Modeling Clinical Judgment and Implicit Guideline Compliance in the Diagnosis of Melanomas Using Machine Learning

Andrea Sboner[1], PhD Constantin F. Aliferis[2], MD, PhD
[1]ITC-irst and Department of ICT, University of Trento, Trento, Italy
[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Abstract
We explore several machine learning techniques to model clinical decision making of 6 dermatologists in the clinical task of melanoma diagnosis of 177 pigmented skin lesions (76 malignant, 101 benign). In particular we apply Support Vector Machine (SVM) classifiers to model clinician judgments, Markov Blanket and SVM feature selection to eliminate clinical features that are effectively ignored by the dermatologists, and a novel explanation technique whereby regression tree induction is run on the reduced SVM model's output to explain the physicians' implicit patterns of decision making. Our main findings include: (a) clinician judgments can be accurately predicted, (b) subtle decision making rules are revealed enabling the explanation of differences of opinion among physicians, and (c) physician judgment is non-compliant with the diagnostic guidelines that physicians self-report as guiding their decision making.

Introduction
Modeling of clinical decision making and judgment is one of the most intriguing areas of research in biomedical informatics. The broad goals of such modeling are typically to (a) understand the laws and factors that govern physician decision making, (b) identify limitations of clinician information processing, and (c) improve the relevant decisions using a variety of approaches (e.g., improved training, improved information availability, computerized decision support, formalized guidelines, etc.)[1,2,3,4]

There exist two main methodological approaches in the related literature. The *cognitive science/information processing approach* uses cognitive models of human decision making performance, relies more heavily on protocol analysis techniques, and may be more descriptive than predictive.[1,2,3,4] The *actuarial approach* uses statistical or decision-theoretic models of decision making, and is more quantitative and prescriptive. The use of statistical approaches in particular goes back to the 1950's exemplified by Brunswick's widely-used and influential "lens" model.[5]

In the present research we focus on the challenging domain of early diagnosis of melanoma, a task that is characterized by significant complexity and importance since: (a) it requires complex perceptual as well as cognitive skills, (b) it is performed by specialists as well as generalists, (c) it affects a large portion of the population, and (d) misdiagnoses are high-stake. Furthermore, our experiments clearly benefit from the fact that there exist different guidelines for specialists and non-specialists in this domain.

We address three interrelated hypotheses regarding the feasibility of:
1. Accurately modeling (and predicting) both specialists' and non-specialists' judgments.
2. Explaining why a clinician makes the decisions she makes, and why she may disagree with another clinician.
3. Comparing the actual strategy for clinical judgment employed by each clinician with the self-reported source strategy (thus detecting guideline compliance and studying self-awareness of compliance).

To investigate these hypotheses we apply machine learning methods from the Support Vector Machine (SVM) family, combined with SVM-based and Markov Blanket multivariate variable selection. The SVM classifiers allow us to analyze datasets with many diagnostic variables and modest samples that are difficult to analyze with more traditional statistical approaches. Moreover, by using advanced variable selection methods and a novel meta-learning strategy we identify small sets of variables that are needed to emulate physician behavior, identify patterns of diagnostic variables that describe physician behaviors, and also explain physician behavior and inconsistencies among physicians. An important component of the work reported here is the comparison of machine learning models of physician behavior with formal guidelines showing that guideline compliance can be automatically assessed and deviations from formal guidelines can be identified and assessed even when the processing of the relevant information is implicit in the clinician's decisions (i.e., it is not recorded which pieces of information the clinicians use to infer a diagnosis and how).

Clinical Background
Skin cancer is the most common type of cancer. In the United States alone it affects more than 1,000,000 people every year. Melanoma is the most dangerous

among skin tumors, as it is responsible for more than 75% of skin cancer deaths. Early diagnosis of melanoma, and its consequent surgical excision, is a key factor for good prognosis of this disease. Unfortunately, diagnosis of this kind of cancer is difficult and requires a well-trained dermatologist because early-stage lesions often have a benign appearance. In fact, the usual clinical practice of melanoma diagnosis involves a visual inspection of the skin, thus requiring perceptual as well as cognitive skills. Dermoscopy is a non-invasive clinical technique that provides physicians with additional diagnostic criteria, making accessible structures that are beneath the skin surface.[6] The visual interpretation of those parameters is highly dependent on the physician's skill and experience, thus, unfortunately, dermoscopy requires experienced physicians to be carried out effectively.[7]

