University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Siim Viigand

Assessment of surgical margins of basal cell carcinoma with Raman microspectroscopy measurements

Mathematics and Statistics Curriculum

(Mathematical Statistics)

Master's thesis (30 ECTS)

Supervisor: Kristi Kuljus, PhD

Alexey Koloydenko, PhD

Tartu 2024

ASSESSMENT OF SURGICAL MARGINS OF BASAL CELL CARCINOMA WITH RAMAN MICROSPECTROSCOPY MEASUREMENTS

Master's thesis

Siim Viigand

Abstract

The aim of the master's thesis is to study and set reference results for assessment of residual tumor margins of basal cell carcinoma. The analysis is based on Raman microspectroscopy measurements of tissue samples extracted during Mohs surgery. Logistic regression, linear discriminant analysis and quadratic discriminant analysis are used to develop classification rules, several set-ups of spectral feature variables are considered. The best classification result is achieved with the logistic regression model with 30 original spectral features. The best model is validated on a test set of extracted tissue samples. Concerns regarding the data can be taken into account and the obtained reference results can be used in future analyses, when models of higher complexity could be studied for classification.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Key Words: basal cell carcinoma, Raman spectroscopy, classification, logistic regression, discriminant analysis

BASALIOOMI HINDAMINE KIRURGILISELT EEMALDATUD KOEKIHTIDES RAMANI MIKROSPEKTROSKOOPIA ANDMETE PÕHJAL

Magistritöö

Siim Viigand

Lühikokkuvõte

Magistritöö eesmärk on uurida basalioomi tuvastamist kirurgiliselt eemaldatud koekihtides. Nahakasvaja olemasolu analüüsitakse koekihtidel teostatud Ramani mikrospektroskoopia mõõtmiste põhjal. Klassifitseermisreeglite leidmiseks on kasutatud logistilist regressiooni, lineaarset ja mittelineaarset diskriminantanalüüsi. Klassifitseerimisreeglite hindamisel on kaalutud erinevaid spektraaltunnuste valikuid. Parimad tulemused saavutati logistilise regressiooniga, kus kasutati kõiki 30 normaliseerimata spektraaltunnust. Saadud tulemusi saab kasutada ning andmestikuga seotud probleeme arvesse võtta tulevikus teostatavates analüüsides, kui on soov uurida keerukamaid mudeleid.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: basalioom, Ramani spektroskoopia, klassifitseerimine, logistiline regressioon, diskriminantanalüüs

Contents

Introduction

Cancer treatment as any other medical procedure is often considered a stressful experience. Refined treatment possibilities can provide more favourable treatment outcomes for patients, while also save time for healthcare workers and facilities. Refining treatment procedures for the most common diseases can have a great overall improvement even with minor adjustments, as these procedures are performed most often.

Basal cell carcinoma is the most common skin cancer type that is most effectively removed with a Mohs micrographic surgery (Chung, 2012). Although the surgery is usually having a great end result, it requires tissue inspection under a microscope. Histopathology is a relatively time consuming method as sample preparation and analysis can take up to multiple hours. Thus, exploring options to analyze the extracted tissue samples faster would make the surgery more successful.

The focus of the master's thesis is on a method developed at the University of Nottingham. The method combines autofluorescence (AF) imaging with Raman microspectroscopy to analyze tumor margins during tissue-conserving surgeries. The method is also referred to as multimodal spectral histopathology (MSH). Although a fully-automated prototype instrument has already been developed and there are multiple articles published on this topic from a biomedical perspective, there is a lack of statistical papers and information about classification methods used in these papers.

The aim of this thesis is to study and set reference results for assessment of residual tumor margins of basal cell carcinoma. The analysis is based on Raman microspectroscopy measurements. In the thesis, all tissue sample pixels are handled as independent observations, which by nature they of course are not. The further statistical analysis would aim to study this topic, while taking neighbours and neighbourhoods into consideration. Basic reference results are obtained with linear and quadratic discriminant analysis and logistic regression. The tissue samples an-

alyzed in the thesis and their respective Raman microspectroscopy measurements have been made available by the University of Nottingham's School of Physics and Astronomy.

In the first section, multimodal spectral histopathology and its background is further described to better understand how the method works and how the data was gathered using AF imaging and Raman microspectroscopy. Previous work on this topic is also covered with the results of already published findings in biomedical articles. Section 2 gives a theoretical overview of the statistical methods used for obtaining the reference results. Multiple logistic regression, linear discriminant analysis and quadratic discriminant analysis have been used for classification in this thesis. In the third section, available data is described to give an overview of the response and explanatory variables in this study. Section 4 covers data preprocessing steps as the provided tissue samples needed extensive cleaning. Finally, in the fifth section the results of the analysis of the tissue samples classification are provided and the reference results are set. Version 4.2.2 of statistical programming language R is used for analysis and data processing.

1 The background of multimodal spectral histopathology and previously published findings

The underlying procedure studied in this thesis is quite specific and falls into the biomedical field. Although the main focus of the thesis is on classification of Raman microspectroscopy measurements, it is good to understand the procedure in general. This section provides a short description of Mohs micrographic surgery and methods utilized to analyze tumor margins. In addition, a short summary of already published findings is also provided.

1.1 Basal cell carcinoma and Mohs surgery

One of the most frequently occurring skin cancer types is basal cell carcinoma (BCC). In 2012, its estimated frequency was around 75% to 80% of all discovered non-melanoma skin cancers. This makes it not only the most common skin cancer, but also the most common across all cancers discovered, as one out of every three new cancers is listed as BCC. The main cause of BCC is related to excessive cumulative exposure to ultraviolet light, intense and occasional ultraviolet exposure, overexposure to X-rays or other sorts of radiation. As a result, DNA in the basal cell layer of the epidermis gets mutated and will eventually cause BCC to develop.(Chung, 2012)

While being so common, BCC has many treatment possibilities, but one of the finest methods is Mohs micrographic surgery (MMS). MMS is a skin cancer surgery that has the highest cure rates for multiple of cutaneous malignancies. MMS makes use of histopathology to study extracted tissue samples, which makes it possible to spare more of the unaffected tissue compared to other BCC treatment choices available. The process in general contains removal of clinically evident tumor and then extracting tissue layers, which are 1-2 mm thin and which will be observed under a microscope. The tissue samples are then validated, that is it is studied if

they are clear of tumor cells. If during that validation residual tumor is found, the process is repeated until future specimen is clear of tumor.(Bittner et al., 2021)

1.2 Raman spectroscopy and autofluorescence imaging

Raman spectroscopy is a method to detect light interactions with matter for determining the type of chemical molecules. Raman spectroscopy measures Raman activity that depends on the polarizability of a chemical bond, meaning essentially how easy it is to displace the electrons in the bond. When there are many electrons that are loosely held, the polarizability is high and Raman signal for the molecule will also be high. Measuring Raman signal across the range of wavenumbers provides information about different properties of molecules and can thus help to separate different cell types.(Barron, 2012)

Raman microspectroscopy is a subclass of the method as it combines Raman spectroscopy with a microscope to provide results at micrometer level ($<0.5\mu$m) (Saletnik, Saletnik, and Puchalski, 2021).

Autofluorescence (AF) imaging is a monitoring technique of cell molecules that become fluorescent when excited by UV or visible radiation of suitable wavelength. Physiological and/or pathological processes can cause changes in the cell and tissue state that alter the amount and distribution of endogenous fluorophores and other chemical-physical properties. AF imaging can be used to monitor these differences and to provide information about the cell and tissue state. The benefit of AF imaging is that it can be performed in real time as it does not need any fixing or staining of tissue sample.(Monici, 2005)

1.3 Published findings

Integrated AF imaging and Raman spectroscopy has been brought up as a great combination to detect tumor margins faster than a standard sectioning and stain-

ing procedure. Autofluorescence imaging will give results with high sensitivity and high speed. It recognizes the tumor cells easily, but it classifies too many other tissue cells as being tumorous (has low specificity). AF imaging can be used as a first step to select and prioritize what sampling points should be measured with Raman microspectroscopy. Raman microspectroscopy can then provide results with high sensitivity and specificity, but in contrast will take considerably longer time to collect information. The master's thesis main focus will be on Raman microspectroscopy results, while tasks related to AF imaging (sampling point selection and region prioritizations) are not analyzed.(Kong et al., 2013)

There are multiple different approaches used for data preparation and BCC classification concerning the method studied in the thesis. A few of the previous published findings and statistical methods used in those articles are brought out below.

The first initial study about feasibility of integrated AF imaging and Raman spectroscopy by Kong et al. (2013) used the spectra provided by Raman spectroscopy as a starting point and derived relevant information through dimensionality reduction methods. Raw spectra was initially summarised by 10 principal components that were further reduced to 5 canonical features using rank-reduced multiclass linear discriminant analysis. These 5 canonical spectral features were then used with the multinomial logistic regression classifier at the target sensitivity of 95%. The final result of cross validated data indicated 95.3% of sensitivity and 94.6% of specificity in discriminating BCC from the rest of the tissue classes. The classifier was later validated against an independent set of skin samples that yielded 100% of sensitivity and 92.9% of specificity. It is also worth mentioning that the background area of the samples was identified separately and was not included into the classification model.

A later study by Boitor et al. (2017) used spectral feature variables in their classification models. Spectral features were calculated as areas under the Raman spectra, thus giving a good numerical overview of spectra intensity in certain wavenumber

regions. It is stated that the best classification model was obtained using an artificial neural network with 20 nodes in a single hidden layer and with 13 such manually selected spectral features. The results of this case study state that after a 5-fold cross-validation, BCC can be discriminated from all other tissue types with 87.7% of sensitivity and 98.4% of specificity. The study by Boitor et al. (2017) also states that distinguishing BCC from dermis and fat is easier than from epidermis, inflammation and muscle tissue classes.

The study by Boitor et al. (2023) reports use of fully-automated device for scanning surgical margins in clinical setting. The article compares integrated AF imaging and Raman spectroscopy results with other frozen section histology alternatives like fluorescence confocal microscopy (FCM). Although there are some differences between these methods, FCM has been shown to detect BCC with a sensitivity of 86% – 92% and specificity of 60% – 90%. These results are comparable to the results observed with MSH.

2 Multiple logistic regression and discriminant analysis

The aim of the thesis is to set reference results that could later be used for comparison when models with higher complexity will be studied. As multiple logistic regression and discriminant analysis are widely used methods in classification, they appear as good approaches to set the reference results. Logistic regression was also used in the feasibility study by Kong et al. (2013), so it further suggests to include this method here. Although commonly applied, the following section provides a short overview of the methods used.

2.1 Multiple logistic regression model

Multiple logistic regression is described based on the book by Hosmer et al. (2013). Let us suppose that we want to describe the relationship between a response variable Y and a set of p explanatory variables $X_1, X_2, \ldots, X_p$. For the convenience, during the multiple logistic regression overview we can assume that the response variable Y is dichotomous (binary) and the explanatory variables $X_1, X_2, \ldots, X_p$ are all continuous. We make this assumption because this will be the set-up in our models later on. Let us also denote a constant 1 and the independent variables as the vector $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)'$. Multiple logistic regression aims to model the conditional probability of $Y = 1$ given the vector $\mathbf{X} = \mathbf{x}$ and is denoted by $\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. The multiple logistic regression model can then be given as a standard logistic function

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}}},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients to be estimated based on the data.

The logit of the multiple logistic regression model is given by the equation

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \boldsymbol{\beta}'\mathbf{x}.$$

The logit transformation illustrates that $g(\mathbf{x})$ has many of the good properties of the linear regression model, most importantly that the logit function $g(\mathbf{x})$ is linear in its parameters. However, in the linear regression model we assume that model errors follow a normal distribution, but this is not the case for binary logistic regression. We can express the value of the outcome variable Y for given $\mathbf{x}$ as $y = \pi(\mathbf{x}) + \varepsilon$. Here the error term ε can take two possible values as the outcome also has only two values: if $y = 1$, then $\varepsilon = 1 - \pi(\mathbf{x})$ with probability $\pi(\mathbf{x})$, and if $y = 0$ then $\varepsilon = -\pi(\mathbf{x})$ with probability $1 - \pi(\mathbf{x})$. It follows that the conditional distribution of the outcome variable follows the Bernoulli distribution with parameter $\pi(\mathbf{x})$.

2.2 Fitting the multiple logistic regression model

Let us assume that we have n independent observations $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$. Maximum likelihood method is usually used to estimate the parameters of the logistic regression model. Thus, fitting the multiple logistic regression model means that we seek the maximum likelihood estimates of the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. To begin with, the likelihood function must be constructed to express the probability of the observed data as a function of the unknown regression coefficients. Then we maximize this function to get the maximum likelihood estimates. We can explore the contribution of the pair $(\mathbf{x}_i, y_i)$ to the likelihood function. Since $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$ and $P(Y = 0|\mathbf{x}) = 1 - \pi(\mathbf{x})$, we can see that the observations $(\mathbf{x}_i, y_i)$, where $y_i = 1$, contribute to the likelihood function by $\pi(\mathbf{x}_i)$ and the observations, where $y_i = 0$, contribute by $1 - \pi(\mathbf{x}_i)$. This can be expressed as

$$\pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}.$$

Since we assumed that the observations are independent, the likelihood function is the product of the contributions of $(\mathbf{x}_i, y_i), i = 1, \ldots, n$:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}.$$

Thus, the log-likelihood is given as follows:

$$l(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n}[y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]].$$

To get the estimates we differentiate $l(\boldsymbol{\beta})$ with respect to $\beta_0, \beta_1, \ldots, \beta_p$ and set the expressions equal to zero. We get $p + 1$ likelihood equations that are expressed as

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0, \quad j = 1, \ldots, p.$$

To solve the likelihood equations, iterative numerical methods are used that adjust the parameters until they maximize the likelihood function. As a result, we get the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. The estimated logistic probabilities for given $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ are given by

$$\hat{\pi}(\mathbf{x}_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}}} = \frac{e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_i}}{1 + e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_i}} = \frac{e^{\hat{g}(\mathbf{x}_i)}}{1 + e^{\hat{g}(\mathbf{x}_i)}}, \quad i = 1, \ldots, n.$$

It is also important to test the overall significance of the model, that is to test the hypotheses

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs}$$

$$H_1 : \text{at least one of the } p \text{ coefficients is different from zero.}$$

The test is based on the statistic G, the difference in deviances between the two models. The deviance is defined as

$$D = -2\ln(\text{likelihood of the fitted model}).$$

This test, also called as the likelihood ratio test, is given as

$$G = -2\ln\left[\frac{\text{likelihood without the variables}}{\text{likelihood with the variables}}\right]$$

$= D(\text{the intercept only model}) - D(\text{the model with the } p \text{ explanatory variables }).$

Under the hypothesis H_0, the statistic G follows a chi-squared distribution with p degrees of freedom, $G \sim \chi^2(p)$. When the statistic G is larger than the quantile of $\chi^2(p)$ at the required significance level, H_0 is rejected.

The same approach can be used to compare a model with p explanatory variables with its reduced version. Suppose the reduced model has q explanatory variables, where $q < p$. We can test the hypotheses

$$H_0 : \text{the slope coefficients of the excluded variables are all equal to zero} \quad \text{vs}$$

$$H_1 : \text{at least one of these coefficients is different from zero.}$$

Perform the likelihood ratio test:

$$G = D(\text{the model with } q \text{ variables}) - D(\text{the model with } p \text{ variables}).$$

Under the hypothesis H_0, $G \sim \chi^2(p-q)$. When the statistic G is larger than the quantile of $\chi^2(p-q)$ at the required significance level, H_0 is rejected.

2.3 Discriminant analysis

In the previous two subsections the multiple logistic regression was described to directly model the conditional probability of the response variable Y, given the p explanatory variables $X_1, X_2, \ldots, X_p$. In discriminant analysis, a less direct approach to estimate these probabilities is taken, namely, we have to estimate the joint distribution of $X_1, X_2, \ldots, X_p$ for the different classes of response variable Y. Both linear and quadratic discriminant analysis are applications of the Bayes discriminant rule. Linear and quadratic discriminant analysis are described based on the book by James et al. (2021).

