DIPLOMARBEIT

Statistical Preliminaries of Using Classifiers to Detect Malignant Melanoma in Infrared Hyperspectral Images.

ausgeführt am

Institut für Chemische Technologien und Analytik
der Technischen Universität Wien

Institute of Chemical Technologies and Analytics
Vienna University of Technology

unter der Anleitung von
Ao.Univ.Prof. Mag.rer.nat. Dr.rer.nat. Johann Lohninger
durch

Elisabeth Renner

e1026129@student.tuwien.ac.at

October 13th, 2016

Datum

Unterschrift

Acknowledgements

First I want to express my gratitude to my supervisor Hans Lohninger, who not only immediately managed to fascinate me for the field of chemometrics but also gets my deepest respect for his sincerity, prudence and thoughtfulness. This work would not have been possible without our long discussions and his support.

Gratitude also goes to Dozent Dr. Christine Hafner. Her patient explanations of dermatology and pathology have been of great value.

I'm thankful to Jonathan Gantner, Benedikt Steindl and Elisabeth Wetzer for interesting and fun discussions on classification algorithms as well as Karin Wieland for introducing me to the world of FTIR microscopy. It goes without saying that I am extremely grateful to the same people - including Stefan Tauber and Christoph Gasser - for the outstanding time I've had and the close friendships that have developed during this thesis.

This brings me to expressing my gratitude to all my friends - especially all the great people I was privileged to study with.

Last, but definitely not least, I want to say thank you to my parents, Sabine and Albin, and my brother Andreas, who have always shown me the world from a special point of view and encouraged me in my passion for science with great devotion.

Thank you.

Kurzfassung

Künstliche Datensätze wurden erzeugt um die Auswirkung von unterschiedlichen Eigenschaften spektraler Daten auf die Leistungsstärke von Klassifikationsalgorithmen zu untersuchen. Die generierten Datensätze, welche spektrale Daten repräsentieren, basieren auf zwei unterschiedlichen Modellen und werden zudem in ihren Eigenschaften (Rauschen, Größe des Trainingsdatensatzes, Dimension des Datenraums und Separierbarkeit der Klassen) variiert. Anschließend wird die Leistung ausgewählter Klassifikationsalgorithmen (k Nearest Neighbor, Partial Least Squares Discriminant Analysis, Random Forest) für die erstellten Datensätze analysiert. Diese Studie betont den Einfluss hoch dimensionaler Datenräume (große Anzahl an gewählten Variablen) auf die Verteilung der Daten im Merkmalsraum und damit auch auf die Leistung der Klassifikationsalgorithmen.

Die gewonnenen Erkenntnisse werden angewandt um mittels IR - Imaging FFPE-Gewebeschnitte zu klassifizieren und malignes Melanom zu erkennen. Verschiedene Transformationen der spektralen Daten aus dem Fingerprint-Bereich werden verwendet um Deskriptoren zu erstellen, welche ein hohes Ausmaß an chemischer Information beinhalten. Mittels der definierten Deskriptoren wird ein Random Forest Modell erstellt, welches die Klassifikation unterschiedlicher Gewebe (Epidermis, Bindegewebe in verschiedenen Formen, Melanom, Ulzeration) an neuen Gewebeschnitten ermöglicht.

Abstract

The effect of various attributes of spectroscopic data on the performance of selected classification algorithms is investigated by creating artificial datasets. Datasets are generated based on two different models and varied in noise, training data size, dimensionality of the data space and class separability. Subsequently the performance of selected classification algorithms (k Nearest Neighbors, Partial Least Squares Discriminant Analysis, Random Forest) is estimated. This study emphasizes the impact of high dimensions (large number of features) on the data distribution and on the classification performance.

The acquired knowledge is applied when classifying tissue types and detecting malignant melanoma in infrared hyperspectral images of paraffin embedded skin tissue sections. Based on various transformations of the spectra in the fingerprint range, selected spectral attributes are identified to encode maximum chemical information. Those features are used for building Random Forest classifiers to enable tissue identification (epidermis, different kinds of connective tissue, malignant melanoma, ulceration) of new samples.

Contents

Introduction

During the last two decades Spectral Histopathology (SHP) has emerged as a promising technique for analyzing biological samples. The prospect of using imaging systems based on infrared (IR) and raman spectroscopy in daily clinical practice has motivated various groups world wide to develop new technologies and algorithms.

Classical histopathology makes use of different dyes to retrieve information based on the obtained stained tissue sections. The main idea of combining chemometrics and IR imaging in SHP is to develop reliable methods for gathering chemical information of the underlying tissue section, assigning labels to tissues and cells (e.g. cancerous vs. non cancerous tissue), identifying biomolecular processes and much more.

Nowadays, big emphasis of SHP studies is put on creating "digital stains" of tissue sections based on IR spectra by the use of classification algorithms. In this work, such a classification is performed on tissue sections of malignant melanoma lesions.

However, it has been noticed that it is worth putting thought into the data distribution of spectroscopic data and its effect on the performance of classification algorithms. Various characteristics such as noise level, dimensionality of the data set, amount of available training data and attributes regarding class separability affect the obtained results of a classification algorithm.

Thus prior to classifying the melanoma sections, artificial data models representing spectroscopic data of various characteristics are generated, classified by selected classification algorithms (Random Forest, k-Nearest Neighbor and Partial Least Squares Discriminant Analysis) and the results analyzed in Part I of this thesis.

The concept of classifiers is explained in Ch. 1. In Ch. 2 the generated data models are outlined and the obtained results are listed and discussed in Ch. 3.

In Part II the gathered knowledge is applied to classify different tissue types in skin tissue sections based on IR images. FTIR microscopy is used to acquire infrared hyperspectral images from paraffin embedded and formalin fixed tissue sections of human skins. The obtained hyperspectral images are pre-processed to account for scattering effects and paraffin contribution as well as variations between the

samples. Subsequently a Random Forest classifier is used to classify the tissue sections and create "digital stains".

Ch. 4 provides the required background on the histopathology of maligant melanoma as well as molecular vibrations and infrared imaging. The methods of data acquisition and processing are explained in Ch. 5. In Ch. 6 the obtained results are listed and discussed.

Part I

Performance Evaluation of Selected Classifiers Using Artificial Spectroscopic Datasets

Chapter 1

Theoretical Background on Classification Algorithms

1.1 Introduction to Classification Algorithms

Classification algorithms are automated procedures which identify patterns and regularities in data sets in order to assign a property (class) to each sample. The patterns are first identified using data with known classes and can then be used to make predictions for future data with unknown class label. Today, classifiers are applied for all different kinds of data[1].

Chemometrics uses multivariate classification algorithms to extract information from analytical chemistry data[1]. In vibrational spectroscopy, classifiers are e.g. used to automatically identify chemical structures by their spectrum. The following description of classifiers mostly states spectroscopic imaging data sets as examples. However, it is worth noting that the same concepts can be used for any kind of data.

In spectroscopic imaging a spectrum is acquired for each pixel. The aim of using classification algorithms in hyperspectral image analysis is to assign initially unknown labels, also called *classes*, to each pixel (e.g. type of tissue, type of molecule etc.), based on selected characteristics of their spectra. Those characteristics are called *features* or spectral descriptors and are explained further in Sec. 1.2.

[1] The concepts of classification algorithms can be understood by an easy example. Let's assume that a vet, who examines birds, dogs and mice, created a database containing the characteristics of the examined animals (so called *features*: height, weight, fur color), but failed to record which animal (thus, *data class*) the properties belonged to. He has only recently started to record the animal type. Thus, he knows the class of the last 50 entries of the dataset, but also wants to assign a class label to the first 500 entries automatically. A classifier manages to analyze the data structure of the 50 entries with known class, recognizes the data structure and uses this structure to assign a class to the 500 unlabeled entries.

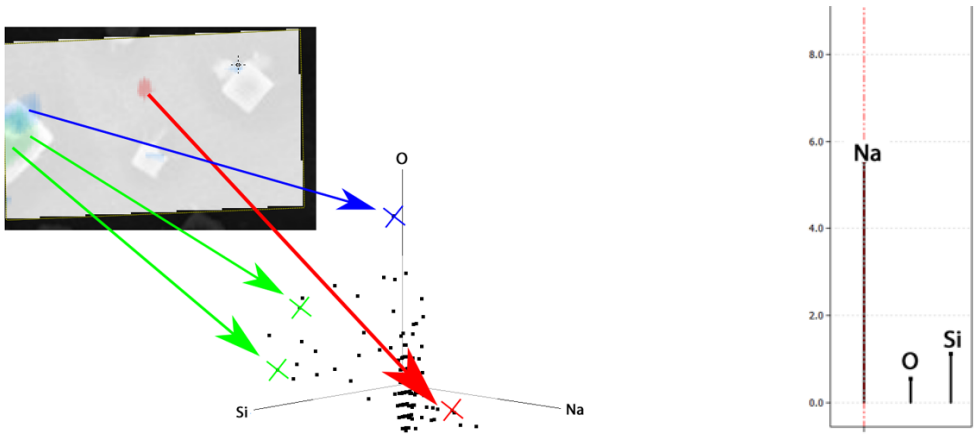


Figure 1.1: Visualization of image pixels as data points in a 3-dimensional data space (3 spectral descriptors) for an EDX(Energy Dispersive X-Ray) spectrum of a sample of particulate matter. Data are retrieved from [2].

There are mainly two groups of algorithms to be distinguished: *classification* and *clustering* methods. Classification algorithms include a set of previously labeled training data. The classes of the training set are also called *ground truths*. The algorithm identifies a pattern in the labeled training set and based on this knowledge a classification model is created. This model can be applied to classify data with unknown labels.

Clustering algorithms find a pattern in the dataset without previous training. Thus, no labeled training data are required. Examples for clustering methods are *k-means Clustering, Hierarchical Clustering* and *Fuzzy Clustering*. However, in this thesis, only classification methods are featured.

To correctly apply classifiers in hyperspectral imaging it is important to emphasize the context of pixel and spectra in the mathematical model. Each feature of a spectrum (e.g. wavenumber in vibrational/ UV-Vis spectroscopy, m/Z in mass spectrometry) is one variable of the underlying mathematical model and is thus equal to one dimension (one axis) of the d-dimensional data space. The data space is also referred to as *feature space*.

Each of the N pixels represents one measurement with respective values of the d features. It thus corresponds to one data point $\mathbf{x}_i$, $i \epsilon [1, N]$ (generally called *sample*) in this d-dimensional data space and can be represented by the *feature vector*

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, ...x_{i,d})\,; \qquad dim(\mathbf{x}_i) = d \times 1 \tag{1.1}$$

where each coordinate $x_{i,j}$ with $j \epsilon [1, d]$ corresponds to the j-th spectral feature of i-th pixel. The classes of the pixels (samples) are described by the *response vector* $\mathbf{y}_i$.

$$\mathbf{y}_i = (y_{i,1}, y_{i,2}, ...y_{i,r})\,; \qquad dim(\mathbf{y}_i) = r \times 1 \tag{1.2}$$

There can be more than one different assignment ($r < 1$) for each pixel. Details on the possible characteristics of the response matrix Y will be explained in more detail in Sec. 1.1.

The feature vectors of all N pixels can be summarized in the *design matrix* $\mathbf{X}$, where each row corresponds to one pixel and each column to one feature. Similarly, labels are summarized the *response matrix* $\mathbf{Y}$.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \dots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,d} \end{pmatrix} \tag{1.3}$$

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,r} \\ y_{2,1} & y_{2,2} & \dots & y_{2,r} \\ \vdots & \vdots & \dots & \vdots \\ y_{N,1} & y_{N,2} & \dots & y_{N,r} \end{pmatrix} \tag{1.4}$$

To sum up, the task of a classifier is mapping the design matrix $\mathbf{X}$ to a predicted response matrix $\hat{\mathbf{Y}} = \hat{\mathbf{f}}(\mathbf{X})$. The function $\hat{\mathbf{f}}$ is estimated during the training phase.

1.1.1 Characteristics of the Response Matrix

In classification problems y_i is a categorical variable. Integer values (mostly 1 and 0 or 1 and -1) are used to represent the classes during computation. In contrast to this, regression models assign a real number to a feature vector $\mathbf{x}_i$. As this work is focused on classification problems the responses are always considered as categorical.

In many cases the prediction is restricted to two classes, C_1 and C_2. However, there might also be $r > 1$ response variables included in the model. In general, the following classification schemes can be distinguished [3].

Binary tasks require to distinguish two classes, C_1 and C_2. Each input sample is assigned to one of the two classes. This can be represented by only one response variable ($r = 1$) which takes e.g. the values $y_i = 1$ for class 1 (e.g. cancer) and $y_i = -1$ for class 2 (e.g. non cancer)

Multi-class tasks require to distinguish between more than two classes C_1 to C_M, with $M > 2$. There are different approaches to multi-class problems. One of the most widely spread is the *One vs. All (OvA.)* approach, which creates M binary classifiers. Each of the M classifiers separates one class from all the other classes. The prediction with the highest confidence score is chosen as the final class assignment.

In contrary, the *One vs. One (OvO)* approach creates $\frac{M \cdot (M-1)}{2}$ classifiers, each describing a binary problem between two classes. The class with the most positive predictions is assigned as the final label.

The classification of polymers (e.g. polyethylen (PE), polypropylen (PP) and polystrol (PS)) based on vibrational spectroscopy is consulted as a brief example. The OvA approach creates the classifiers

- PE versus PP and PS,
- PP versus PE and PS,
- PS versus PE and PP

while the OvO approach creates

- PE versus PP,
- PE versus PS,
- PP versus PS.

Hierachial tasks also require to assign a sample to one of C_M classes with $M > 2$. In this case, the task is solved by a succession of binary classifiers. Each classifier divides the previously separated data into further subsets. The result is a tree with sub- and superclasses in a fixed hierarchy.

Multi-labeled tasks , also called multi-topic tasks, require to assign more than one label to each sample. Multi-labeled tasks are not considered in this thesis.

Some classification algorithms - such as Random Forests - solve multi-class (hierarchial) tasks inherently and a system of several binary classifiers is not required. However, many algorithms (e.g. PLS-DA) depend on such a system of binary

classifications, which can result in ambiguous classifications if the binary classifiers are not tuned well. This led to the development of several more complicated classification schemes for multi-class tasks. Nevertheless, studies show that when using well-tuned classifiers as binary models simple OvA schemes provide results comparable to any of the more complex approaches [4].

1.2 Spectral Descriptors

As already mentioned, in spectroscopy the features of the classification model (variables/columns in the design matrix) are derived from the measured spectra. There are various ways to define those the features and thus map chemical information.

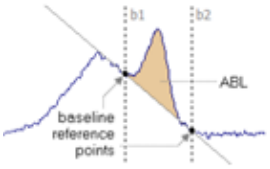
Many researchers simply take all or selected raw intensities (e.g. every 10th wavenumber in an IR spectrum) as features. However, raw intensity data are not very selective in most cases. They are prone to noise and may, depending on the analytical method, be highly correlated if acquired with small spacing.

Furthermore, many spectra cover a large range on the x-axis (wavenumber in IR-spectroscopy, m/Z in mass spectrometry) leading to a high number of features. The phenomenon, that the resulting high dimensional dataspace can be problematic for many mathematical models is described in Sec. 1.5[5].

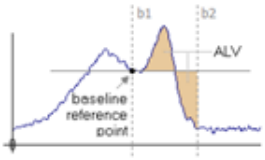
Thus, is has been suggested to derive variables from the spectra, which encode chemical information and use them for further analysis[6]. Using these so called *spectral descriptors* reduces the dimension of data space and improves its structure, because only variables that contain information relevant to the specific problem are used.

The classification of skin tissue sections based on IR images in part Part II is based on such spectral descriptors. Thus, in the following thesis the term *spectral descriptor* will be generally used to define the variables of a classification model.

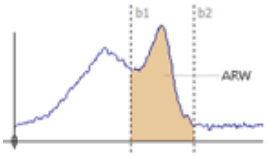
Descriptors which have proven to be useful are illustrated and described in Fig. 1.2. The abbreviations which are here introduced will be used in Part II.



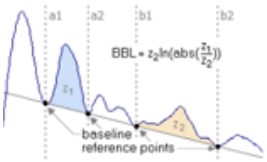
(a) Integral with baseline correction (ABL)



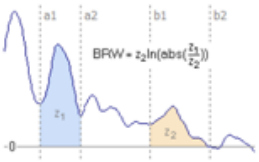
(b) Integral with level baseline correction (ALV)



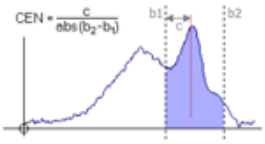
(c) Integral without baseline correction (ARW)



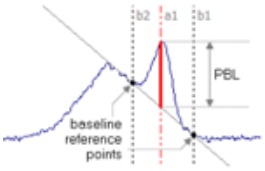
(d) Logarithmic ratio of integrals with baseline correction (ABL)



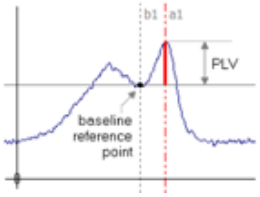
(e) Logarithmic ratio of integrals without baseline correction (BRW)



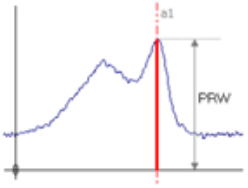
(f) Location of peak center (CEN)



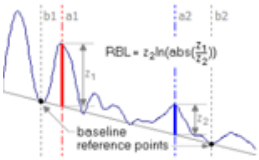
(g) Intensity corrected by baseline (PBL)



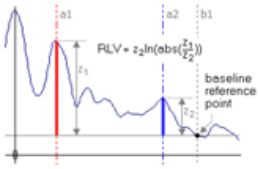
(h) Intensity corrected by level baseline (PLV)



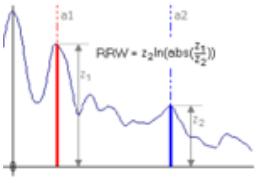
(i) Raw intensity (PRW)



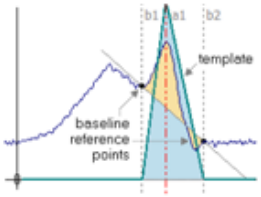
(j) Logarithmic ratio of baseline corrected intensities (ABL)



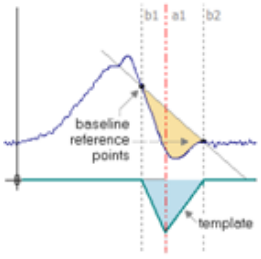
(k) Logarithmic ratio of intensities with level baseline correction (RLV)



(l) Logarithmic ratio of intensities without baseline correction (RRW)



(m) Correlation of the peak area to the template triangle peak (TC)



(n) Correlation of the peak area to the negative template triangle peak (TCI)

Figure 1.2: Illustration of various spectral descriptors. All illustrations, abbreviations and methods are taken from the hyperspectral imaging software package ImageLab (v.1.98, Epina GmbH, Pressbaum, Austria)[7].

1.3 Model Creation, Selection and Assessment

Supervised classifiers use a set of N labeled training samples to create, optimize and assess the classification model. It is very important to note, that the data with which the performance of the model is estimated must differ from the data with which the model is trained. Thus, the N labeled samples are divided into two groups, *training* and *test* data.

$$N = N_{train} + N_{test} \tag{1.5}$$

The following nomenclature is used for training and test datasets.

Training data	$\mathbf{x}_i$	with the ground truth	$y_i,\ i \epsilon [1, N_{train}]$
Test data	$\tilde{\mathbf{x}}_i$	with the ground truth	$\tilde{y}_i,\ i \epsilon [1, N_{test}]$

Furthermore, each classification algorithm depends on various parameters, which have to be optimized before the final classifier can be applied to new datasets. The setting of those parameters for achieving optimized classification performance depends on the individual problem. Examples for such parameters are the number of neighbors for *k-Nearest Neighbor Classifiers* or the number of trees for *Random Forests* (compare Sec. 1.4).

Model selection is the process of optimizing those parameters. It is important to consider the balancing act between sufficiently adapting the model to the training dataset and generalizing the model for application on an independent test set (Sec. 1.3.1). Therefore, optimization of the parameters is controlled by *cross-validation*, which is based on repeatedly taking out validation sets of the training set (Sec. 1.3.3).

After creating the model and optimizing the parameters the final model is applied to an independent test set (with known labels) for estimating the models performance. The test set must not be included in the previous training and validation process (*model assessment*).

Subsequently, the created and assessed model can be applied to new datasets with unknown labels to classify them. The classifier performance, which has been estimated previously using the labeled test set, is important for stating the statistical reliability of the results.

The full process of model creation and assessment is illustrated in Fig. 1.3.

No general rule can be applied for the splitting ratio of test/training and training/validation sets. If N_{test} is too small, the performance measures might be sta-

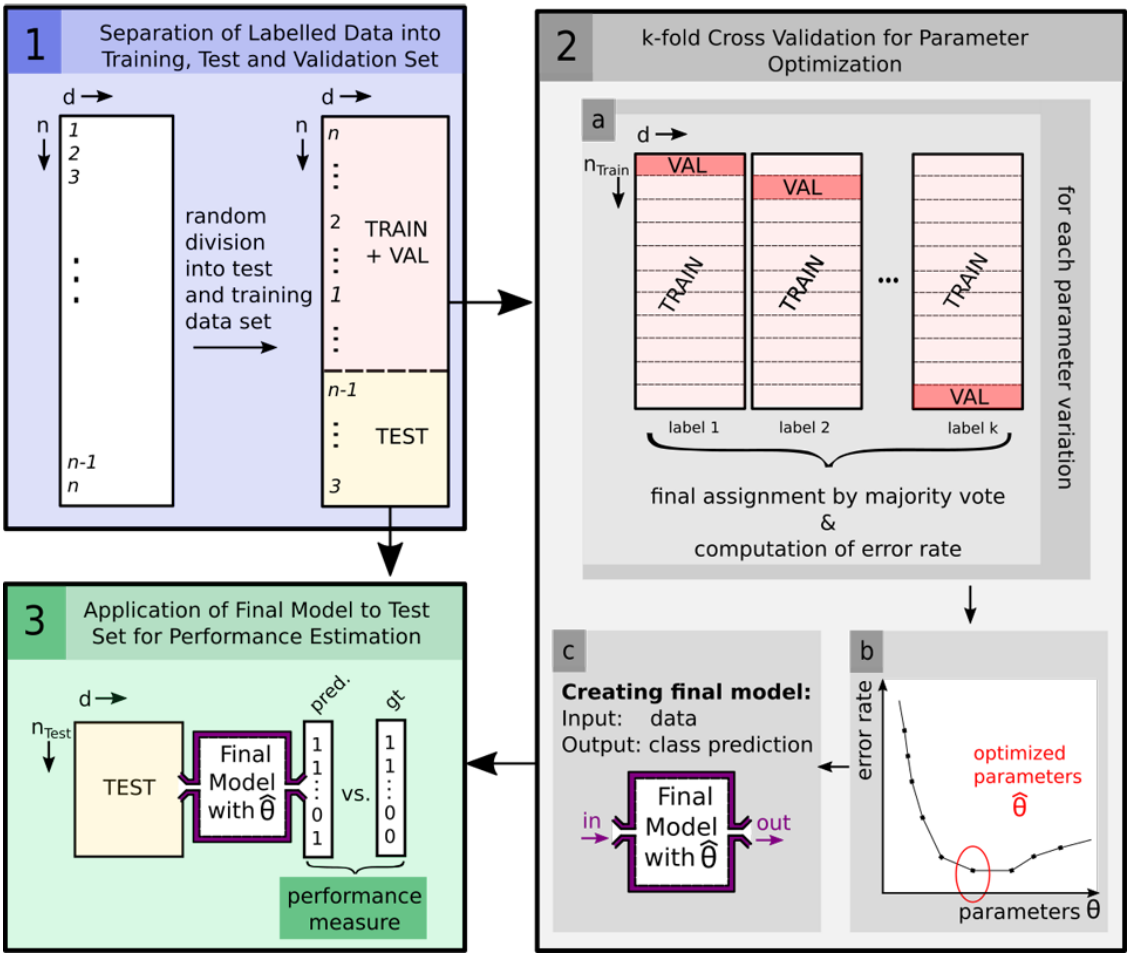


Figure 1.3: Overview of model selection and assessment. 1) The labeled dataset is randomly divided into test and training data. 2) The parameters are optimized and the final model is selected via cross validation. Here, 10-fold cross validation is illustrated as an example. Compare Sec. 1.3.3 for further understanding the cross validation process in box (2a). 3) The final model is assessed by classifying the independent test dataset and comparing the predicted (pred.) labels to the ground truth (gt.). Test data must not have been used during training and model selection.

tistically unstable. If N_{train} is too small the bias of the model might increase. The more difficult it is to recognize the characteristics of the decision boundary in the data space (e.g high noise, high dimensions), the larger N_{Train} has to be.

1.3.1 Model Complexity, Bias and Variance

As mentioned before, each classifier depends on specific parameters which determine the characteristics - the complexity - of the classification model. On the one hand, the model has to be complex enough to correctly map certain characteristics of the data distribution in space (e.g. non linear decision boundaries). On the other hand, in order to ensure possible generalization, the model must not be adapted too strongly to the specific training set.

The quantities *bias* and *variance* are used to exactly describe this balancing act. The *bias* describes the systematic difference of the predicted classes $\hat{\mathbf{f}}(\mathbf{X})$ to the ground truths $f(\mathbf{X})$. [8, 9]

$$Bias[\hat{\mathbf{f}}(\mathbf{X})] = E[\hat{\mathbf{f}}(\mathbf{X}) - f(\mathbf{X})] \tag{1.6}$$

Thus, the bias increases with decreasing model complexity (e.g. k very large for kNN or a linear decision boundary for a non linear problem) and the model fails to correctly map the data distribution.

The *variance* describes the deviation of the prediction on the test set over different training sets.

$$Var[\hat{\mathbf{f}}(\mathbf{X})] = E[\hat{\mathbf{f}}(\mathbf{X})^2] - E[\hat{\mathbf{f}}(\mathbf{X})]^2 \tag{1.7}$$

Thus, the variance increases with increasing model complexity (e.g. k = 1 for kNN or a high-order polynomial decision boundary), as the model tends to adapt very well to the specific set of training data. Training it with a different training data set will result in a different model and finally in different predictions on the test set.

To sum up, the aim is to create a model which is simple enough to predict the labels of an independent test set as reliable as possible (generalization ability, low variance), while being complex enough to correctly recognize the data distributions (low bias).

This compromise is called *variance-bias trade off* and is illustrated in Fig. 1.4. Models are created for different parameter settings, resulting in different model complexities. The performance (in Fig. 1.4: prediction error) estimated on both, an

independent test set and on the training set, is plotted over the model complexity. One can see, that if the training set would be used for estimation of the model performance the model would seem to improve with higher complexity. Computing the performance on an individual test set, though, results in a minimum of the prediction error curve for certain parameter settings. Estimating this minimum by optimizing the complexity parameters is the fundamental task of cross validation (Sec. 1.3.3).

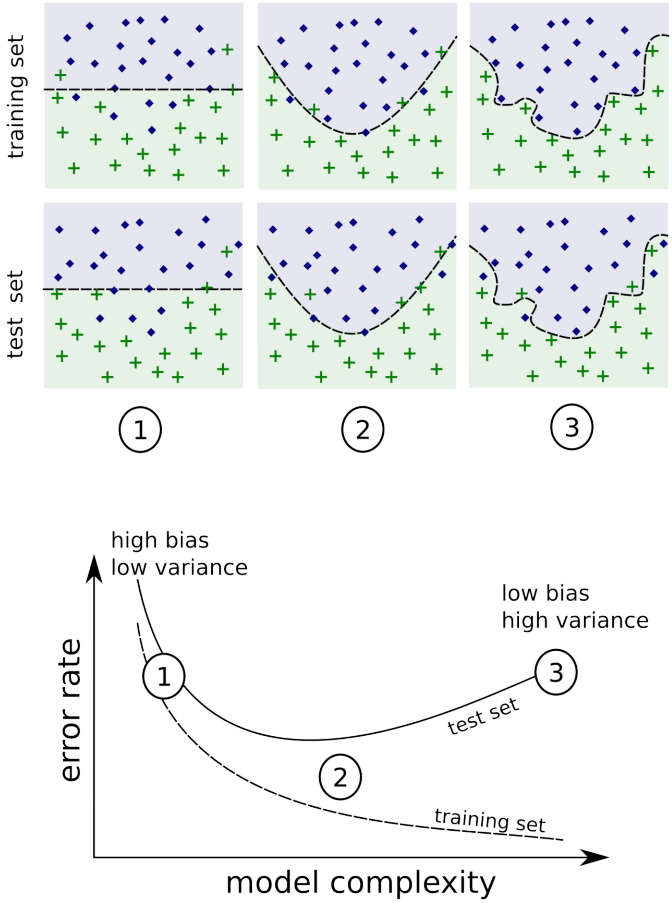


Figure 1.4: Illustration of the variance-bias trade off for different model complexities. The blue (red) data points correspond to class 0 (1) and the blue (red) region is the region in which test data is assigned to class 0 (1). Top row: Scatter plots of training data with estimated decision boundaries for different model complexities. Center row: scatter plots of test data and previously estimated decision boundaries; bottom row: prediction error of applying the classifier on the test set (dashed line; middle row) and the training set (solid line; top row)

1.3.2 Performance Metrics

There are various ways to estimate the performance of a classifier, but all of them are based on the confusion matrix. This matrix lists the correctly and wrongly classified samples of each class separately and is illustrated in Fig. 1.5.

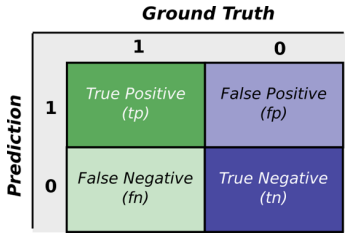


Figure 1.5: Confusion matrix for a binary classifier with $N_c = 2$ classes. To generalize, for any N_c the dimension of the confusion matrix is $(N_c \times N_c)$

In the following thesis, the abbreviations TP, TN, FN, and FP are used for the corresponding counts in the confusion matrix. Different metrics can be computed using those values. As there is no universal performance metric which includes all information in one number the selection of a metric or a combination of metrics, respectively, depends on the individual case. Although, the idea behind the metrics for binary or multi-class problems is similar, the specific execution differs slightly.

The most common metrics are listed for binary and multi-class tasks in Tbl. 1.1 and Tbl. 1.2.

Table 1.1: Metrics for binary classification [3]

Metric	Abbr.	Formula	Description
Accuracy	ACC	$\frac{TP+TN}{TP+FP+TN+FN}$	Overall effectiveness of a classifier, considering both classes
Error Rate	ERR	$\frac{FP+FN}{TP+FP+TN+FN}$	Overall error of a classifier, considering both classes
Sensitivity (Recall, True Positive Rate)	TPR	$\frac{TP}{TP+FN}$	Ability to correctly identify positive samples
Specificity (True Negative Rate)	TNR	$\frac{TN}{TN+FP}$	Ability to correctly identify negative samples
Positive Predictive Value (Precision)	PPV	$\frac{TP}{TP+FP}$	Ability to only classify positive samples as positive
Negative Predicitve Value	NPV	$\frac{TN}{TN+FN}$	Ability to only classify negative samples as negative

The metrics used to assess multi-class methods are based on measures for binary tasks. The confusion matrix is created for each binary classifier, with the entries

TP_i, FP_i, TN_i and FN_i. The measures are computed in analogy to binary tasks by *micro-* or *macro-averaging* (compare Tbl. 1.2).

