University of Natural Resources and Life Sciences, Vienna

Department of Sustainable Agricultural Systems

Division of Livestock Sciences

Mastitis detection from milk mid-infrared (MIR) spectroscopy in dairy cows

Master Thesis

Lisa Maria Rienesl, BSc

Supervisor:

Univ. Prof. DI. Dr. Johann Sölkner

Co-Supervisors:

Dr. Negar Khayatzadeh, MSc.

DI. Dr. Astrid Köck

Vienna, October 2019

Statutory declaration

I declare that I have prepared, developed and written this thesis independently and I have not used any sources, thoughts or literature of others than clearly stated in the text. The master thesis was not used to award an academic degree at any other university.

Place, date	Signature (Lisa Maria Rienesl, BSc)

Abstract

Mid-infrared (MIR) spectroscopy is the method of choice for the standard milk recording system, to determine milk components including fat, protein, lactose and urea. Since milk composition is related to health and metabolic status of a cow, MIR spectra could be potentially used for disease detection. In dairy production, mastitis is one of the most prevalent diseases. The main aim of this study was to develop a calibration equation to predict mastitis events from routinely recorded MIR spectra data. Further objectives were to evaluate the use of test day somatic cell score (SCS) as covariate and to evaluate different calibration settings, such as sample size and time windows (days between diagnosis and test days). The data for this study are from the Austrian milk recording system and its health monitoring system (GMON). Test day data including MIR spectra data was merged with diagnosis data of Fleckvieh, Brown Swiss and Holstein Friesian cows. As prediction variables, MIR absorbance data after taking first derivatives and selection of wavenumbers, corrected for days in milk, were used. The data set contained roughly 600,000 records and was split into calibration and validation sets by farm. Calibration sets were made to be balanced (as many healthy as mastitis cases), while the validation set was kept large and realistic. Prediction was done with Partial Least Squares Discriminant Analysis, key indicators of model fit were sensitivity and specificity. Results were extracted for association between spectra and diagnosis with different time windows in validation. The comparison of different sets of predictor variables (MIR, SCS, MIR + SCS) showed an advantage in prediction for MIR + SCS. For this prediction model, specificity was 0.79 and sensitivity was 0.68 in time window -7 to +7 days (calibration and validation). Corresponding values for MIR were 0.71 and 0.61, for SCS they were 0.81 and 0.62. In general, prediction of mastitis performed better with a shorter distance between test day and mastitis event. For time windows of -21 to +21 days, sensitivities ranged from 0.50 to 0.57 and specificities remained unchanged (0.71 to 0.85). The comparison of different calibration time windows gave better results for the larger time windows. The evaluation of sample sizes in calibration showed slight advantages for the biggest set. Though, there was no regular up trend in sensitivity with an increasing sample size. Additional research to further improve prediction equation, and studies on heritability and genetic correlations among clinical mastitis, SCS and MIR predicted mastitis are planned.

Key words: *MIR spectroscopy, dairy cow, milk, mastitis, somatic cell count, PLS-DA*

