in Table 7 show that PhysicianBoost is able to recognize some malignant lesions that the physician misclassifies, since the sensitivity values are greater than those of physicians, with an average improvement of 0.08. In addition, this improvement is also statistically significant. On the other hand, the specificity however decreases. Anyhow, the average specificity of physicians using the system is 0.73, which is still comparable to that of experts (0.75). For Φ_2 this decrease is not statistically significant.

Table 8 – Performances of model M^Expert when used by physicians. (in bold the differences with Physician-Boost when paired Wilcoxon test p-value <0.05 are reported).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>M^Expert(Φ_1)</td>
<td>0.78 ± 0.10</td>
<td>0.63 ± 0.21</td>
<td>0.89 ± 0.09</td>
</tr>
<tr>
<td>M^Expert(Φ_2)</td>
<td>0.75 ± 0.08</td>
<td>0.57 ± 0.15</td>
<td>0.90 ± 0.10</td>
</tr>
<tr>
<td>M^Expert(Φ_3)</td>
<td>0.75 ± 0.09</td>
<td>0.59 ± 0.17</td>
<td>0.87 ± 0.10</td>
</tr>
</tbody>
</table>

The results in Table 8 show that the standard approach, M^Expert presents very low sensitivity performances when used by the physicians. Since the goal of the CDSS is to support the early diagnosis of melanoma, the sensitivity values are not sufficient for this purpose as they are even lower than those of the corresponding physician, despite the specificity values are always greater.
We also explore the performances of PhysicianBoost with experts.

Table 9 – Performances of PhysicianBoost used by expert (thus renamed ExpertBoost).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert1-Boost</td>
<td>0.79 ± 0.08</td>
<td>0.93 ± 0.10</td>
<td>0.68 ± 0.14</td>
</tr>
<tr>
<td>Expert2-Boost</td>
<td>0.79 ± 0.11</td>
<td>0.92 ± 0.10</td>
<td>0.69 ± 0.16</td>
</tr>
<tr>
<td>Expert3-Boost</td>
<td>0.77 ± 0.11</td>
<td>0.86 ± 0.14</td>
<td>0.70 ± 0.16</td>
</tr>
</tbody>
</table>

Given that sensitivity of experts is already high, we do not notice any further statistical improvement. However, the learning algorithm is able to recognize some malignant lesions misclassified by experts, especially for expert E3. Specificity decreases of about 6% on average. This means that PhysicianBoost is not able to adapt to the experts, decreasing too much their specificity without great improvements in sensitivity.

We also evaluated the performances of a Naïve Bayes model which uses only the objective data. The results were 0.70 ± 0.12, 0.60 ± 0.15, and 0.77 ± 0.16 for accuracy, sensitivity, and specificity, respectively. This result suggests that objective information is not sufficient for diagnostic purposes, hence confirming the importance of dermoscopy evaluations for the early recognition of melanoma.
CHAPTER 5. EXPERIMENTAL RESULTS

5.4.2 Consulting module

The purpose of the consulting module is to share the implicit knowledge included into the previously solved cases among physicians with different experience. As we already noted in section 5.3.1, there is a high level of subjective knowledge in the way experts evaluate the cases, suggested by the comparison of the quite good agreement in clinical decisions ($\kappa = 0.72 \pm 0.11$ and $\kappa = 0.52 \pm 0.18$ for diagnosis and treatment, respectively) with the relatively poor agreement on feature evaluations ($\kappa = 0.39 \pm 0.15$).

The implicit knowledge is included into experts’ reference case base. Hence, a case based reasoning approach can help sharing that knowledge with physicians, by showing physicians the cases in the reference case base.

The fundamental step in case-based reasoning is the retrieval of the previously solved cases [56]. In dermatology, images are typically the main sources of information. In our application, however, the retrieval of the cases is based on a coded description of the images by means of the dermoscopy features. Therefore, despite how good the description is, this description is not sufficient to capture the complexity of the image. Therefore, the actual relevance of the retrieved cases to the case at hand must be assessed by the physician. This is particularly evident in our situation, where, in addition, the data in the case base are few given the wide range of appearance of pigmented skin lesions.

The choice to rely on a coded description of the case is due to the possibility for a physician to use the CDSS without a specific image acquisition device. This is different for other methods that use image-content based retrieval by using image processing analysis coupled with non-image based information.

For sake of simplicity, we introduce some notation and terms.

- Clinical decisions: the clinical diagnosis (benign or malignant) and the choice for excisional biopsy (yes or no);
CHAPTER 5. EXPERIMENTAL RESULTS

- Subjective evaluation: the subjective dermoscopy parameters (reported in Table 2);
- Objective features: the objective parameters of a case such as patient’s age, location of the lesion, etc. (see Table 2);
- Reference dataset \( RDS \): the dataset of all the cases with the objective information and the digital images;
- Physician \( \Phi_i \): the non-expert physician, the index \( i \) varies from 1 to 3;
- Expert \( E_i \): the expert dermatologist, the index \( i \) varies from 1 to 3.
- Physician’s reference case base PCB\(_i\): the dataset made up by the physician \( \Phi_i \)’s subjective evaluations and clinical decisions.
- Expert’s reference case base ECB\(_i\): the dataset made up by the expert’s \( E_i \)’s subjective evaluations and clinical decisions.

As described in section 4.1.2, two approaches are proposed. The first one searches for the similar cases into PCB\(_i\), whereas the second one searches for the similar cases into ECB\(_i\), after estimating which expert is “similar” to the physician in term of subjective evaluations. Actually the similarity measure allows for the definition of a ranking among the cases.