Let us consider a population, where we have K classes (subpopulations), $K \geq 2$. We assume that each observation can only be classified into one of those classes, so the response variable Y can take on K distinct values. Let us also assume that we have n training observations available, where we know the value of the response variable. Let the probabilities

$$P(Y = k) = \pi_k, \quad k = 1, \ldots, K,$$

define the prior probabilities that a randomly chosen population item belongs to the kth class, meaning that these prior probabilities define the proportions of our subpopulations. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, these p variables are used to discriminate between the K classes. The probability density function of $\mathbf{X}$ for the kth class is given as $f_k(\mathbf{x}), k = 1, \ldots, K$. Consider Bayes' theorem that expresses the posterior probability of an observation with $\mathbf{X} = \mathbf{x}$ to belong to the kth class as

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x})}, \quad k = 1, \ldots, K.$$

The posterior probability is the probability that the observation belongs to the kth class, given its measurements of the explanatory variables. This also means that it is reasonable to assign the observation with $\mathbf{X} = \mathbf{x}$ to the class with the highest

posterior probability. In other words, we classify an observation with $\mathbf{X} = \mathbf{x}$ to class k, if

$$\pi_k f_k(\mathbf{x}) = \max_{1 \leq j \leq K} \pi_j f_j(\mathbf{x}).$$

This classification rule is called the Bayes classifier. The master's thesis focus is mostly on the case where $K = 2$, in this case the classification rule for an observation $\mathbf{X} = \mathbf{x}$ simplifies to following:

- if $\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x})$, classify the item to class 1,
- if $\pi_1 f_1(\mathbf{x}) < \pi_2 f_2(\mathbf{x})$, classify the item to class 2,
- if $\pi_1 f_1(\mathbf{x}) = \pi_2 f_2(\mathbf{x})$, choose randomly between the two classes.

To apply the Bayes classifier in practice, one has to estimate $f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})$ and $\pi_1, \ldots, \pi_K$ (if these are not known).

Linear discriminant classifier. In linear discriminant analysis, we assume that $f_k(\mathbf{x})$ is the density of a multivariate Gaussian distribution, with a class-specific mean vector $\boldsymbol{\mu}_k$ and a common covariance matrix $\boldsymbol{\Sigma}$ across all K classes. The probability density function then has the following form:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right), \quad k = 1, \ldots, K.$$

We can now derive the Bayes classifier, where we want to find a maximum of $\pi_k f_k(\mathbf{x}), k = 1, \ldots, K$. Instead of $\pi_k f_k(\mathbf{x})$, we can also consider the maximum over $\ln[\pi_k f_k(\mathbf{x})], k = 1, \ldots, K$, which allows to simplify the rule for the linear discriminant classifier as follows:

$$\begin{aligned} \ln[\pi_k f_k(\mathbf{x})] &= \ln \pi_k + \ln\left(\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) \\ &= \ln \pi_k + \ln\left(\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k. \end{aligned} \tag{1}$$

Since we aim to compare the quantities $\ln[\pi_k f_k(\mathbf{x})]$ depending on k, we can ignore

those parts of equation (1) that do not depend on k. In the end we get the following linear discriminant classification functions:

$$\delta_k(\mathbf{x}) = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k, \quad k = 1, \ldots, K,$$

where we assign an observation $\mathbf{X} = \mathbf{x}$ to the class with the highest value of $\delta_k(\mathbf{x})$.

Quadratic discriminant analysis. We apply a similar approach when the discriminant rule for quadratic discriminant analysis is defined. The only difference is that instead of a common covariance matrix $\boldsymbol{\Sigma}$, we now have class-specific covariance matrices $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$. Let us simplify $\ln[\pi_k f_k(\mathbf{x})]$ for the quadratic discriminant rule as

$$\begin{aligned}\ln[\pi_k f_k(\mathbf{x})] &= \ln\left[\pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)\right] \\ &= \ln \pi_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k).\end{aligned}$$

We once again want to compare the quantities $\ln[\pi_k f_k(\mathbf{x})]$ depending on k, so quadratic discriminant classification functions can be written as

$$q_k(\mathbf{x}) = -\frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}'\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln \pi_k, \quad k = 1, \ldots, K,$$

where an observation with $\mathbf{X} = \mathbf{x}$ is classified to the class with the highest value of $q_k(\mathbf{x})$. When we observe $\delta_k(\mathbf{x})$ and $q_k(\mathbf{x})$, we can see that the linear discriminant function is a linear function of $\mathbf{x}$, while quadratic discriminant function is a quadratic function of $\mathbf{x}$, thus such names.

To apply the classification rules with $\delta_k(\mathbf{x})$ and $q_k(\mathbf{x})$, parameters π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_k$ need to be estimated for each of the K classes. Let

$$\{\mathbf{x}_{1j}, j = 1, \ldots, n_1\}, \{\mathbf{x}_{2j}, j = 1, \ldots, n_2\}, \ldots, \{\mathbf{x}_{Kj}, j = 1, \ldots, n_K\}$$

be the training set from the observed population, where n_k, $k = 1, \ldots, K$, are the observation counts in the respective classes. The parameter estimates for π_k and $\boldsymbol{\mu}_k$ are calculated using the following formulas:

$$\hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{x}_{kj}, \quad k = 1, \ldots, K.$$

The common covariance matrix for the linear discriminant rule is estimated as the pooled sample covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_{pl} = \frac{1}{n - K} \sum_{k=1}^{K} (n_k - 1)\mathbf{S}_k,$$

where

$$\mathbf{S}_k = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} (\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)(\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)', \quad k = 1, \ldots, K.$$

For the quadratic discriminant rule $\boldsymbol{\Sigma}_k$ is estimated as $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}_k$, $k = 1, \ldots, K$, which gives the unbiased estimate.

When we compare logistic regression, linear and quadratic discriminant analysis, there is no clear preference as none of these methods dominates over the others. The results of each method will depend on the actual distribution of the predictors in each of the K classes together with sample size.

3 Data description

The University of Nottingham's School of Physics and Astronomy has built a database consisting of Raman spectroscopy and autofluorescence (AF) imaging data. The following section describes available data and is based on the article by Boitor et al. (2017) if not stated otherwise.

3.1 Tissue samples

The skin tissue database consists of samples obtained during the Mohs micrographic surgery at the Nottingham University Hospitals National Health Service (NHS) Trust and the Nottingham NHS Treatment Centre. During the surgery a thin tissue layer is excised for validation to make sure it is clear of tumor. For the skin tissue database the excised tissue layer is cut into two adjacent sections. One section of 10 μm is used as the standard of reference and stained by hematoxylin and eosin (H&E), while the other section is used for multimodal spectral histopathology measurements.

First, AF imaging is performed on the extracted tissue layer. This process provides an image of size $2cm \times 2cm$ (3200×3200 pixels), which is studied to select a smaller region of interest of size 200×200 pixels. For the region of interest, Raman microspectroscopy measurements are taken. For the training tissue samples the full region of 200×200 pixels is scanned. This is not the case for a final model used in a clinical setting, as collection of 40 000 Raman spectra would take too long time to gather. Instead a smaller group of sampling points are selected based on AF imaging. These tissue samples, where full measurements are taken, are also referred to as raster scans. 37 of these raster scans from batch 1 of the skin tissue database have been made available by the University of Nottingham's School of Physics and Astronomy for the analysis in this thesis. Out of 37 samples, 7 are put aside for validation. Out of the remaining 30 scans, 20 are from different AF

images, while 5 AF images have 2 regions of interest defined and fully scanned.

3.2 Data processing differences

The analysis of the AF images is one part of the integrated AF imaging and Raman spectrocopy method. It is used to select regions and priority points where Raman microspectroscopy is performed. While AF imaging is part of the full instrument process, the master's thesis focus is on the Raman microspectroscopy measurements.

In the available batch of data, there are some differences compared to the article by Boitor et al. (2017). Spectra for batch 1 data is interpolated in the range of $325cm^{-1} - 2081cm^{-1}$, where 1 024 Raman signal values are measured for each pixel. Also, there are some differences compared to the article in the signal to noise ratio (SNR) calculation. An SNR threshold allows to filter and keep the Raman spectra values where the signal is clear enough to make decisions. It has been adjusted compared to the value described in the article. SNR in the thesis is calculated as a comparison between two spectra regions. The signal part is calculated as the maximum range (the highest spectral value minus the lowest spectral value) between wavenumbers $1400cm^{-1}$ and $1512cm^{-1}$. The noise is calculated so that a normal distribution is fitted to the Raman spectra measurements between wavenumbers $1750cm^{-1}$ and $1800cm^{-1}$, from where a standard deviation estimate $\hat{\sigma}$ is obtained. The SNR value is eventually calculated as the signal part divided with two times the estimated standard deviation:

$$SNR = \frac{signal}{2\hat{\sigma}}.$$

The function used to calculate SNR is given in Appendix 1.

3.3 Mapping of the tissue classes

Tissue class mapping for the raster scans has been performed semi-automatically using histopathology and k-means clustering. k-means clustering allows to generate a number of clusters with similar Raman spectra to have large enough regions to map manually. The k-means image will be compared to the adjacent tissue section that was extracted during the tissue preparation process. H&E stained section will then be used to map clustered sections to provide tissue class information for raster scans. It is important to note that this sort of process can generate mislabelling to some degree, but not on a scale to address the issue in this thesis.

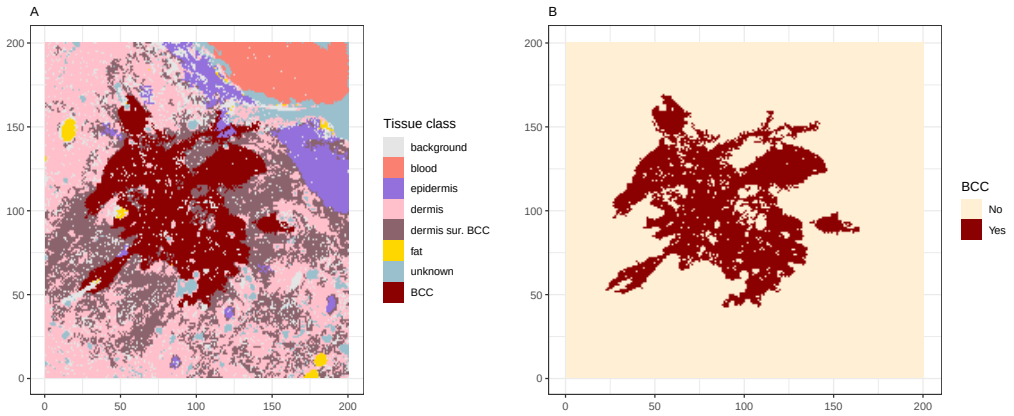


Figure 1. A) Sample with detailed mapping, B) Sample with binary mapping.

For tissue classes, there is available a binary mapping and also a more detailed version of it with multiple tissue classes. The binary mapping covers BCC positive and negative pixels, while the more detailed annotation provides 13 different tissue class types. Examples of mapped tissue samples can be seen in figure 1.

The set of 13 classes is defined as follows: 0 - *substrate/background*, 1 - *blood*, 2 - *epidermis*, 3 - *unknown*, 4 - *inflammation*, 5 - *dermis*, 5.5 - *dermis surrounding BCC*, 6 - *fat*, 7 - *muscle*, 8 - *blue dye*, 8.2 - *unknown*, 9 - *red dye*, 10 - *BCC*.

3.4 Spectral features

Every pixel in the region of interest has 1 024 Raman signal values measured at different wavenumbers. Since 1 024 feature variables are a lot to process without dimension reduction, 30 spectral features have been extracted from the Raman spectra to capture the most relevant information for different tissue classes. To better illustrate Raman spectra, an example for tissue classes *BCC*, *dermis* and *fat* can be seen in figure 2. According to the article by Boitor et al. (2017), *BCC* should have a higher intensity in a region between $788cm^{-1}$ and $1098cm^{-1}$ (black dashed lines) compared to the other tissue classes, while *dermis* should have a higher intensity in a smaller region between $851cm^{-1}$ and $950cm^{-1}$ (grey dashed lines).

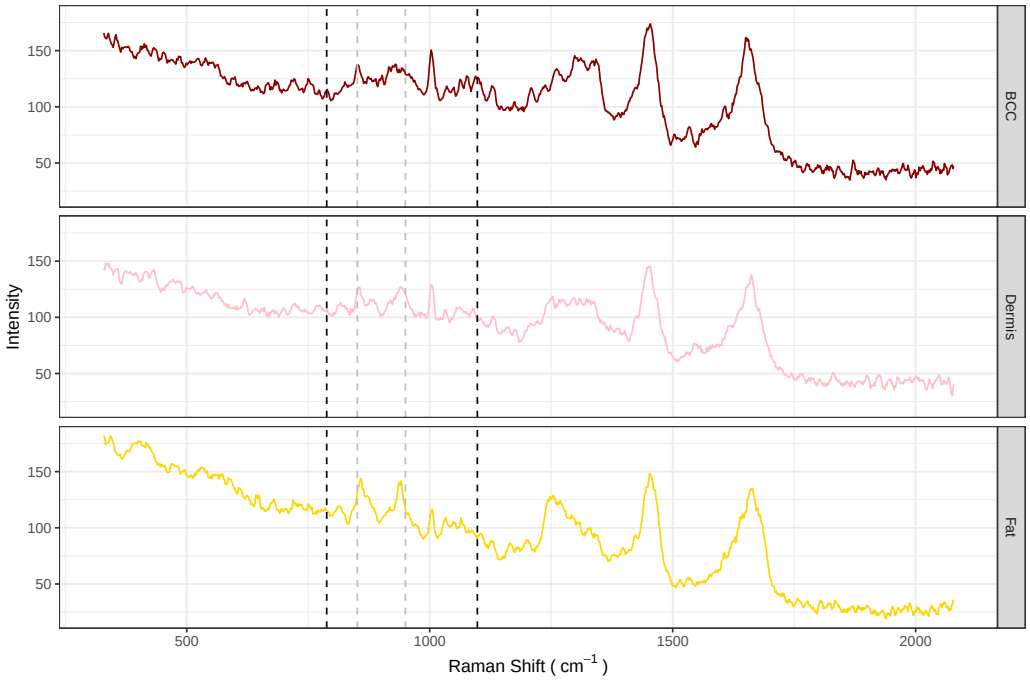


Figure 2. Raman spectra examples for BCC, dermis and fat cells.

Spectral features have been extracted from each Raman spectrum (that is, for each pixel) to provide useful details for the tissue classes. Feature variable calculation means that a certain wavenumber region on a Raman spectrum is taken and the

area between the Raman spectrum and a local baseline for that region is calculated. In other words, the features represent areas under the peaks for the certain Raman spectrum sections. Essentially, dissimilar tissue classes should have different peak/area distributions, so one could tell what sort of tissue class we are dealing with by looking at the Raman spectra, and thus the spectral features that represent them.

For each pixel of a tissue sample, 30 spectral features have been calculated. There is a smaller subset of features suggested by Radu Boitor that could potentially give a good enough tissue class separation. This smaller set of 15 variables will be observed as a potential feature subset to reduce complexity of the classification models.

In addition to the original set of features, the set of L_2-normalized features will be considered. The normalization is performed pixel-wise, each pixel is projected to a unit sphere. Although it is unusual to project each observation (pixel) to a unit sphere, it can help to reduce variability of the original spectral feature values, while maintaining proportions of spectral features within each pixel. Potential viability of this normalization will be further studied in the data cleaning section.

4 Data pre-processing

Tissue samples provided by the University of Nottingham are available in original format. The samples include all 200×200 pixels and their Raman spectroscopy measurements. Since all measurements are included, they may contain obscure values that should be analyzed separately or removed from the dataset. While feature calculation has already been performed for all these samples, any further data cleaning and processing has not been applied. Guidelines for data cleaning can partly be obtained based on what is described by Boitor et al. (2017), but we have also communicated with Radu Boitor personally.

The dataset is cleaned in three steps. First, all incomplete pixels where data is missing or no spectral features were calculated are removed. After that an SNR filter is applied to only include pixels with spectra that provide clear enough signal. Finally, Raman spectra measurements are observed carefully as they may vary a lot in intensity and might not provide reasonable input to classification models.

4.1 Incomplete measurements

There are 30 tissue samples available, each of size 200×200 pixels. First we are going to remove all the pixels where spectral feature calculation is incomplete. This first filtering will remove 43 439 pixels out of 1 200 000 total available ones. Two samples stand out with much higher removed pixel counts: one which has 10 037 pixels removed (out of 40 000) and the other one which has 6 205 pixels removed, see figure 3 for the respective images. We can see that most of the *blue dye* (69.7%) and *red dye* (78.7%) pixels got removed for the first sample, also 42% of the *blood* pixels were removed. For both of these outlying samples there is a high share of *blue dye* pixels that caused an abnormally high exclusion rate. Looking through all the 30 samples, we can observe a similar pattern for the pixels from *blue* and *red dye* classes: for 53.4% out of all *blue dye* pixels no spectral feature values have been

calculated due to unusual spectra behaviour, for the *red dye* class this percentage is even higher – 79.6%. When observing the full dataset of 30 samples, we can see that *blood* pixels don't stand out so much anymore – only 13% are removed. While 13% is still high compared to other tissue classes like *BCC* and *dermis* (both below 1%), it isn't as high as tissue sample 7 in figure 3 suggested.