Methods

We model the physicians' clinical judgments by means of the SVM classifier. SVMs are based on a sound mathematical framework and present several advantages compared to other pattern recognition methods, including the ability to handle large numbers or predictors with relatively small sample.[8]

Our second set of methods includes multivariate variable selection so that we can identify which of the many features that are available to the physicians are used and which are ignored. We use HITON_PC, HITON_MB, and Recursive Feature Selection (RFE) as feature selection algorithms.[9,10] The former methods are based on a theoretical framework that ties together feature selection and graphical models and provides theoretical guarantees for the optimality (i.e., minimal predictor number, maximal predictiveness) of the selected feature set by identifying the Markov Blanket of the response variable (algorithm HITON_MB) or an approximation to the Markov Blanket (algorithm HITON_PC) when computational or distributional considerations dictate so. The RFE method instead selects features based on the contributions (weights) of the features to the SVM hyperplane (i.e., near-zero weight features can be safely ignored without altering the SVM's behavior).

Our third and final method is a meta-learning explanation technique whereby once an SVM model of a physician's judgment is built (after feature selection) we assemble a new training set comprising of the skin lesion feature patterns and the SVM's continuous output for each lesion and learn a regression tree of the SVM model (i.e., not the original data).[11] Thus we can derive tree-like or (equivalently) rule-like human-interpretable descriptions of the SVM model. We emphasize that, non-linear SVM models are "black boxes" and cannot be understood in terms of what patterns of findings lead to a particular diagnosis. We note that although a similar technique was applied by Aphinyanaphongs and Aliferis,[12] the approach here is novel since these researchers do not explain the SVM model per se, but build a decision tree model from the original data. A balanced nested n-fold cross-validation procedure was applied to both optimize the parameters of the SVM models and to estimate the generalization error using area under the ROC curve (AUC) as the metric of choice. SVMs models were built using the LibSVM package.[13] All other types of models and overall experiments were carried out by custom Matlab code on a Windows platform.

Table 1 Feature set.

Objective Features	Type/values
Lesion location	Nominal
Max-diameter	Numeric (mm)
Min-diameter	Numeric (mm)
Evolution	Binary
Age	Numeric (year)
Gender	Binary
Family history of melanoma	{yes, no, not det.}
Fitzpatrick's Photo-type	{1,2,3,4, not det.}
Sunburn	{yes, no, not det.}
Ephelis	{yes, no, not det.}
Lentigos	{yes, no, not det.}
Subjective Features	**Type/values**
Asymmetry	Binary
Irregular Border	Binary
Number of colors	Ordinal [1-6]
Atypical pigmented network	Binary
Abrupt network cut-off	Binary
Regression-Erythema	Binary
Hypo-pigmentation	Binary
Streaks (radial streaming, pseudopods)	Binary
Slate-blue veil	Binary
Whitish veil	Binary
Globular elements	Binary
Comedo-like openings, milia-like cysts	Binary
Telangiectasia	Binary

Data

We collected patient data on a total of 177 pigmented skin lesions, diagnosed as 76 malignant melanomas and 101 benign lesions as determined by histological examination. All pigmented lesions were consecutively clinically and histologically examined in patients presenting at S. Chiara Hospital in Trento (Italy) during the usual activity from June 1999 to September 2002. During the face-to-face visit, objective information, such as patient's age, lesion evolution, etc. as well as digital images of the lesions were acquired. Each case was then submitted through a WWW-based system to six dermatologists (3 expert dermatologists who are routine users of dermoscopy at a University Medical School, and 3 less-experienced dermatologists, seldom using dermoscopy for melanoma diagnosis and working in outpatient clinics). Objective information and digital images of the lesions were available to the physicians

so as to provide them with as much information as possible to perform the dermoscopy evaluations. In addition to dermoscopy evaluation, physicians also rendered their clinical diagnosis: benign or malignant lesion. Table 1 depicts the features used in this study.

Self-reported guidelines

All physicians, after completing their diagnoses, were requested to report in writing the guidelines they use for inspecting skin lesions and in particular to evaluate and interpret the dermoscopy parameters, i.e. the subjective features and to provide references to those guidelines.

Results

1. Modeling physician judgment. Table 2 shows the results of modeling clinician diagnoses. The performances of the optimized SVM classifiers are very high for every physician: the average AUC across the physicians is 0.94 [0.89, 1.00] using all the features.