In the first approach we used the objective features and the subjective evaluation as well as the clinical decisions to compute the similarity measure. The rationale for this choice is to retrieve cases that were previously similarly evaluated in terms of both subjective features and clinical decisions.

Concerning the second approach, we compute first the average agreement for all the \( m_{subj} \) subjective features (Table 2) between physician \( \Phi_i \) and expert \( E_i \):
\[ \bar{\kappa}_g = \frac{1}{m_{\text{subj}}} \sum_{f \in \text{Subj}} \kappa^f (\Phi_i, E_j) \]

where \( \kappa^f (\Phi_i, E_j) \) is the value of the Cohen’s kappa for the feature \( f \) between physician \( \Phi_i \) and expert \( E_j \).

Secondly, we chose the expert \( e \) who presents the maximum agreement with the physician \( i \): \( \kappa^e > \kappa_j \forall j \neq e \). In the end, the closest case is searched for into ECB\(_e\).

Table 10 shows the average agreement across all the subjective features between the physicians and the experts.\(^6\)

**Table 10 – Agreement between physicians and experts. \( \bar{\kappa}_g \) values are reported (in bold the maximum).**

<table>
<thead>
<tr>
<th>( \bar{\kappa} )</th>
<th>( E_1 )</th>
<th>( E_2 )</th>
<th>( E_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Phi_1 )</td>
<td>0.32</td>
<td>0.29</td>
<td>0.32</td>
</tr>
<tr>
<td>( \Phi_2 )</td>
<td>0.32</td>
<td>0.30</td>
<td>0.34</td>
</tr>
<tr>
<td>( \Phi_3 )</td>
<td>0.24</td>
<td><strong>0.39</strong></td>
<td>0.25</td>
</tr>
</tbody>
</table>

The results in Table 10 depict a general low agreement between experts and physicians. Expert \( E_1 \) is the “closest” to both physician \( \Phi_1 \) and \( \Phi_2 \), while expert \( E_2 \) is the “closest” to physician \( \Phi_3 \). In addition, the latter present a slightly greater average agreement.

We compute the similarity measure on ECB\(_e\) in the same way as the first approach, but the clinical decision parameters. The reason of this choice is that we suppose to retrieve cases whose subjective evaluations are similar, but possibly different clinical decisions could have been taken.

Only a qualitative analysis of this module can be performed so far as the module is in a prototypical development phase and thus

\(^6\) The values for \( \bar{\kappa}_g \) were computed using all the data in the reference case base.
some limitation occurs. The principal limitation is due to the rather small dataset to search into that do not cover all the variety of pigmented skin lesion appearance. Furthermore, the well-known “curse of dimensionality” plays here a great role in the similarity measure, since we have a hundreds of cases with a relatively high-dimension feature space. Reducing the feature space by means of automatic feature selection methods can be useful from a classification perspective, but it is reduces also the case information used by dermatologist.

Moreover, the relevance problem, i.e. assessing if the retrieved case are relevant to the problem at hand, is demanded to the physician, as the more complete information is contained into the images. The latter point is also a common problem in CBR systems in medicine [21].

Nevertheless, let first assume that the retrieved case are relevant to that under examination and inspect what can happen with the two retrieval approaches:

1. approach: searching into the PCB. Physicians can compare their own subjective evaluations and clinical decisions on similar cases. Moreover, they can assess the differences with experts by looking at the same cases into the ECB. This is possible because the reference dataset is common to physicians and experts.

2. approach: searching into the ECB. The physician can compare directly expert’s clinical decisions on cases whose feature evaluation is most similar. In this case, if the clinical decisions are the same of the physician, he/she can be reassured about their correct clinical reasoning. On the other hand, if the expert conclusions were different, physicians can figure out why there is such difference.

On the other hand, if the retrieved case is not relevant, the user can figure out the reason for that. In particular we have these two situations, depending on the approach:

1. searching into the PCB. This means that the physician provided similar evaluations to cases judged not relevant
to the problem at hand. A possible result would be the revision of the evaluations, with the help of corresponding experts’ information, for instance;

2. searching into the ECB. This means that the expert evaluated the case similarly to the physician. The physician can critically assess the proper evaluations accordingly to the low actual relevance with expert’s case.

Qualitative assessment
The similarity measure sorts the cases in a decreasing order. This allows the physician for browsing the case base. However, for sake of simplicity we will discuss about the top-ranked case.

To qualitatively assess the consulting module, we retrieved the most similar case using both approaches (see 4.1.2) through a leave-one-out scan along the dataset. This means that each case in the PCB was considered as new on each run and removed from the PCB. Accordingly, the selected “new” case was removed from the ECB also. We then found the closest case with both approaches.

Regarding the second approach, we computed Cohen’s $\kappa$ on the subjective evaluations considering only the training set. For each run, no differences were pointed out with respect to $\kappa$ computed on all dataset (Table 10).

