Reviewer's report

Title: Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system

Version: 3 Date: 2 July 2014

Reviewer: helen Chen

Reviewer's report:

Two strengths are worth noting before addressing the rubric.

a. The potential benefit of using single scattered polarized light spectroscopy, and multiple scattered unpolarized light spectroscopy technology, and the way in which the features were extracted from those scans, appears to be promising for computer aided diagnosis of melanoma.

b. Using the probe only within the region of interest (ROI) without any need for segmentation is an interesting and potentially more efficient approach than those requiring segmentation.

1. Is the question posed by the authors well defined?

a. Major Compulsory Revision: The question posed by the authors regarding the "objectivity" of methods, is not clearly or properly expressed. A quantitative method is not de facto objective, even in comparison to a qualitative or subjective method. Features, model and optimization approaches (automated or not) can exhibit bias, and therefore may not be "objective“. Perhaps "not-subjective" or "quantitative instead of qualitative" are more appropriate words or concepts. The authors do not quantitatively show or estimate the bias or lack thereof, in their model.

2. Are the methods appropriate and well described?

a. Major Compulsory Revision: The results of the algorithms (e.g. 100% sensitivity) are not statistically significant because it only uses 5 runs in 1 experiment. Yet claims (strong claims in fact) are made with these results. As a rule of thumb, at least n>25 or n>30 (Hogg and Tanis) runs or participants are required to make a statistically significant conclusion about a run or participant in general. For example, this paper cites Argenziano et al, for comparison, which had 40 of 51 clinicians complete that study. Argenziano et al provides an instructive example of correct methodology for this and other comments. It should be simple for the author to perform 25 or 30 runs for at least 1 algorithm with the data at hand, make conclusions based on that algorithm, and contextualize the other results as "not statistically significant but indicative".

b. Major Compulsory Revision: Forcing the test and training sets to be balanced, which is completely different from the imbalanced population and sample...
distributions, critically affects the meaning and applicability of the result. 100% sensitivity with a completely unrealistic distribution is a useless result. Training and testing with a realistic distribution may yield a very different result.

With SVM it is important to address the effect of class imbalance (e.g. by using class-specific soft-margin parameter values \(C^+\) and \(C^-\) instead of just \(C\)), but creating a balanced subset of data for training and testing is not a sound approach.

c. Major Compulsory Revision: Even if the distribution were realistic, using a fixed amount of melanomas in the test set provides misleading results -- a fixed amount is easier for any classifier to predict (i.e. to arrive at optimal parameter values for a fixed number or fixed proportion of true positives). In real life, test samples will sometimes have less melanomas, or sometimes more. Argenziano et al demonstrate a sound approach: ensure that the dataset reflects a realistic distribution and randomly sample the whole training set and the whole test set from that population.

d. Major Compulsory Revision: The small sample of melanomas, 19, which is then split in half (e.g. 10) for training and testing, greatly impairs the precision (or positive predictive value, PPV) and the likelihood ratio positive (LR+) of the result. A single false negative, or the lack thereof, changes the result significantly -- e.g. the Precision jumps from 100% to 90% based on a single false negative. There is a balance between three things: having enough data to learn from in training, having enough data for precision in validation testing (to select the optimal parameters), and having enough data for precision in (final) testing (to obtain precise and meaningful results).

In comparison, Argenziano et al, had 33 melanomas in 128 lesions, with the training and test sets divided (by random sampling without replacement) into 20 and 108 samples respectively. i.e. approximately 5/6 or approximately 27 melanomas would show up on average in the test set (as compared with 10). With an automated algorithm for testing rather than people, Li et al have an opportunity to perform the experiment with more combinations of their data.

While the authors appreciate the value of disjoint training and test sets, they do not appear to comprehend how that is achieved in cross-validation. The authors state: "When the number of available samples is limited, people often use all available samples to train the classifier and use the same set as the testing set for cross-validation. However, good cross-validation results using the same data set for training and testing may not guarantee accurate classification of new data samples. In our experiment, we would like to have disjointed training and testing sets to ensure an accurate and robust classifier."

This fails to understand how n-fold or 10-fold or leave-one-out cross-validation work. The training and test sets are *ALWAYS* disjoint. The training from one fold to the next is separate -- the training is not aggregated, it is the results which are aggregated and used to select the hyperparameters that yield the best overall result. There are various sources that discuss cross-validation, how it
prevents overfitting, how it generalizes (i.e. reducing CV's estimated error, reduces the true error) and how it has a slight optimism or underestimation bias.

What cross-validation does not prevent is selective disclosure of results (e.g. running the result many times until you get a desired result based on chance, and only reporting the desired result). Having a "fixed" separate test set precludes that possibility.