Table 2 – Predicting clinical diagnosis: model performances for each physician in terms of AUC. Each column corresponds to no feature selection ("all") or to one of three different feature selection methods. In parentheses the number of features used by each model.

Physicians	All (24)	HITON_PC	HITON_MB	RFE
Exp1	0.94	0.92 (4)	0.92 (5)	0.95 (14)
Exp2	0.92	0.89 (7)	0.90 (7)	0.90 (12)
Exp3	0.98	0.95 (4)	0.95 (4)	0.97 (19)
NonExp1	0.92	0.89 (5)	0.89 (6)	0.90 (22)
NonExp2	1.00	0.99 (6)	0.99 (6)	0.98 (11)
NonExp3	0.89	0.89 (4)	0.89 (6)	0.87 (10)

2. Identifying overlooked/ignored features. As evident in Table 2, the AUC is still high after applying feature selection methods. The average AUCs are 0.92 (range [0.89,0.99]), 0.92 (range [0.89,0.99]), and 0.93 (range [0.87,0.98]) for the HITON_PC, HITON_MB, and RFE methods, respectively. In particular, the methods based on Markov Blanket dramatically reduce the number of features while keeping the AUC very high. This suggests that only few features influence the physicians' diagnoses.

3. Deriving trees that explain the physician judgments. We build a decision tree for each clinician using the reduced set of variables for that clinician. This can help explain why clinicians agree or disagree. The Squared Correlation Coefficient R^2 between SVM and Decision Tree output is very high for all the physician's Decision Tree models (average: 0.99, range [0.94,1.00]). This means that each Decision Tree is able to "explain" well the corresponding SVM model.

As an example, figure 1 shows simplified excerpts of the Decision Trees for the first two experts. Let us suppose that expert 1 assess the presence of slate blue veil and streaks. Following the right branches of the tree, the final clinical diagnosis is malignant. In the same way, expert 2 would render a malignant diagnosis when streaks are present, but only if number of colors is greater than 3 or the lesion is asymmetric.

On the other hand, if expert 1 assesses the presence of streaks and less than 4 colors, the clinical diagnosis is benign only if the skin lesion has not changed (evolution). If this is the case, expert 2 would consider other variables (e.g. asymmetry) in order to decide. The presence of streaks highly suggests a melanoma for both experts.

4. Assessing guideline compliance. The 3 experts and non-expert 1 reported that they employ the so-called *pattern analysis*.[14] With this technique, physicians assess all the dermoscopy parameters altogether to reach a decision. In our dataset then, all subjective features are supposed to be used by those physicians to render their diagnoses.

Non-expert 2, instead, reported that he employs the ABCD rule for dermoscopy.[15] By means of this technique he focuses on asymmetry, irregular border, number of colors (associated also to slate blue veil and whitish veil), and the presence of differential structures (streaks, globular elements, etc.). The guideline requires computing a weighted score of these features. However, the method in practice usually helps in focusing on the most important parameters during a skin lesion examination. The parameter evolution is usually considered, becoming the "ABCDE rule".

Non-expert 3 reported that after a general visual inspection of the lesion, she focuses on slate blue or whitish veil and atypical network and on the presence of many colors to assess the malignancy. If she has not reached a decision so far, then she considers the other parameters: irregular border, streaks, etc. Sometimes she uses the ABCD rule for dermoscopy for confirmation purposes. However, when she has still some doubts, she performs again "general evaluation" of all parameters.

The 3 experts and non-expert 1 stated that they use pattern analysis, i.e. they use all the available subjective features. The low number of features selected by the algorithms shows that this is not the case, especially for expert 3 whose model is the most accurate among experts' models.

Non-expert 2 stated that he uses ABCDE rule (Asymmetry, Border, Color, Differential Structures, and Evolution) but in the built model asymmetry, irregular border and evolution are not present.

Non-expert 3's model includes slate-blue veil, irregular border and number of colors as stated by the physician, but other parameters are missing.