Table 11 shows the mean similarities values for the two methods regarding the top-ranked case.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCB</th>
<th>ECB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Phi_1$</td>
<td>0.85 [0.74-0.96]</td>
<td>0.83 [0.73-0.94]</td>
</tr>
<tr>
<td>$\Phi_2$</td>
<td>0.85 [0.73-0.98]</td>
<td>0.82 [0.71-0.95]</td>
</tr>
<tr>
<td>$\Phi_3$</td>
<td>0.84 [0.72-0.97]</td>
<td>0.84 [0.68-0.97]</td>
</tr>
</tbody>
</table>

For $\Phi_1$ and $\Phi_2$, the mean similarity of the first approach is higher than that of the second approach; where for $\Phi_3$ the second ap-
proach is equal to the first one on average. It is worth noting that this is just a qualitative analysis, as the actual relevance of the retrieved cases should be assessed by the user. In any case, we can argue that this result may be linked to the higher average agreement of $\Phi_3$ with expert $E_2$ compared to the other physicians. Anyhow, to assess if there is a difference in the similarity measure for the two approaches a paired Wilcoxon test was performed. The two approaches for $\Phi_1$ and $\Phi_2$ show a statistically significant difference ($p<0.05$); while for $\Phi_3$ no statistically significant differences were found.

**Some examples**

To gain some insights on what information the CDSS can provide to the physician, some examples are hereinafter proposed. A test case is presented in terms of the objective features and a dermoscopy image. In addition, physician’s subjective evaluations are reported. The top-ranked case following the two approaches are then reported with the same information as the test case and compared. Regarding the first approach, we report both physician’s and expert’s subjective evaluations and clinical decisions. Regarding the second approach we report the “closest” experts’ subjective evaluations.

We selected as exemplar cases the two extreme cases with the highest and lowest similarity among all the physicians, irrespective of the approach used. The situation where the CDSS retrieves the case with highest similarity is the following:

- physician $\Phi_2$, first approach, top-ranked case: similarity 0.98.
Test case 1:

<table>
<thead>
<tr>
<th>Objective Features</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesion location</td>
<td>Trunk</td>
</tr>
<tr>
<td>Max-diameter</td>
<td>6 (mm)</td>
</tr>
<tr>
<td>Min-diameter</td>
<td>5 (mm)</td>
</tr>
<tr>
<td>Evolution</td>
<td>No</td>
</tr>
<tr>
<td>Age</td>
<td>37</td>
</tr>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>Family history of melan.</td>
<td>No</td>
</tr>
<tr>
<td>Fitzpatrick’s Phototype</td>
<td>3</td>
</tr>
<tr>
<td>Sunburn</td>
<td>Not. det</td>
</tr>
<tr>
<td>Ephelis</td>
<td>No</td>
</tr>
<tr>
<td>Lentigos</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>$\Phi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>No</td>
</tr>
<tr>
<td>Border</td>
<td>No</td>
</tr>
<tr>
<td>Number of colors</td>
<td>1</td>
</tr>
<tr>
<td>Atypical pigm. net.</td>
<td>No</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>No</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>No</td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>No</td>
</tr>
<tr>
<td>Streaks</td>
<td>No</td>
</tr>
<tr>
<td>Slate-blue veil</td>
<td>No</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>No</td>
</tr>
<tr>
<td>Globular elements</td>
<td>No</td>
</tr>
<tr>
<td>Comedo-like op., …</td>
<td>No</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>No</td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>No</td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>Benign</td>
</tr>
</tbody>
</table>
In this case, the first approach retrieved the following case:

<table>
<thead>
<tr>
<th>Objective Features</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesion location</td>
<td>Trunk</td>
</tr>
<tr>
<td>Max-diameter</td>
<td>11 (mm)</td>
</tr>
<tr>
<td>Min-diameter</td>
<td>8 (mm)</td>
</tr>
<tr>
<td>Evolution</td>
<td>No</td>
</tr>
<tr>
<td>Age</td>
<td>55</td>
</tr>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>Family history of melan.</td>
<td>No</td>
</tr>
<tr>
<td>Fitzpatrick’s Phototype</td>
<td>3</td>
</tr>
<tr>
<td>Sunburn</td>
<td>Not det.</td>
</tr>
<tr>
<td>Ephelis</td>
<td>No</td>
</tr>
<tr>
<td>Lentigos</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>$\Phi_{2}$</th>
<th>$E_{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Border</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Number of colors</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Atypical pigm. net.</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Streaks</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Slate-blue veil</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Globular elements</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Comedo-like op., ...</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>Benign</td>
<td>Benign</td>
</tr>
</tbody>
</table>

As expected, there are very little differences in feature evaluations between the test case and the retrieved one (shown in bold). Comparing physician’s and expert’s evaluations we can note a different evaluation of the number of color. The physician can for instance review his/her judgment, in the same way as in educational setting where the expert explains the subjective evaluation.

The second approach retrieves the following case as the nearest
In this case the differences in the expert’s subjective evaluations with respect to the test case are lower compared to the first approach, as expected. Actually only a small difference in the “number of colors” is present. Major differences are found in the objective features.

With both approaches no differences in the clinical diagnosis and the treatment decision are present. The similarity values for the two approaches are reported in Table 12.
Table 12 – Similarity values. Overall similarity as well as partial similarities are shown (see 4.1.2).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Overall</th>
<th>Subjective feat. $\text{sim}_{\text{subj}}$</th>
<th>Objective feat. $\text{sim}_{\text{obj}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>2</td>
<td>0.90</td>
<td>0.99</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Those values illustrate the greater difference in the objective features for the second approach. The physician has to be aware of this information to properly assess the expert’s judgment.