Zusammenfassung

Die Mittlere-Infrarot (MIR) Spektroskopie ist die Methode der Wahl in der routinemäßigen Milchleistungsprüfung zur Bestimmung von Milchbestandteilen wie Fett, Protein, Laktose und Harnstoff. Da die Milchzusammensetzung mit der Gesundheit und dem Stoffwechsel einer Kuh zusammenhängt, besteht die Möglichkeit, MIR-Spektren zur Erkennung gewisser Krankheiten zu verwenden. Auf Milchviehbetrieben ist Mastitis eine der häufigsten Krankheiten bzw. auch Abgangsursachen. Aus diesem Grund ist die Thematik wirtschaftlich und nicht zuletzt hinsichtlich Tierwohl höchst relevant. Das Hauptziel dieser Studie war die Entwicklung einer Kalibrierungsgleichung zur Vorhersage von Mastitisereignissen aus den bei der Milchleistungsprüfung routinemäßig aufgezeichneten MIR-Spektren. Weitere Ziele waren die Evaluierung der Verwendung der Zellzahl (SCS) als Covariable und die Evaluierung verschiedener Kalibrierungseinstellungen, wie Stichprobenumfang und Zeitfenster (Tage zwischen Diagnose und Testtag). Die Daten für diese Studie stammen aus der österreichischen Milchleistungsprüfung und dem Gesundheitsmonitoring (GMON). Zunächst wurden die Testtagsdaten aus Milchleistungsprüfung mit den dazugehörigen MIR-Spektren und den Diagnosedaten aus dem GMON verknüpft. In der Studie waren Kühe der Rassen Fleckvieh, Braunvieh und Holstein inkludiert. Als MIR-Vorhersagevariablen wurden nur selektierte Bereiche des Spektrums verwendet, welche die meiste Information beinhalten. Außerdem wurden die ersten Ableitungen herangezogen und nicht die orignalen Spektrenwerte. Der komplette Datensatz enthielt ungefähr 600.000 Einträge und wurde nach Betriebsnummer zufällig in einen Kalibrierungs- und einen Validierungsdatensatz geteilt. Der Kalibrierungsdatensatz wurde danach hinsichtlich Mastitisfällen und gesunder Tiere balanciert (1:1), der Validierungsdatensatz wurde hingegen unbalanciert und realistisch belassen. Die Vorhersage erfolgte mit der Methode Partial Least Squares Discriminant Analysis (PLS-DA). Indikatoren für die Genauigkeit des Vorhersagemodells waren Sensitivität und Spezifizität. Die Ergebnisse aus der Validierung beziehen sich auf das gesamte Zeitfenster von -21 bis +21 Tagen und wurden zusätzlich für kürzere Zeitfenser extrahiert. Beim Vergleich der verschiedenen Vorhersagevariablen (MIR, SCS, MIR + SCS), konnte die Kombiniation von MIR-Spektren und Zellzahl (MIR + SCS) die besten Ergebnisse erzielen. Bei diesem Modell betrug die Spezifität 0,79 und die Sensitivität 0,68 beim Zeitfenster von -7 bis +7 Tagen (in Kalibrierung und Validierung). Entsprechende Werte für MIR alleine waren 0,71 und 0,61 und für SCS alleine 0,81 und 0,62. Im Allgemeinen funktioniert die Vorhersage bzw. Erkennung einer Mastitis besser, wenn die Abstände zwischen Testtag und Mastitis-Ereignis kürzer sind. Beim größten Zeitfenster in der Validierung (-21 bis +21 Tage) lagen die Sensitiväten in einem Bereich von 0,50 bis 0,57 und die Spezifitäten zwischen 0,71 bis 0,85. Der Vergleich verschiedener Kalibrierungszeitfenster ergab bessere Ergebnisse für die größeren Zeitfenster. Die Analysen bezüglich unterschiedlicher Stichprobenumfänge in der Kalibrierung zeigten leichte Vorteile für den größten Datensatz, wobei kein regelmäßiger Aufwärtstrend in der Sensititiät

mit zunehmendem Stichprobenumfang erkennbar war. Weitere Studien zur Verbesserung der Vorhersagegleichung, sowie zur Heritabilität und den genetischen Korrelationen zwischen klinischer Mastitis, SCS und MIR-vorhergesagter Mastitis sind geplant.

Schlüsselwörter: MIR Spektroskopie, Milckkuh, Milch, Mastitis, Zellzahl, PLS-DA

Acknowledgment

This work was conducted within the COMET-Project D4Dairy (Digitalisation, Data integration, Detection and Decision support in Dairying). That is supported by BMVIT, BMDW and the provinces of Lower Austria and Vienna in the framework of COMET-Competence Centers of Excellent Technologies. The COMET program is handled by the FFG.

The data for this thesis was kindly provided by ZuchtData EDV Dienstleistungen GmbH.

First, I would like to express my special thanks to my supervisor Univ. Prof. Dr. Johann Sölkner for his valuable guidance, patience and continuous support during this thesis. Furthermore, I am very grateful to my Co-Supervisors, Dr. Negar Khayatzadeh for the tremendous support in programming, and Dr. Astrid Köck for the insightful and constructive comments.

Special thanks also to my friends and colleagues for their support and for the good time I have spend with them during my studies.

Finally, I would like to say thank you to my family. Especially to my parents, who have awakened my love to animal breeding, to my sisters for their ongoing and great encouragement and to my lovely granny.

Table of content

Statutory declaration ... 1
Abstract ... 2
Zusammenfassung ... 3
Acknowledgment ... 5
Table of content ... 6
List of tables ... 7
List of figures ... 8
List of abbreviations ... 9
1 Introduction ... 10
1.1 General background ... 10
1.2 Aim of the thesis ... 10
1.3 Literature review ... 11
1.3.1 Milk mid-infrared (MIR) spectroscopy ... 11
1.3.2 Mastitis in dairy cattle ... 12
2 Material and Methods ... 15
2.1 Data ... 15
2.1.1 Preliminary tests on pre-treatment of spectra data ... 16
2.1.2 Data preparation for final model tests ... 17
2.1.3 Calibration and validation settings for final model tests ... 18
2.2 Methodology ... 20
3 Results ... 21
3.1 Results of preliminary tests on different pre-treatments of spectra data ... 21
3.2 Results of final model tests ... 23
3.2.1 Comparison of different predictor variables ... 23
3.2.2 Effect of different time windows in calibration set ... 25
3.2.3 Effect of different sample sizes for calibration set ... 27
4 Discussion ... 29
4.1 Discussion of preliminary tests on different pre-treatments of spectra data ... 29
4.2 Discussion of final model tests ... 30
5 Conclusion ... 37
6 References ... 38

