UNIVERSITY OF VAASA

FACULTY OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

Mika Ruohonen

ON THE DETECTION OF CARIES LESIONS IN HUMAN TEETH USING VIS/NIR-SPECTROSCOPY

Master's thesis in Technology for the degree of Master of Science in Technology submitted for inspection, Vaasa, December 14, 2011.

Supervisor	Jouni Lampinen
Instructors	Jarmo Alander Petri Välisuo

FOREWORD

"Most of science is about understanding what is important."
–Professor Asoke K. Nandi, August 17th, 2011,
at the 21st Jyväskylä Summer School.

First I want to express my gratitude towards D.Sc. Petri Välisuo for his invaluable assistance and support during this project. During the numerous times that I felt completely clueless about how to advance on this project he provided me with advice and ideas, and encouraged me to continue the struggle. I want to thank Dr. Vladimir Bochko for helping me with the dimensionality reduction methods and machine learning algorithms. I would like to thank Chief Dental Officer of the City of Vaasa, Ph.D. Jukka Kentala and Acting Chief Dental Officer of City of Vaasa, Dr. Katri Palo for providing me with the samples and background material on dentistry. Without them this project could not have been implemented. I want to also thank B.Sc. Annika Svanh and M.Sc. (Chem.) Katriina Sirviö for helping me with the chemical aspects of my project, especially on my plans to induce caries lesions by acid cycling. My thanks also go to Professor Erkki Hiltunen for helping me with some details of the physics related of my project. Last but not least, I would like to thank my advisor, Professor Jarmo Alander, for giving me an opportunity to work on this project. Being able to work on the project full-time made it possible for me to write this thesis on a topic that was so foreign to me, and to learn about spectroscopy and pattern recognition in the process. While this did not produce the most beautiful of theses, it certainly facilitated learning the craft of scientific research. I want also to thank the organizers of the Field-NIRCE project, especially Professor Paul Geladi. Without this project I would not have had the opportunity to write my thesis about this subject.

TABLE OF CONTENTS **page**

SYMBOLS AND ABBREVIATIONS

Latin symbols

c	The speed of light
g	Anisotropy factor
h	The Planck's constant
n	Refractive index
v_i	A vibrational quantum number

Greek symbols

λ	The wavelength of an electromagnetic wave
μ_a	Absorbance coefficient
μ_s	Scattering coefficient
μ, μ_t	Total attenuation coefficient
ν	The frequency of an electromagnetic wave

Abbreviations

AUC	Area under curve
CV	Cross-validation
FIR	Far infrared
MIR	Mid infrared
NIR	Near infrared
NPV	Negative predictive value
PCA	Principal component analysis
PPV	Positive predictive value
ROC	Receiver operating characteristics
RBF	Radial basis function
SVM	Support vector machine

VAASAN YLIOPISTO
Teknillinen tiedekunta

Tekijä:	Mika Ruohonen
Tutkielman nimi:	Kariesleesion tunnistamisesta ihmishampaassa VIS/NIR-spektrografian avulla.
Valvojan nimi:	Jouni Lampinen
Ohjaajien nimet:	Jarmo Alander, Petri Välisuo
Tutkinto:	Diplomi-insinööri
Koulutusohjelma:	Tietotekniikan koulutusohjelma
Suunta:	Ohjelmistotekniikka
Opintojen aloitusvuosi:	2005
Tutkielman valmistumisvuosi:	2011 **Sivumäärä:** 130

TIIVISTELMÄ:

Lähes 100% useimpien maiden aikuisväestöstä kärsii hammaskarieksesta. Nykyiset menetelmät karieksen tunnistamiseksi kykenevät tunnistamaan karieksen vasta verrattain myöhäisessä kehitysvaiheessa. Minimaalisen invasiivinen hammaslääketiede edellyttää, että karies voidaan tunnistaa jo varhaisessa kehitysvaiheessa ja että sen kehitystä voidaan seurata tiheästi.

Tämän tutkimuksen tavoitteena oli selvittää voidaanko diffuusiin heijastumaan perustuvaa lähi-infrapunaspektroskopiaa käyttää sellaisten kariesleesioiden tunnistamiseen, jotka voidaan tunnistaa manuaalisella tarkastelulla valokuituvalon avulla. Positiiviset tulokset tukisivat mahdollisuutta käyttää heijastuneeseen valoon perustuvaa lähi-infrapunaspektroskopiaa kariesleesioiden tunnistamiseen aikaisessa kehitysvaiheessa.

Yhteensä 24 hammasnäytettä mitattiin kahdella spektrometrillä, jotka yhdessä kattoivat aallonpituudet 200–1706 nm. Vain aallonpituudet 420–1000 nm huomioitiin yksityiskohtaisessa analyysissä. Kukin näyte luokiteltiin joko näytteeksi terveeltä alueelta tai näytteeksi kariesleesiosta viidellä erilaisella binäärisellä luokittelumenetelmällä. Kunkin luokittelijan tarkkuutta arvioitiin ristiinvalidoinnilla. Eräs käytetyistä luokittelumenetelmistä oli binääriluokittelijana käytetty tukivektorikone.

Tämän tutkimuksen tulokset viittaavat siihen, että lähi-infrapunaspektroskopia kykenee parantamaan manuaalisella tarkastelulla tapahtuvan kariesleesioiden tunnistamisen tarkkuutta, ainakin kun tarkastelua suorittava henkilö on aloittelija. Tämä väite perustuu oletukseen, jonka mukaan kaikkien terveen kiilteen alueiden spektrit muistuttavat toisiaan jossain määrin, sekä osittain oletukseen, jonka mukaan kaikki kariesleesiot heijastavat tervettä kiillettä enemmän valoa lähi-infrapuna-alueella. Tekijän kyky diagnosoida kariesleesioita edellä mainitulla manuaalisella menetelmällä, sekä näytteiden kyky esittää spektrin varianssi terveen kiilteen alueilla sekä kariesleesioissa, rajoittavat kuitenkin näiden tulosten luotettavuutta.

AVAINSANAT: Spektroskopia, hahmontunnistus, binäärinen luokittelu, hammaslääketiede, karies

UNIVERSITY OF VAASA
Faculty of Technology

Author:	Mika Ruohonen	
Topic of the Thesis:	On the Detection of Caries Lesions in Human Teeth Using VIS/NIR-Spectroscopy.	
Supervisor:	Jouni Lampinen	
Instructors:	Jarmo Alander, Petri Välisuo	
Degree:	Master of Science in Technology	
Degree Programme:	Degree Programme in Information Technology	
Major of Subject:	Software Engineering	
Year of Entering the University:	2005	
Year of Completing the Thesis:	2011	**Pages:** 130

ABSTRACT:

Dental caries affects nearly 100% of the adult population in most countries. The current methods for diagnosing dental caries are able to detect caries only at a relatively advanced stage. Minimally invasive dentistry requires that caries is detected at an early stage of development, and that its status can be monitored frequently.

The objective of this study was to investigate whether diffuse reflectance near-infrared spectroscopy can be used to detect dental caries lesions that are advanced enough to be detected with manual inspection with fiber-optic illumination. Positive results would support the possibility of using reflectance near-infrared spectroscopy for detecting caries lesions at an early stage.

A total of 24 tooth samples were measured with two spectroscopes that together covered the wavelength range 200–1706 nm, using a general purpose transmission dip probe. Only the wavelength range 420–1000 nm was included in detailed analysis. Five different binary classification methods were used to classify the samples as either healthy or as carious. The performance of each classifier was evaluated with 4-fold cross-validation. One of the classification methods was a binary-classification support vector machine.

The results of this study suggest that diffuse reflectance near-infrared spectroscopy is able to improve the diagnostic accuracy of manual inspection with fiber-optic illumination, at least when the inspection is done by a novice. This claim is contingent on an assumption that all healthy sites of enamel have spectra that somewhat resemble each other, and partly on an assumption that all carious lesions on enamel show increased scattering in the near-infrared range. The reliability of these results is limited by the author's ability to diagnose caries lesions with the said manual method, and by the samples' ability to represent the variance among sites of healthy enamel and among caries lesions, though.

KEYWORDS: Spectroscopy, pattern recognition, binary classification, dentistry, caries

1. INTRODUCTION

Dental caries affects nearly 100% of the adult population in most countries (Karlsson 2010). Diagnosis of dental caries is currently based on visual examination of a dried tooth surface (possibly with help of binocular loupe optics) and on tactile sensation over the surface using a (preferably blunt) dental probe or a dental explorer. Appearance of white spots or discoloration of the surface, or a sticky tooth surface, indicate caries. Diagnosis can be aided by radiographs or by transillumination. (Beighton & Bartlett 2006: 86–87; Baysan 2007; Karlsson 2010: 1–2.)

Minimally invasive dentistry is an approach that seeks to maintain the patient's oral health with preventive measures, and to treat possible disturbances of health as early as possible and with as little intervention (force) as possible (Wilson & Plasschaert 2007). Rather than drilling and filling, minimally invasive dentistry seeks to stop the progression of caries and to reverse the damage that it has already done. This is achieved by using antibacterial rinses, fluoride treatments, and changes in the patient's diet. (Jones, Huynh, Jones & Fried 2003: 2260.) G.V. Black predicted more than a century ago that dentistry would eventually develop towards a preventive approach, which has been advancing for the past twenty years (Wilson & Plasschaert 2007: 1; Karlsson 2010: 1). Minimally invasive dentistry is in the process of becoming the mainstream of dentistry (Wilson & Plasschaert 2007; Jones et al 2003: 2260). Another foreseeable approach is evidence-based dentistry. It emphasizes the use of evidence and case-by-case judgement on clinical decision making rather than opinion and tradition. (Wilson & Plasschaert 2007.)

The current methods for diagnosing caries are able to detect caries only at a relatively advanced stage (Karlsson 2010: 2). Minimally invasive dentistry requires that caries is detected at an early stage of development, and that its status can be monitored frequently (Jones et al 2003: 2260). Accordingly, methods for early detection of caries have been researched for the past twenty years. Many of these methods still require extensive research before they can be used in clinical practice. A set of diagnosis methods known as the optical caries diagnosis methods or dental tissue optics are based on the fact that caries causes changes in the tooth's optical properties at an early stage of development.

(Karlsson 2010: 2) Other novel methods of caries diagnosis include imaging the temperature drop on the tooth surface when air-drying it, and photothermal radiometry, in which the propagation of thermal waves in the tooth caused by pulsed heating of a single point on the surface is imaged (Zakian, Taylor, Ellwood & Pretty 2010; Hellen 2010).

The objective of this study was to measure diffuse reflectance from human teeth using VIS/NIR-spectroscopy, i.e. spectroscopy using visible and near-infrared light, and to investigate whether such measurements can be used to detect dental caries lesions that are advanced enough to be detected with manual inspection with fiber-optic illumination. The measurements were made *in vitro*, in a laboratory. The research hypothesis of this study was that increased scattering in the near-infrared range is the best indication of a dental caries lesion, and that this difference is large enough to enable detection of caries lesions with NIR-spectroscopy (see chapter 4). The research hypothesis was inspired by the results that were obtained in earlier studies of detecting caries lesions with near-infrared light (see, for example, Wu & Fried 2009 and Jones et al 2003).

Beside being limited to *in vitro* measurements, this study is further limited to natural caries lesions on smooth surfaces of extracted tooth. Caries lesions on the biting surface are not studied. This limitation is made because the smooth surfaces are easier to measure spectroscopically than the irregular and grooved biting surfaces. Furthermore, once caries can be diagnosed spectroscopically on the smooth surfaces, it is easier to attribute spectroscopic observations made on the biting surfaces either to caries or to surface irregularities.

1.1. Affiliations

This thesis was made as a part of the FIELD-NIRce project. The project has participants from Finland and Sweden. The Finnish participants are the Novia University of Applied Sciences (in Vaasa), Ketek Oy (in Kokkola) and the Unit of Automation in the University of Vaasa. The Swedish participants are the Department of Chemistry and the Centre for Environmental Research in Umeå University, the Unit of Biomass Technology and Chemistry in the Swedish University of Agricultural Sciences and Umbio AB (in Umeå).

The project is funded by Bothnia-Atlantica, the European Union, Regional Council of Ostrobothnia, Region Västerbotten, and Provincial Government of Västerbotten.

The FIELD-NIRce project aims to construct spectroscopy equipment that is suitable for making measurements outside laboratories or *in the field* and to research the use of such equipment. The wavelength region used in the equipment may be in the ultraviolet, visible, or near-infrared region. The research is divided into three stages. At the first stage the intended measurements are done in a laboratory to assess whether they are feasible under those conditions. At the second stage selected samples are measured both inside and outside a laboratory in order to evaluate the quality of the measurement results outside the laboratory. At the third and final stage the developed measurement device and method are taken into use in the field. This thesis focuses on the first stage, making NIR-spectroscopy measurements in a laboratory.

1.2. Related work

Professor Daniel Fried from University of California, San Francisco, has researched optical diagnosis methods in dentistry with his students. In 2005 he published an article that "discusses the NIR optical properties of sound and demineralized dental enamel and the potential use of polarization sensitive optical coherence tomography and NIR transillumination for the imaging of dental caries" (Fried, Featherstone, Darling, Jones, Ngaotheppitak & Bühler 2005). His student, Robert S. Jones, first researched the use of "near-infrared transillumination at 1310-nm for the imaging of early dental decay" (Jones et al 2003) and later the use of polarization sensitive optical coherence tomography (PS-OCT) with simulated caries lesions (Jones 2006). In the summary of his doctoral dissertation, Jones states that the work he published in three articles in 2002, 2003, and 2004 "established for the first time that interproximal caries can be detected at an earlier stage using NIR transillumination than visible light and x-rays." At the time the dissertation was published it was unknown if NIR transillumination could be used to evaluate how far a caries lesion had progressed. In contrast, PS-OCT could be used to quantify the severity of the lesion. Jones concludes that NIR transillumination could be used for screening for early stage caries lesions, after which PS-OCT could be used for assessing the severity of the lesion.

(Jones 2006: 190–200.). Also Pena (2009) wrote a doctoral dissertation about detecting caries lesions with NIR-imaging under Professor Fried's direction. Tao & Fried (2009) used NIR-imaging to guide the removal of caries lesions by means of a CO_2 laser.

Wu & Fried (2009) used NIR transillumination, fluorescence loss measurements, reflected visible light, and reflected NIR light for imaging artificial caries lesions. They used crossed polarizers after the light source and before the detector. The NIR reflectance imaging produced better results than the other methods for detecting superficial lesions. They hypothesized that the use of NIR transillumination, together with NIR reflectance imaging, could help to evaluate the severity of the lesion, since NIR transillumination can detect only more advanced lesions. Lee, Lee, Darling & Fried (2010) used NIR-imaging to assess the severity of occlusal caries lesions. Zakian, Pretty & Ellwood (2009) used hyperspectral imaging for detecting caries lesions, with good results. Maia, Fonseca, Kyotoku & Gomes (2009) studied NIR-transillumination with sections of teeth. Wist, Moon, Herr & Fatouros (2009) used a technique that was based on "raster scans of the teeth with narrow collimated light beams" to detect carious lesions.

Staninec, Lee, Darling & Fried (2010) present a clinical study for detecting approximal caries lesions located at the contact surfaces between teeth, in vivo, using NIR transillumination. They state that their study is the first of its kind. They used one or two 1310-nm superluminescent diodes (SLD), with a 35-nm bandwidth, and Teflon optical diffusers to provide uniform illumination to the inspected area. The images were captured with a high sensitivity InGaAs camera with a 25-mm objective lens. They imaged only approximal lesions which were visible in bitewing radiographs but not visible in direct visual inspection. A total of 33 lesions were imaged, and all but one were visible in the NIR images. They also noted that "there were many areas on the teeth that appeared to be demineralized in the NIR images that did not show up on the bitewing radiographs."

Ko, Hewko, Sowa, Dong & Cleghorn (2008) present a proof-of-concept study, where they used polarized Raman spectroscopy for detecting early caries lesions in extracted human tooth samples. They focused especially on lesions on the approximal surfaces, i.e. surfaces that face the adjacent teeth. The excitation laser source and the spectrometer were coupled to the measurement set-up via fibre-optics. The excitation laser beam was polarized with a polarizing beam splitter, and the beam of Raman scattered photons from the sample was split into two by another polarizing beam splitter, such that the two resulting beams were orthogonally polarized. The two beams were transmitted to the same spectrometer via a custom-made bifurcated fibre bundle, such that the two beams impinged on the spectrometer's detector one millimeter apart from each other. Spectra for one of the beams was obtained by binning specific rows of the spectrometer's detector, and spectra for the other beam was obtained by binning another set of rows. They were able to detect carious lesions with high accuracy by using the depolarization ratio and the polarization anisotropy at wavelength 959 cm^{-1} with Bayesian analysis. With 47 measurements from healthy sites and 27 measurements from carious sites they had only one false-positive, and otherwise perfect classification. Earlier Ko presented a paper on using Raman spectroscopy to detect caries lesions (Ko, Choo-Smith, Zhu, Hewko, Dong, Cleghorn & Sowa 2006). Hill & Petrou (1997) also studied caries lesions with Raman spectroscopy.

Chung, Fried, Staninec & Darling (2011) used transmission and reflectance imaging of artificial caries lesions with NIR light. Bürmen, Usenik, Fidler, Pernuš & Likar (2011) started the construction of a database of hyperspectral images of teeth by imaging 12 extracted human teeth. The gold standard for the database is an assessment of the images by an expert.

Quantitative light-induced fluorescence (QLF) uses changes in the tooth's autofluorescence to detect changes in the tooth's mineral content. In this method the tooth is illuminated with wavelengths 290–450 nm and imaged with a camera with a 520 nm high pass filter. "A high positive correlation is reported between QLF and absolute mineral loss". (Karlsson 2010: 3–4.) In laser-induced fluorescence (LF) the tooth is illuminated with red laser light at wavelength 655 nm and the resulting NIR fluorescence is measured. More intense fluorescence indicates more extensive pathology. The origin of the fluorescence is unclear; however, it is "believed to originate from bacteria or their metabolites." Two commercial devices that are based on laser-induced fluorescence are available from KaVo Dental Corporation (Charlotte, NC, USA) under the product name DIAGNOdent: DIAGNOdent 2095 (or Classic) and DIAGNOdent pen 2190. The devices have shown good performance in *in vitro* studies, but not in *in vivo* studies. "In general, in vivo studies of LF for occlusal caries detection indicate moderate to high sensitivity and lower specificity. Lack of specificity, the increased likelihood of false-positive readings due to stain and plaque, and the absence of a single threshold are factors underlying the reluctance among authors to recommend the LF method unequivocally for caries detection." (Karlsson 2010: 4–5.) Karlsson (2009) wrote a doctoral dissertation on optical methods for detecting dental caries.

Some related work has also been done at the University of Vaasa in the Department of Electrical Engineering and Energy Technology. B.Sc. Christian Söderbacka has been programming an industrial robot from Fanuc Inc. (Oshino-mura, Yamanashi Prefecture, Japan) for automating the purely executive aspects of spectroscopic measurements. The goal is to be able to lay out a set of samples and have the robot perform the measurements on each of them, and then replace the samples. This would reduce the amount of manual labour required for measuring large batches of samples. M.Sc. Vladimir Chernov has been developing a setup for creating three-dimensional images of teeth *in vitro* by using near-infrared light or multispectral imaging. D.Sc. Petri Välisuo has been working on developing a setup for creating three-dimensional images of dental casts by using visible light. B.Sc. Severi Sutinen and B.Sc. Suvi Karhu have developed a setup and software for taking photographs of human teeth *in vivo* and evaluating the shade of the teeth programmatically.

1.3. Outline of the thesis

Chapter 2 presents background information about teeth, dental caries, and spectroscopy. Chapter 3 introduces the reader to the methods that were used to analyse the measurement results. Chapter 4 describes the samples and measurement setup used in this study, as well as the various methods used to classify the samples as either healthy or as carious. The results are summarized in chapter 5. The results are discussed about in chapter 6 and conclusions are drawn from them in chapter 7. The appendix describes one classification method that was used in this study, namely support vector machine, in more detail than was done in the main text of this thesis.

2. BACKGROUND

This chapter introduces the reader to the anatomy and histology of human teeth, as well as to the process by which dental caries forms. The theoretical basis of spectroscopy is also presented, although in research with spectroscopy a research hypothesis is rarely derived from the theory because such derivation would be exceedingly complicated. The theory is presented to give the reader a qualitative understanding of spectroscopy as a method.

2.1. Human teeth

2.1.1. Anatomy

Human teeth consist of three layers of different types of tissues (Fig. 1). The outermost layer, the crown of the tooth, is composed of enamel. Enamel is hard, mineralized tissue which protects the tooth. It is up to 2–2.5 mm thick on the cusps of the molars (the teeth furthest back in the mouth). The main body of the tooth, dentin, begins underneath the enamel and continues throughout the rest of the tooth. Dentin is a bone-like tissue. It is harder than bone but less hard than enamel. Dentin has a hollow center which contains the third tissue type of the tooth: the pulp. Pulp is a soft connective tissue, which contains blood vessels and nerves. Pulp provides nutrients for the formation of the dentin (by odontoblasts) and acts as a sensory organ for the tooth. (Hellen 2010: 1; Phillips 2006: 9, 11-12; Yaeger 1976.) The interface between enamel and dentin is called the amelo-dentinal junction (AEJ) or dental-enamel junction (DEJ).

Humans have two sets of teeth over their lifetime. The first set consists of 20 deciduous teeth and the second contains 28–32 permanent teeth (Fig. 2). Teeth are divided into four groups according to their anatomical features and location in the mouth (Fig. 3). The first group, incisors, consists of the four anterior teeth (at front of the mouth). Incisors have a single root and and flat, chisel-like incisal (biting) surface. The second group, canines, contains a single tooth at both sides of the incisors at both the maxilla (upper jaw) and the mandible (lower jaw). Canines have a single root and their biting surface forms a relatively sharp point. The third group, premolars, are located distal to (behind) the canines. Premolars have one or two roots and two cusps. Normal anatomy contains two

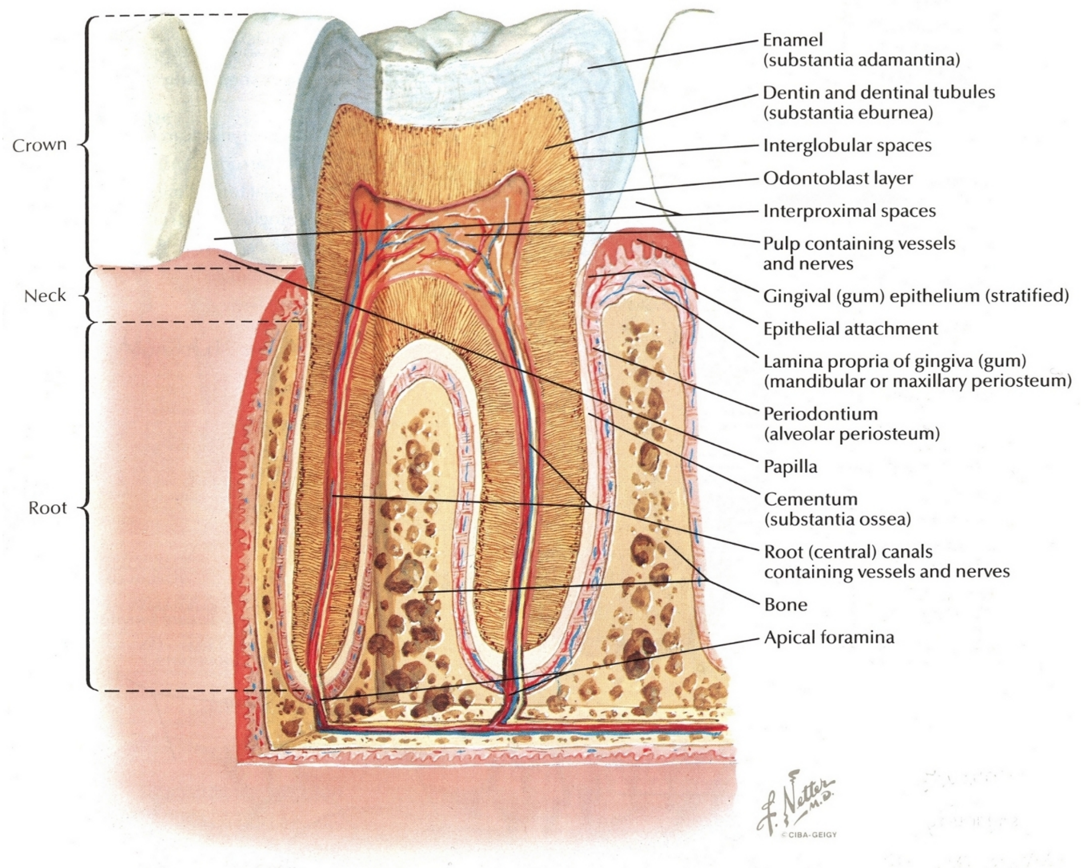

Crown
Neck
Root
Enamel (substantia adamantina)
Dentin and dentinal tubules (substantia eburnea)
Interglobular spaces
Odontoblast layer
Interproximal spaces
Pulp containing vessels and nerves
Gingival (gum) epithelium (stratified)
Epithelial attachment
Lamina propria of gingiva (gum) (mandibular or maxillary periosteum)
Periodontium (alveolar periosteum)
Papilla
Cementum (substantia ossea)
Root (central) canals containing vessels and nerves
Bone
Apical foramina
F. Netter M.D.
© CIBA-GEIGY

Figure 1. The anatomy of a human tooth (Netter 1989: 51).

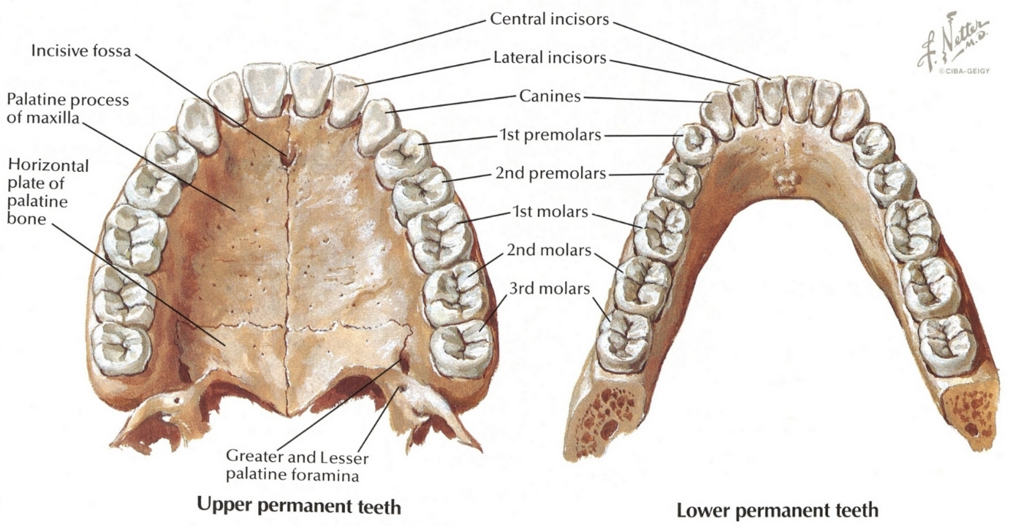


Figure 2. The permanent teeth (Netter 1989: 50).

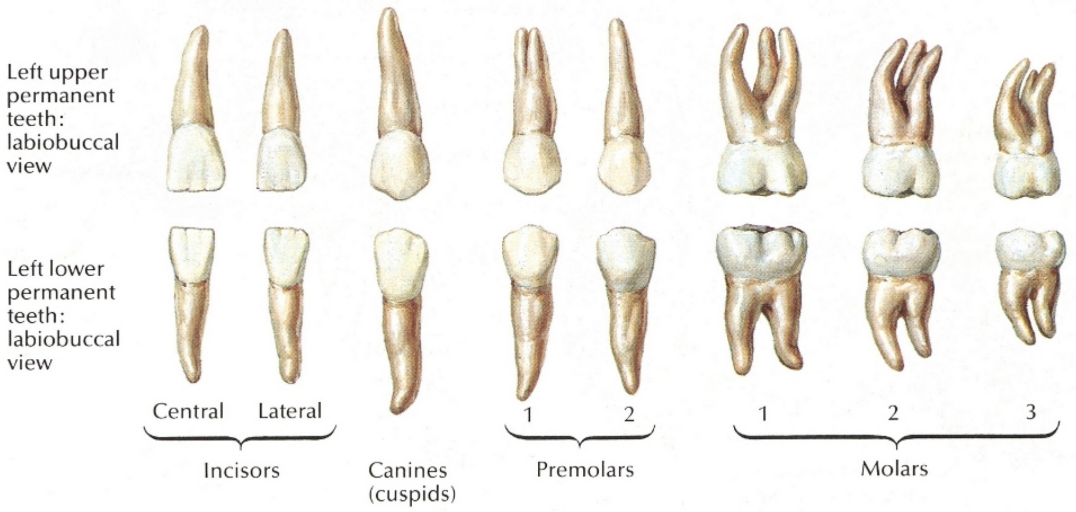

Left upper permanent teeth: labiobuccal view
Left lower permanent teeth: labiobuccal view
Central
Lateral
Incisors
Canines (cuspids)
1
2
Premolars
1
2
3
Molars

Figure 3. The groups of teeth (Netter 1989: 51).

premolars at both sides and at both jaws of the mouth. The fourth group, molars, consists of the three most distal teeth at both sides and both jaws of the mouth, i.e. the teeth furthest back towards the neck. The molars on the maxilla (upper jaw) have three roots and the molars on the mandible (lower jaw) have two roots. The anatomy of the deciduous teeth is similar to the permanent teeth, except that deciduous teeth do not contain premolars, and they contain only two molars at both sides and both jaws of the mouth. (Autti, Bell, Meurman & Murtomaa 2004: 46–47; Alaluusua, Aine, Asikainen, Eriksson, Hurmerinta, Hölttä, Karjalainen, Lukinmaa & Pirinen 2004: 536.)

The teeth are numbered with two-digit figures. The first digit indicates whether the tooth is located at maxilla or mandible, whether it is located on the right half or the left half of the mouth, and whether it is a deciduous tooth or permanent tooth. The significance of the various values of the first digit is explained in Table 1. In short, the four quarters of the mouth are numbered in a counterclockwise manner, starting from the upper right quadrant, with one as the first value for permanent teeth and five as the first value for deciduous teeth. This numbering scheme is seen as a mirror image (advancing in clockwise manner) when looking at the mouth of a patient. The same mirror image scheme is seen in panoramic radiographs of the mouth. The second digit indicates the tooth's distance from the medial line, i.e. from the central line which divides the mouth into right and left

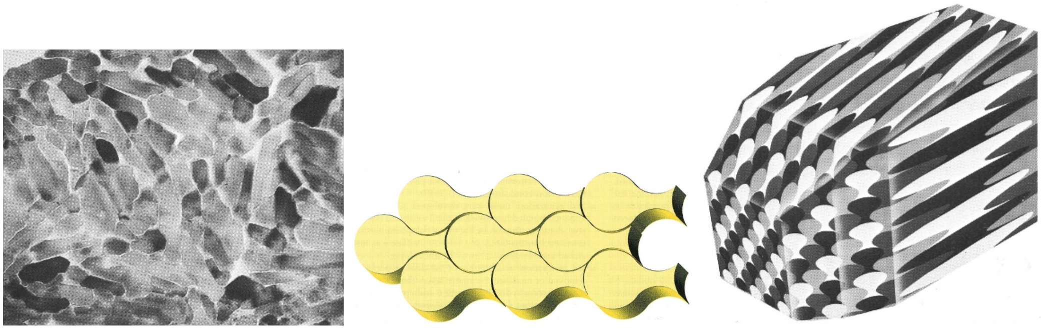
Table 1. The significance of the first digit in the two-digit numbering scheme of the teeth (Alaluusua et al 2004: 536).

First digit	Jaw	Side	Teeth set
1	Maxilla (upper jaw)	Right half	Permanent
2	Maxilla (upper jaw)	Left half	Permanent
3	Mandible (lower jaw)	Left half	Permanent
4	Mandible (lower jaw)	Right half	Permanent
5	Maxilla (upper jaw)	Right half	Deciduous
6	Maxilla (upper jaw)	Left half	Deciduous
7	Mandible (lower jaw)	Left half	Deciduous
8	Mandible (lower jaw)	Right half	Deciduous

halves. Thus incisors have second digits of one and two, canines have three as the second digit, and so on. (Alaluusua et al 2004: 536.)

2.1.2. Histology

Enamel is the hardest tissue in the human body. It consists of 95–97 wt% (≈85 vol%) inorganic material, 1 wt% organic material, and 2–3 wt% water. The organic content is primarily protein. The main inorganic component of enamel is hydroxyapatite (OHAp) in the form of hydroxyapatite crystals. The crystals have a diameter of approximately 30–40 nm and their length may be up to 10 μm (Fig. 4a). The crystals combine to form rods, or more precisely prisms, of enamel. Enamel rods extend from the amelo-dentinal junction to the surface of the tooth with a wavy path. At the horizontal central plane of the crown the rods are approximately horizontal (i.e., at right angle to the surface of the tooth). Above that plane the rods tend to bend upward and below that plane the rods tend to bend downward. The diameter of the rods increases from the amelo-dentinal junction towards the surface, with an average diameter of 4 μm. The shape of the rods resembles a keyhole of a warded lock, with the round part pointing towards the biting surface and the key tooth part pointing towards the root (Fig. 4b). The pattern that the rods form on a section of the enamel depends on the orientation of the section relative to the orientation of



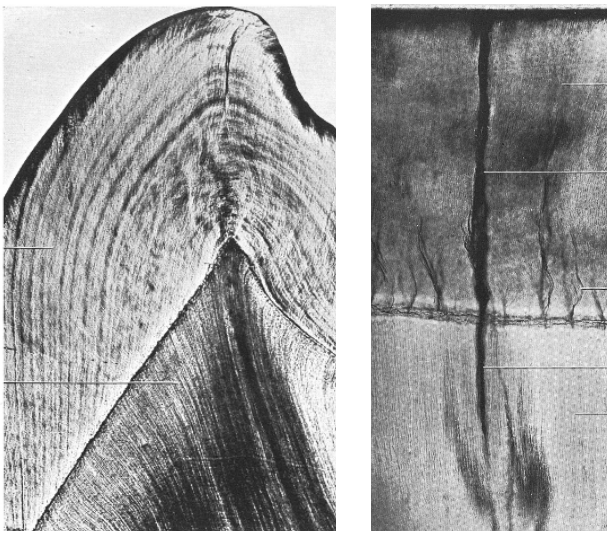
(a) Cross section of hydroxyapatite crystals in enamel (Yaeger 1976: 52).