Table 1.2: Metrics for multi-class tasks: TP_i, TN_i, FP_i and FN_i correspond to the entries of the confusion matrix of the i-th classifier (in case of "One vs. All" the classifier dividing the i-th class from the rest). μ and M indicate micro- and macro averaging. [3]

Metric	Abbr.	Formula	Description
Average Accuracy	ACC	$\frac{\sum_{c=1}^{N_c} \frac{TP_j+TN_j}{TP_j+FP_j+TN_j+FN_j}}{N_c}$	Mean accuracy over all classes
Mean Error	ACC	$\frac{\sum_{c=1}^{N_c} \frac{FP_j+FN_j}{TP_j+FP_j+TN_j+FN_j}}{N_c}$	Mean error over all classes
Micro-Recall	TPR_μ	$\frac{\sum_{c=1}^{N_c} TP_j}{\sum_{c=1}^{N_c} TP_j+FN_j}$	Micro averaging: Sum of confusion matrix counts over all decisions with subsequent rationing
Micro-Precision	PPV_μ	$\frac{\sum_{c=1}^{N_c} TP_j}{\sum_{c=1}^{N_c} TP_j+FP_j}$	
Micro-Specificity	TNR_μ	$\frac{\sum_{c=1}^{N_c} TN_j}{\sum_{c=1}^{N_c} TN_j+FP_j}$	
Micro-Neg. Pred. Value	NPV_μ	$\frac{\sum_{c=1}^{N_c} TN_j}{\sum_{c=1}^{N_c} TN_j+FN_j}$	
Macro-Recall	TPR_M	$\frac{\sum_{c=1}^{N_c} \frac{TP_j}{TP_j+FN_j}}{N_c}$	Macro averaging: Rationing counts of each binary confusion matrix with subsequent averaging
Macro-Precision	PPV_M	$\frac{\sum_{c=1}^{N_c} \frac{TP_j}{TP_j+FP_j}}{N_c}$	
Macro-Specificity	TNR_M	$\frac{\sum_{c=1}^{N_c} \frac{TN_j}{TN_j+FP_j}}{N_c}$	
Macro-Neg. Pred. Value	NPV_M	$\frac{\sum_{c=1}^{N_c} \frac{TN_j}{TN_j+FN_j}}{N_c}$	

The **Receiver Operating Characteristics (ROI)** is a widely used way of illustrating the relationship between sensitivity and specificity and is shown in Fig. 1.6. It is based on the idea, that for each sample the classifier computes a value (comparable to a probability) for which a certain threshold has to be found in order to label the sample. If this threshold is set to its extreme values, all samples will be assigned to one class resulting in $sensitivity = 1$ and $1 - specificity = 0$ (or the other way round). The ROI is obtained by varying this threshold from one extreme to the other.

A metric for the performance of a binary classifier is the area under the resulting ROC (*Area Under the Curve, AUC).* If $AUC = 0.5$ the classifier is not able to separate the two classes. $AUC \to 1$ corresponds to a high performance.

In many works accuracy is used without stating the sensitivity or specificity. It has to be emphasized that this approach can be problematic as the accuracy does not consider whether the correctly classified samples belong to class 1 or 0.

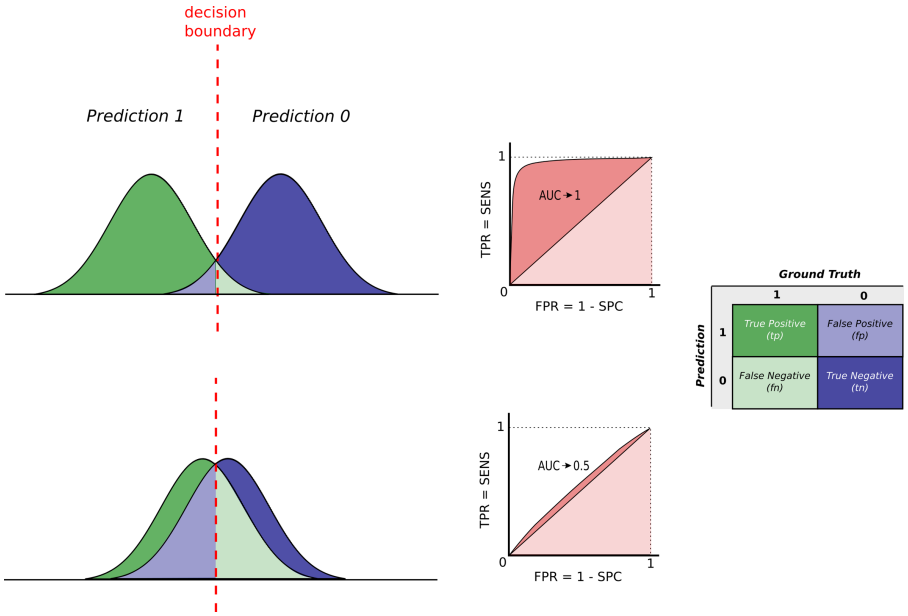


Figure 1.6: Receiver Operating Characteristics with AUC close to 1 (top row, well separable classes) and close to 0.5 (bottom row, non-separable classes).

This problem can be explained well by the example of distinguishing cancerous from non-cancerous tissue. Assuming that there are 50 samples of cancerous tissue and 50 samples of non-cancerous tissue, two different scenarios are pointed out in Tbl. 1.3. Sometimes in medical diagnostics a lower specificity is accepted in order to ensure a high sensitivity. In other words it is preferred to detect all *positives* (cancerous tissue) correctly, while labeling also some non-cancerous tissue as cancerous. However, other applications might require a high specificity. Thus, the accuracy on its own is not sufficient as a performance measure.

1.3.3 Cross Validation (CV)

Cross Validation (CV) is one method to optimize the model parameters for the individual problem. While there exist different variations of CV, in the following k-fold cross validation is explained in more detail.

Basically, during CV models are created and assessed for each potential parameter variation. Finally, the model resulting in the best performance is chosen (*model selection*). CV is still a part of the training phase and the previously assigned test set has thus to be left out completely. Otherwise the same data would be used for model selection and subsequent assessment.

However, evaluating the various models trained during the model validation process

		Ground truth	
		c	n-c
Pred.	c	*40*	*0*
	n-c	*10*	*50*

(a)

		Ground truth	
		c	n-c
Pred.	c	*50*	*10*
	n-c	*0*	*40*

(b)

Table 1.3: Theoretical classification problem to illustrate the insufficiency of the accuracy as the only performance measure. The classifier is supposed to discriminate cancerous from non-cancerous tissue. There are 50 samples of both classes in the test set. a) and b) illustrate the confusion matrix of two different scenarios which result in the same accuracy. However, in case a) 20% of the cancerous samples are classified as non-cancerous while in case b) 20% of the non cancerous tissue is classified as cancerous [a) $ACC = 0.9$, $TPR = 0.8$, $SPC = 1$; b) $ACC = 0.9$, $TPR = 1$, $SPC = 0.8$].

requires again a test set. Thus, the previously assigned training set has again to be subdivided into data with which the models are trained and data to evaluate the models. Latter are referred to as *validation data*. Mostly, the available number of training samples N_{train} is low anyway, wherefore it would not be possible to extract a completely independent validation set, which is sufficiently large to provide statistically stable results while preserving sufficient training data to create a generalized model.

Therefore k-fold CV can be used, which divides the training set into k subsets. Subsequently, for each parameter variation, k models are created, each of which uses (k-1) of those subsets for training and the remaining subset for assessment. Finally, the mean performance of the k models is computed and compared to the mean model performance metrics associated with the other parameters. The model parameters resulting in the best mean performance are chosen. This procedure is illustrated in box (b) of Fig. 1.3.

In many cases the minimum (for some metrics maximum) is located in a flat part of the cross-validation curve. Thus, often the *one-standard error rule* is applied, which selects the least complex model within one standard error of the best model [9]. Selecting a model by means of the one standard error rule is illustrated in Fig. 1.7.

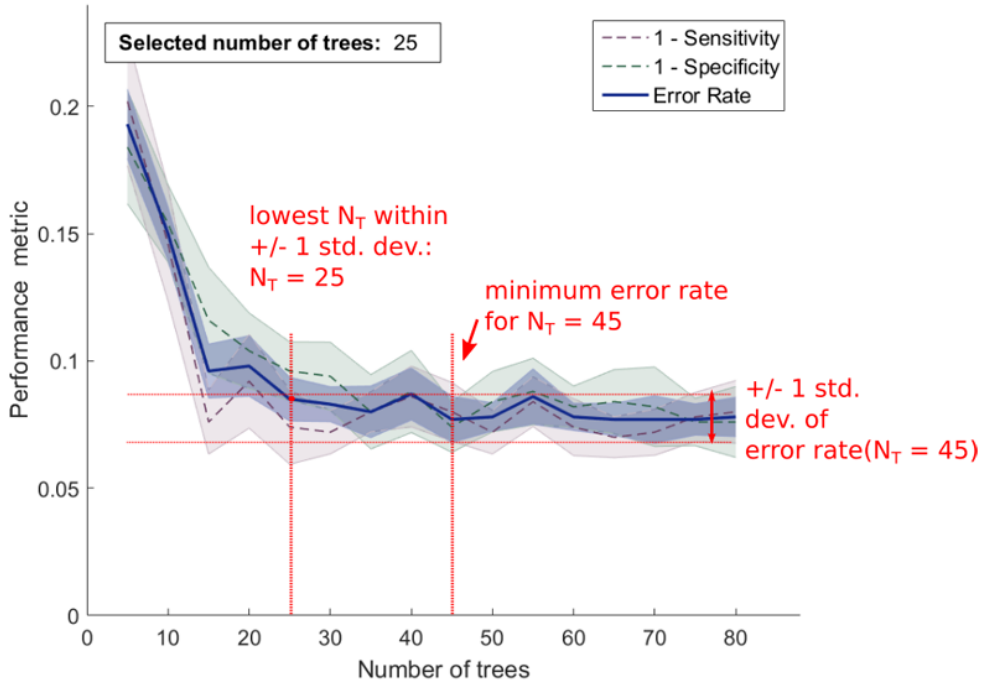


Figure 1.7: Illustration of parameter selection when applying the 1-standard error rule during 10-fold cross validation for a Random Forest Classifier. [Data are taken from a simulated classification problem with $d = 15$ descriptors. $N_{train,pC} = 2500$ for both classes (classes are balanced). The shaded areas show the standard deviation of each parameter setting for prediction error (= 1 - accuracy), 1-specificity and 1-sensitivity.

1.4 Examples of Classification Algorithms

In this thesis the following classification algorithms are considered and thus outlined in the section below:

- *k Nearest Neighbor (kNN)* as it is one of the most trivial and intuitive classifiers and often used as a reference for performance comparisons,
- *Partial Least Squares Discriminant Analysis (PLS-DA)* as it has been a common classifier in chemometrics over the last two decades,
- *Random Forests (RF)*, as they tend to become the current standard for most classification problems.

However, there are many other widely distributed algorithms for supervised learning, such as *Support Vector Machines (SVM), Artificial Neural Networks (ANN)*

or Bayesian Classifiers.

1.4.1 k Nearest Neighbors (kNN)

The k- Nearest Neighbor (kNN) algorithm is seen as the simplest classification algorithm and classifies the test sample according to a majority votw on the class labels of the k nearest training samples.[2]

Because of its simplicity, kNN is often used as a reference for comparison to other algorithms. It is referred to as a *lazy classifier*, as no training phase is required[3]. There are several variations of the algorithm, in the following the simplest implementation is explained in more detail.

Algorithm 1: k- Nearest Neighbors Algorithm

Training phase: In the simplest form it only consists of storing the training data.

Classification Phase: For each test sample $\tilde{\mathbf{x}}_i$ with $i \epsilon [1, N_{test}]$

1. the distance $d_{i,j} = d(\tilde{\mathbf{x}}_i, \mathbf{x}_j)$ between $\tilde{\mathbf{x}}_i$ and each training sample $\mathbf{x}_j$, $j \epsilon [1, N_{train}]$ is computed (mostly the euclidean distance is chosen as a metric, in some cases other metrics proof to be useful[4]),
2. the distances are sorted,
3. the k nearest training samples (smallest distance) are selected,
4. a class is assigned to the test sample $\tilde{\mathbf{x}}_i$ using the ground truths of the k nearest training samples. Mostly a majority vote is used to determine the class of $\tilde{\mathbf{x}}_i$, i.e. the class which most of the k nearest training samples have, is assigned to the test sample $\tilde{\mathbf{x}}_i$.

Parameter selection: The performance strongly depends on the chosen value for k. For a binary classifier k should be odd to avoid a ties assignment. For more than two classes, even an odd k is not sufficient to ensure that there is not more than one class with the majority of the k training samples (e.g. $k = 5$ and 3 classes:

[2] The kNN classifier is well described by the phrase *"If it walks like a duck, quacks like a duck, and looks like a duck, then it probably is a duck"*(Dougherty, Geoff; Pattern Recognition and Classification, An Introduction. page 101)

[3] Classifiers, which require a training phase are called *eager classifiers*.

2 of class 1, 2 of class 2 and 1 of class 3)[10]. The value of k determines the degree of smoothing of the decision boundary. Choosing $k = 1$ results in a low bias but a high variance [9], as the prediction only depends on one training sample close to the test sample. The result is prone to noise.

However, a larger k increases the bias but reduces the variance[11]. Furthermore, a larger k increases the computation time. Thus, as there is no general rule for an optimal k it has to be determined by cross-validation.

Prerequisites: As kNN is based on a distance metric it is essential that each descriptor is standardized ($\mu = 0$ and $\sigma^2 = 1$) to assure equal weight of each descriptor, independent of its average magnitude[9].

Advantages and Disadvantages: As it is based on a distance measure, kNN is prone to high dimensions. As a rule of thumb, at least ten times as many training samples per class than number of descriptors (dimensions) are required to ensure equal performance [10]. However, a high number of training samples also requires high storage and computational capacity. Furthermore kNN is very susceptible to local noise, which often leads to less satisfactory results.

Another disadvantage is that in its simplest implementation the algorithm does not distinguish between the importance of each descriptor for the classification decision. As during pre-processing the descriptors get standardized, a variation in every dimension contributes the same to the final decision, although the decision boundary might only lie in a subspace. Thus, in high dimensional problems kNNs exhibit a high bias. *Adaptive kNNs* address this issue and are briefly discussed later.

On the other hand it excels through its simple implementation and its suitability for parallelization[10] and it is succesful in comparison to other classification algorithms when the decision boundary is irregular[9].

[3] Same common metrics to compute the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ are listed in the following table ($\mathbf{\Sigma}$ is the covariance matrix)[12, 13].

Minkowski distance	$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{s=1}^{d} \lvert \mathbf{x}_{i,s} - \mathbf{x}_{j,s} \rvert^p\right)^{\frac{1}{p}}$
- Euclidean distance	$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{s=1}^{d} \lvert \mathbf{x}_{i,s} - \mathbf{x}_{j,s} \rvert^2\right)^{\frac{1}{2}}$
- Absolute distance (Manhattan distance)	$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{d} \lvert \mathbf{x}_{i,s} - \mathbf{x}_{j,s} \rvert$
Mahalanobis distance	$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$

In 1967 Cover and Hart have shown that for the $k = 1$ classifier, that the asymptotic error rate ($N_{train} \to \infty$) is at most twice the Bayes minimum probability of error. Latter is considered to be the ideal case with perfectly known probability distribution[14].

Adaptions: Over the decades various adaptions of the kNN algorithm have been published. One possibility is the *weighted kNN* (wkNN), which assigns a weight to each of the k Nearest Neighbors to increase the influence of those training samples in the decision, which are closer to the test sample. The weight function is based on the distance $d(\tilde{\mathbf{x}}_i, \mathbf{x}_j)$ (Shepard's method) and can be e.g. an inversion kernel $w_{i,j} \propto \frac{1}{|d(\tilde{\mathbf{x}}_i, \mathbf{x}_j)|}$ or a Gaussian kernel $w_{i,j} \propto \frac{1}{\sqrt{2\pi}} \exp -\frac{d(\tilde{\mathbf{x}}_i, \mathbf{x}_j)^2}{2}$[12].

Other adaptions are the *adaptive kNNs*, which are especially useful for high dimensional problems. One example is the *Discriminative Adaptive Nearest Neighbor (DANN)* algorithm introduced by Hastie and Tibshirani (1996)[15]. It is based on a locally adaptive effective metric for computing neighborhoods, which emphasizes the descriptors with high contribution to the classification outcome. Linear discriminant analysis is used to compute the local metric.

1.4.2 Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA is a supervised learning algorithm based on *Partial Least Square(PLS) Regression* that was first introduced in the 1960's by Herman Wold. Similar to *Principal Component Analysis (PCA)* PLS is based on a change of basis of the design and the response matrix, respectively.

In this section, only a summary of the most important concepts is given. However, to fully understand the concept of PLS-DA it is instructive to shortly mention the basics of PCA and put the two algorithms into comparison. This comparison, the mathematical background and the scheme of algorithms are outlined in Appendix 6.3[1, 11, 16–19].

In contrast to PCA, which conducts a change of basis based on the covariance matrix of the design matrix, PLS-Regression considers both, the design and the response matrix, for the change of basis.

The change of basis applied to the design matrix $\mathbf{X}$ is different to the one applied to the response matrix $\mathbf{Y}$. However, the keypoint of PLS-Regression is that the changes of basis are chosen in such a way that the covariance between the transformed sample descriptors and results becomes a maximum.

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{V}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{W}^T + \mathbf{F} \\ \mathbf{T} &= \mathbf{B}\mathbf{U} \end{aligned} \tag{1.8}$$

$\mathbf{V}$ is the transformation matrix which conducts the change of basis on $\mathbf{X}$ and is called *loadings matrix* as it describes the contribution (= *loading*) of the original variables (basis vectors of original basis) to the new variables (new basis). $\mathbf{T}$ contains the coordinates of the samples referring to the new basis vector (columns of $\mathbf{V}$) and are called *scores*. $\mathbf{E}$ and $\mathbf{F}$ are the error matrices for both transformations and are aimed to be minimized.

In analogy, $\mathbf{W}$ are the loadings of $\mathbf{Y}$ and $\mathbf{U}$ are the corresponding scores. The condition of maximum covariance between $\mathbf{U}$ and $\mathbf{T}$

$$cov(\mathbf{T}, \mathbf{U}) = \frac{1}{N-1}\mathbf{T}^T\mathbf{U} \rightarrow max \tag{1.9}$$

is forced by Eq. (1.8).

Based on this transformation, a discriminant analysis (DA) can be conducted using the computed scores to separate the data into different classes. The total procedure is then called PLS-DA and is summarized in Alg. 2. Details on different algorithms (NIPALS and SIMPLS) to compute the PLS components are outlined in the appendix. In this section PLS-1 (based on NIPALS) is used for classifying the simulated data, as the problem is binary. If a problem requires multi-class classification, the SIMPLS algorithm leads to better results than the PLS-2, which is also based on the NIPALS algorithm[20, 21].

Similar to PCA, it is possible to only use the first f PLS-components (basis vectors in transformed space) for further analysis. This can be very useful, as in high dimensional data spaces the higher PLS components often contain only noise and most of the information is contained in the first f PLS-components (corresponding to large eigenvalues of $cov(\mathbf{T}, \mathbf{U})$.

Algorithm 2: Partial Least Squares Discriminant Analysis

Training Phase:

1. Compute the loadings (**V** and **W**) and scores (**T** and **U**) with SIMPLS or NIPALS - algorithm.
2. Optimize the number of included PLS-components via cross-validation (CV)
   a) Choose the first f PLS-components to compute a linear regression model of the scores.
   b) Estimate the decision boundary by maximizing the *Area under the Curve (AUC)*.
   c) Estimate the performance of the model on validation data.
   d) Repeat steps (a)-(c) for increasing f and choose the model with the best performance on the validation data.

Classification Phase:

1. Compute the scores of the test data with the previously computed transformation matrices **V** and **W**
2. Apply the regression model using the first k PLS-components on the scores.
3. Assign class labels according to the decision boundary

Prerequisites: The covariance matrices have to be computed with standardized data and responses. It is essential that each descriptor is standardized ($\mu = 0$ and $\sigma^2 = 1$) to assure equal weight of each descriptor, independent of its average magnitude. Furthermore, if the classes are not balanced (equal number of training data) a weight is assigned to the scores for computing the decision boundary. Otherwise the decision boundary would be automatically shifted towards the larger class. This is especially crucial when a multi-class task is solved by *OvA*, in which case the classes are often very imbalanced.

Advantages and Disadvantages: An advantage of PLS-DA is that it does not require dimension reduction or feature selection as this happens automatically by computing and choosing the first f PLS-components for the model.

However, it has to be pointed out that PLS-DA is based on *linear discriminant analysis* and thus finds a linear decision boundary. Non-linearly separable datasets cannot be separated by PLS-DA. Furthermore, it is widely distributed in standard chemometric software packages and is thus often used by users without further

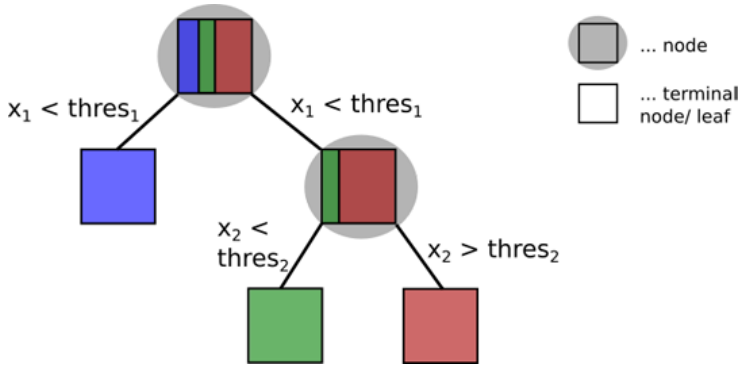


Figure 1.8: Decision tree. For simplified illustration, in this figure each splitting decision considers only one variable. However, normally $m > 1$ variables are used for each splitting decision

knowledge of supervised learning. As PLS-DA is not straightforward to understand this has resulted in misinterpreted models in several publications.

1.4.3 Random Forest (RF)

A Random Forest (RF) is an ensemble classifier based on binary decision trees. Decision trees belong to the most trivial classifiers and consist of a sequence of binary decisions, each carried out considering a subset of descriptors. At each node, the data is split into two subsets, according to a criterion which is found to separate the samples in a best possible way. With every decision the tree grows deeper and the purity of the nodes increases. The growing process is finished when all the samples in a node belong to the same class. This *terminal node* is also called *leaf*. [9, 22]

Decision trees exhibit low bias (if they are grown sufficiently deep), because with each node the model can be adapted better to the training data. However, they also exhibit high variance and are therefore prone to noise.

The idea behind RF is to build a large number N_T of de-correlated trees (ensemble), each of them exhibiting low bias but high variance. By averaging their results the variance is reduced while the bias stays the same. This procedure is called *bootstrap aggregation* or *bagging*.

Each of the N_T trees is grown with a bootstrap sample of size N_{train}, which is drawn

from the training set[4].

The variance of each tree is assumed to be σ^2 and the positive pairwise correlation[5] is $\rho\sigma^2$. The variance of the mean of N_T trees with variance σ^2 is

$$\begin{aligned} var\left[\frac{1}{N_T}\sum_{b=1}^{N_T}\hat{\mathbf{f}}(\mathbf{X^*}_b)\right] &= \frac{1}{N_T^2}var\left[\sum_{b=1}^{N_T}\hat{\mathbf{f}}(\mathbf{X^*}_b)\right] = \\ &= \frac{1}{N_T^2}\left(\sum_{b=1}^{N_T}var\left(\hat{\mathbf{f}}(\mathbf{X^*}_b),\hat{\mathbf{f}}(\mathbf{X^*}_b)\right)\right) + \\ &\quad \frac{1}{N_T^2}\left(\sum_{b1=1}^{N_T}\sum_{b2\neq b1;b2=1}^{N_T}var\left(\hat{\mathbf{f}}(\mathbf{X^*}_{b1}),\hat{\mathbf{f}}(\mathbf{X^*}_{b2})\right)\right) \\ &= \frac{1}{N_T^2}\left(N_T\cdot\sigma^2 + N_T\cdot(N_T-1)\rho\sigma^2\right) \\ &= \sigma^2\rho + \frac{1-\rho}{N_T}\sigma^2 \end{aligned} \tag{1.10}$$

It can be seen in Eq. (1.10) that the variance decreases with increasing N_T. Furthermore, it has to be emphasized that if ρ is not negligible small (much smaller than 1) the first term in Eq. (1.10) limits the effect of averaging the predictions of the tree ensemble. Therefore it is crucial that the positive pairwise correlation ρ of the decision trees is as low as possible. This is ensured by considering a subset $m < d$ of descriptors (rather than all d descriptors) for the splitting condition on each node. Thus, different splitting conditions are found for each node, resulting in differing trees and minimized pairwise correlation.

[4] A bootstrap sample is a sample, which is drawn with replacement. In this case, this means that each tree is grown with N_{train} samples, but within this training set some samples can occur multiple times while others are not represented at all. This ensures that all decision trees are trained with different training data.

[5]
$$corr(\mathbf{A},\mathbf{B}) = \frac{cov(\mathbf{A},\mathbf{B})}{\sigma_A^2\sigma_B^2}$$
If $\sigma_A^2 = \sigma_B^2 = \sigma^2$ and $corr(\mathbf{A},\mathbf{B}) = \rho$ then $cov(\mathbf{A},\mathbf{B}) = \rho\sigma^2$

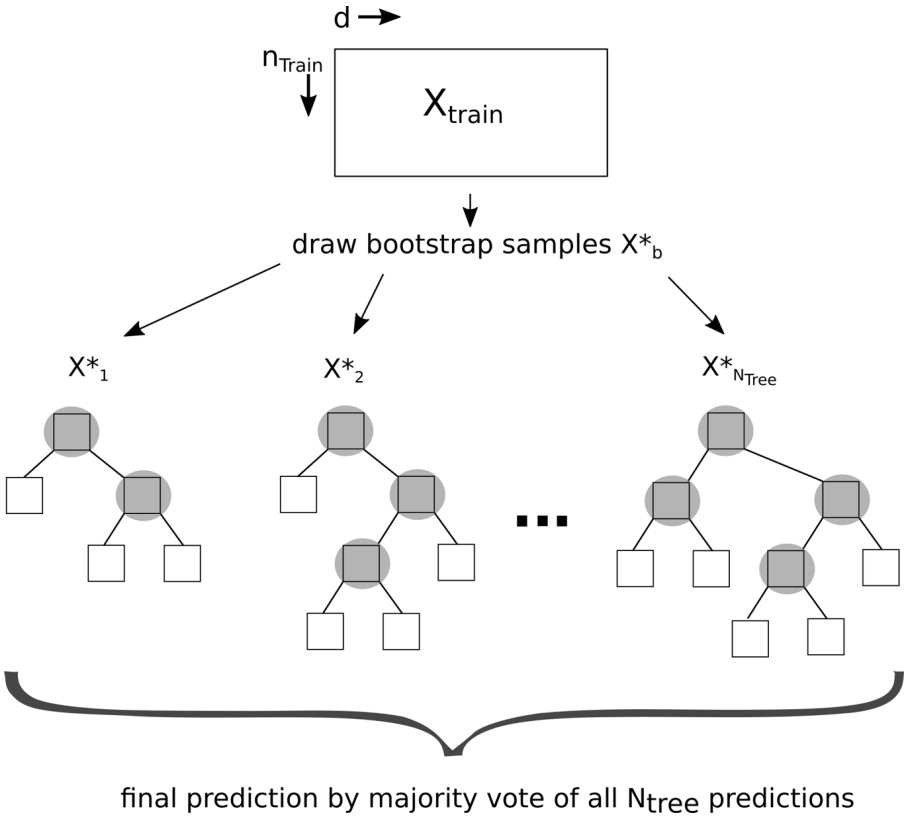


Figure 1.9: Simplified illustration of a Random Forest (RF) classifier consisting of N_T trees.

Algorithm 3: Random Forest [9, 22]

Training Phase: Growing and storing N_T decorrelated, binary decision trees. For each tree

1. a bootstrap sample $\mathbf{X}^*$ of size N_{train} is drawn of the training data,
2. with which a decision tree is grown. The splitting criteria for each node is found by
   a) randomly selecting a subset $m \leq d$ variables of the descriptors,
   b) estimating the best splitting decision for splitting the data resulting in two new daughter nodes.

   The growing process is continued until all terminal nodes reach a specified level of purity.

Classification Phase:

1. The test samples are classified by each of the N_T trees.
2. The final class is assigned by majority voting.

Model parameters: The main model parameter which has to be optimized is the number of trees N_T. If N_T is too small, the variance is still high and the model is prone to noise. Considering only the performance metrics N_T cannot be too large. However, the computation time during training increases with increasing N_T and beyond a certain N_T the performance improves only marginally with increasing N_T. Therefore also for RF it is useful to conduct parameter optimization.

However, RF do not require CV as the parameters are normally optimized by the so-called *out of bag error (OOB)*, a procedure which is similar to CV. It is based on selecting a bootstrap sample of size $N_1 < N_{train}$ for growing the i-th tree. The remaining samples are classified by the i-th tree and the error is computed. After growing all trees the mean error is estimated.

Another parameter is the size $m \leq d$ of the random variable subset used for the splitting decision at each node. The model is not particularly sensible to a change of m. In most applications the default value of $m = \sqrt{d}$ is used. However, it is important to remember that if the information is condensed in a small subset $\ll d$, m should be close to d to ensure including the relevant subspace in the randomly selected variables, as those variables are important for the splitting decision. If the relevant information is distributed over all variables m can be chosen very small.

Prerequisites: RF do not require any crucial pre-processing. The algorithm neither depends on distance measures (as kNN) nor on any regression model between variables and transformed variables, respectively (as PLS-DA). The splitting conditions for the nodes are simply based on comparison of the value of each included variable to a certain threshold. This makes scaling irrelevant as the threshold would be scaled to the same extent.

Advantages and Disadvantages: RF have evolved to one of the *state of the art* classifiers and are nowadays applied in various fields. One major advantage is that RF is rather robust concerning the parameter settings, which makes it suitable for users lacking deeper knowledge. Furthermore it inherently performs multi-class tasks and proofed to be robust in high dimensions. A disadvantage is that in comparison to many other classifiers, such as the algorithms based on discriminant analysis, the computation time during training can be very large. This especially applies to RF using a large number of trees.

1.5 The Curse of Dimensionality

Mathematical models in high dimensional data spaces can become problematic and loose their validity due to the *curse of dimensioninaliy*. This term was introduced by Bellmann (1961, [5]) and describes various phenomena that arise because of sparse data distributions in high dimensions. The volume of the data space (i.e. d dimensional unit hypercube) increases extremely fast with the dimension d. This can be briefly demonstrated by a small example. Let's assume a d-dimensional unit hypercube with uniformly distributed data. The question is, how long the edge a of a smaller cube, which contains the fraction $r = 0.1$ of the total amount of samples, has to be:

Generally, the edge length is $a = \sqrt[d]{0.1}$. In two dimensions, $a = 0.32$. For $d = 3$ one obtains $a = 0.46$ and already $a = 0.79$ for $d = 4$. One can see, that with increasing dimension, the data points are found at a large distance from the origin but close to the edge of the hypercube. For $d = 100$ the edge length is with $a = 0.98$ close to one. That means, that each data point is closer to the edge of a reference volume than to any other data point. In other words, if 100 samples are distrusted uniformally in a unit square (d=2), 10^d samples are needed to achieve the same density in a d-dimensional data space. That many samples are generally not available[15].