List of tables

Table 1 Number of records of the complete data set 15

Table 2 Datasets for preliminary test on pre-treatment of spectra data 16

Table 3 Results of preliminary tests: Effect of different pre-treatments on spectra data 21

Table 4 Results in calibration for different predictor variables (MIR, SCS or MIR + SCS) 23

Table 5 Results in validation: Effect of different predictor variables (MIR, SCS or MIR + SCS) 24

Table 6 Results in validation: Effect of different time windows in calibration 26

Table 7 Results in validation: Effect of different sample sizes calibration 28

List of figures

Figure 1 Typical milk MIR absorption curve (Source: OptiMIR) 11

Figure 2 Average somatic cell count of cows with acute and chronic mastitis (by Astrid Köck) 13

Figure 3 Causes of losses in Austrian dairy cattle in 2018 (modified after Egger-Danner et al., 2018) 14

Figure 4 Sensitivity and specificity of MIR + SCS 30

Figure 5 Course of sensitivity of MIR, SCS and MIR + SCS with different time windows 31

Figure 6 Balanced accuracies of different calibration time windows 32

Figure 7 Sensitivity of a PLS-DA model as a function of the total sample size (Saccenti & Timmerman, 2016)..... 33

Figure 8 Specificity of a PLS-DA model as a function of the total sample size (Saccenti & Timmerman, 2016)..... 34

Figure 9 Sensitivities for different sample sizes in calibration 35

Figure 10 Specificities for different sample sizes in calibration 35

List of abbreviations

bal.acc	balanced accuracy
der	derivative
DIM	days in milk
GMON	Austrian health monitoring system (German: Gesundheitsmonitoring)
MIR	mid-Infrared
NMR	nuclear-magnetic-resonance
log	logarithmic
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
SCC	somatic cell count
SCS	somatic cell score
sens	sensitivity
spec	specificity
test	validation set
train	calibration set

1 Introduction

1.1 General background

This study is part of the project D4dairy, the overall goal of which is to provide digital support to dairy management by a data-driven, networked information system, exploiting the potential of advanced technologies and data analysis to further improve animal health, nutrition, animal welfare and product quality (D4Dairy Consortium, 2019). A subarea of D4dairy is disease detection using Mid-infrared (MIR) spectral data from milk. MIR spectroscopy is the method of choice for standard milk recording systems to measure milk contents including fat, protein, lactose and urea. Besides, MIR spectra data could be used to predict other milk components (De Marchi et al., 2014). Because it is well known that the composition of milk is related to the health and metabolic status of the cow, its changes can be potential indicators (e. g. Hamann & Krömker, 1997). In recent years, MIR spectra data have been used to predict different variables of interest, as mentioned in section 1.3.1. The focus of this study was on detection of mastitis, which is one of the most prevalent diseases in dairy production (section 1.3.2). MIR spectra analysis could be an extra tool, additionally to somatic cell count (SCC) and veterinarian diagnosis, for mastitis prediction, to further improve genetic evaluation of the trait 'Udder health', or to provide farmers with a management tool.

1.2 Aim of the thesis

The main aim of this study was to develop a calibration equation to predict mastitis events from routinely recorded MIR spectra data. A preliminary objective was to test different pre-treatments of spectra data. Further, we aimed to evaluate the effect of different calibration settings and the use of somatic cell score (SCS) as covariate on the sensitivity and specificity of the prediction model.

1.3 Literature review

1.3.1 Milk mid-infrared (MIR) spectroscopy

The spectroscopic technique is based on the interaction between matter and electromagnetic waves. There are different regions of electromagnetic radiation, which are distinguished according to the wavelengths: x- ray region (0.5-10 nm), UV region (10-350 nm), visible region (35-800 nm), near-infrared region (800-2,500 nm), mid-infrared region (2,500-25,000 nm), microwave region (100 µm-1 cm), and radio frequency region (1 cm-1 m) (De Marchi et al., 2014).