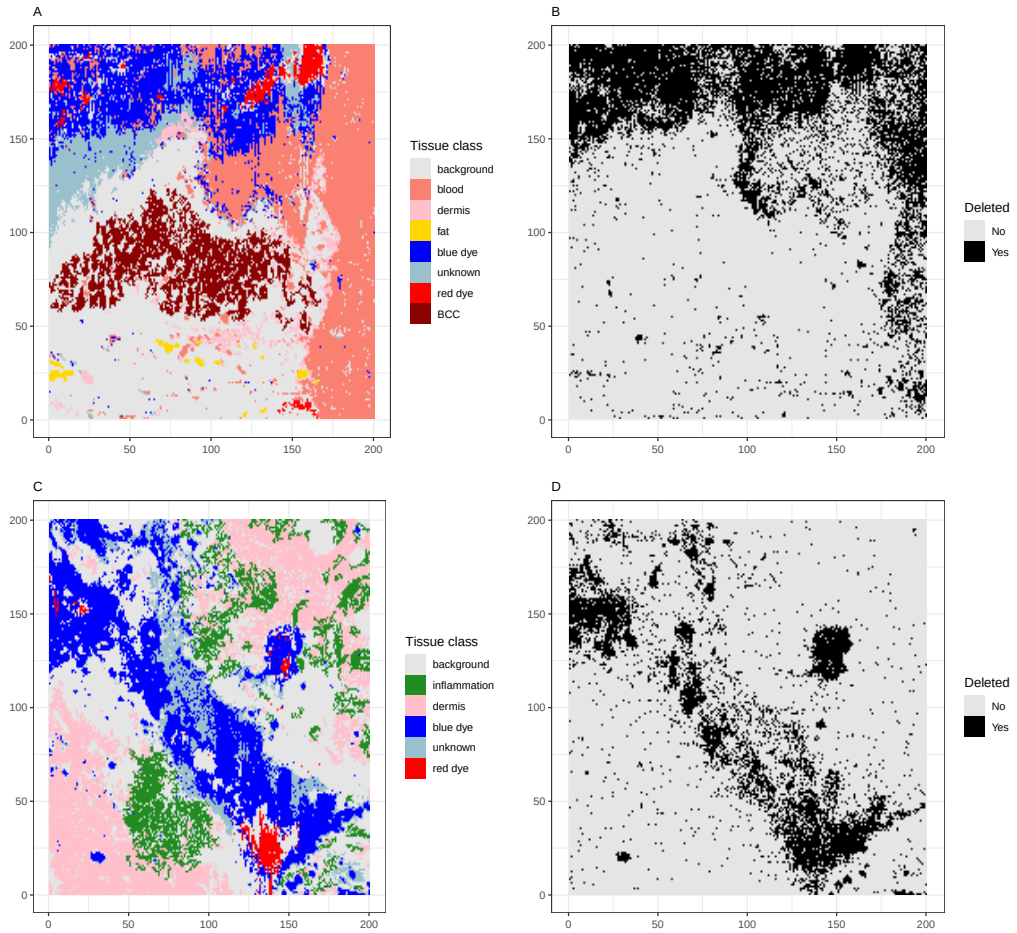


Figure 3. Tissue sample 7: A) detailed mapping, B) deleted pixels; Tissue sample 11: C) detailed mapping, B) deleted pixels.

After the removal of incomplete readings there are some tissue classes that are not so well represented and should also be considered closer: they should either be removed or reclassified under other labels. For example, class number 3 (*unknown*) remains present only in one sample. Since we don't know the actual tissue class

for these pixels (and can not estimate it), it would be wise to reconsider whether these pixels should be included into our training data at all. Similar issue applies for tissue class number 7 – *muscle*. While this class has a clear and understandable meaning, it is present only in 2 samples: one sample which has 6803 *muscle* pixels and another sample with only 7 pixels mapped as *muscle*. In case there is detailed mapping used as target variable, it would be wise to consider reclassification of this class. Similar approach was previously applied and explained by Boitor et al. (2017), where *epidermis*, *muscle* and *inflammation* were grouped under the class EMI.

After the removal of pixels with incomplete feature values, there are now only 254 *red dye* pixels remaining, since 78.7% got removed due to problems with spectral feature calculation. Since there are so few *red dye* pixels left, including these in the training set would create further confusion in modelling. That's why also these 254 pixels will be removed from further analysis. After also removing the classes *unknown* (class 3) and *red dye*, there are 1 156 561 pixels left.

4.2 Signal to noise ratio filtering

Signal to noise ratio filtering leaves us with less noisy Raman spectra where calculated spectral features describe peaks under the spectra better and should give a greater separation of the different tissue classes. For this purpose a suitable SNR threshold value is selected to remove observed spectra where noise overrides useful information, while at the same time keeping enough pixels for informative and accurate analysis. In Boitor et al. (2017) it is suggested that spectra with SNR value lower than 7 should be discarded, no specific explanation is given why such a threshold is selected. Since the SNR threshold value of 7 was based on a different batch of data and would leave out 68% of data pixels in our case, we will investigate this threshold further.

Exclusion rate of feature measurements for different tissue classes at range of SNR

values can be seen in figure 4. We can see that *blood*, *blue dye* and *background/-substrate* SNR values tend to be lower than for the other tissue types, which isn't too concerning if we are going to cut them out. As for others, *fat* definitely seems to be the best tissue type to measure with Raman spectroscopy as it appears to give readings with a really good signal, since drop off only starts at SNR value of 5. Considering this and also taking into account additional expert opinion from Radu Boitor (that for batch 1, we could continue with SNR cut-off at 4 or 5), we are going to continue with SNR values higher than 5 and exclude from the analysis all the pixels with SNR less than 5.

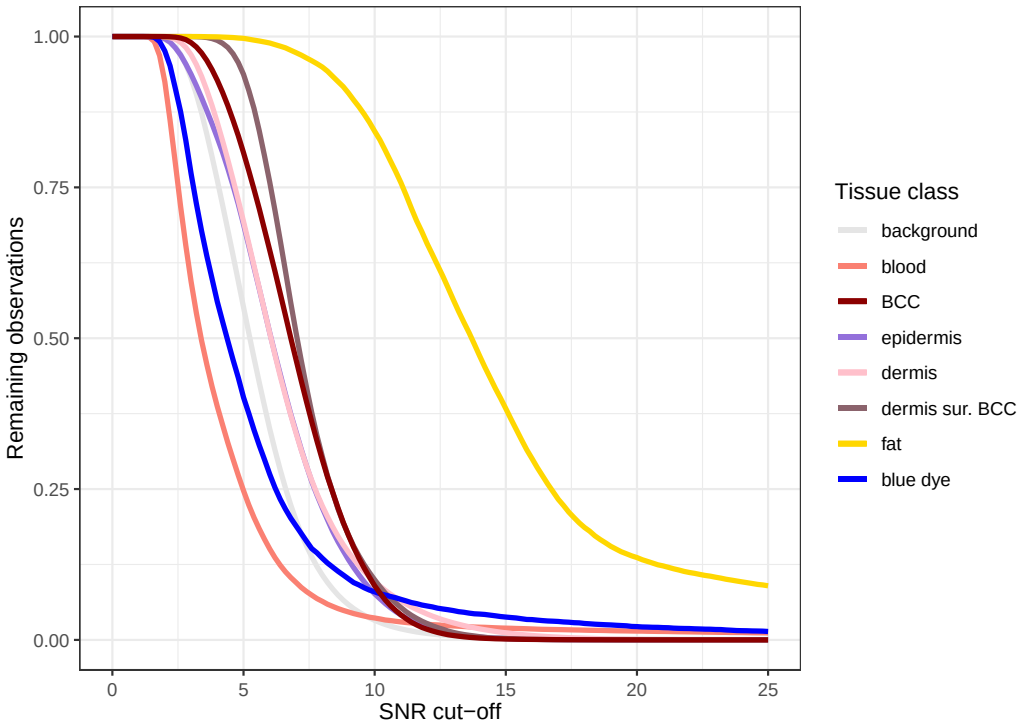


Figure 4. Signal to noise ratio cut-off at different levels.

Using this SNR filter and keeping values higher or equal to 5, we are going to remove additionally 399 145 pixels. This leaves us with 754 767 observations. It appears that 6 tissue samples were so noisy that more than half of all the pixels got removed during the filtering process, for 2 tissue samples more than 34 000 pixels

were removed (out of 40 000). These two samples can be seen in figures 5 and 6.

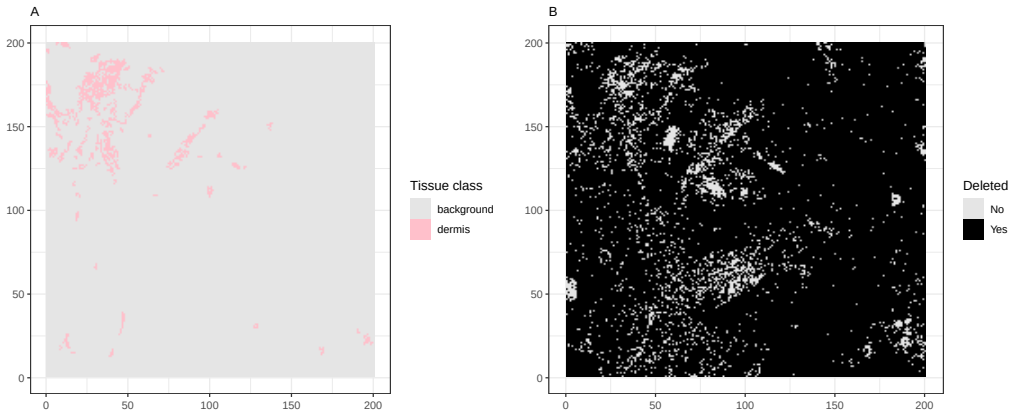


Figure 5. Tissue sample 3: A) detailed mapping, B) deleted pixels.

Observing tissue sample 3 shows that this sample didn't include much information on different tissue classes anyway, having only *substrate* and (little bit) *dermis* available. On the other hand, sample 27 has much more information. Having most of the pixels removed for this sample seems more serious and this might raise some questions about SNR calculation and raster scan quality.

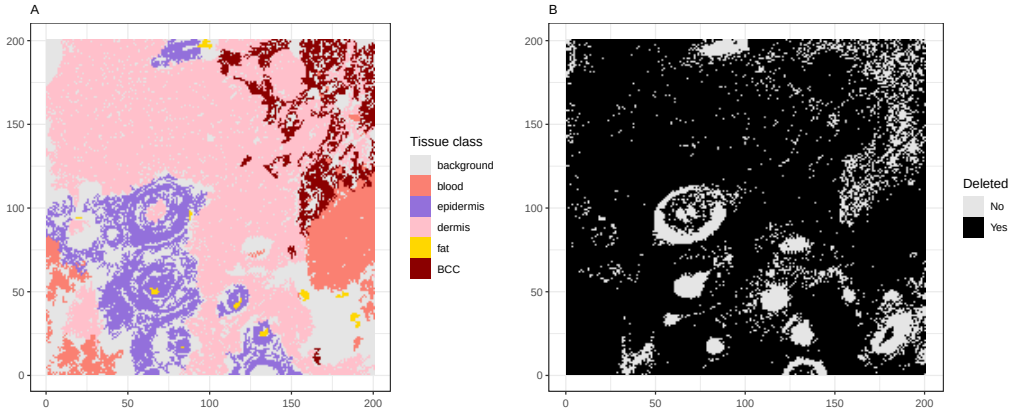


Figure 6. Tissue sample 27: A) detailed mapping, B) deleted pixels.

4.3 Inconsistent spectra intensities

The final pre-processing step concerns Raman spectra with much higher intensity levels (measurement values). It is observed that all tissue classes may have spectrum intensity measurements, where the whole spectrum intensity is much higher than for the majority of the spectra from the same class. This issue will be closer explored in this subsection. An example of this situation can be seen in figure 7, where Raman spectra for 20 pixels from the epidermis class are plotted. For easier analysis of this issue, the median of the intensity is calculated for each spectrum, this should give a general indication of how high the spectrum intensity is.

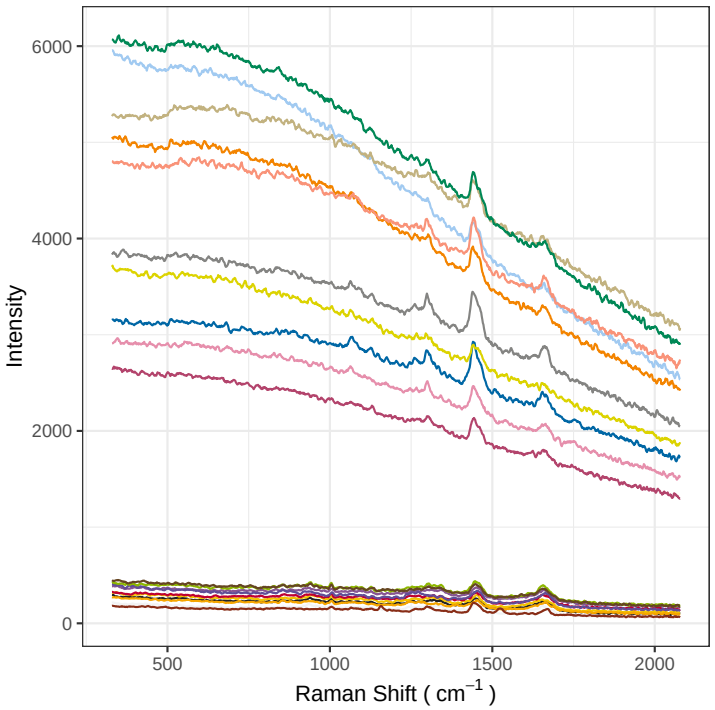


Figure 7. Epidermis Raman spectra for 10 high intensity pixels compared to 10 low intensity pixels.

The increased level of spectra intensities would not be that much of an issue, if it would not impact the values of spectral features that are used for classification. Unfortunately, it can also be seen in figure 7 that peaks in the spectra are much

more emphasized for spectra with higher intensity. This may result in bigger areas and thus, in bigger spectral values for the more intense spectra. To confirm the suspicion, distributions of spectral features can be observed for two different intensity regions. Figure 8 presents estimated densities of some spectral features for *epidermis*. The estimated density of pixels where the median intensity is below 500 is shown in green, whereas the estimated density of pixels where the median intensity is between 2 000 and 5 000 is yellow. It can be seen that higher median intensity will change the distribution of spectral feature values, and tends to give higher values of spectral features. If the intensity of spectra would have no effect on spectral features, one would expect similar distributions across any observed subsamples.

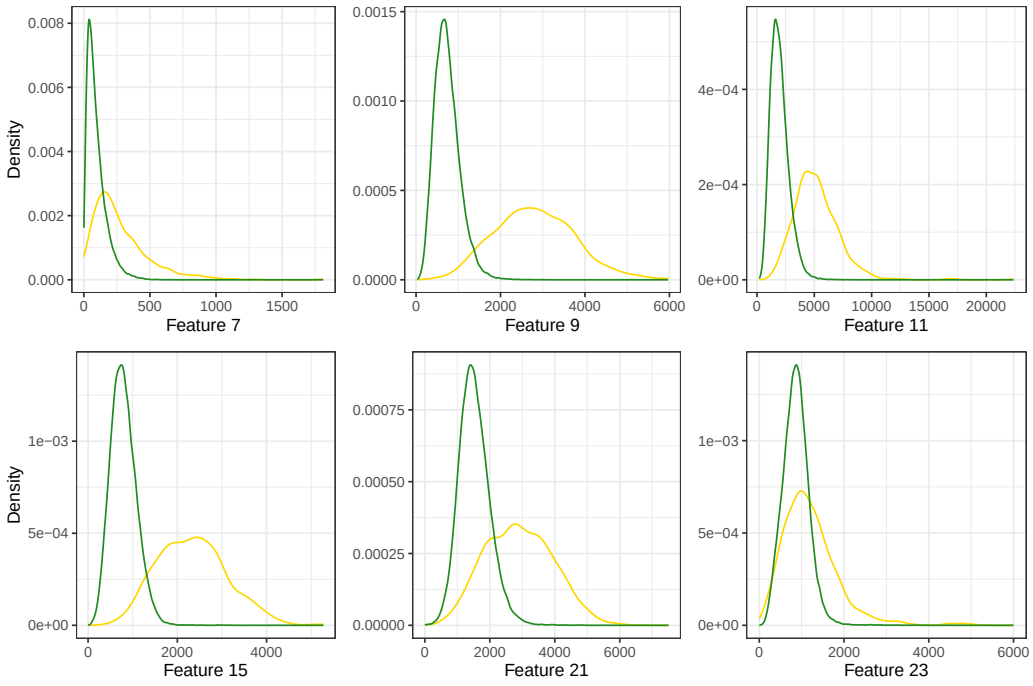


Figure 8. Comparison of estimated densities of 6 spectral features in two subsamples of the class epidermis.

