

WAVELENGTH SELECTION BY GENETIC ALGORITHMS IN NEAR INFRARED SPECTRA FOR MELANOMA DIAGNOSIS

T.E.M. Nordling^{*,**}, J. Koljonen^{*}, J. Nyström^{***,****}, I. Bodén^{****,*****}, B. Lindholm-Sethson^{***,****}
P. Geladi^{*****} and J.T. Alander^{*}

^{*} University of Vaasa/Dept. of Elec. Eng. and Automation, P.O. Box 700, FI65101 Vaasa, Finland

^{**} TKK/Control Engineering Laboratory, P.O. Box 5500, FI02015 Helsinki, Finland

^{***} UmU/Dept. of Chemistry, SE90187 Umeå, Sweden

^{****} UmU/Centre for Medical Engineering and Physics, SE90187 Umeå, Sweden

^{*****} UmU/ Dept. of Surgical and Perioperative Sciences, Surgery, SE90187 Umeå, Sweden

^{*****} SLU/The unit of Biomass Technology and Chem., P.O. Box 4097, SE90403 Umeå, Sweden

tn@kth.se

Abstract: Early, reliable and fast diagnosis of melanoma is particularly important as the number of cases is increasing. In this paper, the potential of using near infrared spectroscopy for melanoma diagnosis is studied. The classification task is complicated by a low signal-to-noise ratio and the high dimensionality of the spectral data. Thus pre-selection of wavelength variables is required. Atypical naevi samples of patients were clinically classified, using the ABCD rule, and their near infrared spectra recorded. A nonlinear clustering model for spectral based classification was calibrated to the spectra and pathologist's classification using a genetic algorithm. The genetic algorithm optimized the spectral based classification by selecting wavelengths correlated to melanoma. Some wavelength selections allowed correct classification of all samples in our dataset. The small size of the dataset and uncertainty in the clinical classification, however, limit the conclusions that can be drawn. Evidence for the existence of spectral regions that contain information needed for melanoma diagnosis is presented.

Introduction

Malignant melanoma or skin cancer is a deadly cancer form. The number of cases is increasing in many parts of the world. In the USA one in 70 persons is estimated to get melanoma during their lifetime. The American Cancer Society reported more than 7000 deaths from malignant melanoma in year 2000. In Sweden the mortality was approximately 360 per year during the same time, as reported by Statistics Sweden.

Melanoma gives dark spots in the skin because of melanin accumulation in cancer cells. No simple definition that distinguishes malignant melanoma from other atypical naevi exists. There are many forms that would confuse even the experienced oncologist or dermatologist. Harmless skin conditions may give dark spots of approximately the same size and shape. In

many patients, dark spots are surgically removed under the philosophy of 'better safe than sorry'. This shows that there is a clear need for early detection and easier classification of melanomas.

The classical subjective diagnosis is based on the ABCD rule where Asymmetry, Border, Colour and Diameter of the lesion are considered [1, 2]. The most objective measurement available is a slow pathological study after surgical removal of a lesion. A simple early diagnosis would help avoid metastases leading to death but also reduce the number of operations. Noninvasive measurements are quick and cheap and can give information on skin status, including melanoma-related changes, but more systematic studies of larger patient groups are needed. It was shown that electrical bioimpedance is an aid in the separation of benign naevi and malignant melanomas [3].

A number of publications propose spectroscopy in the visual and near infrared (NIR) range for studying skin lesions and discriminating between malignant melanomas and harmless discolorations. Bono et al. [4] describe how naevi and melanomas show up below 700 nm in a CCD camera. Melanomas remain visible between 700 and 1040 nm while the naevi disappear. Elbaum et al. [5] use multispectral imaging between 430 and 950 nm for classification of lesions. Langley et al. [6] use confocal imaging at 1064 nm for studying individual cells in lesions.

The strongest peaks in the NIR spectra are caused by water. This is a common motivation for measurement of water content and/or macromolecule water binding in NIR cancer studies. Gniadecka et al. [7] have measured water content and structure in malignant and benign skin tumors. They did not find any difference in water content between malignant and benign skin tumors with an exception of seborrheic keratosis, in which the water content was decreased. However, they found differences in water structure: the free (tetrahedral) water was increased in malignant skin tumors and sun-damaged skin, relative to normal young skin and benign skin tumors.