(b) The shape of enamel rods (Franklin & Brunton 2006: 220).

(c) Packing of rods in enamel (Yaeger 1976: 52).

Figure 4. Hydroxyapatite crystals and enamel rods.

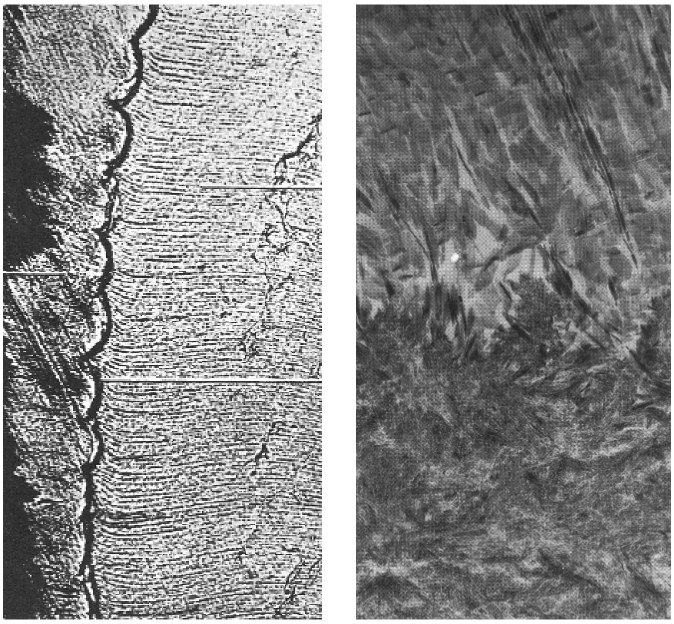
the rods (Fig. 4c). Brownish bands, known as the Retzius bands, can be seen on sections of enamel (Fig. 5a). They reflect the way the enamel was formed layer by layer, and can thus be compared to the growth rings of trees. The outermost layer of enamel lacks the prismatic structure of the deeper layers, and is accordingly called the aprismatic layer. Its thickness varies from few micrometers to about 60 μm. The formation of this layer is attributed to reduced activity of the ameloblasts at the end of the matrix formation process. (Phillips 2006: 10; Yaeger 1976; Hellen 2010: 1–2; Dorozhkin 2009: 411.)



(a) The Retzius bands (Yaeger 1976: 58).

(b) Ground section of dental lamella extending to the dentin (Yaeger 1976: 68).

Figure 5. Histological pictures of human teeth.



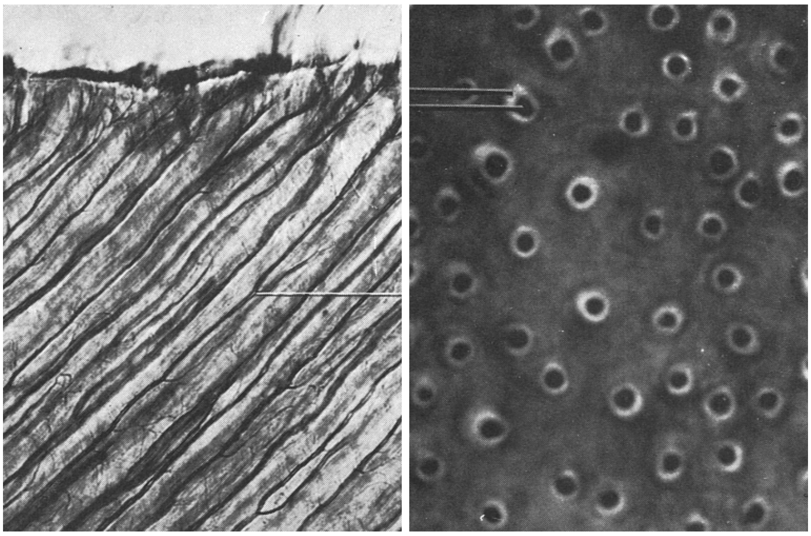
(a) Ground section of amelo-dentinal junction (Yaeger 1976: 69). The junction contains small pits that strengthen the junction.

(b) Ground section of amelo-dentinal junction (Avery 1976: 117).

Figure 6. Ground sections of amelo-dentinal junction.

The enamel may contain thin structures called enamel lamellae, which may be mistaken for cracks (Fig. 5b). Usually they contain mostly organic material with very little mineral material, but they may contain cementum, or even be filled with it. Enamel lamellae extend inwards from the tooth surface, possibly reaching some distance into the dentin. The amelo-dentinal junction is not smooth, but contains small pits on the dentin side, strengthening the junction (Fig. 6). (Yaeger 1976.)

Dentin consists of 70 wt% inorganic material, 20 wt% organic material, and 10 wt% water. As in enamel, the main inorganic component is hydroxyapatite. The hydroxyapatite crystals of dentin are smaller than the crystals in enamel, but otherwise similar. Dentinal tubules, which are found throughout the dentin, are conduits with highly mineralized walls (Fig. 7). The walls of the tubules are called peritubular dentin while the rest of the dentin is intertubular dentin. Each dental tubule contains a single odontoblast cell, usually located on the pulpal surface. Dentin contains 30,000–75,000 dentinal tubules per 1 mm^2 on the pulpal (inner) surface. The tubules traverse the entire thickness of dentin from the pulp to the amelo-dentinal junction with a slightly curved path. They are more densely packed and their diameter is greater at the pulpal surface than at the amelo-dentinal junction. The main organic component of dentin is collagen fibres. The



(a) Dentinal tubules (Avery 1976: 111).

(b) Ground section of dentin (Avery 1976: 111).

Figure 7. Dentinal tubules.

fibers crisscross in the intertubular dentin in a random fashion. During mineralization of the dentin, hydroxyapatite crystals are formed within and around the fibers, such that their long axes are oriented parallel to the fibers. Dentin may contain pockets of poorly mineralized (hypomineralized) tissue, called interglobular areas. They are formed during the mineralization of dentin, when the mineralized sites fail to fuse together at those areas. The dentinal tubules passing through interglobular areas are mineralized normally. Interglobular areas are usually located near the amelo-dentinal junction. A layer that seems to contain small interglobular areas is very often seen in dentin just below cementum on the roots of teeth. It is called the Tomes' granular layer. (Avery 1976; Phillips 2006: 10.)

Hydroxyapatite, $Ca_5(PO_4)_3(OH)$, is one of the calcium orthophosphates. Its chemical formula is often written in the form $Ca_{10}(PO_4)_6(OH)_2$ since each unit cell of hydroxyapatite crystals contains two molecules. All hard tissues of the human body, except a part of the inner ear, consist of calcium orthophosphates with hydroxyapatite as their main component. The ion components of apatites, including hydroxyapatite, can be replaced by other isomorphic ions without disturbing the structure of the crystal. Ion substitutions in the molecules may also lead to a crystal structure where some of the ions are missing,

Table 2. Approximate composition (in wt%) of enamel and dentin (Dorozhkin 2009: 403).

Composition	Enamel	Dentin	Pure hydroxyapatite
Calcium	36.5	35.1	39.6
Phosphorus (as P)	17.7	16.9	18.5
Sodium	0.5	0.6	–
Magnesium	0.44	1.23	–
Potassium	0.08	0.05	–
Carbonate (as CO_3^{2-})	3.5	5.6	–
Fluoride	0.01	0.06	–
Chloride	0.30	0.01	–
Pyrophosphate (as $P_2O_7^{4-}$)	0.022	0.10	–
Total inorganic	97	70	100
Total organic	1.5	20	–
Water	1.5	10	–

creating a non-stoichiometric compound. Chemically pure calcium orthophosphates are white crystals. Natural calcium orthophosphates, such as those that are found in biological systems, always contain impurities, i.e. isomorphic ion components, which cause the crystals to be colored. Such impure hydroxyapatite is sometimes called biological apatite or dahllite. Enamel is mineralized in media which contains significant concentrations of ions that are suitable to be incorporated into hydroxyapatite as impurities, e.g. Na^+, K^+, Mg^{2+}, Na^+, Cl^-, HCO_3^- and F^-. Accordingly, these ions are present in the enamel. The approximate composition of enamel is presented in Table 2. (Dorozhkin 2009: 399–402, 412; Aoba 2004.)

When the OH^- ion of hydroxyapatite is replaced with an F^- ion, the compound becomes fluorapatite (FA), $Ca_5(PO_4)_3F$. Fluorapatite is the least soluble type of calcium orthophosphate, while hydroxyapatite is the second least soluble type. The volume of unit cells of fluorapatite is smaller than that of hydroxyapatite, increasing the electrostatic bond between the fluoride ion and the adjacent ions, thus increasing the chemical stability of the crystal structure. If saliva is supersatured with calcium, phosphate ions, and fluoride, fluorapatite can be precipitated on the surfaces of the (erupted) teeth. There it will protect the tooth's surface from dissolution, i.e. caries. Normally saliva is supersatured with calcium and phosphate ions. This is how fluoridated toothpaste and fluoride in drinking water can help to reduce the progression, and perhaps even the prevalence, of caries. Compounds which contain both hydroxyapatite and fluorapatite are called fluorhydroxyapatites (FHA) or hydroxyfluorapatites (HFA). Their chemical formula is written as $Ca_{10}(PO_4)_6(OH)_{2-x}F_x$, where $0 < x < 2$, or as $Ca_{10}(PO_4)_6(F,OH)_2$. (Dorozhkin 2009: 411–414; Aoba 2004; Tenovuo 2004: 241.)

The composition of the dental tissues varies from tooth to tooth and between different sites of a given tooth. In recently erupted teeth the fluoride concentration of enamel is relatively high at the surface layers, and decreases quickly towards the interior layers. Also, the fluoride concentration is higher at the coronal (biting) surface than at the cervical surface (near the gumline). In older, worn teeth the enamel surface layer contains less fluoride, possibly even less than the interior layers, and the fluoride concentration increases from the coronal surface towards the cervical surface. (Weatherell, Robinson & Hallsworth 1974.)

2.1.3. Dental caries

Dental caries is the demineralization of dental tissue. Its formation depends on three factors: the type of bacteria present in the mouth, the chemical composition of the teeth surfaces, and the types of food consumed. Caries is caused by organic acids (Tab. 3) which are produced when bacteria present in the mouth ferment carbohydrates, e.g. sucrose or table sugar. These acids upset the chemical equilibrium between saliva and the mineral content of teeth, causing minerals in the teeth to dissolve in the saliva. Only bacteria that are aciduric, i.e. able to survive in an acidic environment, and acidogenic, i.e. produce acids, are able to induce caries. (Beighton & Bartlett 2006: 75–78, 82–83.) The acidogenic theory outlined above is not the only theory that has been presented about the causes of caries. However, the other theories have been discarded for lack of evidence. (Soames & Southam 1993: 19.) Caries is most prevalent in the molar teeth. It is most

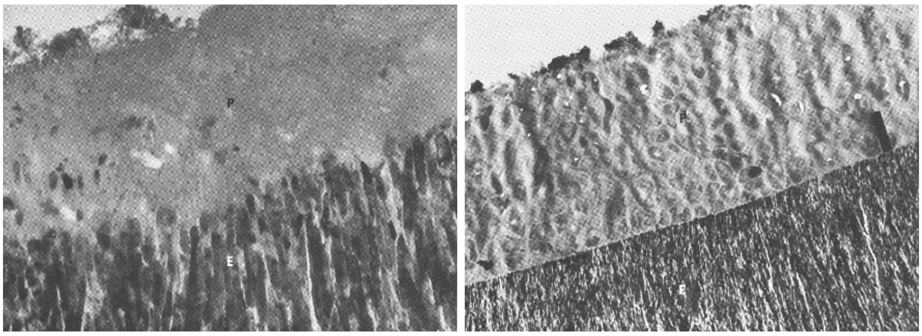
Table 3. Types of organic acid produced by dental plaque (Stösser, Dell, Borutta & Heinrich-Weltzien 2007: 16).

Type	%
Lactic acid	55–80
Acetic acid	20–25
Propionic acid	5–15
Formic acid	1–10
Butyric acid	0–6
Succinic acid	0–4

commonly seen in the occlusal (biting) surface, with the two proximal surfaces as the second- and thirdmost commonly carious surface. (Shafer, Hine & Levy 1974: 395–396.)

Streptococcus mutans is a bacteria that seems to be particularly cariogenic, or able to produce caries, when only a single strain of bacteria is present in the mouth (at least in rats in a laboratory). However, the dental plaque in the human mouth contains over 700 different species of bacteria. The different species of bacteria interact with each other, e.g. by helping other bacteria to bind to the tooth surface. *Fusobacterium nucleatum* is particularly capable of binding with many species of bacteria. The presence or absence of *S. mutans* in the dental plaque has little effect on the formation of caries. (Beighton & Bartlett 2006: 75–78.)

Within seconds of cleaning a tooth surface, glycoproteins from the saliva adhere to the surface, forming a layer called the pellice (Fig. 8a) (Soames & Southam 1993: 20). The pellicle absorbs bacteria, and within an hour the first bacteria (including *S. oralis*, *S. sanguinius*, *Actinomyces naeslundii* as well as *Neisseria* and *Haemophilus* species) are bound to the pellicle (Beighton & Bartlett 2006: 75–78). In one or two days bacteria colonize the pellicle and form a biofilm called bacterial plaque (Fig. 8b) (Yaeger 1976: 63, 67). The composition of the plaque varies according to the age of the plaque and the site where it is located. The patient's diet effects the plaque composition as well. (Beighton & Bartlett 2006: 75–78.)



(a) Pellicle (Yaeger 1976: 64). **(b)** Dental plaque (Yaeger 1976: 65).

Figure 8. Histological images of pellicle and dental plaque.

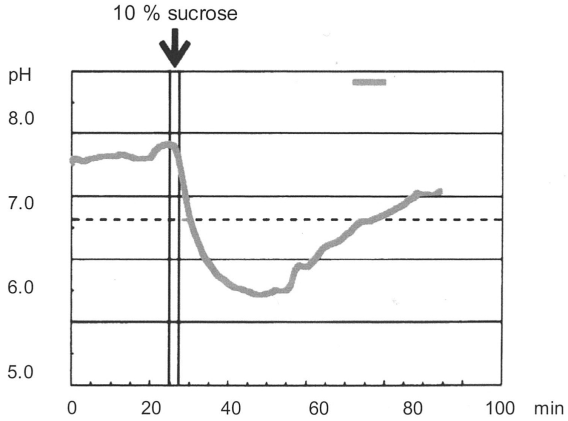


Figure 9. Stephan's curve, a depiction of how the pH-value of plaque changes after food intake (Stösser et al 2007: 16).

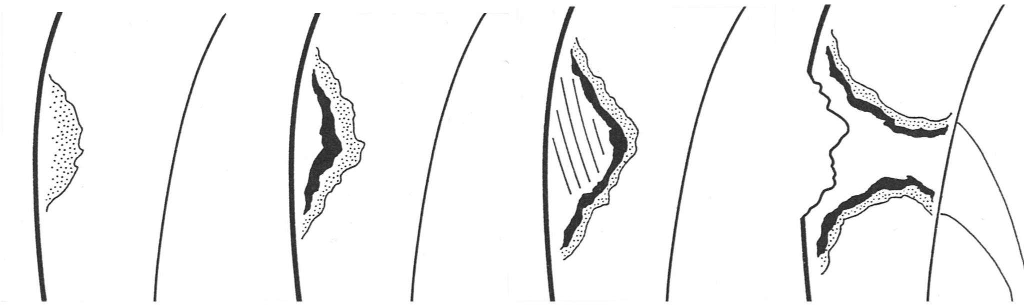
The pH of the plaque drops up to two units in 10 minutes after eating carbohydrates. After 30–60 minutes the pH returns to its normal value. (Soames & Southam 1993: 20–21.) The change of pH as a response of food intake is depicted in a Stephan's curve (Fig. 9), named after R.M. Stephan, who measured the pH in a dental plaque with a microelectrode in 1940 (Shafer et al 1974: 372). Caries formation begins in enamel when the pH drops below the critical pH, i.e. below the lowest pH at which the saliva is satured with the tooth minerals (Dawes 2003). Although the shape of Stephan's curve is always similar, the normal pH value varies between individuals. Therefore the period for which the pH is below the critical value varies from person to person. (Soames & Southam 1993: 21.)

The critical pH depends on the chemical composition of the tooth and the saliva (Dawes

2003). It can be explained by the stoichiometric model of a chemical equilibrium between a mineral and a solution. In this model the amount of mineral dissolved in the solution is depicted by the ion product (IP), whose formula is derived from the ionic formula of the mineral by replacing each ion with its concentration – or more precisely, its activity – in the solution, raised to the power equal to the multiplier of the ion in the ionic formula. For example, the ionic formula of $Ca_5(PO_4)_3(OH)$ is $5Ca^{2+} + 3PO_4^{3-} + OH^-$, and its ion product is $[Ca]^5[PO_4]^3[OH]$, where [Ca] depicts the molar concentration (or activity) of Ca^{2+} ions. When the mineral is placed in the solution, it begins to dissolve until its ion product in the solution reaches a value known as the solubility product constant (K_{sp}). If the ion product is below this value, the solution is unsaturated with the mineral, and if the ion product is above this value, the solution is supersatured with the mineral. Notice that the ion product is zero if any of the components is missing from the solution. The solubility product constant depends on temperature, pH, and the gas environment of the solution (and mineral). In a supersatured solution the ions combine to precipitate solid mineral until the ion product drops to the value of the solubility product constant. In fact, the mineral dissolves in the solution and ions in the solution precipitate (back) into mineral all the time. The rates at which these reactions take place are different and create a net effect of dissolution or precipitation. (Hein, Best, Pattison & Arena 1997: 398–418; Aoba 2004.)

Normally, saliva (and plaque) is supersatured with the tooth mineral, hydroxyapatite. When the pH of saliva decreases, hydroxyl ions in the saliva combine with the hydrogen ions of the acid ($H^+ + OH^- \rightleftarrows H_2O$) and the phosphate is transformed from the form PO_4^{3-} to forms HPO_4^{2-}, $H_2PO_4^-$ and H_3PO_4, which decreases the tooth mineral's ion product in the saliva. (Dawes 2003.) The critical pH is usually around 5.2–5.5 for enamel and 6.0 for dentin (Beighton & Bartlett 2006: 83). The Ca/P molar ratio of the tooth mineral correlates with its solubility. The Ca/P ratio is 1.63 for enamel, 1.61 for dentin, and 1.67 for hydroxyapatite. (Dorozhkin 2009: 402.)

On the buccal and lingual surfaces caries tends to begin close to the gingival margin (Beighton & Bartlett 2006: 84). The development of a caries lesion starts by the formation of a translucent zone beneath the tooth surface (Fig. 10a). Healthy enamel contains 0.1



(a) Translucent zone.

(b) Dark zone.

(c) White spot lesion.

(d) Cavity through the enamel.

Figure 10. Histopathogenesis of enamel caries (Soames & Southam 1993: 26).

vol% pores, whereas the translucent zone of a caries lesion contains 1 vol% pores. In healthy enamel the size of the pores is approximately the size of a water molecule, but the pores grow in size due to caries. As the lesion continues to develop, the translucent zone grows, and a dark zone is formed in the center of it (Fig. 10b). The dark zone contains 2–4 vol% pores. However, some of the pores are smaller than pores in the translucent zone, probably resulting from minerals precipitating (back) to the tooth from saliva, i.e. remineralization. When the lesion continues to grow, the center of the dark zone becomes the body of the lesion (Fig. 10c). The body of the lesion contains 5–25 vol% pores, and the apatite crystals in the body are larger than crystals in healthy enamel. The lesion body is more translucent than normal enamel, and the Retzius bands and the transverse striations of the enamel rods are more visible in the lesion body than in healthy enamel. The lesion body can be visually detected as a white spot (Fig. 11). (Soames & Southam 1993: 25–27.) The increased porosity of the enamel tissue due to demineralization increases the tissue's scattering coefficient, causing the area to appear whiter in reflected light (Beighton & Bartlett 2006: 83–84; Karlsson 2010). If the white spot lesion becomes stained by bacteria, food, or tobacco, the lesion becomes a brown spot. When the lesion is close to becoming a brown spot, it can be detected on a bitewing radiograph. (Soames & Southam 1993: 27.)

The early caries lesion has an 20–50 μm thick layer of apparently healthy tissue covering

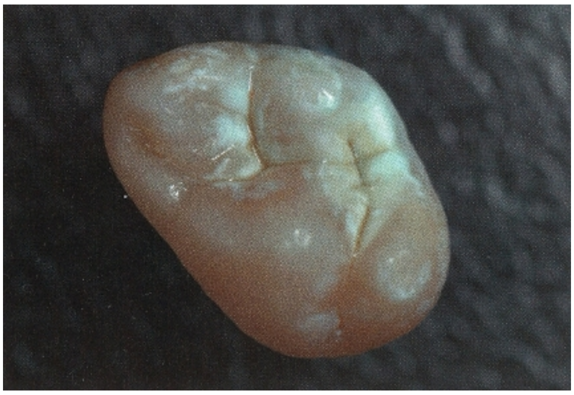
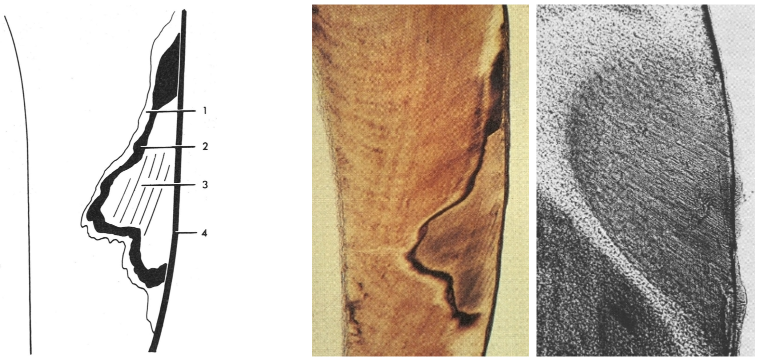


Figure 11. White spot lesions on the occlusal surface of a molar tooth (Beighton & Bartlett 2006: 83).

it (Fig. 12) (Beighton & Bartlett 2006: 83–84). The layer of enamel covering a caries lesion is most likely produced by the reprecipitation of dissolved minerals as they are diffusing out of the tooth. A somewhat similar covering layer is observed on caries lesions on the interior layers of enamel (in the absence of the enamel surface layer), and the composition and structure of the layer covering a caries lesion seems to be different than that of sound enamel. (Weatherell et al 1974.) The covering layer is probably composed of DCPD, or dicalcium phosphate dihydrate, $CaHPO_4{\cdot}2H_2O$ (Aoba 2004). The composition of enamel effects the reprecipitation of the minerals. For example, fluoride (F^-) increases the reprecipitation. Organic debris on the lesion surface might have a similar effect. (Weatherell et al 1974.)

If demineralization continues, the surface layer covering the lesion is lost, and the lesion continues to grow laterally (Beighton & Bartlett 2006: 84). A cavity which forms on a smooth surface is usually roughly cone-shaped, with the apex (top) towards the dentin (Fig. 10d). When a cavity is formed in a fissure (pit) of an occlusal (biting) surface, the cavity usually has a cone-shape, with the base (bottom) pointing towards the dentin. (Shafer et al 1974: 397–399.) Fissure caries begins at the walls of a fissure, forming a ring around it (Fig. 13). Fissure caries lesions are similar to the smooth surface lesions described above, but the ring shape of the lesion results in a cone-shape, with the base pointing upwards. (Soames & Southam 1993: 27.)



(a) Diagram of layers of a caries lesion: 1, translucent zone; 2, dark zone; 3, body of the lesion; 4, surface zone (Soames & Southam 1993: 25).

(b) Ground section of an early caries lesion (Soames & Southam 1993: 25).

(c) Chalky (white spot) lesion of enamel (Shafer et al 1974: 397).

Figure 12. Layers of a caries lesion.

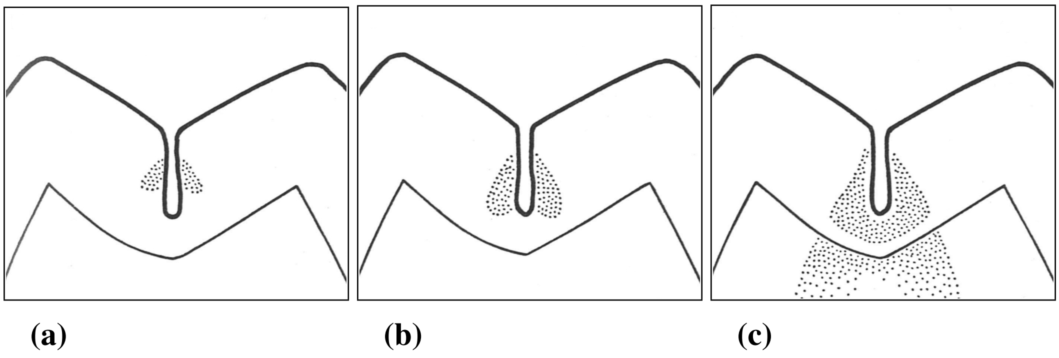


Figure 13. Development of a fissure caries lesion (Soames & Southam 1993: 27).

Although dentine is a softer tissue than enamel, caries progresses in dentine at the same speed as it does in enamel. A carious lesion in dentin is detected by the texture and softness of the lesion surface. In advanced caries the dentin surface is almost wet and can be peeled away. The dentine surface softens earlier in the caries process than the enamel surface. Beside demineralization of the hydroxyapatite crystals, caries also breaks down collagen in the dentine. A residue of the crystals and collagen may be found from the plaque on the lesion. The remineralization of dentine hardens the lesion surface. The recovering lesions are first described as "leathery", and later as "hard". (Beighton & Bartlett 2006: 84–85.)

When the pH returns to normal the caries formation process or *demineralization* process starts to reverse by a *remineralization* process, where the minerals lost from the dental tissue are replaced by minerals in the saliva. The balance between these two processes determines whether the demineralization develops into a clinical dental caries. (Beighton & Bartlett 2006: 83.) The remineralization of enamel causes discoloration of the tissue as brown or yellow. Thus, the discoloration of a tooth is not necessarily associated with an active caries. (Beighton & Bartlett 2006.) Demineralization can be reversed if the lesion has not yet reached the dentin.

2.2. Spectroscopy

Spectroscopy is the art and science of identifying various properties of matter by measuring the light's intensity at various wavelengths before and after it has interacted with matter. Electromagnetic radiation in the approximate wavelength range 390–780 nanometers constitutes the visible light (Fig. 14) (Saleh & Teich 2007: 39). Light at the wavelength region 780–2500 nanometers is classified as near-infrared (NIR) light, although the upper limit is sometimes set at 2000 nm (Bokobza 2002: 11; Saleh & Teich 2007: 39). Occasionally, the lower limit is set at 700 nm (Osborne, Fearn & Hindle 1993: 21). Light is scattered and absorbed in matter in a way that depends on the light's wavelength and the chemical composition and structure of the matter. In an object that comprises several parts of different material, each part has a different effect on the light's path in the object.

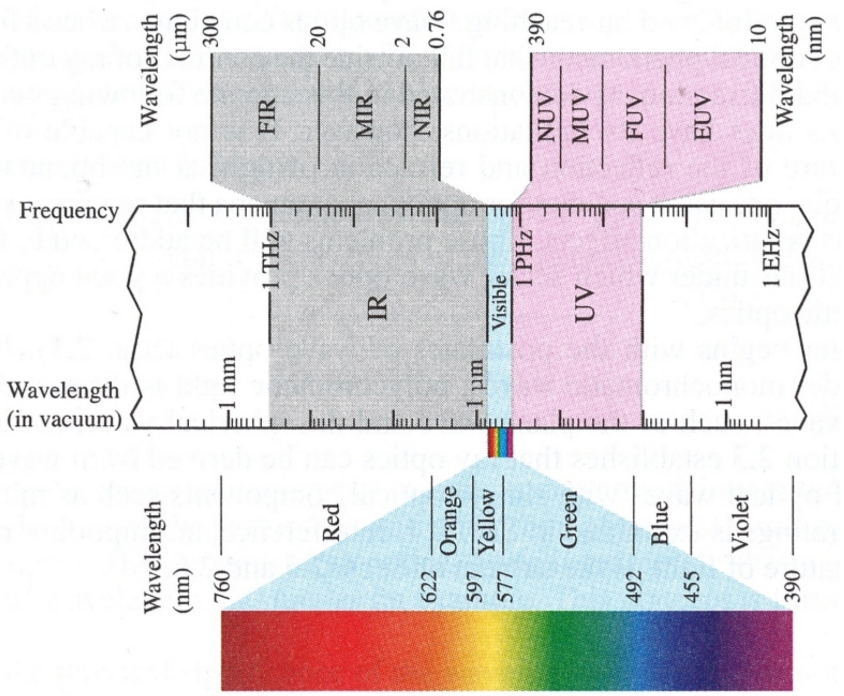


Figure 14. The spectrum of optical wavelengths of electromagnetic radiation (Saleh & Teich 2007: 39).

Light can be modelled with a number of different models at varying levels of detail. The detailed models can explain phenomena which the simpler models can not, at the cost of added complexity. The simplest model is ray optics, where light is described as geometric lines in space. Wave optics complicates the model by describing light as waves, and electromagnetic optics goes a step further by explaining the wave nature of light by the interaction between electric and magnetic fields which constitutes those waves. Some optical phenomena, such as the absorption of light by atoms and molecules, can only be explained by quantum optics, also known as quantum electrodynamics (QED). (Saleh & Teich 2007: 2, 445.)

According to quantum optics, light is composed of packets of energy called photons. Each photon carries an amount of energy E, which determines its wavelength λ according to

$$E = h\nu = \hbar\omega, \qquad \nu = \frac{c}{\lambda}, \qquad \hbar = \frac{h}{2\pi}, \tag{1}$$

where ν is the corresponding frequency, $\omega = 2\pi\nu$ is the corresponding angular frequency, c is the speed of light, and h is the Planck's constant. The photon's frequency ν is also called its mode. When the photon's wavelength is given in micrometers, its energy can be easily calculated in electronvolts using

$$E\ [\mathrm{eV}] = \frac{1.24}{\lambda\ [\mu\mathrm{m}]}. \tag{2}$$

In spectroscopy the photon's wavelength is often expressed as a wave number, which has units of cm^{-1}. Wavelength can be transformed into this unit by expressing it in centimeters and calculating its reciprocal. As the energy of photons increases, their particle nature becomes more prominent. X-rays can usually be considered as sets of particles, whereas light, i.e. photons with their wavelength in the optical region, has both wave and particle nature. (Saleh & Teich 2007: 2, 446–448.)

The intensity I of a monoenergetic light ray is

$$I(\vec{r}, t)\ [\mathrm{W/cm^2}] = E\phi(\vec{r}, t), \tag{3}$$

where E is the energy of a single photon and ϕ is the photon-flux density or the mean number of photons per unit time per unit area. The power delivered by monoenergetic light onto a given area A is

$$P(A, t)\ [\mathrm{W}] = \int_A I(\vec{r}, t)\ dA = E\Phi(A), \tag{4}$$

where Φ is the mean photon flux or

$$\Phi(A)\ [1/\mathrm{s}] = \int_A \phi(\vec{r}, t)\ dA. \tag{5}$$

(Saleh & Teich 2007: 2, 459–462.)

Naturally, the intensity of a polychromatic, or polyenergetic, light ray is an integral over the various wavelengths present in the ray, i.e.

$$I_{\mathrm{poly}}(\vec{r}, t)\ [\mathrm{W/cm^2}] = \int_0^\infty E(\nu)\phi(\vec{r}, t, \nu)\ d\nu. \tag{6}$$

The intensity of a monochromatic component of a light ray as a function of wavelength is called the spectrum, or intensity spectral density, of the light. It has a unit of W/(cm^2·Hz). (Saleh & Teich 2007: 410.)

$$I(\vec{r}, t, \nu)\ [\mathrm{W/(cm^2 \cdot Hz)}] = E(\nu)\phi(\vec{r}, t, \nu). \tag{7}$$

Generally, the intensity of a light ray varies over time. The mean number of photons $\bar{n}$ over a given area A in given time interval T varies accordingly.

$$\bar{n}\ [\text{count}] = \int_{t}^{t+T} \Phi(A, \tau)\ d\tau. \tag{8}$$

The light ray's ability to convey information despite this variance can be described by its signal-to-noise ratio (SNR), which is defined as

$$\text{SNR [no unit]} = \frac{\bar{n}^2}{\sigma_n^2}. \tag{9}$$

Term σ_n^2 is the variance of the number of photons,

$$\sigma_n^2\ [\text{count}] = \sum_{n=0}^{\infty} (n - \bar{n}) p(n), \tag{10}$$

where $p(n)$ is the fraction of the measurements which reported the number of photons to be n, i.e. the probability distribution of the various results. (Saleh & Teich 2007: 2, 459–462.) Notice that SNR is a function of the wavelength.

The intensity I of a monoenergetic light ray traveling in matter is attenuated over the length of the ray's path x through the matter exponentially, according to equation

$$I(\nu, x) = I_0(\nu) e^{-\mu(\nu)x} \Big|_{\nu=\nu_0}, \tag{11}$$

where I_0 is the ray's original intensity (as a number of photons), x the path length, and μ the attenuation coefficient of the matter (in units of cm^{-1}). It follows that

$$\ln \frac{I(\nu, x)}{I_0(\nu)} = -\mu(\nu)x \Bigg|_{\nu=\nu_0}. \tag{12}$$

The ray attenuates because photons interact with the matter by becoming absorbed in the matter or by scattering, i.e. changing the direction of their path. (Hendee & Ritenour 2002: 51–56.) Spectroscopy is based on studying how the intensity of light changes during interaction with matter.

The mean distance that the photons travel in the matter before interacting is the mean free path, which equals $\mu^{-1}(\nu)$. From Eq. (11) we see that a photon travels distance x in the matter without interaction with probability $e^{-\mu(\nu)x}$. The intensity of a polychromatic light attenuates in a more complex fashion, since each wavelength has its own attenuation coefficient. (Hendee & Ritenour 2002: 51–56.) Sometimes the attenuation coefficient is defined with number ten as the base number of the exponential attenuation, i.e.

$$I(\nu, x) = I_0(\nu)10^{-\mu'(\nu)x} \tag{13}$$

(Prasad 2003: 105). The base number of the attenuation coefficient must thus be considered when using it.