As a second example we show the case for which the first retrieved case has the lowest of all the experimental evaluations:
- physician 3, the second approach retrieved a case with similarity 0.68.
### Test case 2:

<table>
<thead>
<tr>
<th>Objective Features</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesion location</td>
<td>Trunk</td>
</tr>
<tr>
<td>Max-diameter</td>
<td>Not det.</td>
</tr>
<tr>
<td>Min-diameter</td>
<td>Not det.</td>
</tr>
<tr>
<td>Evolution</td>
<td>Not det.</td>
</tr>
<tr>
<td>Age</td>
<td>40</td>
</tr>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>Family history of melan.</td>
<td>Yes</td>
</tr>
<tr>
<td>Fitzpatrick’s Phototype</td>
<td>Not det.</td>
</tr>
<tr>
<td>Sunburn</td>
<td>Not det.</td>
</tr>
<tr>
<td>Ephelis</td>
<td>Not det.</td>
</tr>
<tr>
<td>Lentigos</td>
<td>Not det.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>( \Phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>No</td>
</tr>
<tr>
<td>Border</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of colors</td>
<td>2</td>
</tr>
<tr>
<td>Atypical pigm. net.</td>
<td>Yes</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>No</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>No</td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>Yes</td>
</tr>
<tr>
<td>Streaks</td>
<td>Yes</td>
</tr>
<tr>
<td>Slate-blue veil</td>
<td>No</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>No</td>
</tr>
<tr>
<td>Globular elements</td>
<td>Yes</td>
</tr>
<tr>
<td>Comedo-like op., ...</td>
<td>No</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>No</td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>No</td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>Benign</td>
</tr>
</tbody>
</table>
In this case, the highest similarity for the first approach is the following case (in bold the differences with the test case):

<table>
<thead>
<tr>
<th>Objective Features</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesion location</td>
<td>Legs</td>
</tr>
<tr>
<td>Max-diameter</td>
<td>Not det.</td>
</tr>
<tr>
<td>Min-diameter</td>
<td>Not det.</td>
</tr>
<tr>
<td>Evolution</td>
<td>Not det.</td>
</tr>
<tr>
<td>Age</td>
<td>34</td>
</tr>
<tr>
<td>Gender</td>
<td>female</td>
</tr>
<tr>
<td>Family history of melan.</td>
<td>Not det.</td>
</tr>
<tr>
<td>Fitzpatrick’s Phototype</td>
<td>Not det.</td>
</tr>
<tr>
<td>Sunburn</td>
<td>Not det.</td>
</tr>
<tr>
<td>Ephelis</td>
<td>Not det.</td>
</tr>
<tr>
<td>Lentigos</td>
<td>Not det.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>$\Phi_3$</th>
<th>$E_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Border</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Number of colors</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Atypical pigm. net.</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Hypo-pигmentation</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Streaks</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Slate-blue veil</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Globular elements</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Comedo-like op.,....</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>Benign</td>
<td>Malignant</td>
</tr>
</tbody>
</table>

The retrieved case here present markedly differences in subjective evaluations with respect to the test case, both for the physician and the expert. Moreover, expert’s clinical decisions were different from those of the physician.

In this case the physician is expected to carefully assess the reliability of the retrieval procedure, provided also the low similarity value and also the rather different evaluations physician and expert propose for the retrieved case (4 subjective parameters, namely asymmetry, number of colors, abrupt network cut-off, and globular elements).
The second approach retrieves the following case as the nearest one, which is the case with the lowest similarity (0.68):

<table>
<thead>
<tr>
<th>Objective Features</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesion location</td>
<td>Trunk</td>
</tr>
<tr>
<td>Max-diameter</td>
<td>Not det.</td>
</tr>
<tr>
<td>Min-diameter</td>
<td>Not det.</td>
</tr>
<tr>
<td>Evolution</td>
<td>Not det.</td>
</tr>
<tr>
<td>Age</td>
<td>34</td>
</tr>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>Family history of melan.</td>
<td>Not det.</td>
</tr>
<tr>
<td>Fitzpatrick’s Phototype</td>
<td>Not det.</td>
</tr>
<tr>
<td>Sunburn</td>
<td>Not det.</td>
</tr>
<tr>
<td>Ephelis</td>
<td>Not det.</td>
</tr>
<tr>
<td>Lentigos</td>
<td>Not det.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjective Features</th>
<th>$E_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>Yes</td>
</tr>
<tr>
<td>Border</td>
<td>No</td>
</tr>
<tr>
<td>Number of colors</td>
<td>2</td>
</tr>
<tr>
<td>Atypical pigm. net.</td>
<td>Yes</td>
</tr>
<tr>
<td>Abrupt network cut-off</td>
<td>No</td>
</tr>
<tr>
<td>Regression-Erythema</td>
<td>No</td>
</tr>
<tr>
<td>Hypo-pigmentation</td>
<td>No</td>
</tr>
<tr>
<td>Streaks</td>
<td>No</td>
</tr>
<tr>
<td>Slate-blue veil</td>
<td>No</td>
</tr>
<tr>
<td>Whitish veil</td>
<td>No</td>
</tr>
<tr>
<td>Globular elements</td>
<td>No</td>
</tr>
<tr>
<td>Comedo-like opp, milia…</td>
<td>No</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>No</td>
</tr>
<tr>
<td>Excisional biopsy</td>
<td>No</td>
</tr>
<tr>
<td>Clinical diagnosis</td>
<td>Benign</td>
</tr>
</tbody>
</table>

Five subjective evaluations were different from the test case. The clinical decisions are the same. However, given the low similarity value, physician is expected to carefully assess the reliability of the retrieval procedure.