In the previous section it was mentioned that L_2-normalized features could potentially be used to reduce the effect of variation of spectra intensities, and decrease

the variability of the spectral feature values. Figure 9 illustrates that normalization will reduce, but not completely eliminate such behaviour, meaning that spectra with higher intensity can still influence the behaviour of spectral features and they are not just acting as the scaled versions of each other that keep similar feature proportions no matter of the intensity level. Observe that only 638 pixels are available to estimate the density of *epidermis* for the higher intensity subset, but it should still be enough to raise a reasonable concern and be a target for validation if bigger dataset is available.

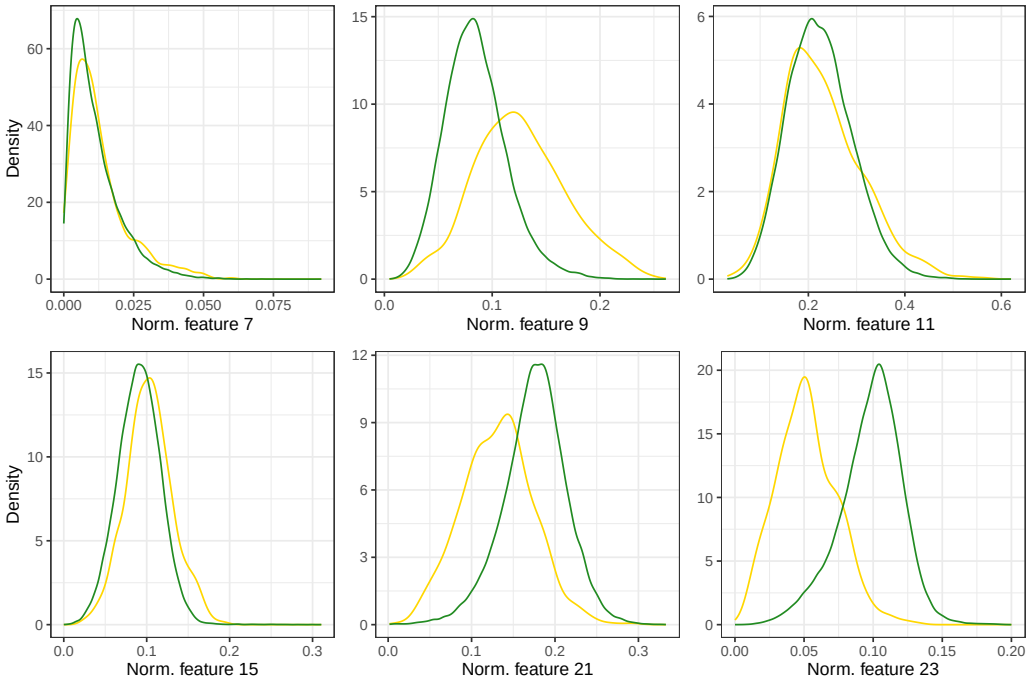


Figure 9. Comparison of estimated densities of 6 normalized spectral features in two subsamples of the class epidermis.

To conclude: pixels with spectrum intensity that is abnormally high should be removed from the training set. To find out the reasonable cut-off value, we can observe how many spectra have considerably higher intensity. In figure 10 we can see that the median intensity of the Raman spectra increases exponentially. This means that there are some measurements for raster scans where the whole Raman

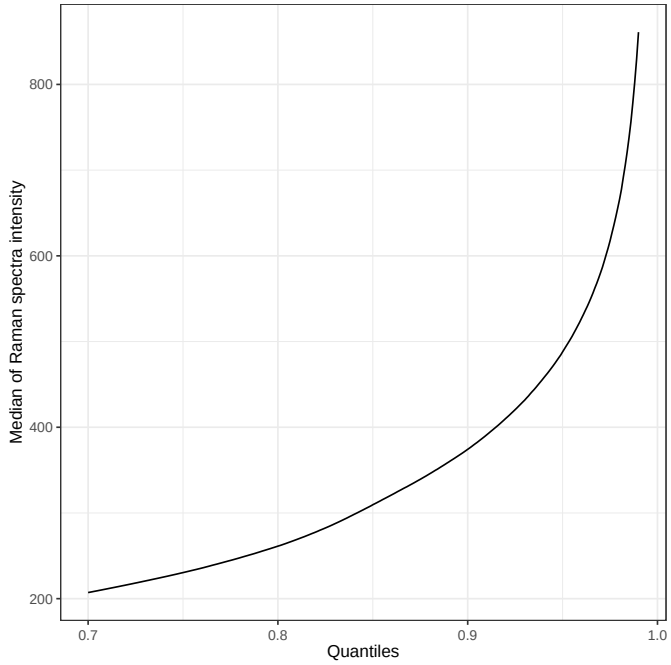


Figure 10. Median of Raman spectra intensity for dataset.

spectra are considerably higher than for most of the cases. Figure 10 shows data up until to the 99-percentile, thus the final 1% of Raman spectroscopy measurements has the median intensity value above 800, while 70% of the medians are all below 200. Although there are some differences across the tissue type classes, the general picture is the same, that is the median value inside the different tissue classes may vary a great deal.

Differences in Raman spectra intensities can be caused by several reasons, but if not dealt with, they may change classification results. It can be observed in figure 11 that higher median values definitely stand out for the region where *red dye* was present. While most of these really extreme cases are already removed in the step of missing feature value removal, there are still quite a few of such spectra with high intensity remaining. These pixels with high intensity are in the neighbourhood of already removed *red dye* pixels.

To be sure that we have the training data of best quality for our classification

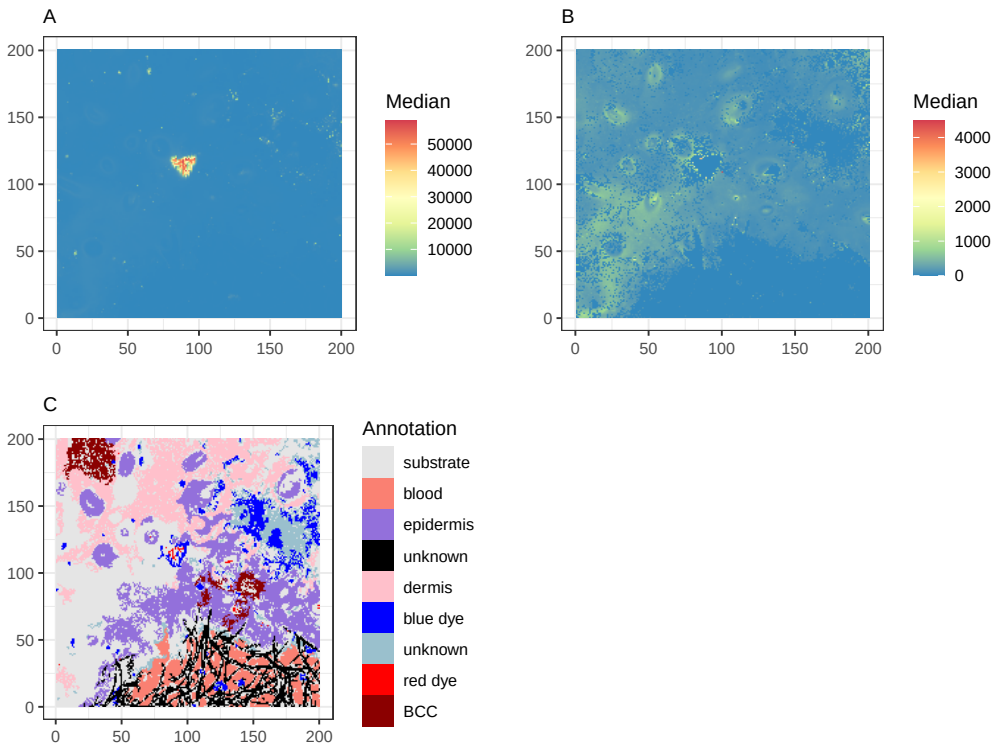


Figure 11. A) Raman spectra median values of the complete tissue sample B) Raman spectra median values of the already cleaned set C) detailed mapping of the tissue sample.

models, we suggest to remove all the pixels where the median intensity exceeds 500. This would keep 95% of the current pixels, but would throw out a lot of pixels with high intensity spectra, what could eventually alter the results. This reduces the dataset to 719 562 pixels.

4.4 Further dataset limitations

After data pre-processing we are left with a reasonable set of pixels, but we should also consider tissue samples in more general terms. There are two samples that contain only *BCC* class and if left in, they will heavily change the prior probabilities if we are going to estimate discriminant analysis parameters. These two tissue

samples will be left out from the final dataset to avoid overly optimistic *BCC* rate. This also means that we are going to reduce the number of *BCC* pixels from 139 855 to 86 014.

In the end there are 655 432 pixels left with the tissue classes shown in table 1. Table 1 also illustrates that *dermis surrounding BCC* and *fat* have the most reliable Raman microspectroscopy measurements as these classes have smallest number of pixels removed. All bigger and more common tissue classes are also relatively well kept, while *blood*, *blue dye* and *unknown (8.2)* are classes with the highest percentage of pixels removed.

Table 1. Frequency table of remaining and initial pixels.

Class	**Remaining**	**Initial**	**Remaining %**
Substrate	177 494	369 169	48.1%
Blood	10 830	56 626	19.1%
EMI	62 510	100 627	62.1%
Dermis	256 412	395 165	64.9%
Dermis sur. BCC	36 355	38 930	93.3%
Fat	18 464	23 703	77.9%
Blue dye	2 514	19 379	13.0%
Unknown (8.2)	4 859	17 300	28.1%
BCC	86 014	174 838	49.2%
Red dye / Unknown (3)	0	4 263	0%
Total	655 432	1 200 000	54.6%

5 Discrimination of basal cell carcinoma

This section describes the process of finding the best reference model for classification, the classification methods tested in the thesis are logistic regression analysis, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). For all the three approaches classification rules have been found with the training set, where different feature variable selections have been considered. The models have been trained with the following variable selections:

1) all the 30 original spectral feature variables,

2) 15 original spectral feature variables chosen by the expert opinion,

3) selected number of principal components of the original features.

Furthermore, the models have also been trained with pixel-wise normalized spectral features, for pixel-wise normalized variables the same variable selection procedure has been applied. To find the best method, model accuracy and specificity at the target sensitivity of 95% are compared.

The pixel-wise normalization of the feature variables means that for each pixel, each variable value has been divided with the Euclidean length of the feature vector, thus L_2-normalization is used. This means that each pixel is projected to a sphere with radius 1. This normalization is considered for comparison, since it has been a standard approach in the previous studies.

Model training is done on the training tissue samples, where 10-fold cross-validation is performed to get the initial results. Cross-validation folds are generated randomly out of all pixels as they are considered independent. After inspecting the results of the initial modeling stage, the most promising approaches can be selected (regarding classification method and variable selection). The models with the most promising set-up are then trained on the full training set and then validated against

the validation tissue samples. In this way, the final reported results should indicate well enough how the chosen model will perform in classification with new data.

In the following subsections all the models are fitted to the

1) full cleaned dataset,

2) dataset where the pixels of the *substrate/background* class are removed.

These classification datasets are considered with three possible feature variable selections mentioned above. An example of R-code used to work out the classification rules using 10-fold cross-validation is given in Appendix 1. To reiterate, all 30 spectral features are used in classification to include as much information as possible. Although for more complex models with dependent pixel structure (that could potentially be studied in the follow up works) 30 features are quite a lot. Observe also that a few of the feature variables are highly correlated. The correlation matrices of both the raw features and pixel-wise normalized features are given in Appendix 2. To reduce the number of variables, the set of 15 spectral features is used. These 15 spectral features have been suggested by Radu Boitor based on biomedical knowledge. For selection of the third subset of feature variables, principal component analysis is performed and the number of principal components to cover at least 99% of the total variance is used for model fitting.

5.1 Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction technique. The purpose is to derive a reduced set of orthogonal linear projections of a collection of correlated variables. These new variables (linear combinations of the original variables) are ordered by decreasing variance. Selection of these new variables that cover a required amount of the total variance can then be used in analysis instead. (Izenman, 2009)

A short description of PCA is based on the book by Izenman (2009). Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$' denote the vector of p original variables with $E\mathbf{X} = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$. PCA is based on the spectral decomposition of $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as

$$\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}',$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}$ (ordered from largest to smallest) and $\mathbf{V}$ is the matrix of the corresponding normalized eigenvectors (in columns), $\mathbf{V}\mathbf{V}' = \mathbf{I}_p$. The principal components $\boldsymbol{\xi} : p \times 1$ are defined as

$$\boldsymbol{\xi} = \mathbf{V}'(\mathbf{X} - \boldsymbol{\mu}).$$

Thus, to perform PCA in practice, we need to calculate the sample mean and covariance matrix ($\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}, \hat{\boldsymbol{\Sigma}} = \mathbf{S}$), and find the spectral decomposition of $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{V}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{V}}'$. Then the sample principal components for given $\mathbf{X} = \mathbf{x}$ can be calculated as $\hat{\boldsymbol{\xi}} = \hat{\mathbf{V}}'(\mathbf{x} - \bar{\mathbf{x}})$.

To illustrate the nature and behaviour of principal components in the training dataset, PCA is performed with all the 30 original spectral features and with pixel-wise normalized features. Table 2 shows the PCA results, where 30 original features were used. The first row of the table (Var) shows how much of the total variance is explained by each principal component (PC). The second row (Cum) gives the cumulative proportion of variance explained by the respective number of PCs. It can be seen that the first and second principal component cover 86.6% of the total variance. Considering the amount of initial spectral features these two are very influential. As PCA is heavily influenced by the scale of the variables, it can be immediately seen that some of the spectral features (areas under the peaks) have much higher variance and drive the large proportion of variance explained by the first PCs. For example, the first PC is mostly represented by three features –

feature 25, feature 26 and feature 11 (coefficients in the first eigenvector 0.6258, 0.3797 and 0.3504, respectively). The rest of the feature variables have a much smaller impact on PC1. The same can be said about the second PC, where the same features in a different combination dominate. A few of the first principal components were also studied visually, but *BCC* class couldn't be separated as it merges into the rest of the tissue classes. However, it is interesting to note the shape of the *BCC* class pixels on the scatter plot of the first two PCs. It can be observed in figure 12 that *BCC* pixels seem to form two separate groups with different directions.

Table 2. PCA of original spectral features.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**...**	**PC8**	**PC9**	**...**
Var	0.618	0.248	0.067	0.024	0.012	...	0.004	0.003	...
Cum	0.618	0.866	0.933	0.957	0.970	...	0.989	0.992	...

Table 3. PCA of pixel-wise normalized spectral features.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**...**	**PC15**	**PC16**	...
Var	0.520	0.236	0.090	0.046	0.032	...	0.004	0.003	...
Cum	0.520	0.756	0.846	0.892	0.924	...	0.989	0.992	...

PCA results with pixel-wise normalized spectral features are presented in table 3. We can see that 16 principal components are needed to explain 99% of the total variance. The same cumulative proportion of variance for the original spectral features only required 9 PCs. Although the first two PCs do not explain as much of the total variance as for the original features, the cumulative variance still grows quickly. Once again, visual observation of the scatter plot of the first two PCs doesn't suggest a simple discrimination boundary for the *BCC* class. This might also suggest that finding a good discrimination rule might be complicated.

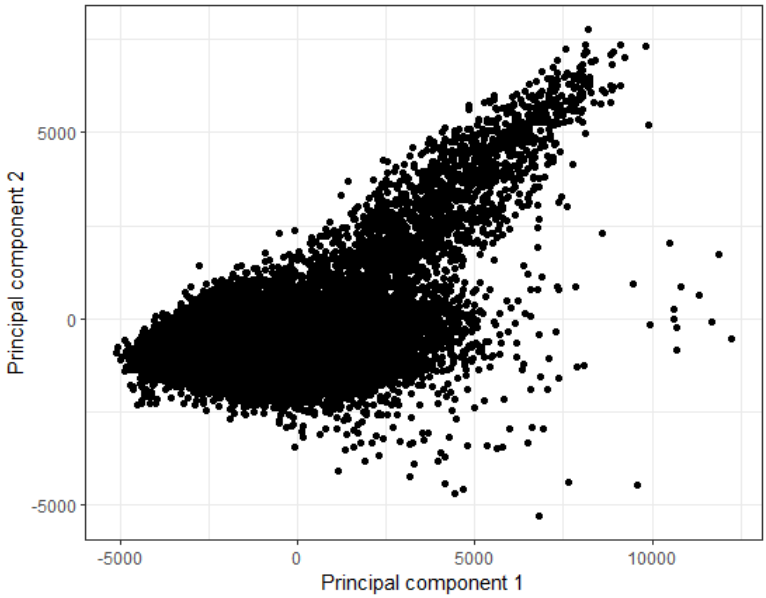


Figure 12. Scatter plot of the first two PCs for BCC class.