Attenuation coefficient is the sum of the absorption coefficient (μ_a) and scattering coefficient (μ_s), which describe the probabilities that a photon is absorbed or scattered, respectively. In other words,

$$\mu(\nu) = \mu_a(\nu) + \mu_s(\nu). \tag{14}$$

(Hendee & Ritenour 2002: 51–56.) The absorption coefficient is the product of the density of absorbing entities (ρ_a), which are atoms or molecules, and the entities' absorption cross section (σ_a), which in turn is the product of the entities' absorption efficiency (Q_a) and their geometric cross-sectional area (σ_g), i.e.

$$\mu_a(\nu) = \rho_a\sigma_a(\nu) = \rho_a Q_a(\nu)\sigma_g \tag{15}$$

(Wang & Wu 2007: 5; Välisuo 2011: 13). The absorption coefficient may also be defined as the product of the molar extinction coefficient ($\epsilon(\nu)$, in units of liter per centimeter, per mole, or L·mol^{-1}·cm^{-1}) and the molar concentration (in units of mole per liter) (Prasad 2003: 105). The scattering coefficient is derived similarly from the entities' density, scattering efficiency (Q_s) and their geometric cross-sectional area (σ_g), which yield the scattering cross section (σ_s) (Wang & Wu 2007: 8).

2.2.1. Atomic absorbance

A photon may be absorbed by either a single atom or by a molecule. When a photon is absorbed by an atom, its energy is transferred to one of the atom's electrons. A photon can be absorbed in this way only if all of its energy can be transferred to the electron. The conditions under which it is possible are called the selection rules. In such absorption the electron's energy level, i.e. its potential energy, is raised by the photon's energy and the photon disappears, i.e. is annihilated. Electrons that are bound to an atom have a finite set of possible amounts of energy, which are called energy levels. Thus the photon's energy can be transferred to the electron only if the electron's energy ends up at one of the possible energy levels, or if the photon has enough energy to free the electron from the atom. In the latter case the atom becomes an ion. This phenomenon is called ionization and photons that are able to induce it constitute ionizing radiation.

The energy level of an electron that is bound to an atom depicts the amount of potential energy that binds the electron to the atom. It is given as a negative value. If the electron receives an amount of energy equal to its energy level, the electron is unbound from the atom. The smallest amount of energy that can achieve that is called the ionization energy. It is 13.60 eV for a hydrogen atom, which corresponds to wavelength 91.235 nm in a vacuum. If the received amount of energy is greater, the rest of the energy becomes kinetic energy of the electron. (Young & Freedman 2000: 1463–1464.)

The energy levels that are possible for an atom's electrons are assigned four integers, which are called quantum numbers. The first integer, the principal quantum number (n), can be thought of as the distance at which the electron encircles the atom's nucleus. It has the greatest effect on the electron's energy level. Values of the principal quantum number are called shells of the electrons. In theory, the electron's energy level could be calculated from the quantum numbers using the Schrödinger equation. However, it leads to complex equations that have been solved exactly only for hydrogen, the simplest possible atom. (Young & Freedman 2000: 1548–1572.)

The second quantum number, the orbital angular-momentum quantum number (l) or the azimuthal quantum number, can be thought of as the degree to which the electron prefers one orbit around the atom's nucleus over other orbits of the same radius (Young & Freedman 2000: 1548–1572; Saleh & Teich 2007: 485). It is often called the orbital quantum number. It has also a direction which defines the electron's preferred orbit. Because of the uncertainty principle we can not know that direction exactly. However, the magnitude of one component of that direction is defined by the third quantum number, the orbital magnetic quantum number (m_l). The last quantum number, the spin quantum number (m_s), defines the electron's spin angular momentum. Its only possible values are spin up ($m_s = +1/2$) and spin down ($m_s = -1/2$). Unless the atom is in a magnetic field, the last three quantum numbers do not effect its energy level, apart from the coupling effects. The fact that electrons with different quantum numbers can have the same energy level is called degeneracy. (Young & Freedman 2000: 1548–1572.)

Selection rules are an important part of the theoretical description of how photons are absorbed by atoms or molecules. The selection rules of atomic absorbance are described next. When an atom absorbs a photon, one of the atom's electron's absorbs the photon's energy, which usually causes the electron to move from one shell to another ($n \rightarrow n+1$). Because of the principle of conservation of angular momentum, the electron absorbs the photon's angular momentum as well. Thus, the absorption causes the electron's orbital quantum number to change by one ($l \rightarrow l \pm 1$) and the electron's orbital magnetic quantum number can change at most by one ($\Delta m_l \leq 1$). These limitations are called the selection rules. (Young & Freedman 2000: 1560–1561, 1567–1568.) The Pauli exclusion principle states that each electron in a given atom must have a unique set of quantum numbers (n, l, m_l, m_s) (Young & Freedman 2000: 1560–1561, 1567–1568; Saleh & Teich 2007: 486). This 4-tuple can be called the quantum state of the electron. A change from one quantum state to another is called an allowed transition if it obeys the selection rules. Other transitions are called forbidden transitions. An electron can absorb a photon only if the atom has a free quantum state such that the electron can absorb the photon by moving to a new, free quantum state through an allowed transition. (Young & Freedman 2000: 1560–1561, 1567–1568.)

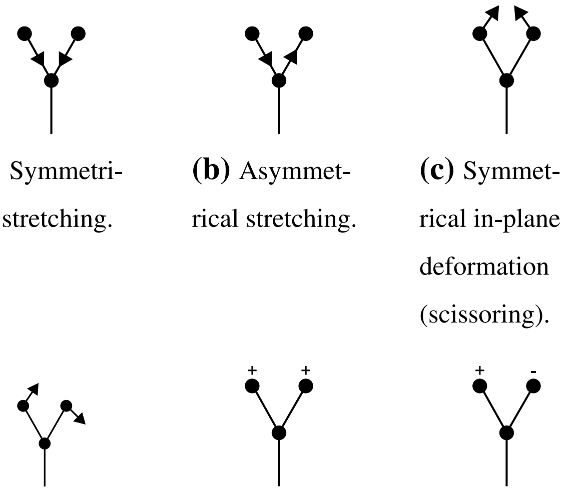
An electron that has the quantum state with the lowest energy level out of the available states is said to be in the ground level. Electrons in other states are in excited levels. An atom whose electrons are all in the ground state is itself in the ground state. Other atoms are in excited levels. (Young & Freedman 2000: 1456–1457.) Most of the sample's atoms and molecules are in the ground state, as described by the Maxwell-Boltzmann distribution. Thus, most of the photons absorbed by atoms or molecules are absorbed in a fundamental transition, i.e. in a transition from the ground state to the lowest excited state.

(Young & Freedman 2000: 1467–1468; Osborne et al 1993: 19–20.) Due to the selection rules the amount of energy that the photon must have in order to make the fundamental transition possible depends on the structure of the atom and, e.g. on the temperature of the sample. The amount of energy that the photon has is defined by the photon's wavelength (Eq. 1). Thus the number of photons, with a given wavelength λ, that are absorbed in the sample can be used to identify the structure of the atom. These numbers of photons can be calculated from the change in the light's intensity spectra.

2.2.2. Molecular absorbance

Molecules have energy levels that resemble the energy levels of atoms. Whereas atoms have four quantum numbers defining their quantum state, the number of quantum numbers for a molecule depends on the shape of the molecule, and on the number of atoms in it. The simplest of molecules, which contain only two atoms, have two quantum numbers. The first quantum number for a molecule, $l \in \mathbb{N} \cup \{0\}$, defines the molecule's rotational energy level. The molecule's rotational energy level depicts the speed at which the molecule rotates around its center. (Young & Freedman 2000: 1587–1588.) The rest of the molecule's quantum numbers define the molecule's vibrational energy level. Vibrational energy level depicts how the distances between the molecule's atoms change back and forth over time, or vibrate, around their average values, and the energy associated with this movement. The total energy level of a molecule is approximately the sum of the rotational energy level and the vibrational energy level. (Young & Freedman 2000: 1589–1590; Bokobza 2002: 17–22.) Much like atoms, molecules can raise their energy level by absorbing a photon which has an amount of energy that is equal to the difference between the molecule's current energy level and the next energy level.

In a molecule that has two atoms, the only vibrational degree of freedom is the distance between the two atoms. A molecule that has N atoms has $3N - 6$ vibrational degrees of freedom. Linear molecules, were the atoms form a single line, have $3N - 5$ vibrational degrees of freedom. The atoms may, for example, all move symmetrically away from the center of the molecule, two of the atoms may move towards each other, or the molecule may bend (Fig. 15). Each vibrational degree of freedom has a vibrational quantum number (v_i), and the molecule's vibrational energy level is a function of all of them,



(a) Symmetrical stretching.

(b) Asymmetrical stretching.

(c) Symmetrical in-plane deformation (scissoring).

(d) Asymmetrical in-plane deformation (rocking).

(e) Symmetrical out-of-plane deformation (wagging).

(f) Asymmetrical out-of-plane deformation (twisting).

Figure 15. Modes of vibration for a triatomic molecule or group AX_2, i.e. ways in which a molecule which has three atoms may vibrate (Reproduced from Osborne et al 1993: 21).

$G(v_1, v_2, \ldots)$. Each vibrational degree of freedom may be excited to a higher energy level by absorbing a photon whose energy ($h\nu$) matches the difference between the current vibrational energy level and the next vibrational energy level of that vibrational degree of freedom. (Bokobza 2002: 17–22.)

An event where the molecule's vibrational quantum number changes from zero to one is called the fundamental transition, and other events where the molecule's vibrational quantum number changes by one are called hot bands. Transitions where the vibrational quantum number changes by more than one are called overtones. If the quantum number changes by two the transition is first overtone, if it changes by three the transition is second overtone, and so on. (Bokobza 2002: 14–16.) The probability of overtone transitions decreases as the change of the vibrational quantum number increases such that overtones above second overtone are very rarely observed by (NIR-) spectroscopy (Osborne et al 1993: 19).

Events where more than one of the molecule's vibrational quantum numbers are changed by an absorption of a photon are called combination transitions. If $\sum_i \Delta v_i = 2$, the transition is called a binary combination; if $\sum_i \Delta v_i = 3$ it is a tertiary combination, and so on. A transition where one of the quantum numbers decreases and another increases is called a difference transition. The energy of the photon that is absorbed during such transition matches the difference between the molecule's vibrational energy level before and after the transition. (Bokobza 2002: 18–21.) Combination transitions and difference transitions have a very low probability of occurrence (Osborne et al 1993: 20).

2.2.3. Scattering

The direction of photons may change when they encounter particles. This phenomenon is called scattering. Assuming that the scattering particle is a homogeneous sphere, that the light is monochromatic, and that the wavefront of the incident light is much wider than the particle and much wider than the wavelength of the light, a model of scattering can be derived from Maxwell's equations. The resulting theory is called the Mie theory. If the scattering particle is much smaller than the light's wavelength, scattering can be modelled by a simpler model called Rayleigh theory. (Wang & Wu 2007: 17, 20, 26.)

The anisotropy factor

$$g = \langle \cos\theta \rangle = 2\pi \int_0^\pi p(\theta)\cos(\theta)\sin(\theta)\,d\theta \tag{16}$$

depicts the material's tendency to scatter the photons forward or backward. The phase function ($p(\theta)$) gives the probability that the polar angle ($0 \leq \theta \leq \pi$) between the photon's direction before and after the scattering event is θ. (Välisuo 2011: 13–14; Wang & Wu 2007: 46–47.) If $g = 0$ the material is fully isotropic, and all directions are equally likely after a scattering event. If $g \approx 1$ the photons tend to maintain their direction and if $g \approx -1$ the photons tend to invert their direction in scattering events. (Välisuo 2011: 13–14.)

As a photon undergoes several scattering events, the changes in its direction cumulate. If the anisotropy factor g is not precisely 1 or -1 and if the photon's path in the matter is long enough, the angle between the photon's eventual direction and its original direction will ultimately have equal probability for all values, making the scattering effectively isotropic at that pathlength. This phenomenon can be modelled by assuming that the material scatters light isotropically and by reducing the value of the scattering coefficient to reflect the pathlength required to make the scattering isotropic. Such a scattering coefficient is called the reduced scattering coefficient, or transport scattering coefficient, and denoted as

$$\mu_s' = \mu_s(1-g). \tag{17}$$

(Välisuo 2011: 14; Wang & Wu 2007: 93–94.) The total attenuation coefficient calculated from the absorption coefficient and the reduced scattering coefficient is called the reduced interaction coefficient,

$$\mu' = \mu_t' = \mu_a + \mu_s'. \tag{18}$$

Its reciprocal is called the transport mean free path, $l_t' = 1/\mu_t'$. (Wang & Wu 2007: 93–94.)

According to the Rayleigh theory the scattering coefficient (μ_s) is proportional to λ^{-4} (Wang & Wu 2007: 18). Scattering events that can be modelled by the Rayleigh theory, i.e. scattering from particles that are smaller than the wavelength, constitute Rayleigh scattering. Other scattering events constitute Mie scattering. The scattering coefficient resulting from Mie scattering alone is proportional to $\lambda^{-1.5}$. The total scattering coefficient can be modelled as the sum of the scattering coefficient that would result from the Rayleigh theory alone and the coefficient that would result from Mie scattering alone, i.e.

$$\mu_s = \mu_{s,\text{Mie}} + \mu_{s,\text{Rayleigh}} = k\lambda^{-4} + k'\lambda^{-1.5}. \tag{19}$$

(Välisuo 2011: 9, 36.)

If the energy of the photon does not change as a result of the scattering event, the scattering is called elastic scattering. It may occur as the scattering particle first absorbs the incident photon, and then emits a photon with identical energy to another direction. (Prasad 2003: 93–95) Another possible source of elastic scattering is that the incident photon, as electromagnetic radiation, may induce a vibrating motion on an electron of the scattering particle, such that the electron vibrates with the same frequency as the incident radiation. As an accelerating (and decelerating) electric charge, the vibrating electron will then emit electromagnetic radiation with the same frequency, and with a phase shift of π, thus effectively scattering the incident photon to another direction without changing the photon's energy. (Erkkilä 1992: 5.)

In inelastic scattering the photon's energy changes – or more precisely, the scattering particle absorbs the incident photon, and then emits a photon with a different amount of energy, and hence different wavelength (λ) and frequency (ν). The energy level of the scattering particle changes by an amount that is equal to the difference between the energy level of the emitted photon (E_f) and the energy level of the incident photon ($\Delta E = E_f - E_i$). (Prasad 2003: 93–95.) This kind of scattering is called Raman scattering, and the change in the photon's energy is called the Raman shift. If the scattering particle is originally at its ground level, Raman scattering usually occurs as the particle absorbs a photon and then returns to its lowest excited level, i.e. to the energy level that corresponds to its fundamental transition. The emitted photon will then have less energy than the incident photon($E_f < E_i, \nu_f < \nu_i$), and the difference in the photons' energies equals the amount of energy that the particle would have absorbed from a single photon in its fundamental transition, except that the value is negated. Such scattering is called Stokes Raman scattering. (Furukawa 2002: 86–87.)

2.3. Summary

A human tooth consists of three layers. The outermost layer consists of enamel, which is hard, mineralized tissue which protects the tooth. The main body of the tooth, dentin, begins underneath the enamel and continues throughout the rest of the tooth. Dentin has a hollow center which contains the third tissue type of the tooth: pulp. Pulp is a soft connective tissue, which contains blood vessels and nerves.

Dental caries is demineralization of dental tissue. Its formation depends on three factors: the type of bacteria present in the mouth, the chemical composition of the teeth surfaces, and the types of food consumed. Dental caries is caused by organic acids which are produced when bacteria present in the mouth ferment carbohydrates, e.g. sucrose or table

Table 4. Division of infrared region (Osborne et al 1993: 21).

Region	Characteristic transitions	Wavelength range	Wavenumber range
Near infrared (NIR)	Overtones and combinations	700–2500 nm	14300–4000 cm^{-1}
Mid infrared (MIR)	Fundamental vibrations	2500–5$\times$10^{4} nm	4000–200 cm^{-1}
Far infrared (FIR)	Rotations	5$\times$10^{4}–10^{6} nm	200–10 cm^{-1}

sugar. The pH of the plaque drops up to two units in 10 minutes after eating carbohydrates. The formation of dental caries begins in enamel when the pH drops below the so called critical pH. After 30–60 minutes the pH returns to its normal value. Caries is most common in the grooves of biting surfaces of molars and on surfaces between teeth (Jones et al 2003: 2260). These surfaces are also more difficult to diagnose for caries than other surfaces of the teeth. The most difficult site to diagnose caries in is near an existing restoration (filling). (Beighton & Bartlett 2006: 86.)

Spectroscopy is the art and science of identifying various properties of the matter by measuring the light's intensity at various wavelengths before and after it has interacted with the matter. Electromagnetic radiation in the approximate wavelength range 390–790 nanometers constitutes the visible light. Wavelength regions for the various types of infrared light are presented in Table 4. NIR-Spectroscopy (NIRS) refers to spectroscopy using NIR-light, and Raman scattering refers to spectroscopy that is based on Raman scattering.

During the interaction an atom or a molecule in the matter may absorb a photon if that photon has a suitable wavelength. What a suitable wavelength for a given atom or molecule is depends on the structure and energy level of the atom or molecule. For example, the temperature of the matter influences the energy levels of the atoms or molecules. The absorption of photons of specific wavelengths can be observed in the intensity spectra of the light after the interaction. Thus, the spectra contains information about the structure of the atoms and molecules, among other things.

The direction of photons may change when they encounter atoms or molecules. This phenomenon is called scattering. The probability that a given photon scatters to a given direction depends on the photon's wavelength and on certain properties of the matter. Observation of scattering of large numbers of photons of various wavelengths may provide information about the properties of the matter. In real life measurements both absorption and scattering are always present to some extent.

3. INTRODUCTION TO ANALYSIS OF SPECTROSCOPIC RESULTS

This chapter will introduce the reader to the methods that were used in this project for analysing the results of the spectroscopic measurements. Most of these methods belong to the field of pattern analysis or pattern recognition.

3.1. Overview of the analysis

When the spectroscopic measurements have been made, the resulting data is analysed in order to find a method which can detect which measurements are made from the caries lesions, i.e. to be able to tell from a given measurement whether it depicts a site of healthy enamel or a site of carious enamel. In spectroscopy, the development of such a method is called calibration. More precisely, spectroscopical measurements seek to assess some property of the sample being measured, whereby that property can be described by a scalar or a vector. Calibration refers to the steps that are required to be able to approximate the value of the target property based on the measurement results of a given sample, made on a given spectroscope and on a given measurement setup. (Osborne et al 1993: 99–100.) The scalar or vector that describes the assessed property of a sample can be called a dependent variable. We will outline the steps of calibration after a short introduction to its background.

A piece of information can be used for approximating the value of the dependent variable if, and only if, it is available every time the approximation method is used (if some information is available only occasionally, we may create several approximation methods that are used either as alternatives or as subsequent steps of the approximation, such that each of these methods can be used only when all information that it uses is available). Any piece of information that is associated with a sample and can be used for classifying the samples is called a feature or a variable of the sample. Typically, all the samples are required to have the same (sub-) set of features. For example, the intensity at a single wavelength is one feature of a sample if the intensity at that wavelength is measured for all samples during the calibration and during later use of the approximation method.

Each feature is modelled as a numerical value, after which each sample can be represented as a vector of those values. The features have to be in the same order for every sample to make the vectors comparable. Since the samples are encoded in this way, a natural form of presenting a set of samples is a matrix. Usually the matrix is orientated so that each row contains a single sample, and the columns correspond to the features. Let $\vec{\mathbf{x}}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$ be a sample, where $x_{i,j}$ is a value of a feature. The samples can then be represented as a matrix

$$\mathbf{X} = \begin{bmatrix} \vec{\mathbf{x}}_1 \\ \vec{\mathbf{x}}_2 \\ \vdots \\ \vec{\mathbf{x}}_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}. \tag{20}$$

The problem of finding a method to approximate the value of the dependent variable based on the available features is a task for a pattern recognition algorithm. A pattern recognition algorithm seeks to find one or more relations between the samples' features and another set (or vector) of values, namely the dependent variable, or between two or more subsets of the samples' features. Such a relation is called a pattern and the search for them is called pattern analysis. (Shawe-Taylor & Cristianini 2004: 3–6.) Let the dependent variable be $\vec{\mathbf{y}}_i \in \mathbb{R}^k$, where $k \in \mathbb{N}$. The pattern being searched for is a relation $P(\vec{\mathbf{x}}_i, \vec{\mathbf{y}}_i)$ that holds for all samples of the object that is being studied, or at least for as large a fraction of them as possible.

Classification is a process where a set of samples is divided into two or more subsets, which are called classes, according to some property of the samples. When the samples are divided into two classes, the process is called binary classification. When there are more than two classes the process is called multiclass classification. (Shawe-Taylor & Cristianini 2004: 19–20.) A binary classification task might be, for example, to divide a set of steaks into sets of fresh and spoiled steaks.

If a pattern is being searched for between the samples' features and a dependent variable, which has at most a countable number of possible values, the dependent variable can be interpreted as an indication of the class into which the sample is classified, and thus the task of finding such pattern is a task of building a classification method. The value that indicates the sample's class is called the sample's label. In binary classification the labels are usually positive unit and negative unit (1 and -1), i.e. $\vec{\mathbf{y}}_i = y_i \in \{1, -1\}$. (Shawe-Taylor & Cristianini 2004: 3–6, 19–21.) Classification typically refers to the use of a classification method that was built using a pattern recognition algorithm.

If a pattern exists between two or more subsets of the samples' features for most of the samples, then the samples for which that pattern does not hold may be considered as significantly different from the other samples. Such samples are called outliers. The task of finding such a pattern is called anomaly detection or novelty detection. If the dependent variable has an uncountable number of possible values, the task of finding the pattern is called regression. (Shawe-Taylor & Cristianini 2004: 19–21.) This kind of dependent variable might be, for example, the age of a steak, which may in general have any real value.

Pattern analysis produces a prediction function $g(\vec{\mathbf{x}}_i) \approx \vec{\mathbf{y}}_i$. The prediction function tries to approximate the pattern as closely as possible. The output of the prediction function is called a prediction for that sample. If the pattern exists (only) between two or more subsets of the samples' features, prediction function gives a measure of how much the given sample $\vec{\mathbf{x}}_i$ differs from that pattern, i.e. how much it looks like an outlier. In that case, the value of the prediction function should be close to zero for most samples. (Shawe-Taylor & Cristianini 2004: 10, 19.) Pattern analysis, i.e. search for a pattern, is sometimes called training of the prediction function, or training of the classification method or regression method where that prediction function is used.

A loss function is defined as a function that gives the cost of making a given prediction for a sample when the sample's dependent variable has a given value. It can be denoted as $L(g(\vec{\mathbf{x}}_i), \vec{\mathbf{y}}_i)$. The value of the loss function is also called a loss. A prediction which has a smaller cost is considered better, i.e. the cost or loss is a measure of how bad the prediction is. The conditional risk of a given prediction for a given sample is the expected cost of making that prediction for that sample. Formally, the conditional risk is

$$R(\vec{\mathbf{y}}_i, \vec{\mathbf{x}}_i) = \int_{\mathcal{Y}} L(g(\vec{\mathbf{x}}_i), \vec{\mathbf{y}}_i) P(\vec{\mathbf{y}}_i | \vec{\mathbf{x}}_i), \tag{21}$$

where $\mathcal{Y}$ is the set of dependent value's possible values, i.e. its domain, and $P(\vec{\mathbf{y}}_i | \vec{\mathbf{x}}_i)$ is

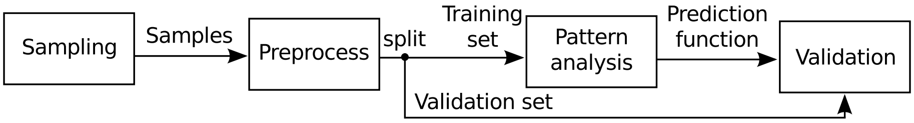


Figure 16. An outline of the calibration steps.

the probability that $\vec{\mathbf{y}}_i$ is the correct value for the dependent variable for sample $\vec{\mathbf{x}}_i$. The latter term maybe easier to understand through the right hand side of the Bayes formula:

$$P(\vec{\mathbf{y}}_i|\vec{\mathbf{x}}_i) = \frac{P(\vec{\mathbf{x}}_i|\vec{\mathbf{y}}_i)P(\vec{\mathbf{y}}_i)}{P(\vec{\mathbf{x}}_i)}, \tag{22}$$

where $P(\vec{\mathbf{y}}_i)$ is the probability of obtaining a sample whose dependent variable has the given value, $P(\vec{\mathbf{x}}_i)$ is the probability of obtaining the given sample, and if the dependent variable indeed has value $\vec{\mathbf{y}}_i$, then the sample in question is $\vec{\mathbf{x}}_i$ with probability $P(\vec{\mathbf{x}}_i|\vec{\mathbf{y}}_i)$. The overall risk of a given prediction function is the expected conditional risk of making a prediction with that function for a sample from a given source. Formally,

$$R = \int_{\mathcal{X}} R(g(\vec{\mathbf{x}}), \vec{\mathbf{x}})P(\vec{\mathbf{x}}), \tag{23}$$

where $\mathcal{X}$ is the set of all possible samples from the given source. The prediction function should be selected so that the overall risk is minimized. (Duda, Hart & Stork 2001: 24–26.) Shawe-Taylor & Cristianini (2004: 22, 29) depict the accuracy of a prediction function by a pattern function, which is defined as

$$f(\vec{\mathbf{x}}_i, \vec{\mathbf{y}}_i) = |\vec{\mathbf{y}}_i - g(\vec{\mathbf{x}}_i)|. \tag{24}$$

Trivially, a pattern function is a kind of loss function. When the prediction function is accurate (for the given sample) f gives a value that is close to zero.

We will now outline the steps of calibration (Fig. 16). Usually some kind of signal processing is performed on the samples' features before the pattern recognition algorithm is used. These processing steps are called preprocessing. Their goal is to improve classification results by making the pattern easier to identify. In order to be able to search for the pattern, one must first gather a set of samples for which the value of the dependent variable is known. This might mean, for example, measuring samples from fresh and spoiled steaks, and recording the class, or label, of each sample. This set is then divided into two subsets. The first of these subsets is called training set and it is used in pattern analysis for producing a prediction function $g(\vec{\mathbf{x}}_i)$. The second subset is called validation set. The prediction function $g(\vec{\mathbf{x}}_i)$ is used to predict the value of the dependent variable $\vec{\mathbf{y}}_i$ for the samples in the validation set, using only the samples' features. For example, we might use pattern analysis on the training set to approximate the pattern between the samples' spectra and their freshness status. Then we could use the spectra of a steak sample in the validation set to predict whether the sample was measured from a fresh steak or a spoiled steak. By comparing the dependent variables' known values $\vec{\mathbf{y}}_i$ for the samples in the validation set with the predictions $\vec{\mathbf{y}}'_i$ for those values we get a measure of the accuracy of the pattern's approximation, or the accuracy of the prediction function. This step is called validation of the pattern. In our example we would compare the known freshness status of the validation set's samples with the predictions of that status. The training set can be seen as a representation of the samples that were gathered for the calibration, while the validation set represents the samples from which a prediction will be calculated at some point after the calibration.

A pattern analysis task can be either supervised, semisupervised or unsupervised, depending on how much information about the dependent variable's value is provided with the training set. In supervised pattern analysis the training set contains the correct value for the dependent variable, e.g. the label, for the samples in that set. In semisupervised pattern analysis the training set contains partial information about the dependent variable, e.g. the correct ordering of the samples based on that variable's values, but not the values themselves. In unsupervised pattern analysis the training set does not contain any information about the dependent variable. This type of pattern analysis can be used only for finding patterns between two or more sets of the samples' features. Search for a classification method using unsupervised pattern analysis is called clustering. (Shawe-Taylor & Cristianini 2004: 19–21.) In this project we will use only supervised pattern analysis.

The objective of a pattern recognition algorithm is to find a prediction function that gives as accurate predictions as possible – not only for the samples in the training set, but also for all other similar samples, or samples from the same source. The phrase "all other similar samples" might mean, for example, all other steaks whose freshness we want to evaluate. The prediction function's ability to correctly predict the class of similar samples that were not in the training set is called the quality of generalization. This term refers to the idea that the prediction function generalizes the relation between the samples $\vec{\mathbf{x}}_i$ and the dependent variables $\vec{\mathbf{y}}_i$, which can be observed in the training set. (Shawe-Taylor & Cristianini 2004: 14.)

The set of prediction functions that the algorithm considers as possible solutions, i.e. the set of patterns that it considers possible, is limited by the assumptions that the algorithm makes about the pattern that it is searching for. If the algorithm makes too many inaccurate assumptions about the pattern and ends up ruling out the best patterns that actually exist between the samples and their labels, the algorithm is said to underfit the data. In the opposite case the algorithm has too few assumptions about the pattern and it ends up selecting a prediction function that seems to relate the samples of the training set to their labels, but which does not depict the correct pattern. In this case the algorithm is said to overfit the data. The prediction function may then seem to perform very well on the training set, but it will most likely perform poorly on samples that were not in the training set. This happens because the algorithm does not know what kind of prediction function it is searching for, so it selects some function that seems to make good predictions on the samples of the training set. The predictions may be based on, e.g. some set of random measurement errors, i.e. noise, that happens to correlate with the labels in the training set, but which, being random, does not correlate with the labels of other similar samples. (Shawe-Taylor & Cristianini 2004: 14–15.) If the algorithm ends up selecting the correct pattern, the assumptions that it used are of little concern.

The way the training set is selected from the samples that can be obtained from the source should have only small effects on the results of a pattern recognition algorithm. In other words, a pattern recognition algorithm should produce similar results for all training sets of similar samples. This implies that the algorithm does not overfit or underfit the data too easily. Algorithms that fulfill this objective are called (statistically) stable. Algorithms should also be robust, i.e. be able to detect the patterns, or good approximations for them, even if the samples contain random errors, i.e. noise or wrong labels, as long as the errors are neither too large nor too frequent. Naturally, the algorithms should also be efficient enough to enable their use with the kinds of quantities of samples and features that are met in real-life pattern recognition problems. (Shawe-Taylor & Cristianini 2004: 12–13.) Algorithms that are able to correctly classify all samples in the training set are called consistent (Shawe-Taylor & Cristianini 2004: 212).

3.2. Preprocessing

Preprocessing aims to improve the accuracy of the prediction function that can be obtained by the calibration by using signal processing methods on the samples. Preprocessing tries to filter the information in the samples' features such that the parts of the information that are most useful for selecting the prediction function are kept, while the least useful parts are removed. Particularly information that relates to the properties of the spectrometer that was used to make the measurements is removed, leaving only information about the samples.

Although the prediction accuracy can be expected to improve with increasing number of features, in practice it often begins to decrease at some point when more and more features are added (Duda et al 2001: 107–111). In order to avoid the problems caused by having too many features, the number of features may be reduced by preprocessing, such that the parts of the information that are most useful for the prediction are preserved. Reduction of the number of features is called dimensionality reduction. It can be done by removing some of the samples' features, which is called feature selection. Another method to reduce the number of features is to replace all existing features with a new set of features, which effectively projects the samples to a new coordinate system. Such methods are called feature generation methods. (Theodoridis & Koutroumbas 2006: 702–704)

Generally, empirical measurements of physical phenomena contain many relatively small random errors that are collectively called noise. Noise can be depicted as a vector which is added to the features. Preprocessing is used to reduce the amount of noise in the samples, or to remove it completely if possible, before running the pattern analysis. This is called noise cancellation. It requires either information or assumptions about the probability distribution of the noise, i.e. the probabilities with which the noise obtains various values. An often used assumption about noise is that it is normally distributed, i.e. that its probability distribution is the the Gaussian distribution with a zero mean. The Gaussian distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \sigma > 0, \tag{25}$$

where μ is the mean value and σ is the standard deviation of the distribution. The Gaussian distribution is symmetrical with respect to its mean value. Thus if we could extract the noise from several measurements which contain normally distributed noise, the average of those samples of noise can be expected to equal the mean value μ.

In this project we will assume that the noise is normally distributed. Thus, the samples are

$$\vec{\mathbf{x}}_i = \vec{\mathbf{x}}_i^* + \xi, \tag{26}$$

where $\vec{\mathbf{x}}_i^*$ is the true measurement value and ξ is the noise. In this case, taking multiple measurements of the same sample and calculating the average of the measurement results should reduce the amount of noise, because noise from different measurements can be expected to cancel each other out and the true measurement value should stay the same.

$$\vec{\mathbf{x}}_i = \frac{1}{K}\sum_{k=1}^{K}(\vec{\mathbf{x}}_{i,k}^* + \xi_k) = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\vec{\mathbf{x}}_{i,k}^*}_{\text{expected to equal } \vec{\mathbf{x}}_i^*} + \underbrace{\frac{1}{K}\sum_{k=1}^{K}\xi_k}_{\text{expected to equal } \mu = 0} \quad . \tag{27}$$

In complete darkness the measurement result of an optical device like a spectrometer should be zero. However, this is not necessarily the case. The measurement result obtained from the spectrometer in darkness is called the dark current of the spectrometer. The dark current is added to all measurement results obtained with that spectrometer. As information about the spectrometer rather than the sample, the dark current should be removed from the samples before pattern analysis. In this context it is also known as the black reference.

The numerical range in which the measurement results are reported depends on the spectrometer. Thus, a given sample may produce different numerical values for the features, or the spectra, in different spectrometers. In general, a wider range implies better measurement resolution. Since the numerical range is information about the spectrometer rather than the sample, it should be removed by transforming the samples into a fixed numerical range, typically range $[0, 1]$. Such transformation is called normalization. Normalization is done by measuring a sample called a white reference, or simply a reference, which approximates a sample that reflects all incident light, and by dividing the samples' spectra with the spectra of the white reference.

3.2.1. Savitzky-Golay method

Noise cancellation, or reduction of the amount of noise in the sample, can also be done by smoothing the spectra. Smoothing removes rapid changes in the spectra, which are assumed to represent noise. Savitzky-Golay is a method that can be used for smoothing a spectra. It is also known as polynomial smoothing or digital smoothing polynomial (DISPO). Derivation of the spectra emphasizes rapid changes in the spectra, including noise. Thus if the spectra is going to be derivated, it is recommendable to smooth it first. Savitzky-Golay is able to both smooth and derivate a spectra simultaneously.