Several phenomena are based on this sparsity of a high dimensional space, e.g. distance measures loose their validity and the data of multivariate normal distributions are found to be located on a shell of $r = \sqrt{d}$ rather than around the mean.

Chapter 2

Methods for Artificial Dataset Generation and Classification

Most papers about *Spectral Histopathology (SHP)* focus on sample preparation, data acquisition and data pre-processing. Undoubtedly those are crucial aspects and essential to allow subsequent classification of the tissue. However, considerations on suitable classification procedures are often kept in the background. This approach is not only limited to SHP but widely distributed in analytics.

Therefore, one aim of this thesis is to theoretically examine how spectroscopic data are distributed in the d-dimensional data space and the effect of this distribution on classification problems. For that purpose artificial datasets with different characteristics are generated. The generated data sets differ in

- noise,
- dimension,
- size of training dataset and
- separability (linearly or non-linearly).

Subsequently, different classification algorithms (kNN, PLS-DA and RF) are applied to those datasets and their performance is evaluated. Although the main focus of this thesis lies on vibrational spectroscopy, the used concepts should also be applicable to other spectroscopic methods and are therefore kept as general as possible.

In Sec. 2.1 the concept and the computation of the artificial datasets are explained. In Sec. 2.2 the parameter settings of the examined classifiers are listed and justified. The results of the experiments are listed and discussed in Ch. 3.

2.1 Generation of Artificial Datasets

2.1.1 General Assumptions

On the one hand, the generated datasets should represent spectroscopic data and be comparable for different conditions (noise, dimensions, etc.). On the other hand, they should be kept simple enough to enable understanding as well as generalization for different applications. To meet these requirements, the following assumptions and restrictions are made:

Decorrelated data: The effect of correlated features on the performance of each of the considered classifiers was mentioned in Ch. 1. Furthermore it has to be emphasized that highly correlated features often lead to incorrect model interpretation and misleading feature importance ranking [23].

As for kNN decorrelated features are necessary the features of the artificial datasets are assumed to be uncorrelated.

Absorption spectroscopy results in highly correlated data, as adjacent wavenumbers have similar intensities[6]. However, the features can always be decorrelated prior to classification tasks (e.g. Principal Component Analysis, Orthogonal Nonnegative Matrix Factorization (ONMF), Discrete Wavelet Transform (DWT), Generalized Principal Component Analysis (GPCA)) [24]. Furthermore, if specific descriptors are selected rather than e.g. intensity values for equally spaced wavenumbers, feature correlation can also be decreased.

Beer-Lambert law: The measurements are assumed to obey the Beer-Lambert law, which describes the linear relationship between the absorbance $A(\bar{\nu})$ and the analyte density n, $[cm^{-3}]$ (and concentration c, $[cm^{-3}]$, respectively):

$$A(\bar{\nu}) = \sigma(\bar{\nu}) \cdot n \cdot \xi \tag{2.1}$$

$\sigma(\bar{\nu})$, $[cm^2]$ is the frequency dependent absorption cross section and ξ, $[cm]$ the path length (see also Sec. 4.2.1) [25].

Under certain circumstances the linear relationship between absorbance and concentration (or density) looses its validity and becomes non-linear. Factors which limit the linearity are [26]

- scattering effects,

[6] Data of mass spectrometry are correlated in a different way which is not considered here

- flourescence, phosphorescence
- for solutions: high concentrations of the analyte (due to inter-molecular interactions of the solvent resulting in different charge distributions and potential change of the refractive index)
- for solid samples: high sample thickness
- instrumental deviations (polychromatic radiation, non-linearity of the detector etc.)

Binary classification task with balanced classes: It is assumed that the classification task is binary (two classes). As discussed in Sec. 1.1 a multi-class problem can be solved by a system of binary classifiers.

Furthermore both classes are assumed to have the same number of training samples. In case this condition cannot be met, various resampling techniques can be applied to ensure balanced classes.

Gaussian noise: All models are based on Gaussian noise, which means that the samples of each class are distributed normally with the class specific mean vector $\mu_{\mathbf{c}}$, $c = 0, 1$ and the covariance matrix $\mathbf{\Sigma}$, which is assumed to be the same for both classes.

$$\begin{aligned} \mathbf{X} &\sim \mathcal{N}_d(\mu, \mathbf{\Sigma}) \\ \mu_c \, &\epsilon \mathbb{R}^d \\ \mathbf{\Sigma}_{i,j} &= cov(\mathbf{x}_i, \mathbf{x}_j), \; \mathbf{\Sigma} \, \epsilon \mathbb{R}^{d \times d} \end{aligned} \tag{2.2}$$

As already mentioned, the features are assumed to be uncorrelated and the variance is assumed to be equal for all features. Thus, $\mathbf{\Sigma} = \sigma^2 \cdot \mathbf{I}$ is diagonal as $\mathbf{\Sigma}_{i,j} = 0$ for $i \neq j$ and $\mathbf{\Sigma}_{i,i} = \sigma^2$.

2.1.2 Artificial Data Models

Two different approaches are used for the generation of the artificial datasets. Both are based on the idea of creating linearly (LS) and non-linearly (NLS) separable data in different dimensions, with different noise levels and different sizes of the training dataset.

Model 1 (*Ball Model*) is intuitive and widely used in theoretical examinations of classification algorithms. For the LS dataset the model consists basically of two

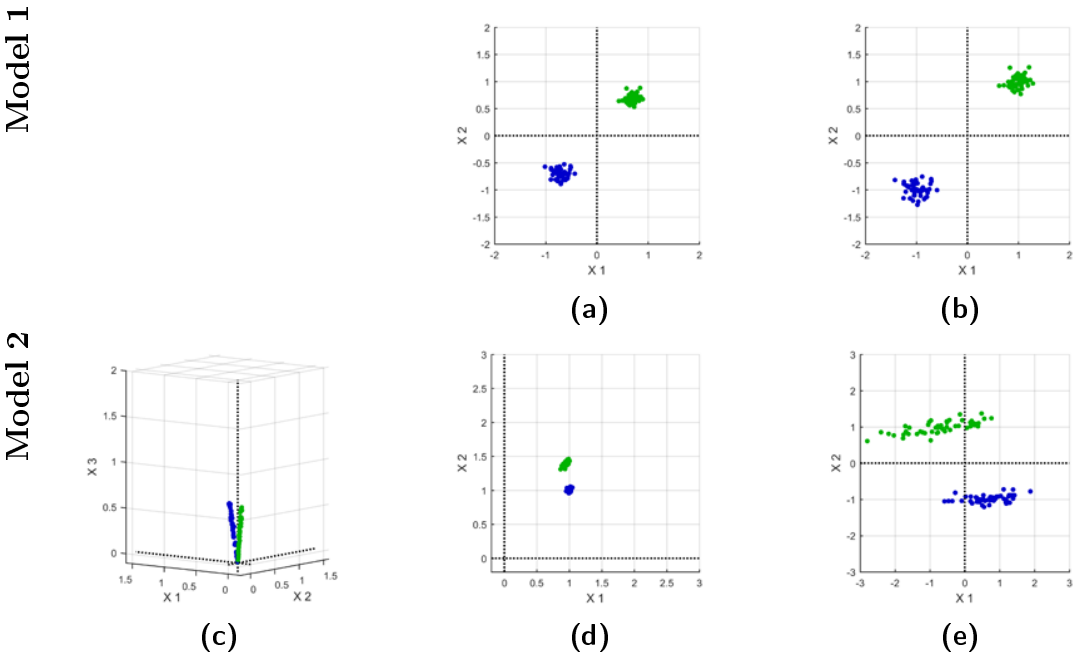


Figure 2.1: Scatter plots of artificial datasets for noise level $\eta = 0.05$. Row 1 illustrates Model 1 ($d = 2$, LS) a) raw data; b) standardized data. Row 1 illustrates Model 2 ($d = 3$, LS) a) raw data; b) normalized to internal standard x_3; c) standardized.

d-dimensional multivariate normal distributions with different mean vectors. The NLS dataset consists of a d-dimensional multivariate normal distributions for class 1 surrounded by a noisy (d-1)-sphere. Model 1 does not represent raw hyperspectral data, but it is suitable for understanding concepts in high dimensional spaces and is comparable to hyperspectral data after normalization. It is further explained in Sec. 2.1.

Model 2 (*ray Model*) is created to represent hyperspectral data considering the restrictions mentioned in Sec. 2.1. The idea is, that each substance in the measured compound consists of a specific combination of the features. However, different concentrations of the substances result in different absorbance values. Before normalization of the spectra, this can be represented by rays in the d-dimensional space. All rays intersect in the origin and the position of a data point along its class ray represents the concentration. Model 2 is further explained in Sec. 2.1.

In most applications the information about a concentration or sample thickness is equalized by the normalization during pre-processing. The most prominent approaches for normalization are vector normalization (sum of feature squares equals

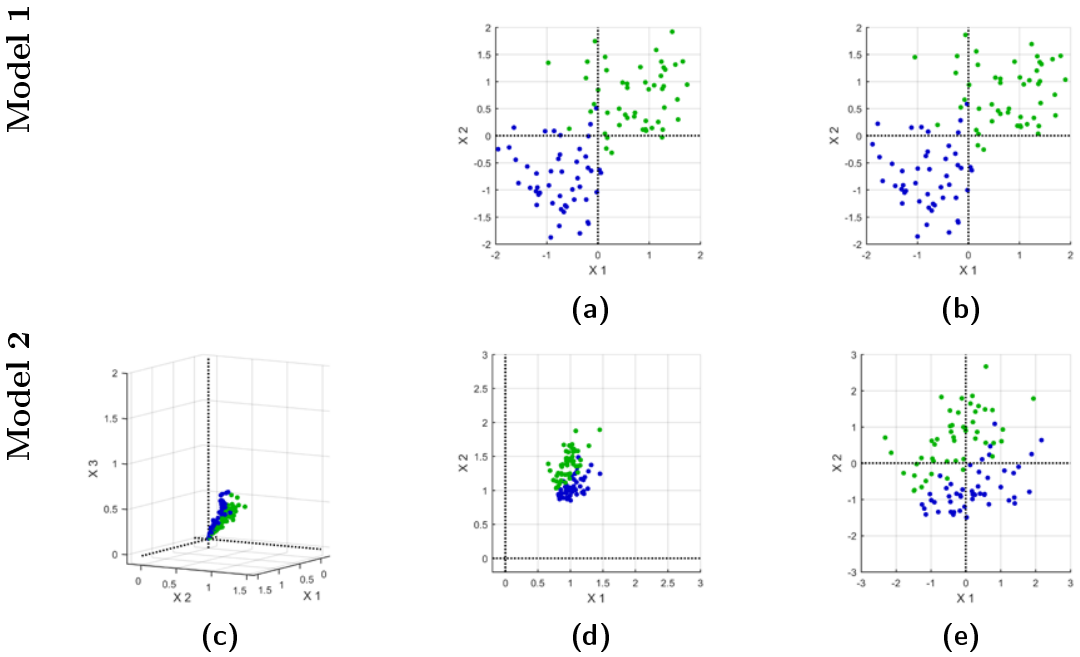


Figure 2.2: Scatter plots of artificial datasets for noise level $\eta = 0.3$. Row 1 illustrates Model 1 ($d = 2$, LS) a) raw data; b) standardized data. Row 1 illustrates Model 2 ($d = 3$, LS) a) raw data; b) normalized to internal standard x_3; c) standardized.

1) or normalization to an internal standard[7]. Therefore, the Model 2 datasets are normalized before classification. Here, normalization to an internal standard is chosen due to better comparability to Model 1. However, similar results are expected for vector normalization.

Model 1 data do not contain any information about concentration and thus do not have to be normalized. As by normalization to an internal standard the data space is reduced by one dimension, the d-dimensional Model 1 can be compared to the (d+1)-dimensional Model 2.

As mentioned in Ch. 1, kNN and PLS-DA require standardization during preprocessing while for RF standardization does not make any difference. Thus, both models are standardized before classification. In Fig. 2.1 the data distributions for the 2D Model 1 and the 3D Model 2 are illustrated before and after standardization for noise level $\eta = 0.05$ and in Fig. 2.2 for $\eta = 0.3$

All data are created using MATLAB 2015b (The MathWorks, Inc., Natick, Mas-

Symbol	Description	Values
d	dimension	1 to 30
N_{pC}	number of training samples per class	100,200,300,...,4900,5000
η	noise level	0.05, 0.1, 0.2, 0.3
	separability	LS (linearly), NLS (non-linearly)

Table 2.1: Parameters for generation of Model 1 data

sachusetts, United States).

Model 1: Ball Model

In the following the method to generate Model 1 datasets is outlined briefly. Tbl. 2.1 lists the parameters, which are varied in order to obtain different datasets.

In the **linearly separable case (LS)** the model basically consists of two classes with class means $\mu_{\mathbf{c}}$ (c = 0 or 1), which are located at opposite directions of the data space.

$$\begin{aligned} \mu_{\mathbf{0}} &= (1,1,1,...,1)/\sqrt{d}, \qquad \mu_{\mathbf{0}} \epsilon \mathbb{R}^d \\ \mu_{\mathbf{1}} &= (-1,-1,...,-1)/\sqrt{d}, \qquad \mu_{\mathbf{1}} \epsilon \mathbb{R}^d \end{aligned} \tag{2.3}$$

According to the Gaussian noise model, the data of each class are distributed normally in d dimensions with the class mean $\mu_{\mathbf{c}}$ and $\mathbf{\Sigma}$. The covariance matrix Σ is the same for both classes. σ^2 depends on the noise level parameter η and the Euclidean distance between the class means $a = \|\mu_{\mathbf{0}} - \mu_{\mathbf{1}}\|$.

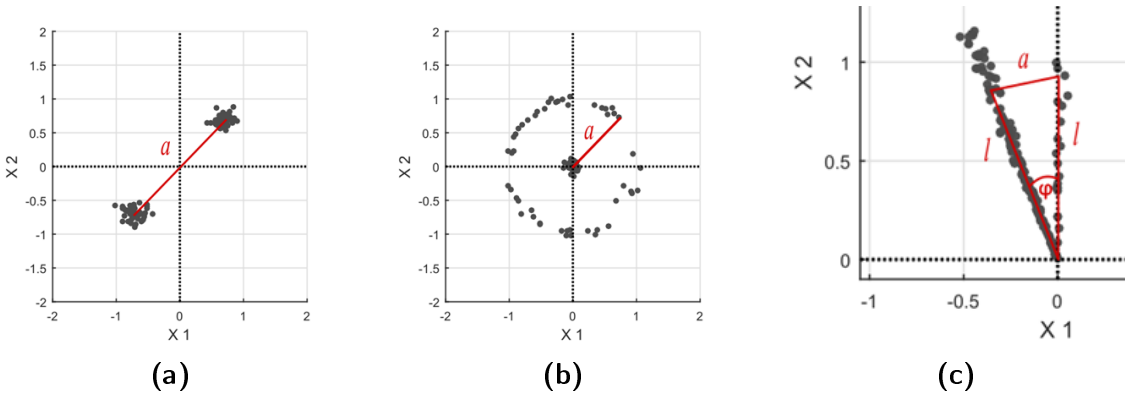
$$\sigma = \eta \cdot a \tag{2.4}$$

In the **non-linearly separable case (LS)** class 0 data are distributed normally around the origin.

$$\mu_{\mathbf{0}} = (0,0,...,0), \qquad \mu_{\mathbf{0}} \epsilon \mathbb{R}^d \tag{2.5}$$

Class 1 data are distributed normally with mean vectors $\mu_{\mathbf{1,j}}$ on a d-sphere with radius a. For each of the $j = 1$ to N_{pC} class 1 samples, a mean vector $\mu_{\mathbf{1,j}}$ is chosen

[7] An *interal standard* is a specific feature, which is assumed to have the same intensity for each substance of the compound, such as the amide 1 band in most biological samples in IR spectroscopy based images



(a) (b) (c)

Figure 2.3: Illustration of the reference length a for setting the standard deviation σ according the specified noise level η. (Here: $\eta = 0.05$ a) Model 1: linearly separable; b) Model 1: non-linearly separable; c) Model 3, before rotation to first hyperoctant: valid for linearly and non-linearly separable case

randomly from a uniform distribution on the sphere. The sample coordinates are then assigned by a normal distribution with the previously selected $\mu_{\mathbf{1,j}}$ and $\mathbf{\Sigma}$.

Model 2: Ray Model

In the following the algorithm to generate Model 2 datasets is outlined briefly. Basically the model consists of one class 0 ray, which is oriented along the diagonal $\hat{\mathbf{x}}_0 = (1, 1, ..., 1)/\sqrt{d}$ of the d-dimensional hyperoctant. This ray represents the substance which should be detected. The model assumes that class 1 consists of $N_{ray1} > 1$ substances, which are all represented by a specific direction in the d-dimensional data space (ray) with the corresponding unit vector $\hat{\mathbf{x}}_{c1,i}$, $i \, \epsilon \, [1, N_{ray1}]$, $\hat{\mathbf{x}}_{c1,i} \neq \hat{\mathbf{x}}_0$. Class 1 rays are all chosen to form the angle φ with the class 0 ray.

$$\cos\varphi = \hat{\mathbf{x}}_{c1,i} \cdot \hat{\mathbf{x}_0}^T \tag{2.6}$$

In addition to this condition, the N_{ray1} class 1 rays are subject to certain constraints which ensure the respective separability (linearly or non-linearly) and a ray distribution which is as even as possible. Those constraints are outlined in more detail below.

The algorithm first creates the noisy data for all rays orientated parallel to $\hat{\mathbf{x}}_{ini} = (0, 0, ..., 0, 1)$, $\hat{\mathbf{x}}_{ini} \, \epsilon \, \mathbb{R}^d$ by a multivariate normal distribution in d dimensions. Subsequently each class 1 ray is rotated in a previously chosen rotation plane by the defined angle φ (Fig. 2.4) Finally, all data (class 0 and class 1) are rotated to the first hyperoctant. The steps are summarized in Alg. 4 and discussed further be-

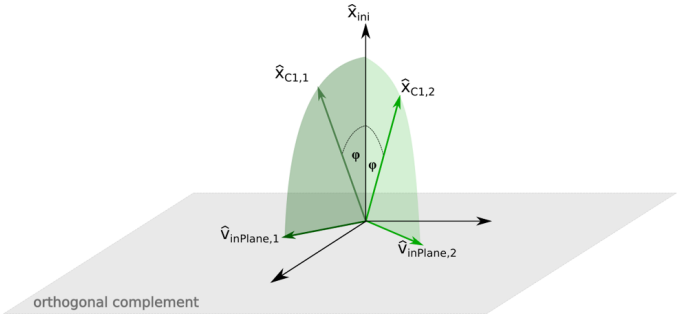


Figure 2.4: Illustration of the rotation of the class 1 cones by the angle φ in 3D.

low. Tbl. 2.2 lists the parameters, which are varied in order to obtain different datasets.

Algorithm 4: Overview on generation of model 2 data

for ***each dimension*** **do**
- compute the rotation matrix M_{all} to rotate all data to first hyperoctant;
- compute the rotation matrices M_i, $i \, \epsilon \, [1, N_{ray1}]$ to rotate class 1 rays;

for ***each*** N_{pC} ***and noise level*** η **do**
- create multivariate normally distributed data with $\mu = (0, 0, ..., 0, 1)l$ for uniformly distributed l;
- rotate all class 1 cones to a boundary ray of the main cone (defined by the fixed aperture angle φ, $\cos(\varphi) = \langle \hat{\mathbf{x}}_\mathbf{0}, \hat{\mathbf{x}}_{c1,i} \rangle$) using the computed rotation matrices M_i;
- rotate all (class 1 and class 0) data points to the first hyperoctant using computed rotation matrix M_{all};

end

end

Generating data with different intensities and Gaussian noise: This model aims to simulate compounds with different concentrations and Gaussian noise. The concentration is considered by varying the length l of the mean vector μ for the normal distribution. For each data point l is selected randomly from a uniform distribution $U(0, l_{max})$. As the data distribution takes place before the rays are rotated, the initial mean vector for the j-th data point of the c-th class is:

Symbol	Description	Values
d	dimension	2 to 15
N_{pC}	number of training samples per class	10,50,100,200,300,...,1900,2000
η	noise level	0.05, 0.1, 0.2, 0.3
	separability	LS (linearly) a and b, NLS (non-linearly)
$[N_{ray1}$	number of class 1 rays	20]
$[\theta$	opening angle of class 1 rays hull	$\pi/18]$

Table 2.2: Parameters for generation of Model 2 data. The parameters in brackets are kept constant in this model, could be modified though.

$$\begin{aligned} \mu_{\mathbf{c,j}} &= (0,0,0,...,1)l \\ l &\sim U\,(0,l_{max}) \\ c &\,\epsilon\, 0,1 \\ j &\,\epsilon[1,N_{pC}] \end{aligned} \tag{2.7}$$

The coordinates for each data point are then computed by a multivariate normal distribution with the corresponding $\mu_{\mathbf{c,j}}$ and $\mathbf{\Sigma_{c,j}}$. Similar to Model 1, the standard deviation for the Gaussian distribution is related to the inter-class distance and the specified noise level η. In contrast to the *ball model*, the inter-class distance a of the the *ray model* depends on the signal intensity, which is here represented by the length of the mean vector l.

Furthermore the inter-class distance depends on the chosen angle φ between the class 0 and class 1 rays. The relationship between the standard deviation of the noise and vector length l is assumed to be linear and modeled by Eq. (2.8)

$$\begin{aligned} a &= l/2 \cdot \sin\frac{\varphi}{2} \\ \sigma &= a \cdot \eta \end{aligned} \tag{2.8}$$

The proportionality between the standard deviation and the feature vector length is based on the assumption that the noise intensity is proportional to the signal intensity. This assumption would result in different σ_i for each dimension (spectral descriptor), depending on its intensity. However, for simplification the standard deviation σ of the noise is assumed to be equal for all dimensions in this model. Instead it is chosen to be proportional to the feature vector length.

It is worth noting that noise originates from various sources and that the assumption of proportionality between signal and noise intensity is often violated.

Rotation matrices to obtain tilted class 1 rays: The matrices M_i, $i \epsilon [1, N_{ray1}]$ for rotating the N_{ray1} class 1 rays from $\hat{\mathbf{x}}_{ini} = (0, 0, ..., 0, 1)^T$, $\hat{\mathbf{x}}_{ini} \epsilon \mathbb{R}^d$ to the desired $\hat{\mathbf{x}_{\mathbf{c1,i}}}$ are computed individually for all different dimensional datasets. However, for each dimensionality of the feature space the same set of rotation matrices is used for different noise levels η and training set sizes N_{pC}.

First the rotation plane is defined by randomly choosing $\hat{\mathbf{v}}_{InPlane,i}$ from a uniform distribution in the orthogonal complement to $\hat{\mathbf{x}}_{ini} = \hat{\mathbf{x}}_0$. The orthogonal complement to the plane spanned by $\hat{\mathbf{x}}_{ini}$ and $\hat{\mathbf{v}}_{InPlane,i}$ is the subspace about which the j-th class 1 ray is rotated by θ. Any basis of this (d-2)-dimensional orthogonal complement is taken as a simplex to compute the rotation matrix M_i using the algorithm of Aguilera - Perez [27] (see 6.3).

Supplementary conditions have to be satisfied regarding the rotation of the class 1 rays to ensure linearly (LS) and non-linearly (NLS) separable classes, respectively.

Constraints for linearly separable data: Before randomly choosing $\hat{\mathbf{v}}_{InPlane,i}$ and thus defining the rotation planes a reference vector $\hat{\mathbf{v}}_{ref}$ is chosen randomly in the d-1 dimensional orthogonal complement of $\hat{\mathbf{x}}_{ini} = \hat{\mathbf{x}}_0$. Subsequently, $\hat{\mathbf{v}}_{InPlane}$ is rejected if $\langle \hat{\mathbf{v}}_{InPlane}, \hat{\mathbf{v}}_{ref} \rangle < 0.7$ for the LS-a ray model (linearly separable A).

The LS-b ray model defines the reference vector $\hat{\mathbf{v}}_{ref}$ as $\hat{\mathbf{v}}_{InPlane}$ for all class 1 rays. Thus, all class 1 rays are rotated in the same manner, resulting in one class 0 and one class 1 ray. After normalization to an internal standard, this case corresponds to the linearly separable ball model.

Constraints for non linearly separable data: For each randomly chosen $\hat{\mathbf{v}}_{InPlane,i}$ (and therefore defined rotation plane) rotations are conducted by both, φ and $-\varphi$. If the number of class 1 rays is lower than the number of hyperoctants (2^d), each hyperoctant can only be occupied once.

Rotation matrix to rotate all data to the first hyperoctant: After tilting the class 1 rays all data are rotated to the first hyperoctant, so that the class 0 ray ($\hat{\mathbf{x}_0}$) becomes orientated along the direction $\hat{\mathbf{x}_0} = (1, 1, ..., 1)/\sqrt{d} \, \epsilon \, \mathbb{R}^d$. The corresponding rotation matrix $\mathbf{M_{all}}$ describes a rotation in the plane spanned by $\hat{\mathbf{x}_{ini}} = (0, 0, ..., 0, 1)$ and $\hat{\mathbf{x}_0} = (1, 1, ..., 1)/\sqrt{d}$ and is computed by the Aguilera - Perez algorithm[27] (see 6.3). The required simplex is again spanned by the basis vectors of the (d-2)-dimensional orthogonal complement of the rotation plane

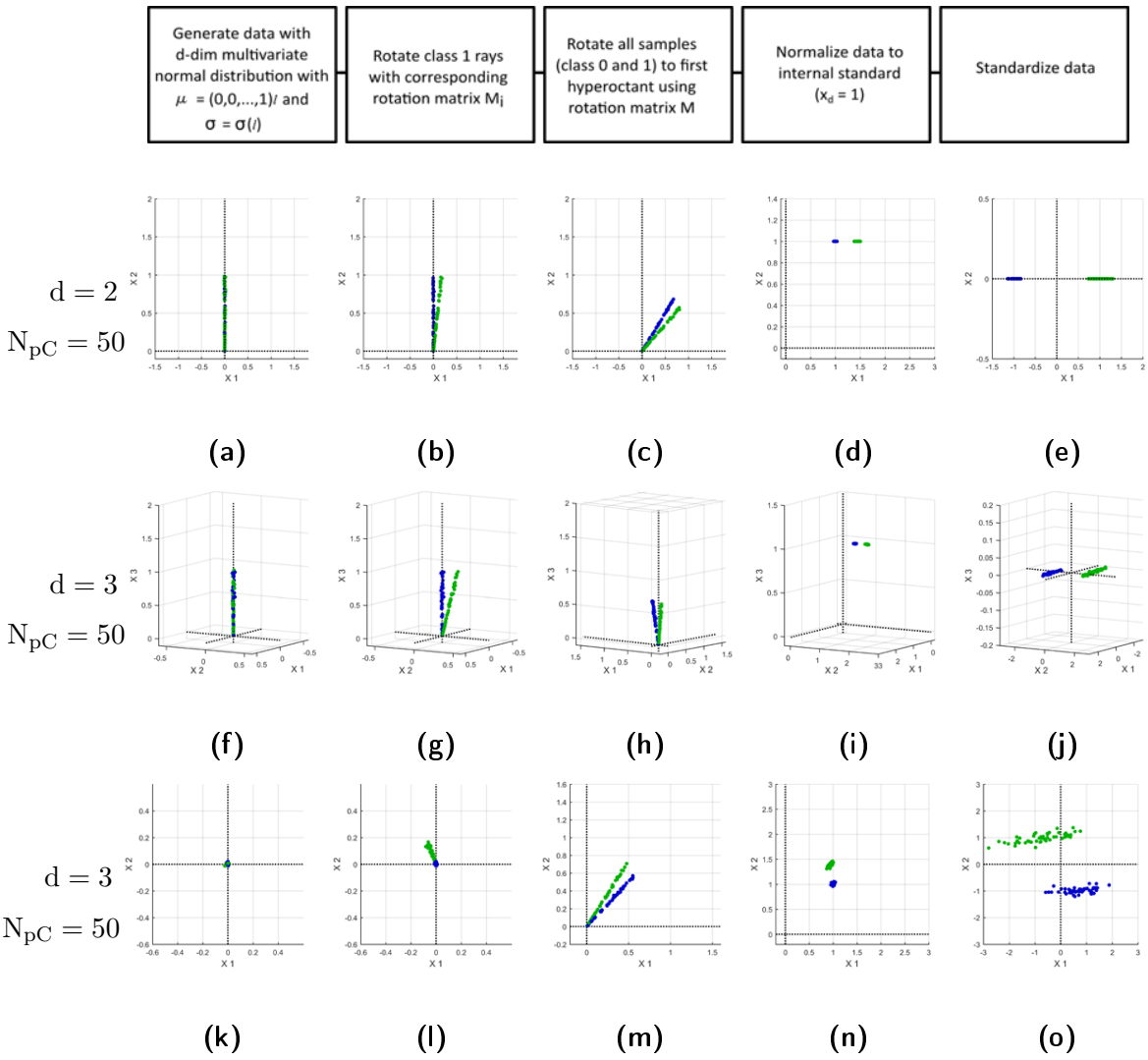


Figure 2.5: Visualization of data generation and pre-processing for linearly separable data of model 2 (ray model, LS-b). a)-c) data generation steps and d)-e) pre-processing steps for for d = 2; f)-h) data generation steps and i)-j) pre-processing steps for d = 3 (3 dimensional view); k)-m) data generation and n)-o) pre-processing steps for d = 3 (2 dimensional view; $X_1 - X_2$ plane)

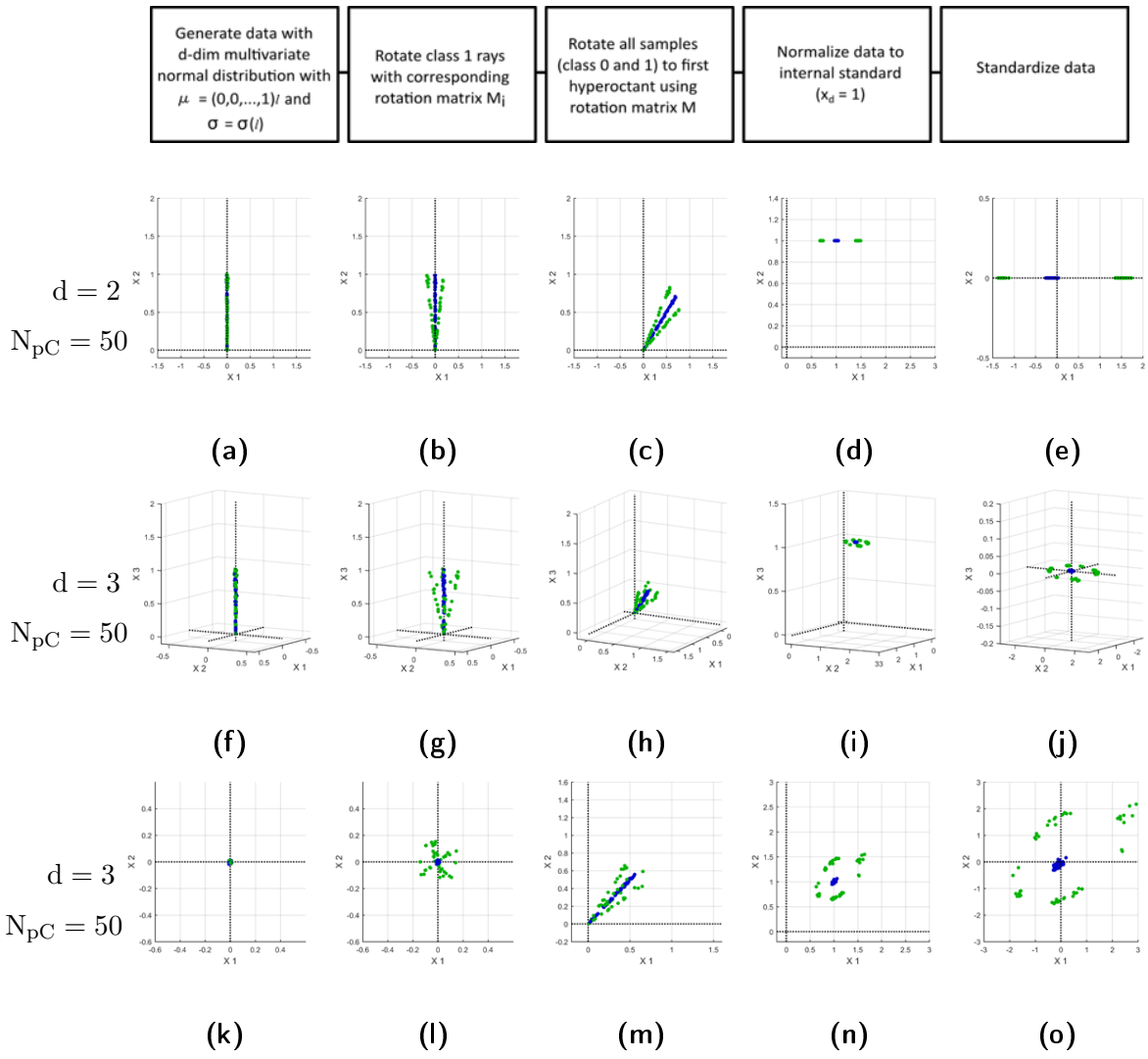


Figure 2.6: Visualization of data generation and pre-processing for non-linearly separable data of model 2 (ray model, NLS). a)-c) data generation steps and d)-e) pre-processing steps for for d = 2; f)-h) data generation steps and i)-j) pre-processing steps for d = 3 (3 dimensional view); k)-m) data generation and n)-o) pre-processing steps for d = 3 (2 dimensional view; $X_1 - X_2$ plane)

2.2 Classifier Settings

Classifier training and testing is carried out using MATLAB 2015b (The MathWorks, Inc., Natick, Massachusetts, United States). All classifiers are tested on a test set with 5000 samples per class. Test data exhibit the same noise level and dimension as the respective training set. Test data for the ray model are generated using the same rotation matrices as are used for training data generation.