The similarity values for the two approaches are reported in Table 13.
Table 13 – Similarity values. Overall similarity as well as partial similarities are shown (see 4.1.2).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Overall</th>
<th>Subjective feat.</th>
<th>Objective feat.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$sim_{subj}$</td>
<td>$sim_{obj}$</td>
</tr>
<tr>
<td>1</td>
<td>0.74</td>
<td>0.87</td>
<td>0.64</td>
</tr>
<tr>
<td>2</td>
<td>0.68</td>
<td>0.62</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Those values illustrate that the overall similarity is rather low for both the approaches. This fact points out one of the limits of the consulting module, i.e. the coverage of the problem domain. Adding new cases into the case base can give some improvement. However, the physician has to be aware of this problem for the CDSS to be helpful.
Chapter 6

6. Conclusions

The research on Clinical Decision Support Systems (CDSSs) is one of the principal topics of Medical Informatics (MI). Medicine is the most challenging domain where artificial intelligence methodologies were applied since the beginning of the computer era. Most of the recent research activity in CDSSs deals with knowledge representation and uncertainty reasoning, since their basic paradigm is helping physician in the complex knowledge management process of nowadays medical activity. In this context, CDSSs have been recognized as important tools to improve the quality of patient care.

The rationale of CDSSs is that proper management of clinical knowledge is required to support physician’s daily decisions, as their personal knowledge might be outdated, or time-constraints and limited information may impair their reasoning procedure.

To overcome this problem Evidence-based Medicine (EBM) paradigm has recently emerged. EBM aims at providing the health practitioners with up to date scientifically-based medical evidence, so as to disseminate “the best practice”. The ultimate goal is the improvement of the quality of care delivery. Accordingly to this purpose, MI has properly developed tools for the exploitation of EBM paradigm through an intensive research on computerized clinical guidelines. This research is based on the long time experience on knowledge-based systems; an activity that was pursued since the beginning of MI.

CDSSs built so far typically deal with explicit knowledge, i.e. knowledge that can be formally expressed. Tacit knowledge may still play a great role and may affect the performances of CDSSs. Different skill levels and expertise of health care providers lead to
a high grade of inter-user variability in the assessment of patient’s conditions. This variability is even more evident when no clear and definite biomedical knowledge is available and therefore the interpretation of symptoms is deeply dependent on physician’s expertise and skill. To our knowledge, the research on CDSSs has not properly addressed this problem so far [11].

In this work we proposed a novel approach to create a clinical decision support system whenever inter-user variability in clinical assessments is present. To our knowledge, this thesis is the first study to investigate a user-tailored CDSS, which exploits physician’s expertise and support the sharing of experts’ implicit knowledge. Our approach fits the field of multi-modal reasoning, where multiple problem-solving strategies are employed [18]. We applied our novel method to the problem of the early recognition of melanoma, where inter-user variability is a major concern [16,17]. The user-tailored CDSS is made up by two modules:

- a critiquing module, PhysicianBoost, based on a novel machine learning approach that deals with the specific skill and expertise of the user;
- a consulting module, based on case-based reasoning methodology, to support the sharing of implicit knowledge among physicians.

Both modules of the CDSS rely on a reference dataset, composed by a number of known cases. Objective information (such as patient’s age) as well as the “gold standard” data (i.e. in the melanoma scenario, the histological diagnosis) are present in the dataset.

Subjective information is provided by a group of expert dermatologists, who evaluated those cases independently from each other. They also provide the clinical diagnosis as well as their choice for excisional biopsy. The reference dataset together with dermatologists’ subjective evaluations and clinical decision is the
reference case base. Each new user of the system is required to evaluate the reference dataset in the same way experts did. Accordingly to other studies [17], we assessed the high level of inter-user variability on this dataset regarding the assessment of the subjective features (dermoscopy parameters). This result confirmed that the development of CDSS has to account for this. In the next sections a critical appraisal of the main characteristics of the two modules is reported.

6.1. Critiquing module

A critiquing module is supposed to provide unsolicited advices to the user. We employed a novel machine learning-based approach to develop this module.

The proposed method for the critiquing module, called PhysicianBoost, builds a model for each physician (a tailored model), starting from his/her feature evaluations, and then combines its output with clinical decision through a suitable combination rule. By using this innovative machine learning based approach, our solution tackles both the inter-user variability problem as well as the provision of user-adapted advices. We tested our approach on a set of 177 cases of pigmented skin lesions.

PhysicianBoost performances, compared both with unaided physicians’ ones and with those of standard solutions, showed that inter-user variability is properly addressed, provided the limited data at our disposal. In fact, the results we obtained showed that our approach (PhysicianBoost) is able to improve the diagnostic performances of physicians in terms of sensitivity, i.e. the recognition of the malignant lesions. On the other hand, though decreasing, specificity is comparable to that of expert dermatologists.

As expected, we showed that variability in feature assessments affects the performances of the model built on “clean” dataset (MExpert) when used by non expert physicians, especially concerning sensitivity. This shows that, at least for this dataset, the standard approach of creating decision support systems is not feasible when inter-user variability in feature assessment is a big deal.
CHAPTER 6. CONCLUSIONS

Some limitations of this approach are worth mentioning. The underlying assumption of the user-tailored system is that physician’s subjectivity evaluation of parameters is stable over time, i.e. intra-user variability is lower with respect to inter-user variability. This assumption is reasonable as reported by other authors [17]. In any case, we expect that a physician will improve over time his/her expertise by using the system. The Naïve Bayes classifier can partially account for this trend by incrementally updating the likelihood values in the user’s model. Moreover, the consulting module of the CDSS may induce the user to re-evaluate some cases in the reference dataset, accordingly to the new acquired skills. Again, in this case, the Naïve Bayes model is easily updated. Therefore, intra-user variability is not expected to limit CDSS potentiality.