In the Savitzky-Golay method a polynomial of given degree, say d, is fitted to the spectra, by least-squares techniques, in a moving window of given odd width, say L, and the point, or feature, in the middle of the window is given the value of the polynomial at that point. To elaborate, for feature $x_{i,j}$ in sample $\vec{\mathbf{x}}_i$ the moving window with width L refers to L consecutive samples such that $x_{i,j}$ is in the middle of them, i.e. to samples $x_{i,(j+t)}$, where $t \in [-(L-1)/2, (L-1)/2]$. Let the feature $x_{i,j}$ depict the intensity at wavelength λ_j. Coefficients a_i of polynomial $p(\lambda_j) = a_0 + a_1\lambda_j + a_2\lambda_j^2 + \ldots + a_d\lambda_j^d$ are selected such that the values of this polynomial approximate the values $x_{i,(j+t)}$, i.e. $p(\lambda_{j+t}) \approx x_{i,(j+t)}$. The Savitzky-Golay method replaces the value of $x_{i,j}$ in sample $\vec{\mathbf{x}}_i$ with the value of $p(\lambda_j)$. Each feature $x_{i,j}$ is associated with a different window and a different polynomial. The Savitzky-Golay method can be used to derivate the spectra while it is being smoothed by derivating the polynomials. In that case, the degree of the polynomial should be greater or equal to four. (McClure 2008: 100-102; Press, Teukolsky, Vetterling & Flannery 1992: 650–651.)

Calculating the coefficients for the polynomials is computationally complex. If the feature's wavelengths are regularly separated, i.e. the difference between two consecutive wavelengths is constant, the coefficients can be calculated efficiently with matrix operations. For irregularly separated wavelengths the matrix operations produce coefficients that effectively add noise to the sample before smoothing it. However, if the sample already contains so much noise that the noise added by the matrix operations would be negligible in comparison, the matrix operations can be used even for irregularly separated wavelengths. (Press et al 1992: 653–654.)

3.2.2. Decimation

Decimation or down-sampling is a simple method to reduce the number of features in the samples. In decimation only every K^{th} feature in the samples is kept, and the rest are removed. Formally it is defined as

$$x'_{i,j} = x_{i,jK}. \tag{28}$$

(Porat 1997: 461–462.) Since decimation removes some of the sample's features, it can be considered a feature selection method (see page 48).

Decimation may cause aliasing, which would impair the quality of the samples. Thus, the samples need to be filtered with a decimation filter, which is a low-pass filter whose cutoff frequency is π/K, i.e. spectral components that correspond to frequencies less than the cutoff frequency need to be removed from the samples. (Porat 1997: 469–470.) This can be achieved with a finite impulse response (FIR) filter which has an appropriate vector of coefficients, i.e. an appropriate window. FIR-filter convolves each sample with the vector of the filter's coefficients, i.e.

$$z_{i,j} = \sum_{k=0}^{j} x_{i,k} \cdot c_{j-k}, \tag{29}$$

where $\vec{\mathbf{z}}_i$ is the filtered sample, $\vec{\mathbf{x}}_i$ is the unfiltered sample, and $\vec{\mathbf{c}}$ is the vector of coefficients. The vector of coefficients is also called the filter's window (Porat 1997: 168). Let $L = |\vec{\mathbf{c}}| = 2N + 1, N \in \mathbb{N}$ be the length of the filter's window. The sample is padded with zeros at the start and at the end for the duration of the calculation so that all (original) features can be placed at the center of a window. The values of $(L-1)/2$ first features and the values of the same number of last features are affected by these zeros. These values are not used in the results of the decimation, which leaves $(m-(L-1))$ usable values. When only every K^{th} feature in the samples is kept, only $\lfloor(m-(L-1))/K\rfloor$ features remain.

There are several well-known equations, called windows, for calculating the coefficients for a low-pass filter. In this study we will use the Hamming window. We still have to choose the number of coefficients to use, or the size of the filter's window. The more coefficients the filter uses, the better the filter is able to remove the desired spectral components, and the more features at the start and at the end of the samples are affected by zero padding. (Porat 1997: 163–173.)

The performance of the FIR-filter can be evaluated by calculating the Fourier transform of the vector of its coefficients. This produces a periodic function whose prime period is $[-\pi, \pi]$, and which maps each value in this interval to a complex number. Let that function be $f(\omega) : \mathbb{R} \to \mathbb{C}$. We can plot the magnitude of the complex numbers $f(\omega)$ as a function of the angular frequency ω. We can then compare the shape of the plot to the ideal shape of the filter. An ideal low-pass filter would have a unit magnitude until a cut-off frequency, after which the ideal magnitude would be zero. Generally, the ideal shape can not be achieved. The longer the filter's window is, i.e. the more coefficients the filter has, the better approximation of the ideal shape can be achieved. However, having a longer window increases the number of features that are affected by the added zeros.

3.3. Pattern analysis

In one of the methods to analyse the samples that is used in this project, the analysis begins with further dimensionality reduction with principal component analysis (PCA), followed by development of a classification method with support vector machines (SVM). In that method some of the dimensionality reduction is done in the preprocessing step and some in the pattern analysis step, because the latter dimensionality reduction method (PCA) requires training with the training set. When the validation step uses cross-validation, as in our case, such dimensionality reduction methods can only be trained after the samples have been divided into a training set and a validation set. Preprocessing is done to all samples before cross-validation, and thus these kinds of dimensionality reduction methods have to be considered as parts of the pattern analysis step, rather than the preprocessing step. In some other analysis methods used in this project the analysis method implicitly performs feature selection.

3.3.1. Principal component analysis

Principal component analysis (PCA) is a dimension reduction method that seeks to produce a model of the samples, such that the sum of squared differences between the actual (original) samples and the samples depicted by the model is minimal, when the model can use at most a given number of variables to depict the samples. In other words, PCA seeks to represent the samples with a new set of features, such that criterion value J is minimal, where

$$J = \sum_{k=1}^{n} ||\mathcal{M}(\vec{\mathbf{x}}'_k) - \vec{\mathbf{x}}_k||^2, \tag{30}$$

$\vec{\mathbf{x}}_k$ is an original sample, $\vec{\mathbf{x}}'_k$ is the corresponding sample represented with the new set of features, using at most a given number of features, and $\mathcal{M}(\cdot)$ is a transformation from the new set of features to the original set of features (a model). It is also known as the Karhunen-Loève transform (KLT). (Duda et al 2001: 115, 568.) PCA can trivially be considered a feature generation method, because it replaces all existing features with a new set of features (see section 3.2).

Each sample is a vector of features. The set of such vectors can be seen as the original model of the samples. PCA transforms these vectors by projecting them to a new cartesian coordinate system, such that the mean sample

$$\vec{\mathbf{m}} = \frac{1}{n} \sum_{k=1}^{n} \vec{\mathbf{x}}_k \tag{31}$$

defines the origo of the new coordinate system in units of the original coordinate system, and unit vectors called principal components define the basis vectors of the new coordinate system in units of the original coordinate system. The samples' coordinates in the new coordinate system are called the scores of the samples. They become the samples' features after the transformation. If the new coordinate system contains the same number of dimensions as the original coordinate system, the transformation, i.e. the projection, does not lose any information – apart from the rounding errors due to finite machine precision. It can be shown that in order to keep the criterion value J minimal for a given number of principal components, the principal components must be the eigenvectors of the scatter matrix $\mathbf{S}$, such that the eigenvectors correspond to the largest eigenvalues, where

$$\mathbf{S} = \sum_{k=1}^{n} (\vec{\mathbf{x}}_k - \vec{\mathbf{m}})(\vec{\mathbf{x}}_k - \vec{\mathbf{m}})^T. \tag{32}$$

The principal components are sorted such that the first principal component is the eigenvector that corresponds to the largest eigenvalue, the second principal component is the eigenvector that corresponds to the second largest eigenvalue, and so on. (Duda et al 2001: 115–117, 568.)

The transformation from the new set of features to the original set of features is

$$\mathcal{M}(\vec{\mathbf{x}}'_k) = \vec{\mathbf{m}} + \vec{\mathbf{x}}'_k \cdot \mathbf{B}, \tag{33}$$

where i^{th} row of $\mathbf{B}$ is the i^{th} principal component and the number of rows in matrix $\mathbf{B}$ corresponds to the number of new features. The transformation from the original set of features to the new set of features is

$$\mathcal{M}^{-1}(\vec{\mathbf{x}}_k) = \mathbf{B}^T \cdot (\vec{\mathbf{x}}_k - \vec{\mathbf{m}}) \tag{34}$$

(Duda et al 2001: 568).

3.3.2. Support vector machine

Support vector machine (SVM) is a supervised pattern recognition algorithm. It can be used for regression or classification. When it is used for classification, the resulting algorithm is called a support vector classifier (SVC). (Chang & Lin 2001: 3.) SVC uses a model $\mathcal{M}$ to depict the samples, so that $\vec{\mathbf{x}} = \mathcal{M}(\vec{\mathbf{p}})$, where $\vec{\mathbf{p}} = (p_1, p_2, \ldots, p_d)$ are the model's parameters that correspond to sample $\vec{\mathbf{x}}$. The model is selected so that there is a linear relation between the model's parameters and the sample's class – or at least so that the relation is as linear as possible. (Shawe-Taylor & Cristianini 2004: 16–17, 33, 212–213.) The model must be such that the set of all possible parameter vectors $\vec{\mathbf{p}}$ forms a Hilbert space, i.e. a vector space that is an inner product space, separable and complete (Shawe-Taylor & Cristianini 2004: 48–50).

The vector space F that is formed by the parameter vectors is called a feature space. A function that gives the parameters that correspond to a given sample, i.e. $\vec{\mathbf{p}} = \mathcal{M}^{-1}(\vec{\mathbf{x}}) = \phi(\vec{\mathbf{x}})$, is called a feature map or an embedding map. The latter term refers to an idea that the function *embeds* the data into the feature space. (Shawe-Taylor & Cristianini 2004: 27, 33.) The parameter vector $\vec{\mathbf{p}}$ is called the projection of the sample $\vec{\mathbf{x}}$ to the feature space. The problem of finding the optimal prediction function for a linear relation is a well-studied problem that can be solved with quadratic programming or with least squares approximation (Suykens & Vandewalle 1999; Shawe-Taylor & Cristianini 2004: 29). Therefore, the task of building a classification method using SVC is essentially a problem of selecting, or building, the model $\mathcal{M}$, whose parameters have a linear relation with the sample's class. We will later see that the parameters $\vec{\mathbf{p}}$ do not have to be calculated explicitly; we only need the inner products of pairs of parameters, i.e. $\vec{\mathbf{p}}_i \cdot \vec{\mathbf{p}}_j = \vec{\mathbf{p}}_i^T \vec{\mathbf{p}}_j = \langle \vec{\mathbf{p}}_i, \vec{\mathbf{p}}_j \rangle$. Therefore, the model may have infinite number of parameters, provided that the inner products, which are scalars, can be calculated.

Suppose for now that an appropriate model has been found. Then function

$$g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}) = \vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}) + w_0 = \vec{\mathbf{w}}^T \phi(\vec{\mathbf{x}}) + w_0 \tag{35}$$

with proper vector $\vec{\mathbf{w}}$ and value w_0 presents that relation, and $\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}))$, so that $\operatorname{sign}(0) = 1$, produces optimal predictions for the samples. Variable w_0 may also be denoted as variable b. (Theodoridis & Koutroumbas 2006: 93–103; Shawe-Taylor & Cristianini 2004: 213; Chang & Lin 2001: 4.) Scalar w_0 is called the bias of the classifier (Boser, Guyon & Vapnik 1992). The prediction function $\operatorname{sign}(g_{\vec{\mathbf{w}}})$ can be visualized by a hyperplane in the feature space (Fig. 17), where for every point $\vec{\mathbf{p}}$ on the hyperplane,

$$\vec{\mathbf{w}} \cdot \vec{\mathbf{p}} + w_0 = 0. \tag{36}$$

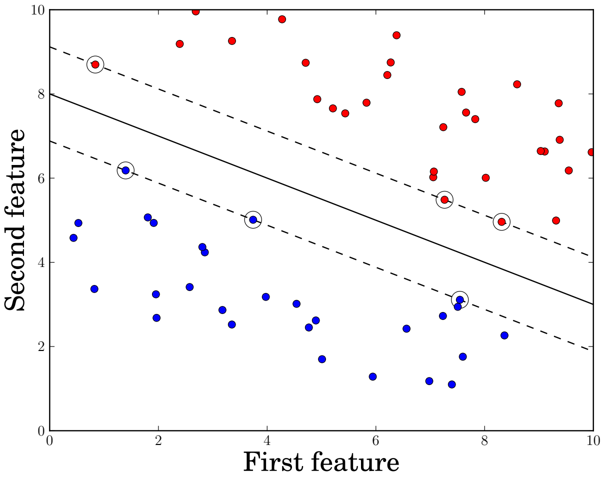


Figure 17. A hyperplane, in two dimensions, that visualizes a linear prediction function in feature space. The projections of the samples to the feature space $\phi(\vec{x})$ are presented as dots, so that the dot's color indicates the sample's class. The solid black line is the decision function's hyperplane, the dashed black lines are the hyperplanes that are defined by the support vectors, and the circled samples are the support vectors.

The vector $\vec{w}$ is the normal vector of the hyperplane. It points towards the class whose label is positive (unit). (Theodoridis & Koutroumbas 2006: 93–94.)

The location and orientation of the hyperplane are selected so that the hyperplane splits the samples' projections into the two classes, and that the hyperplane is as far away from the samples' projections as possible. More precisely, the hyperplane is selected so that the minimum distance between a projection of a sample of a given class and the hyperplane is maximized and that this distance is the same for both classes. When the hyperplane can be selected so that the samples' projections are perfectly separated into two classes, the classes are said to be separable, and the samples whose projections are closest to the hyperplane are called support vectors. (Theodoridis & Koutroumbas 2006: 93–103.)

The problem of finding the optimal location and orientation for the hyperplane is a constrained optimization problem (Theodoridis & Koutroumbas 2006: 100), and it can be solved with its Lagrangian (Boyd & Vandenberghe 2004: 215–216). Solving the Lagrangian produces the optimal Lagrange multipliers $\vec{\lambda}^\star$ and $\vec{\nu}^\star$ (Boyd & Vandenberghe

2004: 223). The normal vector of the hyperplane $\vec{\mathbf{w}}$ can be presented as a sum of products of the samples' projections, their correct labels, and the corresponding optimal Lagrange multipliers. Thus, the prediction function can be written as

$$\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})) = \operatorname{sign}\left(\vec{\mathbf{w}}^{\star} \cdot \phi(\vec{\mathbf{x}}) + w_0\right) = \operatorname{sign}\left(\left(\sum_{i=1}^{n} \lambda_i^{\star} y_i \phi(\vec{\mathbf{x}}_i)\right) \cdot \phi(\vec{\mathbf{x}}) + w_0\right) \quad (37)$$

(Boser et al 1992). Because of the inner product's property called linearity, the prediction function can be further rewritten as

$$\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})) = \operatorname{sign}\left(\sum_{i=1}^{n} \lambda_i^{\star} y_i (\phi(\vec{\mathbf{x}}_i) \cdot \phi(\vec{\mathbf{x}})) + w_0\right). \quad (38)$$

(Boser et al 1992; Greenberg 1998: 434.)

A function which produces the inner product of the samples' projections to the feature space is called a kernel and is denoted as κ or K. In other words,

$$\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \phi(\vec{\mathbf{x}}_i) \cdot \phi(\vec{\mathbf{x}}_j) = \langle \phi(\vec{\mathbf{x}}_i), \phi(\vec{\mathbf{x}}_j) \rangle. \quad (39)$$

When the kernel is used in the definition of the prediction function we do not necessarily need to calculate explicitly the samples' projections to the feature space – we only need to be able to the calculate the value of the kernel function, which implicitly contains the projections. (Boser et al 1992; Shawe-Taylor & Cristianini 2004: 34.) In fact, we do not even have to know explicitly what the feature map $\phi(\cdot)$ is. With the kernel the prediction function can be written as

$$\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})) = \operatorname{sign}\left(\sum_{i=1}^{n} \lambda_i^{\star} y_i \kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}) + w_0\right). \quad (40)$$

One commonly used kernel is the polynomial kernel, or

$$\kappa_d(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j + R)^d = \sum_{s=0}^{d} \binom{d}{s} R^{d-s} (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j)^s. \quad (41)$$

For a given kernel κ_1, the derived polynomial kernel is $\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = p(\kappa_1(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j))$, where $p(\cdot)$ is any polynomial with positive coefficients. Another commonly used kernel is the Gaussian kernel, which is defined as

$$\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp\left(-\frac{\|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\|^2}{2\sigma^2}\right), \quad \sigma > 0. \quad (42)$$

Table 5. The sets into which samples of the validation set are divided in order to evaluate the performance of a binary classifier, and values that are calculated from the sizes of these sets. A matrix of the sizes of these sets, organised as presented here, is called a confusion matrix. (Fawcett 2006.)

	Carious sample	Healthy sample	
Classified as carious	True positive (TP)	False positive (FP)	→ PPV
Classified as healthy	False negative (FN)	True negative (TN)	→ NPV
	↓	↓	
	Sensitivity	Specificity	

(Shawe-Taylor & Cristianini 2004: 292, 296.) Radial Basis Function (RBF) -kernel is defined as $\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp\left(-\gamma|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j|^2\right)$ (Chang & Lin 2001: 34). It is equal to the Gaussian kernel when $\gamma = 1/(2\sigma^2)$.

Beside the quadratic programming approach outlined here, the problem of selecting the optimal hyperplane for the decision function can also be solved by using least squares approximation. This leads to a variant of SVM called least squares SVM or LS-SVM. (See, for example, Suykens & Vandewalle 1999, or Shawe-Taylor & Cristianini 2004: 27–32.) More detailed description of support vector machine is presented in appendix 1.

3.4. Classification performance measures

In a binary classification task there are two options for the sample's class. In this project we can call the class of carious samples the signal present class and the class of healthy samples the no signal present class. The performance of a binary classifier is evaluated by dividing the samples of the validation set into four sets based on their correct class and their predicted class. The names of these sets are obtained by combining the following two terms. The samples which were classified correctly are called true cases and incorrectly classified cases are called false cases. The samples whose prediction is (membership in) the signal present class are called positive cases, and samples whose prediction was the no signal class are called negative cases. (Fawcett 2006.)

A 2×2 matrix of the sizes of these sets is called the confusion matrix. It depicts the performance of the classifier. The cell $(1,1)$ of the confusion matrix contains the size of the true positive set, cell $(1,2)$ contains the size of the false positive set, cell $(2,1)$ contains the size of the false negative set, and cell $(2,2)$ contains the size of the true negative set. Let the abbreviations TP, FP, FN and TN denote these values. These four values are used to calculate five new values, which describe different aspects of the performance of the classifier (Tab. 5). Sensitivity is the probability that a carious sample (signal present) is classified as such, i.e.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{43}$$

Sensitivity is also called hit rate, true positive rate, and recall. Specificity is the probability that a healthy sample (no signal present) is classified as such, i.e.

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{44}$$

Value $1 -$ specificity is called false positive rate or false alarm rate. Positive predictive value (PPV) is the probability that a sample which is classified as carious (signal present) is correctly classified, i.e.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{45}$$

PPV is also called precision. Negative predictive value (NPV) is the probability that a sample which is classified as healthy (no signal) is correctly classified, i.e.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}. \tag{46}$$

Accuracy is the probability that a sample is classified correctly, i.e.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{47}$$

(Fawcett 2006; Bradley 1997.)

Validation produces a confusion matrix for each prospective prediction function, and for each pair of a training set and a validation set. Since each sample in a validation set belongs to exactly one category in the matrix, the sum of the (four) values in the matrix equals the number of samples in the validation set. A confusion matrix, or the five values calculated from it, can be interpreted as a (point) vector which describes the classification performance. By varying the value of one of the classifier's parameters over some range and calculating a confusion matrix for each of those values, we obtain a set of different classification performances, i.e. different confusion matrices, that can be reached with that classification technique.

The performance of a binary classifier is often analysed this way by plotting the so-called receiver operating characteristic (ROC) -curve. It is compiled by resolving the optimal sensitivity-value for each specificity-value, and by plotting the sensitivity-value as a function of value $1 -$ specificity, i.e. as a function of false positive rate. (Duda et al 2001: 48–50.) Points on the ROC-curve are called operating points (Bradley 1997). The ideal operating point is the upper left corner of the ROC-graph. Because the number of samples in the validation set is finite, the analysis produces only a finite number of points on the ROC-curve. (Fawcett 2006.) The maximum number of points in the ROC-curve is equal to the number of samples in the validation set. Other points in the ROC-curve have to be interpolated. The ROC-curve describes the available trade-offs between classifying all healthy samples correctly and classifying all carious samples correctly (Fawcett 2006), and thus the ROC-curve can be considered as a Pareto-front. Trivially the ROC-curve always contains a point where the sensitivity is unit and specificity is zero, and a point where the sensitivity is zero and specificity is unit. These points are obtained by predicting the same class for all samples.

More generally, an analysis which produces a two-dimensional scatter plot where the y-axis is the sensitivity-value and the x-axis is the value $1-$specificity, is called an operating characteristic -curve. The points on the plot are obtained by varying the value of *some* single parameter of the classifier, and calculating the sensitivity and specificity for each of the values. Usually, the parameter whose value is varied is such that at one of its extreme values only a negligible number of samples are classified as positive, and at the other extreme value only a negligible number of samples are classified as negative. (Duda et al 2001: 50–51.) Points on an operating characteristic -curve are also called operating points (Bradley 1997). If the parameter whose value is being varied is the decision threshold, the resulting operating characteristic -curve can be considered as a ROC-curve (Duda et al 2001: 49; Fawcett 2006). In the case of SVM classification, varying the decision threshold corresponds to varying the bias, which may be denoted with variable w_0 or b.

The ROC-curve can be composed by calculating the first point with such decision threshold or bias, that all samples are classified into the same class. Then the decision threshold is changed by a value that is just enough to change the classification of one sample, and a new point on the ROC-curve is calculated. Notice that the classification of several samples may change simultaneously when the decision threshold is changed this way. This is continued, keeping the direction of the changes identical, until all samples are classified into the opposite class. (Fawcett 2006: 866–867.)

As the decision threshold or bias changes monotonically, the sensitivity and the specificity change monotonically as well, because after the decision threshold has reached a point where a sample is classified correctly (or incorrectly), that sample will not later become incorrectly (or correctly) classified (Fawcett 2006). Thus, an ROC-curve is monotonically increasing. However, that does not generally hold for operating characteristic -curves. Also, it is generally impossible to reach an operating point where either sensitivity or specificity is unit by varying a given classification parameter. When the former property does not hold for the parameter being varied, the points have to be sorted according to the specificity before a curve can be interpolated with them. When the latter property does not hold, the range of the parameter can not be resolved by the sensitivity and specificity values.

Suppose that the samples' predictions are given at random, such that the sample is predicted to belong to the signal present class with probability $p_{\text{pred,s}}$ and to the no signal present class with probability $p_{\text{pred,ns}}$. Suppose, further, that the predictions are made for a population of samples where the signal present class has prevalence $p_{\text{prev,s}}$, i.e. this fraction of the population does belong to the signal present class, and where the no signal present class has prevalence $p_{\text{prev,ns}}$. Notice that $p_{\text{pred,s}} + p_{\text{pred,ns}} = p_{\text{prev,s}} + p_{\text{prev,ns}} = 1$. It can be shown by a direct calculation, that in this case sensitivity is $p_{\text{pred,s}}$, specificity is $p_{\text{pred,ns}}$, PPV is $p_{\text{prev,s}}$, NPV is $p_{\text{prev,ns}}$, and accuracy is $p_{\text{pred,s}}p_{\text{prev,s}} + p_{\text{pred,ns}}p_{\text{prev,ns}}$. In this case optimal accuracy is always achieved by predicting all samples as belonging to the class with higher prevalence. This can be shown by viewing the probabilities $p_{\text{pred,s}}$ and $p_{\text{pred,ns}}$ as a weak composition with a length of two for 100 points, where each point represents one percent probability ($p_{\text{pred,s}} + p_{\text{pred,ns}} = 100\%$). Each point contributes one percent of the prevalence of the class into which it is assigned into the accuracy. The optimal accuracy is obtained when each point contributes a maximal value, which happens when all of them are assigned to the class with higher prevalence. For random predictions, varying the decision threshold corresponds to varying the prediction probabilities of the two classes. Because sensitivity and specificity are equal to these probabilities, the ROC-curve for random predictions is always a diagonal line from point $(0, 0)$ to point $(1, 1)$.

Usually the ROC-curve is concave and lies entirely above the diagonal line, apart from the end points (Duda et al 2001: 50). The area between the ROC-curve and the diagonal line describes how much information about the pattern the classifier is using when making the predictions (cf. Duda et al 2001: 48–49). This area multiplied by two is called the Gini coefficient (Fawcett 2006). The ROC-curve may also lie beneath the diagonal line if the classifier is not making random predictions, and the pattern that was found in the training set by pattern analysis is inverted in the validation set. "Any classifier that produces a point [below the diagonal line] can be negated to produce a point [above the diagonal line]" (Fawcett 2006: 863). This can be interpreted such, that "a classifier below the diagonal [has] useful information, but it is applying the information incorrectly" (Fawcett 2006: 863). However, the very fact that the pattern seems to be inverted between the training set and the validation set suggests that this pattern, inverted or not, does not necessarily hold for all samples from the same source.

If we vary the values of $k > 1$ classification parameters, we obtain a k-dimensional surface of the classification performances. Points on the surface can be called extended operating points. The surface has to be interpolated from a finite number of points, as with the operating curves. We must also consider for each classification parameter the properties that we considered earlier for the operating curve. The SVM training procedure which was partly described earlier selects a point on such a surface, so that the classification accuracy and generalization are maximized. However, sometimes we are willing to accept certain kinds of classification errors, e.g. false positives, in order to make the probability of another kind of classification error, e.g. false negative, smaller. For example, if a severe disease can be treated with an inexpensive method with few side effects, then we may want to treat a patient for that disease just in case, even if we are not totally certain that the patient indeed has that disease. Vaccinations can be considered as an example of this principle. Conversely, we want to spare the patient from a treatment with major side effects until we are certain that the patient will benefit from the treatment despite the side effects.

In the Neyman-Pearson (NP) paradigm the decision threshold or bias is selected such that the specificity is given a lower limit which it must achieve or exceed, and the sensitivity is maximized with that constraint, or vice versa – the specificity might be allowed to fall a bit short of the requirement, though. Such an operating point is readily obtained from the ROC-curve. (Scott 2007; Bradley 1997.) If we can assign a cost C_{FP} for incorrectly classifying a sample as carious (false positive) and a cost C_{FN} for incorrectly classifying a sample as healthy (false negative), we can easily calculate the risk of selecting a given value for the bias w_0 (see section 3.1, page 45). The risk or expected cost is

$$R = C_{\mathrm{FP}}(1 - \mathrm{Specificity}(w_0)) + C_{\mathrm{FN}}(1 - \mathrm{Sensitivity}(w_0)). \tag{48}$$

An optimal choice for the bias would be a value which minimises the risk. (cf. Duda et al 2001: 24–26; cf. Bradley 1997.) It is in general different than the value chosen by SVM, which produces the best accuracy.

Points in the ROC-graph which have equal expected cost form an iso-performance line. The slope of that line is

$$s = \frac{C_{\mathrm{FP}} p_{\mathrm{prev,ns}}}{C_{\mathrm{FN}} p_{\mathrm{prev,s}}} = \frac{\mathrm{Sensitivity}_2 - \mathrm{Sensitivity}_1}{\mathrm{FPR}_2 - \mathrm{FPR}_1}, \tag{49}$$

where FPR is the false positive rate, and $(\mathrm{Sensitivity}_1, \mathrm{FPR}_1)$ and $(\mathrm{Sensitivity}_2, \mathrm{FPR}_2)$ are two points on the line. The optimal operating point is the point which is as close to the upper left corner of the ROC-graph as possible, and whose tangent is the iso-performance line of minimal expected cost. (Fawcett 2006.) In other words, the operating points can be sorted according to their expected cost by projecting them onto the vector $(1, s)$ rotated by 90 degrees (counterclockwise), where s is the slope of the iso-performance lines. The operating point whose projection has the greatest value has the minimal expected cost.

An overall estimate of the classification performance with various values of decision threshold or bias is obtained by calculating the area under the ROC-curve, i.e. the area between the ROC-curve and the x-axis. This value is known as the area under curve (AUC). It provides a good way to evaluate the classification performance with a single value. This value is independent of the decision threshold or bias, the class prevalences, and the costs C_{FP} and C_{FN}. It also describes how much information about the sample's class the classifier is able to the extract from a sample. (Bradley 1997; Fawcett 2006.) In contrast, the accuracy of a classifier which predicts all samples as belonging to the class with higher prevalence is equal to the prevalence, although such a classifier does not extract any information from an individual sample. The Gini coefficient can be calculated from AUC with Gini $+ 1 = 2 \times$ AUC (Fawcett 2006).

3.5. Cross-validation

In this project we will use cross-validation for evaluating the accuracy of the prediction function. In cross-validation several different pairs of training and validation sets are created by partitioning the samples that were collected for calibration, and the corresponding values of the dependent variable, in different ways. The accuracy of the resulting prediction function is evaluated in each of these cases, and the average of the accuracies is used as the true accuracy of the prediction function for such samples. In k-fold cross-validation the samples are divided into k subsets, such that one of the subsets is used as the validation set, while the other subsets form the training set. When each subset has been used as the validation set once, the average of the resulting performances is calculated. In leave-one-out cross-validation (LOOCV) a single sample is used as the validation set, while the other samples form the training set. The average of the performances is calculated after each sample has been used as the validation set once. This corresponds to the k-fold cross-validation when k equals the number of samples. Leave-one-out cross-validation is also known as the jackknife method. This name reflects the impression that this method is useful in many ways. (Duda et al 2001: 472, 483–486.)

3.6. Summary

When the spectroscopic measurements have been made, the resulting data is analysed in order to find a method which can detect which measurements are made from caries lesions, i.e. to be able to tell from a given measurement whether it depicts a site of healthy enamel or a site of carious enamel. In spectroscopy, development of such method is called calibration. A piece of information can be used in the detection method if, and only if, it is available every time that detection method is used. Such pieces of information are called features. The problem of finding the detection method with the available features is a task for a pattern recognition algorithm. The detection method is searched for by using a subset of the available samples, which is called a training set. The performance of the resulting detection method is evaluated by using another subset of the available samples, which is called a validation set.

In this project we will use cross-validation for evaluating the accuracy of the detection method. In cross-validation several different pairs of training and validation set are created by partitioning the available samples in different ways. The accuracy of the detection method is evaluated in each of these cases, and the average of the accuracies is used as the true accuracy of the detection method.

Usually, some kind of signal processing is performed on the samples' features before the pattern recognition algorithm is used. These processing steps are called preprocessing. Their goal is to improve the classification results by making the pattern easier to identify. In order to avoid the problems caused by having too many features, the number of features may be reduced by preprocessing, such that the parts of the information that are most useful for the prediction are preserved. Reduction of the number of features is called dimensionality reduction.

In complete darkness the measurement result of an optical device like a spectrometer should be zero. However, this is not necessarily the case. The measurement result obtained from the spectrometer in darkness is called the dark current of the spectrometer.

The numerical range in which the measurement results are reported depends on the spectrometer. Thus, a given sample may produce different numerical values for the features, or the spectra, in different spectrometers. The samples are transformed into a fixed numerical range, typically range $[0, 1]$ by removing the spectra of a black reference sample from each sample, and by dividing the samples with the spectra of a white reference sample.

Savitzky-Golay is a method that can be used for smoothing and derivating a spectra. It fits a polynomial of given degree to the spectra by least-squares techniques in a moving window of given odd width. Decimation is a simple method to reduce the number of features in the samples. In decimation only every K^{th} feature in the samples is kept, and the rest are removed. Decimation may cause aliasing, which would impair the quality of the samples. Thus, the samples need to be filtered with a decimation filter, which is a low-pass filter whose cutoff frequency is π/K.

Principal component analysis (PCA) is a dimension reduction method that seeks to produce a model of the samples, such that the sum of squared differences between the actual (original) samples and the samples depicted by the model is minimal, when the model can use at most a given number of variables to depict the samples.

Support vector machine (SVM) is a pattern recognition algorithm. It uses a specific kind of matrix, called a kernel, to calculate the inner products of pairs of the samples' projections to a higher-dimensional space, and then searches for a hyperplane that would separate the samples in that space. When binary classification is done with SVM, the prediction is

$$\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})) = \operatorname{sign}\left(\sum_{i=1}^{n} \lambda_i^{\star} y_i \kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}) + w_0\right), \tag{50}$$

where $\vec{\lambda}^{\star}$ is the dual variable of the Lagrangian of the problem, $y_i \in \{1, -1\}$ is the correct label for sample $\vec{\mathbf{x}}_i$ of the training set, κ is the kernel function, w_0 is the bias of the classifier, and $\vec{\mathbf{x}}$ is the sample for which the prediction is calculated.

Confusion matrix is a 2×2 matrix that describes the performance of a binary classifier. From a confusion matrix we can calculate, for example, sensitivity, i.e. the probability that a carious sample is classified correctly, and specificity, i.e. the probability that a healthy sample is classified correctly. The performance of a binary classifier is often analysed by plotting the so-called receiver operating characteristic (ROC) -curve. It is compiled by resolving the optimal sensitivity-value for each specificity-value, and by plotting the sensitivity-value as a function of value $1 -$ specificity, i.e. as a function of false positive rate. An overall estimate of the classification performance is obtained by calculating the area under the ROC-curve, i.e. the area between the ROC-curve and the x-axis. This value is known as the area under curve (AUC).

4. MATERIALS AND METHODS

The research hypothesis of this study was that dental caries lesions could be detected with reflectance NIR-spectroscopy. The scattering coefficient of a dental caries lesion in enamel is greater than that of healthy enamel, i.e. a caries lesion scatters more light than healthy enamel (Karlsson 2010). At the early stages of development a dental caries lesion appears translucent or white, and at later stages of development it appears brown (see section 2.1.3). Because the brown discoloration, or change in the color of the enamel, can be detected visually, the change occurs in the wavelengths of visible light. Thus, it seems that a dental caries lesion reflects more light than healthy enamel, and once it has reached a late enough stage of development, it begins to absorb light in the visible range.

Stains or developmental defects in the enamel may also cause changes in the enamel's color, though. Stains tend to make the enamel darker in color. However, "stains are not visible in the NIR" (Wu & Fried 2009: 212). Therefore, wavelengths in the near-infrared range may be able to differentiate caries lesions from stains more accurately than wavelengths in the visible range. Accordingly, our hypothesis was that the increased scattering in the near-infrared range is the best indication of a dental caries lesion.