Mid-infrared (MIR) spectroscopy is the method of choice during routine milk recording, for quality control and determination of standard milk contents including fat, protein, lactose and urea. It is a fast and non-destructive method to quantify milk chemical properties (Grelet et al., 2015). The used instrument is called spectrometer, it records the quantity of radiation absorbed in transmittance at specific wavelengths in mid-infrared region. Trough calibration models on representative samples, the spectral data are then transformed into estimates of concentration or other physico-chemical parameters (IDF, 2012). Figure 1 shows a typical MIR absorption curve of a milk sample.

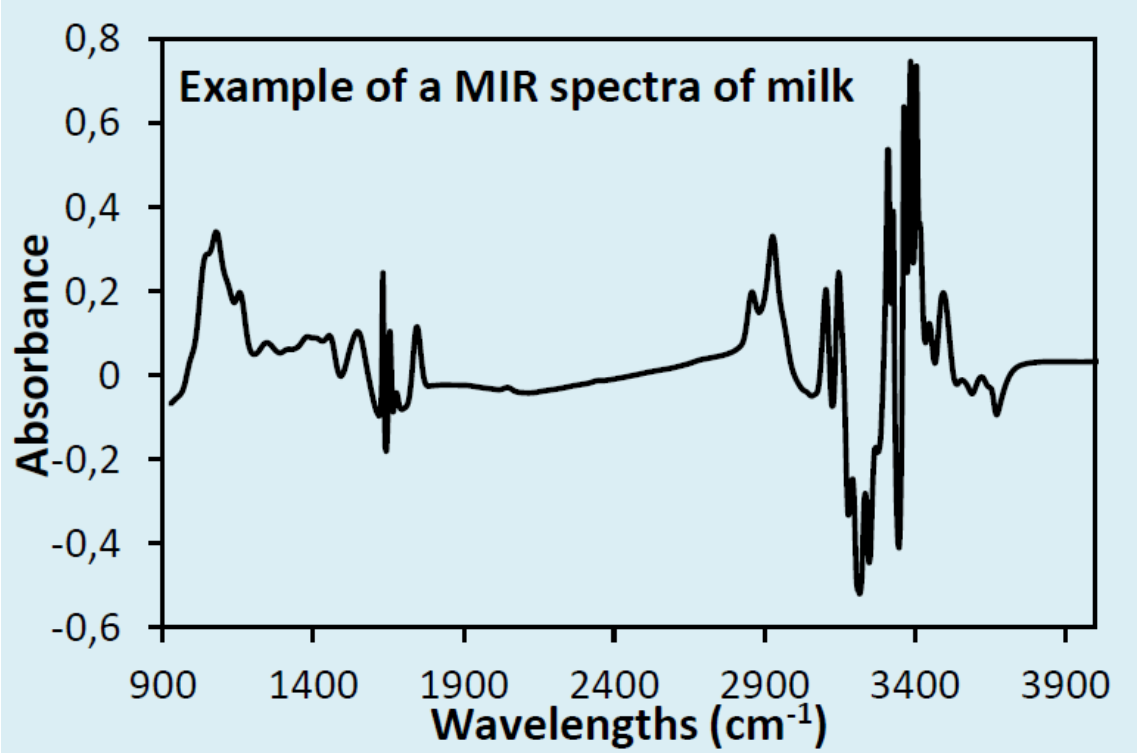

Figure 1 Typical milk MIR absorption curve (Source: OptiMIR)

As MIR spectroscopy is a rapid and cost-effective tool for recording phenotypes on a large scale, it has been used to predict various milk traits and other variables of interest. Visentin et al. (2015) investigated the prediction of milk technical traits, like rennet coagulation time, curd-firming time and curd firmness, which are important factors for cheese production. Further, there are studies on fatty acid composition of milk (Soyeurt et al., 2007; 2011), methane emissions (Vanlierde et al., 2015), feed intake (Wallén et al., 2018), energy intake and efficiency (McParland et al., 2014) or ration composition (Klaffenböck et al., 2017). Since the last decade, fitness and health traits gain more and more importance in breeding programs, not only in Austria, which was a pioneer in that field, also in many other countries in Europe and across the world. Hence, there are several studies on the prediction health traits and diseases with MIR spectroscopy, e.g. subclinical ketosis (De Roos et el., 2007) and clinical ketosis (Belay et al., 2017), mastitis (Soyeurt et al., 2012; Dale & Werner, 2017) and lameness (Mineur et al., 2017).

In animal breeding, not least for genomic selection, accurate and efficient tools to collect phenotypes play a key role (Houle et al., 2010; Pryce et al., 2010). MIR spectroscopy has been evaluated as an appropriate tool for collecting data at the population level for phenotypic and genetic purposes, and it opens many opportunities and a wide research field (De Marchi et al., 2014).