The above results are somewhat contradictory to some other studies that have found a higher concentration of water in tumor-bearing tissues. (See [7] for references)

Hirosawa et al. [8] have used Fourier transform NIR spectroscopy for *in vivo* investigation of rat mammary gland tumors. They have found small differences between the peaks of normal and tumor tissues in the second derivative of the spectra.

In the present paper, a group of patients was studied with the traditional clinical diagnosis methods (ABCD) and NIR spectroscopy. For some patients more than one lesion was studied. NIR spectra of adjacent healthy skin were taken as a reference. This paper introduces classification optimized by genetic algorithms (GAs), as a means of classifying skin lesions and identifying spectral regions important for melanoma diagnosis.

Materials and Methods

We use NIR spectra recorded from malignant melanoma and non-malignant naevi skin spots of patients to calibrate a classifier. (The spectra can be seen in Figure 2.) This study is based on 20 samples of atypical naevi (Table 1), diagnosed by classical means (ABCD) as eight superficial spreading melanoma type (SSM), two nodular malignant melanoma (N), three lentigo maligna melanoma (L) and seven benign naevi (G). However, at least one N-case (m52a12) may as well be of SSM type. A reliable diagnosis is difficult to achieve with the ABCD rule, since it only has a stated accuracy of 80% [9].

The NIR spectrum is recorded directly at the interesting skin spot of the patient, as assigned by the doctor. A reference spectrum of normal healthy skin (R) is recorded approximately 2 cm from the spot. The same reference spectrum was used if two spots were closely located on a patient; hence the dataset contains 17 independent reference spectra.

The spectra were recorded using a Bruker Matrix-F, FT-NIR spectroscope with a Peltier cooled InGaAs detector, in the wavelength range 833 to 1442 nm. The spectra contained increasing noise in the upper end and were therefore cut at 1442 nm. The resolution of the spectroscope was set at 4 cm⁻¹, which after the built-in Fourier transformation yields an approximate spacing of 0.4 nm.

Suitable pre-processing has in many works been used to increase the signal-to-noise ratio [10, 11]. We therefore tried several alternatives, e.g. removal of the mean of each spectrum, Standard Normal Variate Transformation, detrending by a quadratic polynomial or taking the first two derivatives. We also tried subtracting the reference spectra from the spectra recorded at atypical naevi. It proved to be much easier to decrease the signal-to-noise ratio than to increase it, thus no pre-processing is done in this paper before the wavelength selection. Some pre-processing methods, however, showed a bit of promise, considering that comparison of different pre-processing techniques is difficult, since no self-evident and objective definition of the signal-to-noise ratio exists in this case.

Table 1: The dataset, calibration (C) and test (T) subsets, and the clinical classification of the samples.

Number	Sample ID	Set	Class	Diagnosis
1	m100a12	C	SSM	Atypical naevi transitional to SSM II
2	m100a22	C	SSM	Atypical naevi transitional to SSM II
3	m36a12	C	L	Lentigo maligna melanoma Clark II
4	m38a12	C	SSM	Melanoma in situ of SSM-type
5	m38a22	C	SSM	Melanoma in situ of SSM-type
6	m39a12	C	SSM	Melanoma in situ of SSM-type
7	m41a12	C	N	Nodular malignant melanoma
8	m42a12	C	G	Benign naevi
9	m42a22	C	G	Benign naevi
10	m43a12	T	G	Intravenous naevus
11	m44a12	T	SSM	SSM-type Clark III
12	m45a22	T	G	Merged naevus
13	m46a12	C	G	Actinic keratosis
14	m47a12	C	SSM	SSM-type Clark III
15	m48a12	C	L	Lentigo maligna melanoma
16	m49a12	C	G	Cavern bluffer
17	m50a12	C	G	Dysplastic naevus
18	m51a12	T	L	Lentigo maligna melanoma
19	m52a12	C	N	Nodular malignant melanoma
20	m52a22	T	SSM	Malignant melanoma of SSM-type
21	r100a12	C	R	Healthy skin
22	r36a62	C	R	Healthy skin
23	r38a32	C	R	Healthy skin
24	r38a42	C	R	Healthy skin
25	r39a12	C	R	Healthy skin
26	r41a12	C	R	Healthy skin
27	r42a12	C	R	Healthy skin
28	r43a12	T	R	Healthy skin
29	r44a12	T	R	Healthy skin
30	r45a22	T	R	Healthy skin
31	r46a12	C	R	Healthy skin
32	r47a12	C	R	Healthy skin
33	r48a12	C	R	Healthy skin
34	r49a12	C	R	Healthy skin
35	r50a12	C	R	Healthy skin
36	r51a12	C	R	Healthy skin
37	r52a12	T	R	Healthy skin