Alternatively, given the limited number of melanomas, the authors may consider bootstrapping, which Steyerberg indicates is a preferred method over cross-validation. Bootstrapping is a well-known method that samples with replacement (i.e. training and testing sets are not disjoint, but that is part of how it models uncertainty, and it requires a lot of runs).

e. Major Compulsory Revision: What is the probability of error (or uncertainty) in the algorithm? With SVM, this may be reported using Platt's method which is reported by the LIBSVM library used by the authors (with the -b option). Even if the revised test results (revised according to the prior comments) produce 100% sensitivity with statistical significance, the probability of error should be reported, particularly with a 100% result.

f. Discretionary Revision: Since NB had the best overall result, why was SVM selected? It would help to understand the rationale.

g. Discretionary Revision: Argenziano et al are cited for the accuracy of dermatologists in the introduction. Since this article focuses on a 100% sensitivity result, it makes sense to report the sensitivity of dermatologists at some other point in the article? One cannot compare 100% sensitivity with 75-85% accuracy, since the latter tells me nothing about the sensitivity (it could be 100% as well; but it isn't in this case).

3. Are the data sound?

The data was collected by a dermatologist (BTH).

a. Major Compulsory Revision: How was ground truth established for this data? Were all of the lesions in the data set sent for pathology so that those labelled benign are in fact benign?

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?

The reporting is adequate. The authors do not indicate the availability of their data, i.e. their data deposition. I don't know what the data deposition standard is (e.g. publicly available data set via web) for this journal or this area/topic.

a. Discretionary Revision: At times the article refers to 100% sensitivity and high specificity. It would be much better served to be more explicit about the specificity up-front.
5. Are the discussion and conclusions well balanced and adequately supported by the data?

Due to major issues with the method, the conclusions from the data are not supported. Due to a lack of understanding of real-life use and issues, the discussion and conclusions regarding benefits and applications are not supported.

a. Major Compulsory Revision: The claim that the 100% sensitivity result "guarantees" that every melanoma will be found is erroneous. Aside from the probability of error in comment 2.e., a guarantee would have to ensure that changes in the device, the device's calibration, the protocols to use and/or calibrate the device, the environmental lighting/temperature/voltage and the person using the device, have no effect on the outcome. This article only speaks to variation in lighting and makes claims about variation in skin colour and age which are not sufficiently supported/cited.

This claim also assumes that the sample data set reflects the population. In some areas or countries the population may differ. The authors acknowledge limitations in their sample with respect to gender and age, but suggest (without sufficient support) that the method renders those factors irrelevant. It does not render them irrelevent if the types of melanomas they get differ and those melanomas did not occur in this data set.

b. Major Compulsory Revision: The authors claim that: "the influence of unbalanced gender and age as well as different features of human skin (i.e. the skin pigmentation or color variations) has been smoothed out due to the way that we calculated the pixel intensity. The pixel intensity differences of lesion skin and normal skin nearby of the same subject effectively removed the background information (e.g. age, gender, skin color) and correlated the intensity value with the pathological change only."

This claim needs to be supported by either a citation, evidence, proof or a convincing argument. The only sentence in this paper that appears to be related and weakly supportive (because its support is not direct or clear), is:

"We also found that the intensity distribution of V scan has demonstrated similar differences between all benign and all melanoma skin lesions."

c. Discretionary Revision: "the well-trained SVM classifier". Given comments 2a-2e re distribution and small sample size, I suggest avoiding the words "well-trained".

d. Major Compulsory Revision: The authors claim:

"Since the quantitative approach enables automatic diagnosis of melanoma, the diagnosis is more cost-effective and time-saving than clinical diagnosis."

Time savings imply collection by a non-dermatologist, which is contrary to how this data was collected. There is nothing time saving about a dermatologist
having to do something extra -- especially for every patient, which is necessary to accrue the claimed benefits of the study in this paper.

6. Are limitations of the work clearly stated?

Aside from a couple of limitations stated, the authors do not acknowledge the limitations of their work per the aforementioned comments 1a, 2a-2e, 3a, 5a-d.

a. Major Compulsory Revision: The claim that: "With this classifier, 82% of the benign lesion patients can avoid painful and costly biopsies" does not consider real life limitations. Specifically:
1. how does a doctor combine the machine's result with their own thinking? especially considering medical liability? from an ethical and liability perspective can they and should they rely on a machine? these are not light or simple issues! it is strongly recommended that any language implying real-life benefits be muted to only state the facts. it is strongly recommended that the authors describe the limitations of their approach (as described above).
2. as soon as this device does a false negative and it is discovered, can it survive the liability? if their is an end user license agreement absolving the device maker of any liability, then will doctors use it or trust it?
3. wouldn't doctors only trust this device once the FDA has approved it?

b. Discretionary Revision: With this small sample of only 19 melanomas in 168 lesions, how do we know that a "difficult" melanoma was included. That is, a melanoma that is difficult for a doctor to assess? What is the physician's sensitivity and specificity and precision on this particular data set?

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?

Not examined.

8. Do the title and abstract accurately convey what has been found?

Sufficient.

9. Is the writing acceptable?

The writing is acceptable. There are a couple of grammatical errors that are of little consequence.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:

I declare that I have no competing interests