The performance of a classifier is estimated in terms of accuracy, sensitivity and specificity. Here, the accuracy is equal to the mean of sensitivity and specificity, as only balanced classes are featured. Furthermore the required training and testing times are recorded. The results are visualized using the MATLAB script *Hatchfill*[28].

Classification parameters are estimated using 10-fold cross validation based on the accuracy as performance measure. The *one standard error rule* is applied to estimate the optimum model parameters. CV is stopped automatically when the accuracy does not change more than $\epsilon = 0.05$ for five consecutive parameter variations.

kNN: The MATLAB function *fitknn* with the Euclidean distance metric is used to build the kNN model. The number of nearest neighbors k is estimated by CV in the interval $k = 3$ to 21.

PLS-DA: The PLS components are computed using a script of Yi Cao (2009)[29], featuring the NIPALS algorithm. The number of PLS components used for the discriminant analysis is estimated by CV. The threshold for the discriminant analysis is found by maximizing the AUC.

RF: The MATLAB function *treebagger* is used to implement the Random Forest. N_{Train} samples are selected randomly with replacement (bagging) for each tree. The number of randomly chosen descriptors for each split is $\sqrt{d}$, with d being the number of features (dimension). Considering the number of trees, preliminary experiments are conducted on similar datasets which all result in a selection of 30-50 trees. Cross validation curves show, that choosing more than the selected number of trees increases the computation time but does not change the performance (in contrast to other classifiers). Based on this knowledge, all RF are created with 75 trees in order to save capacities.

Chapter 3

Results and Discussion of Classifying Artificial Spectroscopic Datasets

3.1 Cross Validation

As already mentioned in Sec. 2.2, the classification parameters are selected using 10-fold CV applying the 1-standard error rule. For kNN and PLS-DA the CV is conducted for all different models during the training. A preliminary study showed, that in case of the RF the performance does not decrease if selecting a higher N_{tree} than necessary. Thus, to save computation time $N_{tree} = 75$ was chosen for all cases. This behavior of a RF is illustrated in Fig. 3.1, which shows the CV curve for a linearly separable dataset of the ball model, with $d = 15$, $N_{train} = 1000$ and $\eta = 0.05$.

3.2 Classifier Performance

For each separation characteristic (LS or NLS), model type (ball or ray) and noise level η (e.g. linearly separable ball model, $\eta = 0.2$) the accuracy, sensitivity and specificity are illustrated as heat maps with varying dimension d (x-axis) and number of training data per class N_{pC} (y-axis).

Those heatmaps allow the interpretation of the performance trend with increasing d and N_{pC}. It is evident that classifier performance is mainly limited by the structure of the data in the dataspace, but also dependencies on the selected classifier can be explained.

The heatmaps are color coded in the range of 0.5 to 1.0 with increments of 0.05. The relative pixel area $A_{px>th}$ covered by each threshold level is estimated and listed for better comparison between the individual cases.

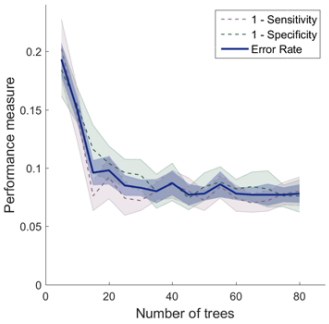


Figure 3.1: Cross validation of an RF classifier for the linearly separable ball model, $d = 15$ and $N_{pC} = 1000$. Accuracy, sensitivity and specificity are illustrated. The shaded areas represent one standard error. Here, $N_{tree} = 25$ is selected using the 1-standard error rule.

3.2.1 Ball Model

Linearly separable ball model (Fig. 3.2): The accuracy of the linearly separable data set is high ($ACC > 0.99$ for $\eta = 0.05, 0.1, 0.2$ and $ACC > 0.94$ for $\eta = 0.3$) for all classifiers and cases. The decrease in accuracy for higher noise level can be explained, as the data cluster of the two classes overlap for noisy data Fig. 2.2.

In Fig. 3.2j to Fig. 3.2l (linearly separable ball model, $\eta = 0.3$) stripes of slightly different accuracy are present at certain dimensions. This is only an artifact due to machine accuracy and the chosen threshold for color mapping, as for all classifiers $ACC \sim 0.95$ (different color coding for $ACC > 0.95$ and $ACC < 0.95$).

Non-linearly separable ball model (Fig. 3.3): The curse of dimensionality becomes apparent when interpreting the obtained accuracies for the NLS ball model. As a reminder, class 0 of the model is created by a multivariate normal distribution with $\mu = (0, ..., 0)$ and a covariance matrix $\mathbf{\Sigma}$ according to the corresponding noise level. The mean vectors of class 1 are placed on a sphere around the class 0 cluster. The class 1 data points are then created according to a multivariate normal distribution with those mean vectors and $\mathbf{\Sigma}$. As mentioned in Sec. 1.5, normally distributed data points in high dimensions are not accumulated close to the mean vector anymore (as in low dimensions), but rather distributed in a shell with $r = \sqrt{d}$ around the mean. Thus, class 0 and class 1 overlap highly in high dimensions and are not distinguishable by a classifier anymore.

Nevertheless, performance differences between the classifiers are apparent. The accuracy obtained for the PLS-DA classification of the non-linear data set is $ACC = 0.66$ for $d = 1$ and decreases with higher dimensions, approaching $ACC = 0.5$.

The reason is obvious, PLS-DA attempts to find a linear decision boundary for a non-linear problem, which is not possible at all. However, the classifier "tries it's best" by defining a decision boundary that correctly assigns all of class 0 as positive (TP), while assigning as few class 1 data points as possible to class 0 (FP). However, it becomes apparent that PLS-DA is not suitable for non-linear classification problems.

Although the accuracy obtained by the RF classification also decreases with higher dimensions, the results illustrate that RF is more suitable for such problems than kNN. For kNN $ACC \rightarrow 0.5$ already at $d = 5$ for a low number N_{pC} and at $d = 10 - 12$ for higher N_{pC}. This can be explained, as the data space becomes sparse for high dimensions and distance measures loose their validity (Sec. 1.5). The RF is not based on a distance measure. Furthermore it selects a random subset of descriptors and creates thus a reduced dataspace for each split.

3.2.2 Ray Model

Non-linearly separable ray model (Fig. 3.4): This model shows similar characteristics to the NLS ball model and the results can be reasoned similarly. However, the classifier performance on the NLS ray model is slightly higher than on the NLS ball model (for $d = 7$, $\eta = 0.3$ and $N_{pC} = 1500$ approx. $ACC(ball\ model) = 0.8$ and $ACC(ray\ model) = 0.85$), This can be explained, by the chosen dependence of the standard deviation of the noise from the interclass distance. In the ray model, this interclass distance is taken before the rays are rotated to the diagonal of the first hyperoctant. When subsequent normalization to the X_d-axis is performed, for certain class 1 ray orientations the resulting class distances in the projected space become larger than before the projection.

Linearly separable ray model A (Fig. 3.6): Although this model is linearly separable, it shows similar characteristics to the NLS sets of the ball and ray model. The main characteristics of the NLS model are not due to the type of separability but rather in the curse of dimensionality and its effect on the multivariate normal distribution. In the LS-A ray model, all class 1 rays are rotated to one part of the orthogonal subspace of the class 0 ray (restriction that $\langle \hat{\mathbf{v}}_{InPlane}, \hat{\mathbf{v}}_{ref} \rangle > 0.7$). Nevertheless, this results in a spherical distribution of class 1 data around class 0 data. Even though only a part of the sphere is occupied by class 1 mean vectors, the result is similar to the NLS case as the curse of dimensionality has the same effect.

Linearly separable ball model B (Fig. 3.5): This model corresponds to the LS ball model and leads to similar results. Differences in accuracy between the dimensions for $\eta = 0.3$ arise from the random selection of the rotation matrix, which can result in differently rotated class 1 rays. As a reminder, class 1 rays are rotated with respect to the class 0 ray, when all rays are still orientated along the initial vector $\mathbf{x}_{ini} = (0, 0, ..., 0, 1)$. Subsequently all data points are rotated in such a way, that the class 0 ray is orientated along the diagonal of the first hyperoctant $(1, 1, ..., 1)/\sqrt{d}$. Normalization to the variable $x_d = 1$ can skew the class separation in different ways, which results in better or worse class separation.

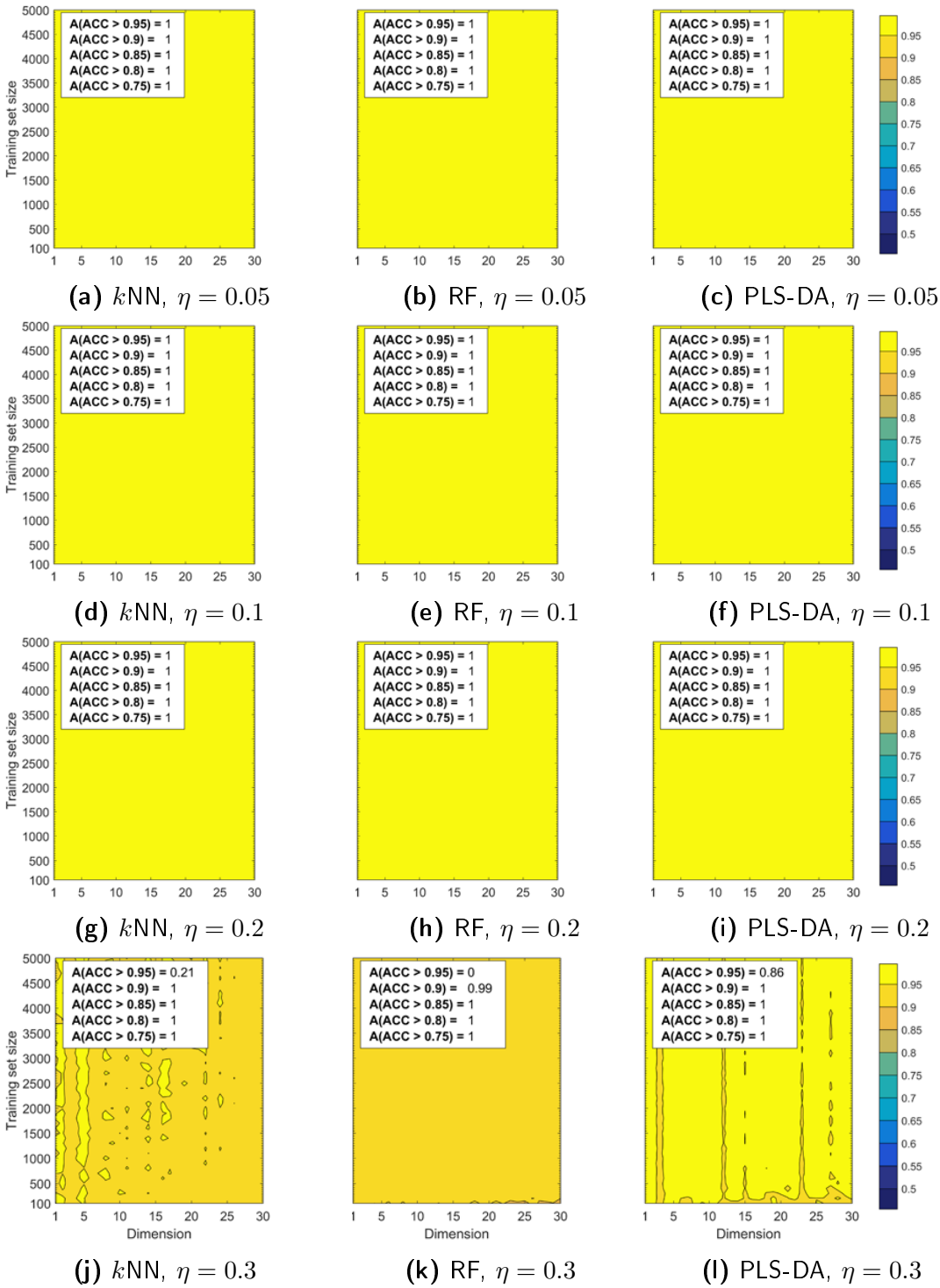
3.3 Sensitivity, Specificity and Accuracy

Sensitivity and specificity plots are not shown for all cases as the general behavior can be illustrated in two examples (Fig. 3.7 and Fig. 3.8). It becomes apparent, that the kNN classifier maintains a high sensitivity by accepting a low specificity (or the other way round). That means, for higher dimensions kNN starts assigning all samples to one class. On the contrary, the RF for the same problem results both in reduced sensitivity and reduced specificity which show similar behavior with dimension and N_{train}. Furthermore, the above mentioned behavior of the PLS-DA classification on NLS datasets becomes apparent.

3.4 Computation Time

In addition to the performance measures, the required computation time for training and testing is recorded. The numeric values are not listed in this thesis, as they depend on the available processing power. However, a ratio of the required computation time can be stated. This comparison is conducted using the same CV settings for all classifiers (e.g. 10-fold cross validation for all classfiers). kNN requires hardly any training time but a higher testing time. In contrast to that RF requires a lower testing than training time. Hence, the sum of training and testing time is used for comparison.

Obviously for all classifiers the computation time increases with increasing dimension and training data set. However, it turned out that the required computation time differs by one order of magnitude between the classifiers. The time required for kNN is approximately ten times the time needed for training and testing a PLS-DA classifier. A RF needs approximately a hundred times the time PLS-DA requires.



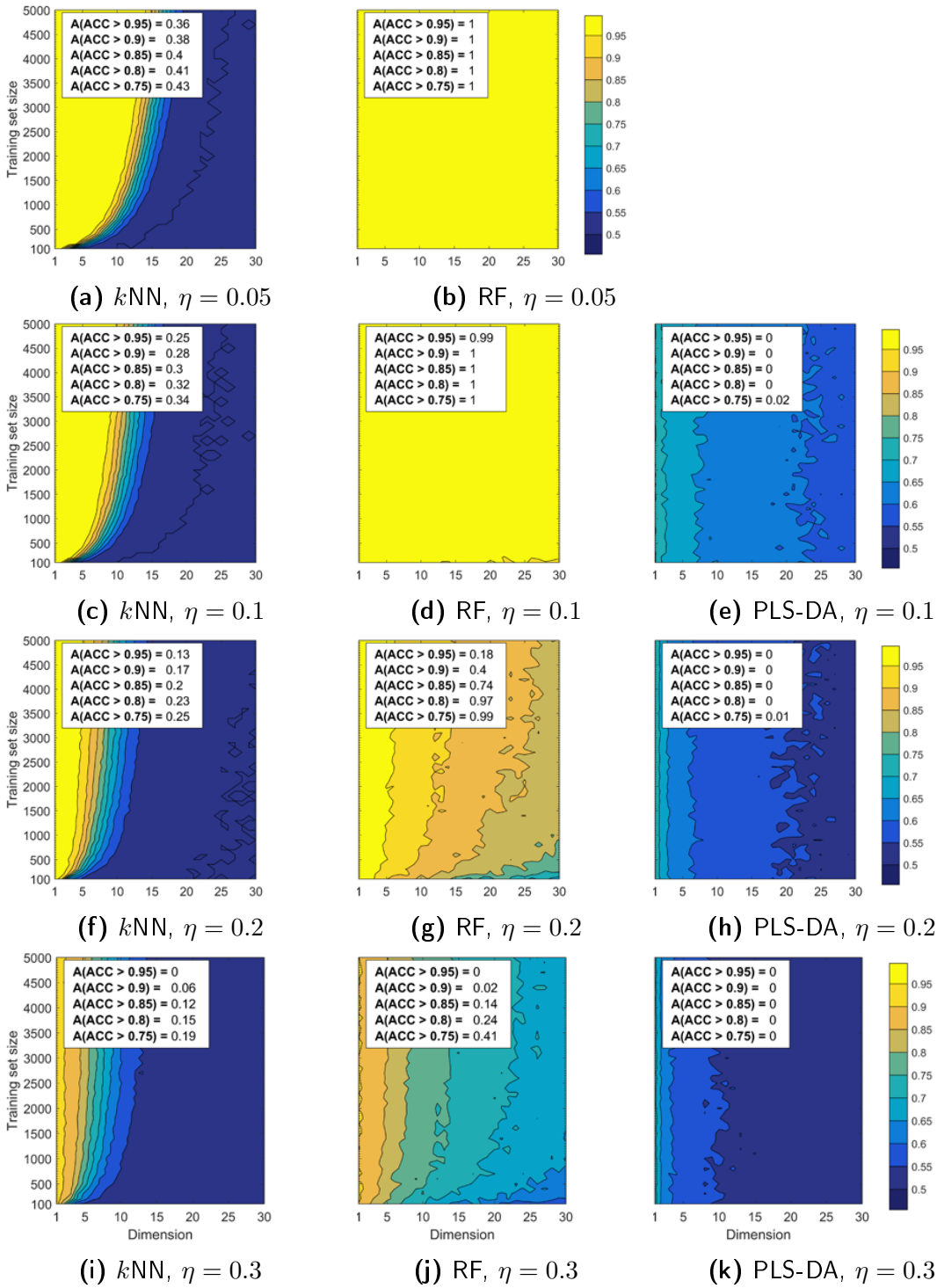
(a) kNN, $\eta = 0.05$ **(b)** RF, $\eta = 0.05$ **(c)** PLS-DA, $\eta = 0.05$

(d) kNN, $\eta = 0.1$ **(e)** RF, $\eta = 0.1$ **(f)** PLS-DA, $\eta = 0.1$

(g) kNN, $\eta = 0.2$ **(h)** RF, $\eta = 0.2$ **(i)** PLS-DA, $\eta = 0.2$

(j) kNN, $\eta = 0.3$ **(k)** RF, $\eta = 0.3$ **(l)** PLS-DA, $\eta = 0.3$

Figure 3.2: Classification accuracy for the data sets of the linearly separable *ball model*. The inset specificies the relative areas at different levels of accuracy.



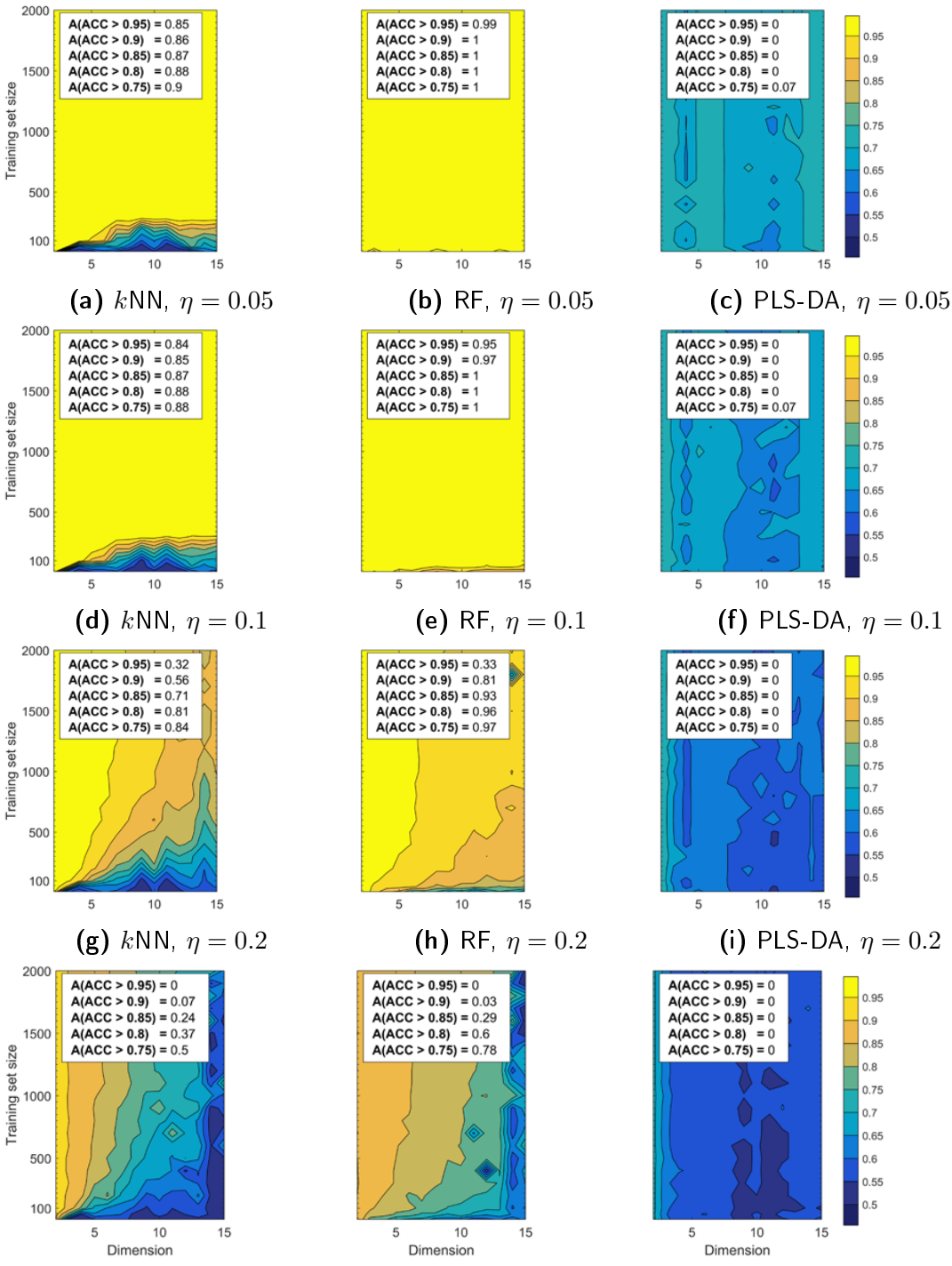
(a) kNN, $\eta = 0.05$ **(b)** RF, $\eta = 0.05$

(c) kNN, $\eta = 0.1$ **(d)** RF, $\eta = 0.1$ **(e)** PLS-DA, $\eta = 0.1$

(f) kNN, $\eta = 0.2$ **(g)** RF, $\eta = 0.2$ **(h)** PLS-DA, $\eta = 0.2$

(i) kNN, $\eta = 0.3$ **(j)** RF, $\eta = 0.3$ **(k)** PLS-DA, $\eta = 0.3$

Figure 3.3: Classification accuracy for the data sets of the non-linearly separable *ball model*. The inset specifies the relative areas at different levels of accuracy.



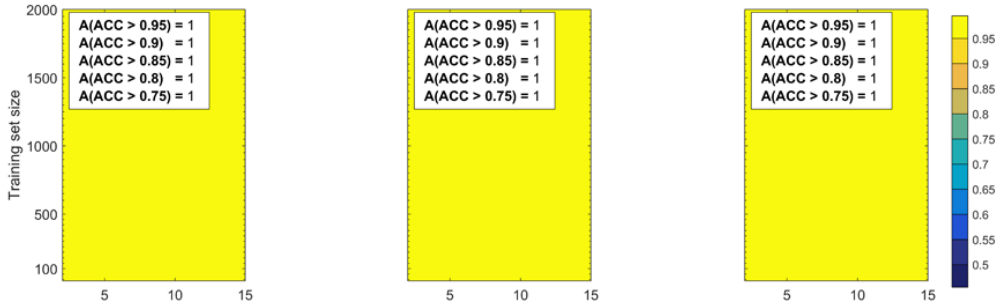
(a) kNN, $\eta = 0.05$ **(b)** RF, $\eta = 0.05$ **(c)** PLS-DA, $\eta = 0.05$

(d) kNN, $\eta = 0.1$ **(e)** RF, $\eta = 0.1$ **(f)** PLS-DA, $\eta = 0.1$

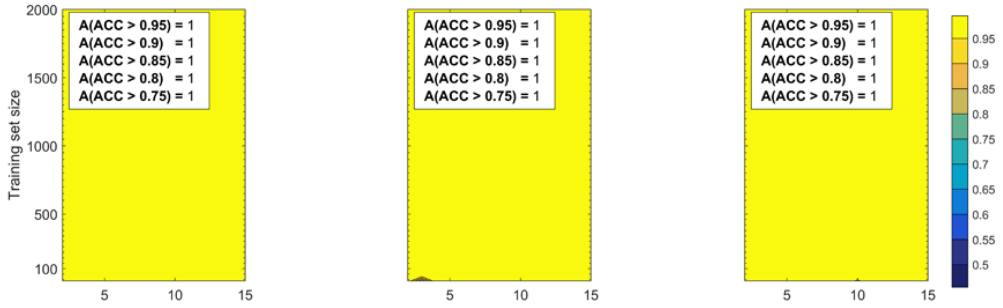
(g) kNN, $\eta = 0.2$ **(h)** RF, $\eta = 0.2$ **(i)** PLS-DA, $\eta = 0.2$

(j) kNN, $\eta = 0.3$ **(k)** RF, $\eta = 0.3$ **(l)** PLS-DA, $\eta = 0.3$

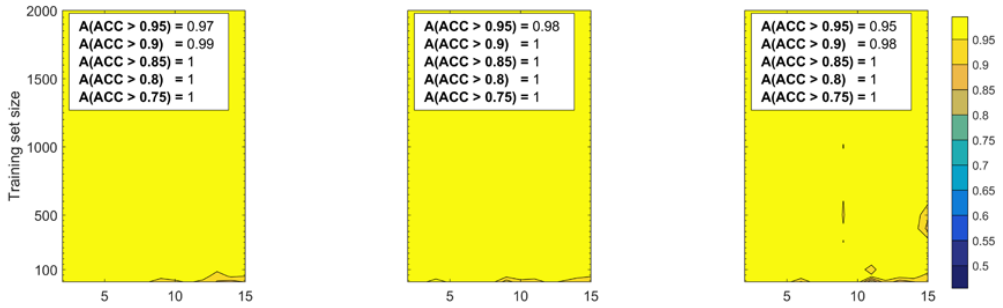
Figure 3.4: Classification accuracy for the data sets of the non-linearly separable *ray model*. The inset specificies the relative areas at different levels of accuracy.