This study was limited to natural caries lesions on smooth surfaces of extracted tooth. Caries lesions on the biting surface are not studied. This limitation is made because the smooth surfaces are easier to measure spectroscopically than the irregular and grooved biting surfaces. Furthermore, once caries can be diagnosed spectroscopically on the smooth surfaces, it is easier to attribute spectroscopic observations made on the biting surfaces either to caries or to surface irregularities.

In this chapter we will first look at the samples that were used in this study and the way they were measured spectroscopically. Then the analysis of the measurement results is described. The end of the chapter presents notes on the implementation of the analysis.

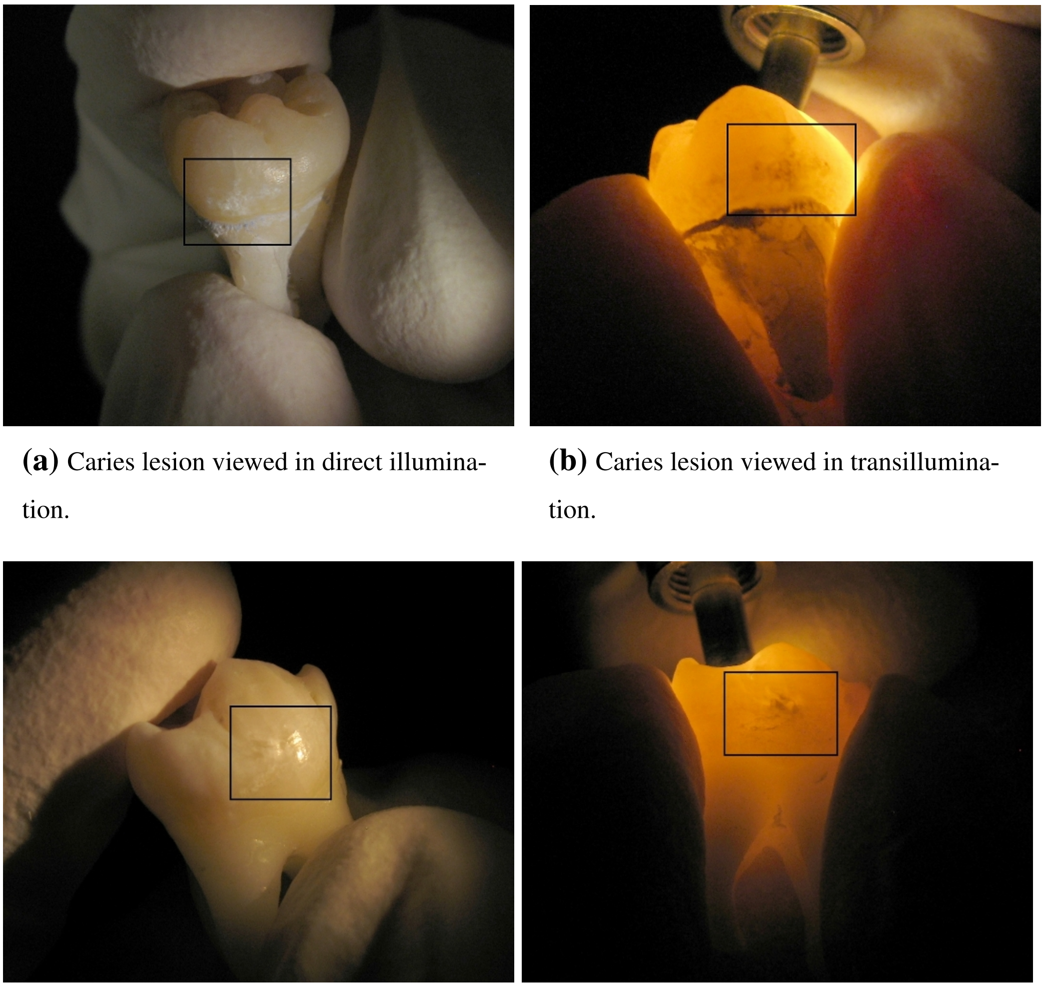
4.1. Samples

In total 28 extracted human teeth were obtained from the dental services of the City of Vaasa. Spectroscopic measurements where taken from 24 of them. The teeth were given an alphanumeric identifier and stored in an improvised container, immersed in denatured alcohol in order to disinfect them and to keep them hydrated. Before inspection and measurements the teeth were gently dried with a cue tip. The teeth were inspected by the author with fiber-optic illumination in order to detect healthy areas of enamel and areas of enamel that contained caries lesions.

The areas of healthy enamel and the caries lesions were detected by first illuminating the surface of the tooth from the same side as it was observed from with illuminating fiber optic (TEQS Hard Cladding Multimode Fiber FT600EMT from ThorLabs Inc., Newton, NJ, USA) held at some distance from the surface, i.e. in direct light. Then the illuminating fiber optic was placed in contact with the tooth surface and moved some distance away from the inspected area in order to view the inspected area using the light scattered outwards from the tooth, i.e. in backlight. The objective was to detect white-spot caries lesions at various levels of development. More advanced caries lesions were easy to detect without a need for fiber optic illumination. An area was diagnosed as a caries lesion if it either appeared translucent, indicating very early stage caries, or if it appeared whiter than the surrounding enamel in direct illumination and cast a shadow that extends below the surface of the tooth in backlight (Fig. 18). Most of the areas that were diagnosed as caries lesions were diagnosed as an early-stage white-spot caries lesion or a white-spot caries lesion.

This diagnostic method was introduced to the author by acting chief dental officer, Dr. Katri Palo. During this demonstration the author made notes about the diagnoses provided by Dr. Palo. However, it was not possible to make complete, detailed notes about the diagnoses at the time, which is why the author needed to repeat the diagnoses at the time of the measurements. After the measurements had been made, the diagnoses made at that time where compared with the notes about the diagnoses provided by Dr. Palo. All measurements taken from three tooth samples where discarded, because the diagnoses made by the author for those samples during the measurements were considered to be too different from the diagnoses provided by Dr. Palo.

Analysis was performed using in total 111 measurements sets taken from 21 tooth sam-



(a) Caries lesion viewed in direct illumination.

(b) Caries lesion viewed in transillumination.

(c) Caries lesion viewed in direct illumination.

(d) Caries lesion viewed in transillumination.

Figure 18. The teeth were inspected by the author with fiber-optic illumination in order to detect healthy areas of enamel and areas of enamel that contained caries lesions. An area was diagnosed as a caries lesion if it either appeared translucent, indicating very early stage caries, or if it appeared whiter than the surrounding enamel in direct illumination and cast a shadow that extends below the surface of the tooth in backlight.

ples. Each set contained 100 individual measurements taken from the same point. The measurement sets contained 69 sets from healthy sites of enamel and 55 sets from carious sites of enamel.

4.2. Measurements

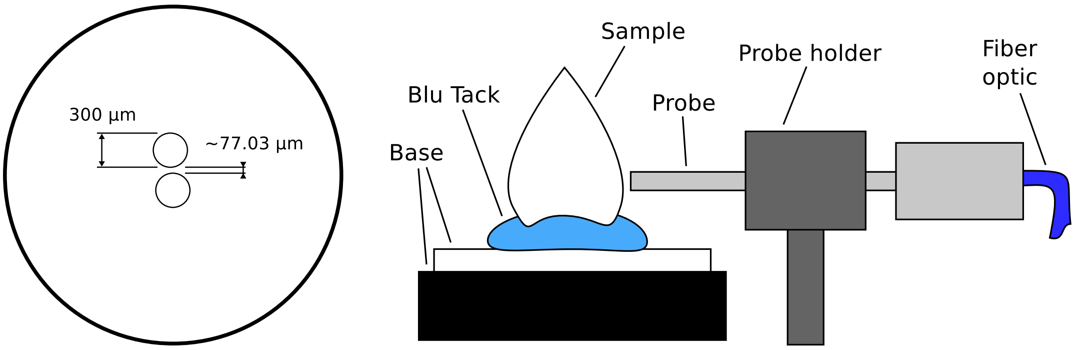
The construction of a custom probe for this project was deemed unfeasible. Thus, the measurement setup had to be constructed using the equipment that was readily available in our laboratory. An emphasis was placed on keeping the measurement geometry as fixed as possible for all samples. Another criterion was that the probe design should resemble a design which could be implemented and used for *in vivo* measurements.

The probe that was selected for the measurements is a general purpose transmission dip probe model T300-RT-VIS/NIR (Ocean Optics Inc., Dunedin, FL, USA). It is designed for measuring the transmission spectra of liquid samples. It contains two 300 μm optical fibers that can transfer light at wavelength range 400–2500 nm. One of the fibers is connected to a light source, while the other is connected to a spectrometer. The fibers are housed in a stainless steel assembly with a diameter of 3.175 mm. For measuring liquid samples, the assembly is placed in a ferrule with a diameter of 6.35 mm, and a measurement tip is attached at the end of the ferrule with screw threads. The ferrule contains a lens which focuses the two fibers of the assembly to the same focal point. The measurement tip contains a flat mirror. When measuring a liquid sample, light enters the sample from the illuminating fiber and is focused to a focal point by the lens in the ferrule. After traversing the sample the light is reflected back by the mirror, focused to the second fiber optic by the ferrule's lens, and then transmitted to the spectrometer. (Ocean Optics Inc. 2011.) The probe was used without the measurement tip or ferrule in place.

The distance between the two optical fibers was measured by taking two photographs of the probe head, so that on both pictures one of the fibers was coupled to a light source, and by then combining the two pictures. The pictures were taken with Fujifilm IS Pro -camera (FUJIFILM Corporation, Tokyo, Japan) with 1/200 seconds exposure time, f/18 aperture, and ISO 800 film speed. The diameter of the probe head is 3.174 mm in real life and 270 pixels in the picture. Based on this, one pixel corresponds to 3.175/270 mm = 0.011759259 mm, and one millimeter corresponds to 270/3.175 px = 85.039370079 px. The diameters of the fiber optic heads are 300 μm in real life, which corresponds to 25.512 pixels in the picture's scale. The pictures of the fiber optic heads were thresholded with such pixel values that the sizes of the fiber optic heads seemed to correspond to the real life size. The picture of the upper fiber optic was thresholded with value 185 and the picture of the lower fiber optic was thresholded with value 240. The sizes of the fiber optic heads ended up being 25–26 pixels, which corresponds to 293.98–305.74 μm in real life. The locations of the fiber optic heads were estimated by fitting a circle with a radius of 26 pixels to the picture and recording the coordinates of the circle's center. The magnitude of the difference vector between these two points was 32.06 pixels, which corresponds to 377.03 μm. The margin of error of the location of each circle can be set at one pixel. The margin of error of the distance between the centers is thus two pixels, which corresponds to 23.518 μm. Since the radius of the fiber optics was 150 μm, this suggests that the gap between the two fiber optics is 77.03 μm $\pm$ 23.518 μm (Fig. 19a).

This probe represents a measurement geometry where the location of the illumination source and the location of the measurement point are fixed relative to each other, and their locations relative to the sample are defined by the probe. Such a design can be implemented for *in vivo* measurements. It also keeps the measurement geometry constant, provided that the contact between the probe and the sample is similar for all samples. This is facilitated by the fact that the probe can easily be clamped to an optical bench built using components that were available in our laboratory.

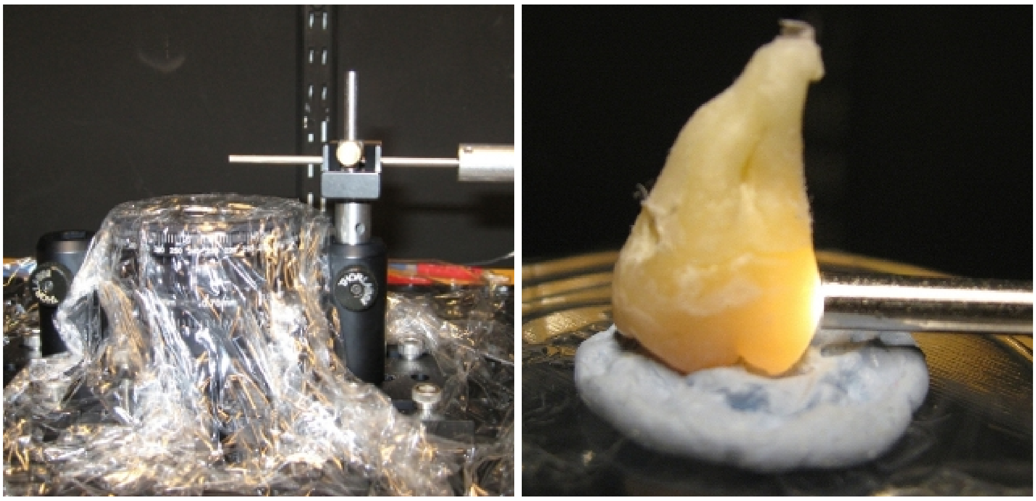
A diagram of the measurement setup is presented in Figure 19b. An optical bench (Fig. 20a) was constructed for making measurements. The components for the bench (Thorlabs Inc., Newton, New Jersey, USA) were available in our laboratory. The bench was built on a breadboard, i.e. on a metal plate with screw holes. The bench contained a table for the sample. The table's height could be slightly adjusted, although this feature was not used during the measurements. The bench also featured two post holders on opposite sides of the table. Either of them could hold a post which had a clamp on top of it, which in turn held the probe. The posts could be adjusted vertically and rotated around



(a) The geometry of the measurement probe head.

(b) Diagram of the measurement setup.

Figure 19. Diagrams of the measurement setup.



(a) The optical bench for making the measurements.

(b) An example of a contact between a sample and the probe.

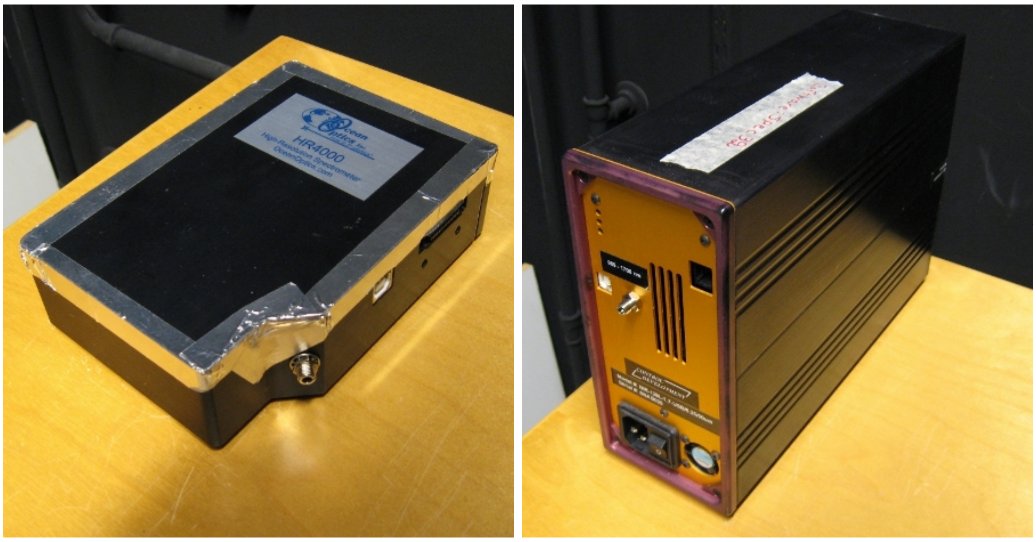
Figure 20. Photos of the measurement setup.

their longitudinal axis. This was used for positioning the probe and sample in contact with each other. The bench's table was covered with plastic wrapping as a measure of infection control.

The table was a little too low for obtaining a good contact between the sample and the probe. This was solved by building a platform for the sample out of the base of a plastic cup. A piece of Blu Tack® was used to adhere the sample to the platform, as well as to allow the sample to be orientated such, that the surface of the site to be measured was perpendicular to the tip of the measurement probe (Fig. 20b). All measurements were done in a dark environment, more precisely in a relatively small windowless room, where the walls were painted black, the room lights turned off and the door closed. Some ambient light was still present, however.

Two spectrometers were used with each sample. The first spectrometer (Fig. 21a) used was HR4000 (Ocean Optics Inc., Dunedin, FL, USA). It is a grating spectrometer that measures light intensity at wavelength range 200–1100 nm with a 3648-element silicon CCD array. The spectrometer was used with SpectraSuite software on Microsoft Windows XP. (Ocean Optics Inc. 2008.) The second spectrometer (Fig. 21b) used was SNAB035 (Control Development Inc., South Bend, IN, USA). A label in front of the spectrometer reports the model as NIR-128L-1.7-USB/6.25/50um, which does not correspond to any model in the manufacturer's current list of NIR spectrometers. However, the label can be interpreted to indicate that the spectrometer's model is either NIR128L-1.7TS or NIR128L-1.7T1 by interpreting 6.25 as the linear dispersion in nanometers per pixel and 50 as the slit width in micrometers. Both of these models seem to be grating spectrometers, since properties of the grating are reported for both models. The SNAB035 spectrometer measures light intensity at wavelength range 909–1706 nm with a 128-element InGaAs detector. The spectrometer was used with CDI Spec32 software on Microsoft Windows XP. (Control Development Inc. 2011.)

Preprocessing requires that the spectrometer's dark current, also known as the black reference, and a white reference, also known as a reference sample, are measured. They are required for calibrating the spectroscope (see chapter 3). The dark current was mea-



(a) The HR4000 spectrometer.

(b) The SNAB035 spectrometer.

Figure 21. The spectrometers used for making the measurements.

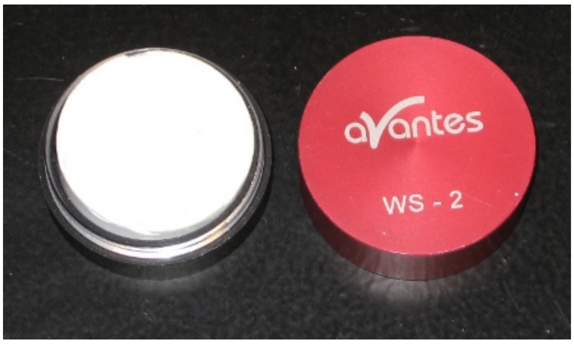


Figure 22. The ceramic disc used as the (white) reference sample.

sured with the normal measurement setup, except that there was no sample in front of the probe. A white reference tile WS-2 (Avantes Inc., Eerbeek, The Netherlands) was measured as the white reference (Fig. 22). It reflects approximately 98% of incident light at wavelengths 350–1800 nm (Avantes Inc. 2009: 135).

The integration time was set to 20 ms on both spectrometers. This time was selected by measuring the white reference with the HR4000 spectrometer, and leaving a margin between the maximum intensity in the spectra of the white reference and the level at which the spectrometer saturates. The margin was left so that a spectra could be measured even if the sample produced a higher intensity than the white reference, for example due to autofluorescence. Later, it became obvious that the integration time should have been longer for the SNAB035 spectrometer. Measurements that where made with the HR4000 spectrometer produced consistently better results, and thus only those measurements were used in more extensive analysis.

All spectra were stored as raw, unprocessed spectra in ASCII format to allow as flexible analysis as possible. The author wrote a Python software to parse the spectra from the ASCII files, and to perform analysis on them. Analysis of the samples is discussed in the following sections.

4.3. Common preprocessing

Several different classification methods were used during this study. Different methods used different kinds of preprocessing methods. This section describes the preprocessing methods that were common to all classification methods. The description of each of the classification methods includes a description of the additional preprocessing methods that were used with that method.

Each sample was measured 100 times, such that each of these spectra were measured immediately after one another, without disturbing the measurement set up. These spectra were averaged out such that the reported intensity at wavelength λ was the average of the intensities at that wavelength over the spectra, i.e.

$$I''(\nu) = \frac{1}{K}\sum_{i=1}^{K} I_i(\nu)\Bigg|_{K=100}, \qquad \nu = \frac{c}{\lambda}. \tag{51}$$

The resulting spectra was used as a single sample in subsequent processing. The performance of the noise cancellation can be evaluated by subtracting the spectra after noise cancellation from the original spectra. The difference should approximate the noise of the original spectra. The histogram of the difference depicts the probability distribution of the information that was removed from the samples by the noise cancellation procedure.

In this project the dark current of the spectrometers $(I_b(\nu))$ was measured with the same set up as the samples, in dark environment, except that there was no sample to measure (see page 75). The dark current or black reference was measured 100 times, and the measurement results were averaged. The dark current was removed from the samples by subtracting it from each sample, i.e.

$$I'(\nu) = I''(\nu) - I_b(\nu). \tag{52}$$

The white reference sample $(I_w(\nu))$ was also measured 100 times and averaged. The sample spectra were then normalized by the equation

$$I(\nu) = \frac{I'(\nu)}{I_w(\nu) - I_b(\nu)}. \tag{53}$$

If this normalization resulted in division by zero the intensity was given value zero, even if the numerator was not zero, e.g. due to autofluorescence.

Samples which contained a feature whose value was smaller than zero or greater than the unit were removed as outliers. This resulted in two samples being removed, because both contained one or more features with values that were greater than the unit. Features that corresponded to wavelengths below 420 nm or above 1000 nm were removed in order to remove the features that seemed to contain more noise than signal due to poor illumination at these wavelengths.

4.4. Classification with intensity thresholds

A simple classifier was constructed to test our hypothesis. The classifier used a number of rules, so that every rule had the following format: if the sample's normalised intensity at wavelength λ was greater than (or smaller than) threshold t, the sample was classified as carious; otherwise, the sample was classified as healthy. If, and only if, one or more of the rules classified the sample as carious, the sample was classified as carious. Before using this method the samples were first smoothed by using the Savitzky-Golay method with window length $L = 61$ and degree $d = 6$.

The classifier was first trained with a training set of samples. During the training the classifier first used exhaustive search to select the rule which gave the best classification accuracy on the training set. The range of available wavelengths was divided into a number of equally spaced steps, and the wavelength at each of these steps was considered as an option in the search. For each given wavelength, the classifier sorted the samples' intensities at that wavelength and considered the midpoint between each two consecutive samples as a possible threshold. The classifier calculated the classification accuracy on the training set for each possible threshold, using that threshold first as a lower limit for classifying the sample as carious and then using it as the upper limit, and chose the threshold and type of limit that gave the best accuracy with that wavelength. After repeating this for each of the considered wavelengths, the classifier chose the rule which gave the best accuracy.

Then the classifier used this same method to select another rule, so that the new rule gave the best possible accuracy when used together with the previously selected rule(s). This was continued until the maximum allowed number of rules was reached, or until the classifier was unable to find a new rule that would improve the classification accuracy.

This method was used so that the maximum number of available rules was set at five rules, and the available wavelength range was divided into 1000 steps. When this method was used with k-fold cross-validation, each of the k training set produced a set of rules for classifying the samples. After each CV-folder was processed, the average set of rules was constructed from the sets of rules of the training sets, and those rules where used to classify all available samples as a further evaluation of the accuracy that could be achieved with this method. The average set of rules used medians of the rules of the CV-folders. Median was used because it approximates the mean of the values while reducing the effect of outlier values.

4.5. Classification with difference in endpoint intensities

Next an even simpler classification method was tried. In this method the sample was classified as carious if, and only if, the difference of the normalised intensity at the last available wavelength and the normalised intensity at the first available wavelength was greater than a given threshold. Two different methods of selecting the threshold were tried. In the first method the threshold was given a constant value of zero. In the second method the threshold which gave the best classification accuracy within the training set was selected. Notice that if the threshold was zero or greater, the sample was classified as carious if, and only if, the normalised intensity at the last available wavelength was greater than the intensity at the first available wavelength. Before using this method the samples were first smoothed by using the Savitzky-Golay method with window length $L = 61$ and degree $d = 6$.

4.6. Classification with one-class Mahalanobis distance

The third classification method that was used calculated the sample's Mahalanobis distance to the mean spectra of the healthy samples, and classified the sample as carious if the distance was greater than a given threshold. The threshold was selected by finding the threshold which produced the best accuracy in the training set. Before using this method the samples were first smoothed by using the Savitzky-Golay method with window length $L = 61$ and degree $d = 6$.

The Mahalanobis distance for sample $\vec{\mathbf{x}}$ is

$$d = \left[(\vec{\mathbf{x}} - \vec{\mu})^T \Sigma^{-1} (\vec{\mathbf{x}} - \vec{\mu})\right]^{1/2}, \Sigma = \vec{\sigma}^2 I, \tag{54}$$

where $\vec{\mu}$ is the mean spectra of the healthy samples, $\vec{\sigma}^2$ is the variance spectra of the healthy samples, and I is an unit matrix. Thus Σ is a diagonal matrix, such that its element $\sigma_{i,j}$ is the variance of the intensities at the i^{th} feature among the healthy samples if $i = j$, and $\sigma_{i,j} = 0$ if $i \neq j$. When we look at the samples' m-dimensional feature space, the mean spectra of the healthy samples is a point and the spectra that have equal Mahalanobis distance from that point form a hyperellipsoid. (Theodoridis & Koutroumbas 2006: 25–26.)

4.7. Classification with two-class Mahalanobis distance

The fourth classification method that was used calculated the sample's Mahalanobis distance to the mean spectra of the healthy samples and its Mahalanobis distance to the mean spectra of the carious samples, and classified the sample as carious if the difference between the former and the latter was greater than a given threshold. In other words, the sample is classified as carious if it is closer to the mean carious spectra than to the mean healthy spectra by a given margin, in the sense of the Mahalanobis distance. The threshold was selected by finding the threshold which produced the best accuracy in the training set. Before using this method the samples were first smoothed by using the Savitzky-Golay method with window length $L = 61$ and degree $d = 6$.

The difference in the Mahalanobis distances for sample $\vec{\mathbf{x}}$ is $d = d_h - d_c$, where

$$d_h = \left[(\vec{\mathbf{x}} - \vec{\mu_{\mathbf{h}}})^T \Sigma^{-1} (\vec{\mathbf{x}} - \vec{\mu_{\mathbf{h}}})\right]^{1/2}, \Sigma = \vec{\sigma}^2 I, \tag{55}$$

$$d_c = \left[(\vec{\mathbf{x}} - \vec{\mu_{\mathbf{c}}})^T \Sigma^{-1} (\vec{\mathbf{x}} - \vec{\mu_{\mathbf{c}}})\right]^{1/2}. \tag{56}$$

Here, $\vec{\mu_{\mathbf{h}}}$ is the mean spectra of the healthy samples and $\vec{\mu_{\mathbf{c}}}$ is the mean spectra of the carious samples, and $\vec{\sigma}^2$ is the variance spectra of all training samples. (Theodoridis & Koutroumbas 2006: 25–26.)

4.8. Classification with a support vector machine

As a representative of more advanced classification methods, support vector machine was also used to classify the samples. With this method the samples were preprocessed by smoothing and derivating them by using the Savitzky-Golay method with window length $L = 141$ and degree $d = 6$, and by then decimating them, keeping only every eighth feature. Before decimation the samples were filtered with a decimation filter. The coefficients of the decimation filter were calculated with the Hamming window, using window size 61. The number of coefficients was selected empirically by trying different numbers of coefficients, and by selecting a compromise between having a minimal number of coefficients and having optimal results. The performance of the selected decimation filter is presented in Figure 23.

After preprocessing the dimensionality of the samples was reduced with principal component analysis (PCA), using various numbers of principal components in the range 1–20. The classification performance was evaluated for each of them. The samples were then classified into two classes, healthy samples and carious ones, using support vector machine (SVM). A linear kernel, a polynomial kernel, an RBF-kernel, and a sigmoid-kernel where tried with different degrees.

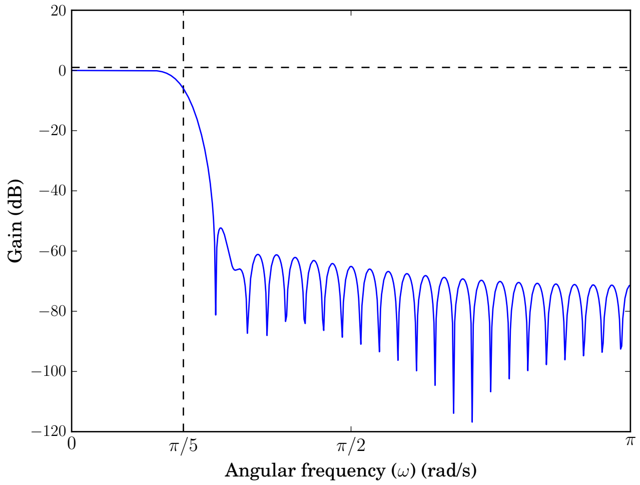


Figure 23. Decimation filter's frequency-response in decibels. The plot is symmetrical with respect to the y-axis. Only the positive half of the plot is shown here. Ideally, the gain would be unit – equal to the horizontal dashed line – until the cutoff frequency, which is represented by the vertical dashed line, after which the ideal gain would be zero.

4.9. Validation

With each method the performance of the classifier was evaluated with 4-fold cross-validation. The samples were assigned to the folders such that each folder contained as balanced a mix of healthy samples and carious samples as possible. If necessary, the last folder had a less balanced mix than the other folders. All CV folders contained 17 measurement samples from healthy sites and 10 measurement samples from carious sites. This accounted for $4 \cdot (17 + 10) = 108$ measurement samples. The last remaining measurement sample was from a healthy site, and for each CV folder it was assigned to the training set.

4.10. Implementation of the analysis

The analysis was implemented in its entirety in `Python` programming language, using several libraries. The main library used is `SciPy` (Jones, Oliphant, Peterson et al 2001–), which in turn depends on a library called `NumPy`. Plots were made with the matplotlib library (Hunter 2007). Together, these libraries provide functionalities that resemble the Matlab program (MathWorks Inc., Natick, MA, USA). Altogether, the implementation contained approximately 18 000 lines of source code. However, a considerable portion, if not most of these lines, contain obsolete or unused source code, or comments describing the implementation details.

The SVM classification was implemented with the scikits.learn library. It provides a Python-API for SVM classification, and uses LibSVM library (Chang & Lin 2001) for implementing the classification. The PCA transformation was implemented with the Modular toolkit for Data Processing (MDP) -library (Zito, Wilbert, Wiskott & Berkes 2008).

The algorithm presented in Fawcett (2006: 866–867) was used as the basis for generating the ROC-curves. The library used for SVM classification did not provide the samples' classification margins in the feature space, i.e. their projections on the normal vector of the decision hyperplane in the feature space. These margins are required for building the ROC-curve. Therefore, the margins were resolved by searching. First, the smallest bias value that was large enough to predict all samples as negative was resolved. Then successively smaller bias values were searched for, so that the new value was as close as possible to the previous value, and at each step the prediction changed for at least one sample in the validation set.

The searching was done with a windowing method. A start value or the previous bias value was used as one edge of the window, and the other edge was moved so that the window included the next bias value that was being searched for, i.e. that the predictions for the validation samples differed on at least one sample at the opposing edges of the window. The absolute value of bias at the window's edge was increased by doubling the value, and it was decreased by halving the value. Then the bias value at the center of the window was evaluated and turned into one of the window's edges, depending on which half of the window contained the sought bias value. This was repeated until the window could not be made any smaller, because the center of the window appeared to be identical to one of the window's edges due to rounding errors in the floating-point operations. The value at the center of the window was calculated by adding the values at the window's edges and by dividing the result by two. On each iteration it was checked that the center value was inside the window. The edge whose value was initially changed was used as the sample's margin, and the size of the window was used as an upper limit of error in the margin value.

4.11. Summary

This study used a total of 24 extracted human teeth, which were obtained from the dental services of the City of Vaasa. Measurements taken from three teeth where discarded because they were suspected of being mislabeled. The teeth were stored immersed in denatured alcohol in order to disinfect them and to keep them hydrated. They were gently dried with a cue tip before being inspected with fiber-optic illumination for healthy areas of enamel and for caries lesions in the enamel.

Diffuse reflectance spectras of the tooth samples were measured with a general purpose transmission dip probe model T300-RT-VIS/NIR (Ocean Optics Inc., Dunedin, FL, USA), using a spectrometer HR4000 (Ocean Optics Inc., Dunedin, FL, USA) in wavelength range 200–1100 nm and a spectrometer SNAB035 (Control Development Inc., South Bend, IN, USA) in wavelength range 909–1706 nm. The integration time was set to 20 ms on both spectrometers. It became obvious that the integration time should have been longer for the SNAB035 spectrometer. Measurements that where made with the HR4000 spectrometer produced consistently better results, and thus only those measurements were used in more extensive analysis.

In total 111 samples where produced, consisting of 69 samples of healthy sites and 42 samples of carious sites. Two samples, both from carious sites, where removed as outliers because both contained one or more features with values that were greater than unit.

The spectra were normalized with the help of a black and white reference, and features that corresponded to wavelengths below 420 nm or above 1000 nm were removed in order to remove the features that seemed to contain more noise than signal due to poor illumination at these wavelengths.

The samples were then classified using several different classifiers with different preprocessing methods. The first classifier used a number of rules, so that every rule had the following format: if the sample's normalised intensity at wavelength λ is greater than (or smaller than) threshold t, the sample is classified as carious; otherwise the sample is classified as healthy. If, and only if, one or more of the rules classified the sample as carious, the sample was classified as carious.

In the second classifier the sample was classified as carious if, and only if, the difference of the normalised intensity at the last available wavelength and the normalised intensity at the first available wavelength was greater than a given threshold.

The third classification method that was used calculated the sample's Mahalanobis distance to the mean spectra of the healthy samples, and classified the sample as carious if the distance was greater than a given threshold. The fourth classification method calculated the sample's Mahalanobis distance to the mean spectra of the healthy samples and its Mahalanobis distance to the mean spectra of the carious samples, and classified the sample as carious if the difference between the former and the latter was greater than a given threshold.

As a representative of more advanced classification methods, support vector machine was also used to classify the samples. With this method the samples were preprocessed by smoothing and derivating them by using the Savitzky-Golay method with window length $L = 141$ and degree $d = 6$, and by then decimating them, keeping only every eighth feature. After preprocessing the dimensionality of the samples was reduced with principal component analysis (PCA), using various numbers of principal components in the range 1–20.

The performance of the classifier was evaluated with 4-fold cross-validation. All CV folders contained 17 samples from healthy sites and 10 samples from carious sites. Each training set contained one additional sample from a healthy site.

5. RESULTS

Figure 24 presents the samples after averaging 100 consecutive measurements into a single sample, removing the dark current and dividing the samples by the white reference spectra. The two samples that contained a feature whose value was above unit were also removed. In total 109 samples – 69 healthy samples and 40 carious samples – were used in the analysis. They were measured from 21 different teeth.