One of the critical points for the application of machine learning methods in medicine is their explanation ability. It is well known that physicians require a CDSS to explain its reasoning to be acceptable [82]. We argue that the probabilities of the Naïve Bayes are easily understood by physicians. As suggested by Lavrač [40], Naïve Bayes classifier is one of the preferred methods by physicians given the explanation in terms of features for or against a given diagnosis, and typically has the advantage to be stable and achieve good performances. However, no specific investigation was carried out to assess interpretability or acceptance of the results by final users.

PhysicianBoost training phase requires unbalancing the dataset in order to focus on the most difficult cases for the physician. This is accomplished by a weighting procedure that unbalances the classes (step 5, Box 1). Actually, on this application problem, the effect of physician’s error weighting is not very relevant. Similar overall performances were obtained forcing the weights $w_i$ to be all equal. Anyhow, the user-tailored system still outperforms standard solutions.


6.2. Consulting module

The consulting module aims at sharing the implicit knowledge included into the previously solved cases. This module applies a case-based reasoning methodology to perform its task. Case-based reasoning means to retrieve previously solved problems which are similar to the current one and adapt their solution for the problem at hand. This paradigm presents some advantages in medicine especially to deal with the subjective knowledge which is typically contained in previously solved cases [21].

We focused on the first task of CBR, i.e. the retrieval of the most similar case. To our knowledge this is the first system that employs CBR aiming at sharing experts’ implicit knowledge.

Two approaches are proposed to retrieve the similar cases: the first one search for the most similar case within the physician’s case base proposing expert’s evaluations and clinical decision on the retrieved cases; the second approach first assess who is the expert most similar to the physician in terms of feature evaluations and then search the most similar cases into that expert’s case base. With both approaches, the user has to decide the relevance of the information proposed. Whenever the relevancy of retrieved case is, physicians can have insights into personal, tacit knowledge of experts, as described in section 5.4.2.

In our application, a content-based image retrieval would be an alternative solution. However, this approach is not feasible in this context, because images in the reference case base were acquired with different acquisition devices leading to very diverse images from a computerized image processing perspective (see section 3.2.1). Similarly to other authors [62], case retrieval is based on a coded description. In addition to overcome image processing drawbacks, this choice also allows a physician without digital image acquisition devices to use the system. These aspects would foster the actual implementation and use of the system in the clinical practice.
Both approaches suffer from the “curse of dimensionality” that may impair the retrieval capabilities of the system as long as it is performed on a small dataset with respect to the features space. It is expected that increasing the number of cases into the case base, the CDSS would retrieve more pertinent cases and cover a greater part of the domain space. The similarity measures computed by the system, however, give some hints to the user for assessing the relevance of the retrieved cases. Hence, the CDSS leaves the user the freedom to assess the information retrieved [58].

The ultimate assessment should evaluate how clinical decisions are affected by the consulting systems in a long-term training framework.

6.3. Future directions
The model we propose for the development of a CDSS for the early diagnosis of melanoma showed encouraging results. However, several other investigations can be pursued.

From a methodological viewpoint, there is room to further investigate the learning method in terms of explanation capabilities, weighting procedure, etc. as described in section 6.1. For instance, a future development can be the implementation of nomograms to further enhance the explanation ability of the critiquing system [95]. In addition, the consulting module, as well as the critiquing one, should surely benefit from the increase in the number of cases into the case base. This can ensure a proper coverage of pigmented skin lesions domain, and possibly include other kinds of dermatologic lesions (such as seborrheic keratosis) as well as other relevant clinical information (such as the “ugly duckling” sign).

The combination rule we propose in the critiquing module is highly focused on the diagnosis of malignant lesions. The rationale for this choice is the need for some sort of “safety condition”,

79
avoiding the possibility that patients are not properly treated due to wrong CDSS advices. However, other combination rules can be applied to address different purposes. For instance, an expert dermatologist would prefer a system that reassures him/her about the diagnosis of a benign lesion, in order to reduce the amount of unnecessary surgical intervention. The critiquing system can be modified accordingly.

In addition, the common reference case base can help to enhance the model of the medical diagnostic reasoning. The comparison on a set of similar cases of experts’ and novices’ decisions may provide insights into the medical reasoning process.

Among the aspects that deserve to be further investigated is the deployment of the CDSS to test the actual effectiveness and usefulness in a real clinical setting.

The ultimate goal of our system is to provide the physician with a tool that is able to support them in delivering the best quality of care to patients. Any effort in this direction should be appreciated by physicians as well as patients.
Appendix A

7. Web-based tele-dermatology system

A prototypical web-based system was developed in this thesis for tele-dermatology purposes. It is intended to support a consultation between an expert dermatologist, the consultant, and other dermatologists or general practitioners for a second opinion on pigmented skin lesions as well as a wider spectrum of dermatology diseases. The consultation modality is that of store-and-forward tele-dermatology [96]. The paradigm of store-and-forward dermatology requires that digital images and patient’s information are transferred to the consultant for subsequent evaluation in a time-independent manner. The consultant can review the images and the information at any time after the transfer of the data. On the other hand, real-time tele-dermatology requires an interactive communication between the consultant and the requesting doctor.

7.1. Architecture and components

The tele-dermatology service is based on a web-based 3-tier architecture. Figure 7 shows the general architecture of the system. The tele-dermatology system was developed by using eXtended Markup Language (XML), eXtensible Stylesheets Language (XSL), Microsoft COM+ Objects, Active Server Page (ASP), and JavaScript.
APPENDIX A. WEB-BASED TELE-DERMATOLOGY SYSTEM

Figure 7: Architecture of the web-based tele-dermatology system.