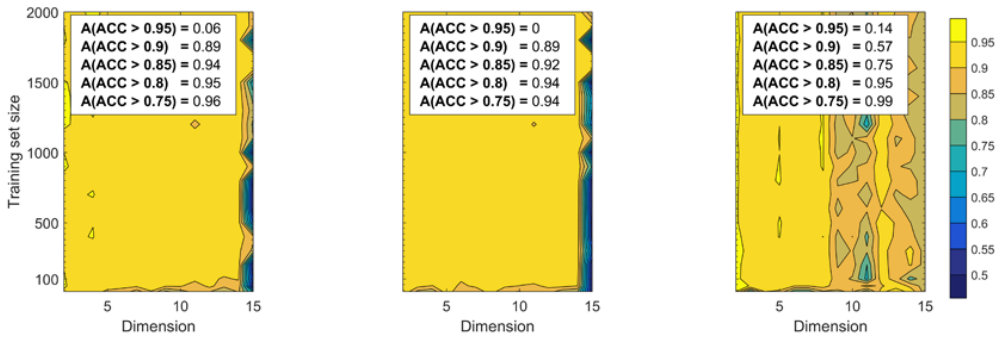
(a) kNN, $\eta = 0.05$ **(b)** RF, $\eta = 0.05$ **(c)** PLS-DA, $\eta = 0.05$



(d) kNN, $\eta = 0.1$ **(e)** RF, $\eta = 0.1$ **(f)** PLS-DA, $\eta = 0.1$

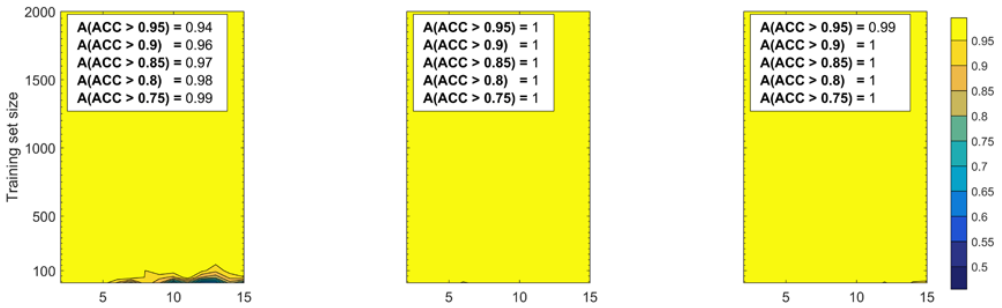


(g) kNN, $\eta = 0.2$ **(h)** RF, $\eta = 0.2$ **(i)** PLS-DA, $\eta = 0.2$

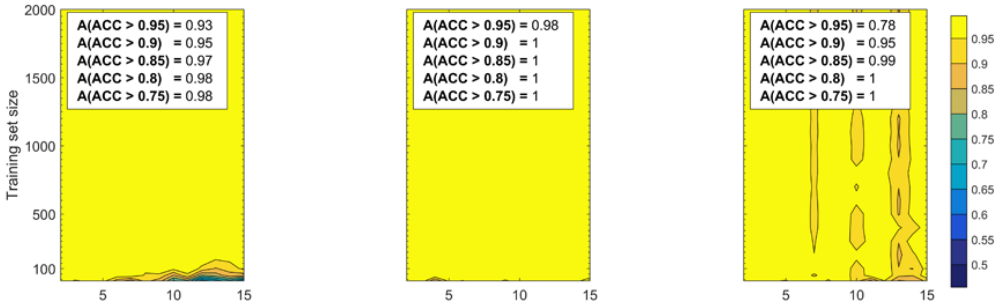


(j) kNN, $\eta = 0.3$ **(k)** RF, $\eta = 0.3$ **(l)** PLS-DA, $\eta = 0.3$

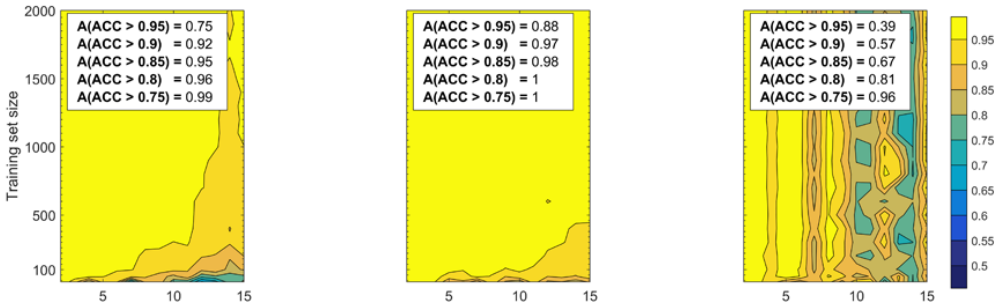
Figure 3.5: Classification accuracy for the data sets of the linearly separable *ray model* B.



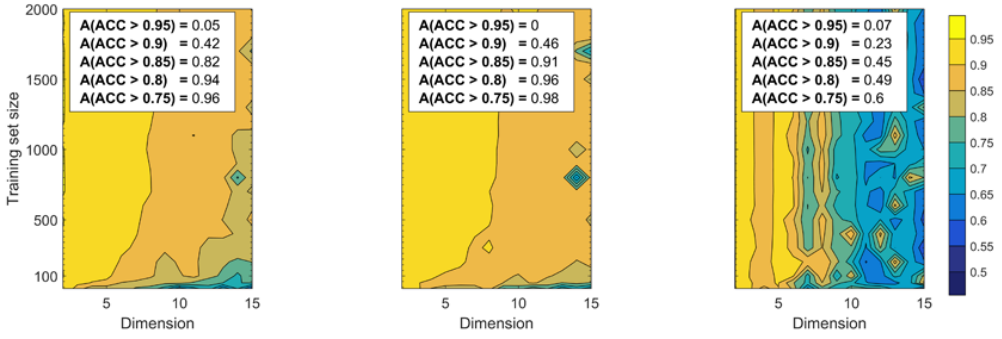
(a) kNN, $\eta = 0.05$ **(b)** RF, $\eta = 0.05$ **(c)** PLS-DA, $\eta = 0.05$



(d) kNN, $\eta = 0.1$ **(e)** RF, $\eta = 0.1$ **(f)** PLS-DA, $\eta = 0.1$

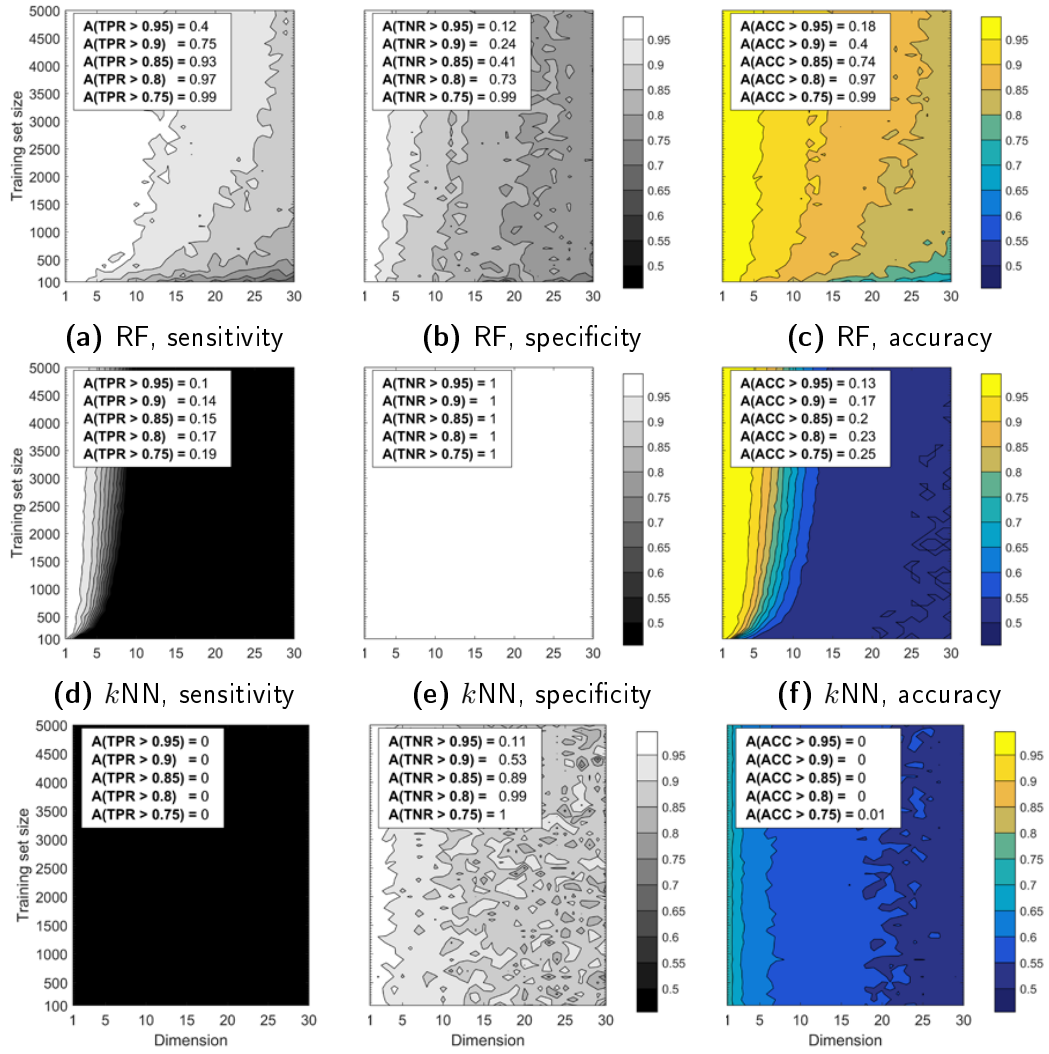


(g) kNN, $\eta = 0.2$ **(h)** kNN, $\eta = 0.2$ **(i)** kNN, $\eta = 0.2$



(j) kNN, $\eta = 0.3$ **(k)** RF, $\eta = 0.3$ **(l)** PLS-DA, $\eta = 0.3$

Figure 3.6: Classification accuracy for the data sets of the linearly separable *ray model* A. The inset specificies the relative areas at different levels of accuracy.

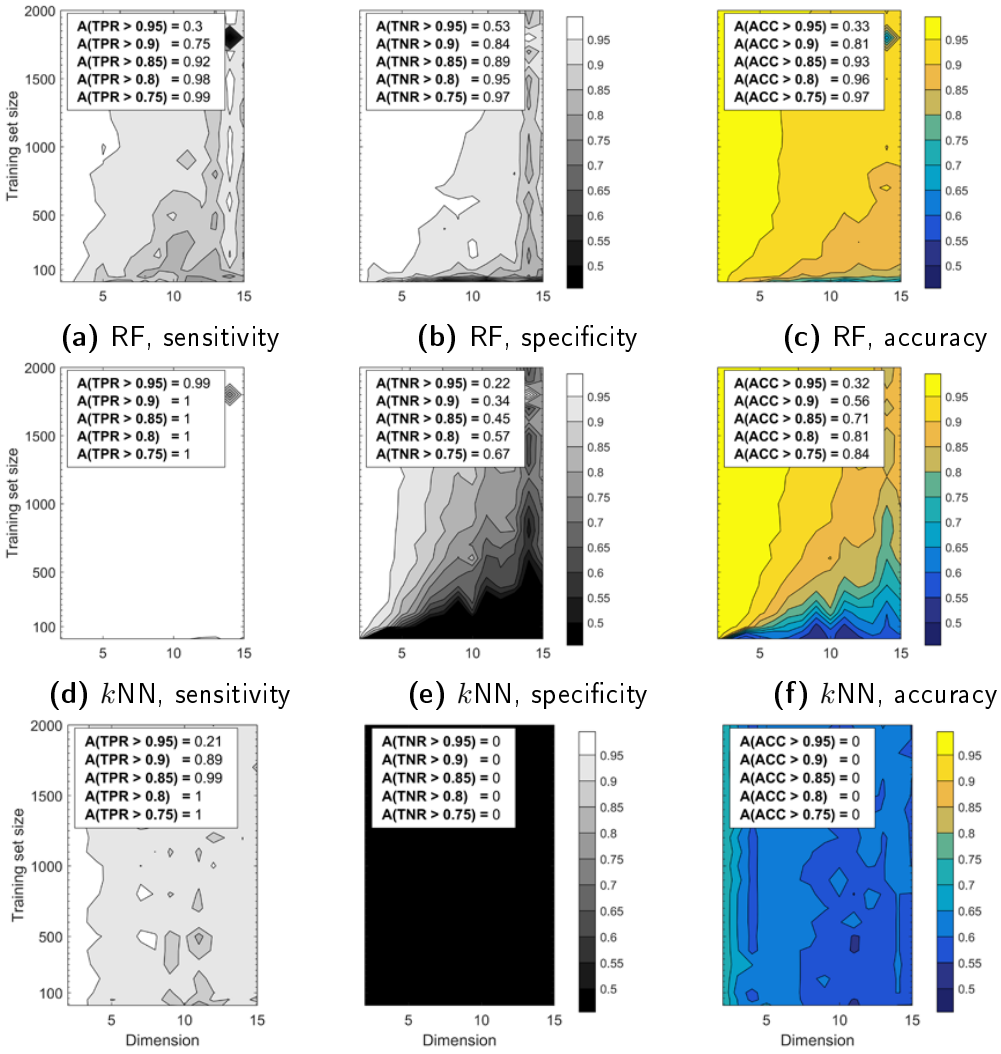


(a) RF, sensitivity **(b)** RF, specificity **(c)** RF, accuracy

(d) kNN, sensitivity **(e)** kNN, specificity **(f)** kNN, accuracy

(g) PLS-DA, sensitivity **(h)** PLS-DA, specificity **(i)** PLS-DA, accuracy

Figure 3.7: Comparison of RF, kNN and PLS-DA in terms of sensitivity, specificity and accuracy taking the non-linearly separable *ball model*, $\eta = 0.2$ as an example. The inset specificies the relative areas at different levels of sensitivity/ specificity/ accuracy.



(a) RF, sensitivity **(b)** RF, specificity **(c)** RF, accuracy

(d) kNN, sensitivity **(e)** kNN, specificity **(f)** kNN, accuracy

(g) PLS-DA, sensitivity **(h)** PLS-DA, specificity **(i)** PLS-DA, accuracy

Figure 3.8: Comparison of RF, kNN and PLS-DA in terms of sensitivity, specificity and accuracy taking the non-linearly separable *ray model*, $\eta = 0.2$ as an example. The inset specifies the relative areas at different levels of sensitivity/ specificity/ accuracy.

Part II

Infrared Spectral Histopathology as a Tool for Malignant Melanoma Diagnosis

Chapter 4

Theoretical Background on Infrared Hyperspectral Imaging and Malignant Melanoma

4.1 Cutaneous Malignant Melanoma

Malignant melanoma (MM) is a malignancy from melanocytes, which are cells of the epidermis. Although it only contributes to a small percentage of all skin cancer cases, it accounts for most skin cancer deaths as it is the most aggressive one[8]. Besides from the skin, malignant melanoma is rarely found in all other organs in which melanocytes are located, e.g. eyes, mouth or intestines[30]. However, in this thesis only malignant melanoma of the skin (*cutaneous melanoma*) is examined, which is in the following referred to as malignant melanoma (MM).

In 2012 approximately 100 000 Europeans were diagnosed with MM, accounting for 3 % of all new cancer cases that year. Furthermore, MM was responsible for approximately 22 200 deaths in the same year, which made up 1% of all cancer deaths in Europe. Worldwide, the countries with the highest recorded rates are Australia and New Zealand[31].

To enable understanding and interpretation of the classification procedure and results in Ch. 5 and Ch. 6, the following section focuses on the histology of the human skin as well as the changes brought by progression of malignant melanoma. In addition the basics of etiology and diagnosis of MM are mentioned briefly.

[8] The three major types of human skin cancer are malignant melanoma, basal cell carcinoma and squamous cell carcinoma. Malignant melanoma is the most lethal and thus requires a lot of attention.

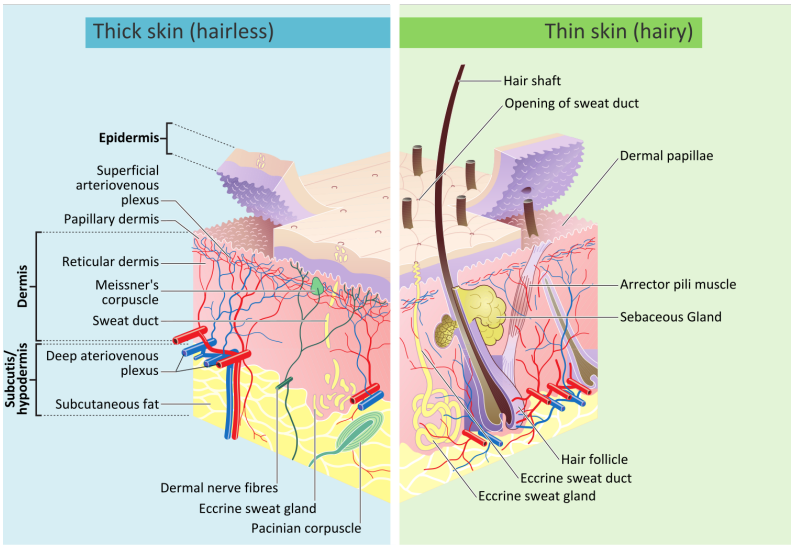


Figure 4.1: Model of the cutaneous covering for both, non-hairy and hairy skin [33]

4.1.1 Histology of the Human Skin

The human cutaneous covering is part of the integumentary system[9] and consist of the skin (cutis) and the subcutaneous tissue (subcutis), which is located below the skin. The skin is further divided into two layers, the epidermis and the dermis. The epidermis is a multilayered keratinized squamous epithelium and is the outermost layer of the cutaneous covering. The dermis, a dense irregular connective tissue which mainly consists of collagen and elastin, is located between the epidermis and the subcutis. The dermis is strongly connected to the subcutaneous tissue[32].

The epidermis is further divided into the horny layer (stratum corneum), granular layer (stratum granulosum), spineous layer (stratum spinosum) and basal layer (stratum basale). The thin basal membrane is located directly below the basal layer and separates epidermis and dermis distinctly (dermal-epidermal junction).

The epidermis mainly consists of keratinozytes, which proliferate continuously in the lowest parts of the epidermis, the basal and spineous layer. After the mitosis one daughter cell migrates to the upper epidermal layers while the other one stays in the lower layers and undergoes a further cell division. As they move outwards progressive differentiation takes place, which results in the distinguishable layers of the epidermis. This differentiation is reflected by increasing keratinization of the cells until they are shedded as dead cells from the stratum corneum[34, 35].

[9] The organ system that acts as a barrier to protect the body from damage, such as UV-radiation, bacteria, viruses, chemicals and other pathogens, is called integumentary system. It consists of the skin, subcutaneous tissue, assorted glands, nails and hair.

Other cell populations, that occur within the epidermis are mainly melanocytes, Langerhans' cells and Merkel cells[10][32].

Melanocytes are rather large cells and are located in the stratum basale. They synthesize *melanin* from tyrosin, the pigment which is responsible for skin and hair color. Melanin production is carried out in melanosoms, organells which are related with lysosomes. Subsequently melanin is stored in the melanosoms and passed to neighboring keratinocytes, respectively. The compound of one melanocyte and about 35 neighboring keratinoctes is labeled as *epidermal melanin unit* [36]. Melanin absorbs solar radiation and thus acts as a protecting filter. Enhanced sun exposure provokes melanin production.

The dermis is responsible for the tear strength and the plasticity of the skin and mainly contains connective tissue (CT). It is based on a matrix in which polysaccharides and proteins (collagen and elastin) are linked to macromolecules. In contrast to the epidermis, the dermis contains blood and lymphatic vessels. Hair follicles and sebaceous glands (glandula sebacea) are located in the dermis.[32].

The subcutaneous tissue (subcutis), also called hypodermis, establishes the connection of the skin to the superficial fascia of the body and consists mainly of adipose tissue, which is held together by connective tissue.

4.1.2 Etiology and Pathology of Malignant Melanoma

Abnormally increased proliferation of melanocytes can be malignant or benign. Benign lesions are called *melanocytic nevi*[11], well known as *moles*, and are mostly already present at birth. In many cases it is difficult for the dermatologist to distinguish a malignant melanoma from a benign nevus. However, some nevi are potential precursors to melanoma and have to be monitored[37]. Melanomas which do not arise from pre-existing nevi are referred to as *de novo*.

Etiology: As epidermal cells are the outermost layer of the human body they are strongly exposed to pathogens and thus prone to gene mutations. A high risk factor is excessive UV-exposure due to ionization of cellular molecules and subsequent damage of the DNA. There is a higher incidence in white people with blond or red hair, freckles and poorly tanning skin. MM is rare in black, asian or orientalic

[10] Langerhans' cells belong to the specific immune-reaction system and are initially not differentiated. After contact to an antigen differentiation takes place and the resulting dendrite cell is further presented to T-lymphocytes in lymph nodes. Merkel cells are mechanoreceptors (sensory cells) and are located in the basal cell and in hair follicles, especially in sensitive areas of the skin[32].

[11] A lesion with a local excess of one or more cell types of the skin is called *nevus*

people and before puberty, respectively. Furthermore, 10 to 15 % of melanomas are familial[38].

Diagnostics: The *'ABCDE'* rule is applied to decide whether a lesion is potentially harmful and has to be examined closer by a histopathologist. In such a case a skin biopsy is taken for further examinations. This rule summarizes possible signs for melanoma based on the appearance of the lesion. Suspicious features are **a**symmetrical shape, **b**order irregularity, **c**olor variability, **d**iameter greater than 6 mm and **e**volution of the lesion. Further symptoms that are not covered by the 'ABCDE' rule are e.g. itching or bleeding of the lesion[39].

Types: There are different forms of malignant melanoma[38, 39]:

- **Superficial spreading melanomas** are common on lower limbs of young/middle-aged adults and related to intermittent high-intensity UV-radiation
- **Nodular melanomas** appear with no prior *in situ* phase and are related to intermittent high-intensity UV-radiation,
- **Lentigo maligna melanoma** appears mostly on exposed skin of elderly and is related to long-term cumulative UV exposure,
- **Acral lentiginous melanomas** occur on palms, soles and nail beds. There is no indication to a relation with UV-exposure.

Pathophysiology and phases[30, 37] Various characteristics apply to all different types of MM. Firstly, all melanocytic lesions origin in the dermal-epidermal junction. Due to cell mutation cancer cells can emerge from melanocytes and subsequently proliferate. Nests of MM cells are created and expand horizontally at the beginning. Later they grow downwards into the dermis and deeper layers.

Furthermore, MM cells tend to be larger than melanocytes and nevus cells[12] with a larger, pleomorphic[13] nuclei. Also the cytoplasm is amophilic, meaning that it can be stained by both, acidic and basic dies. Melanin pigment can, but does not have to be present in the cancer cells. If melanin is present the lesion is called *melanotic*, if it is absent *amelanotic melanoma*. Due to the immune response, inflammatory infiltrate is normally present in the tumor.

For the therapy regime and prognosis it is important to assess the progress and the metastatic potential of the tumor. The *radial growth phase (RGP)* and the *vertical*

[12] Nevus cells are a type of melanocyte, which are larger in size. Nevus cells are the main constituent of melanocytic nevi.
[13] varying in size, shape and staining properties

growth phase (VGP) is an attempt to distinguish between a precancer of MM and MM with metastatic potential.

MM in the RGP are restricted to the dermal-epidermal junction and are also called *melanoma in situ*. The thickness is smaller than 1 mm. As no blood or lymphatic vessels pass the dermal epidermal junction, no metastases can emerge in this phase. The melanoma can be removed completely by surgery. Thus the RGP is considered a precancer.

In the VGP the cancer starts invading the dermis and continuously grows into deeper layers, in all of which there are blood and lymphatic vessels. Cancer cells can enter the bloodstream or the lymphatic system and secondary tumors (metastasis) can be formed in another organ. The risk for metastasis becomes higher with deeper penetration.

To sum up, the depth of the cancer is an important factor for the disease management. Two metrics have been developed to measure the depth. The *Breslow thickness* measures the distance from the granular layer to the deepest reaches of the tumor. In case of an ulcerated lesions, the base of the ulcer is taken as the top boundary.

Another widely used scheme are the *Clark's levels*, which categorize the MM based on the deepest layer which is invaded by the cancer cells.

- **Level I:** Cancer cells are restricted to the the epidermis (all cancer cells are located above the basal membrane). This corresponds to the RGP (melanoma in situ).
- **Level II:** The melanoma starts growing into the papillary dermis (upper part of the dermis)
- **Level III:** The tumor cells reach the border between papillary and reticular dermis.
- **Level IV:** The reticular dermis is invaded by tumor cells
- **Level V:** The subcutaneous fat is invaded by tumor cells.

Staging: The severity of the cancer is described by staging systems. Mostly the TNM-system (Europe) or the AJCC-system (American Joint Committee on Cancer, USA) is used. A major factor for the staging is the tumor thickness. An explanation of staging systems is required to understand the histopathological diagnosis of the used samples and is provided in Appendix 6.3.

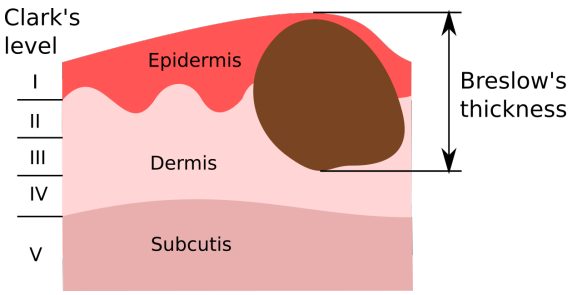


Figure 4.2: Visualisation of Clark's levels and Breslow thickness, in analogy to [38]

4.2 Fourier Transform Infrared (FT-IR) Imaging

Infrared spectroscopy is a commonly used analytic technique for non-destructive chemical analysis of various samples. Information about the structural properties of a sample can be derived based on its absorption of infrared radiation.

4.2.1 Theory of Infrared Spectroscopy

Infrared radiation is electromagnetic radiation with *wavelengths* between $\lambda = 0.77\,\mu\text{m}$ and $\lambda = 1000\,\mu\text{m}$, which corresponds according to Eq. (4.1) to *photon energies* between $E = 1.24\,\text{meV}$ and $E = 1.61\,\text{eV}$.

$$\begin{aligned} E &= h \cdot \nu \\ c &= \lambda \cdot \nu \\ \bar{\nu} &= \nu / c = 1/\lambda \end{aligned} \tag{4.1}$$

ν is the *frequency* of the IR radiation, $\bar{\nu}$ the corresponding *wavenumber*, h is Planck's constant ($h = 6.626\,070\,04 \times 10^{-34}\,\text{m}^2\,\text{kg}\,\text{s}^{-1}$) and c the speed of light ($c = 299\,792\,458\,\text{m}\,\text{s}^{-1}$). Eq. (4.1) shows that wavenumber, wavelength, frequency and photon energy all depend on each other. As a rule, in IR spectroscopy the wavenumber $\bar{\nu}$ with the unit $[\bar{\nu}] = \text{cm}^{-1}$ is used to refer to the photon energy.

IR radiation is divided into three ranges depending on the wavenumber, near-IR, mid-IR and far-IR (Tbl. 4.3). In biological applications one is interested and conducts measurements in the mid-IR region.

The mid-IR region covers the energy range of *molecular vibrational* states in biological molecules. Hence, infrared radiation is able to excite molecular vibrations either by absorption or by inelastic scattering. IR spectroscopy exploits the effect of absorption.

Table 4.1: Infrared regions and there wavelength/wavenumber intervals. [40]

Nomenclature	Wavenumber	Wavelength
far-IR	$\bar{\nu} \epsilon [10, 400]$ cm^{-1}	$\lambda \epsilon [1000, 25]$ µm
mid-IR	$\bar{\nu} \epsilon [400, 4000]$ cm^{-1}	$\lambda \epsilon [25, 2.5]$ µm
near-IR	$\bar{\nu} \epsilon [4000, 13\,000]$ cm^{-1}	$\lambda \epsilon [2.5, 0.77]$ µm

The molecular vibrational states are quantum states with molecule specific energy levels E_i. By absorbing a photon with the exact energy of the specific quantum transition

$$h\nu = |\Delta E| = |E_i - E_j| \tag{4.2}$$

the molecule can be excited to a higher energy level E_j. Based on the energy of the absorbed photon, conclusions about the absorbing molecule and the specific vibrational states can be drawn (functional groups, inter- and intramolecular interactions).

Molecular vibration is characterized by a periodic motion of the atoms in the molecule independent of the translational and rotational motion of the whole molecule. The vibrational motion can be approximated by the model of a *quantum harmonic oscillator*. The *eigenfrequencies* and therewith the energy levels of the harmonic oscillator depend on the atomic masses and the bond strengths[41].

This model implies that each vibrational motion can be decomposed into independent *normal modes*[14]. The number of normal modes depends on the number N of atoms in the molecule. A molecule with N atoms has $3N - 6$ normal modes. To describe the position of each atom in the molecule, 3 coordinates are required. The whole molecule has $3N$ degrees of freedom as it takes $3N$ coordinates to describe its spatial configuration. However, 3 of those degrees of freedom correspond to the translational and rotational information of the whole molecule, respectively, and have to be subtracted. For a non-linear molecule the resulting number of degrees of freedom (number of normal modes) is

$$3N - 6 \tag{4.3}$$

A linear molecule is invariant to rotation about the axis in the molecule. Thus, the number of normal modes in a linear molecule is

[14] orthogonal, eigenstates of the harmonic osciallator; corresponds to standing waves

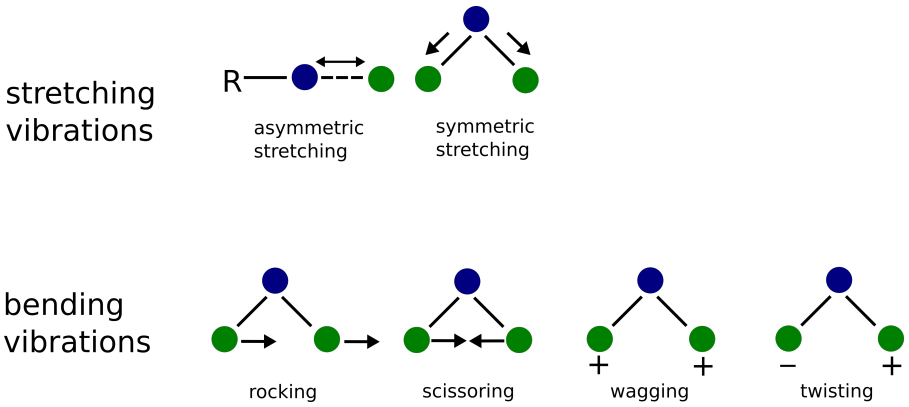


Figure 4.3: Different modes of molecular vibrations

$$3N - 5 \tag{4.4}$$

Based on their characteristic the normal modes are organized into *stretching* (change in bond length) or *bending* (change in bond angle) vibrations. Stretching can be either *symmetrical* (in-phase stretching) of *asymmetrical* (out-of-phase). Bending vibrations are further divided into *scissoring*, *wagging*, *rocking* and *twisting*[40, 42].

Furthermore, it is important to know, that one selection rule for allowed vibrational state transitions is a non-zero change in the electric dipole moment. This rule prohibits certain excitations, such as the symmetric CO_2 stretching.

To sum up, IR radiation with a certain wavenumber gets absorbed by certain molecules, which are excited into higher vibrational states. If the sample is irradiated with IR radiation of wavenumber $\bar{\nu}$ and known intensity $I(\bar{\nu}, 0)$, the absorbance can be quantified by measuring the intensity of transmitted radiation $I(\bar{\nu})$. *Transmittance* $T(\bar{\nu})$ is defined as the ratio of transmitted and initial intensity:

$$T(\bar{\nu}) = \frac{I(\bar{\nu})}{I(\bar{\nu}, 0)} \tag{4.5}$$

The *absorbance* $A(\bar{\nu})$ is defined as

$$A(\bar{\nu}) = \log_{10} \frac{1}{T(\bar{\nu})} = \log_{10} \frac{I(\bar{\nu}, 0)}{I(\bar{\nu})} \tag{4.6}$$

The Beer-Lambert law (Eq. (4.7)) states that the absorbance $A(\bar{\nu})$ is proportional to the sample thickness ξ (pathlength), the concentration c of the substance and

the wavenumber dependent absorptivity $\kappa(\bar{\nu})$, which contains the information about the excitable vibrational states.

$$A(\bar{\nu}) = \kappa(\bar{\nu})\xi c \tag{4.7}$$

4.2.2 FT-IR Microscopy

As already mentioned IR spectroscopy measures the wavenumber dependent absorptivity of a sample. For this purpose, first a background spectrum is acquired to estimate the initial intensity $I(\bar{\nu}, 0)$ of the beam without absorption in a sample. Subsequently the beam intensity $I(\bar{\nu})$ after interaction with the sample is measured and the ratio is computed.

The first IR spectrometers included a dispersive element and a single-element detector. The dispersive element (e.g. prism or grating) is turned in order to scan over all wavenumbers. As this procedure is time consuming, today normally *Fourier transform infrared (FTIR) spectrometers* are used instead. Those are based on a Michelson-Morley interferometer. The beam of the polychromatic light source is split into two beams by a semi-permeable mirror (beamsplitter). One beam is reflected by a fixed mirror (fixed pathlength), the other one is reflected by a periodically moving mirror (varying pathlength). The two beams are recombined and the resulting wavenumber dependent interferogram is acquired[43].