We have defined all parts of the spectrum that help distinguish the four classes, i.e. SSM, N, L and G/R, as signal. All other parts are considered noise, since they are uninteresting. The method of this paper is based on the following two assumptions: Firstly, existence of signal, i.e. wavelength ranges for which a difference in intensity between the four classes can be observed. Secondly, measurement noise that is random, varying or at least does not completely hide the signal. If these two assumptions are true, then it is possible to perform classification based on the NIR spectra of melanoma and reference samples of healthy skin. The first assumption is based on a hypothesis that the concentrations of molecules that can be observed in the NIR range change as the skin cells undergo malignant transformation.

A genetic algorithm (GA) is used for wavelength selection, since GAs are 'building block' algorithms, i.e. construct solutions by combining good blocks, with a capability of getting stuck to local optima. We would like to claim that the global optimum is uninteresting in all cases characterized by heavy measurement noise, since the model then would be optimized to the random noise of the particular dataset. The ability to get stuck to local optima might, in our opinion, thus be one reason for the success of using GAs for wavelength selection in spectroscopy [12, 13].

The employed GA is a modified version of the Core GA used in Nordling et al. [14]. The fitness function is minimizing the mean probability of wrong classification of the samples in the dataset. Spectral ranges selected by the GA, i.e. intensity of certain wavelengths, and the sample classifications, done by traditional clinical means, are given as the input and the desired output, respectively, for the multivariate calibration model. We have termed our calibration model adaptive hierarchical clustering (AHC), since the classical hierarchical clustering automatically is stopped when two groups of different classes are about to be combined into one cluster.

Uniform crossover of binary chromosomes, where 1 denotes inclusion and 0 exclusion of the wavelength variable, is used together with either a gene or chromosome based mutation operator, (see [14] for more information). Many different values, which are considered reasonable, are used for the GA parameters, since the brute force algorithm that was proposed by Nordling et al. [14] is employed. In short, the idea of the brute force algorithm is to use many different GA parameters to generate many solution candidates and then to select one based on the number of solution candidates that the wavelengths are selected in.

In the AHC model, all samples are regarded as points in a space having one dimension for each wavelength variable included in the data given to the model. In hierarchical clustering the samples are first ranked based on increasing distance to the closest neighbour sample. Next the class of the two samples separated by the shortest geometrical distance is checked. If they belong to the same class, then they are united into a cluster. This cluster will in the next round be represented by its centroid, i.e. the point exactly in the middle between the two points on a line through them. The centroid is used because it is in general more stable than the nearest neighbour. After this the ranking is redone and the two samples, or cluster centroids, with the shortest distance are again looked for. This continues until the two samples with the shortest distance are of different class. The centroids of the clusters that exist at this point, some clusters might consist of only one sample, are used for prediction of the class of a new test sample. The test sample is assigned to the same class as the cluster having the closest centroid.

Adaptive hierarchical clustering is used, since samples of class SSM, for example, can form two clusters in such a way that their centroids, if they were merged as one cluster, would be in the middle of a cluster of another class. In other words, another cluster may be located in between the two SSM clusters i.e. a nonlinearly separable classification is allowed. Many different classes of melanoma exist and we have by constructing the AHC model in this way made sure that it would work even if any of the commonly used melanoma classes contain two or more subtypes. When samples of the same class are clustered into several clusters, it actually suggests the existence of subtypes.

The fitness measure, 'probability of wrong classification', $P(\sim A)$, is based on a ratio between the

closest centroid of the same class, a , and the closest centroid of different class, b :

$$P(\sim A) = a/(a + b). \quad (1)$$

It will be zero if the test sample is located exactly at the same point as a centroid of a cluster of the same class and one if the test sample is located on a centroid of a different class. Our dataset is so small that any fitness measure based on the number of correctly classified samples would lead to discrete steps that would cause an improper ranking of the chromosomes and hence low convergence of the GA. 'Probability of wrong classification' gives a continuous measure, avoiding all problems with improper ranking. It is intuitive, simple and works in practice.