1.3.2 Mastitis in dairy cattle

Bovine mastitis is defined as 'inflammation of the mammary gland' and can have an infectious or non-infectious aetiology. The causes of the disease include pathogens as bacteria, mycoplasmas, yeasts and algae (Blowey & Edmondson, 2000; Bradley, 2002). Mastitis can either occur in a clinical or subclinical form. The symptoms for a clinical mastitis are a visible inflamed quarter changes in the appearance of the milk, which are the cow's inflammatory response to the infection. The subclinical form does not show external changes that indicate the occurrence of mastitis, although the infection is present in the udder (Blowey & Edmondson, 2010). According to their epidemiology, mastitis pathogens can be classified into two types, contagious or environmental (Blowey & Edmondson, 2000; Cervinkova et al., 2013). Contagious pathogens have their primary reservoir in the infected mammary gland and are spread from cow to cow. The main reservoir of environmental pathogens is a contaminated environment, as bedding, soil or manure. Consequently, environmental pathogens are strongly influenced by management practices and hygiene (Garcia, 2004). Typical contagious pathogens are: Streptococcus agalactiae, Staphylococcus aureus and Mycoplasma bovis. The most common environmental pathogens are Echerichia coli, Klebsiella spp, Enterobacter and environmental streptococci (Garcia, 2004; Cervinkova et al., 2013).

Somatic cells are primarily milk-secreting epithelial cells that have been shed from the lining of the udder gland, and white blood cells (leukocytes) that occurred in the mammary gland in response to injury or infection (Dairyman's Digest, 2009). Thus, Somatic cells are related to mastitis, especially to contagious infections (Blowey & Edmondson, 2000). Further, SCC is a useful predictor for infections of mammary gland and can be used to monitor the level or occurrence of mastitis in herds or individual cows (Sharma et al., 2011). Figure 2 displays the average somatic cell count (SCC) of cows with acute and chronic mastitis.

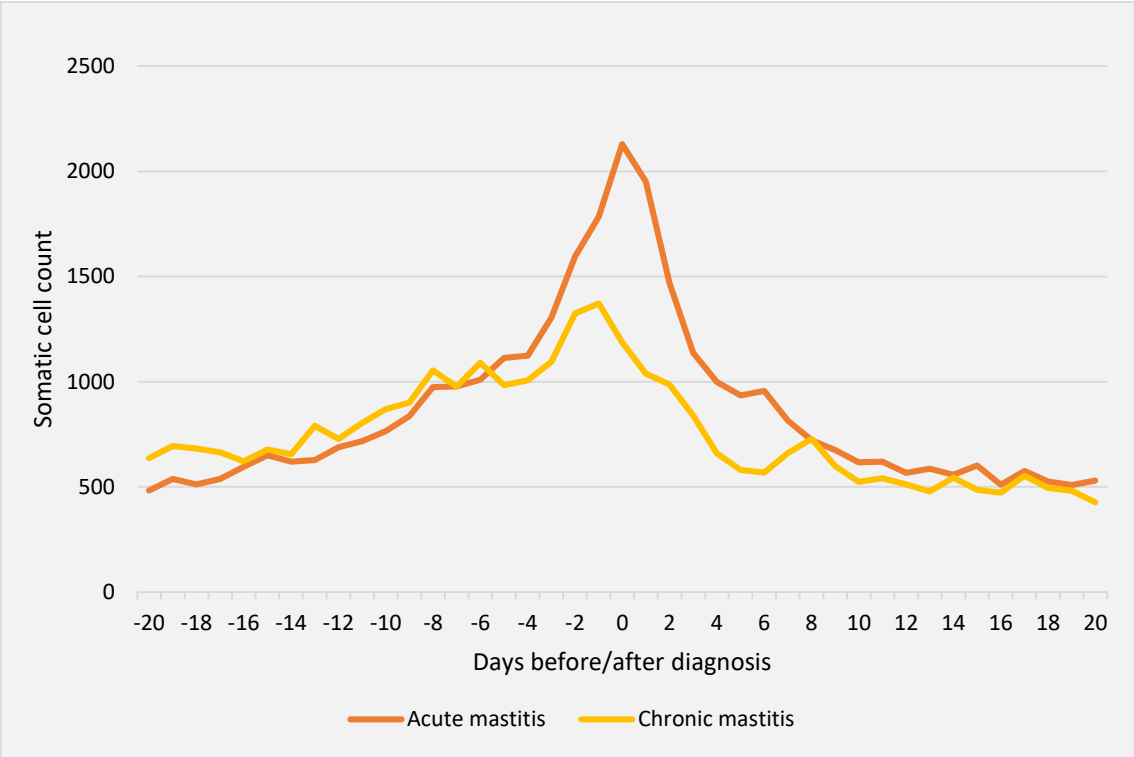

Figure 2 Average somatic cell count of cows with acute and chronic mastitis (slide produced and provided by Astrid Köck)