FTIR-microscopy is the combination of IR spectroscopy and microscopy. In contrary to conventional FTIR-spectrometers it enables focusing of the IR beam to the dimensions of the sample, which results in an increased *signal to noise ratio (SNR)* due to the higher photon throughput. Depending on the application different detectors are used, e.g. *single element detectors* or *focal plane array (FPA) detectors*. The latter consist of an array of $n \times m$ detector elements. Advances in detector technologies include the *microbolometer*, which does not require cooling, is less expensive than common detector technologies and enables real time imaging.

4.2.3 Spectral Characteristics of Biological Samples in Mid-IR

Biological samples consist of a mixture of numerous biomolecules and the spectra therefore exhibit overlapping bands. The resulting bands are specific to differences in protein, carbohydrate, lipid composition and DNA conformational changes[44].

Mostly it is impossible to assign a certain vibrational mode to a peak. However, the various bands are rather assigned to a certain functional group. In the following

only the fingerprint region is considered. The relevant bands for this study are listed in Tbl. 4.2 and illustrated in Fig. 4.4, which shows the acquired mean spectra of various tissue types in the analyzed tissue sections.

Normally lipid bands also occur in the fingerprint range but as most lipids are washed out during paraffin embedding (Sec. 4.2.4), the lipid bands are not mentioned in this section. Furthermore, bands which are overlapped by the paraffin bands are not mentioned either.

Characteristic for all biological samples are the amide peaks (A,B,I,...,IX), which are due to vibrations of the peptide group in the protein backbone. The most prominent band is the amide I peak (between 1620 and 1700 cm^{-1}). It is mainly assigned to C=O stretching, partly also to C-N stretching and N-H bending. The location of the peak maximum is specific to the secondary conformation of proteins. Furthermore, the amide II peak is also prominent in the fingerprint range. It has a lower intensity than the amide I peak and is largely due to C-N stretching and N-H in-plane bending.

Further, several bands are assigned to phosphate vibrational modes, which originate mainly in it phosphodiester groups of nucleic acids. An increase in those bands indicates an increase in nucleic acids, which can be found in malignant tissue[46].

As mentioned in Sec. 4.1.1, the protein collagen is the main constituent of the extracellular matrix (and hence the connective tissue in the cutis and subcutis). It is thus relevant for correctly classifying dermis and subcutaneous tissue as well as identifying strings of connective tissue within the lesion. Collagen consists of three polypeptid chains (with α-helix secondary structure), which form together a triple helix. There are more than 30 different types of collagen. The IR spectra of many collagen types were analyzed by Belbachir *et. al.*[45]. In figure Fig. 4.4 collagen bands can be seen in the spectra labeled *Connective Tissue A-C*.

Weak bands could also be assigned to the pigment melanin. However, not all cancer cells of malignant melanoma need to contain melanin. In order to obtain general results, which are not based on the melanin contribution, those bands are not considered in this study and are not listed in Tbl. 4.2.

4.2.4 Effects of Formalin Fixation and Paraffin Embedding

There are various methods to prepare excised tissue sections for microscopy and long-term storage[30, 51, 52]. All methods need to both, preserve the excised tissue from autolysis and harden it to enable fine sectioning with the microtom into thin slices ($3 - 8\,\mu\mathrm{m}$).

Table 4.2: Band assignments for measured spectra of different tissue types. A constant baseline of (a) $A = (n-1) \cdot 0.05$ and (b) $A = (n-1) \cdot 0.002$ is added to the n-th spectrum for improved visualization. The characteristic intensity and shape of the bands are described using abbreviations (vs: very strong, s: strong, m: medium, w: weak, vw: very waek/ vb: very broad, b: broad, sp: sharp, sh: shoulder)[45–50]

Label	$\bar{\nu}$, [cm^{-1}]	Vibrational mode	Molecule	Type
a	1620-1700	amide I,C=O stretching,N-H bending, C-H stretching	proteins	vs, vb
b	1520-1550	amide II , N-H bending, C-N stretching	proteins	s,b
c	1400	symmetric CH_3 bending	proteins	vw
d	1338	CH2 wagging vibrations from glycine backbone and proline sidechains	collagen	w,b
e	1280	CH_2 wagging vibrations from glycine backbone and proline sidechains	collagen	vw
f	1220-1280	amide III and PO_2^- asymmetric stretching, C-N stretching, N-H bending	proteins, DNA/RNA	w,b
g	1205	CH_2 wagging vibrations from glycine backbone and proline sidechains	collagen	vw, sh
h	1150	C-C and C-O stretching	proteins	w, b
i	1080	PO_2^- symmetric stretching	DNA/RNA	w, b
i	1080	vibrational modes of carbohydrate residues (e.g. C-O, C-C-O and C-C skeletal stretching)	proteins	w, b
i	1035	vibrational modes of carbohydrate residues (e.g. C-O, C-C-O and C-C skeletal stretching)	proteins	w, b

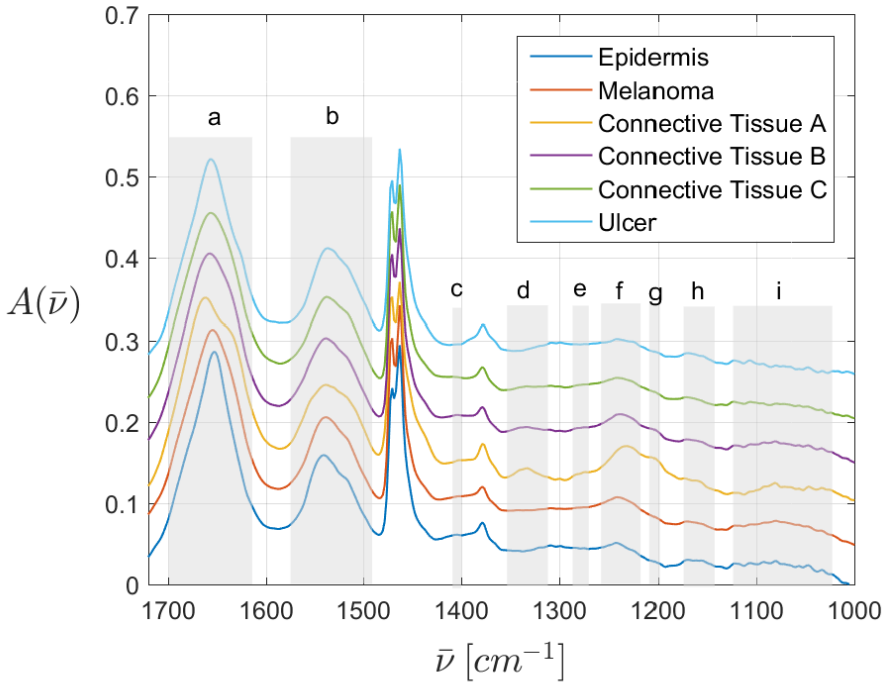


Figure 4.4: Acquired average spectra of selected tissue types. *Connective Tissue A-C* are three different types of connective tissue solely identified by spectral differences.

The two most common techniques are *formalin fixation and paraffin embedding (FFPE)* and *cryopreservation (snap freezing)* of the tissue, respectively. Both methods have advantages and disadvantages and are selected based on the specific purpose.

In case of cryopreservation the excised tissue is hardened by snap freezing in liquid nitrogen cooled isopentane (−160 °C) and sectioned. In case of light microscopy the tissue section is subsequently stained. The frozen tissue sections have to be stored. This method is often used to determine tumor margins during surgery, as the processing takes less time than formalin fixation and paraffin embedding. Furthermore, as no organic solvents that could cause loss of some cellular components are used, snap freezing is often preferred for molecular based studies. However, snap frozen sections have to be constantly frozen and thus storage is more complicated and expensive.

Fixatives are chemicals which inactivate lysosomal enzymes and inhibit the growth of molds and bacteria. Formalin fixation and paraffin embedding (FFPE) is based on first fixating the tissue in (most commonly) 10% formalin, which is equivalent to a 4% aqueous solution of formaledhyde[15]. Hydrated formalin cross-links the primary and secondary amine groups of proteins. However, selected lipids are preserved by reaction of formalin with the double bonds of the hydrocarbon chains of the lipids.

After formalin fixation the tissue section has to be hardened for sectioning, which is achieved by embedding it into molten paraffin wax. For this to happen, the tissue block has to be dehydrated first. Dehydration is carried out by consecutive immersion of the tissue in solutions of increasing alcohol concentration until 100% alcohol. Several molecular changes are induced by this step such as denaturation of the protein tertiary structure or significant precipitation of lipid molecules that are not preserved through the primary fixation step. Subsequently the tissue is permeated by molten paraffin wax. When the paraffin block is cooled down to room temperature, thin sections can be cut without destruction of cellular structure.

For light microscopic analysis the tissue sections are then deparaffinized to be subsequently stained. Most commonly *hematoxylin and eosin stain (H & E* is used.

As paraffin wax shows strong signals in the mid-IR region, FFPE tissue sections are often chemically deparaffinized for IR analysis as well. However, chemical dewaxing is time consuming, uses again chemicals which have potential effect on the tissue sections and does not always remove the paraffin completely. Thus, in several studies tissue is used without dewaxing, in which spectral regions affected by paraffin

[15] Formalin contains 40% w/w formaldehyde in water with the addition of 10% methanol.

Table 4.3: Assignment of paraffin bands in measured samples [42, 54]

Band position [cm^{-1}]	Remarks
1375	Methyl symmetrical C-H bending
1462	Methylene scissoring
1471	Methyl asymmetrical C-H bending
2848	Methylene symmetrical C-H stretching
2920	Methylene asymmetrical C-H stretching
2954	Methyl asymmetrical C-H stretching

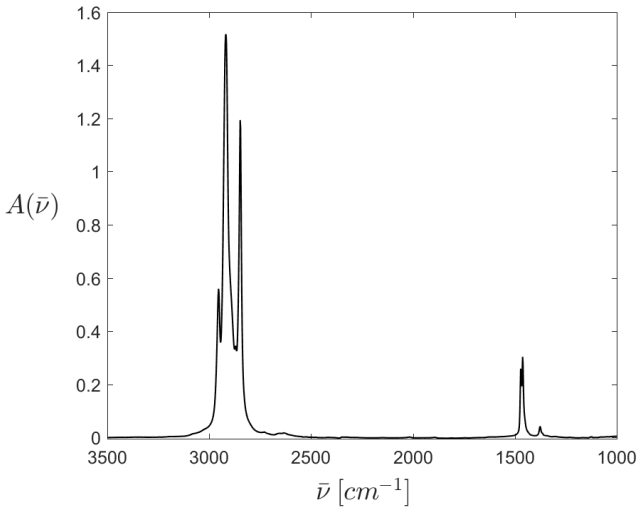


Figure 4.5: Mean spectrum of pure paraffin. Data taken from measuring pure paraffin regions of a FFPE embedded tissue section.

are either left out, or the paraffin contribution is neutralized by *Extendend Multiplicative Signal Correction (EMSC)* based algorithm or the spectra are digitally dewaxed by independent component analysis[53].

Here, FFPE tissue sections are used for IR imaging. To sum up, the main effects of the FFPE procedure on the spectra are the addition of the paraffin wax bands and the loss of the lipid bands. Fig. 4.5 shows the average of pure paraffin spectra, which were acquired at regions with no tissue present. The corresponding band assignments are listed in Tbl. 4.3.

4.2.5 Data Pre-Processing

Infrared spectra are distorted by undesirable effects, which should be corrected during pre-processing in order to guarantee optimal and reliable information output. There is no general procedure for pre-processing of IR images. Selected methods rather depend on the individual case (sample characteristics, data acquisition etc.). However, some concepts always have to be considered and can be divided into *noise reduction* and *spectral correction* methods. Denoising accounts for random errors (noise), while spectral correction accounts for systematic errors (baseline effects, scattering etc.)

Noise reduction: Spectra of IR images inhibit noise due to various effects (detector noise, electronic noise etc.), which is to be reduced prior to further pre-processing and analysis. In current SHP studies various noise removal tools are implemented, such as *Savitzky Golay smoothing, wavelet denoising, noise adjusted PCA (NAPC)* or *maximum noise fraction (MNF) transform*[55]. In this thesis MNF is featured[44, 56, 57].

Green *et al.* [57] introduced in 1988 the *Maximum Noise Fraction Transform (MNF)* for noise filtering of hyperspectral images. It assumes an additive noise model, i.e. each sample (pixel) spectrum can be decomposed into a signal and a noise term.

$$\begin{aligned} \mathbf{x_i} &= \mathbf{x_{s,i}} + \mathbf{x_{n,i}} \\ \mathbf{\Sigma} &= \mathbf{\Sigma_s} + \mathbf{\Sigma_n} \end{aligned} \tag{4.8}$$

In analogy to PCA, MNF is based on a linear transformation to a new basis. As a reminder, in PCA data are transformed to the eigenspace of the sample covariance matrix $\mathbf{\Sigma}$, resulting in features ordered by decreasing variance. MNF transforms the data to the eigenspace of $\mathbf{\Sigma_n}\mathbf{\Sigma}^{-1}$, resulting in features which are ordered by decreasing *noise fraction* rather than variance. The noise fraction of the i-th feature is defined as

$$\frac{var(\mathbf{x_{n,i}})}{var(\mathbf{x_{s,i}})} \tag{4.9}$$

The first k components (components with the highest noise fraction) are assumed to contain only noise and can be excluded for back transformation to the original data space. In order to compute the transformation, the signal and noise covariance matrices ($\mathbf{\Sigma_s}$ and $\mathbf{\Sigma_n}$) have to be estimated. For this purpose the fact is exploited, that at any region in the image the signal is strongly correlated, while the noise

only exhibits weak spatial correlation. It has to be emphasized that information about the noise structure of the dataset (salt & pepper noise/gaussian noise, vertical stripes, horizontal stripes, diagonal stripes, fringes,...) is required for computing $\mathbf{\Sigma_n}$. Thus, it is crucial to select the correct noise structure.

Spectral correction methods are routines to remove or neutralize adverse attributes of the spectra, such as baseline distortions (due to scattering, changing conditions during data acquisition, instrumental factors etc.) and spectral components due to an additional, disturbing compound (such as water vapor, paraffin, absorption by a supporting substrate, etc.)[55]. In this study spectra have to be corrected for two effects:

- **Paraffin contributions:** As described in Sec. 4.2.4, paraffin wax displays strong bands in the fingerprint region. If the samples are not chemically dewaxed the paraffin contribution has to be either neutralized (e.g. by using EMSC) or digitally removed[53, 58].
- **Resonant Mie scattering**: Morphological variations in biological tissues and cells result in strong scattering effects, which express themselves in baseline distortions as well as band maxima shifts (shifting to lower wavenumbers). It has been shown that the broad baseline oscillation can be explained by *Mie scattering*, the scattering of electromagnetic radiation on homogenous, spherical absorbing particles. The simplest model to approximate the scattered radiation $Q_{scatter}$ was introduced by Hendrik C. Van de Hulst (Eq. (4.10))[59–62]

$$\begin{aligned} Q_{scatter} &= 2 - \frac{4}{\rho}\sin\rho + \frac{4}{\rho^2}(1-\cos\rho) \\ \rho &= 2\pi d(n-1)\bar{\nu} \end{aligned} \tag{4.10}$$

  d is the particle diameter and $n = n_{spl}/n_{air}$ the ratio of the real refractive indices of the sample (n_{spl}) and the surrounding medium (air,$n_{air} = 1$). However, this model fails to explain the shift of the band maxima. Bassan *et. al.*[62] have shown that the latter can be explained by *resonant Mie scattering*, which considers a complex refractive index of the sample:

$$n(\bar{\nu}) = n_r(\bar{\nu}) - i \cdot \kappa(\bar{\nu}) \tag{4.11}$$

  As a result of considering the complex refractive index, maximal absorptivity leads to anomalous dispersion which is expressed in the peak maximum shift. The imaginary part (absorptivity $\kappa(\bar{\nu})$) is related to the real part $n_r(\bar{\nu})$ by

the Kramers-Kronig transform. Finally the real refractive index, which is required to compute the scatter curves in Eq. (4.10) can be written as

$$n_r(\bar{\nu}) = a + b \cdot n_{kk} \tag{4.12}$$

with the average refractive index a and the scaling parameter b and the result of the Kramers-Kronig transform n_{kk}. A detailed explanation of the physical background and the correction model can be found in Bassan *et. al.* [63].

In principle, pure baseline effects can be corrected by various methods, the most simple algorithms are based on polynomial or spline interpolation. In case of IR spectra it is common to choose 4th-order polynomials for base line approximation. Considering these baseline effects is crucial for valid analysis of the spectra, especially if the analysis is based only on intensity values at selected wavenumbers. However, if baseline corrected spectral descriptors are used baseline correction of the spectra is less crucial.

Another approach for spectral correction is the *Extended Multiplicative Signal Correction (EMSC)*[64, 65] which achieves both the neutralization of an additional compound as well as the correction of the baseline[65].

It basically assumes that each measured absorbance spectrum $A(\bar{\nu})$ can be approximated by the average spectrum $\bar{x}(\bar{\nu})$, an n-th order polynomial to explain baseline distortions and the residual $e(\bar{\nu})$

$$A(\bar{\nu}) = a\bar{x}(\bar{\nu}) + \sum_{i=0}^{n} b_i \bar{\nu}^i + e(\bar{\nu}) \tag{4.13}$$

In order to account for irrelevant effects specific orthogonal subspace models are introduced in Eq. (4.13), resulting in Eq. (4.14). Here, the spectra are corrected for paraffin wax contributions by incorporating the subspace $\sum_{i=1}^{n_P} c_i p(\bar{\nu})_i$ and for resonant Mie scattering by including $\sum_{i=1}^{n_R} d_i r(\bar{\nu})_i$.

$$A(\bar{\nu}) = a\bar{x}(\bar{\nu}) + \sum_{i=0}^{n} b_i \bar{\nu}^i + \sum_{i=1}^{n_P+1} c_i p(\bar{\nu})_i + \sum_{i=1}^{n_R+1} d_i r(\bar{\nu})_i + e(\bar{\nu}) \tag{4.14}$$

Subsequently for each spectrum the model parameters a, b, c and d are fitted my linear least squares regression and the corrected spectrum is computed by

$$A(\bar{\nu})_{corr} = \frac{1}{a}(A(\bar{\nu}) - \sum_{i=0}^{n} b_i \bar{\nu}^i - \sum_{i=1}^{n_P+1} c_i p(\bar{\nu})_i - \sum_{i=1}^{n_R+1} d_i r(\bar{\nu})_i) \tag{4.15}$$

The subspace models are created as followed:

- **Paraffin model:** Images with pure paraffin are measured and analyzed by principal component analysis. The first n_P eigenvectors and the average paraffin spectrum are to be included in the model.[53, 58, 66].
- **Resonant Mie scattering**: $Q_{scatter}$ curves are created for different d and n. Subsequently a principal component analysis is conducted on the resulting curves. The first n_R eigenvectors and the average scatter curve are to be included in the model[62][16].

It has to be emphasized that one advantage of EMSC is that it not only neutralizes the spectra from paraffin and corrects scattering effects, but also normalizes the spectra.

Normalization: If the spectra are corrected by other methods than EMSC, they usually have to be normalized to account for different pathlengths. Common approaches are the correction to an internal standard (specific wavenumber) or vector normalization and have been mentioned in Sec. 2.1. Infrared spectra of biological compounds are often normalized to the amide I peak (after baseline correction). For second derivative spectra vector normalization is commonly applied [40, 55]

4.3 Current Status of Using SHP for Diagnostics of Malignant Melanoma

Infrared microscopy emerged as a non-destructive, label free and sensitive method to analyze various materials. Due to low spatial resolution it could not be used for analyzing biological tissue on microscopic scale until the 1990s, though. However, recent developments in imaging technology and data processing have led to higher

[16] In case of the Mie scatter model this approach is straight forward and scatter curves with parameters $d = 2$ to $20\,\mu m$ and $n = 1$ to 1.5 are created.

In case of RMieS the parameters $d = 2$ to $20\,\mu m$, $a = 1.1$ to 1.5 and $b = 0$ to $(a-1)$ are varied to obtain different $Q_{scatter}$ curves.

However, resonant Mie scattering requires an absorption spectrum $A(\bar{\nu})$ (and thus $\kappa(\bar{\nu})$) to compute the corresponding real refractive index $n_r(\bar{\nu})$. This is crucial, as a non optimal reference spectrum will not adjust the spectra correctly. To account for this fact, after first correction with an average spectrum, each pixel is iteratively corrected with the processed spectrum as the new reference spectrum. With this approach, spectra can be corrected reliably even when the differs from the true pixel spectrum. However, the iterations of the RMieSEMSC algorithm require long computation times (several hours for a few iterations of average sized image)

resolution and faster image acquisition and have made IR imaging a powerful tool in *spectral histopathology (SHP)*. Detailed descriptions on current developments, studies and methods can be found in various review papers[44, 46, 47, 50, 60, 67–69].

Most studies on FTIR analysis of skin biopsies focus on cluster analysis as a chemometric tool, both on deparaffinized and paraffin embedded tissue. Studies on paraffin embedded tissue are conducted e.g. by Tfayli *et. al.* (discrimination of nevus and melanoma on paraffin-embedded skin biopsies using hierarchical cluster analysis (HCA) [48]), Sebiskveradze *et. al.* (description of an innovative fuzzy C-means (FCM)-based clustering algorithm, allowing the automatic and simultaneous estimation of the optimal FCM parameters[70]) and Ly *et. al.* (combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies [66]).

Belbachir *et. al.* analyzed and characterized different collagen types based on their spectra.[45]. Some studies focus on identifying diagnostic parameters and feature selection, respectively, as well as suitable computational or statistical methods [71, 72]. Various publications consider digital dewaxing methods on IR images and their effects[53, 73, 74].

All in all, the power of IR analysis for spectral histopathology is exhibited by all studies. However, some studies showed that in contrast to FTIR, Raman spectroscopy manages to distinguish the sublayers of the epidermis (on dewaxed tissue)[75].

Chapter 5

Methods for Data Acquisition and Tissue Classification

5.1 Sample Preparation and Characteristics

Tissue sections of five melanoma samples have been obtained from the *Department of Pathophysiology and Allergy Research (Center for Pathophysiology, Infectiology and Immunology, Medical University Vienna Vienna, Austria).* The samples are formalin fixed and paraffin embedded (FFPE) and mounted on a $CaFl_2$ sample carrier. The adjacent FFPE tissue sections are H & E stained in order to obtain labeled test and training data. Tbl. 5.1 summarizes the characteristics of the tissue sections.

5.2 Data Acquisition

The images were collected in transmission mode on a FTIR-microscope *Bruker Hyperion 3000* with a liquid nitrogen cooled 64×64 pixel FPA detector, featuring a sample area of 175×175µm. A 15-fold objective and 4×4 binning was used, resulting in a pixel size of $2.7 \cdot 4 = 10.8$µm. For each measurement position 4 scans were accumulated to enhance the S/N ratio. A background scan on the $CaFl_2$ slide was conducted every 20 scans using 32 accumulations. Spectra were acquired between $\bar{\nu} = 3845$ and $879\,\mathrm{cm}^{-1}$ with a spectral resolution of $2\,\mathrm{cm}^{-1}$. The sample chamber was purged with dry air to reduce spectral components of water vapor.

The measured image sections (3-4 sections for each sample) were selected according to the adjacent H & E stained sections and cover areas between approx. 1 and $10\,\mathrm{mm}^2$. The image acquisition took less than 1 h for smaller and 3 to 4 h for larger images. The resulting file sizes range from less than 500 MB to approx. 2 GB.

Table 5.1: Summarized histological diagnosis of the measured samples. Training data are taken from samples 1-3 only.

Idx.	Type	Breslow depth	TNM	Mitotic activity	Remarks
1	Nodular	5 mm	pT4b	$> 1\,\mathrm{mm}^{-1}$	Superficial ulceration
2	Nodular	15 mm	pT4a L0 V0 Nx Mx R0		Moderately pigmented, focally-arranged melanin pigment, pressure atrophy
3	Superficial spreading	2.1 mm	pT3a, pN2c, local R0	$> 1\,\mathrm{mm}^{-1}$	Solitary satellite lesion (pN2c)
4	Partly nodular / polypoid	4.05 mm	pT4b	$> 1\,\mathrm{mm}^{-1}$	Superficial ulceration
5	Superficial spreading with nodular component	2.95 mm	pTx	$> 1\,\mathrm{mm}^{-1}$	Superficial ulceration, irregularities in pigmentation and fibrosclerosis

5.3 Pre-processing

Data are processed using ImageLab (v.1.98, Epina GmbH, Pressbaum, Austria) and MATLAB 2015b (The MathWorks, Inc., Natick, Massachusetts, United States).

Most processing steps are based on the application of spectral descriptors, which have been introduced in Sec. 1.2. In the following, the spectral descriptors are referred to by the abbreviations which have been assigned in Sec. 1.2.

Selecting spectral region of interest: Even though the amide A peak (between 3300 and cm^{-1}) also inhibits interesting features (e.g. epidermis can be distinguished from dermis by the amide A peak shift due to collagen contributions) spectra are cut to the fingerprint region ($\bar{\nu} = 1800$ to $\bar{\nu} = 1000\,\mathrm{cm}^{-1}$) prior to any further pre-processing.

Maximum noise transform: MNF is chosen for noise removal, with a salt and pepper noise structure. The resulting MNF-components are manually analyzed and all components which enable visual identification of a tissue structure are se-

Table 5.2: Spectral descriptor for exclusion of bad pixel

Nb.	Spec.	Type	Reference $\bar{\nu}$ [cm-1]	Baseline	
				Reference $\bar{\nu}$ [cm-1]	Neighbors
DC001		TCI	1655	1714, 1591	5

lected for back transformation. In most cases, the selected components are those with the lowest noise fraction and the highest auto-correlation factor, respectively. Depending on the image, between 15 and 30 components are used for back transformation.

Exclusion of bad pixel: Next all pixel which do not exhibit a distinct contribution of a biological spectrum are assigned as background pixel. As a criterion for distinct tissue contributions the correlation of the amide I peak area with a positive triangle template peak is computed and multiplied by the signal area (spectral descriptor DC001; see descriptor specification in Tbl. 5.2). The resulting values are plotted in an intensity histogram. The logarithm of the intensity histogram counts is interpolated by a penalized spline in the interval from $DC001 = 0$ to $DC001 = 6$.

The first minimum of the resulting curve is estimated by a minimum search and is taken as a threshold. Fig. 5.1 illustrates this histogram and the interpolated curve for one tissue section.

This procedure excludes pixel

- with pure paraffin spectrum,
- with vanishingly low contribution of a biological spectrum (due to scattering or adipose tissue of the subcutis, as most lipids are washed out during sample preparation),
- of the empty $CaFl_2$ slide.

Selection of training, test and paraffin data: Subsequently regions for training and test data are defined. This is explicitly carried out before spectral correction by EMSC as the mean spectrum of the training data is used during the RMieS algorithm (see below). Due to the spectral and biomolecular characteristics six classes are defined, which are listed in Tbl. 5.3.

According to the H & E stains, which are adjacent to the measured tissue sections, several 10×10 pixel areas were selected from the tissue sections and assigned to a

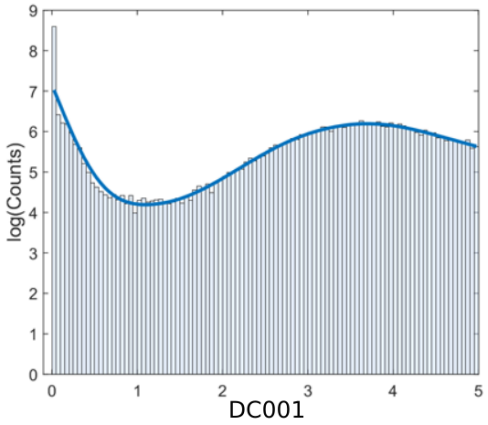


Figure 5.1: Logarithmic intensity histogram of D001 (descriptor to identify background pixel). The histogram counts are interpolated by a penalized spline (blue curve)

Table 5.3: Defined classes and the component that is expected to provide spectral contributions, which are class specific.

No.	Label	Expected class specific component	Color
1	Epidermis	Keratin	Dark blue
2	Malignant melanoma	DNA/RNA concentration	Red
3	Connective tissue A	Collagen	Yellow
4	Connective tissue B	Collagen	Purple
5	Connective tissue C	Collagen	Green
6	Ulceration	Erythrocytes	Light blue

certain class. However, it is important to emphasize that the stained tissue section and the IR measured section exhibit slight differences. On the one hand, because during sample preparation they are distorted in slightly different ways. On the other hand, the cellular features of different layers may also underlie variations as the sample thickness is 8 µm. Thus, it is difficult to overlay the images reliably. However, using the H & E stains together with chemical images of the (denoised) data helps to correctly choose training data.

Fig. 5.2 illustrates the process of training data selection on a region of S2, which exhibits melanoma, epidermis and two distinguishable types of connective tissue. A chemical image of the amide I peak intensity is useful to identify structures, which can be found on the stained sections. This procedure also allows the selection of training sets when the stained and the IR imaged sections are not directly adjacent (if further tissue sections were taken in between the one measured by IR and the stained tissue section).

Certain regions of samples S1 to S3 are selected and merged together as training data, leaving samples S4 and S5 as test sets only. This is important, as the aim is to develop a classifier which leads to reliable results on unknown tissue section of other patients. Between the patients there is always a slight variation of the cellular components, which must not make any difference in the classification performance. To test this requirement, the classifier is trained on selected spectra from S1-S3 and tested on other regions of S1-S3 as well as regions of S4 and S5.

For each class there are all together about 3000 pixels collected as potential training data. A random subset of 50, 200 and 400 pixels per class was selected as training sets for different classification scenarios.

As the training spectra will also be used as reference for the spectral correction (see below), it is important to ensure during the selection process that they are not dominated by scattering effects.

Signal correction using EMSC: The spectra are corrected by means of EMSC using the RMieS-EMSC algorithm with an additional orthogonal subspace model to account for the paraffin contribution. The developed script for spectral correction is based on the MATLAB EMSC Toolbox provided by Martens *et. al.*[64, 76, 77].

As baseline distortions are corrected by the subspace model of the scattering curves, no high order polynomials are required. Here, a polynomial of order 1 ($n = 1$ constant and linear baseline) is chosen for the polynomial correction.