The summation and weighting of the probabilities of wrong classification for several test samples can be done in different ways. We have used the mean, mainly because it is simple. Instead a weight function giving a higher weight with increasing probability could be motivated by the medical fatality of making a wrong classification.

We divided our dataset into two sets in order to get a test set that is independent of the calibration set (Table 1). The wavelength selection and the AHC model calibration were first done on the training set, and the test set was predicted by using the previously achieved settings.

The performance of the calibrated model is measured by per cent of correctly predicted samples in the test set

$$\%CP = \frac{\text{nr. of correctly predicted samples}}{\text{total nr. of samples in test set}} \cdot 100. \quad (2)$$

The mean probability of wrong classification

$$MPW = \frac{1}{N} \sum_{i=1}^N P_i(\sim A_i), \quad (3)$$

gives a measure of how well the model was calibrated to the calibration set, as well as an estimate of the probability of predicting a sample in the test set wrongly.

Results and discussion

It is natural to start by examining how well the AHC model calibrated on the calibration set will predict the class of each sample in the test set, when the full spectra are given as input. No pre-selection of wavelengths by the GA is in other words performed. The calibrated AHC model predicted 7 of 9 samples in the test set correctly and the MPW (3) in the test set was 0.34. The two wrongly predicted samples were m43a12 G predicted as SSM, and m51a12 L predicted as SSM. The probabilities of wrong classification were 0.61 and 0.92. The geometrical distance to a centroid of the

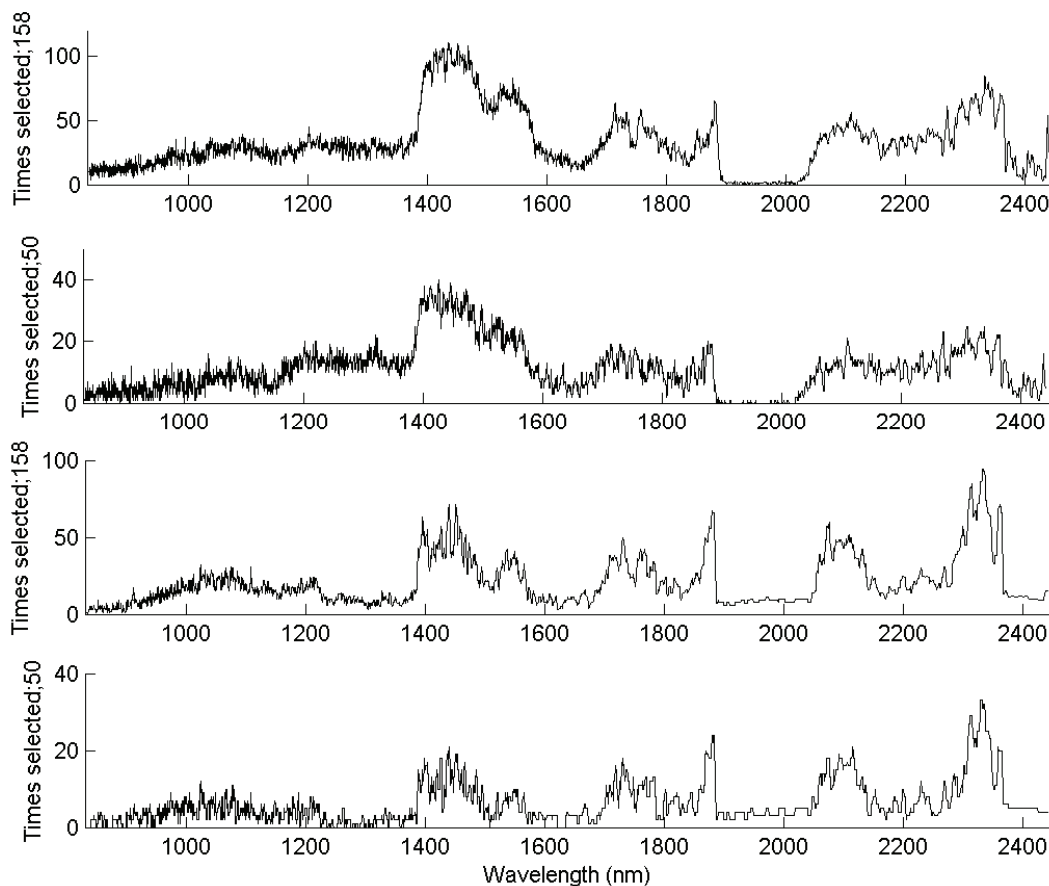


Figure 1: Selection frequencies of wavelengths in the top 158 and 50 solution candidates of the test set by decreasing %CP (top) and the top 158 and 50 solution candidates of the calibration set based on increasing MPW (bottom).

wrong class was, especially in the later case, much shorter, henceforth the wrong classification.