5.2 Classification on full cleaned dataset

Multiple different combinations of classification methods and feature selections are studied. A 10-fold cross-validation on the training data is performed across all combinations. The binary response variable (BCC versus other tissue classes) is used in classification models. The results obtained with different models are compared to select out for validation a few of the set-ups that perform best. To measure the performance of the combinations, sensitivity, specificity, accuracy and balanced accuracy are used. Sensitivity shows the proportion of correctly classified positive cases across all positive cases, while specificity shows the proportion of correctly classified negative cases across all negative cases. Accuracy is the share of correctly classified observations and balanced accuracy is the arithmetic mean of sensitivity and specificity. In the context of the thesis, sensitivity is given as

$$\text{sensitivity} = \frac{\text{number of pixels correctly classified as } BCC}{\text{number of } BCC \text{ pixels}},$$

and specificity is given as

$$\text{specificity} = \frac{\text{number of pixels correctly classified as not } BCC}{\text{number of not } BCC \text{ pixels}}.$$

Sensitivity is the most important measure as we want to classify correctly as many BCC pixels as possible: not discovering the remaining cancer cells could have serious consequences for the patients. That's why the target sensitivity of 95% is set by the researchers at the University of Nottingham, and this target is also considered in the thesis. A target sensitivity can be reached with the adjustment of the posterior probability threshold. An observation $\mathbf{X} = \mathbf{x}$ is classified as BCC when

$$P(Y = BCC|\mathbf{X} = \mathbf{x}) > threshold.$$

Observe that the default threshold value of the Bayes classification rule is 0.5 when we have a binary response variable Y. An increase in the sensitivity comes at the cost of the specificity. Too extreme requirements for one or the other may result in a threshold probability value that is very high or very low, meaning that it is difficult to discriminate the classes. The probability threshold of 0.5 is still used in the first step of model validation, then the two possible errors have the same cost.

Table 4 illustrates the results when the threshold is not adjusted (it is 0.5). It can be seen that LDA and logistic regression tend to give a higher specificity, while QDA provides a much higher sensitivity. LDA and logistic regression behave rather similarly, which is expected as they both are linear in their parameters. While LDA has a bit lower sensitivity compared to logistic regression for the set-ups with the original features, the methods provide almost the same results for the set-ups with the normalized features. It is hard to compare QDA to the rest of the classification methods using accuracy (proportion of correctly classified pixels). That's why the balanced accuracy ((sensitivity + specificity)/2) is also provided. It illustrates how well a method classifies both classes (BCC and non-BCC). The best accuracy of 90.27% is obtained for the logistic regression model with the 30 normalized features,

Table 4. Results at the probability threshold of 0.5.

Features	Measures (%)	LDA	QDA	LR
Original 30	**Accuracy**	88.47	58.80	89.87
	Balanced Acc.	65.03	73.36	69.01
	Sensitivity	33.23	93.09	40.71
	Specificity	96.82	53.62	97.30
Normalized 30	**Accuracy**	89.66	79.42	90.27
	Balanced Acc.	72.29	81.97	71.78
	Sensitivity	48.74	85.43	46.70
	Specificity	95.84	78.51	96.85
Expert 15	**Accuracy**	88.08	63.25	88.81
	Balanced Acc.	62.66	75.16	64.47
	Sensitivity	28.18	91.31	31.45
	Specificity	97.13	59.01	97.48
Normalized Expert 15	**Accuracy**	89.01	81.93	89.46
	Balanced Acc.	69.01	81.99	68.14
	Sensitivity	41.89	82.06	39.24
	Specificity	96.13	81.91	97.04
Original PCA 99%	**Accuracy**	86.03	59.56	86.64
	Balanced Acc.	50.74	73.72	54.13
	Sensitivity	2.87	92.92	10.05
	Specificity	98.60	54.52	98.21
Normalized PCA 99%	**Accuracy**	89.52	79.29	90.10
	Balanced Acc.	71.54	82.28	71.05
	Sensitivity	47.16	86.28	45.23
	Specificity	95.92	78.28	96.87

while the best balanced accuracy of 82.28% is achieved with QDA using principal components of normalized features.

Models with normalized features provide a bit better results in general compared to the same models with the original features. All feature-method combinations provide higher accuracy and balanced accuracy compared to their original feature counterparts. While accuracy improvements are not so noticeable, balanced accuracy increases a lot.

Table 5 gives the comparison between the different models when the posterior probability threshold is adjusted to reach the target sensitivity of 95%. The probability

Table 5. Results at the target sensitivity of 95%.

Features	Measures	LDA	QDA	LR
Original 30	**Threshold**	0.0423	0.112	0.056
	Accuracy (%)	57.52	53.19	67.01
	Specificity (%)	51.85	46.87	62.78
Normalized 30	**Threshold**	0.0277	0.0131	0.0382
	Accuracy (%)	60.95	56.25	65.75
	Specificity (%)	55.81	50.39	61.33
Expert 15	**Threshold**	0.0451	0.11	0.0526
	Accuracy (%)	53.93	53.62	60.51
	Specificity (%)	47.73	47.37	55.30
Normalized Expert 15	**Threshold**	0.0296	0.026	0.0377
	Accuracy (%)	56.18	56.93	60.93
	Specificity (%)	50.31	51.18	55.79
Original PCA 99%	**Threshold**	0.0797	0.228	0.0688
	Accuracy (%)	57.04	53.54	59.89
	Specificity (%)	51.30	47.28	54.58
Normalized PCA 99%	**Threshold**	0.0324	0.0284	0.0391
	Accuracy (%)	62.20	56.87	65.27
	Specificity (%)	57.25	51.11	60.78

threshold was reduced for all of the set-ups as the threshold of 0.5 didn't provide high enough sensitivity for any of the method and feature combinations. Logistic regression with all 30 original spectral features provides the best accuracy in conjunction with specificity, when the target sensitivity of 95% is set. However, to reach that high sensitivity, the probability threshold needs to be decreased a lot. When the conditional probability of the logistic regression model exceeds 0.056, that is $P(Y = BCC|\mathbf{X} = \mathbf{x}) > 0.056$, an observation is going to be classified as *BCC*. This is a really low probability and the classification rule is only going to reject the cases where it is almost certain that the pixel is not *BCC*. It can be also noted that the methods which performed best at the threshold value of 0.5 are not the best at the target sensitivity of 95%. Though, the set-ups with the normalized features are still usually better as it was with the threshold of 0.5. The only exception is logistic regression, when the original 30 features are compared to

the normalized 30 features. For logistic regression the original 30 spectral features performed better and it also gives the best model at the target sensitivity. Observe also that specificity of 60% is only exceeded with the logistic regression classifier.

Setting the high target sensitivity for pixel-wise normalized features causes really low probability threshold classification rules. To reach the target sensitivity of 95%, all the methods must have the probability threshold reduced to below 0.04. Analyzing the behaviour of original and normalized spectral features in the classification methods suggests that, the pixel-wise normalization helps to improve the methods at the default probability threshold. However, to reach the target sensitivity of 95%, their probability threshold must be considerably lower compared to the original feature methods.

Table 6. Confusion matrix of LDA with 30 original features for the detailed mapping.

Act \Pred	**0**	**1**	**2**	**5**	**5.5**	**6**	**8**	**8.2**	**10**
0 (%)	54.05	0.18	5.95	26.80	0	0.53	0.83	1.85	9.82
1 (%)	23.55	47.08	0.54	11.94	0	0.47	7.95	8.25	0.20
2 (%)	26.89	0.6	45.49	13.14	0	0.3	0.07	0.11	13.40
5 (%)	17.66	0.02	1.7	78.25	0	0.08	0.05	0.17	2.05
5.5 (%)	55.67	0	0.12	42.38	0	0	0	0.01	1.82
6 (%)	15.52	1.4	2.31	7.25	0.01	71.92	0	0.01	1.59
8 (%)	12.69	0	0.20	1.67	0	0	46.90	37.71	0.84
8.2 (%)	41.35	0.72	0.56	9.08	0	0.04	14.96	31.08	2.22
10 (%)	37.58	0	7.68	3.42	0	0	0.03	0.08	51.20

We can conclude that the best model is the logistic regression model with all 30 original features. It has the model accuracy of 67.01% and specificity of 62.78% at the target sensitivity level. These quantities are still quite low and better alternatives should be explored. The detailed mapping of tissue classes can be utilized to better understand the problem. Table 6 gives the confusion matrix of LDA with the original 30 features and the detailed mapping. Table 6 shows that for given true class, the percentages in the respective row present the proportions of pixels

predicted to each class. The percentages in column 0 show that substantial amount of pixels is always predicted to the *background* class. In the article by Kong et al. (2013), it is stated that the *background* class could be left out from the classification of Raman microspectroscopy measurements as it can be pre-filtered by AF imaging. This means that sampling points used for Raman measurements should never fall on pixels of the *background/substrate* class. Although it is quite a strong restriction, discrimination rules without the *background* class will be studied closer in the next subsection.

5.3 Classification on dataset without the background class

After the removal of *substrate/background*, a similar analysis to the previous subsection is carried out. Although PCA is also performed again, the results are similar to the dataset with the *background* class included and are not separately analyzed here. The only difference regarding PCA is that 99% of the total variance is explained by 8 and 15 components respectively in the analysis with the original and normalized spectral features.

Table 7 gives the results at the threshold probability of 0.5. For most of the set-ups, there is a clear improvement in model accuracy, balanced accuracy and sensitivity. LDA and logistic regression have much higher sensitivity values compared to table 4. The same can be said about QDA for specificity. While there is a decrease in the accuracy for three feature-method combinations, the models without the *background* class have now better balanced accuracy, meaning that the increase in the sensitivity is bigger than the loss in specificity for these cases. The best accuracy of 91.11% (improvement of 0.84) is once again for the logistic regression model with the 30 original spectral features. The best balanced accuracy at the probability threshold of 0.5 is for QDA with the 30 normalized features – 86.23% (improvement of 3.95).

Table 7. Results at the threshold probability of 0.5 on the dataset where the background class is removed.

Features	Measures (%)	LDA	QDA	LR
Original 30	**Accuracy**	88.01	68.41	91.11
	Balanced Acc.	73.92	78.06	81.99
	Sensitivity	51.89	93.13	67.74
	Specificity	95.94	62.99	96.24
Normalized 30	**Accuracy**	90.32	83.93	91.02
	Balanced Acc.	81.86	86.23	82.52
	Sensitivity	68.64	89.82	69.23
	Specificity	95.08	82.64	95.81
Expert 15	**Accuracy**	87.24	70.66	89.49
	Balanced Acc.	71.54	79.20	77.69
	Sensitivity	47.01	92.53	59.25
	Specificity	96.07	65.86	96.13
Normalized Expert 15	**Accuracy**	89.36	84.74	90.01
	Balanced Acc.	79.08	85.72	79.78
	Sensitivity	63.02	87.25	63.79
	Specificity	95.14	84.19	95.77
Original PCA 99%	**Accuracy**	82.13	69.14	85.66
	Balanced Acc.	56.96	78.39	69.06
	Sensitivity	17.62	92.84	43.11
	Specificity	96.29	63.94	95.00
Normalized PCA 99%	**Accuracy**	90.13	83.64	90.69
	Balanced Acc.	81.40	86.09	81.67
	Sensitivity	67.75	89.92	67.57
	Specificity	95.04	82.26	95.76

The selection of the best method (reference method) is based on the results at the target sensitivity of 95%. There are now several set-ups for which specificity is above 60% and logistic regression even exceeds 70% for some of the cases. These results are given in table 8. For most of the models, there is a slight increase in the probability threshold values. The best classification result is still provided by logistic regression with the 30 original spectral features. The best method has 77.6% of accuracy and 73.78% of specificity at the target sensitivity. This means that out of all *BCC* pixels in the training set, it classified correctly 95% of

Table 8. Results at the target sensitivity of 95% on the dataset where the background class is removed.

Features	Measures	LDA	QDA	LR
Original 30	**Threshold**	0.064	0.11	0.0891
	Accuracy (%)	68.04	62.55	77.6
	Specificity (%)	62.13	55.43	73.78
Normalized 30	**Threshold**	0.0336	0.0475	0.0652
	Accuracy (%)	72.50	70.02	76.59
	Specificity (%)	67.56	64.54	72.54
Expert 15	**Threshold**	0.0635	0.167	0.0725
	Accuracy (%)	63.45	63.38	70.27
	Specificity (%)	56.53	56.44	64.84
Normalized Expert 15	**Threshold**	0.0326	0.058	0.0571
	Accuracy (%)	65.76	68.64	70.84
	Specificity (%)	59.34	62.85	65.53
Original PCA 99%	**Threshold**	0.1249	0.228	0.0974
	Accuracy (%)	67.51	62.58	69.49
	Specificity (%)	61.48	55.47	63.89
Normalized PCA at 99%	**Threshold**	0.036	0.0685	0.0627
	Accuracy (%)	72.28	69.14	75.12
	Specificity (%)	67.29	63.47	70.76

them. At the same time this classification rule also correctly captured 73.78% of pixels that were not related to tumor. It is still worth mentioning that the rule with such target sensitivity classifies all the pixels with the posterior probability $P(Y = BCC|\mathbf{X} = \mathbf{x}) > 0.0891$ as *BCC*. This means once again that the tumor free pixels are classified correctly, when it is utmost certain that they are indeed free of tumor.

5.4 Selection and validation of reference models

Validation process is carried out for two classification models that performed well on the cross-validated training data. Selected classification rules are:

1) logistic regression with the 30 original spectral features on the dataset, where the *background* class is removed,

2) linear discriminant analysis with the 30 normalized spectral features on the dataset, where the *background* class is removed.

Two different classification rules are validated to cover different methods and both the cases with the original and the pixel-wise normalized features. The logistic regression model has three of the best classifiers, LDA with the normalized features is the next best (that is not logistic regression).

The logistic regression model with the 30 original spectral features is trained on the cleaned training data without the *substrate/background* class. Due to the fact that there are so many training pixels, only 1 explanatory variable is not significant and is removed from the logistic regression model. Software R output of the logistic regression model is given in Appendix 3. This model is used for the classification of the validation data. The validation data is pre-processed with the same rules as the training data. Table 9 presents the validation results for the logistic regression model across all 7 validation samples. It can be immediately seen that sensitivity is close to 100% for almost all the samples. Thus, we can draw the conclusion that the threshold probability of 0.0891 that yielded 95% of the sensitivity for the training dataset is too low for the validation data. Although it is good to classify correctly all *BCC* pixels, a lot of tumor free pixels are also classified as tumor positives. Sample 2 is a tumor free validation sample. Thus, sensitivity couldn't be calculated. The logistic regression classification rule for sample 2 is still going to classify 10.29% of the classified pixels as *BCC*. In practice, this would mean that another tissue layer is going to be extracted, since the patient is not considered tumor-free.

High sensitivity results stand out also for LDA with the 30 normalized spectral features. The linear discriminant classifier is also trained on the cleaned training

Table 9. Validation results for logistic regression on 30 original spectral features.

	Accuracy (%)	**Sensitivity (%)**	**Specificity (%)**
Sample 1	80.67	98.6	62.17
Sample 2	89.71	-	89.71
Sample 3	68.16	99.46	57.78
Sample 4	64.85	98.39	58.6
Sample 5	66.21	99.94	32.45
Sample 6	52.87	94.72	42.5
Sample 7	78.35	99.75	70.44
Overall	71.48	99.1	59.53

data without the *background* class and the pixels, for which the posterior probability exceeds 0.0336, are classified as *BCC*. Table 10 shows the results of the LDA classifier on the validation samples. LDA shows lower accuracy for 6 out of 7 validation samples compared to the logistic regression model results, it is expected considering the training data results.

Table 10. Validation results for LDA on 30 pixel-wise normalized spectral features.