5.1. Classification with intensity thresholds

The first classification method was allowed to use at most five simple intensity threshold -rules for classifying the samples (see section 4.4). It selected the rules with exhaustive

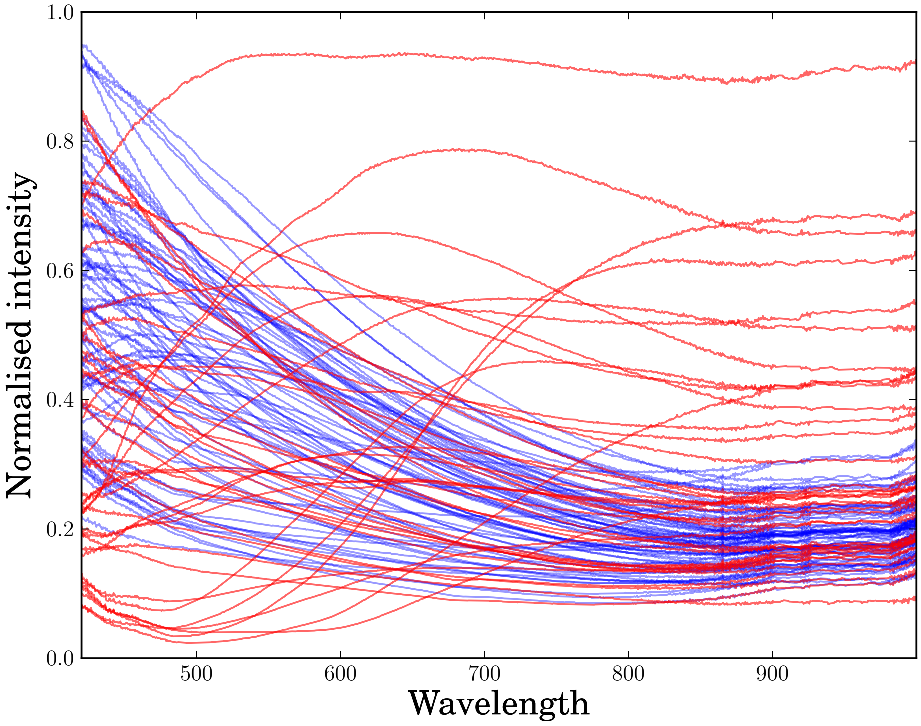


Figure 24. Samples after minimal preprocessing, i.e. after averaging 100 measurements and removing samples which contained a feature whose value was below zero or greater than unit. Blue curves represent samples from healthy sites while red curves represent samples from the carious sites.

Table 6. Classification rules that were selected by the first classifier. The rules are presented in the order they were selected. The sample is classified as carious if, and only if, one or more of these rules is fulfilled. The wavelengths and the accuracies are rounded to integers and the threshold values are rounded to three decimal places. The last three columns present, respectively from left to right, the classification accuracy within the training set of the corresponding CV folder when only the corresponding rule is used, the classification accuracy within the training set when the corresponding rule and all preceding rules in the same CV folder are used, and, in the rightmost column, the classification accuracy within the validation set of the corresponding CV folder when only the corresponding rule is used.

CV folder	Rule Nro	Wavelength	Threshold	Acc.	Comb. acc.	Valid. acc.
1	1	420 nm	≤ 0.261	82%	82%	70%
1	2	609 nm	≥ 0.561	70%	87%	63%
2	1	783 nm	≥ 0.324	80%	80%	63%
2	2	420 nm	≤ 0.246	77%	87%	81%
3	1	420 nm	≤ 0.268	77%	77%	85%
3	2	702 nm	≥ 0.380	76%	85%	70%
4	1	420 nm	≤ 0.273	79%	79%	78%
4	2	752 nm	≥ 0.353	72%	83%	85%
4	3	877 nm	≤ 0.112	66%	84%	63%

search with one thousand wavelength options to choose from. The classifier chose two rules for the first three CV-folders and three rules for the fourth CV-folder. This implies that for most of the samples two rules was the optimal number of rules. The rules that were selected by the classifier are presented in Table 6 and evaluations of the corresponding classification performance are presented in Table 7.

Each set of rules selected by the classifier includes wavelength 420 nm, so that the normalised intensity at this wavelength is given an upper limit which must not be exceeded for the sample to be classified as carious. Even the threshold selected for this wavelength

Table 7. Classification accuracies that were obtained by the first classifier. The values are rounded to integers.

CV folder	Accuracy	Sensitivity	Specificity	PPV	NPV
1	70%	30%	94%	75%	70%
2	81%	50%	100%	100%	77%
3	81%	60%	94%	86%	80%
4	89%	80%	94%	89%	89%
Mean	81%	55%	96%	87%	79%

is similar in all CV-folders: the thresholds were 0.261, 0.246, 0.268, and 0.273. In all but one CV-folder the classifier selected this wavelength for its first rule. In one folder this was the second rule. In all CV-folders the classifier selected another rule with a longer wavelength. These wavelengths had the range 609–784 nm. In three of the four folders this was the second rule, and in one it was the first rule. In all folders this rule presented a lower limit for the intensity.

The average set of rules, which was constructed from the rules that the classifier selected for the CV-folders, contains two rules. The fact that three rules were selected for the fourth folder was ignored as an outlier. The parameters of the two constructed rules are medians of the corresponding parameters in the rules that the classifier selected. The first rule of the average set uses wavelength 420 nm and sets an upper limit of 0.265 for the normalised intensity. The second rule uses wavelength 727 nm and sets a lower limit of 0.366 for the intensity. When the classifier was used on all available samples with these rules the resulting accuracy was 84%, sensitivity 63%, specificity 97%, PPV 93%, and NPV 82%. These values were rounded to integers. The values in the confusion matrix were 25 true positives, 2 false positives, 67 true negatives, and 15 false negatives. The samples, as classified by this classifier with these rules, are presented in Figure 25.

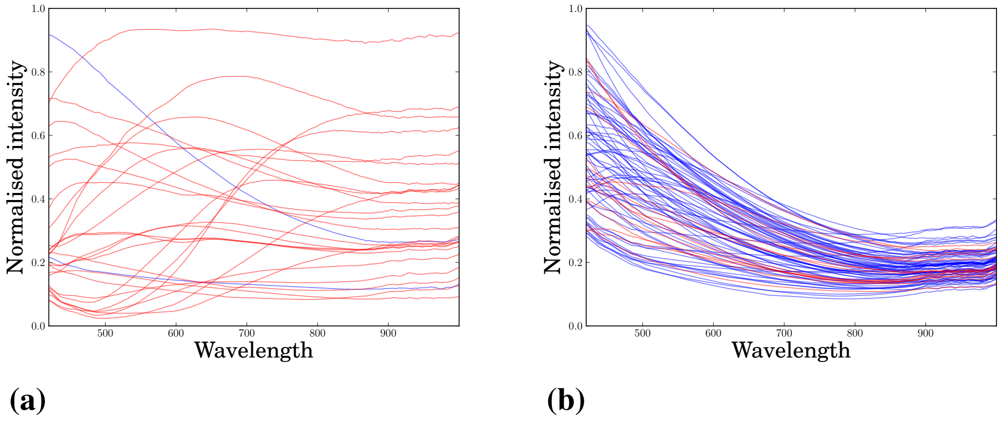


Figure 25. Samples that were classified (a) as carious and (b) as healthy by this classifier with the average set of rules that was constructed from the rules that the classifier selected for the CV-folders. Blue curves represent samples from healthy sites, while red curves represent samples from the carious sites. The set of rules contained two rules, so that wavelength 420 nm had an upper limit of 0.265 for the normalised intensity, and wavelength 727 nm had a lower limit of 0.366 for the intensity. The resulting accuracy was 84%, sensitivity 63%, specificity 97%, PPV 93%, and NPV 82%. These values were rounded to integers.

5.2. Classification with difference in endpoint intensities

The second classification method was first used so that the threshold was given a constant value of zero. With this threshold the sample was classified as carious if, and only if, the normalised intensity at the last available wavelength was greater than the normalised intensity at the first available wavelength. The accuracies that were achieved with this threshold in the CV-folders are presented in Table 8.

Next, the classifier selected the threshold which gave the best accuracy within the training set, and used that for classifying the validation set. The selected thresholds and the results achieved with this method are presented in Table 9.

Table 8. Classification accuracies that were obtained by the second classifier when the threshold was given a constant value of zero. The values are rounded to integers.

CV folder	Accuracy	Sensitivity	Specificity	PPV	NPV
1	74%	30%	100%	100%	71%
2	81%	50%	100%	100%	77%
3	85%	60%	100%	100%	81%
4	81%	50%	100%	100%	77%
Mean	81%	48%	100%	100%	77%

Table 9. Thresholds selected by the second classifier for the CV-folders and the corresponding classification accuracies. The threshold values are rounded to four decimal places and the accuracy values are rounded to integers.

CV folder	Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV
1	-0.0127	74%	30%	100%	100%	71%
2	-0.0842	67%	50%	76%	56%	72%
3	-0.0146	85%	60%	100%	100%	81%
4	-0.0146	81%	50%	100%	100%	77%
Mean	-0.0315	77%	48%	94%	89%	75%

5.3. Classification with one-class Mahalanobis distance

The thresholds selected by the third classifier for the CV-folders and the corresponding classification accuracies are presented in Table 10.

5.4. Classification with two-class Mahalanobis distance

The thresholds selected by the fourth classifier for the CV-folders and the corresponding classification accuracies are presented in Table 11.

Table 10. Thresholds selected by the third classifier for the CV-folders and the corresponding classification accuracies. The threshold values are rounded to two decimal places and the accuracies are rounded to integers. The third classifier was based on the Mahalanobis distance from the mean spectra of the healthy samples.

CV folder	Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV
1	79.98	74%	50%	88%	71%	75%
2	89.69	63%	0%	100%	0%	63%
3	86.37	74%	40%	92%	80%	73%
4	67.30	67%	80%	59%	53%	83%
Mean	80.84	69%	43%	85%	51%	74%

Table 11. Thresholds selected by the third classifier for the CV-folders and the corresponding classification accuracies. The threshold values are rounded to two decimal places and the accuracies are rounded to integers. The fourth classifier was based on the difference between the Mahalanobis distance from the mean spectra of the healthy samples and the Mahalanobis distance from the mean spectra of the carious samples.

CV folder	Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV
1	-13.98	59%	60%	59%	46%	71%
2	0.08	67%	20%	94%	67%	67%
3	4.84	81%	50%	100%	100%	77%
4	1.49	74%	70%	76%	64%	81%
Mean	18.49	69%	48%	81%	64%	73%

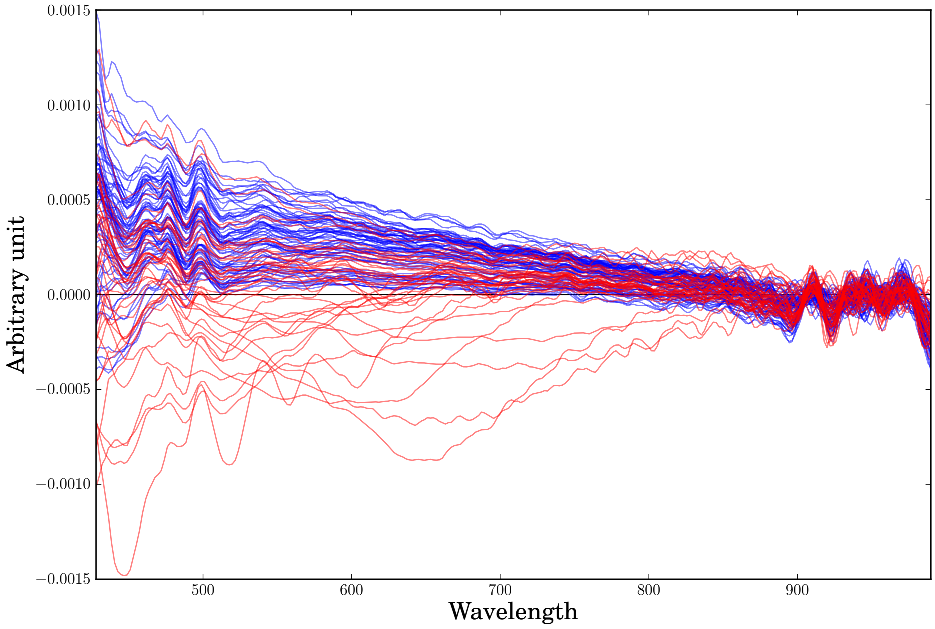


Figure 26. Samples after preprocessing for SVM – i.e. after averaging 100 measurements, removing outliers, smoothing and derivating with Savitzky-Golay method, and decimating – but before PCA transformation. Blue curves represent samples from healthy sites and red curves represent samples from carious sites.

5.5. Classification with support vector machine

The classification with support vector machine had several parameters: the number of principal components from principal component analysis (PCA), the type of kernel used, and the degree of the kernel. Choices for the values of these parameters were preceded by the choice of the preprocessing methods, which were outlined in section 4.8.

Samples after preprocessing for the support vector machine are presented in Figure 26. The first three principal components that were identified by the principal component analysis (PCA) in the training set of the first CV-folder are presented in Figure 27. Figure 28 presents the fraction of variance in the samples of the first CV-folder's training set that can be represented by the principal components as a function of the number of principal components. The graph was similar for the other three CV-folders. Six principal components could represent more than 98% of the variance in the samples in all four CV-folders.

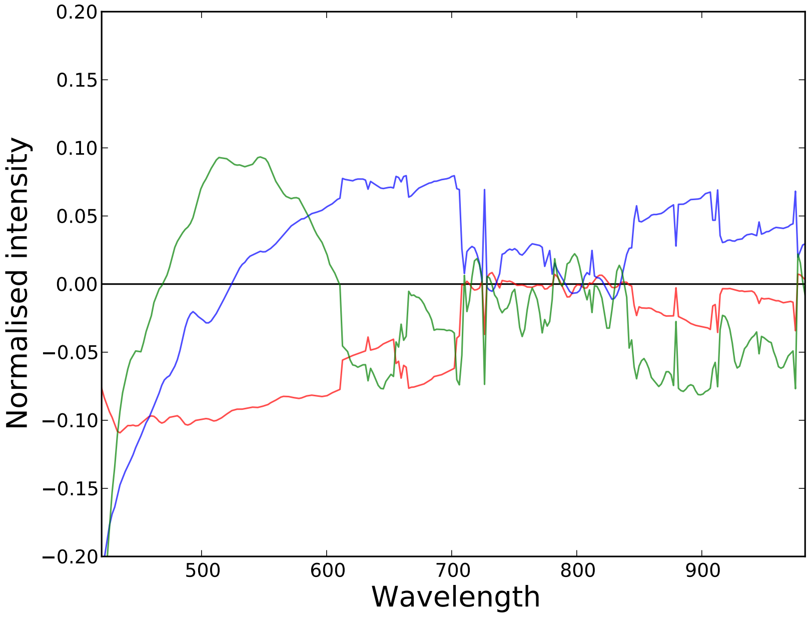


Figure 27. The first three principal components identified by the principal component analysis (PCA) in the training set of the first CV-folder, presented as red, blue and green curves, respectively.

The analysis was run with different kernels for the SVM, using six principal components from PCA and different degrees for the kernels. The resulting mean AUCs are tabulated in Table 12. A polynomial kernel of fifth degree was selected for more extensive analysis because it seemed to produce the best results. With this kernel the sensitivity was 0% and specificity was 100% for all CV folders.

The number of principal components in the range 1–20 had extremely small effect on the mean area under curve (AUC) after cross-validation when using the polynomial kernel of fifth degree. With every number of principal components the mean of the CV-folders' AUCs was between 81.2% and 81.5%. Five principal components seemed to produce somewhat better results than fewer principal components, and the addition of further principal components did not seem to improve the results. Thus, the results for five principal components are used as representative results. The ROC-curves for the four CV folders, using five principal components, are presented in Figure 29.

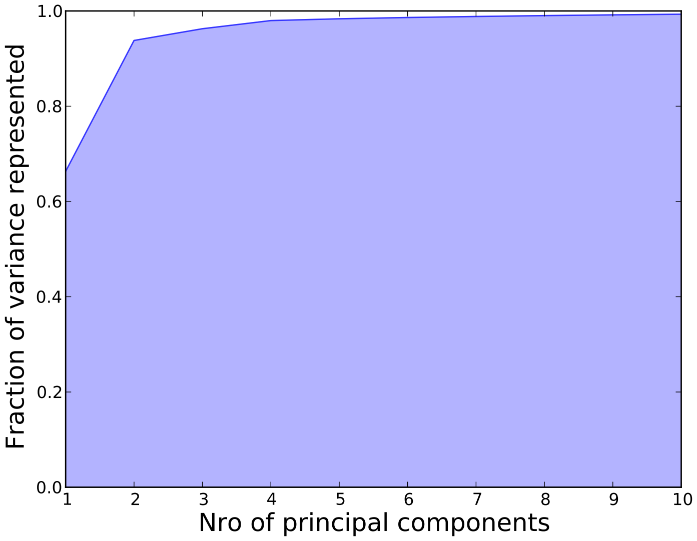


Figure 28. The fraction of variance in the samples of the first CV-folder's training set that can be represented by the principal components as a function of the number of principal components. The graph was similar for the other three CV-folders. Six principal components could represent more than 98% of the variance in the samples in all four CV-folders.

Table 12. Mean AUCs for the CV folders using various kernels for the SVM with six principal components from principal component analysis (PCA).

Kernel	Degree	Mean AUC
Linear	–	80.7%
Polynomial	3	81.2%
Polynomial	5	81.5%
Polynomial	10	63.2%
RBF	3	80.9%
RBF	10	80.9%
Sigmoid	3	50%
Sigmoid	10	50%

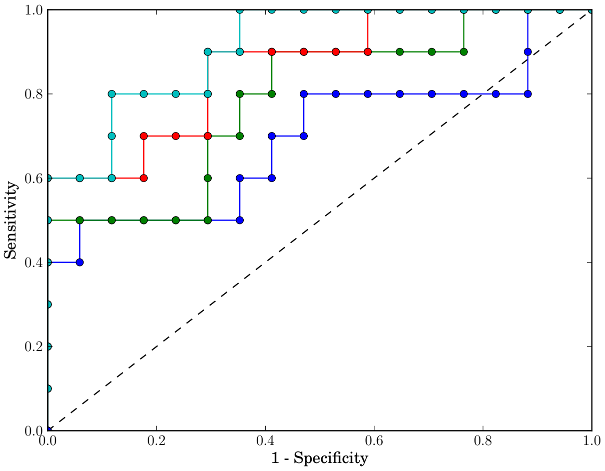


Figure 29. ROC-curves for the CV folders using five principal components and SVM with a polynomial kernel of fifth degree as the classifier. The folders' ROC-curves are depicted by the blue, green, red and cyan curve, respectively. The sensitivity may also be described as the true positive rate and the value $1 -$ specificity may be described as false positive rate (Fawcett 2006).

Table 13 presents classification accuracies that could be reached with five principal components by varying the decision threshold. We can see in the table that in all CV folders the `libsvm`-library chose a decision threshold that was far greater than the threshold that would have produced the optimal classification accuracy within the training set. In fact, the chosen decision threshold was always large enough to cause all samples to be classified as samples from a healthy site, i.e. samples from the negative class. We can also see that the resulting classification accuracy within the validation set was always greater than the accuracy that would have been achieved with the threshold that was optimal for the training set. The negative class had a higher prevalence than the positive class in all CV folders (see section 4.9). Therefore, the results are consistent with a situation where the classifier predicts the sample's class at random and seeks to maximize the resulting classification accuracy (see page 61). Because all validation sets contained 17 samples from healthy sites and 10 samples from carious sites, the decision threshold chosen by the `libsvm`-library produced the same accuracy in all CV folders.

Table 13. The optimal decision thresholds and the corresponding accuracies for the four CV folders, using five principal components and SVM with a polynomial kernel of fifth degree as the classifier. The optimal threshold gave the optimal classification accuracy within the training set. Perfect accuracy was achieved in all training sets. The last two columns present the decision threshold that was selected by the `libsvm`-library, and the corresponding classification accuracy. The accuracies were rounded to integers.

PCA	CV	Optimal threshold	Accuracy	Threshold	Accuracy
5	1	3.789×10^{-36}	59%	1.000	63%
5	2	1.017×10^{-35}	30%	1.000	63%
5	3	1.004×10^{-35}	52%	1.000	63%
5	4	1.317×10^{-35}	44%	1.000	63%

6. DISCUSSION

The first classification method, which was based on intensity threshold -rules, showed how some of the carious samples were easily detected as such. Classification accuracy of approximately 84% was achieved with this method. The second classification method, which was based on the difference between the intensity at the first available wavelength and the intensity at the last available wavelength, showed an even simpler method for detecting the clearly carious samples. However, the rest of the carious samples resemble the healthy samples (Fig. 30), making it much more difficult to classify them as carious. There were 25 carious samples that could be detected as carious by the first classification method, and 15 carious samples that were not detected.

There were many more healthy samples than carious samples available for analysis, namely 69 healthy samples compared to 40 carious samples. Thus, the classifiers that were searching for optimal accuracy may have classified the carious samples incorrectly rather than classifying the healthy samples incorrectly.

The fact that the first classifier selected similar rules in every CV-folder (see section 5.1) suggests that these rules describe a phenomenon that is consistently present in the samples in all CV-folders. These rules are partly consistent with our research hypothesis. The rules show that carious samples tend to have higher intensity than the healthy samples at the long wavelengths, namely in the range 609–784 nm, and lower intensity than the healthy samples at the shorter wavelengths, namely at 420 nm – which was the shortest wavelength included in the detailed analysis. According to our hypothesis (see chapter 4) the increased scattering in the near-infrared range is the best indication of a dental caries lesion. However, according to the rules the increased absorbance in the visible range is an even better indication of a lesion. We will soon argue why this is most likely caused by stains in the samples, which mislead the author to erroneously diagnose those samples as carious.

The carious samples that resemble healthy samples (Fig. 30) might be mislabeled, i.e. they may be healthy samples, that were misdiagnosed as carious while making the measurements. The samples were diagnosed by the author, who has no prior experience in

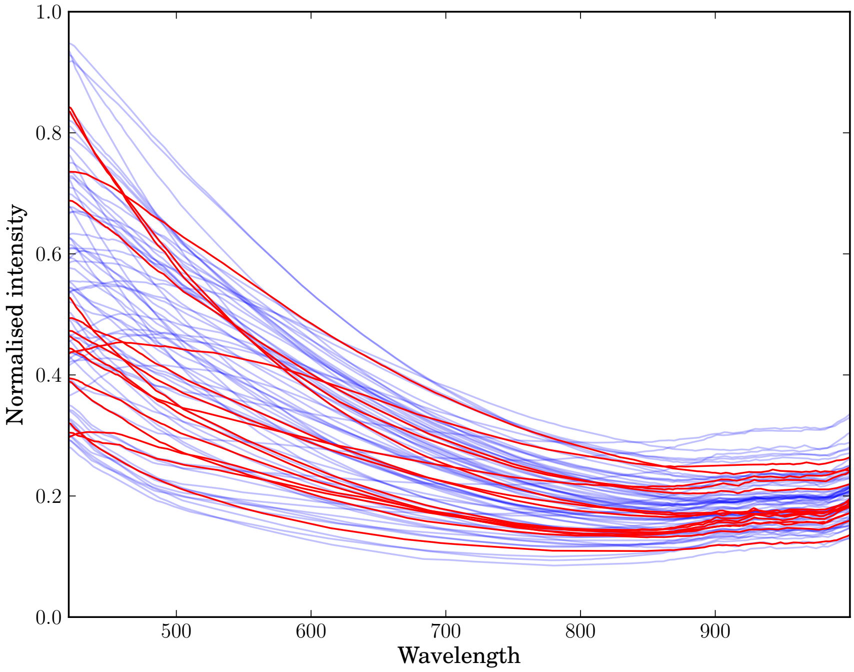
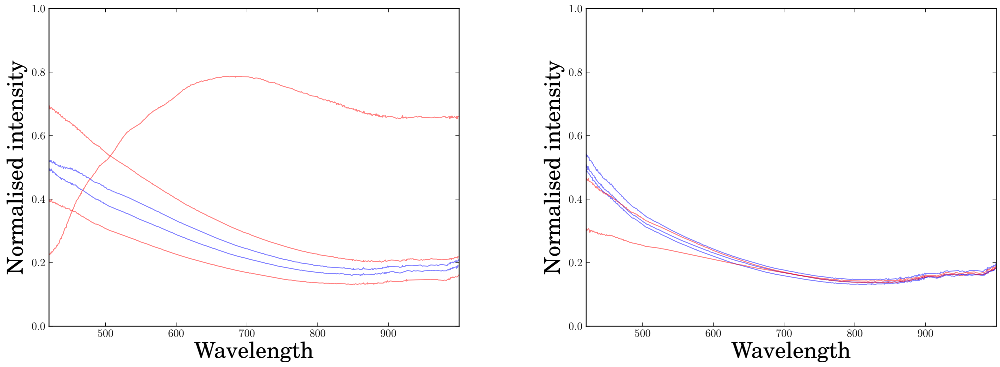


Figure 30. Samples that where classified as healthy by the first classification method, i.e. the intensity threshold -rules, using the average rule set. Blue curves represent samples from healthy sites while red curves represent samples from the carious sites. Here, the red curves are emphasized. This picture illustrates how some of the carious samples resemble healthy samples.

diagnosing carious lesions. Support for this possibility can be seen in the samples that were measured from two particular teeth (Fig. 31). The samples show how a supposed caries lesion may have a higher or lower intensity than the healthy samples at all wavelengths, or it may be clearly different than the healthy samples. Alternatively, a carious sample and a healthy sample may be very similar, even when measured from the same tooth.

A total of 109 samples were used in the analysis. If only the 15 carious samples that were not detected as carious by the first classifier (Fig. 30), i.e. the false negatives, were misdiagnosed or mislabeled, that would imply that the author's diagnostic accuracy during the measurements was 86%. This might be considered as a reasonable accuracy



(a) **(b)**

Figure 31. Two sets of samples, so that each set was measured from a particular tooth. Blue curves represent samples from healthy sites while red curves represent samples from the carious sites. This picture presents support for the idea that some of the samples may be mislabeled. The five samples in (a) are the samples that were measured from a particular tooth, after only the common preprocessing steps (see section 4.3). They show how a supposed caries lesion may have a higher or lower intensity than the healthy samples at all wavelengths, or it may be clearly different than the healthy samples. The five samples in (b) are the samples that were measured from a different tooth, after only the common preprocessing steps. They show how a carious sample and a healthy sample may be very similar, even when measured from the same tooth.

for the first few sessions of diagnosing caries lesions. The rules that the first classifier would select in this case are presented in Table 14 and the resulting accuracies in the CV-folders are presented in Table 15. The median rule set of these rules is that the normalized intensity of the carious samples is equal to or less than 0.264 at wavelength 420 nm, or equal to or greater than 0.367 at wavelength 727 nm. When applied to all samples, these rules produced accuracy of 98%, sensitivity of 100%, specificity of 98%, PPV of 93% and NPV of 100%. In other words, in that case the classifier would be as accurate as the author in detecting caries lesions. Measurements taken from three teeth were discarded during the preprocessing because they were suspected of being mislabeled. This resulted in total of 20 samples being discarded. If all these samples are also considered as misdiagnosed,

Table 14. Classification rules that were selected by the first classifier, when 15 samples that were suspected of being misdiagnosed as carious were relabeled as healthy. The rules are presented in the order they were selected. The sample is classified as carious if, and only if, one or more of these rules is fulfilled. The last two columns present, respectively, the classification accuracy within the training set of the corresponding CV folder when only the corresponding rule is used, and the classification accuracy within the training set when the corresponding rule and all preceding rules in the same CV folder are used. The wavelengths and the accuracies are rounded to integer values and the threshold values are rounded to three decimal places.

CV folder	Rule Nro	Wavelength	Threshold	Acc.	Comb. acc.
1	1	420 nm	≤ 0.261	95%	95%
1	2	609 nm	≥ 0.561	84%	100%
2	1	783 nm	≥ 0.324	93%	93%
2	2	420 nm	≤ 0.246	89%	99%
3	1	420 nm	≤ 0.268	94%	94%
3	2	702 nm	≥ 0.380	89%	99%
4	1	420 nm	≤ 0.273	94%	94%
4	2	752 nm	≥ 0.353	88%	99%

the author's diagnostic accuracy would still have been 73%.

Based on previous research on detecting dental caries lesions with transilluminating NIR-light, it seems that wavelengths in the NIR range are more useful in diagnosing caries lesions than wavelengths in the visible range (Jones et al 2003; Wu & Fried 2009; Staninec et al 2010). Longer wavelengths have been used for detecting dental caries lesions based on fluorescence, though (Karlsson 2010). A reasonable starting point for this study is to assume that the same wavelengths which are useful in detecting caries lesions with NIR transillumination are also useful in the same task with NIR reflectance spectroscopy. The results of this study partly support this hypothesis. Wavelengths in the NIR range, particularly in the range 609–784 nm, were useful. However, a longer wavelength, at

Table 15. Classification accuracies that were obtained by the first classifier, when 15 samples that were suspected of being misdiagnosed as carious were relabeled as healthy. The values are rounded to integers.

CV folder	Accuracy	Sensitivity	Specificity	PPV	NPV
1	86%	57%	95%	80%	87%
2	100%	100%	100%	100%	100%
3	96%	100%	95%	88%	100%
4	100%	100%	100%	100%	100%
Mean	96%	89%	98%	92%	97%

approximately 420 nm, was found to be even more useful. This contradicts the results for NIR transillumination, suggesting that the optimal wavelength set is different for NIR reflectance spectroscopy than for NIR transillumination.

The sensitivity spectrum of the spectroscope that was used to make the measurements might affect the perceived optimal set of wavelengths. Particularly, if the spectroscope is more sensitive at the visible range than at the NIR range, the signal-to-noise ratio might be more favourable for the wavelengths at the visible range than at the NIR range, making them appear more useful. The sensitivity spectrum is affected by the sensitivity of the spectroscope's sensor at various wavelengths, by the intensity of the light source at various wavelengths, and by the optical properties of the fiber optics used. Calibration of the spectroscope with a white reference can be expected to partly correct for such differences. It can not fully compensate for a low intensity of the light source at a given wavelength range, though.

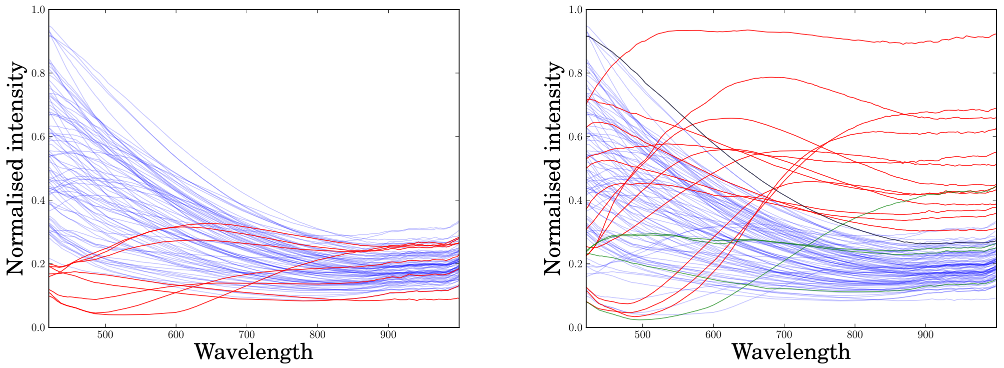
However, there is a more probable explanation for the apparent usefulness of wavelength 420 nm. A stain might mislead the author to erroneously diagnose the sample as carious when making the measurements. In such case, the sample's spectra would resemble the spectra of a healthy sample, except for the absorption in the visible range, and the sample would be labeled as carious. This hypothesis was tested by first relabeling the samples that were previously suspected of being mislabeled as carious (Fig. 30). Then samples

Table 16. Classification rules that were selected by the first classifier, when 23 samples that were suspected of being misdiagnosed as carious were relabeled as healthy. The wavelengths and the accuracies are rounded to integers and the threshold values are rounded to three decimal places. The last column presents the classification accuracy within the training set of the corresponding CV folder.

CV folder	Wavelength	Threshold	Acc.
1	783 nm	≥ 0.324	98%
2	752 nm	≥ 0.353	98%
3	779 nm	≥ 0.320	98%
4	783 nm	≥ 0.324	98%

whose spectra showed absorbance in the visible range and whose spectra did not show increased scattering in the NIR range were considered as false positives due to misleading stains, and were consequently relabeled as healthy. More precisely, a sample was thought to represent a stain if the normalized intensity of its spectra was below 0.206 at wavelength 420 nm, and below 0.313 at wavelength 815 nm. This way eight samples became suspected stains (Fig. 32a). New rules were then selected for the first classification method in the same manner as before (see section 4.4). The resulting rules are presented in Table 16, and the resulting accuracies are presented in Table 17. The new rules indicated that wavelengths in the range 750–785 nm were most useful in detecting caries lesions. The median of the rules for the four CV-folders was that a sample is carious if its normalized intensity at wavelength 781 nm is equal or greater than 0.324. When applied to all samples, after the relabeling described above, this resulted in classification accuracy of 95% (Fig. 32b).

The samples are therefore consistent with the hypothesis that stains mislead the author to erroneously diagnose eight samples as carious, causing the wavelength 420 nm to appear more useful in diagnosing caries lesions than it actually is. The author's diagnostic accuracy would then fall to 79%, not considering the discarded samples. With this additional hypothesis, the results are consistent with the research hypothesis.



(a) **(b)**

Figure 32. (a) The red curves represent the eight samples that where considered as false positives due to misleading stains. The blue curves represent samples from healthy sites. A sample was thought to represent a stain if the normalized intensity of its spectra was below 0.206 at wavelength 420 nm, and below 0.313 at wavelength 815 nm. (b) The red and blue curves represent samples that were classified as carious or as healthy by the new rule(s) of the first classifier, respectively. There were 91 correctly classified healthy samples and 13 correctly classified carious samples. The four green curves represent samples that were diagnosed as carious but classified as healthy (i.e. false negatives), and the one black curve represents the sample that was diagnosed as healthy but classified as carious (i.e. a false positive). Overall the classification accuracy was 95%.

Table 17. Classification accuracies that were obtained by the first classifier, when 23 samples that were suspected of being misdiagnosed as carious were relabeled as healthy. The values are rounded to integers.