The probability to by chance get 7 or more correctly predicted, if we assume that the four classes are equally probable, is only 0.001. Thus 7 of 9 correct is significantly better than what can be expected if no signal is present. The presence of a signal is by our definition increasing the probability of correct prediction from 0.25, which is the probability to predict a sample correctly by pure chance.

The *MPW* in the calibration set was 0.55, i.e. it is more likely to predict a sample wrongly than correctly. This contradicts the result of the test set; hence the calibration is examined closer. A key element of the AHC model is clustering. The full spectra model, however, have 27 clusters. This is so high, considering that the calibration set contains only 28 samples, that we started by verified that the clustering algorithm works properly on another dataset. Only one cluster of two samples is formed and the rest of the clusters contain just one sample. This can be explained by the fact that the clustering is stopped immediately when one sample of a different class is about to be added to a cluster. Obviously the calibration set contains two samples of different class that are the second two closest samples, so that the clustering is stopped after the formation of the first cluster with two samples.

The clustering is stopped early in order to make sure that no sample that is close to the border of two clusters of different class can end up being wrongly classified, i.e. closer to the centroid of a cluster of different class than its own centroid. The centroid of a cluster moves a bit each time a sample is added to the cluster and we did not want to add a condition, checking that none of the old samples of any cluster is becoming closer to another centroid, at each sample addition. We tried a model where the clustering was forced to continue, but it performed poorly (results not shown).

A consequence of the low degree of clustering is that, in most cases, when a sample is removed from the calibration set for calculation of its probability of wrong classification, then no sample remains to define the cluster it belongs to. Now given that samples of different class that are close to each other exist, the probability of wrong classification will become high. If we had a bigger calibration set, then it is likely that more clusters would contain more than one sample and the present problem would disappear. The high *MPW* in the calibration set should therefore at least partly be interpreted as an artefact of having a small dataset.

Let us now get to the point by employing the GA for pre-selection of wavelengths. Because GAs are stochastic optimisation algorithms, the solutions of two runs look different. We have therefore generated many

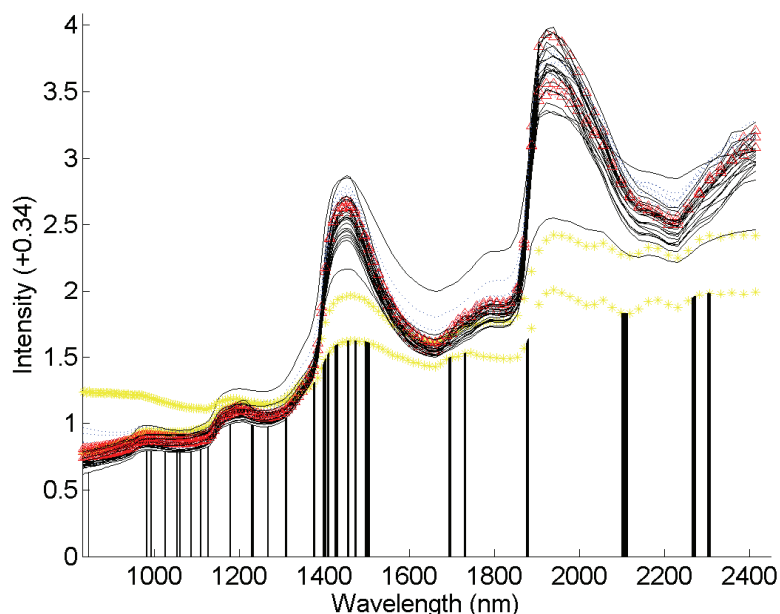


Figure 2: Solution candidate with the 3rd lowest *MPW* in the calibration set and 9/9 correct in the test set. The spectrum of each sample is marked according to class, black solid line = superficial spreading melanoma (SSM), yellow star = nodular malignant melanoma (N), red triangle = lentigo maligna melanoma (L) and blue dotted line = benign naevi (G) or reference (R).

solution candidates by running the same algorithm 1917 times for different GA parameters, in 1580 of these 7 or more were correctly classified in the test set.