	Accuracy (%)	**Sensitivity (%)**	**Specificity (%)**
Sample 1	75.79	99.13	51.7
Sample 2	75.37	-	75.37
Sample 3	67.57	99.28	57.06
Sample 4	58.49	97.95	51.13
Sample 5	71.3	99.68	46.9
Sample 6	48.55	95.33	36.96
Sample 7	77.59	98.31	69.92
Overall	69.51	98.91	56.79

As it was already mentioned, the target sensitivity of 95% is high, which caused the probability thresholds to be too conservative for the validation data. In hindsight, if the target sensitivity for the training data would be 90%, the probability threshold for the logistic regression model with the 30 original spectral features will be 0.1959, which is a more reasonable value. Then the overall sensitivity for the validation

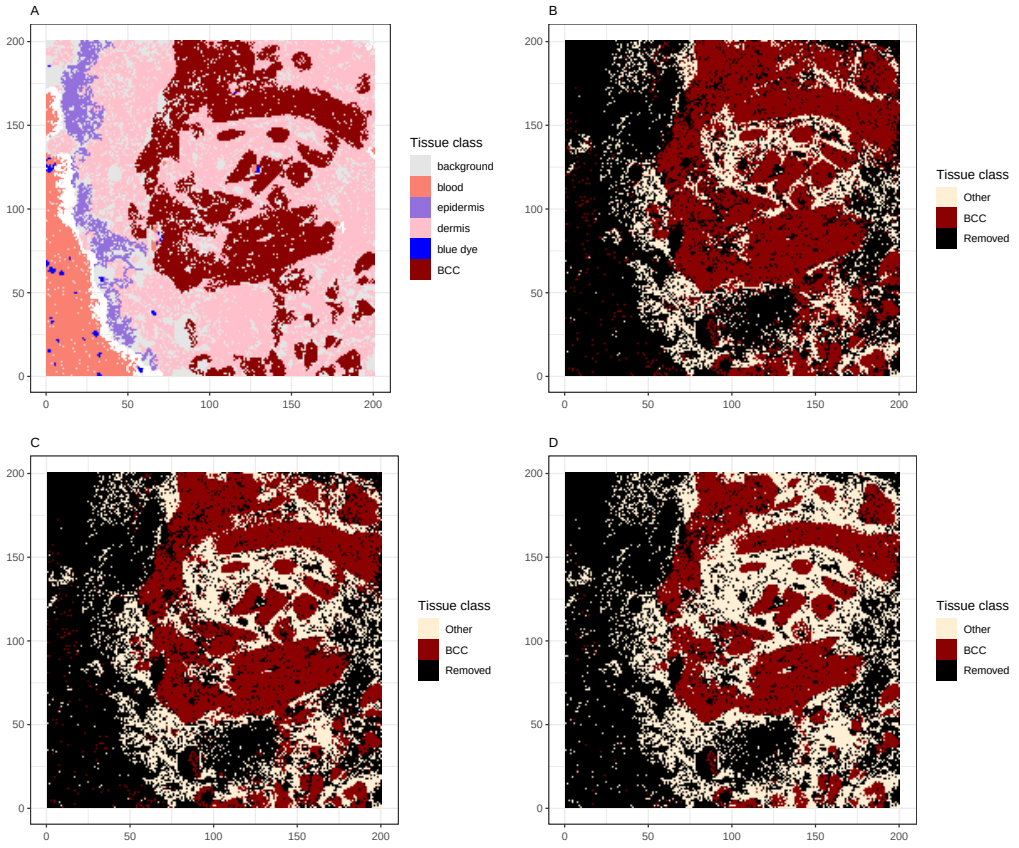


Figure 13. A) Detailed mapping of the validation sample 1 B) Logistic regression with 30 original features – trained to sensitivity of 95% C) trained to sensitivity of 90% D) probability threshold of 0.5.

data is 96.62% and specificity is 75.56%. However, we observe currently full raster scans (200×200 pixels) for which the target sensitivity of 95% may appear too high as it is unlikely to miss whole *BCC* regions. The actual model used in the clinical setting must be able to recognize a smaller group of *BCC* pixels, so the high sensitivity could then be justified.

Visualized classification results of validation sample 1 for the logistic regression model can be seen in figure 13. In addition to the detailed mapping of the validation sample, 3 different classification results are displayed: target sensitivity of 95%, target sensitivity of 90% and the default model with 0.5 as the posterior probability

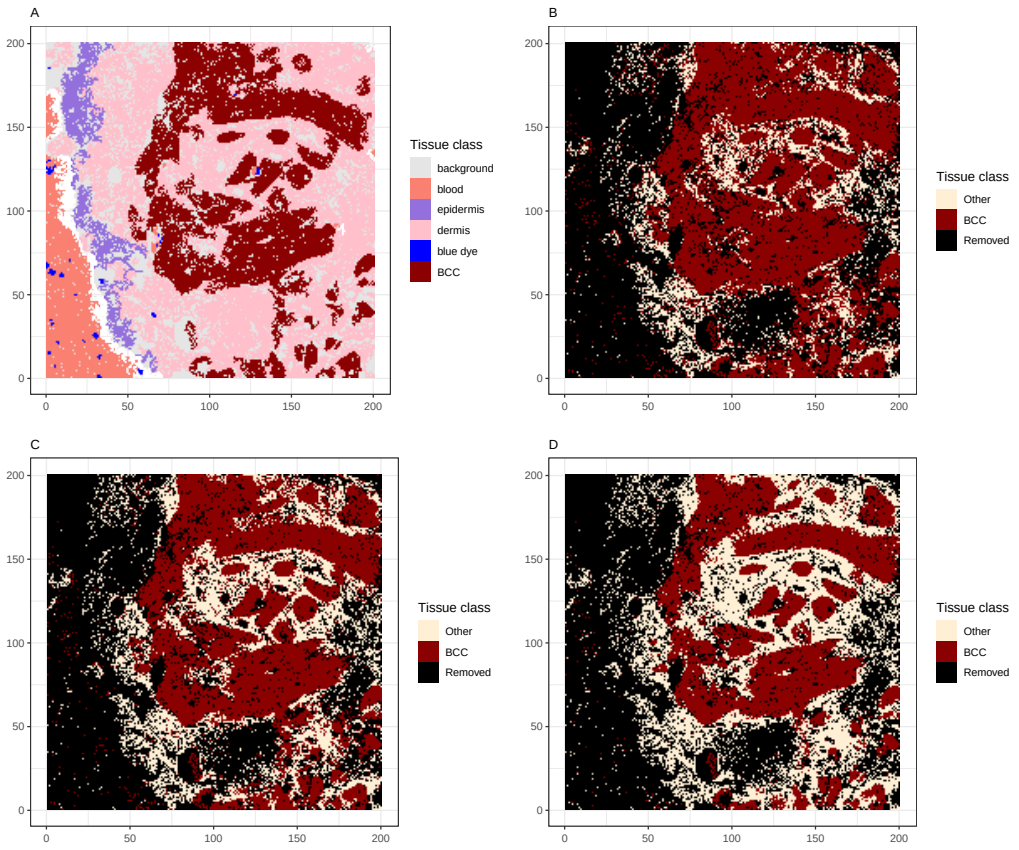


Figure 14. A) Detailed mapping of the validation sample 1 B) LDA with 30 normalized features – trained to sensitivity of 95% C) trained to sensitivity of 90% D) probability threshold of 0.5.

threshold. It can be seen that the regions grow when the probability threshold is lowered to achieve higher sensitivity. This means that the pixels near the *BCC* class pixels are more likely to share common properties with the *BCC* class. It also suggest that observing models that take neighbourhood and dependent pixel structure into account could improve the classification results. While considering that, it should also be considered in the data pre-processing set to preserve pixel neighbourhoods. Visualized classification results for the other 6 validation samples for the logistic regression model with the target sensitivity of 95% can be seen in Appendix 4. The same kind of visualization of classification results for LDA can be

seen in figure 14. While the default probability threshold and the target sensitivity of 90% generate rather similar visualizations to the logistic regression model, the observed target sensitivity of 95% classifies much more pixels as *BCC*.

Conclusions

The objective of this thesis was to study and set reference results for assessment of residual tumor margins of basal cell carcinoma. The analysis is based on Raman microspectroscopy measurements extracted during Mohs surgery. Tissue samples dataset was provided by the University of Nottingham's School of Physics and Astronomy. In the first part, the background of multimodal spectral histopathology and a theoretical overview of three classification methods were given. Logistic regression, linear discriminant analysis and quadratic discriminant analysis were used to set the reference results. In the second part, the training dataset of tissue samples was analyzed and the classification methods were compared and validated.

Extensive pre-processing of the data was required as Raman microspectroscopy measurements varied greatly in quality and intensity. Pixels of the tissue samples were assumed to be independent, which allowed us to pre-process the data pixelwise, that is without considering the neighbourhoods. This sort of data pre-processing should be reconsidered when the assumption of independent pixels is dropped, which makes the quality of the data even more important.

Classification models were estimated on the cleaned tissue samples dataset and later validated on a test set. Different classification methods and feature combinations were studied. Best classification models were selected based on the performance at the target sensitivity of 95%. The best result was achieved with the logistic regression model with all 30 original spectral features. At the target sensitivity of 95% the logistic regression model reached specificity of 73.78%. However, results on the validation samples indicated that too many pixels were classified as basal cell carcinoma as sensitivity and specificity for validation samples were 99.1% and 59.53%, respectively.

All in all, the master's thesis provides an overview of multiple issues and concerns regarding the data that should be resolved before any additional classification methods are applied. When further statistical analysis is considered in connection with

this application, the thesis could be used as an introduction to the field.

References

Barron, A. (2012). *Physical Methods in Chemistry and Nano Science*. Connexions, Rice University, pp. 297–308. URL: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Physical_Methods_in_Chemistry_and_Nano_Science_(Barron) (Retreived on Mar. 16, 2024).

Bittner, G. C., Cerci, F. B., Kubo, E. M., and Tolkachjov, S. N. (2021). Mohs micrographic surgery: a review of indications, technique, outcomes, and considerations. In: *Anais Brasileiros de Dermatologia* 96.3, pp. 263–277. DOI: https://doi.org/10.1016/j.abd.2020.10.004.

Boitor, R., Kong, K., Shipp, D., Varma, S., Koloydenko, A., Kulkarni, K., Elsheikh, S., Schut, T. B., Caspers, P., Puppels, G., Wolf, M. van der, Sokolova, E., Nijsten, T. E. C., Salence, B., Williams, H., and Notingher, I. (2017). Automated multimodal spectral histopathology for quantitative diagnosis of residual tumour during basal cell carcinoma surgery. In: *Biomed. Opt. Express* 8.12, pp. 5749–5766. DOI: https://doi.org/10.1364/BOE.8.005749.

Boitor, R., Varma, S., Sharma, A., Elsheikh, S., Kulkarni, K., Eldib, K., Jerrom, R., Odedra, S., Patel, A., Koloydenko, A., Williams, H., and Notingher, I. (2023). Ex vivo assessment of basal cell carcinoma surgical margins in Mohs surgery by autofluorescence-Raman spectroscopy: A pilot study. In: *JEADV Clinical Practice*. DOI: https://doi.org/10.1002/jvc2.336.

Chung, S. (2012). Basal cell carcinoma. In: *Archives of plastic surgery* 39.2, pp. 166–170. URL: https://doi.org/10.5999/aps.2012.39.2.166.

Hosmer, D. W. J., Lemeshow, S., and Studivant, R. X. (2013). *Applied Logistic Regression*. John Wiley and Sons.

Izenman, A. J. (2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Second. Springer.

Kong, K., Rowlands, C. J., Varma, S., Perkins, W., Leach, I. H., Koloydenko, A. A., Williams, H. C., and Notingher, I. (2013). Diagnosis of tumors during tissue-conserving surgery with integrated autofluorescence and Raman scattering microscopy. In: *Proceedings of the National Academy of Sciences* 110.38, pp. 15189–15194. DOI: https://doi.org/10.1073/pnas.1311289110.

Monici, M. (2005). Cell and tissue autofluorescence research and diagnostic applications. In: Biotechnology Annual Review 11, pp. 227–256. DOI: https://doi.org/10.1016/S1387-2656(05)11007-2.

Saletnik, A., Saletnik, B., and Puchalski, C. (2021). Overview of Popular Techniques of Raman Spectroscopy and Their Potential in the Study of Plant Tissues. In: *Molecules* 26.6. DOI: https://doi.org/10.3390/molecules26061537.

Appendix 1. Examples of R-code

```
# 1. Function for SNR calculation ----
# package for fitdistr
library(MASS)

# wn - vector of 1024 wavenumbers
# spectra - 200 x 200 x 1024 array,
# Raman measurements for each pixel

snr <- function(wn, spectra){

  # Wavenumber areas
  wn1 = 1400
  wn2 = 1512
  wn3 = 1750
  wn4 = 1800
  # Find vector indices closest to desired wavenumbers
  index_wn1 <- which.min(abs(wn - wn1))
  index_wn2 <- which.min(abs(wn - wn2))
  index_wn3 <- which.min(abs(wn - wn3))
  index_wn4 <- which.min(abs(wn - wn4))

  # Order indices
  if (index_wn1 > index_wn2){
    index_wn1_c <- index_wn1
    index_wn1 <- index_wn2
    index_wn2 <- index_wn1_c
  }
  if (index_wn3 > index_wn4){
    index_wn3_c <- index_wn3
    index_wn3 <- index_wn4
    index_wn4 <- index_wn3_c
  }

  # Create empty 200x200 matrix for SNR values
  SNR <- rep(NA, 40000)
  SNR <- matrix(SNR, nrow=200)

  for (i in 1:200){
    for (j in 1:200){
      temp_sp <- spectra[i,j,]
      signal_raw <- temp_sp[index_wn1:index_wn2]
```

```
      # Find max signal
      signal <- max(signal_raw)-min(signal_raw)
      noise <- temp_sp[index_wn3:index_wn4]
      # Fit normal distribution
      fit <- fitdistr(noise, "normal")
      noise_sd <- fit$estimate[2]
      # Calculate SNR
      noise_final <- noise_sd*2
      SNR[i,j] <- signal/noise_final
    }
  }
  # Return SNR matrix
  return(SNR)
}


#####################################################
# 2. Set-ups training ----

# df_folds - DataFrame: 2 - Binary response variable,
# 8:37 - original spectral features,
# 38:67 - normalized spectral features
# 72 - fold number

# df_folds is training dataset

# Different selection of feature variables
df_raw <- df_folds[,c(2, 8:37, 72)] # 30 original
# df_raw <- df_folds[,c(2, 38:67, 72)] # 30 normalized
# df_raw <- df_folds[,c(2, 14, 22, 30, 16, 28, 17, 32, 33,
#                       12, 34, 25, 19, 20, 23, 31, 72)]
# # 15 original expert
# df_raw <- df_folds[,c(2, 44, 52, 60, 46, 58, 47, 62, 63,
#                       42, 64, 55, 49, 50, 53, 61, 72)]
# # 15 normalized expert
# pca.results<-prcomp(df_folds[,c(8:37)]) # PCA original
# pca.results<-prcomp(df_folds[,c(38:67)]) # PCA normalized
# pca_df <- pca.results$x[,1:9]
# # Nr of components for 99% - currently 9 for original
# df_raw <- as.data.frame(cbind(df_folds[,2],
#                         pca_df, df_folds[,72]))
# names(df_raw) <- c("BCC_ann", "PC1", "PC2", "PC3", "PC4",
#         "PC5", "PC6", "PC7", "PC8", "PC9", "group_number")
```

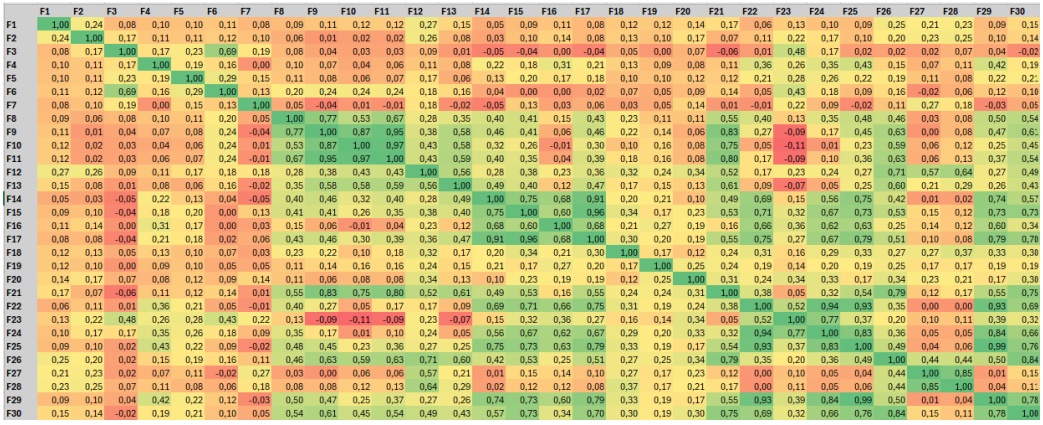
```
# All features
folds = 10
# Create empty confusion matrix
confusion_matrix <- matrix(c(0,0,0,0), nrow = 2)

for(i in 1:folds){
  ## Respective folds data, fold number is removed
  training <- df_raw[df_raw$group_number != i, -32]
  test <- df_raw[df_raw$group_number == i, -32]
  assign(paste0("test_",i), test)
  # Save for target sensitivity

  ## Train model
  model <- lda(BCC_ann~., data=training)
  # model <- qda(BCC_ann~., data=training)
  # model <- glm(BCC_ann~., data = training,
  #              family =binomial(link='logit'))
  ## Feature selection for logistic regression
  # bw_model <- step(model, direction = "backward")

  ## Predictions
  pred.prob <- predict(model, test)
  assign(paste0("pred.prob_",i), pred.prob)
  # Save for target sensitivity
  predicted <-
    as.factor(ifelse(pred.prob$posterior[,2]>=0.5,1,0))
  ## For logistic regression
  # pred.prob <- predict(bw_model,newdata=test,
  #                      type='response')
  # predicted <- ifelse(pred.prob > 0.5,1,0)

  # Add results to total confusion matrix
  temp_table <- table(test$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  # Visual indicator for completed folds
  print(paste0("Fold ", i))
}

# Accuracy
round((confusion_matrix[1,1]+
         confusion_matrix[2,2])/sum(confusion_matrix),4)
# Sensitivity
```

```
round(confusion_matrix[2,2]/(confusion_matrix[2,2]+
                               confusion_matrix[2,1]),4)
# Specificity
round(confusion_matrix[1,1]/(confusion_matrix[1,1]+
                               confusion_matrix[1,2]),4)

################################################
# 3. Function for target sensitivity ----

# Iteration until target sensitivity is found
# Logistic regression uses different format
# predicted <- ifelse(pred.prob_1 > threshold,1,0)

target_sensitivity <- function(threshold){
  confusion_matrix <- matrix(c(0,0,0,0), nrow = 2)
  #1 fold
  predicted <- as.factor(ifelse(
    pred.prob_1$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_1$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #2 fold
  predicted <- as.factor(ifelse(
    pred.prob_2$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_2$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #3 fold
  predicted <- as.factor(ifelse(
    pred.prob_3$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_3$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #4 fold
  predicted <- as.factor(ifelse(
    pred.prob_4$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_4$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #5 fold
  predicted <- as.factor(ifelse(
    pred.prob_5$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_5$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #6 fold
  predicted <- as.factor(ifelse(
    pred.prob_6$posterior[,2]>=threshold,1,0))
```

```
  temp_table <- table(test_6$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #7 fold
  predicted <- as.factor(ifelse(
    pred.prob_7$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_7$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #8 fold
  predicted <- as.factor(ifelse(
    pred.prob_8$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_8$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #9 fold
  predicted <- as.factor(ifelse(
    pred.prob_9$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_9$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  #10 fold
  predicted <- as.factor(ifelse(
    pred.prob_10$posterior[,2]>=threshold,1,0))
  temp_table <- table(test_10$BCC_ann, predicted)
  confusion_matrix = confusion_matrix + temp_table
  # Return sensitivity
  return(round(confusion_matrix[2,2]/
          (confusion_matrix[2,2]+confusion_matrix[2,1]),4))
}
```