Mastitis (clinical or subclinical) is one of the most prevalent diseases in dairy production and causes economic harm for farmers and not least, affects animal welfare (Halasa et al., 2007; Sharma et al., 2011; Heikkilä et al., 2011; Guimarães et al., 2017). The economic losses are due to direct and indirect costs, which include: Milk production losses, costs for treatment and drugs, discarded milk, veterinary service, reduced product quality, extra labour, secondary diseases and higher culling and replacement rates (Blowey & Edmondson, 2000; Halasa et al., 2007).

Furthermore, mastitis is also one of the most frequent reasons for animal losses in dairy farms. Figure 3 shows the causes of losses in Austrian dairy cattle in 2018, for the breeds Fleckvieh, Brown Swiss and

Holstein Frisian. Accounting between 12 to 14 %, udder diseases are the third most common cause of losses, after fertility (22-27 %) and sale for breeding (15-16 %).

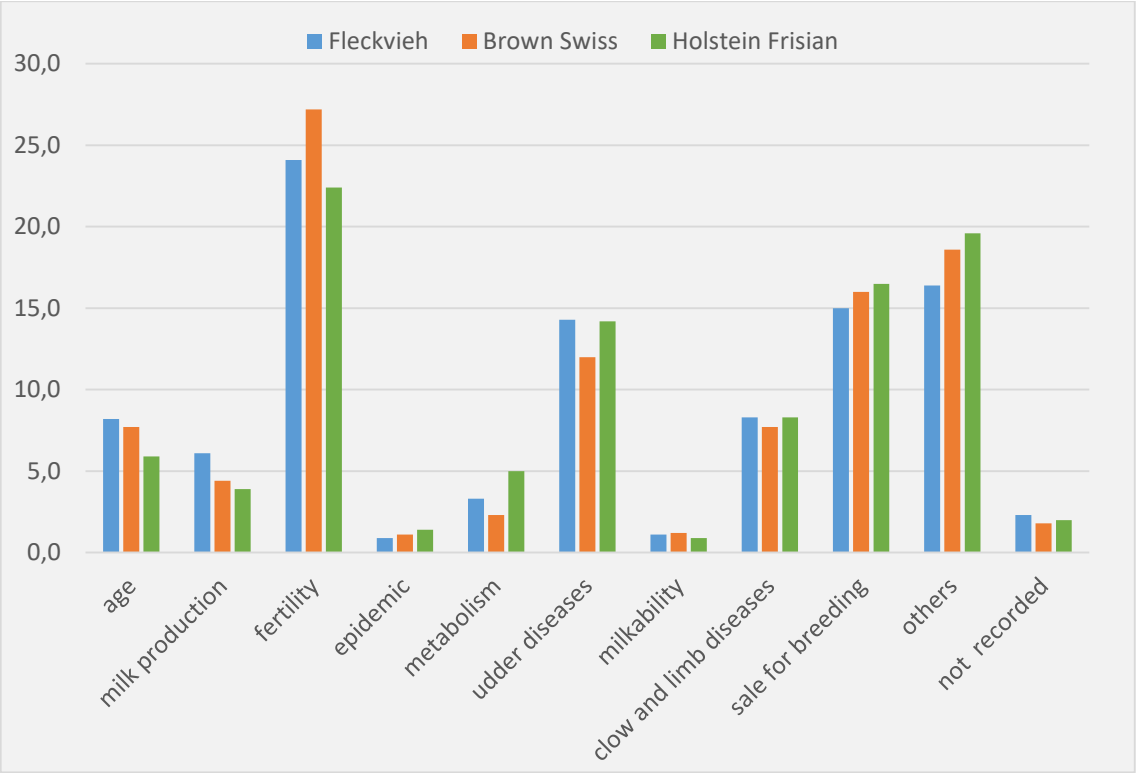

Figure 3 Causes of losses in Austrian dairy cattle in 2018, in %, all lactations (modified after Egger-Danner et al., 2018)

2 Material and Methods

2.1 Data

The data for this study was from the Austrian milk recording system and its health monitoring system (GMON), for the period of July 2014 to December 2018 and was provided by Zuchtdata GmbH. The test day milk data consisted of information on breed, herd, region, calving date, parity, days in milk, milk yield, somatic cell count (SCC), fat, protein and MIR spectra data for the respective test days. The GMON data included recorded mastitis diagnosis for acute and chronic mastitis, which were not distinguished for the prediction model. All data (test day and GMON) used in this study were derived from validated farms with complete disease diagnosis data recording. Test day records of Fleckvieh, Brown Swiss and Holstein Friesian cows between 3 and 305 days of lactation were included. Table 1 shows the number of records of the complete data set.