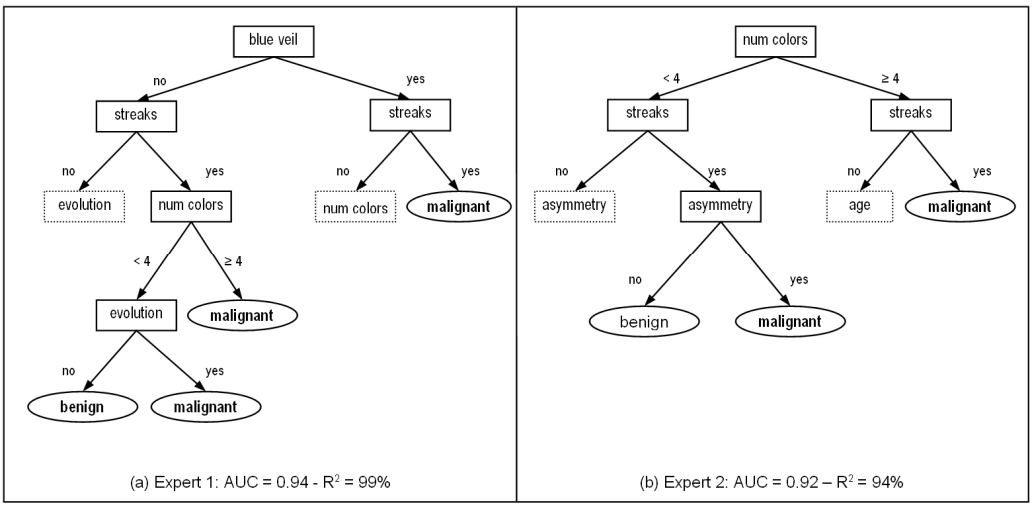

Figure 1 Simplified excerpts of Decision Trees for Expert 1 (a) and Expert 2 (b). Due to space limitations the complete Decision Trees are available upon request as supplemental material. The dotted shapes mean that further branches are present. Moreover, some splitting nodes leading to the same final diagnosis (with different degree of certainty) are grouped together in the leaves. The AUC with respect to the clinical diagnosis, and the R^2 correlation coefficient compared with the SVM predictions, are reported.

In summary these results suggest that physicians are not compliant to the diagnostic guidelines that they themselves identify as appropriate for the task and state that they follow. This is true for both specialists and non-specialists. It is worth noting that actual guidelines are not tree-structured as each physician's empirical model. We can thus perform less direct comparisons, e.g., which features are used and which ones are discarded, between free-text guidelines and physician trees. A more direct comparison is possible among physician models and is described in the next section.

5. Explaining disagreement among physicians & diagnostic errors. We measured the agreement between physicians for each subjective feature as well as for the clinical diagnosis using Cohen's kappa.[16] Our results showed a high level of disagreement regarding the subjective parameters, (average κ=0.32, with κ ranging from 0.10 for comedo-like openings to 0.48 for streaks and slate-blue veil) in accordance with other studies.[17] The disagreement on the clinical diagnosis is lower (κ=0.63). This suggests that the inference mechanism is different among physicians, in addition to their subjective interpretation of the lesions.

As an example of patient case-by-case comparison of physician models, consider a case that was differently diagnosed by expert 1 and expert 2. Expert 1 assessed the presence of whitish veil, streaks, irregular pigmented network and abrupt cut-off of the network, and 3 colors. Moreover, the lesion changed over time. According to both the decision tree corresponding to expert 1 and his self-reported guidelines, the clinical diagnosis is malignant. Expert 2 assessed the presence of irregular border, irregular pigmented network, abrupt cut-off of the network, streaks, globular elements and 3 colors. The clinical diagnosis is benign and in accordance with the Decision Tree for expert 2. This is rather surprising given that the two experts evaluated similarly 4 parameters (streaks, irregular pigmented network and abrupt cut-off of the network, and 3 colors). This suggests that the differences between these two physicians for that patient are due to the inference mechanism rather than to the subjective interpretation of the features. This is confirmed by the different features that are present in their Decision Trees.

Conclusions & Limitations

The experiments presented demonstrate the feasibility of using machine learning techniques to better understand aspects of physician diagnostic judgment in a non-trivial medical domain. It was found that such techniques can accurately model (and predict) both specialists' and non-specialists' judgments. These models of clinician judgment can be analyzed with feature selection techniques to identify diagnostic features that are redundant, and can be converted to human-interpretable decision trees that more vividly capture the underlying patterns of decision making. These representations can then be used to compare the diagnostic process of different physicians explaining their differences of opinion according to which variables they focus on and how they weight them. We were also able to demonstrate that non-compliance to gold-standard guidelines is widespread among physicians, both specialists and non-specialists.