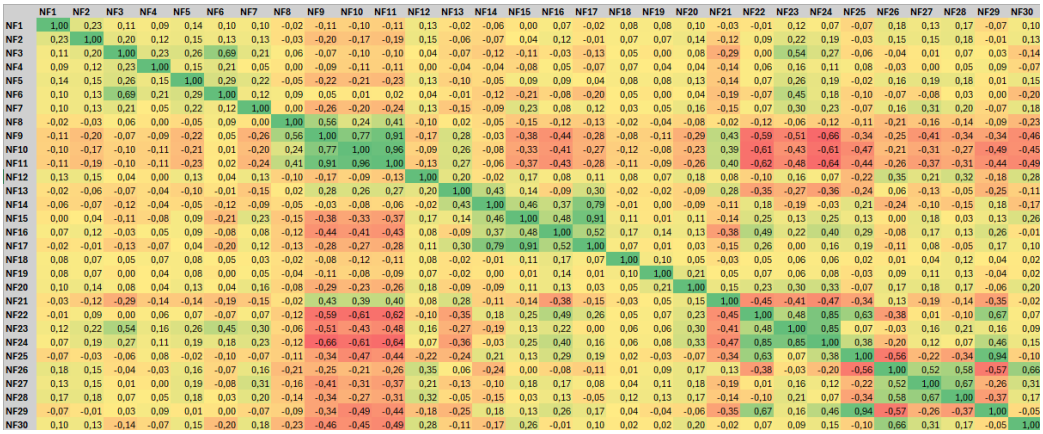
Appendix 2. Correlation matrices

Appendix 1 includes correlation matrices for original spectral features and pixel-wise L_2 normalized spectral features. Correlation matrices are conditionally formatted so red color indicates high negative correlations and green high positive correlations.



	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24	F25	F26	F27	F28	F29	F30
F1	1,00	0,24	0,08	0,10	0,10	0,11	0,08	0,09	0,11	0,12	0,12	0,27	0,15	0,05	0,09	0,11	0,08	0,12	0,12	0,14	0,17	0,06	0,13	0,10	0,09	0,25	0,21	0,23	0,09	0,15
F2	0,24	1,00	0,17	0,11	0,11	0,12	0,10	0,06	0,01	0,02	0,02	0,26	0,08	0,03	0,10	0,14	0,08	0,13	0,10	0,17	0,07	0,11	0,22	0,17	0,10	0,20	0,23	0,25	0,10	0,14
F3	0,08	0,17	1,00	0,17	0,23	0,69	0,19	0,08	0,04	0,03	0,03	0,09	0,01	-0,05	-0,04	0,00	-0,04	0,05	0,00	0,07	-0,06	0,01	0,48	0,17	0,02	0,02	0,02	0,07	0,04	-0,02
F4	0,10	0,11	0,17	1,00	0,19	0,16	0,00	0,10	0,07	0,04	0,06	0,11	0,08	0,22	0,18	0,31	0,21	0,13	0,09	0,08	0,11	0,36	0,26	0,35	0,43	0,15	0,07	0,11	0,42	0,19
F5	0,10	0,11	0,23	0,19	1,00	0,29	0,15	0,11	0,08	0,06	0,07	0,17	0,06	0,13	0,20	0,17	0,18	0,10	0,10	0,12	0,12	0,21	0,28	0,26	0,22	0,19	0,11	0,08	0,22	0,2
F6	0,11	0,12	0,69	0,16	0,29	1,00	0,13	0,20	0,24	0,24	0,24	0,18	0,16	0,04	0,00	0,00	0,02	0,07	0,05	0,09	0,14	0,05	0,43	0,18	0,09	0,16	-0,02	0,06	0,12	0,16
F7	0,08	0,10	0,19	0,00	0,15	0,13	1,00	0,05	-0,04	0,01	-0,01	0,18	-0,02	-0,05	0,13	0,03	0,06	0,03	0,05	0,14	0,01	-0,01	0,22	0,09	-0,02	0,11	0,27	0,18	-0,03	0,05
F8	0,09	0,06	0,08	0,10	0,11	0,20	0,05	1,00	0,77	0,53	0,67	0,28	0,35	0,40	0,41	0,15	0,43	0,23	0,11	0,11	0,55	0,40	0,13	0,35	0,48	0,46	0,03	0,08	0,50	0,54
F9	0,11	0,01	0,04	0,07	0,08	0,24	-0,04	0,77	1,00	0,87	0,95	0,38	0,58	0,46	0,41	0,06	0,46	0,22	0,14	0,06	0,83	0,27	-0,09	0,17	0,45	0,63	0,00	0,08	0,47	0,6
F10	0,12	0,02	0,03	0,04	0,06	0,24	0,01	0,53	0,87	1,00	0,97	0,43	0,58	0,32	0,26	-0,01	0,30	0,10	0,16	0,08	0,75	0,05	-0,11	0,01	0,23	0,59	0,06	0,12	0,25	0,45
F11	0,12	0,02	0,03	0,06	0,07	0,24	-0,01	0,67	0,95	0,97	1,00	0,43	0,59	0,40	0,35	0,04	0,39	0,18	0,16	0,08	0,80	0,17	-0,09	0,10	0,36	0,63	0,06	0,13	0,37	0,54
F12	0,27	0,26	0,09	0,11	0,17	0,18	0,18	0,28	0,38	0,43	0,43	1,00	0,56	0,28	0,38	0,23	0,36	0,32	0,24	0,34	0,52	0,17	0,23	0,24	0,27	0,71	0,57	0,64	0,27	0,49
F13	0,15	0,08	0,01	0,08	0,06	0,16	-0,02	0,35	0,58	0,58	0,59	0,56	1,00	0,49	0,40	0,12	0,47	0,17	0,15	0,13	0,61	0,09	-0,07	0,05	0,25	0,60	0,21	0,29	0,26	0,43
F14	0,05	0,03	-0,05	0,22	0,13	0,04	-0,05	0,40	0,46	0,32	0,40	0,28	0,49	1,00	0,75	0,68	0,91	0,20	0,21	0,10	0,49	0,69	0,15	0,56	0,75	0,42	0,01	0,02	0,74	0,57
F15	0,09	0,10	-0,04	0,18	0,20	0,00	0,13	0,41	0,41	0,26	0,35	0,38	0,40	0,75	1,00	0,60	0,96	0,34	0,17	0,23	0,53	0,71	0,32	0,67	0,73	0,53	0,15	0,12	0,73	0,73
F16	0,11	0,14	0,00	0,31	0,17	0,00	0,03	0,15	0,06	-0,01	0,04	0,23	0,12	0,68	0,60	1,00	0,68	0,21	0,27	0,19	0,16	0,66	0,36	0,62	0,63	0,25	0,14	0,12	0,60	0,34
F17	0,08	0,08	-0,04	0,21	0,18	0,02	0,06	0,43	0,46	0,30	0,39	0,36	0,47	0,91	0,96	0,68	1,00	0,30	0,20	0,19	0,55	0,75	0,27	0,67	0,79	0,51	0,10	0,08	0,79	0,78
F18	0,12	0,13	0,05	0,13	0,10	0,07	0,03	0,23	0,22	0,10	0,18	0,32	0,17	0,20	0,34	0,21	0,30	1,00	0,17	0,12	0,24	0,31	0,16	0,29	0,33	0,27	0,27	0,37	0,33	0,30
F19	0,12	0,10	0,00	0,09	0,10	0,05	0,05	0,11	0,14	0,16	0,16	0,24	0,15	0,21	0,17	0,27	0,20	0,17	1,00	0,25	0,24	0,19	0,14	0,20	0,19	0,25	0,17	0,17	0,19	0,19
F20	0,14	0,17	0,07	0,08	0,12	0,09	0,14	0,11	0,06	0,08	0,08	0,34	0,13	0,10	0,23	0,19	0,19	0,12	0,25	1,00	0,31	0,24	0,34	0,33	0,17	0,34	0,23	0,21	0,17	0,30
F21	0,17	0,07	-0,06	0,11	0,12	0,14	0,01	0,55	0,83	0,75	0,80	0,52	0,61	0,49	0,53	0,16	0,55	0,24	0,24	0,31	1,00	0,38	0,05	0,32	0,54	0,79	0,12	0,17	0,55	0,75
F22	0,06	0,11	0,01	0,36	0,21	0,05	-0,01	0,40	0,27	0,05	0,17	0,17	0,09	0,69	0,71	0,66	0,75	0,31	0,19	0,24	0,38	1,00	0,52	0,94	0,93	0,35	0,00	0,00	0,93	0,69
F23	0,13	0,22	0,48	0,26	0,28	0,43	0,22	0,13	-0,09	-0,11	-0,09	0,23	-0,07	0,15	0,32	0,36	0,27	0,16	0,14	0,34	0,05	0,52	1,00	0,77	0,37	0,20	0,10	0,11	0,39	0,32
F24	0,10	0,17	0,17	0,35	0,26	0,18	0,09	0,35	0,17	0,01	0,10	0,24	0,05	0,56	0,67	0,62	0,67	0,29	0,20	0,33	0,32	0,94	0,77	1,00	0,83	0,36	0,05	0,05	0,84	0,66
F25	0,09	0,10	0,02	0,43	0,22	0,09	-0,02	0,48	0,45	0,23	0,36	0,27	0,25	0,75	0,73	0,63	0,79	0,33	0,19	0,17	0,54	0,93	0,37	0,83	1,00	0,49	0,04	0,06	0,99	0,76
F26	0,25	0,20	0,02	0,15	0,19	0,16	0,11	0,46	0,63	0,59	0,63	0,71	0,60	0,42	0,53	0,25	0,51	0,27	0,25	0,34	0,79	0,35	0,20	0,36	0,49	1,00	0,44	0,44	0,50	0,84
F27	0,21	0,23	0,02	0,07	0,11	-0,02	0,27	0,03	0,00	0,06	0,06	0,57	0,21	0,01	0,15	0,14	0,10	0,27	0,17	0,23	0,12	0,00	0,10	0,05	0,04	0,44	1,00	0,85	0,01	0,15
F28	0,23	0,25	0,07	0,11	0,08	0,06	0,18	0,08	0,08	0,12	0,13	0,64	0,29	0,02	0,12	0,12	0,08	0,37	0,17	0,21	0,17	0,00	0,11	0,05	0,06	0,44	0,85	1,00	0,04	0,11
F29	0,09	0,10	0,04	0,42	0,22	0,12	-0,03	0,50	0,47	0,25	0,37	0,27	0,26	0,74	0,73	0,60	0,79	0,33	0,19	0,17	0,55	0,93	0,39	0,84	0,99	0,50	0,01	0,04	1,00	0,78
F30	0,15	0,14	-0,02	0,19	0,21	0,10	0,05	0,54	0,61	0,45	0,54	0,49	0,43	0,57	0,73	0,34	0,70	0,30	0,19	0,30	0,75	0,69	0,32	0,66	0,76	0,84	0,15	0,11	0,78	1,00