CV folder	Accuracy	Sensitivity	Specificity	PPV	NPV
1	96%	80%	100%	100%	96%
2	93%	60%	100%	100%	92%
3	93%	80%	96%	80%	96%
4	96%	80%	100%	100%	96%
Mean	95%	75%	99%	95%	95%

We argued above that some of the samples are most likely misdiagnosed, based on their spectra, spectra of other samples, and the results of previous research. It follows from those arguments that the results of the spectroscopic measurements can be used to improve the author's diagnostic accuracy. In other words, diffuse reflectance near-infrared spectroscopy is able to detect dental caries lesions with a higher accuracy than what the author could reach with a manual inspection with fiber-optic illumination. Thus, NIR spectroscopy seems to improve the diagnostic accuracy of a manual inspection, at least when the inspection is done by a novice. This claim is contingent on an assumption that all healthy sites of enamel have spectra that somewhat resemble each other, and partly on an assumption that all carious lesions on enamel show increased scattering in the near-infrared range. If more weight were placed on the diagnoses of the author, these assumptions would have to be questioned. However, in this study it is more plausible that the author made a number of misdiagnoses.

In this study the combination of PCA and SVM did not find any information in the samples' spectra that would be useful in detecting caries lesions – at least not with the preprocessing methods that were used. PCA seeks to represent a spectra as a linear combination of component spectra. It assumes that the component spectra can be recognised by the variance they cause in the samples' spectra. More precisely, the changes in the contributions from the components, or changes in their weights, are assumed to be the greatest sources of variance in the samples' spectra. It seems that in this case this assumption did not hold.

Figure 33 presents the first three principal components that are identified by principal component analysis (PCA) in the training set of the fourth CV-folder, when the samples undergo the preprocessing that was used with the first classifier. When they are compared to the results of the first classifier, the first principal component seems to recognise the significance of the short wavelengths around 420 nm, while the second principal component seems to recognise also the significance of the longer wavelengths. The first principal component can represent 97.4% of the variance in the samples, and the first two principal components can represent 99.8% of the variance in the samples. All three spectra are quite jagged, which demonstrates how the principal components try to represent the en-

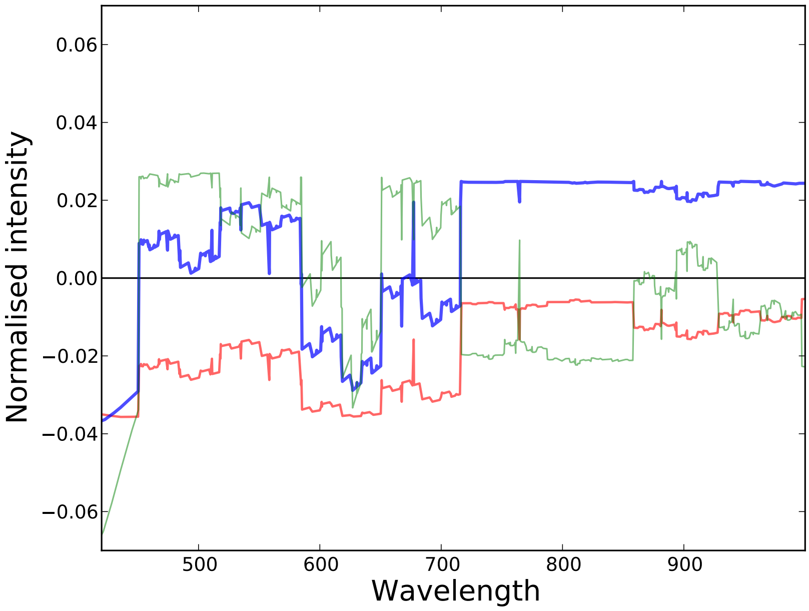
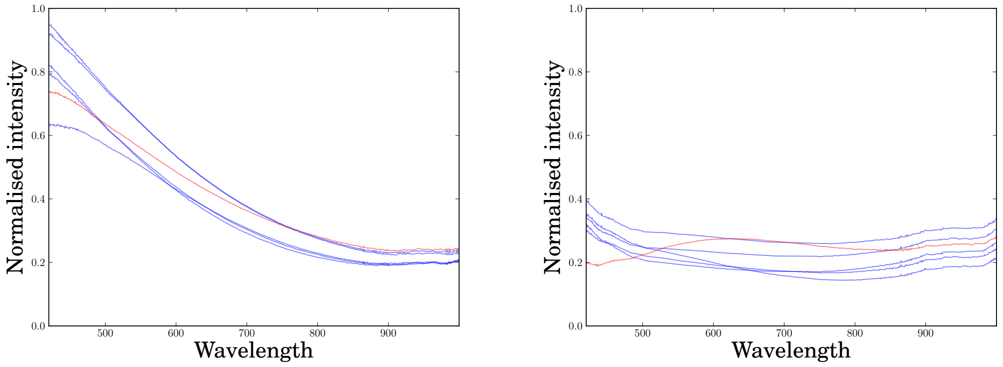


Figure 33. The first three principal components that are identified by principal component analysis (PCA) in the training set of the fourth CV-folder, when the samples undergo the preprocessing that was used with the first classifier. The components are presented as red, blue and green curves, respectively.

tire spectra of the samples, including the features that are not useful for the classification. Considering the useless features in the pattern analysis, i.e. in the training of the SVM, confuses the analysis and makes it more difficult to find the classes' defining features (see section 3.2).

This highlights the importance of using a preprocessing procedure that is appropriate for the pattern to be found and for the pattern analysis method used to search for it. Perhaps PCA works well in chemometrical applications, where the sample spectra are indeed linear combinations of the spectra of the components in a solution or compound, possibly affected by the environmental conditions, like the temperature, and by their coupling effects. This kind of samples do fit the assumptions that PCA makes. Evidently this method is less suited for the task of caries detection, though. It is possible, or probable, that appropriate preprocessing would have helped the combination of PCA and SVM to reach better results in this study. However, since the much simpler first classifier worked so



(a) **(b)**

Figure 34. Two sets of samples, where each set was measured from a particular tooth. Blue curves represent samples from healthy sites while red curves represent samples from the carious sites. This picture presents support for the idea that the spectra of healthy enamel is different in different teeth. The six samples in (a) are the samples that were measured from a particular tooth, and the six samples in (b) are the samples that were measured from a different tooth, all after only the common preprocessing steps (see section 4.3).

well, it might be difficult to justify the use of PCA and SVM.

The composition of the dental tissues varies from tooth to tooth and between different sites of a given tooth (see section 2.1.2). This may cause differences in the spectra of healthy enamel, which complicates the detection of caries lesions. Indications of the presence of variance between the teeth can be seen in the samples (Fig. 34). In this study the extracted teeth that were measured had only small differences in their colors. In a clinical setting much larger color variations may be encountered, for example, due to smoking. The effects of such variations in the spectroscopic detection of caries lesions needs to be considered in future work. These variations may undermine the accuracy of caries lesion detection with NIR spectroscopy, especially its ability to improve the accuracy of a manual inspection.

In future work the diagnosis or labeling of the samples, i.e. the ground truth, could be improved by using an appropriate reference or gold standard method for diagnosing the samples or by forming the caries lesions artificially *in vitro*. More reliable labeling can be expected to lead to more reliable results, and consequently more reliable conclusions. One option for the reference method is serial histological sectioning, which grinds the tooth in steps of 0.1–0.2 mm, taking a photograph at each step (Zakian et al 2010: 790). This obviously destroys the tooth, making it impossible to use the same tooth for further research or to use this method in clinical practice.

Alternatively, caries can be induced *in vitro* by painting tooth samples with acid-resistant varnish and by leaving windows on which the caries is to be induced, and then immersing the samples in demineralization solution. The demineralization solution needs to have a controlled pH-value of approximately 4.5 and temperature of 37°C. (Wu & Fried 2009.) More realistic demineralization may be obtained by cycling the samples in de- and remineralization solutions or by adding microorganism, e.g. yeast extract, to the demineralization solution (see Marquezan, Correa, Sanabe, Filho, Hebling, Guedes-Pinto & Mendes 2009). The CO_2 concentration of the gas environment effects the caries formation as well (Aoba 2004: 252–253).

When a number of teeth are kept in the demineralization solution for a given time period, they will undergo varying degrees of demineralization because the precise chemical composition of a tooth varies between teeth. The amount of mineral dissolved from the tooth could be measured from the solution in an attempt to control the measurements for this variable. A participant of the FIELD-NIRce project, Ketek Oy, has a mass spectrometer that we might be able to use for these measurements.

7. CONCLUSIONS

The results of this study suggest that diffuse reflectance near-infrared spectroscopy is able to improve the diagnostic accuracy of manual inspection with fiber-optic illumination, at least when the inspection is done by a novice. This claim is contingent on an assumption that all healthy sites of enamel have spectra that somewhat resemble each other, and partly on an assumption that all carious lesions on enamel show increased scattering in the near-infrared range. The reliability of these results is limited by the author's ability to diagnose caries lesions with the said manual method, and by the samples' ability to represent the variance among sites of healthy enamel and among caries lesions, though. More reliable diagnoses can be expected to lead to more reliable results, and consequently more reliable conclusions.

In this study, the best classification results where obtained with a classifier that used a number of rules (see section 4.4), so that every rule had the following format: if the sample's normalised intensity at wavelength λ was greater than (or smaller than) threshold t, the sample was classified as carious; otherwise, the sample was classified as healthy. At least one such rule had to consider the sample as carious in order to classify the sample as carious. The classifier first selected the rule(s) that gave the best classification accuracy in a training set, and then applied these rules to a validation set. This classifier reached an accuracy of approximately 84%. Its sensitivity was 63% and specificity 97%, which shows how there were more healthy samples than carious samples available.

Further analysis suggested that the author had erroneously diagnosed 23 samples as caries lesions while measuring the samples, eight of them due to a misleading stain. When these samples were relabeled, the classification method that was described above selected rules that were consistent with the research hypothesis, and which reached a classification accuracy of 95%. This lead to the conclusion that NIR spectroscopy is able to improve the diagnostic accuracy of manual inspection.

An interested reader might want to explore the publications of Professor Daniel Fried from University of California, San Francisco, and those of his students, particularly Dr Robert S. Jones (see section 1.2).

REFERENCES

Alaluusua, Satu, Liisa Aine, Sirkka Asikainen, Anna-Leena Eriksson, Kirsti Hurmerinta, Päivi Hölttä, Sára Karjalainen, Pirjo-Liisa Lukinmaa & Sinikka Pirinen (2004). Pedodontia. In: *Therapia Odontologica – Hammaslääketieteen käsikirja*, pp. 529–584. Ed.: J. Meurman, H. Murtomaa, Y. LeBell & H. Autti. Academica-Kustannus Oy. ISBN 952-5046-04-4.

Aoba, Takaaki (2004). Solubility properties of human tooth mineral and pathogenesis of dental caries. *Oral Diseases* 10:5, 249–257. doi:10.1111/j.1601-0825.2004.01030.x.

Autti, Heikki, Yrsa Le Bell, Jukka H. Meurman & Heikki Murtomaa (2004). Ongelmalähtöinen diagnostiikka. In: *Therapia Odontologica – Hammaslääketieteen käsikirja*, pp. 27–90. Ed.: J. Meurman, H. Murtomaa, Y. LeBell & H. Autti. Academica-Kustannus Oy. ISBN 952-5046-04-4.

Avantes Inc. (2009). Product Catalog 2009–2010, p. 135. `<URL: http://www.avantes.com/Download-document/171-Catalogus-2009-2010-pag-111-135-H5-Accessories.html>`.

Avery, James K. (1976). Dentin. In: *Orban's Oral Histology and Embryology*, pp. 105–141. 8th edition, Ed.: S.N. Bhaskar. The C. V. Mosby Company. ISBN 0-8016-4608-1.

Baysan, Aylin (2007). The diagnosis of caries – Initial clinical examination. In: *Minimally Invasive Dentistry – The Management of Caries*, pp. 29–33. Ed.: Nairn H.F. Wilson. Quintessence Publishing Co, Ltd. ISBN 978-1-85097-105-4.

Beighton, D. & D. Bartlett (2006). Dental caries and pulpitis. In: *Clinical Textbook of Dental Hygiene and Therapy*, pp. 75–92. Ed.: Robert Ireland. Blackwell Munksgaard. ISBN 978-1-4051-3540-5.

Bokobza, L. (2002). Origin of near-infrared absorption bands. In: *Near-Infrared Spectroscopy – Principles, Instruments, Applications*, pp. 11–41. Ed.: H. Siesler, Y. Ozaki, S. Kawata & H. M. Heise. Wiley-Vch Verlag GmbH. ISBN 3-527-30149-6.

Boser, Bernhard E., Isabelle M. Guyon & Vladimir N. Vapnik (1992). A training algorithm for optimal margin classifiers. In: *Proceedings of The Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152. ACM. doi:10.1145/130385.130401.

Boyd, Stephen & Lieven Vandenberghe (2004). *Convex Optimization.* Cambridge University Press. ISBN 978-0-521-83378-3. `<URL: http://www.stanford.edu/~boyd/cvxbook/>`.

Bradley, Andrew P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30:7, 1145–1159. doi:10.1016/S0031-3203(96)00142-2.

Bürmen, Miran, Peter Usenik, Aleš Fidler, Franjo Pernuš & Boštjan Likar (2011). A construction of standardized near infrared hyper-spectral teeth database: A first step in the development of reliable diagnostic tool for quantification and early detection of caries. In: *Proceedings of SPIE*, vol. 7884, p. 78840E. doi:10.1117/12.875059.

Chang, Chih-Chung & Chih-Jen Lin (2001). LIBSVM: A library for support vector machines. Software available at `<URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm>`.

Chung, S., D. Fried, M. Staninec & C.L. Darling (2011). Near infrared imaging of teeth at wavelengths between 1200 and 1600 nm. In: *Proceedings of SPIE*, vol. 7884, p. 78840X. doi:10.1117/12.878894.

Control Development Inc. (2011). Near Infrared Spectrometers. `<URL: http://www.controldevelopment.com/pdfs/Spectrometers/NIR_Spec.pdf>`.

Dawes, Colin (2003). What is the critical pH and why does a tooth dissolve in acid? *Journal of the Canadian Dental Association* 69:11, 722–724.

Dorozhkin, S.V. (2009). Calcium orthophosphates in nature, biology and medicine. *Materials* 2:2, 399–498. doi:10.3390/ma2020399.

Duda, Richard O., Peter E. Hart & David G. Stork (2001). *Pattern Classification.* 2nd edition. John Wiley & Sons, Inc. ISBN 0-471-05669-3.

Erkkilä, Heli (1992). Fraktaalirakenteiden tutkiminen TiO_2-hydraateista röntgensäteilyn pienkulmasironnalla. Master's thesis, Department of Physics, University of Turku.

Fawcett, Tom (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27:8, 861–874. doi:10.1016/j.patrec.2005.10.010.

Franklin, P. & P. Brunton (2006). Restorative materials. In: *Clinical Textbook of Dental Hygiene and Therapy*, pp. 208–225. Ed.: Robert Ireland. Blackwell Munksgaard. ISBN 978-1-4051-3540-5.

Fried, Daniel, JD Featherstone, C.L. Darling, R.S. Jones, P. Ngaotheppitak & C.M. Bühler (2005). Early caries imaging and monitoring with near-infrared light. *Dental Clinics of North America* 49:4, 771.

Furukawa, Yukio (2002). Near-infrared FT-Raman spectroscopy. In: *Near-Infrared Spectroscopy – Principles, Instruments, Applications*, pp. 85–114. Ed.: H. Siesler, Y. Ozaki, S. Kawata & H. M. Heise. Wiley-Vch Verlag GmbH. ISBN 3-527-30149-6.

Greenberg, Michael D. (1998). *Advanced Engineering Mathematics.* 2nd edition. Prentice-Hall, Inc. ISBN 0-13-321431-1.

Hein, Morris, Leo R. Best, Scott Pattison & Susan Arena (1997). *Introduction to General, Organic, and Biochemistry.* 6th edition. Brooks/Cole Publishing Company. ISBN 0-534-25878-6.

Hellen, Adam (2010). Quantitative Evaluation of Simulated Enamel Demineralization and Remineralization using Photothermal Radiometry and Modulated Luminescence. Master's thesis, University of Toronto, Graduate Department of Dentistry.

Hendee, William R. & E. Russell Ritenour (2002). *Medical Imaging Physics*. 4th edition. Wiley-Liss, John Wiley & Sons, Inc. ISBN 0-471-38226-4.

Hill, W. & V. Petrou (1997). Detection of caries and composite resin restorations by near-infrared Raman spectroscopy. *Applied Spectroscopy* 51:9, 1265–1268.

Hunter, John D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9:3, 90–95.

Jones, Eric, Travis Oliphant, Pearu Peterson et al (2001–). SciPy: Open source scientific tools for Python. <URL: http://www.scipy.org/>.

Jones, Robert S., Gigi D. Huynh, Graham C. Jones & Daniel Fried (2003). Near-infrared transillumination at 1310-nm for the imaging of early dental decay. *Optics Express* 11:18, 2259–2265. doi:10.1364/OE.11.002259.

Jones, Robert Simon (2006). *Near-infrared Optical Imaging of Early Dental Caries*. Ph.D. thesis, University of California, San Francisco.

Karlsson, Lena (2009). *Optical Based Technologies for Detection of Dental Caries*. Ph.D. thesis, Karolinska Institutet.

Karlsson, Lena (2010). Caries detection methods based on changes in optical properties between healthy and carious tissue. *International Journal of Dentistry* 2010. doi: 10.1155/2010/270729.

Ko, AC, L.P. Choo-Smith, R. Zhu, M. Hewko, C. Dong, B. Cleghorn & MG Sowa (2006). Application of NIR Raman spectroscopy for detecting and characterizing early dental caries. In: *Proceedings of SPIE*, vol. 6093, p. 60930L. doi:10.1117/12.647159.

Ko, Alex C.T., Mark Hewko, Michael G. Sowa, Cecilia C.S. Dong & Blaine Cleghorn (2008). Early dental caries detection using a fibre-optic coupled polarization-resolved Raman spectroscopic system. *Optics Express* 16:9, 6274. doi:10.1364/OE.16.006274.

Lampinen, Jouni (2000). Multiobjective Nonlinear Pareto-optimization: A Pre-Investigation Report. Lappeenranta University of Technology, Laboratory of Information Processing, Lapperanta, Finland.

Lee, Chulsung, Dustin Lee, Cynthia L. Darling & Daniel Fried (2010). Nondestructive assessment of the severity of occlusal caries lesions with near-infrared imaging at 1310 nm. *Journal of Biomedical Optics* 15:4, 047011. doi:10.1117/1.3475959.

Maia, A.M.A., D.D.D. Fonseca, B.B.C. Kyotoku & A.S.L. Gomes (2009). Evaluation of sensibility and specificity of NIR Transillumination for early enamel caries detection-an in vitro study. In: *European Conference on Lasers and Electro-Optics 2009 and the European Quantum Electronics Conference (CLEO Europe-EQEC 2009)*, p. 1. IEEE. doi:10.1109/CLEOE-EQEC.2009.5192541.

Marquezan, Marcela, Fernanda Nahas P. Correa, Mariane Emi Sanabe, Leonardo Eloy Rodrigues Filho, Josimeri Hebling, Antonio Carlos Guedes-Pinto & Fausto Medeiros Mendes (2009). Artificial methods of dentine caries induction: A hardness and morphological comparative study. *Archives of Oral Biology* 54:12, 1111–1117. doi:10.1016/j.archoralbio.2009.09.007.

McClure, W. Fred (2008). Analysis using Fourier transforms. In: *Handbook of Near-Infrared Analysis*, pp. 93–121. 3rd edition, Ed.: Donald A. Burns & Emil W. Ciurczak. CRC Press, Taylor & Francis Group. ISBN 978-0-8493-7393-0.

Netter, Frank H. (1989). *Atlas of Human Anatomy.* Ciba-Ceigy Corporation. ISBN 0-914168-19-3.

Ocean Optics Inc. (2008). HR4000 and HR4000CG-UV-NIR Series High-Resolution Fiber Optic Spectrometers – Installation and Operation Manual. Dunedin, FL, USA.

Ocean Optics Inc. (2011). General Purpose Transmission Dip Probes. `<URL: http:`

`//www.oceanoptics.com/products/t300t200transdipprobes.asp>`.

Osborne, B.G., T. Fearn & P.H. Hindle (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. 2nd edition. Pearson Education Limited. ISBN 0-470-22128-3.

Pena, William A. (2009). *Optical Imaging of Early Dental Caries in Deciduous Teeth with Near-IR Light at 1310nm*. Ph.D. thesis, University of California, San Francisco.

Phillips, S. (2006). Oral embryology, histology and anatomy. In: *Clinical Textbook of Dental Hygiene and Therapy*, pp. 3–24. Ed.: Robert Ireland. Blackwell Munksgaard. ISBN 978-1-4051-3540-5.

Porat, Boaz (1997). *A Course in Digital Signal Processing*. John Wiley & Sons Inc. ISBN 0-471-14961-6.

Prasad, Paras N. (2003). *Introduction to Biophotonics*. John Wiley & Sons Inc. ISBN 0-471-28770-9.

Press, William H., Saul A. Teukolsky, William T. Vetterling & Brian P. Flannery (1992). *Numerical Recipes in C – The Art of Scientific Computing*. 2nd edition. Cambridge University Press. ISBN 0-521-43108-5.

Saleh, Bahaa E. A. & Malvin Carl Teich (2007). *Fundamentals of Photonics*. 2nd edition. John Wiley & Sons, Inc. ISBN 978-0-471-35832-9.

Scott, Clayton (2007). Performance measures for Neyman–Pearson classification. *IEEE Transactions on Information Theory* 53:8, 2852–2863. doi:10.1109/TIT.2007.901152.

Shafer, William G., Maynard K. Hine & Barnet M. Levy (1974). *A Textbook of Oral Pathology*. 3rd edition. W.B. Saunders Company. ISBN 0-7216-2918-0.

Shawe-Taylor, John & Nello Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press. ISBN 0-521-81397-2.

Soames, J. V. & J. C. Southam (1993). *Oral Pathology*. 2nd edition. Oxford University Press. ISBN 0-19-2622145.

Staninec, Michal, Chulsung Lee, Cynthia L. Darling & Daniel Fried (2010). In vivo near-IR imaging of approximal dental decay at 1310 nm. *Lasers in Surgery and Medicine* 42:4, 292–298. ISSN 1096-9101. doi:10.1002/lsm.20913.

Stösser, Lutz, Marcus Dell, Annerose Borutta & Roswitha Heinrich-Weltzien (2007). Bacterial and enzymatic tests – Parameters of caries–novel tests. In: *Minimally Invasive Dentistry – The Management of Caries*, pp. 15–28. Ed.: Nairn H.F. Wilson. Quintessence Publishing Co, Ltd. ISBN 978-1-85097-105-4.

Suykens, J.A.K. & J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural Processing Letters* 9:3, 293–300. doi:10.1023/A:1018628609742.

Tao, Y.C. & D. Fried (2009). Near-infrared image-guided laser ablation of dental decay. *Journal of Biomedical Optics* 14:5, 054045. doi:10.1117/1.3253390.

Tenovuo, Jorma (2004). Sylki ja suun puolustusmekanismit. In: *Therapia Odontologica – Hammaslääketieteen käsikirja*, pp. 239–244. Ed.: J. Meurman, H. Murtomaa, Y. LeBell & H. Autti. Academica-Kustannus Oy. ISBN 952-5046-04-4.

Theodoridis, Sergios & Konstantinos Koutroumbas (2006). *Pattern Recognition*. 3rd edition. Academic Press. ISBN 978-0-12-369531-4.

Välisuo, Petri (2011). *Photonics Simulation and Modelling in Design of Spectrocutometry*. Ph.D. thesis, University of Vaasa.

Wang, Lihong V. & Hsin-i Wu (2007). *Biomedical Optics – Principles and Imaging*. John Wiley & Sons Inc. ISBN 978-0-471-74304-0.

Weatherell, JA, C. Robinson & AS Hallsworth (1974). Variations in the chemical composition of human enamel. *Journal of Dental Research* 53:2, 180. ISSN 0022-0345. doi:10.1177/00220345740530020501.

Wilson, Nairn & Alphons Plasschaert (2007). Dental caries, minimally invasive dentistry and evidence-based clinical practice. In: *Minimally Invasive Dentistry – The Management of Caries*, pp. 1–6. Ed.: Nairn H.F. Wilson. Quintessence Publishing Co, Ltd. ISBN 978-1-85097-105-4.

Wist, A.O., P. Moon, S.L. Herr & P.P. Fatouros (2009). Rapid communication: Investigation of a new light imaging technique to detect incipient caries in teeth. *Journal of Clinical Laser Medicine & Surgery* 12:3, 165–170. doi:10.1089/clm.1994.12.165.

Wu, J. & D. Fried (2009). High contrast near-infrared polarized reflectance images of demineralization on tooth buccal and occlusal surfaces at $\lambda = 1310$-nm. *Lasers in Surgery and Medicine* 41:3, 208–213. ISSN 1096-9101. doi:10.1002/lsm.20746.

Yaeger, James A. (1976). Enamel. In: *Orban's Oral Histology and Embryology*, pp. 45–105. 8th edition, Ed.: S.N. Bhaskar. The C. V. Mosby Company. ISBN 0-8016-4608-1.

Young, Hugh D. & Roger A. Freedman (2000). *University Physics with Modern Physics*. Pearson Education, Addison Wesley. ISBN 0-8053-8684-X.

Zakian, Christian, Iain Pretty & Roger Ellwood (2009). Near-infared hyperspectral imaging of teeth for dental caries detection. *Journal of Biomedical Optics* 14:6, 064047. doi:10.1117/1.3275480.

Zakian, C.M., A.M. Taylor, R.P. Ellwood & I.A. Pretty (2010). Occlusal caries detection by using thermal imaging. *Journal of Dentistry* 38:10, 788–795. ISSN 0300-5712. doi:10.1016/j.jdent.2010.06.010.

Zito, T., N. Wilbert, L. Wiskott & P. Berkes (2008). Modular toolkit for Data Processing (MDP): A Python data processing framework. *Frontiers in neuroinformatics* 2. doi: 10.3389/neuro.11.008.2008.

APPENDIX. Detailed description of support vector machine

Support vector machine (SVM) is a supervised pattern recognition algorithm. It can be used for regression or classification. When it is used for classification, the resulting algorithm is called a support vector classifier (SVC). (Chang & Lin 2001: 3.) SVC uses a model $\mathcal{M}$ to depict the samples, such that $\vec{\mathbf{x}} = \mathcal{M}(\vec{\mathbf{p}})$, where $\vec{\mathbf{p}} = (p_1, p_2, \ldots, p_d)$ are the model's parameters that correspond to sample $\vec{\mathbf{x}}$. The model is selected such that there is a linear relation between the model's parameters and the sample's class – or at least so that the relation is as linear as possible. (Shawe-Taylor & Cristianini 2004: 16–17, 33, 212–213.) The model must be such that the set of all possible parameter vectors forms a Hilbert space, i.e. a vector space that is an inner product space, separable and complete (Shawe-Taylor & Cristianini 2004: 48–50).

The vector space F that is formed by the parameter vectors is called a feature space. A function that gives the parameters that correspond to a given sample, i.e. $\vec{\mathbf{p}} = \mathcal{M}^{-1}(\vec{\mathbf{x}}) = \phi(\vec{\mathbf{x}})$, is called a feature map or an embedding map. The latter term refers to an idea that the function *embeds* the data into the feature space. (Shawe-Taylor & Cristianini 2004: 27, 33.) The parameter vector $\vec{\mathbf{p}}$ is called the projection of the sample $\vec{\mathbf{x}}$ to the feature space. The problem of finding the optimal prediction function for a linear relation is a well-studied problem that can be solved with quadratic programming or with least squares approximation (Suykens & Vandewalle 1999; Shawe-Taylor & Cristianini 2004: 29). Therefore, the task of building a classification method using SVC is essentially a problem of selecting, or building, the model $\mathcal{M}$ whose parameters have a linear relation with the sample's class. We will later see that the parameters $\vec{\mathbf{p}}$ do not have to be calculated explicitly; we only need the inner products of pairs of parameters, i.e. $\vec{\mathbf{p}}_i \cdot \vec{\mathbf{p}}_j = \vec{\mathbf{p}}_i^T \vec{\mathbf{p}}_j = \langle \vec{\mathbf{p}}_i, \vec{\mathbf{p}}_j \rangle$. Therefore, the model may have an infinite number of parameters, provided that the inner products, which are scalars, can be calculated.

Suppose for now that an appropriate model has been found. Then prediction function

$$g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}) = \vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}) + w_0 = \vec{\mathbf{w}}^T \phi(\vec{\mathbf{x}}) + w_0 \tag{57}$$

with proper vector $\vec{\mathbf{w}}$ and value w_0 presents that relation, and $\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}))$, so that $\operatorname{sign}(0) = 1$, produces optimal predictions for the samples. Variable w_0 may also be denoted as variable b. (Theodoridis & Koutroumbas 2006: 93–103; Shawe-Taylor & Cristianini 2004: 213; Chang & Lin 2001: 4.) Scalar w_0 is called the bias of the classifier (Boser et al 1992). The prediction function $g_{\vec{\mathbf{w}}}$ can be visualized by a hyperplane in the feature space (Fig. 17), where for every point $\vec{\mathbf{p}}$ on the hyperplane,

$$\vec{\mathbf{w}} \cdot \vec{\mathbf{p}} + w_0 = 0. \tag{58}$$

The vector $\vec{\mathbf{w}}$ is the normal vector of the hyperplane. It points towards the class whose label is positive (unit). (Theodoridis & Koutroumbas 2006: 93–94.)

The direction of the normal vector is orthogonal to the hyperplane, or conversely, every direction that is orthogonal to the normal vector is parallel to the hyperplane. Thus, the normal vector defines the orientation of the hyperplane. Value $-w_0$ defines how far the hyperplane is translated from the origo in the direction of vector $\vec{\mathbf{w}}$. The latter statement can be shown by noting that the inner product $\vec{\mathbf{a}} \cdot \vec{\mathbf{b}}$ gives the projection of vector $\vec{\mathbf{a}}$ onto the direction of vector $\vec{\mathbf{b}}$ and by rewriting equation 58 as

$$\vec{\mathbf{w}} \cdot \vec{\mathbf{p}} = -w_0, \tag{59}$$

where Eq. 59 holds for every point $\vec{\mathbf{p}}$ on the hyperplane. Thus, the projection of every point on the hyperplane onto the direction of the hyperplane's normal is $-w_0$, and the distance of the translation is $|w_0|$.

Let $y_i \in \{1, -1\}$ be the (correct) label for sample $\vec{\mathbf{x}}_i$. Then $y_i g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}_i)$ depicts the distance between the hyperplane and the sample such that the distance is positive if, and only if, the sample is on the correct side of the hyperplane, in which case it would be correctly classified. This value can be interpreted as the margin by which the sample is correctly classified. The smallest of such margins in the training set is called the functional margin of the training set. For a (training) set of samples S and a prediction function g the functional margin can be denoted by

$$m(S, g) = \min_{(\vec{\mathbf{x}}_i, y_i) \in S} y_i g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}_i). \tag{60}$$

If $\vec{\mathbf{w}}$ is a unit vector, i.e. $\|\vec{\mathbf{w}}\| = 1$, then the functional margin corresponds to a geometric distance between a sample and the hyperplane, and thus it can be called geometric margin.

The hyperplane is selected such that the margin is as large as possible, because this gives the prediction function maximal generalization performance. This task can be presented as an optimization problem to

$$\begin{aligned} \text{maximize} \quad & m(S,g) && (61) \\ \text{subject to} \quad & y_i(\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}_i) + w_0) \geq m(S,g), \quad i = 1, 2, \ldots, n && (62) \\ & \|\vec{\mathbf{w}}\| = 1. && (63) \end{aligned}$$

The selection of the prediction function is stable (i.e. does not overfit easily) if the margin is large enough, but it is not robust, because a single (possibly noisy) sample in the training set can substantially change the resulting prediction function. (Shawe-Taylor & Cristianini 2004: 102, 212–213.)

If it is possible to select the hyperplane such that it classifies all samples in the training set correctly, i.e. makes the prediction function consistent, then the functional margin $m(S,g)$ is greater than zero, the hyperplane separates the two classes of samples – or more precisely, the two classes of projections of the samples to the feature space ($\vec{\mathbf{p}} = \phi(\vec{\mathbf{x}})$) – from each other, and the classes are said to be separable. In that case, the samples whose projections $\phi(\vec{\mathbf{x}})$ are the closest to the hyperplane are called support vectors. For all support vectors $\vec{\mathbf{x}}^*$, $yg_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}^*) = m(S,g)$. We can draw a hyperplane that includes the support vectors of one of the two classes (Fig. 17). Both such hyperplanes are parallel to the decision function's hyperplane, and the distances from those hyperplanes to the decision function's hyperplane are identical. Sometimes the margin is defined as the distance between those two hyperplanes. Because the margin is the same for both classes, this would effectively only add a constant factor of two to the value of the margin. This would appear in the definition of the problem of selecting the optimal vector $\vec{\mathbf{w}}$ for the prediction function as a factor of $1/2$, but it would not change the result. (Theodoridis & Koutroumbas 2006: 94–97.)

Traditionally, the following observations are made to refine the problem of selecting the hyperplane. (Shawe-Taylor & Cristianini (2004: 213–214) present a more direct approach.) The distance between a sample's projection $\phi(\vec{\mathbf{x}})$ and the hyperplane can be calculated by projecting the vector $\phi(\vec{\mathbf{x}})$ to the hyperplane, i.e.

$$z = \frac{|\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}) + w_0|}{||\vec{\mathbf{w}}||} = \frac{|g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})|}{||\vec{\mathbf{w}}||}. \tag{64}$$

The vector $\vec{\mathbf{w}}$ can be scaled without affecting the hyperplane. Therefore, when the classes are separable, the vector $\vec{\mathbf{w}}$ can be scaled such that the prediction becomes either positive unit or negative unit for the support vectors, i.e. so that $g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}^*) = y$ for the support vectors, where y is the label of the sample's class and $y \in \{1, -1\}$. As a result, $m(S, g) = 1$. Then the task is to select the hyperplane such that the length of the vector $\vec{\mathbf{w}}$ is minimal and all samples (of the training set) are correctly classified, since that corresponds to the largest margin before scaling $\vec{\mathbf{w}}$ (we can see in Eq. 64 that $|g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})|$ and $\vec{\mathbf{w}}$ are inversely proportional). This task can be presented as an optimization problem to

$$\text{minimize} \quad J(\vec{\mathbf{w}}, w_0) = ||\vec{\mathbf{w}}||^2 \tag{65}$$

$$\text{subject to} \quad y_i(\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}_i) + w_0) \geq 1, \quad i = 1, 2, \ldots, n. \tag{66}$$

If the margin is defined as the distance between the two hyperplanes which are defined by the support vectors of each class, then $||\vec{\mathbf{w}}||^2$ becomes $\frac{1}{2}||\vec{\mathbf{w}}||^2$ in equation 65. (Theodoridis & Koutroumbas 2006: 93–95.)