To create the PCA model of paraffin, 5000 spectra were randomly selected on regions of pure paraffin of the training samples. Data are standardized prior to

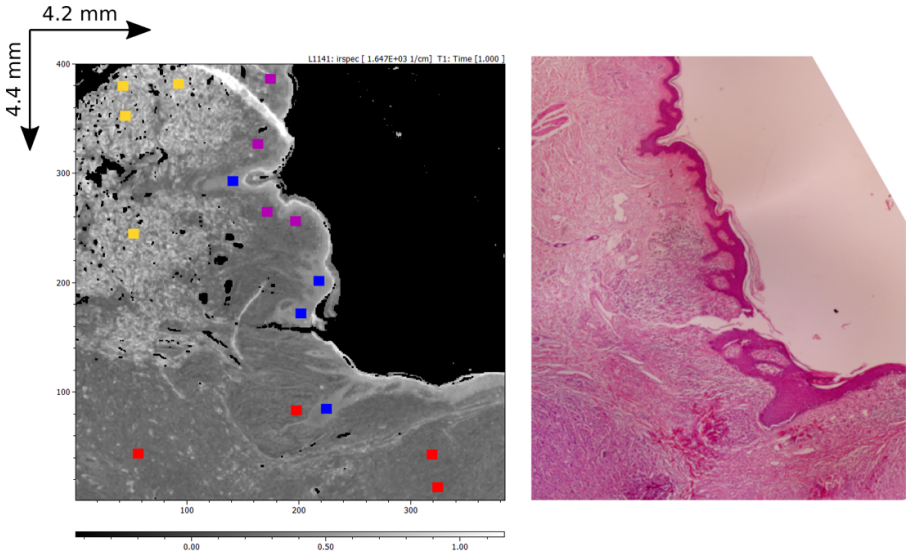


Figure 5.2: Example for training set assignment. Left) Chemical image of $\bar{\nu} = 1655\,\mathrm{cm}^{-1}$ (amide I) with assigned training sets: Blue: epidermis, red: melanoma, yellow: connective tissue A and purple: connective tissue B. Right) H & E stain of the adjacent tissue section.

PCA. The first three components of the paraffin model explain 97 % of the total variance and are selected for the subspace model, together with the mean paraffin spectrum. The components as well as the subspace model are smoothed by a moving average with a kernel size of 5.

The initial reference spectrum for computing the resonant Mie scatter subspace is retrieved from averaging the training data, as explained and reasoned below. The mean spectrum is smoothed by a moving average with a kernel size of 5. Scatter curves for different parameters ($d = 2$ to $20\,\mu\mathrm{m}$, $a = 1.1$ to 1.5 and $b = 0$ to $(a-1)$[63]) are computed using an algorithm for the Kramers Kronig transform based on Lucarini *el. al.*[78]. After the subsequent (standardized) PCA the first 6 principal components and the average scatter spectrum are chosen for the subspace model. The first 6 PCs explain 99 % of the variance.

As mentioned in Sec. 4.2.5, literature suggests to use the RMieS-EMSC correction iteratively, taking the corrected spectrum of a pixel as new references spectrum for a subsequent correction. It is stated that at least a few iterations are required, in order to completely correct the spectra independently of a non-optimal initial reference spectrum. However, the aim is to improve the classification process considering performance and required computation time. As the iteration cycles of the RMieS-EMSC algorithm require high computation time, an approach without iterations but an improved initial reference spectrum is attempted.

The mean spectrum is computed for the training samples of each class. The re-

sulting mean class spectra are again averaged to obtain a mean spectrum of all tissue types. This 2-step averaging process is chosen to ensure equal weights for all classes. If the mean spectrum is closer to one than to the other classes, features of this spectrum will be enhanced during correction. This is also the reason, why the average training data spectrum is used as a reference rather than the mean spectrum of the individual tissue section, as it cannot be ensured that all tissue types are equally frequent in each image. Furthermore, taking the mean spectrum of training data from several samples also enables correction of differences between patients, location of the lesion etc. A similar approach was used by Bird *et al.* for correcting spectra of human lymph node tissue[79].

Second Derivative: After the spectral correction, the second derivative of the spectra is computed and appended to the data cube. This allows to use both the descriptors of the original spectrum and the features of the second derivative spectrum for classification.

5.4 Classification

Based on the considerations and results in part I of this thesis, a *Random Forest* is selected to classify the tissue sections. It is attempted to use a low number of descriptors to avoid the curse of dimensionality and noise removal has been conducted on the data prior during pre-processing. Different classifiers featuring different amounts of training data (N_{pC} = 50, 200 or 400) are trained and tested subsequently.

The descriptor set is created by manually selecting spectral descriptors with the *Spectral Descriptor Tool* of ImageLab in order to exhibit intensity histograms which seem advantageous for class discrimination. 18 spectral descriptors are selected and listed in Tbl. 5.4, using the abbreviations which are introduced in Sec. 1.2.

However, it has to be emphasized that the selected descriptors are correlated. While this correlation is irrelevant for the performance of the RF classifier, it is important to be considered in other chemometric methods. To add this information, multicollinearities are analyzed by stating the variance inflation factor (VIF). This information is not specifically relevant for the RF classifier[17].

[17] High descriptor correlation does not imply that no complementary information can be found in the descriptors. While adding perfectly correlated variables to a model does not improve class separation, class separation can be enhanced due to noise reduction by including correlated variables in a model[80].

Parameter selection: For each classifier the optimum number of trees is estimated via cross validation using the out of bag error. The number of randomly selected descriptors per leaf is set to the square root of the number of descriptors ($\sqrt{d}$, with d being the dimension of the data space). For creating each split N_{train} samples are selected randomly with replacement.

5.5 Performance Estimation

As the stained tissue sections exhibit clear distortions from the measured samples and are not always directly adjacent it does not seem meaningful to specify performance measures such as accuracy, sensitivity and selectivity. Staining of the measured tissue sections subsequent to data acquisition was intended but the samples detached from the slides during the staining process.

The uncertainty in correctly choosing the test data could influence the performance measures considerably. Therefore, the results are interpreted visually and explained qualitatively.

Table 5.4: Selected spectral descriptors for classification

	Spec.	**Type**	**Reference $\bar{\nu}$ [cm-1]**	**Baseline**	
				Reference $\bar{\nu}$ [cm-1]	**Neighbors**
DC002		TCI	1354	1375..1333	0
DC003		ABL	1352..1294	*	0
DC004		TC	1335	1354..1321	5
DC005		PLV	1236	1267	5
DC006		RLV	1275/1294	1236	0
DC007		TC	1230	1298..1115	5
DC008		ARW	1142..1007	*	
DC009		PLV	1070	1132	0
DC010	2. dv	PRW	1674		
DC011	2. dv	TCI	1662	1680..1647	5
DC012	2. dv	ABL	1658..1687	*	0
DC013	2. dv	ABL	1662..1630	*	0
DC014	2. dv	PRW	1651		
DC015	2. dv	ABL	1628..1639	*	0
DC016	2. dv	PRW	1633		
DC017	2. dv	TC	1633	1630..1641	0
DC018	2. dv	TCI	1624	1618..1631	5
DC019	2. dv	BBL	$\frac{1500..1522}{1562..1523}$	*	0

Chapter 6

Results and Discussion of Skin Tissue Classification

6.1 Tissue Spectra

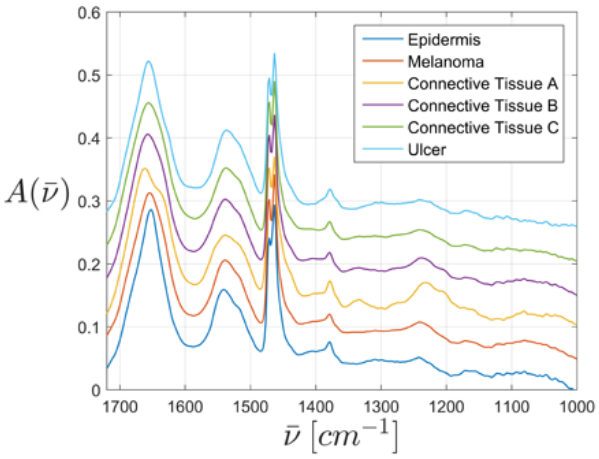
Fig. 6.1 illustrates the obtained average class spectra and their second derivative after spectral correction (paraffin neutralization and resonant Mie scatter correction using EMSC). The apparent spectral differences have been mentioned and explained in Sec. 4.2.3.

6.2 Properties of Spectral Descriptors and Random Forest Classifier

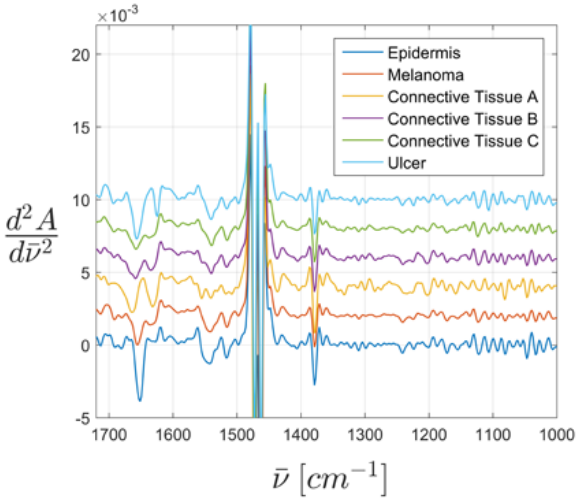
As mentioned in Sec. 5.4, the applied spectral descriptors are selected manually with the attempt to encode as much chemical information as possible.

Although this is not relevant for the Random Forest classifier, the generated descriptor sets are tested for multi-collinearities by means of the *variance inflation factor* (VIF). Consecutive deselection of the descriptors with the highest VIF results in descriptor sets B and C, which would be suitable for statistical methods, that required decorrelated data. Tbl. 6.1 summarizes the resulting VIF values for the individual descriptor sets and refers to the original descriptor set (18 descriptors) as set A. For descriptor set C, all VIF are less than 15 for all remaining variables. However, it has to be pointed out that all RF models in this study have been trained with the original descriptor set (Set A).

The used number of trees N_{tree} is chosen by analyzing the out of bag error rate (*OOB estimate*). Fig. 6.2 illustrates this OOB curve for RF models. In most cases $N_{tree} = 20$ is selected.



(a)



(b)

Figure 6.1: Average spectrum (a) and 2nd derivative spectrum (b) of the acquired training data for each class after spectral correction. A constant baseline of (a) $A = (n-1) \cdot 0.05$ and (b) $A = (n-1) \cdot 0.002$ is added to the n-th spectrum for improved visualization.

Table 6.1: Analysis of multi-collinearities by means of the variance inflation factor. (VIF)

DC:	**002**	**003**	**004**	**005**	**006**	**007**	**008**	**009**	**010**	**011**	**012**	**013**	**014**	**015**	**016**	**017**	**018**	**019**
A	82.8	18.2	7.9	19.2	189	110	156	67.0	12.6	14.5	46.2	12.9	100	75.5	7.2	17.4	31.8	88.2
B	13.6	17.1	11.7	29.4	13.7	x	x	6.4	14.8	6.3	9.20	103	89.9	x	x	x	5.2	11.5
C	13.4	11.9	10.1	x	13.2	x	8.2	2.7	13.8	5.6	x	x	15.3	x	x	x	4.3	10.2

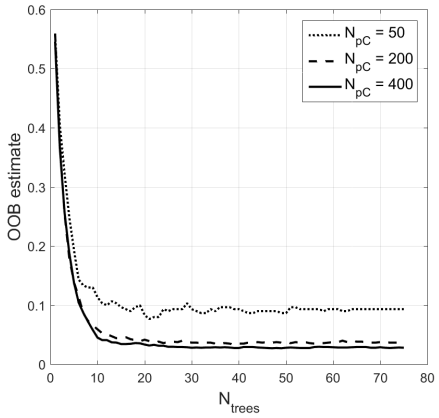
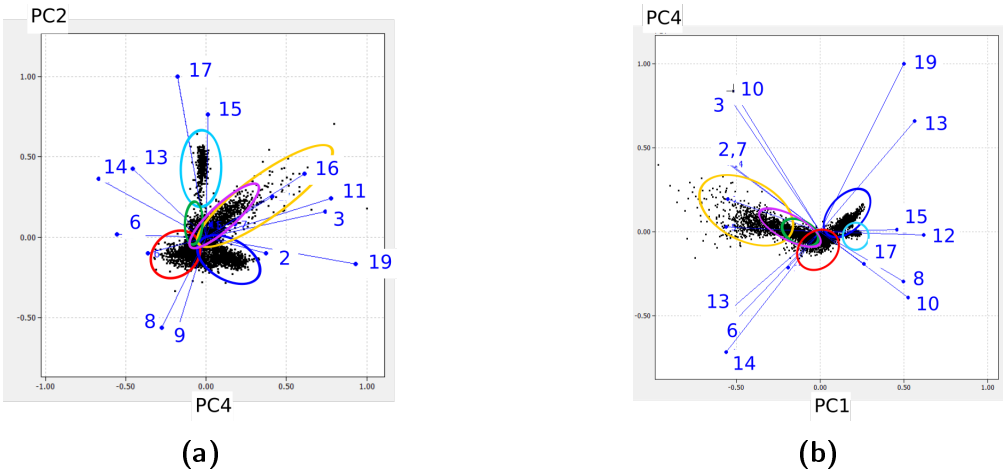


Figure 6.2: Typical *out of bag error estimate* curve for selection of N_{tree} illustrating errors for different training set sizes.

The descriptors are analyzed qualitatively by means of a PCA biplot (score and loading plot, Fig. 6.3). It becomes apparent, that in the created model melanoma is largely characterized by the fact, that it exhibits no extreme values for any descriptor. The effect of this will be discussed later.



(a) **(b)**

Figure 6.3: Biplot (scores and loadings) resulting from Principal Component Analysis of the training data. The descriptor labels correspond to the labels in Tbl. 5.4. The approximate regions in the score plots that are occupied by distinct classes are encircled (dark blue: epidermis, red: melanoma, yellow: connective tissue A, purple: connective tissue B, green: connective tissue C, light blue: ulceration tissue). (a) PC4 vs. PC 2; (b) PC1 vs. PC 4

6.3 Qualitative Assessment of Classified Tissue Sections

In this section, selected obtained digital stains for different classifiers are compared to corresponding H & E stains and qualitatively discussed in terms of performance. All results have been discussed with and approved by a dermatologist. The featured stains are selected to represent both, the strengths and the faults of the applied method.

It has to be mentioned again, that no quantitative performance estimation is conducted as no confident pixel assignment is possible between the two tissue sections. This is due to distortions between the IR measured and the H & E stained samples. Thus, the error in selecting a correct test set could vastly influence the obtained performance measure. This is also the reason for some differences between the H & E stains and the "digital" stains in the following figures.

It is suggested that if it is not possible to stain the tissue section after IR analysis, both adjacent tissue sections are H & E stained, enabling a subsequent pixel interpolation between the three obtained images. Furthermore, it should be ensured that there are no further tissue sections between the IR measured sample and the H & E stain.

In general, correct tissue assignment is achieved by all differently trained classifiers (all training data sizes and descriptor sets). The classifiers distinguish well between epidermis, melanoma, different connective tissues and ulceration. Especially the clear differentiation between strings of connective tissue within the lesion should to be pointed out. Furthermore, ulcerating tissue is detected with high assignment probability.

Correct tissue assignment is achieved for both, samples of which certain areas have been used as training data (S1, S2 and S3; Fig. 6.4, Fig. 6.5, Fig. 6.10 and Fig. 6.9) and samples of which no data have been taken for training (S4, S5; Fig. 6.6, Fig. 6.7 and Fig. 6.8).

It can be seen in all sections, that apart from melanoma tissue, also blood vessels and transitions between two tissues (especially the basal layer between epidermis and connective tissue) tend to be detected as melanoma (red pixel assignment, compare boxes in Fig. 6.9 and Fig. 6.10). Those false positive melanoma assignments can be explained by various effects.

Firstly, no training data and class is assigned to the walls of blood vessel (endothelium; muscle in case of arteries), thus those spectra cannot be assigned to any known class. The same applies to the basal membrane, the transition of epidermis and connective tissue. Obviously melanocytes are present in the basal membrane, which

could exhibit a similar spectrum as melanoma cells. However, due to the extent of the wrongly marked pixels and the fact, that false positive melanoma labeling can be obtained as well at the horny layer of the epidermis, those assignments are rather explained by the fact that the spectra at the transition become indistinct due to lateral resolution limits. Insufficient scatter correction has also been considered as a possible reason, but has been discarded as the original spectra before pre-processing do not show strong scatter effects for the respective pixels.

Such indistinct spectra are assigned to melanoma, as in contrast to other tissue no descriptor is a clear indicator for melanoma. Melanoma is rather recognized as the pixel with "values in between epidermis and connective tissue" for most descriptors. This is also illustrated well in the PCA biplots (Fig. 6.3). The problem is assumed to be solved when a descriptor can be found, which features a maximum or minimum value for melanoma. However, the bands which should be indicative for melanoma are either bands due to lipids, which have been washed out during FFPE tissue preparation, or peaks due to phosphate. Latter are mostly overlapped by protein bands (e.g. amide III), which are prominent in other tissue types as well. However, there is no doubt that a suitable descriptor can be estimated for FFPE prepared tissue sections.

Fig. 6.9 and Fig. 6.10 illustrate the tendency of false positive melanoma assignments for differently trained classifiers. Regions of interest are marked by the boxes. In Fig. 6.9 a tissue section of S3 with no present melanoma is classified. However, at the transition of epidermis (blue) to dermis (connective tissue: purple, green) several pixel are identified as melanoma (upper box). Furthermore, blood vessels are wrongly marked as melanoma (lower box). It is interesting that the false positive melanoma labeling decreases with increasing training set size.

Fig. 6.10 presents the obtained stains of the transition from melanoma to cutis and subcutis in S2, featuring the classifier with 400 training data per class. Because adipose tissue is predominant in the subcutis, most of it is masked during pre-processing as background (due to lipid loss during paraffin embedding). Obtained digital stains featuring different probability threshold values (0.5, 0.6 and 0.7) for positive class assignment are illustrated. It can be seen, that for a threshold of 0.7, most of the melanoma tissue is still classified as melanoma, while the false positive melanoma assignments (here: blood vessels) decrease. However, the general rate of non assignable pixels obviously increases.

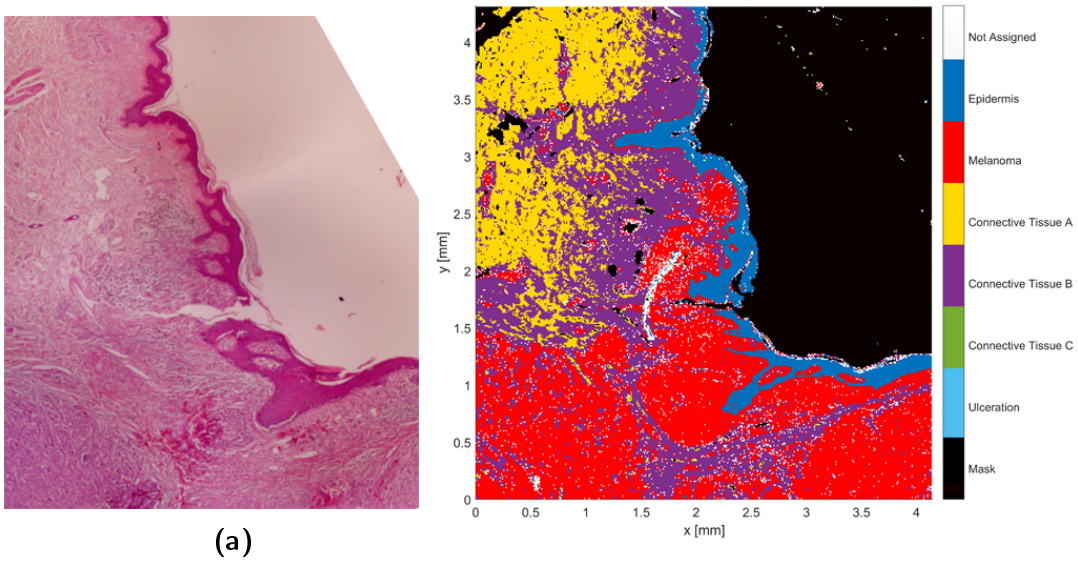


Figure 6.4: Sample 2 (included in training); epidermis, melanoma, different types of cutanous tisssue. $N_{pC} = 400$

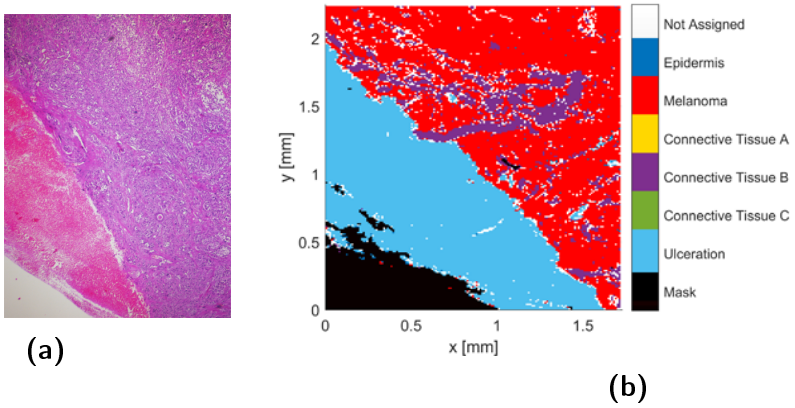
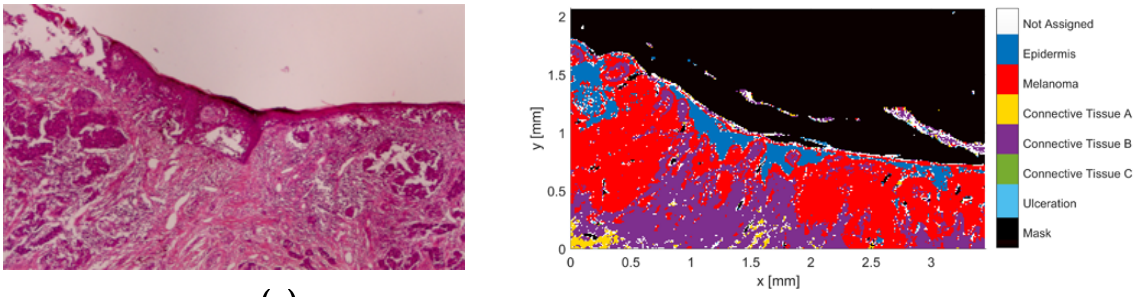
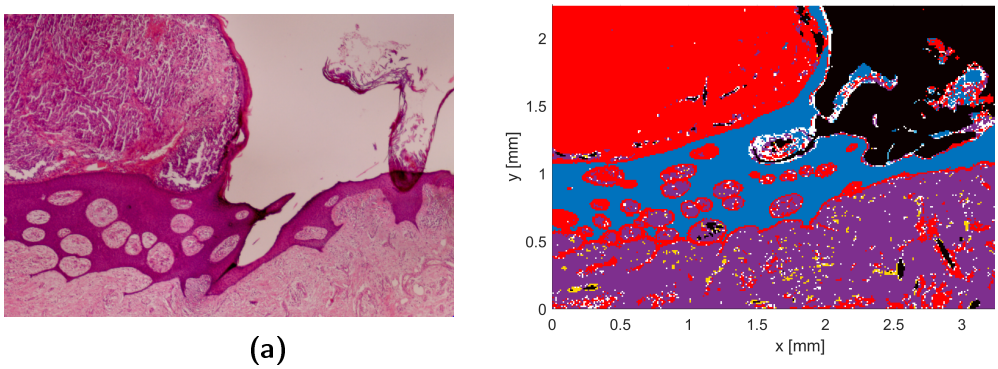


Figure 6.5: Sample 5 (included in training); melanoma, strings of connective tissue, ulceration. $N_{pC} = 400$



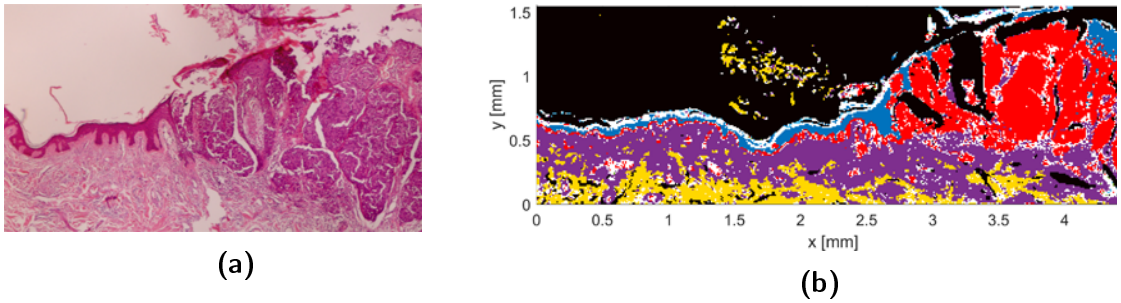
(a) (b)

Figure 6.6: Sample 7 (not included in training); melanoma, residuals of epidermis, connective tissue. $N_{pC} = 400$



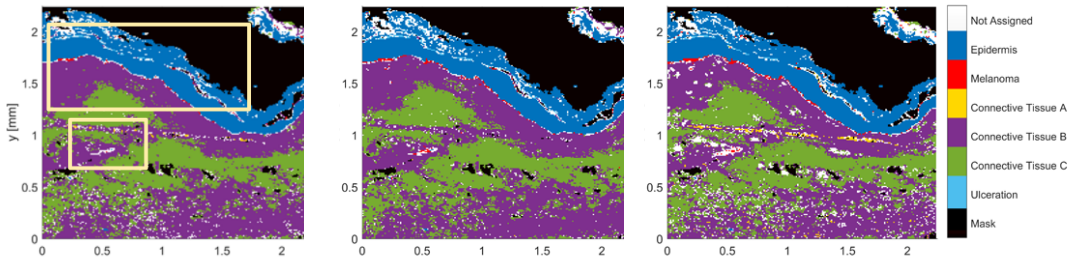
(a) (b)

Figure 6.7: Sample 6 (not included in training); melanoma, elongated epidermis, connective tissue with a large amount of blood vessels. $N_{pC} = 400$



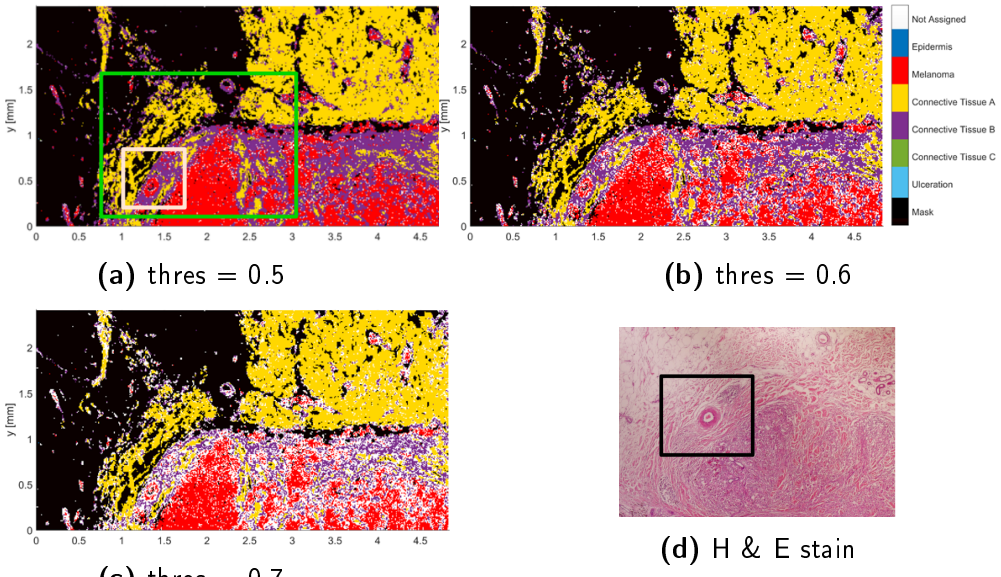
(a) (b)

Figure 6.8: Sample 7 (not included in training); epidermis, melanoma, connective tissue. $N_{pC} = 400$



(a) $N_{pC} = 400$ **(b)** $N_{pC} = 200$ **(c)** $N_{pC} = 50$

Figure 6.9: Sample 3 (included in training); epidermis, cutis and subcutanous tissue. Dependency of false positive melanoma assignments on size of the training data set. The boxes mark interesting regions for analyzing the false positive melanoma assignments (transition from epidermis to dermis and blood vessels) and are valid for all subfigures.



(a) thres = 0.5 **(b)** thres = 0.6

(c) thres = 0.7 **(d)** H & E stain

Figure 6.10: Sample 2 (included in training); transition from melanoma to subcutaneous tissue and cutis. $N_{pC} = 400$. Different threshold values (0.5,0.5,0.7) for the classification probability are featured. The white/black box indicates an intersting region for evaluation the false positive melanoma assignments (blood vessel). The green box indicates the respective sector of this image which can be seen in the H & E stain. The boxes are valid for all subfigures

Conclusion

In Part I the performance of classification algorithms on different artificial datasets has been demonstrated. The datasets are designed to represent spectroscopic data and are created according to two different models. Model 1 (ball model) does not include any information about concentration but the data generation is more simple and straightforward. Model 2 (ray model) is more complex, but contains information about compound concentrations. All models are further varied in class separability, noise level, dimension and size of the training dataset.

Applying the classifiers to datasets of the different models showed that after normalization model 2 leads to similar results as model 1. This justifies the use of the simpler model 1 for artificial dataset generation.

Consequences of the *curse of dimensionality* on the data distribution are reflected in the results. Furthermore, the commonly known limitations of kNN in high dimensional dataspaces are demonstrated clearly. In the case of non linearly separable classes, it can be seen that kNN does not manage correct class assignment for more than 5-15 dimensions, depending on the noise level and the training data size (assuming uncorrelated variables). In addition the expected inability of PLS-DA to discriminate non linearly separable datasets and the importance of a sufficient amount of training data are reflected in the results.

While the classifier selection has proven not to be relevant for certain datasets (low dimensions, linearly separable, low noise), for other datasets the advantages of the Random Forest classifier are notable.

Based on these findings a Random Forest model is chosen for classifying infrared hyperspectral images of skin tissue sections in Part II. Selected spectral descriptors and different sizes of training datasets are used to train the classification models. The model distinguishes epidermis, melanoma, ulcerating tissue and different kinds of connective tissue. Correct tissue identification for all tissue types is achieved. The model is applied successfully to create "digital stains" of samples which have not been included in the training data.

An open problem are false positive melanoma assignments for isolated pixels at the transition from epidermis to dermis (basal layer) and for pixels related to blood vessels. This effect is explained by the mixture of different class spectra at those

transitions and is expected to be reduced by identification and application of improved descriptors, which are more selective towards melanoma cells.

Further improvement and generalization of the model is expected for a higher number of training data gathered from different patients. Additionally, in further investigations, samples featuring nevi should be analyzed and included in the model as an additional class.

Recent high technological advances in infrared imaging, based on quantum cascade lasers, lead to strongly reduced data acquisition times and are therefore ideally suited for clinical applications. Together with improved classification methods, this guarantees exciting years to come for the field of spectral histopathology.