Correlation matrix of original spectral features



	NF1	NF2	NF3	NF4	NF5	NF6	NF7	NF8	NF9	NF10	NF11	NF12	NF13	NF14	NF15	NF16	NF17	NF18	NF19	NF20	NF21	NF22	NF23	NF24	NF25	NF26	NF27	NF28	NF29	NF30
NF1	1,00	0,23	0,11	0,09	0,14	0,10	0,10	-0,02	-0,11	-0,10	-0,11	0,13	-0,02	-0,06	0,00	0,07	-0,02	0,08	0,08	0,10	-0,03	-0,01	0,12	0,07	-0,07	0,18	0,13	0,17	-0,07	0,10
NF2	0,23	1,00	0,20	0,12	0,15	0,13	0,13	-0,03	-0,20	-0,17	-0,19	0,15	-0,06	-0,07	0,04	0,12	-0,01	0,07	0,07	0,14	-0,12	0,09	0,22	0,19	-0,03	0,15	0,15	0,18	-0,01	0,13
NF3	0,11	0,20	1,00	0,23	0,26	0,69	0,21	0,06	-0,07	-0,10	-0,10	0,04	-0,07	-0,12	-0,11	-0,03	-0,13	0,05	0,00	0,08	-0,29	0,00	0,54	0,27	-0,06	-0,04	0,01	0,07	0,03	-0,14
NF4	0,09	0,12	0,23	1,00	0,15	0,21	0,05	0,00	-0,09	-0,11	-0,11	0,00	-0,04	-0,04	-0,08	0,05	-0,07	0,07	0,04	0,04	-0,14	0,06	0,16	0,11	0,08	-0,03	0,00	0,05	0,09	-0,07
NF5	0,14	0,15	0,26	0,15	1,00	0,29	0,22	-0,05	-0,22	-0,21	-0,23	0,13	-0,10	-0,05	0,09	0,09	0,04	0,08	0,08	0,13	-0,14	0,07	0,26	0,19	-0,02	0,16	0,19	0,18	0,01	0,15
NF6	0,10	0,13	0,69	0,21	0,29	1,00	0,12	0,09	0,05	0,01	0,02	0,04	-0,01	-0,12	-0,21	-0,08	-0,20	0,05	0,00	0,04	-0,19	-0,07	0,45	0,18	-0,10	-0,07	-0,08	0,03	0,00	-0,20
NF7	0,10	0,13	0,21	0,05	0,22	0,12	1,00	0,00	-0,26	-0,20	-0,24	0,13	-0,15	-0,09	0,23	0,08	0,12	0,03	0,05	0,16	-0,15	0,07	0,30	0,23	-0,07	0,16	0,31	0,20	-0,07	0,18
NF8	-0,02	-0,03	0,06	0,00	-0,05	0,09	0,00	1,00	0,56	0,24	0,41	-0,10	0,02	-0,05	-0,15	-0,12	-0,13	-0,02	-0,04	-0,08	-0,02	-0,12	-0,06	-0,12	-0,11	-0,21	-0,16	-0,14	-0,09	-0,23
NF9	-0,11	-0,20	0,07	-0,09	-0,22	0,05	-0,26	0,56	1,00	0,77	0,91	-0,17	0,28	-0,03	-0,38	-0,44	-0,28	-0,08	-0,11	-0,29	0,43	-0,59	-0,51	-0,66	-0,34	-0,25	-0,41	-0,34	-0,34	-0,46
NF10	-0,10	-0,17	-0,10	-0,11	-0,21	0,01	-0,20	0,24	0,77	1,00	0,96	-0,09	0,26	-0,08	-0,33	-0,41	-0,27	-0,12	-0,08	-0,23	0,39	-0,61	-0,43	-0,61	-0,47	-0,21	-0,31	-0,27	-0,49	-0,45
NF11	-0,11	-0,19	-0,10	-0,11	-0,23	0,02	-0,24	0,41	0,91	0,96	1,00	-0,13	0,27	-0,06	-0,37	-0,43	-0,28	-0,11	-0,09	-0,26	0,40	-0,62	-0,48	-0,64	-0,44	-0,26	-0,37	-0,31	-0,44	-0,49
NF12	0,13	0,15	0,04	0,00	0,13	0,04	0,13	-0,10	-0,17	-0,09	-0,13	1,00	0,20	-0,02	0,17	0,08	0,11	0,08	0,07	0,18	0,08	-0,10	0,16	0,07	-0,22	0,35	0,21	0,32	-0,18	0,28
NF13	-0,02	-0,06	-0,07	-0,04	-0,10	-0,01	-0,15	0,02	0,28	0,26	0,27	0,20	1,00	0,43	0,14	-0,09	0,30	-0,02	-0,02	-0,09	0,28	-0,35	-0,27	-0,36	-0,24	0,06	-0,13	-0,05	-0,25	-0,11
NF14	-0,06	-0,07	-0,12	-0,04	-0,05	-0,12	-0,09	-0,05	-0,03	-0,08	-0,06	-0,02	0,43	1,00	0,46	0,37	0,79	-0,01	0,00	-0,09	-0,11	0,18	-0,19	-0,03	0,21	-0,24	-0,10	-0,15	0,18	-0,17
NF15	0,00	0,04	-0,11	-0,08	0,09	-0,21	0,23	-0,15	-0,38	-0,33	-0,37	0,17	0,14	0,46	1,00	0,48	0,91	0,11	0,01	0,11	-0,14	0,25	0,13	0,25	0,13	0,00	0,18	0,03	0,13	0,26
NF16	0,07	0,12	-0,03	0,05	0,09	-0,08	0,08	-0,12	-0,44	-0,41	-0,43	0,08	-0,09	0,37	0,48	1,00	0,52	0,17	0,14	0,13	-0,38	0,49	0,22	0,40	0,29	-0,08	0,17	0,13	0,26	-0,01
NF17	-0,02	-0,01	-0,13	-0,07	0,04	-0,20	0,12	-0,13	-0,28	-0,27	-0,28	0,11	0,30	0,79	0,91	0,52	1,00	0,07	0,01	0,03	-0,15	0,26	0,00	0,16	0,19	-0,11	0,08	-0,05	0,17	0,10
NF18	0,08	0,07	0,05	0,07	0,08	0,05	0,03	-0,02	-0,08	-0,12	-0,11	0,08	-0,02	-0,01	0,11	0,17	0,07	1,00	0,10	0,05	-0,03	0,05	0,06	0,06	0,02	0,01	0,04	0,12	0,04	0,02
NF19	0,08	0,07	0,00	0,04	0,08	0,00	0,05	-0,04	-0,11	-0,08	-0,09	0,07	-0,02	0,00	0,01	0,14	0,01	0,10	1,00	0,21	0,05	0,07	0,06	0,08	-0,03	0,09	0,11	0,13	-0,04	0,02
NF20	0,10	0,14	0,08	0,04	0,13	0,04	0,16	-0,08	-0,29	-0,23	-0,26	0,18	-0,09	-0,09	0,11	0,13	0,03	0,05	0,21	1,00	0,15	0,23	0,30	0,33	-0,07	0,17	0,18	0,17	-0,06	0,20
NF21	-0,03	-0,12	-0,29	-0,14	-0,14	-0,19	-0,15	-0,02	0,43	0,39	0,40	0,08	0,28	-0,11	-0,14	-0,38	-0,15	-0,03	0,05	0,15	1,00	-0,45	-0,41	-0,47	-0,34	0,13	-0,19	-0,14	-0,35	-0,02
NF22	-0,01	0,09	0,00	0,06	0,07	-0,07	0,07	-0,12	-0,59	-0,61	-0,62	-0,10	-0,35	0,18	0,25	0,49	0,26	0,05	0,07	0,23	-0,45	1,00	0,48	0,85	0,63	-0,38	0,01	-0,10	0,67	0,07
NF23	0,12	0,22	0,54	0,16	0,26	0,45	0,30	-0,06	-0,51	-0,43	-0,48	0,16	-0,27	-0,19	0,13	0,22	0,00	0,06	0,06	0,30	-0,41	0,48	1,00	0,85	0,07	-0,03	0,16	0,21	0,16	0,09
NF24	0,07	0,19	0,27	0,11	0,19	0,18	0,23	-0,12	-0,66	-0,61	-0,64	0,07	-0,36	-0,03	0,25	0,40	0,16	0,06	0,08	0,33	-0,47	0,85	0,85	1,00	0,38	-0,20	0,12	0,07	0,46	0,15
NF25	-0,07	-0,03	-0,06	0,08	-0,02	-0,10	-0,07	-0,11	-0,34	-0,47	-0,44	-0,22	-0,24	0,21	0,13	0,29	0,19	0,02	-0,03	-0,07	-0,34	0,63	0,07	0,38	1,00	-0,56	-0,22	-0,34	0,94	-0,10
NF26	0,18	0,15	-0,04	-0,03	0,16	-0,07	0,16	-0,21	-0,25	-0,21	-0,26	0,35	0,06	-0,24	0,00	-0,08	-0,11	0,01	0,09	0,17	0,13	-0,38	-0,03	-0,20	-0,56	1,00	0,52	0,58	-0,57	0,66
NF27	0,13	0,15	0,01	0,00	0,19	-0,08	0,31	-0,16	-0,41	-0,31	-0,37	0,21	-0,13	-0,10	0,18	0,17	0,08	0,04	0,11	0,18	-0,19	0,01	0,16	0,12	-0,22	0,52	1,00	0,67	-0,26	0,31
NF28	0,17	0,18	0,07	0,05	0,18	0,03	0,20	-0,14	-0,34	-0,27	-0,31	0,32	-0,05	-0,15	0,03	0,13	-0,05	0,12	0,13	0,17	-0,14	-0,10	0,21	0,07	-0,34	0,58	0,67	1,00	-0,37	0,17
NF29	-0,07	-0,01	0,03	0,09	0,01	0,00	-0,07	-0,09	-0,34	-0,49	-0,44	-0,18	-0,25	0,18	0,13	0,26	0,17	0,04	-0,04	-0,06	-0,35	0,67	0,16	0,46	0,94	-0,57	-0,26	-0,37	1,00	-0,05
NF30	0,10	0,13	-0,14	-0,07	0,15	-0,20	0,18	-0,23	-0,46	-0,45	-0,49	0,28	-0,11	-0,17	0,26	-0,01	0,10	0,02	0,02	0,20	-0,02	0,07	0,09	0,15	-0,10	0,66	0,31	0,17	-0,05	1,00

Correlation matrix of pixel-wise L_2-normalized spectral features

Appendix 3. Output of the logistic regression model

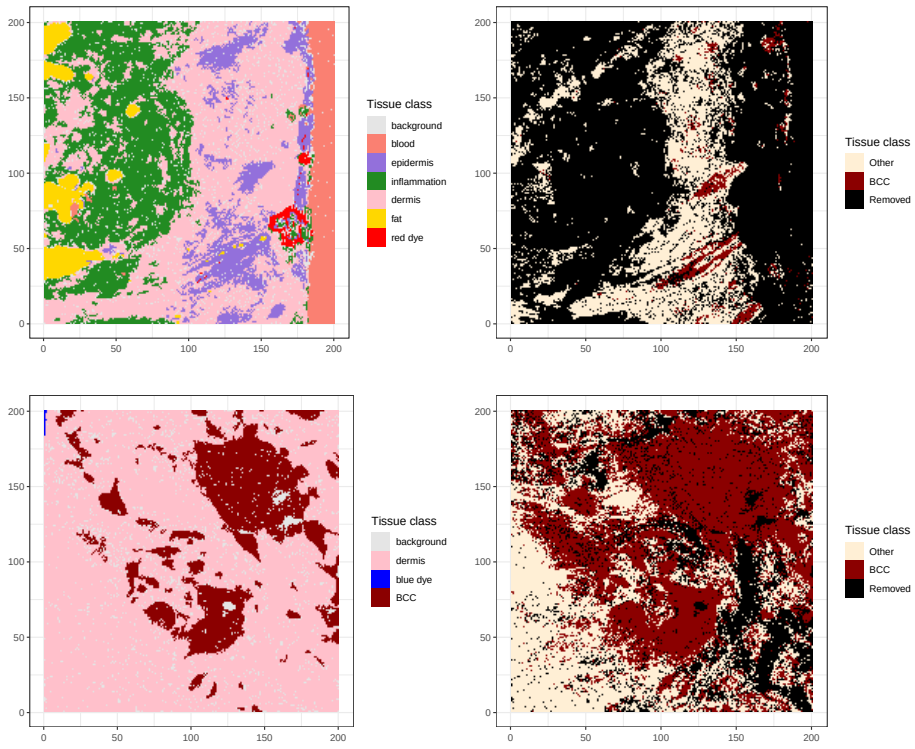
R software output of the logistic regression model with original features on cleaned training data without the *background* class.

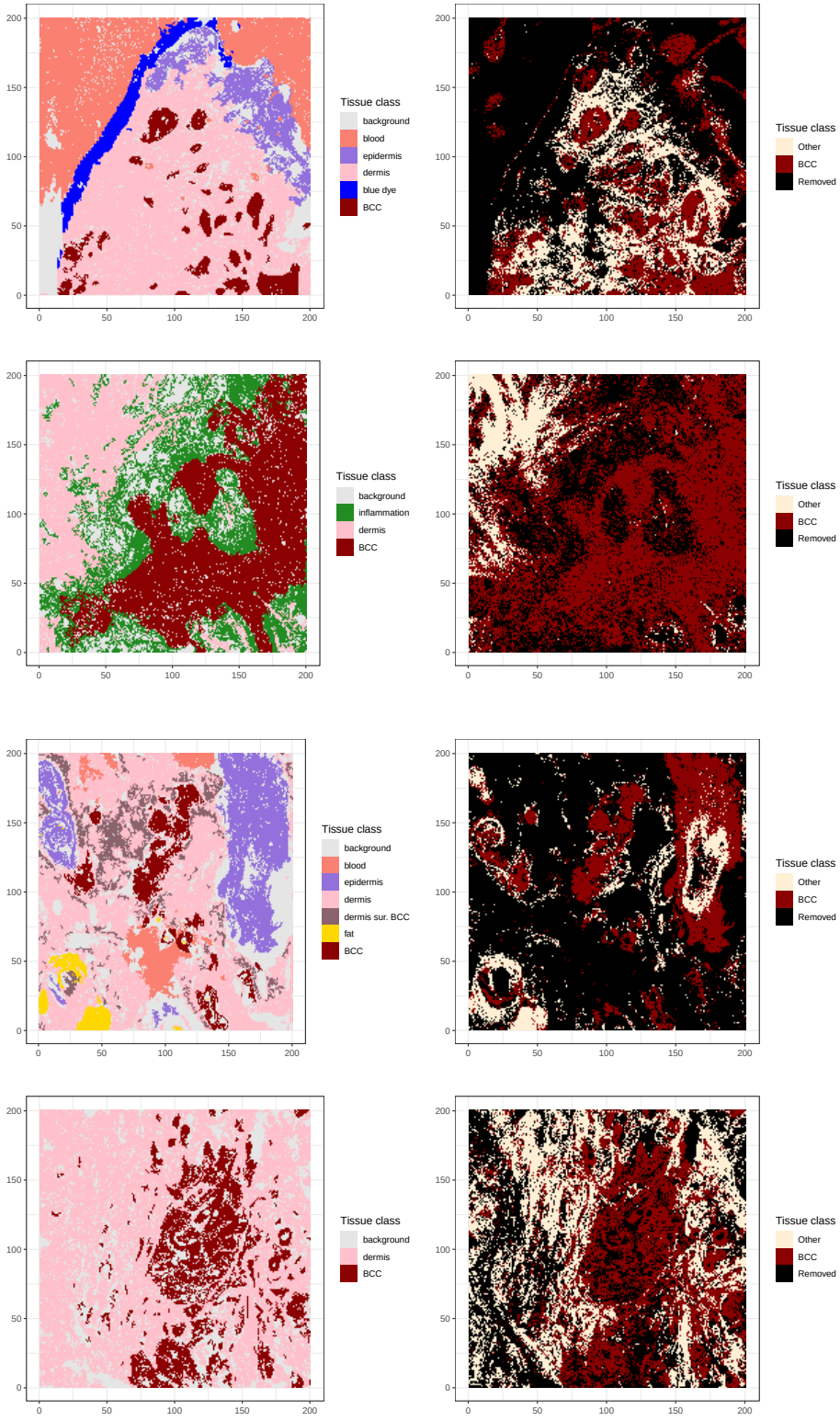
```
# Call:
#   glm(formula = BCC_ann ~ . - Feature_25, family = binomial(link = "logit"),
#       data = training_df)
#
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
# (Intercept)  9.215e-01  2.710e-02  34.009  < 2e-16 ***
#   Feature_1   -1.338e-03  1.785e-04  -7.498 6.48e-14 ***
#   Feature_2    4.956e-04  1.588e-04   3.120  0.00181 **
#   Feature_3   -2.530e-03  8.555e-05 -29.571  < 2e-16 ***
#   Feature_4   -1.337e-03  1.830e-04  -7.307 2.73e-13 ***
#   Feature_5    1.542e-03  9.876e-05  15.616  < 2e-16 ***
#   Feature_6   -3.154e-03  6.761e-05 -46.648  < 2e-16 ***
#   Feature_7    2.083e-03  7.533e-05  27.657  < 2e-16 ***
#   Feature_8    2.766e-03  8.022e-05  34.479  < 2e-16 ***
#   Feature_9   -4.830e-03  1.127e-04 -42.857  < 2e-16 ***
#   Feature_10  -2.178e-03  1.179e-04 -18.469  < 2e-16 ***
#   Feature_11   2.203e-03  9.691e-05  22.735  < 2e-16 ***
#   Feature_12   1.643e-03  8.713e-05  18.859  < 2e-16 ***
#   Feature_13  -2.711e-03  1.112e-04 -24.370  < 2e-16 ***
#   Feature_14  -7.116e-03  6.197e-04 -11.483  < 2e-16 ***
#   Feature_15   1.646e-03  6.270e-04   2.626  0.00865 **
#   Feature_16  -3.594e-03  7.830e-05 -45.902  < 2e-16 ***
```

```
#   Feature_17   3.397e-03  6.172e-04   5.504 3.70e-08 ***
#   Feature_18  -4.024e-03  1.466e-04 -27.452  < 2e-16 ***
#   Feature_19  -1.561e-03  1.215e-04 -12.849  < 2e-16 ***
#   Feature_20   4.161e-03  1.073e-04  38.781  < 2e-16 ***
#   Feature_21  -1.588e-03  2.606e-05 -60.919  < 2e-16 ***
#   Feature_22  -4.539e-03  1.181e-04 -38.419  < 2e-16 ***
#   Feature_23  -2.692e-03  1.104e-04 -24.388  < 2e-16 ***
#   Feature_24   3.278e-03  8.346e-05  39.284  < 2e-16 ***
#   Feature_26  -1.324e-03  2.345e-05 -56.483  < 2e-16 ***
#   Feature_27   1.054e-03  7.486e-05  14.086  < 2e-16 ***
#   Feature_28   1.742e-03  7.850e-05  22.186  < 2e-16 ***
#   Feature_29  -2.068e-03  3.863e-05 -53.545  < 2e-16 ***
#   Feature_30   5.634e-03  9.050e-05  62.253  < 2e-16 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
# Null deviance: 450549  on 477937  degrees of freedom
# Residual deviance: 233590  on 477908  degrees of freedom
# AIC: 233650
#
# Number of Fisher Scoring iterations: 8
```

Appendix 4. Figures of the validation samples and visualized classification results

Appendix 3 includes the figures of the validation samples 2-7. Detailed mapping and their respective classification results for the logistic regression model with the 30 original spectral features at the target sensitivity of 95% can be seen.




Tissue class
background
blood
epidermis
dermis
blue dye
BCC
Tissue class
Other
BCC
Removed
Tissue class
background
inflammation
dermis
BCC
Tissue class
Other
BCC
Removed
Tissue class
background
blood
epidermis
dermis
dermis sur. BCC
fat
BCC
Tissue class
Other
BCC
Removed
Tissue class
background
dermis
BCC
Tissue class
Other
BCC
Removed

Non-exclusive licence to reproduce thesis and make thesis public

I, Siim Viigand,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, Assessment of surgical margins of basal cell carcinoma with Raman microspectroscopy measurements, supervised by Kristi Kuljus and Alexey Koloydenko.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Siim Viigand
22.05.2024