The network infrastructure is the Internet. The client side of the system is a standard web-browser (Microsoft Internet Explorer). On the server side, we used Microsoft Internet Information Service 6.0 as web-server and Microsoft SQL Server 2000 as multimedia database.

7.1.1 Data layer

The data layer is made up by a multimedia relational database management system (RDBMS), which stores data and images. It is also responsible to properly authenticate the user into the system. It is located behind a firewall to ensure security and privacy of data therein. A backup procedure on a different physical storage system is provided to ensure data recovery in case of a failure of the RDBMS.

The multimedia RDBMS communicates with Web Server to exchange the textual data, and with ImageSrv, a server object we developed to exchange digital images, by means of Microsoft COM+ technology.
7.1.2 Middle layer

The middle layer is the core of the system. It is composed by a Web Server and by ImageSrv, a server object we developed. The Web Server enables the communication between the user and the data layer, performing several controls and transformation onto the transmitted data. The dynamic part of the system was developed through Vbscript and the Active Server Pages (ASP). Concerning the communication with the data layer, the relational structure of the data in the RDBMS is mapped to a hierarchical structure by means of XML.

Figure 8 reports an example of the XML data structure we used. The general structure of the XML data has three main sections:

- **transport** section: this section includes details about the transmission of the data. For example, it contains the identifier of the schema that is contained in the **document** section, the modality the user is looking at the data, etc.
- **identifier** section: this section contains the identification data used by the system: user identification, access number etc., patient identification number and demographic data.
- **document** section: this is the core of the XML data structure since it contains the actual data that are used by the system to exchange information with the user. In fact, the actual content of this section depends on the data that are requested by the user.

The XML data structure is populated by means of standard SQL query to the multimedia RDBMS. On the other hand the ASPs are also responsible for properly storing back the information contained into the XML data structure into the relational format of the RDMS.
Concerning the communication with the client side, the web server exchanges the XML data structure, the XSL stylesheet that allows suitably displaying the information to the user.

The use of XML and XSL has great advantage in developing web based systems. The separation of the data (XML) from their visual representation (XSL) enables higher flexibility in the representation of the data which can also be tailored to the specific user.

In addition, the XML data structure includes a semantic representation of the data. This allows performing some computation on
APPENDIX A. WEB-BASED TELE-DERMATOLOGY SYSTEM

the client side, reducing network traffic and server computational overload.

The ImageSrv object is a Microsoft COM+ component we developed in Visual C++ in order to store the images, sent by the client, into the database as Binary Large Objects (BLOB), and conversely, to retrieve them from the database. Moreover, ImageSrv also creates thumbnails of full-size images. This solution is chiefly useful to reduce the overload due to full-size image downloading. Actually, the user downloads first the set of thumbnails for a certain patient and then, by clicking one of the thumbnails with the mouse, s/he can download and view the full-size image.

7.1.3 Presentation layer
The presentation layer is in charge to interact with the user. Dynamic HTML pages combine the XML data structure with the XSL stylesheet to provide the user with the proper modality of data visualization. If the user inserts or modifies the data, part of the data input control are performed on the client side through JavaScript code. Moreover, data input is supported through the use of pre-defined lists, check buttons, etc.
User screen is divided into three frames as illustrated in Figure 9.
On the left, the *system frame* includes the main buttons that allows the user to interact with the system: change password, logout the system, etc. On top, the *identification frame* reports the user identification information as well as patient’s demographic information, if applicable. The central frame, *data frame*, contains the main information, displaying for instance specific patient’s data. The data frame is the dynamic part of the system, where information is displayed accordingly to the XSL stylesheet that interprets the data in the XML DOM.
The identification frame contains also XML data structure through a XML Document Object Model (XMLDOM). This enables the presentation layer, for instance, to change user interface without requesting new information to the web server.

Typically, the user will interact with the system in two ways: the first one is the retrieval and visualization of information, the second one for inserting or modifying data. The XSL stylesheets handle the two interaction modalities, changing the user-interface accordingly, without the need to download new data.

Regarding data input and updating, Microsoft Data Island technology allows easily linking the new information provided by user via HTML objects to the XMLDOM.

Images can be viewed by the user in a separate window. As already mentioned, first thumbnails of the images are downloaded and then, upon user request, a full-size image can be retrieved. Simple manipulation function such as zooming in and out is provided. Figure 11 illustrate an example of the user interface for the image visualization.
The user can acquire digital images by means of any device that is able to store them in JPG format. The images are sent by means of form-based file upload in HTML.

Security
A Secure Socket Layer (SSL) connection ensures secure communication between the client browser and the tele-dermatology web-server. The SSL protocol encrypts the data transferred from the client to the server, and the other way around, by means of a public key infrastructure.

The user must log onto the system by providing his/her username and password. The security mechanism of the database authenticates the user to the system.

In this prototype, patient data are de-identified. Moreover, images of the pigmented lesions cannot disclose patient identity.
**Interaction modality**

The web-based tele-dermatology service is intended to be used by a requesting physician, who collects and stores into the system patient’s demographic and clinical data as well as digital images of the dermatologic problem. The physician can then ask advices to the consultant dermatologist in terms of diagnosis as well as treatment, through an integrated “e-mail” service. The consultant automatically receives the request and can then answer the question by looking at the data and images collected through the system. The communication and messaging between the consultant and the requesting physician are integrated within the system.
Appendix B

8. Clinical problem

Skin cancer is the most common type of cancer. In the United States they affect more than 1,000,000 people every year. Among skin tumors, melanoma is the most dangerous as it is responsible of more than 75% of skin cancer deaths. About 55,100 new melanoma diagnoses are estimated for 2004 in the United States, and about 7,910 patients are estimated to die for the disease [97,98]. Malignant melanoma develops from melanocytes, skin cells that produce the protective pigment melanin (see Figure 12).