Wavelength selection has clearly taken place, as indicated by the differences in selection frequency of the wavelength variables (Figure 1). Some wavelength variables are selected much more frequently than the others, which indicates that a signal is present or noise that can be used for calibration. The former is more likely, since we have no indication that the measurement noise would be anything but random.

We have arranged the top 158 and 50 solution candidates of the test set by decreasing %*CP* (2) in the upper half of Figure 1. The lower half contains the top 158 and 50 solution candidates of the calibration set based on increasing *MPW*. The frequency of selection look a little different for the wavelength variables of the two sets, as can be expected, assuming that the two sets contain different noise. Nonetheless, the frequency functions remain similar. This indicates that the AHC model calibrated on the calibration set can be used to successfully predict the class of the samples in the independent test set.

The two major intervals of selected wavelengths appear to be 1400 to 1470 nm and 2300 to 2350 nm. Minor peaks exist around 1550, 1730, 1760, 1880 and 2100 nm. It is interesting to note that the first major interval falls within the first water peak 1400 to 1500 nm, while wavelengths of the second water peak 1900 to 2000 nm clearly should not be selected. This might be in accordance with the finding of Gniadecka et al. [7] that the water structure is different in malignant tumours.

The solution candidate with the lowest *MPW* in the calibration set predicts 8 of 9 samples in the test set correctly. The candidate with third lowest *MPW* in the calibration set predicts all test set samples correctly. Visual inspection of the selected wavelength variables (Figure 2), however, reveals many short intervals of selected wavelengths, which might indicate over-fitting, i.e. inclusion of noise unique to our dataset.

We attempted to construct a final solution based on the intervals that are indicated as important in Figure 1. The starting point was a chromosome with only wavelengths included in more than 25% of the top 158 solution candidates with lowest *MPW* in the calibration set. This chromosome only needed minor smoothing of the selected intervals. The wavelength

ranges that finally were included was 1390–1470, 1540–1560, 1720–1735, 1755–1770, 1865–1890, 2070–2120 and 2290–2350 nm. This wavelength selection gave 7 of 9 correctly classified in the test set. As good as the full spectrum model, but this time m43a12 G was predicted as N and m52a22 SSM as G. The first error was the same sample, but it was assigned a different class. Obviously this sample is close to the border of three clusters. The *MPW* of the test set was now 0.36, slightly worse, but the probabilities of wrong prediction had decreased to 0.60 and 0.59 respectively. This indicates that inclusion of only the interesting wavelength intervals give a slightly better prediction than the full spectrum model.

A closer study of the wavelengths selected in the five candidates with lowest *MPW* in the calibration set reveals that they have quite few wavelengths in common, even though the intervals from which the wavelengths are selected mainly stays the same. This is explained by the fact that the characteristic spectral peaks of molecules are broad in the NIR range, hence many wavelength variables correspond to the same constituent. The predictive ability of the five candidates varies between 7 and all correct. Both the selection of different wavelengths and the variation in predictive ability are indications of influence by noise. The true predictive ability of the method presented here or the correct wavelength selection can therefore not be determined based on the dataset that we have used. A larger dataset is needed.

Conclusions

This work shows that NIR spectra contain data that can be used for classification of atypical naevi into different established clinical classes. It is also shown that wavelength selection by genetic algorithms can be used to find the interesting spectral ranges and also improve the classification. The dataset used in this study is, however, too small and the traditional clinical diagnosis contains too much uncertainty for precise determination of which wavelengths should be selected.

It is advisable to create a larger dataset, where the diagnosis, in case of the ABCD rule, is done so that the classification uncertainty of each sample is estimated. The uncertainty estimates can then be used to improve the method by fuzzy programming. To actually decrease uncertainty by performing a biopsy after lesion removal would be even better. Then the statistical analysis of the results should be improved by selecting the independent calibration and test set on random from the total dataset. Different pre-processing methods should be evaluated. It might be possible to improve the model by using the minimum geometrical distance to halfplanes defined by the points of the clusters instead of centroids. Every intersection of these halfplanes forms a convex set representing a cluster. Different weight functions for summation of the probabilities of wrong classifications also need to be examined.