The training set – or more precisely its projection to the feature space – may be such, that it is impossible to select a hyperplane for the decision function that would classify all samples (of the training set) correctly. In that case, the functional margin $m(S, g)$ is negative and the classes are said to be nonseparable. Then each sample is given a new variable called the slack variable ξ_i, which depicts the distance between the decision function's hyperplane and the sample such that the distance is positive if, and only if, the sample is on the *wrong* side of the hyperplane (in which case it would be incorrectly classified). This value can be interpreted as the margin by which the sample is *incorrectly* classified. If the sample is correctly classified, then the value of its slack variable is zero. We will reformulate the task of selecting the optimal hyperplane for the decision function such that it includes the cases where the classes are nonseparable. The goal is to select the hyperplane such that the length of the vector $\vec{\mathbf{w}}$ is minimal and the slack variables are as small as possible, while the ideal functional margin would be unit. (Theodoridis &

Koutroumbas 2006: 98–99.) Minimizing the length of the vector $\vec{\mathbf{w}}$ and the values of the slack variables are contradictory objectives: ultimately, the quality of the solution can be advanced on one of these objectives only by reducing its quality on the other objective. We may approach this kind of optimization task either by Pareto optimization, or by assigning each objective a (relative) weight (Lampinen 2000: 15). In SVC the normal approach is to assign weights to the objectives. The weight may be assigned to the slack variables as constant C (Theodoridis & Koutroumbas 2006: 100) or to the length of the vector $\vec{\mathbf{w}}$ as constant λ (Shawe-Taylor & Cristianini 2004: 31). Here we use the former notation.

The task of selecting the decision function's hyperplane can now be presented as

$$\text{minimize} \quad J(\vec{\mathbf{w}}, w_0, \vec{\xi}) = \|\vec{\mathbf{w}}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{67}$$

$$\text{subject to} \quad y_i(\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}_i) + w_0) \geq 1 - \xi_i, \quad i = 1, 2, \dots, n \tag{68}$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, n \tag{69}$$

(Theodoridis & Koutroumbas 2006: 100). We can rewrite this problem in another form as

$$\text{minimize} \quad f_0(\vec{\mathbf{w}}, w_0, \vec{\xi}) = \|\vec{\mathbf{w}}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{70}$$

$$\text{subject to} \quad f_i(\vec{\mathbf{w}}, w_0, \xi_i, \vec{x}_i, y_i) = 1 - \xi_i - y_i(\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}_i) + w_0) \leq 0, \quad i = 1, 2, \dots, n \tag{71}$$

$$f_{i+n}(\vec{\mathbf{w}}, w_0, \xi_i, \vec{x}_i, y_i) = -\xi_i \leq 0, \quad i = 1, 2, \dots, n. \tag{72}$$

The standard form of an optimization problem is

$$\text{minimize} \quad f_0(\vec{\mathbf{w}}) \tag{73}$$

$$\text{subject to} \quad f_i(\vec{\mathbf{w}}) \leq 0, \quad i = 1, 2, \dots, n \tag{74}$$

$$h_i(\vec{\mathbf{w}}) = 0, \quad i = 1, 2, \dots, p. \tag{75}$$

The function f_0 is called the cost function or the objective function, vector $\vec{\mathbf{w}}$ is called the optimization variable, and the value of $f_0(\vec{\mathbf{w}})$ is called the objective value of $\vec{\mathbf{w}}$. The functions f_i are called the constraint functions. The set of vectors $\vec{\mathbf{w}}$ for which the cost function and the constraint functions are defined is called the domain of the optimization problem. Let $\operatorname{dom} f_i$ be the domain of function f_i, i.e. the set of optimization variables for which the function is defined. Then the domain of the optimization problem is

$$\mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom} f_i. \tag{76}$$

Vector $\vec{\mathbf{w}}$ (a candidate solution) is feasible if it satisfies the constraints (otherwise it is unfeasible). The set of feasible vectors is the feasible set, which is a subset of the domain. The problem is called feasible if, and only if, there is at least one feasible candidate solution for it. Vector $\vec{\mathbf{w}}^{\star}$ is called optimal, or solution of the optimization problem, if, and only if, $f_0(\vec{\mathbf{w}}^{\star}) \leq f_0(\vec{\mathbf{w}})$ for every feasible vector $\vec{\mathbf{w}}$. If the problem is unfeasible, then $f_0(\vec{\mathbf{w}}^{\star}) = \infty$. If the objective value can be made arbitrarily small by selecting an appropriate feasible solution candidate, then $f_0(\vec{\mathbf{w}}^{\star}) = -\infty$ and the problem is said to be unbounded below. (Boyd & Vandenberghe 2004: 1, 127–128.) Since the model $\mathcal{M}$ that is used to depict the samples is required to be such that the optimization variables $\vec{\mathbf{w}}$ (points in the feature space) form a complete vector space with an inner product, the domain in Eq. 70 is $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$. We will see that the optimal prediction function can be selected by using the saddle point of the Lagrange dual function of the optimization problem. We'll first look at how it works in the standard form, and then apply it to the problem of selecting the optimal hyperplane for the decision function.

This kind of problem can be represented by its Lagrange dual function as

$$L^*(\vec{\lambda}, \vec{\nu}) = \inf_{\vec{\mathbf{w}} \in \mathcal{D}} L(\vec{\mathbf{w}}, \vec{\lambda}, \vec{\nu}) = \inf_{\vec{\mathbf{w}} \in \mathcal{D}} \left(f_0(\vec{\mathbf{w}}) + \sum_{i=1}^{n} \lambda_i f_i(\vec{\mathbf{w}}) + \sum_{i=1}^{n} \nu_i h_i(\vec{\mathbf{w}}) \right), \tag{77}$$

where $L^* : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ (Boyd & Vandenberghe 2004: 215–216). Infimum is defined as

$$\inf_{\vec{\mathbf{w}} \in \mathcal{D}} f(\vec{\mathbf{w}}) = \max \{ \, \xi \in \mathbb{R} \mid \forall \vec{\mathbf{w}} \in \mathcal{D} : \xi \leq f(\vec{\mathbf{w}}) \, \}. \tag{78}$$

Function $L(\vec{\mathbf{w}}, \vec{\lambda}, \vec{\nu})$ is called the Lagrangian of the problem. It adds a weighted sum of the constraint functions to the objective function. Scalars λ_i and ν_i are called Lagrange multipliers, and vectors $\vec{\lambda}$ and $\vec{\nu}$ are called the dual variables or Lagrange multiplier vectors. (Boyd & Vandenberghe 2004: 215–216.) Sometimes vector $\vec{\lambda}$ is denoted as $\vec{\alpha}$ (Boser et al 1992). Let $\vec{\mathbf{w}}^\star$ be the solution to the optimization problem presented in equations 73–75. Then

$$L^*(\vec{\lambda},\vec{\nu}) \leq f_0(\vec{\mathbf{w}}^\star), \quad \forall(\vec{\lambda},\vec{\nu}) \in \mathbb{R}^n \times \mathbb{R}^p, \quad \forall i \in \{1,2,\ldots,n\} : \lambda_i \geq 0. \tag{79}$$

In other words, the dual function gives a lower bound for the optimal value with any vectors $\vec{\lambda}$ and $\vec{\nu}$ where all components of $\vec{\lambda}$ are zero or greater, i.e. the optimal value can not be smaller (better) than the value of L^*. The vector pair $(\vec{\lambda},\vec{\nu})$ is called dual feasible if, and only if, $L^*(\vec{\lambda},\vec{\nu}) > -\infty$. (Boyd & Vandenberghe 2004: 215–216.)

In order to get as much information as possible about the optimal value $f_0(\vec{\mathbf{w}}^\star)$ (and the solution $\vec{\mathbf{w}}^\star$), we need to find the maximal value of the dual function. This task can be presented as an optimization problem

$$\begin{aligned} &\text{maximize} \quad L^*(\vec{\lambda},\vec{\nu}) && (80)\\ &\text{subject to} \quad \lambda_i \geq 0, \quad i = 1,2,\ldots,n. && (81)\end{aligned}$$

This problem is called the Lagrange dual problem. In order to distinguish it from the original optimization problem (Eq. 73–75), the original problem is called the primal problem. Let $(\vec{\lambda}^\star,\vec{\nu}^\star)$ be the solution to the dual problem. Then $L^*(\vec{\lambda}^\star,\vec{\nu}^\star)$ is the optimal value of the dual problem. The pair $(\vec{\lambda}^\star,\vec{\nu}^\star)$ is called the dual optimal or optimal Lagrange multipliers. Now

$$L^*(\vec{\lambda}^\star,\vec{\nu}^\star) \leq f_0(\vec{\mathbf{w}}^\star). \tag{82}$$

This property is called weak duality. The difference between the optimal value of the primal problem and that of the dual problem, i.e. $f_0(\vec{\mathbf{w}}^\star) - L^*(\vec{\lambda}^\star,\vec{\nu}^\star)$, is called the optimal duality gap. If the optimal duality gap is zero, i.e. $L^*(\vec{\lambda}^\star,\vec{\nu}^\star) = f_0(\vec{\mathbf{w}}^\star)$, then strong duality holds between the primal problem and the dual problem. This happens only if the primal problem satisfies certain conditions. There are several alternative conditions that are sufficient to establish strong duality. They are called constraint qualifications. For example, strong duality holds if the primal problem is convex and Slater's condition holds. (Boyd & Vandenberghe 2004: 223–227, 234–236.)

An optimization problem is convex if, and only if, the cost function and the constraint functions are convex, i.e.

$$\begin{aligned}&\forall\vec{\mathbf{w}},\vec{\mathbf{w}}'\in\mathbb{R}^n,\forall\alpha,\beta\in\mathbb{R}:\alpha\geq0,\beta\geq0,\alpha+\beta=1\rightarrow\\&f_i(\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}')\leq\alpha f_i(\vec{\mathbf{w}})+\beta f_i(\vec{\mathbf{w}}').\end{aligned}\tag{83}$$

Slater's condition states that there exists a strictly feasible solution candidate $\vec{\mathbf{w}}$ in the relative interior of the domain, i.e.

$$\exists\vec{\mathbf{w}}\in\operatorname{relint}\mathcal{D}:f_i(\vec{\mathbf{w}})<0,i=1,2,\ldots,n\wedge h_i(\vec{\mathbf{w}})=0,i=1,2,\ldots,p.\tag{84}$$

Relative interior of the domain refers to the interior relative to the affine hull of the domain, i.e.

$$\operatorname{relint}\mathcal{D}=\{\vec{\mathbf{w}}\in\mathcal{D}\mid\exists r\in\mathbb{R}:r>0,(B(\vec{\mathbf{w}},r)\cap\operatorname{aff}\mathcal{D})\subseteq\mathcal{D}\},\tag{85}$$

where the affine hull of the domain is

$$\operatorname{aff}\mathcal{D}=\left\{\sum_{i=1}^{k}\theta_i\vec{\mathbf{w}}_i\,\middle|\,\forall i\in\{1,2,\ldots,k\}:\vec{\mathbf{w}}_i\in\mathcal{D},\sum_{i=1}^{k}\theta_i=1\right\}\tag{86}$$

and $B(\vec{\mathbf{w}},r)$ is a ball with radius r centered at $\vec{\mathbf{w}}$, i.e. $B(\vec{\mathbf{w}},r)=\{\vec{\mathbf{v}}\in\mathcal{D}\mid\|\vec{\mathbf{v}}-\vec{\mathbf{w}}\|<r\}$. (Boyd & Vandenberghe 2004: 1–2, 21–23, 226–227.)

The objective function f_0 (Eq. 70) is convex if, and only if,

$$\begin{aligned}&\forall(\vec{\mathbf{w}},w_0,\vec{\xi}),(\vec{\mathbf{w}}',w_0',\vec{\xi}')\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^n,\\&\forall\alpha,\beta\in\mathbb{R}:\alpha\geq0,\beta\geq0,\alpha+\beta=1\rightarrow\\&\|\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}'\|^2+C\sum_{i=1}^{n}(\alpha\xi_i+\beta\xi_i')\leq\\&\alpha\left(\|\vec{\mathbf{w}}\|^2+C\sum_{i=1}^{n}\xi_i\right)+\beta\left(\|\vec{\mathbf{w}}'\|^2+C\sum_{i=1}^{n}\xi_i'\right).\end{aligned}\tag{87}$$

The value of a norm is always zero or greater (Greenberg 1998: 434). Therefore, there is a one-to-one correspondence between the norm and its square. Thus, we can show that the condition holds by showing that it holds when the squaring of the norm is omitted.

This inequality can now be written in the form

$$\begin{aligned}&\|\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}'\|+C\sum_{i=1}^{n}(\alpha\xi_i+\beta\xi_i')\leq\\&\alpha\|\vec{\mathbf{w}}\|+\beta\|\vec{\mathbf{w}}'\|+C\sum_{i=1}^{n}(\alpha\xi_i+\beta\xi_i')\end{aligned}\tag{88}$$

$$\Leftrightarrow\quad\|\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}'\|\leq\alpha\|\vec{\mathbf{w}}\|+\beta\|\vec{\mathbf{w}}'\|.\tag{89}$$

For all norms, $\|\alpha\vec{\mathbf{w}}\|=|\alpha|\|\vec{\mathbf{w}}\|$ (Greenberg 1998: 434). Because $\alpha\geq 0,\beta\geq 0$, we can write the inequality in form

$$\|\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}'\|\leq\|\alpha\vec{\mathbf{w}}\|+\|\beta\vec{\mathbf{w}}'\|.\tag{90}$$

This inequality corresponds to the triangle inequality, which holds for all norms (Greenberg 1998: 425). Thus, the objective function is convex. The constraint functions f_i (Eq. 71) are convex if, and only if,

$$\begin{aligned}&\forall(\vec{\mathbf{w}},w_0,\vec{\xi}),(\vec{\mathbf{w}}',w_0',\vec{\xi'})\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^n,\\&\forall\alpha,\beta\in\mathbb{R},\forall i\in\{1,2,\ldots,n\,\}:\alpha\geq 0,\beta\geq 0,\alpha+\beta=1\rightarrow\\&1-(\alpha\xi_i+\beta\xi_i')-y_i((\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}')\cdot\phi(\vec{\mathbf{x}}_i)+(\alpha w_0+\beta w_0'))\leq\\&\alpha(1-\xi_i-y_i(\vec{\mathbf{w}}\cdot\phi(\vec{\mathbf{x}}_i)+w_0))+\beta(1-\xi_i-y_i(\vec{\mathbf{v}}\cdot\phi(\vec{\mathbf{x}}_i)+w_0)).\end{aligned}\tag{91}$$

We can rewrite this inequality as

$$\begin{aligned}&1-(\alpha\xi_i+\beta\xi_i')-y_i((\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}')\cdot\phi(\vec{\mathbf{x}}_i)+(\alpha w_0+\beta w_0'))\leq\\&\underbrace{(\alpha+\beta)}_{=1}-(\alpha\xi_i+\beta\xi_i')-\alpha y_i(\vec{\mathbf{w}}\cdot\phi(\vec{\mathbf{x}}_i)+w_0)-\beta y_i(\vec{\mathbf{v}}\cdot\phi(\vec{\mathbf{x}}_i)+w_0)\end{aligned}\tag{92}$$

$$\begin{aligned}\Leftrightarrow\quad&(\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{v}})\cdot\phi(\vec{\mathbf{x}}_i)+(\alpha w_0+\beta w_0')\leq\\&\alpha(\vec{\mathbf{w}}\cdot\phi(\vec{\mathbf{x}}_i)+w_0)+\beta(\vec{\mathbf{v}}\cdot\phi(\vec{\mathbf{x}}_i)+w_0)\end{aligned}\tag{93}$$

$$\Leftrightarrow\quad(\alpha\vec{\mathbf{w}}+\beta\vec{\mathbf{w}}')\cdot\phi(\vec{\mathbf{x}}_i)\leq\alpha\vec{\mathbf{w}}\cdot\phi(\vec{\mathbf{x}}_i)+\beta\vec{\mathbf{w}}'\cdot\phi(\vec{\mathbf{x}}_i)\tag{94}$$

Now the inequality corresponds to a property (namely linearity) which holds for all inner products (Greenberg 1998: 434). Thus, the constraint functions, and the optimization problem as a whole, are convex. Since the domain of the optimization problem is $\mathcal{D}=\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^n$, aff $\mathcal{D}=\mathcal{D}$ because all linear combinations of points in that space are in that space, and $\operatorname{relint} \mathcal{D} = \mathcal{D}$ because every point in that space is surrounded by a ball (of any positive radius) of points which are also in that space. Thus, Slater's condition holds if, and only if, a strictly feasible solution candidate $\vec{\mathbf{w}}$ exists in the domain. Such a candidate exists when the classes are separable, i.e. when the functional margin $m(S, g)$ is greater than zero, or when the slack variables are given large enough values. Thus, strong duality holds for the problem of selecting the optimal hyperplane for the prediction function (at least) if the slack variables are given large enough values.

Since the problem of selecting the optimal prediction function does not contain equality constraints (Eq. 75), we omit equality constraints from the representation of the optimization problem. If strong duality holds between the primal problem and the dual problem, then there is a saddle point $\vec{\mathbf{w}}$ of the Lagrangian, such that

$$\underbrace{\sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} \inf_{\vec{\mathbf{w}} \in \mathcal{D}} L(\vec{\mathbf{w}}, \vec{\lambda})}_{L^*(\vec{\lambda}^\star, \vec{\nu}^\star)} = \underbrace{\inf_{\vec{\mathbf{w}} \in \mathcal{D}} \sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} L(\vec{\mathbf{w}}, \vec{\lambda})}_{f_0(\vec{\mathbf{w}}^\star)} . \tag{95}$$

In other words, we arrive at the same solution candidate $\vec{\mathbf{w}}$ regardless of whether we evaluate the infimum before the supremum or the supremum before the infimum (strong duality is a sufficient condition for the existence of a saddle point, but not necessarily a necessary condition). The weak duality can be represented as

$$\sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} \inf_{\vec{\mathbf{w}}} L(\vec{\mathbf{w}}, \vec{\lambda}) \leq \inf_{\vec{\mathbf{w}}} \sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} L(\vec{\mathbf{w}}, \vec{\lambda}). \tag{96}$$

Next we will see that the left hand side (LHS) of equations 95 and 96 correspond to the dual problem (L^*) and that the right hand side (RHS) of those equations correspond to the primal problem (f_0). Notice that the dual function was defined in equation 77 as $L^*(\vec{\lambda}, \vec{\nu}) = \inf_{\vec{\mathbf{w}} \in \mathcal{D}} L(\vec{\mathbf{w}}, \vec{\lambda})$. Now the solution $L^*(\vec{\lambda}^\star, \vec{\nu}^\star) = \sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} \inf_{\vec{\mathbf{w}} \in \mathcal{D}} L(\vec{\mathbf{w}}, \vec{\lambda})$, which follows from the definition of the dual problem (Eq. 80 and 81). Notice also that the Lagrangian was defined in equation 77 as $L(\vec{\mathbf{w}}, \vec{\lambda}) = f_0(\vec{\mathbf{w}}) + \sum_{i=1}^{n} \lambda_i f_i(\vec{\mathbf{w}})$. If solution candidate $\vec{\mathbf{w}}$ is unfeasible, then $\exists i : f_i(\vec{\mathbf{w}}) > 0$. In that case, $\sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} L(\vec{\mathbf{w}}, \vec{\lambda}) = \infty$ by letting $\lambda_i \to \infty$ and letting the other Lagrange multipliers λ_j be zero. If solution candidate $\vec{\mathbf{w}}$ is feasible, then $\forall i : f_i(\vec{\mathbf{w}}) \leq 0$ and $\sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} L(\vec{\mathbf{w}}, \vec{\lambda}) = f_0(\vec{\mathbf{w}})$ by letting all Lagrange multipliers be zero, which follows from the defition of $L(\vec{\mathbf{w}}, \vec{\lambda})$. Thus, $f_0(\vec{\mathbf{w}}^\star) = \inf_{\vec{\mathbf{w}} \in \mathcal{D}} \sup_{\vec{\lambda} \geq \vec{\mathbf{0}}} L(\vec{\mathbf{w}}, \vec{\lambda})$. (Boyd & Vandenberghe 2004: 237–238.)

If the objective function and the constraint functions are differentiable, and strong duality holds between the primal and the dual problem, then the Karush-Kuhn-Tucker (KKT) conditions hold for every pair of primal and dual optimal points. The KKT conditions are

$$\vec{\mathbf{0}} = \frac{\partial L(\vec{\mathbf{w}}^\star, \vec{\lambda}^\star)}{\partial \vec{\mathbf{w}}} = \nabla f_0(\vec{\mathbf{w}}^\star) + \sum_{i=1}^{n} \lambda_i^\star \nabla f_i(\vec{\mathbf{w}}^\star), \tag{97}$$

$$\lambda_i^\star \geq 0, \quad i = 1, 2, \ldots, n \quad \text{and} \tag{98}$$

$$\lambda_i^\star f_i(\vec{\mathbf{w}}^\star) = 0, \quad i = 1, 2, \ldots, n. \tag{99}$$

Equation 97 states that at the optimal $\vec{\mathbf{w}}^\star$ the derivative of the Lagrangian over the optimization variable $\vec{\mathbf{w}}$ is zero. The condition in Eq. 99 is called complementary slackness, and it implies that $\forall i : \lambda_i^\star = 0 \vee f_i(\vec{\mathbf{w}}^\star) = 0$, which follows from Eq. 97 and 98, and from the constraint $\forall i : f_i(\vec{\mathbf{w}}^\star) \leq 0$ in Eq. 74. (Boyd & Vandenberghe 2004: 242–243.) If for each i either one of the terms is zero and the other is non-zero, i.e. $\forall i : (\lambda_i^\star = 0 \vee f_i(\vec{\mathbf{w}}^\star) = 0) \wedge (\lambda_i^\star \neq f_i(\vec{\mathbf{w}}^\star))$, then strict complementarity holds. If $\lambda_i = 0$, then the value of the constraint function $f_i(\vec{\mathbf{w}})$ does not affect the value of the Lagrangian $L(\vec{\mathbf{w}}, \vec{\lambda})$, and thus f_i is called inactive. Constraint functions for which $f_i(\vec{\mathbf{w}}) = 0$ are called active constraints. (Theodoridis & Koutroumbas 2006: 811–812.)

If the optimization problem is convex – like the problem of selecting the optimal hyperplane for the decision function – then the KKT conditions are not only necessary, but also sufficient for the point that satisfies them to be a solution. In other words, every pair of optimization variable $\vec{\mathbf{w}}$ and dual variable $\vec{\lambda}$ that satisfy the KKT conditions are solutions to the optimization problem. (Boyd & Vandenberghe 2004: 244.)

We can now represent the problem of selecting the optimal hyperplane for the decision function by using the Lagrange dual problem with the additional constraint set by the KKT condition. The Lagrangian of the problem is

$$\begin{aligned} L(\vec{\mathbf{w}}, w_0, \vec{\xi}, \vec{\lambda}, \vec{\mu}) = \|\vec{\mathbf{w}}\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \mu_i(-\xi_i) \\ + \sum_{i=1}^{n} \lambda_i (1 - \xi_i - y_i(\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}_i) + w_0)). \end{aligned} \tag{100}$$

(Theodoridis & Koutroumbas 2006: 100.) The additional term associated with the dual variable $\vec{\mu}$ comes from the additional constraint functions in Eq. 72. Sometimes the dual variable $\vec{\mu}$ is denoted by $\vec{\nu}$ (Suykens & Vandewalle 1999).

The primal problem is convex, and thus every point $(\vec{\mathbf{w}}^*, w_0^*, \vec{\xi}^*, \vec{\lambda}^*, \vec{\mu}^*)$ that satisfies the KKT conditions is a solution to the dual problem, i.e. $(\vec{\mathbf{w}}^*, w_0^*, \vec{\xi}^*, \vec{\lambda}^*, \vec{\mu}^*) = (\vec{\mathbf{w}}^\star, w_0^\star, \vec{\xi}^\star, \vec{\lambda}^\star, \vec{\mu}^\star)$, and (by strong duality) yields also a solution to the primal problem.

The KKT conditions for this problem are

$$\frac{\partial L}{\partial \vec{\mathbf{w}}} = \vec{\mathbf{0}} \Leftrightarrow \vec{\mathbf{w}} = \sum_{i=1}^{n} \lambda_i y_i \phi(\vec{\mathbf{x}}_i) \tag{101}$$

$$\frac{\partial L}{\partial w_0} = 0 \Leftrightarrow \sum_{i=1}^{n} \lambda_i y_i = 0 \tag{102}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Leftrightarrow C - \mu_i - \lambda_i = 0, \quad i = 1, 2, \ldots, n \tag{103}$$

$$\mu_i \geq 0, \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, n \tag{104}$$

$$\lambda_i \underbrace{(1 - \xi_i - y_i(\vec{\mathbf{w}} \cdot \phi(\vec{\mathbf{x}}_i) + w_0))}_{f_i(\vec{\mathbf{w}}, w_0, \xi_i, \vec{\mathbf{x}}_i, y_i)} = 0, \quad i = 1, 2, \ldots, n \tag{105}$$

$$\mu_i \xi_i = 0, \quad i = 1, 2, \ldots, n \tag{106}$$

For points where $\xi_i > 0$, $\mu_i = 0$ by condition 106, and thus $\lambda_i = C$ by condition 103. (Theodoridis & Koutroumbas 2006: 101.)

The optimization problem can now be presented in the Wolfe dual representation as

$$\text{maximize} \quad L(\vec{\mathbf{w}}, w_0, \vec{\xi}, \vec{\lambda}, \vec{\mu}) \tag{107}$$

$$\text{subject to} \quad \vec{\mathbf{w}} = \sum_{i=1}^{n} \lambda_i y_i \phi(\vec{\mathbf{x}}_i) \tag{108}$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0 \tag{109}$$

$$C - \mu_i - \lambda_i = 0, \quad i = 1, 2, \ldots, n \tag{110}$$

$$\lambda_i \geq 0, \mu_i \geq 0, \quad i = 1, 2, \ldots, n. \tag{111}$$

(Theodoridis & Koutroumbas 2006: 100–101.) The maximization of L in this optimization problem corresponds to the condition $\lambda_i f_i = 0$ in Eq. 105, because $\forall i : \lambda_i \geq 0$ and

$f_i \leq 0$. It also ensures that condition $\mu_i \xi_i = 0$ in Eq. 106 holds, because the maximization requires that $\sum \mu_i(-\xi_i)$ is maximized, which occurs when $\mu_i \xi_i = 0$ because $\forall i : \mu_i \geq 0$ and $\xi_i \geq 0$. The condition in Eq. 108 ensures that $\vec{\mathbf{w}} = \vec{\mathbf{w}}^\star$. Thus, this optimization problem corresponds to

$$\sup_{\vec{\lambda},\vec{\mu}} \inf_{\vec{\mathbf{w}},w_0,\vec{\xi}} L(\vec{\mathbf{w}}, w_0, \vec{\xi}, \vec{\lambda}, \vec{\mu}). \tag{112}$$

Since strong duality holds, this is a saddle point, and thus equal to $f_0(\vec{\mathbf{w}}^\star, w_0^\star, \xi_i^\star)$.

By substituting the equality conditions in Eq. 108–111 into the Lagrangian, the optimization problem becomes

$$\text{maximize} \quad \left(\sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j (\phi(\vec{\mathbf{x}}_i) \cdot \phi(\vec{\mathbf{x}}_j)) \right) \tag{113}$$

$$\text{subject to} \quad 0 \leq \lambda_i \leq C, \quad i = 1, 2, \ldots, n \tag{114}$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0 \tag{115}$$

(Theodoridis & Koutroumbas 2006: 101). The only variable whose value we can select in this optimization problem is $\vec{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. The solution to this optimization problem is $\vec{\lambda}^\star$. With Eq. 108 we can write the decision function in the form

$$g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}) = \vec{\mathbf{w}}^\star \cdot \phi(\vec{\mathbf{x}}) + w_0 = \left(\sum_{i=1}^{n} \lambda_i^\star y_i \phi(\vec{\mathbf{x}}_i) \right) \cdot \phi(\vec{\mathbf{x}}) + w_0 \tag{116}$$

(Boser et al 1992). Every inner product has a property called linearity, which states that $(\alpha \vec{\mathbf{w}}_1 + \beta \vec{\mathbf{w}}_2) \cdot \vec{\mathbf{u}} = \alpha(\vec{\mathbf{w}}_1 \cdot \vec{\mathbf{u}}) + \beta(\vec{\mathbf{w}}_2 \cdot \vec{\mathbf{u}})$, where α and β are scalars and $\vec{\mathbf{w}}_1$, $\vec{\mathbf{w}}_2$ and $\vec{\mathbf{u}}$ are vectors in a given vector space (Greenberg 1998: 434). Thus, the decision function can be written in the form

$$g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}) = \sum_{i=1}^{n} \lambda_i^\star y_i (\phi(\vec{\mathbf{x}}_i) \cdot \phi(\vec{\mathbf{x}})) + w_0 \tag{117}$$

(Boser et al 1992).

We saw in equation 66 that $y_i g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}_i)$ depicts how well sample $\vec{\mathbf{x}}_i$ of the training set is classified and that this value should be at least an unit. In equation 71 we saw that constraint function f_i depicts the degree to which sample $\vec{\mathbf{x}}_i$ *fails* to fulfil this requirement, when the slack variable ξ_i is considered, and naturally this value should be less or equal to zero. Support vectors are by definition the (projections of) samples that are classified correctly with the least margin. Therefore, and because $\vec{\mathbf{w}}$, w_0 and $\vec{\xi}$ are optimized, $f_i = 0$ for the support vectors, and only for the support vectors, i.e. they fulfil the requirements just barely. By this and the condition in equation 105, the Lagrange multiplier λ_i can be non-zero (greater than zero) only if $\phi(\vec{\mathbf{x}}_i)$ is a support vector. Particularly samples for whom $\xi_i > 0$ and thus $\lambda_i = C$ are support vectors. We can now see from Eq. 108 that the vector $\vec{\mathbf{w}}$ is a linear combination of the support vectors $\phi(\vec{\mathbf{x}}_i)$. However, the optimization process seeks to maximize the squared norm of the vector $\vec{\mathbf{w}}$ (as a term in the objective function, Eq. 107). Thus, the optimization process will seek to select large values, particularly values that are greater than zero, for the Lagrange multipliers λ_i that correspond to support vectors.

A function which produces the inner product of the samples' projections to the feature space is called a kernel and denoted as κ or K. In other words,

$$\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \phi(\vec{\mathbf{x}}_i) \cdot \phi(\vec{\mathbf{x}}_j) = \langle \phi(\vec{\mathbf{x}}_i), \phi(\vec{\mathbf{x}}_j) \rangle. \tag{118}$$

When the kernel is used in the definition of the decision function we do not necessarily need to calculate explicitly the samples' projections to the feature space – we only need to be able to the calculate the value of the kernel function, which implicitly contains the projections. (Boser et al 1992; Shawe-Taylor & Cristianini 2004: 34.) In fact, we do not even have to know explicitly what the feature map $\phi(\cdot)$ is. With the kernel the decision function can be written as

$$g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}) = \sum_{i=1}^{n} \lambda_i^{\star} y_i \kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}) + w_0. \tag{119}$$

The prediction is $\operatorname{sign}(g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}))$.

One commonly used kernel is the polynomial kernel or

$$\kappa_d(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j + R)^d = \sum_{s=0}^{d} \binom{d}{s} R^{d-s} (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j)^s. \tag{120}$$

For a given kernel κ_1, derived polynomial kernel is $\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = p(\kappa_1(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$, where $p(\cdot)$ is any polynomial with positive coefficients. Another commonly used kernel is the Gaussian kernel, which is defined as

$$\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp\left(-\frac{\|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\|^2}{2\sigma^2}\right), \quad \sigma > 0. \tag{121}$$

(Shawe-Taylor & Cristianini 2004: 292, 296.) Radial Basis Function (RBF) -kernel is defined as $\kappa(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp(-\gamma|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j|^2)$ (Chang & Lin 2001: 34). It is equal to the Gaussian kernel when $\gamma = 1/(2\sigma^2)$.

In order to maximize the functional margin the value of the bias w_0 is selected such, that the hyperplane lies halfway between the two hyperplanes defined by the support vectors of each class. This value is obtained by calculating how far the two hyperplanes are translated from origo in the direction of the hyperplane (Eq. 59), and by using the average of these two values, i.e.

$$\begin{aligned} w_0^\star &= -\frac{1}{2}(\vec{\mathbf{w}}^\star \cdot \vec{\mathbf{x}}_A + \vec{\mathbf{w}}^\star \cdot \vec{\mathbf{x}}_B) \\ &= -\frac{1}{2}\sum_{i=1}^{n} \lambda_i^\star y_i(\kappa(\vec{\mathbf{x}}_A, \vec{\mathbf{x}}_i) + \kappa(\vec{\mathbf{x}}_B, \vec{\mathbf{x}}_i)) \end{aligned} \tag{122}$$

where $\vec{\mathbf{x}}_A$ is any support vector from the class whose label is unit, and $\vec{\mathbf{x}}_B$ is any support vector from the class whose label is negative unit. (Boser et al 1992.)

For problems with large training sets the optimization problem is solved in parts. First, a subset of the training set is used to select a hyperplane for the decision function. Next, only the samples which became support vectors in that optimization process are kept in the subset. Samples from the training set, which are incorrectly classified by the selected hyperplane by a large enough margin, i.e. $y_i g_{\vec{\mathbf{w}}}(\vec{\mathbf{x}}) < 1-\epsilon$, are added to the subset. A new hyperplane is selected for the decision function based on the new subset of samples. This process is repeated until all samples of the training set are correctly classified. (Boser et al 1992; Theodoridis & Koutroumbas 2006: 102.)

Beside the quadratic programming approach presented here, the problem of selecting the optimal hyperplane for the decision function can also be solved by using least squares approximation. This leads to a variant of SVM called Least squares SVM or LS-SVM. (See, for example, Suykens & Vandewalle 1999 or Shawe-Taylor & Cristianini 2004: 27–32.)