Appendices

Appendix A

Comparison of Partial Least Square to Principal Component Regression

Principal Component Analysis (PCA) performs a unitary change of basis (rotation) to the *eigenspace* of the sample covariance matrix $\mathbf{\Sigma}$, which is defined by Eq. (A.1).

$$\begin{aligned} \mathbf{X} &\epsilon \mathbb{R}^{(n \times d)} \\ \mathbf{\Sigma} &\epsilon \mathbb{R}^{(d \times d)} \\ \mathbf{\Sigma} &= \frac{1}{n-1} \mathbf{X}^T \mathbf{X} \end{aligned} \tag{A.1}$$

$\mathbf{\Sigma}$ is a symmetric matrix with the sample variances as diagonal and the corresponding co-variances as off-diagonal elements. Thus, by a change of basis to its eigenspace $\mathbf{\Sigma}$ can be diagonalized.[18]

[18] A vector space which is spanned by the eigenvectors of the operator $\mathbf{A}$ is called its *eigenspace*. $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ if the relation

$$\begin{aligned} \mathbf{A}\mathbf{x}_i &= \lambda_i \mathbf{x}_i \\ 1 &\leq i \leq d \end{aligned}$$

is fulfilled. $\mathbf{x}_i$ is the i-th eigenvector of A and λ_i is the corresponding i-th eigenvalue. For a d-dimensional operator $\mathbf{A}$ there exists at least 1 and at most d linearly independent eigenvectors. Thus, the eigenspace is a subspace of the original d-dimensional vector space. If, and only if, A is symmetric (complex generalization: hermetic) d linearly independent eigenvectors are found. In this case the operator can be diagonalized by a change of basis to its eigenspace.

$$\mathbf{A} = \mathbf{V}^T \mathbf{D}_A \mathbf{V}$$

$\mathbf{V} \epsilon \mathbb{R}^{(d \times d)}$ is the transformation matrix to the eigenspace of $\mathbf{A}$, with the eigenvector $\mathbf{x}_i \epsilon \mathbb{R}^{(d \times 1)}$ as the i-th column. As $\mathbf{A}$ is symmetric, d linearly independent eigenvectors exist, which can thus be orthogonalized and thus $\mathbf{V}$ is an orthogonal matrix. $\mathbf{D}_A$ is a diagonal matrix with λ_i as the i-th diagonal element, if $\mathbf{x}_i$ is the i-th column of $\mathbf{V}$.

The eigenspace of the $\mathbf{\Sigma}$ is called *Principal Component Space*. The eigenvectors are referred to as *principal components* (PC). Thus, a change of basis to the eigenspace of $\mathbf{\Sigma}$ results in its diagonal representation, meaning that data were decorrelated as all $\sigma_{i,j} = 0$ for $i \neq j$. The remaining diagonal elements are the sample variances along the axes of the new coordinate system - the principal components. PC-1 is orientated along the direction with the largest variance, PC - 2 is along the direction with the largest variance which is orthogonal to PC - 1 and so on. To sum up, the idea behind PCA is to transform the data to its principal component space to decorrelate them and to orientate the new axes along the orthogonal direction of the largest variance.

This can be very useful, as in high dimensional data spaces the higher PCs often only contain noise and all the information (large variances) is contained in the first f PCs. If only the first f PCs are considered for further analysis PCA is thus very powerful for the reduction of dimensionality and exploratory data analysis.

One drawback of PCA is that for the change of basis only the sample covariance matrix is considered, without taking into account the relation of descriptors to a response (e.g. classification output).

Although in many cases the directions of the highest variance of the descriptors coincide with the directions for the best response prediction, this does not apply to all cases. To solve this, Would adapted the PCA algorithm and created PLS, a *supervised learner*, as it considers both, the predictors and the response, for the change of basis.

While PCA is based on the diagonalisation of $\mathbf{\Sigma}$, PLS is based on the covariance matrix of data $\mathbf{X}$ and response $\mathbf{Y}$

$$cov(\mathbf{X}, \mathbf{Y}) = \frac{1}{n-1}\mathbf{X^T Y} \tag{A.2}$$

In the following the idea and computation of PLS is outlined by direct comparison to PCA. In both cases the NIPALS algorithm, which is illustrated briefly below, is used to solve the eigenvalue problem. An alternative approach for computation of the PLS components is the SIMPLS algorithm [21]. For a univariate response Matrix (1 class), the SIMPLS algorithm is equivalent to NIPALS-PLS1 which is explained below.

PCA

Idea: Perform a change of basis

$$\mathbf{T} = \mathbf{XV} \tag{A.3}$$

with the transformation matrix $\mathbf{V}$ that results in uncorrelated transformed data $\mathbf{T}$, i.e. $cov(\mathbf{T}, \mathbf{T})$ is diagonal. The coordinates of the data with respect to the transformed coordinate space are called *scores.* The basis vectors of the new basis (expressed by the initial coordinate system) are called *loadings* and are the column vectors of the transformation matrix $\mathbf{V}$.

The diagonal elements describe the variance along the new basis vectors and are ordered by decreasing value (decreasing variance). The diagonalization (eigenvalue) problem is described by

$$\begin{aligned} \mathbf{X^T X} \cdot \mathbf{v_i} &= \lambda_i \mathbf{v_i} \\ cov(\mathbf{T}, \mathbf{T}) &= \mathbf{V^T \Sigma V} \\ \mathbf{X} &= \mathbf{T V^T} + \mathbf{E} \end{aligned}$$

PLS

Idea: Perform changes of basis for features and response variables to new data spaces by the transformation matrices $\mathbf{V}$ and $\mathbf{W}$ that maximize the covariance $cov(\mathbf{T}, \mathbf{U})$ between the resulting feature and response scores. The changes of basis are described by ($\mathbf{E}$ and $\mathbf{F}$ are the residual matrices)

$$\begin{aligned} \mathbf{X} &= \mathbf{T} \cdot \mathbf{V^T} &+\mathbf{E} \\ \mathbf{Y} &= \mathbf{U} \cdot \mathbf{W^T} &+\mathbf{F} \end{aligned}$$

and the scores $\mathbf{T}$ and $\mathbf{U}$ are related by the linear regression ($\mathbf{B}$ is a diagonal matrix)

$$\mathbf{U} = \mathbf{TB} \tag{A.4}$$

The eigenvalue problem is described by

$$\mathbf{X^T Y Y^T X w_i} = \lambda_i \mathbf{w_i} \tag{A.5}$$

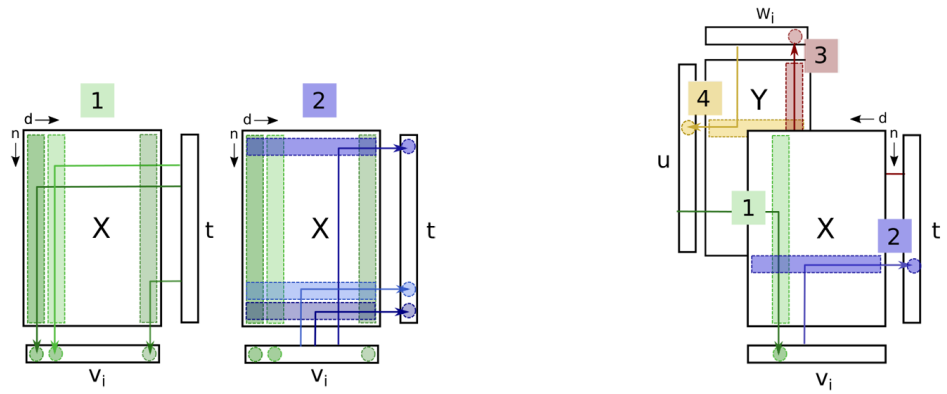


Figure A.1: Illustriation of NIPALS - algorithm for solving the eigenvalue problem of a) PCA and b) PLS-DA

NIPALS for PCA

0. Initialization of $\mathbf{t}$ with an arbitrary column of $\mathbf{X}$ and normalization

$$\mathbf{t} = \mathbf{x_j} \qquad \|\mathbf{t}\| \to 1$$

1. Estimation of $\mathbf{v}$ by projection of $\mathbf{X^T}$ onto $\mathbf{t}$ and normalization

$$\mathbf{v} = \mathbf{X^T t} \qquad \|\mathbf{v}\| \to 1$$

2. Estimation of iteratively adapted $\mathbf{t}$ by projection of $\mathbf{X}$ onto $\mathbf{v}$ and normalization

$$\mathbf{t_{old}} = \mathbf{t}$$
$$\mathbf{t} = \mathbf{Xv} \qquad \|\mathbf{t}\| \to 1$$

3. Repeat steps 1 and 2 until $\|\mathbf{t} - \mathbf{t_{old}}\| < \epsilon$, with the user defined threshold ϵ.

NIPALS for PLS

0. Initialization of $\mathbf{u}$ with an arbitrary column of $\mathbf{X}$ and normalization

$$\mathbf{u} = \mathbf{x_j} \qquad \|\mathbf{u}\| \to 1$$

1. Estimation of $\mathbf{v}$ by projection of $\mathbf{X^T}$ onto $\mathbf{u}$ and normalization

$$\mathbf{v} = \mathbf{X^T u} \qquad \|\mathbf{v}\| \to 1$$

2. Estimation of $\mathbf{t}$ by projection of $\mathbf{X}$ onto $\mathbf{v}$ and normalization

$$\mathbf{t} = \mathbf{Xv} \qquad \|\mathbf{t}\| \to 1$$

3. Estimation of $\mathbf{w}$ by projection of $\mathbf{Y^T}$ onto $\mathbf{t}$ and normalization

$$\mathbf{w} = \mathbf{Y^T t} \qquad \|\mathbf{w}\| \to 1$$

4. Estimation of adapted $\mathbf{u}$ by projection of $\mathbf{Y}$ onto $\mathbf{w}$ and normalization

$$\mathbf{u_{old}} = \mathbf{u}$$
$$\mathbf{u} = \mathbf{Yw} \qquad \|\mathbf{u}\| \to 1$$

5. Repeat steps 1 and 4 until $\|\mathbf{u} - \mathbf{u_{old}}\| < \epsilon$, with the user defined threshold ϵ.

Appendix B

General Rotations in d Dimensions

A rotation matrix $\mathbf{M}(\varphi)$ is the operator which describes a rotation of any d - dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_d)^T$ by the angle φ, resulting in $\mathbf{x}' = (x'_1, x'_2, ..., x'_d)^T$. Rotation matrices are orthogonal operators and thus preserve the length of the rotated vectors [81].

$$\begin{aligned} \mathbf{x}' &= \mathbf{M}\,\mathbf{x} \\ \mathbf{M}\,\mathbf{M^T} &= \mathbf{1} \\ \mathbf{M^T} &= \mathbf{M^{-1}} \end{aligned} \tag{B.1}$$

The plane spanned by $\mathbf{x}$ and $\mathbf{x}'$ is referred to as the *rotation plane*. The orthogonal complement to the rotation plane, a (d-2)-dimensional subspace, is the subspace around which the rotation by φ takes place. In 2D, the (d-2)-dimensional subspace corresponds to a point, in 3D to an axis, in 4D to a 2-dimensional plane etc. [82, 83]. This subspace can have an arbitrary orientation with respect to the axes of the Euclidean coordinate system. If the rotation plane is spanned by any two of the main axes X_a and X_b of the Euclidean coordinate system (and is this a *coordinate plane*), the rotation is referred to as a *standard rotation* $\mathbf{R_{Std;\,a,b}}$ and the matrix is given by Eq. (B.2).

$$\mathbf{R_{Std;\,a,b}} \rightarrow r_{ij} = \begin{cases} r_{a,a} = \cos(\varphi) \\ r_{b,b} = \cos(\varphi) \\ r_{a,b} = -\sin(\varphi) \\ r_{b,a} = \sin(\varphi) \\ r_{i,i} = 1 \quad if\ i \neq a, b \\ r_{i,j} = 0 \quad elsewhere \end{cases} \tag{B.2}$$

To estimate the rotation matrix around an arbitrarily orientated subspace (with a rotation plane other than a coordinate plane) the rotation can be decomposed into

a sequence of rotations in coordinate planes[19]

The orthogonal subspace around which the rotation takes place (orthogonal complement of rotation plane) is spanned by the so called *simplex*[20]

Firstly, the data are rotated in a way that the (d-2)-simplex is rotated to be aligned with any (d-2)-dimensional subspace spanned by the main axis. In this work it is aligned to the subspace spanned by the X_1 to X_{d-2} axis. This transformation is represented by a succession of $N_k = \frac{(n-1)\cdot d}{2} - 1$ standard rotations $\mathbf{R_k}(\phi_k)$, $k \epsilon [1, N_k]$ by the angle ϕ_k.

$$\mathbf{R_{SubSp}}(\phi_1, ..., \phi_{N_k}) = \mathbf{R_{N_k}}(\phi_{N_k}) \, ... \, \mathbf{R_1}(\phi_1) \tag{B.4}$$

Each performed rotation transforms the simplex and thus updates the vertices matrix to $\mathbf{vert}^{(k-1)}$ (for the (k-1)-th rotation). The angle ϕ_k for the required subsequent rotation around is computed based on those updated vertices. (See [27] for further information).

This succession of transformations rotates the rotation plane of the requested rotation to the $X_{d-1} - X_d$ plane. Subsequently, a rotation by φ is performed in this

[19] Rotation matrices (orthogonal matrices with $det(\mathbf{M}) = 1$) form the group SO(N). Thus, a succession of k rotations $\mathbf{M_1}$ to $\mathbf{M_k}$ can be expressed as the multiplication of the individual rotation matrices [81].

$$\mathbf{M} = \mathbf{M_k} \, \mathbf{M_{k-1}} \, ... \, \mathbf{M_2} \, \mathbf{M_1}$$

[20] A simplex is the simplest form of a n-dimensional polytope and can be seen as the generalization of a triangle in two and a tetrahedron in three dimensions, respectively. A n-simplex is described by (n+1) coordinate points, so called *vertices* [84].

Dim	Rotation subspace (dimension)	Simplex	No. of vertices
3	axis (1)	vector	2
4	plane (2)	triangle	3
5	cube (3)	tetrahedron	4

For example, a rotation in 4-D takes place around a plane. The simplex representing that plane is a triangle. The vertices (coordinates) of that triangle can be summarized in a matrix:

$$\mathbf{vert}^{(0)} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix} \tag{B.3}$$

If the triangle does not have any offset from the origin, no translation prior to rotation is required and one vertex of the triangle already coincidence with the origin.

rotation plane

$$\mathbf{R_{Std;d-1,d}}(\varphi) = \begin{matrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & & \ddots & \vdots \\ \vdots & & 1 & & 0 \\ \vdots & \ddots & & \cos(\varphi) & -\sin(\varphi) \\ 0 & \dots & 0 & \sin(\varphi) & \cos(\varphi) \end{matrix} \tag{B.5}$$

Finally, the back transformation (transpose of Eq. (B.4)) is conducted in order to rotate the simplex back to its original orientation [27].

$$\begin{aligned} \mathbf{R^{-1}_{SubSp}}(\phi_1, ..., \phi_{N_k}) = \mathbf{R^{T}_{SubSp}}(\phi_1, ..., \phi_{N_k}) = \\ \mathbf{R^{T}_{1}}(\phi_1) \dots \mathbf{R^{T}_{N_k}}(\phi_{N_k}) \end{aligned} \tag{B.6}$$

The complete rotation can be written as a product of the individual rotations.

$$\mathbf{M} = \mathbf{R^{-1}_{SubSp}}(\phi_1, ..., \phi_{N_k}) \ \mathbf{R_{Std;\,d-1,d}}(\varphi) \ \mathbf{R_{SubSp}}(\phi_1, ..., \phi_{N_k}) \tag{B.7}$$

Appendix C

Staging System For Melanoma

The TNM system defines the tumor staging based on its progression in size (T: tumor), affected lymph nodes (N: nodes) and distant metastasis (M: metastasis)[85].

The indicator T is defined by the Breslow depth of the melanoma:

- Tis: Melanoma in situ. Cancer cells only in the epidermis
- T1: Breslow depth less than 1mm
- T2: Breslow depth between 1 and 2 mm
- T3: Breslow depth between 2 and 4 mm
- T4: Breslow depth larger than 4mm

Additionally, the T system often states information about the presence of ulceration (a for an ulcerated and b for not ulcerated lesion.)

The indicator T is defined by the presence of cancer cells in the neighboring lymph nodes and lymphatic ducts.

- N0: No melanoma cells present
- N1: Cancer cells in one lymph node
- N2: Cancer cells on 2 or 3 lymph nodes
- N3: Cancer cells in more than 3 lymph nodes

An additional label of the N system gives information about the characteristics of the cancer cells in the lymph nodes. If the cancer cells in the lymph nodes can only be recognized using a microscope, they are classified by the index a. If the lymph node metastasis can be seen macroscopically, the class label b is used. The melanoma is classified by the label **c**, if cancer cells can be found in the lymphatic ducts of the skin.

Table C.1: Melanoma staging and survival rates [85–88]

TNM	AJCC Stage				5 year	10 year Survival
0	0	Tis	N0	M0	100 %	100 %
I	Ia	T1a	N0	M0	97%	95%
	Ib	T1b/T2a	N0	M0	92%	86%
II	IIa	T2b/T3a	N0	M0	81%	67%
III	IIb	T3b/T4a	N0	M0	70%	57%
	IIc	T4c	N0	M0	53%	40%
	IIIa	T1-4a	N1-2a	M0	78%	68%
	IIIb	T1-4a/T1-4b	N0	M0	59%	43%
	IIIc	T1-4b	N0	M0	40%	24%
IV	IV	Any T	Any N	M1	15-20%	10-15%

The M system indicates, whether the tumor forms metastasis in other parts of the body:

- M0: No metastases
- M1a: Melanoma cells present in skin or lymphatic organs at distant body sites
- M1b: Melanoma cells are found in the lung
- N3: Melanoma cells are found in other organs or the lactate dehydrogenase level of the blood is high.

Based on the T, N and M indicator a certain stage is assigned to the tumor. An overview of the stage assignment and respective survival rates are given in Tbl. C.1

Bibliography

[1] R. G. Brereton, *Applied Chemometrics for Scientists*. John Wiley & Sons, 2007.

[2] J. Ofner and H. Lohninger, "Multisensor hyperspectral imaging as a versatile tool for image-based chemical structure determination," *Spectroscopy Europe*, vol. 26, no. 5, pp. 6–10, 2014.

[3] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, 2009.

[4] R. Rifkin, "In Defense of One-Vs-All Classification," *Journal of machine learning research*, vol. 5, pp. 101–141, 2004.

[5] R. E. Bellman, *Adaptive Control Processes*. Princeton University Press, 1961.

[6] H. Lohninger and J. Ofner, "Multisensor hyperspectral imaging as a versatile tool for image-based chemical structure determination," *Spectroscopy Europe*, vol. 26, no. 5, p. 6, 2014.

[7] H. Lohninger, "Help File ImageLab." (2016-10-09). Available at `http://www.imagelab.at/help/spectral_descriptors.htm`.

[8] M. J. Zaki and M. J. Wagner, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2 ed., 2009.

[10] G. Dougherty, *Pattern Recognition and Classification. An Introduction*. Springer, 2013.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[12] K. Hechenbichler and K. Schliep, "Weighted k-nearest-neighbor techniques and ordinal classification," 2004.

[13] S. Theodoridis and K. Koutroumbas, *Pattern Recognition.* Elsevier, 2 ed., 2003.

[14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967.

[15] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 607–616, 1996.

[16] J. A. Westerhuis, E. J. J. van Velzen, H. C. J. Hoefsloot, and A. K. Smilde, "Discriminant Q2 (DQ2) for improved discrimination in PLSDA models," *Metabolomics*, vol. 4, pp. 293–296, 2008.

[17] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. Velzen, J. P. M. Duijnhoven, and F. A. Dorsten, "Assessment of PLSDA cross validation," *Metabolomics*, vol. 4, no. 1, pp. 81–89, 2008.

[18] T. D. Randal, "An introduction to partial least squares regression," in *Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL*, pp. 2–5, SAS Institute Inc, 1995.

[19] H. Lohninger, "Fundamentals of Statistics." (2016-09-10). Available at `http://www.statistics4u.com/fundstat_eng/`.

[20] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: Taking the magic away," *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, 2014.

[21] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[22] L. Breiman, "Random Forests," *Machine learning*, vol. 45, no. 5, pp. 5–32, 2001.

[23] L. Tolosi and T. Lengauer, "Classification with correlated features: Unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011.

[24] H. Tamhankar and J. E. Fowler, "Spectral-decorrelation strategies for the compression of hyperspectral imagery," in *Proceedings of the International Geoscience and Remote Sensing Symposium*, pp. 1041–1044, 2007.

[25] R. J. D. Helmut H. Telle, Angel González Ureña, *Laser Chemistry: Spectroscopy, Dynamics and Applications.* John Wiley & Sons, 2007.

[26] A. Mehta, "Ultraviolet-Visible (UV-Vis) Spectroscopy – Limitations and Deviations of Beer-Lambert Law." (2016-09-16). Available at `http://pharmaxchange.info/press/2012/05/ultraviolet-visible-uv-vis-spectroscopy-%E72%80%93-limitations-and-deviations-of-beer-lambert-law/`.

[27] A. Aguilera and R. Pérez-Aguila, "General n-dimensional rotations," in *Proc. WSCG SHORT Commun. Papers*, pp. 1–8, UNION Agency - Science Press, 2004.

[28] N. Tandon, "Hatchfill." 2011; (2016-06-23) Available at `https://de.mathworks.com/matlabcentral/fileexchange/30733-hatchfill`.

[29] Y. Cao, "Partial Least-Squares and Discriminant Analysis." 2011; (2016-04-10) Available at `https://de.mathworks.com/matlabcentral/fileexchange/18760-partial-least-squares-and-discriminant-analysis`.

[30] H. Mohan, *Textbook of Pathology*. Jaypee, 7 ed., 2015.

[31] European Network of Cancer Registry (ENCR), "Malignant Melanoma of the Skin (MM) Factsheet." April 2015, (2016-09-18). Available at `http://encr.eu/images/docs/factsheets/ENCR_Factsheet_Malignant_Melanoma_2015.pdf`.

[32] A. Faller and M. Schünke, *Der Körper des Menschen. Einführung in Bau und Funktion*. Thieme, 15th ed., 2008.

[33] Madhero88 and M.Komorniczak, "Skin Layers," 2012. (2016-10-01). Available at `https://en.wikipedia.org/wiki/File:Skin_layers.png`.

[34] J. A. McGrath, R. A. J. Eady, and F. M. Pope, "Anatomy and Organization of Human Skin," in *Rook's Textbook of Dermatology* (T. Burns, S. Breathnach, N. Cox, and C. Griffiths, eds.), pp. 45–128, Oxford, UK: Wiley-Blackwell, 8th ed., 2004.

[35] R. L. Eckert and E. A. Rorke, "Molecular biology of keratinocyte differentiation," *Environmental Health Perspectives*, vol. 80, pp. 109–116, 1989.

[36] M. Walensi, "Melanozyt." (2016-09-18). Available at `http://flexikon.doccheck.com/de/Melanozyt`.

[37] V. Liu and M. C. Mihm, "Pathology of malignant melanoma," *Surgical Clinics of North America*, vol. 83, pp. 31–60, 2003.

[38] J. Hunter, J. Savian, and M. Dahl, *Clinical Dermatology*. Blackwell Science, 3rd ed., 2002.

[39] N. Y. Z. Chiang and J. Verbov, *Dermatology. A handbook for medical students & junior doctors*. British Association of Dermatologists, 2nd ed., 2014.

[40] G. Bellisola and C. Sorio, "Infrared spectroscopy and microscopy in cancer research and diagnosis.," *American journal of cancer research*, vol. 2, no. 1, pp. 1–21, 2012.

[41] W. Demtröder, *Experimentalphysik 3. Atome, Molehüle und Festkörper*. Springer-Lehrbuch, 2010.

[42] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*. Wiley, 2004.

[43] C. K. Wagner, *Application of microfluidic devices for time resolved FTIR spectroscopy*. PhD thesis, Vienna University of Technology, 2012.

[44] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. a. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, "Using Fourier transform IR spectroscopy to analyze biological materials.," *Nature protocols*, vol. 9, no. 8, pp. 1771–91, 2014.

[45] K. Belbachir, R. Noreen, G. Gouspillou, and C. Petibois, "Collagen types analysis and differentiation by FTIR spectroscopy," *Analytical and Bioanalytical Chemistry*, vol. 395, no. 3, pp. 829–837, 2009.

[46] Z. Movasaghi, S. Rehman, and D. I. ur Rehman, "Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues," *Applied Spectroscopy Reviews*, vol. 43, no. 2, pp. 134–179, 2008.

[47] F. Lyng, I. Ramos, O. Ibrahim, and H. Byrne, "Vibrational Microspectroscopy for Cancer Screening," *Applied Sciences*, vol. 5, no. 1, pp. 23–35, 2015.

[48] A. Tfayli, O. Piot, A. Durlach, P. Bernard, and M. Manfait, "Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy," *Biochimica et Biophysica Acta - General Subjects*, vol. 1724, no. 3, pp. 262–269, 2005.

[49] B. De Campos Vidal and M. L. S. Mello, "Collagen type I amide I band infrared spectroscopy," *Micron*, vol. 42, no. 3, pp. 283–289, 2011.

[50] R. K. Dukor, "Vibrational spectroscopy in the detection of cancer," in *Handbook of vibrational spectroscopy*, pp. 3335–3361, John Wiley & Sons, 2002.

[51] F. Lyng, E. Gazi, and P. Gardner, "Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging," in *Biomedical Applications of Synchrotron Infrared Microspectroscopyy, RSC Analytical Spectroscopy Monographs*, no. 11, pp. 147–185, Royal Society of Chemistry, 2011.

[52] J. M. Nowacek, "Fixation and Tissue Processing," in *Special Stains and H & E*, pp. 151–152, Dako, 2nd ed., 2010.

[53] A. Tfayli, C. Gobinet, V. Vrabie, R. Huez, M. Manfait, and O. Piot, "Digital dewaxing of Raman signals: Discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies," *Applied Spectroscopy*, vol. 63, no. 5, pp. 564–570, 2009.

[54] G. Socrates, *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. John Wiley & Sons, 3rd ed., 2001.

[55] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 100–114, 2012.

[56] C. Gordon, "A generalization of the maximum noise fraction transform," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 608–610, 2000.

[57] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "Transformation for Ordering Multispectral Data in Terms of Image Quality With Implications for Noise Removal.," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, no. 1, pp. 65–74, 1988.

[58] T. N. Q. Nguyen, P. Jeannesson, A. Groh, O. Piot, D. Guenot, and C. Gobinet, "Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal colon," *Journal of Biophotonics*, vol. 532, no. 5, pp. 521–532, 2016.

[59] P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, M. Brown, N. Clarke, and P. Gardner, "RMieS-EMSC correction for infrared spectra of biological cells: Extension using full Mie theory and GPU computing," *Journal of Biophotonics*, vol. 3, no. 8-9, pp. 609–620, 2010.

[60] M. Diem, M. Miljković, B. Bird, T. Chernenko, J. Schubert, E. Marcsisin, A. Mazur, E. Kingston, E. Zuser, K. Papamarkakis, and N. Laver, "Applications of Infrared and Raman Microspectroscopy of Cells and Tissue in Medical Diagnostics: Present Status and Future Promises," *Spectroscopy: An International Journal*, vol. 27, no. 5-6, pp. 463–496, 2012.

[61] P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke, and P. Gardner, "FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm.," *The Analyst*, vol. 137, no. 6, pp. 1370–7, 2012.

[62] P. Bassan, *Light scattering during infrared spectroscopic measurements of biomedical samples*. Ph.d., University of Manchester, 2011.

[63] P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, and P. Gardner, "Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples.," *The Analyst*, vol. 135, no. 2, pp. 268–277, 2010.

[64] N. K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 92–99, 2012.

[65] H. Martens and E. Stark, "Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 9, no. 8, pp. 625–635, 1991.

[66] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait, "Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies.," *The Analyst*, vol. 133, no. 2, pp. 197–205, 2008.

[67] G. Lu and B. Fei, "Medical hyperspectral imaging: a review.," *Journal of biomedical optics*, vol. 19, no. 1, p. 10901, 2014.

[68] M. Pilling and P. Gardner, "Fundamental developments in infrared spectroscopic imaging for biomedical applications," *Chem. Soc. Rev.*, vol. 45, pp. 1935–1957, 2016.

[69] R. K. Sahu and S. Mordechai, "Fourier transform infrared spectroscopy in cancer detection," *Future Oncology*, vol. 1, no. 5, pp. 635–47, 2005.

[70] D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson, and O. Piot, "Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections.," *Laboratory investigation; a journal of technical methods and pathology*, vol. 91, no. 5, pp. 799–811, 2011.

[71] Z. Hammody, S. Argov, R. K. Sahu, E. Cagnano, R. Moreh, and S. Mordechai, "Distinction of malignant melanoma and epidermis using IR micro-spectroscopy and statistical methods," *Analyst*, vol. 133, no. 3, pp. 372–378, 2008.

[72] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data.," *BMC bioinformatics*, vol. 10, no. 1, p. 213, 2009.

[73] D. Sebiskveradze, C. Gobinet, E. Ly, M. Manfait, P. Jeannesson, M. Herbin, O. Piot, and V. Vrabie, "Effects of digital dewaxing methods on K-means-clusterized IR images collected on formalin-fixed paraffin-embedded samples of skin carcinoma," in *8th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2008*, 2008.

[74] C. Gobinet, D. Sebiskveradze, V. Vrabie, A. Tfayli, O. Piot, and M. Manfait, "Digital dewaxing of Raman spectral images of paraffin-embedded human skin biopsies based on ICA and NCLS," *European Signal Processing Conference*, no. Eusipco, pp. 1–5, 2008.

[75] S. M. Ali, "A Comparison of Raman , FTIR and ATR-FTIR Micro Spectroscopy for Imaging Human Skin Tissue Sections .," *Analytical Methods*, vol. 5, pp. 2261–2291, 2013.

[76] H. Martens, "The EMSC toolbox for MATLAB." (2016-09-01). Available at `http://www.models.life.ku.dk/source/emsctoolbox`.

[77] H. Martens, J. Pram Nielsen, and S. Balling Engelsen, "Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures," *Analytical Chemistry*, vol. 75(3), pp. pp.394–404, 2003.

[78] V. Lucarini, J. J. Saarinen, K.-E. Peiponen, and E. M. Vartiainen, *Kramers-Kronig Relations in Optical Materials Research*. Springer, 2005.

[79] B. Bird, M. Milos, and M. Diem, "Two step resonant Mie scattering correction of infrared micro-spectral data: human lymph node tissue," *Journal of Biophotonics*, vol. 3, no. 8-9, pp. 597–608, 2010.

[80] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 2003.

[81] M. Bartelmann, B. Feuerbacher, T. Krüger, D. Lüst, A. Rebhan, and A. Wipf, *Theoretische Physik*. 2014.

[82] S. R. Hollasch, *Four-Space Visualization of 4D Objects*. Master, Arizona State University, 1991.

[83] D. Banks, "Interactive manipulation and display of surfaces in four dimensions," in *1992 Symposium on Interactive 3D Graphics*, pp. 197–207, 1992.

[84] E. W. Weisstein, "Simplex." (2016-08-10) Mathworld - a Wolfram based web resource. Available at http://mathworld.wolfram.com/Simplex.html.

[85] Cancer Research UK, "Stages of Melanoma." (2016-10-01). Available at http://www.cancerresearchuk.org/about-cancer/type/melanoma/treatment/stages-of-melanoma.

[86] American Cancer Society, "Survival rates for melanoma skin cancer, by stage ." (2016-09-20). Available at http://www.cancer.org/cancer/skincancer-melanoma/detailedguide/melanoma-skin-cancer-survival-rates-by-stage.

[87] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, and A. Trotti, *Melanoma of the skin. AJCC Cancer Staging Manual.* 7th ed., 2010.

[88] National Cancer Institute, "SEER Stat Fact Sheets: Melanoma of the Skin." (2016-09-20). Available at http://seer.cancer.gov/statfacts/html/melan.html.