An important -- more theoretical than practical, but still worth-mentioning -- limitation is that the derived models of clinician judgment are "paramorphic",[1] meaning that clinicians most likely do not apply decision trees or any of the employed machine learning models when making diagnostic decisions. On the other hand from a *functional* perspective if a decision tree captures a clinician's behavior almost perfectly, predicting her diagnoses and using a limited set of features, then effectively the physician's judgment cannot be empirically distinguished from such a diagnostic model. This in turn fully justifies describing the physician on a functional (input-output) level using that machine learning model.

It is worth noting that the models we created are not supposed to improve on human diagnosis. In fact, they are strictly focused on modeling physician logic. An interesting future research path is the creation of models that can help physician's diagnostic process.

In the present preliminary report we provide only a few examples of the application of our methodology. A thorough comparison among physicians and among physicians and guidelines as well as data derived computer models will be provided in a forthcoming journal paper.

A more practical limitation of the specific techniques employed here is that they are better suited to perceptual diagnosis and thus will require extensions to be used for analysis of other, more complex processes of clinical decision making (e.g., sequential gathering of diagnostic information with periodic re-assessment of the patient's status and related therapeutic options). This is clearly an exciting area for future research.

Moreover, the experimental design implemented in the reported work controls for interactions of the clinical diagnostic process with organizational factors. Such factors may further affect the physicians' decision making in this domain thus and deserve further study.[18]

Despite the above limitations, the techniques discussed here open up several more potentially valuable methodological and medical possibilities that we are in the process of exploring. These include the construction of computerized as well as hybrid computer/human classifiers in this domain, explaining disagreement of clinicians with automated decision support tools, using the analysis to improve physician diagnostic skills and reduce errors, and deriving high-performance but simplified versions of the established guidelines in this domain.

Acknowledgements

The first author would like to thank Riccardo Bellazzi and Paolo Traverso for their support in Italy, and all the dermatologists involved in the study: A. Bergamo, P. Bauer, P. Carli, A. Chiarugi, M. Cristofolini – Lega per la Lotta contro i Tumori, P. Nardini, and C. Salvini. Both authors are indebted to Alexander Statnikov for his enthusiastic help with the code and computer infrastructure needed to run the experiments.

References

1 Dowie J, Elstein AS, editors. Professional Judgment, New York: Cambridge University Press; 1988.

2 Elstein AS, Shulman LS, Sprafka SA, editors. Medical Problem Solving. Cambridge: Harvard University. Press; 1978.

3 Patel VL, Arocha JF, Kaufman DR. A primer on aspects of cognition for medical informatics. J Am Med Inform Assoc. 2001;8(4):324-343.

4 Arocha JF, Wang D, Patel VL. Identifying reasoning strategies in medical decision making: a methodological guide. J Biomed Inform 2005;38(2):154-171.

5 Brunswick E. Representative design and Probabilistic Theory in a Functional Psychology. Psychological Review. 1955;62:193-217.

6 Pehamberger H, Binder M, Steiner A, Wolff K. In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma. J Invest Dermatol. 1993;100(3):356S-362S.

7 Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol. 2002;3(3):159-165.

8 Scholkopf B, Burges CJC, Smola AJ, editors. Advances in kernel methods: support vector learning. Cambridge: The MIT Press; 1999.

9 Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annu Symp Proc. 2003:21-25.

10 Guyon I., Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389-422.

11 Breiman L, Friedman JH, Olshen RA. Stone CJ, Classification and Regression Trees. Boca Raton: Chapman & Hall; 1993.

12 Aphinyanaphongs Y, Aliferis CF. Learning Boolean Queries for Article Quality Filtering. In Proc. 11th World Congress on Medical Informatics (MEDINFO), September 7-11, 2004, San Francisco, California, USA.

13 Chang CC, Lin CJ. LIBSVM: a library for support vector machines. National Taiwan University. 2003.

14 Pehamberger H, Steiner A, Wolff K. In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions. J Am Acad Derm. 1987;17(4):571-583.

15 Nachbar F, Stolz W, Merkle T et al. The ABCD rule of dermatoscopy. J Am Acad Derm. 1994;30:551-559.

16 Cohen J, A Coefficient of Agreement for Nominal Scales Educ Psychol Meas 1960;20:37-46.

17 Argenziano G, Soyer HP, Chimenti S, et al. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. J Am Acad Dermatol. 2003;48(5):679-693.

18 Lorenzi NM, Reily RT. editors. Managing Technological Change: Organizational Aspects of Health Informatics 2nd ed. New York: Springer Verlag, 2004.