The next planned step is to collect a large reference database of NIR spectra recorded of normal healthy skin. This will be used for a background elimination study.

Acknowledgements

We are grateful to Med. Dr. Anders Nilson and colleges for selecting, appointing and examining the patients. Unizon NIRCe is gratefully acknowledged for funding this research. J. Nyström, I. Bodén and B. Lindholm-Sethson would further like to thank the European Union Structure Foundation for financial support.

References

- [1] GOLDSTEIN B. and GOLDSTEIN A. (2001): 'Diagnosis and management of malignant melanoma', *American Family Physician*, **63**, pp.1359–68, 1374
- [2] FERANT A., JOHNSON J., DEMASTES-SHERIDAN C., and CAFFREY T. (2000): 'Early detection and treatment of skin cancer', *American Family Physician*, **62**, pp. 357–6, 381–2
- [3] ÅBERG P. (2004): 'Skin cancer as seen by electrical impedance', PhD Thesis, Karolinska Institutet, Stockholm
- [4] BONO A., TOMATIS S., BARTOLI C., CASCINELLI N., CLEMENTE C., CUPERA C. and MARCHESINI R. (1996): 'The invisible colors of melanoma. A telephotospectrometric diagnostic approach on pigmented skin lesions', *European Journal of Cancer*, **332A**, pp. 727-729

- [5] ELBAUM M., KOPF A., RABINOVITZ H., LANGLEY R., KAMINO H., MIHRM M., SCHER A., PECK G., BOGDAN A., GUTKOWICZ-KRUSIN D., GREENEBAUM M., KEEM S., OLIVIERO M. and WANG S. (2001): 'Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: a feasibility study', *Journal of the American Academy of Dermatology*, **44**, pp. 207-218
- [6] LANGLEY G., RAYADHYAKSHA M., DWYER P., SOBER A., FLOTTE T. and ANDERSON R. (2001): 'Confocal scanning laser microscopy of benign and malignant melanocytic skin lesions *in vivo*', *Journal of the American Academy of Dermatology*, **45**, pp. 365-376
- [7] GNIADOCKA M., NIELSEN O. F. and WULF H. C. (2003): 'Water content and structure in malignant and benign skin tumors', *Journal of Molecular Structure*, **661-662**, pp. 405-410
- [8] HIROSAWA N., SAKAMOTO Y., KATAYAMA H., TONOOKA S. and YANO K. (2002): 'In vivo investigation of progressive alterations in rat mammary gland tumors by near-infrared spectroscopy' *Analytical Biochemistry*, **305**(2), pp. 156-165
- [9] MENZIES, S. and STOLZ W. (2004): 'Surface microscopy features of melanoma', in THOMPSON J.F., MORTON D.L. and KROON B.B.R. (Ed): 'Textbook of Melanoma', (Martin Dunitz, United Kingdom), pp. 227
- [10] GELADI P. & DÅBAKK E. (1999): 'Computational methods of analysis and chemometrics in near-infrared spectroscopy', in LINDON J., TRANTER G. and HOLMES J. (Ed.): 'Encyclopedia of Spectroscopy and Spectrometry', (Academic Press, London), 1999, pp. 343-349
- [11] GELADI P. (2002): 'Some recent trends in the calibration literature', *Chemometrics and Intelligent Laboratory Systems*, **60**, pp. 211-224
- [12] JOUAN-RIMBAUD D., MASSART D-L., LEARDI R. and DE NOORD O.E. (1995). 'Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration'. *Analytical Chemistry*, **67**, pp. 4295-4301
- [13] ABRAHAMSSON C., JOHANSSON J., SPARÉN A. and LINDGREN F. (2003): 'Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets', *Chem. Intel. Lab. Sys.*, **69**, pp. 3-12
- [14] NORDLING T.E.M., KOLJONEN J., ALANDER J.T. and GELADI P. (2004): 'Genetic algorithms as a tool for wavelength selection', Proc. of the 11th Finnish Artificial Intelligence Conference (STeP 2004), Vantaa, Finland, 2004, Vol. 3, pp. 99-113

For more details on this study see
<ftp://ftp.uwasa.fi/cs/report05-4>.