Table 1 Number of records of the complete data set

Variables	**Records**
Farms	**7,914**
Animals (Cows)	**69,028**
Fleckvieh	52,287
Brown Swiss	7,260
Holstein Friesian	9,481
Test day records	**635,588**
healthy	627,593
mastitis	7,995
acute	5,644
chronic	2,351

MIR spectra consist of 1,060 data points, which are the absorbance values of infrared light at different wavenumbers (925.66 cm^{-1} to 5,010.16 cm^{-1}). MIR spectra from different instruments and different periods were standardized into a common basis (Grelet et al., 2015). According to Grelet et al. (2016), it is recommended to use selected parts of the spectra for the prediction model: 968.1 to 1,577.5 cm^{-1}, 1,731.8 to 1,762.6 cm^{-1}, 1,781.9 to 1,808.9 cm^{-1}, and 2,831.0 to 2,966.0 cm^{-1}. These spectra areas (212 data points) contain most of the information, whilst other areas are less informative, because of strong water absorbance, or not repeatable among MIR instruments. Moreover, in some studies (Soyeurst et al., 2011 and 2012; Grelet et al., 2016; Lainé et al., 2017; Mineur et al., 2017; Ho et al., 2019) first or second derivative of spectra values (Savitzky-Golay method) were taken for the predictions, while in other studies (Visentin et al., 2015; McDermott et al., 2016; Visentin et al., 2016)

untreated spectra values were used, since no improvement was found with applying mathematical pre-treatments. Therefore, a preliminary aim of this thesis was to examine pre-treatments of the spectra data, before starting the main model tests.

2.1.1 Preliminary tests on pre-treatment of spectra data

Merging of the data sets, primary data preparation and pre-treatments of spectra data were done in SAS (SAS Institute Inc., 2017). Data preparation for the preliminary tests was done independently from the one of the final model tests and data sets differed in some settings. Mastitis diagnoses were linked with 'adjacent' milk recording test days. Test day records in the range of 20 days before and 15 days after diagnosis were considered as mastitis cases. For the healthy group, only records of cows without mastitis diagnosis 20 days before and 15 days after test day were used.

In a next step, six individual datasets with different pre-treatments on spectra data were prepared (Table 2). For the data set original untreated spectra data was used. The first derivative of each spectra value was calculated by applying the formula dx(n)=x(n)-x(n+4), and second derivative by using the formula d2x(n)=dx(n)-dx(n+4). Then, of each dataset either the full spectra with 1,060 wavelengths were used, or only 212 selected spectra data points, according to Grelet et al. (2016).

Table 2 Datasets for preliminary test on pre-treatment of spectra data

Dataset	Pre-treatment of spectra data
original full	untreated full spectra (1,060 wavelengths)
original select	untreated selected spectra (212 wavelengths)
1st der. full	1st derivative of full spectra (1,060 wavelengths)
1st der. select	1st derivative of selected spectra (212 wavelengths)
2nd der. full	2nd derivative of full spectra (1,060 wavelengths)
2nd der. select	2nd derivative of selected spectra (212 wavelengths)

Each of these six datasets was further randomly split by farm into half a calibration and a validation set. The calibration set, additionally, got balanced in terms of mastitis and healthy cases (1:1) by using random down sampling, which resulted in a final sample size of 2,098 (1,049 mastitis and 1,049 healthy cases). The validation set was kept unbalanced, but total sample size randomly reduced to roughly 20,000 records, which resulted in a mastitis proportion of mastitis of around 5 %.

A prediction model was run with all datasets, by using the methodology, explained in section 2.2. The accuracy of the different models was then compared by applying a t-test (p-value of 0.05).

2.1.2 Data preparation for final model tests

Merging of the data sets and primary data preparation were again done in SAS (SAS Institute Inc., 2017). Mastitis diagnoses were linked with 'adjacent' milk recording test days. Test day records in the range of 21 days before and 21 days after diagnosis were considered as mastitis cases. For the healthy group, only spectra from cows without mastitis diagnosis 21 days before and 30 days after test day were used.