![Figure 12 – Structure of the skin.](image)

The incidence of malignant melanoma has been constantly increasing worldwide during last century, especially for fair-skinned populations, at a rate of about 3-7% yearly (see Figure 13). In the U.S., in 1935, one’s estimated lifetime risk of disease was 1 in 1500; whereas in 2002 the lifetime risk was estimated at 1 in 68 [99].

92
Figure 13 - Age-adjusted (2000 U.S. standard population) melanoma incidence (per 100,000) 1973-2000. [100]

Fortunately, although the growing incidence of this disease over the years, the mortality rates have been rising at a rate of increase lower than that for melanoma incidence (see Figure 14).
Moreover, despite the rising incidence of the disease, the overall survival has shown a marked improvement over the years. The five-year survival improved from about 40% in the 1940s to about 90% now [99].

The main reason for this improvement in survival can be attributed to the earlier detection of melanoma. As a matter of fact, the median tumor thickness decreased over time, though the number of thick tumors remains steady. Breslow thickness directly measures, by using an ocular micrometer, the depth of tumor cells that have invaded the skin from the granular layer of the epidermis [102]. This parameter is one of the most important prognostic factors and the first criterion for determining the T staging. Figure 15
APPENDIX B. CLINICAL PROBLEM

reports the survival curves for the first three stages which are mainly related to the Breslow thickness.

Better treatments cannot account for this improvement as there have been no major changes in melanoma treatment. Tumor thickness is the main prognostic factor because malignant melanoma is capable of deep invasion and then it can spread widely over the body via the lymphatic and the blood vessels. In fact, in the first phase of melanoma development, it typically spread the skin surface. In a second phase it grows in depth, reaching the vessels by which it can spread over the body. Very often, there is a prolonged horizontal growth phase during which time the tumor expands centrifugally beneath the epidermis but does not invade the underlying dermis. This horizontal growth phase may provide lead time for early detection. Melanoma is more easily cured if treated before the onset of the vertical growth phase with its metastatic potential.
This is the reason why the early diagnosis of melanoma, and the subsequent surgical excision, are the key factors for the successful prognosis of this disease. Unfortunately, diagnosis of this kind of cancer is difficult and requires a well-trained dermatologist, because early lesions often have a benign appearance. The usual clinical practice of melanoma diagnosis is a visual inspection of the skin. A simple rule has been established to help non-expert physicians as well as the population through self-screening, in the diagnosis of malignant melanoma [104]. During the inspection of a pigmented skin lesion, asymmetry, border, color and diameter are evaluated. This procedure is called ABCD rule. This simple rule, however, does not completely address the problem of the recognition of early lesions.

Unfortunately, not all pigmented skin lesions can be diagnosed correctly by their clinical appearance. Additional criteria are required for the clinical diagnosis of such lesions. In vivo epiluminescence microscopy (ELM, also called dermoscopy, dermatoscopy, or surface microscopy) has become an increasingly popular method of inspecting lesions [14,15]. It is a non-invasive method that allows in vivo examination of skin lesions. Covering the lesion with immersion oil and a glass slide highly reduces reflectance of light from the surface. This reduction of scattered light permits the investigators to look through the surface structures of the skin, making accessible structures that are beyond to the skin surface. This feature greatly increases the morphological details that are visualized, providing additional diagnostic criteria to the dermatologist (see Figure 16) [14,15]. However the accepted “gold standard” for the ultimate diagnosis is still the histological examination and therefore the lesion must be surgically excised for undergoing to this kind of examination.
ELM inspection is driven by protocols, which require the assessments of specific features of lesions. Pattern analysis has been one of the first methods developed for the diagnosis of PSL, describing the entire set of lesion characteristics that has to be assessed [105]. Unfortunately, pattern analysis relies on features that can be difficult for non-expert to properly recognize. There is a lack of proper definition of these parameters, introducing subjectivity.

Some studies highlighted that only well-trained dermatologists can benefit from this technique, while non-expert physicians can even getting worse [16,84]. Moreover, a consensus among expert dermatologists on the dermoscopic features can help to improve the diagnosis suggesting that inter-rater variability is the main factor for the different performances of dermoscopy [16,17]. To reduce inter-rater variability and help non-trained physician to benefit this technique several simplified methods have been developed. One of these is the ABCD rule of dermoscopy [106] whereby physicians assess Asymmetry of the lesion, irregularity of the Border, presence and the distribution of Colors and presence of some Differential structures (brown globules, black dots, radial streaming and pseudopods, etc.). However, recently it seems that pattern analysis is the best method for training novices [107].

Several studies have shown that the diagnostic accuracy of a trained dermatologist is about 75%-80% for early melanomas, but
APPENDIX B. CLINICAL PROBLEM

it reduces to about 20%-30% for non-specialists (for instance general practitioners) [108,109].
Those aspects gave rise to several studies investigating the possibility of computerized systems that support physicians in the recognition of early melanoma.
9. Bibliography


[61] I. Bichindaritz Solving safety implications in a case based decision support system in medicine. In Workshop on CBR in the Health Sciences, 9-18. 2003