Based on the results of the preliminary tests on data pre-treatment (section 2.1.1), selected parts of the spectra, according to Grelet et al., (2016), were used for the prediction model. Before selecting the specific areas, first derivative (dx(n)=x(n)-x(n+4)) of full spectra was taken. Even though, the preliminary tests on data treatment did not show clear advantages, we decided to take first derivative of the spectra, pursuant to other relevant studies (Soyeurt et al., 2011; Soyeurt et al., 2012; Grelet et al., 2016; Lainé et al., 2017; Dale & Werner, 2017; Ho et al., 2019).

Further data preparation was done in Rstudio (R Development Core Team, 2008). The 212 selected spectra variables were corrected for days in milk (DIM), according to Vanlierde et al. (2015): Each first derivative value of the selected spectra was multiplied by a constant (i.e., 1), a linear ($\sqrt{3} * x$) and a quadratic [$\sqrt{5}/4 * (3x^2 - 1)$] modified Legendre polynomial (Gengler et al., 1999), where $x = -1 + 2[(DIM - 3)/(305 - 3)]$. This modification resulted in 636 (212 constant, 212 linear, 212 quadratic) spectra variables, which were finally used for the prediction model. The somatic cell count (SCS) was logarithmically transformed to the somatic cell score (SCS), by applying the formula: $SCS = \log2 (SCC /100{,}000) + 3$ (Fürst et al., 2019). Further, all values were centered and scaled.

The 635,588 records of the complete data set where randomly split by farm into half a calibration (train) and half a validation (test) data set (except for last model test, where splitting was done in other ratios also). In this way, cows in the validation set were from different herds than those in the calibration set. In final calibration data sets, the numbers of healthy and mastitis cases were always balanced (1:1) by using random down sampling. Further, different settings were applied on calibration sets for testing various factors.

2.1.3 Calibration and validation settings for final model tests

To evaluate various effects in the model, different settings, as explained in detail below, were applied on the calibration subsets of each model test. The settings for validation data set remained the same for all model tests; mastitis cases with maximum -21 to +21 days between diagnosis and test day were considered. In order to have a realistic validation data set, no further settings or restrictions were applied. There were only changes in sample size of validation data for the last model test, which was about to examine the effect of different sample sizes of calibration data set (Table 7). The results for validation set (Table 5, 6 and 7) are displayed for the overall time window of -21 to +21 days and additionally split into different shorter time windows. This was to demonstrate the difference in accuracy of prediction, when test day is before or after mastitis event, and how accuracy of prediction changes with the distance of days between mastitis event and test day record.

Effect of SCC restrictions in calibration set

The first objective was to test the effect of SCC restrictions for mastitis diagnosis in calibration set: Animals were considered as healthy, if Diagnosis = 0 and SCC <= 100,000; animals were considered to have mastitis, if Diagnosis = 1 and SCC >=400,000. Observations that did not fulfil these conditions were deleted. Another subset was created without SCC restrictions. This comparison was done for all model tests.

Comparison of different predictor variables

Second objective was to compare different predictor variables in the model: MIR (636 DIM corrected spectral data points), SCS alone and MIR plus SCS as covariate. The prediction was either done with 636 MIR variables, with SCS as a single predictor variable or with 636 MIR variables including SCS as covariate. For all calibration subsets, the maximum days between diagnosis and test day were set to -7 to +7 days, according to Soyeurt et al. (2012).

Effect of different time windows in calibration set

Further, we aimed to test the effect of variant time windows in calibration set. Therefore, 3 calibration subsets, which differed only in the number of days between diagnosis and test day date, were created: -7 to + 7 days, -14 to +14 days and -21 to +21 days. As predictor variables only MIR spectra (636 DIM corrected spectral data points) were used in the model.

Effect of different sample sizes for calibration set

Last objective was to examine whether sample size of calibration set influenced accuracy of the prediction model. Divergent to all other model tests, splitting into calibration and validation data set was not only done 50 : 50 %, there were also other ratios. This was necessary to get bigger sample sizes for calibration data set; which in turn reduced the size of validation set.

The proportions of calibration and validation set from complete data set with 635,588 records were as follows:

- small 1: 25 % train : 75 % test (enlarged sample size in validation)
- small 2: 25 % train : 50 % test (sample size in validation equal to previous model tests)
- medium: 50 % train : 50 % test (sample size in validation equal to previous model tests)
- big: 75 % train : 25 % test (reduced sample size of validation)

The different calibration subsets were again balanced in terms of mastitis and healthy cases by using random down sampling. Maximum days between diagnosis and test day were set to -7 to +7 days for calibration set. MIR spectra (636 DIM corrected spectral data points) were used as predictor variables.

2.2 Methodology

Prediction models were done with Partial Least Squares Discriminant Analysis (PLS-DA), using the R package 'caret' (Kuhn, 2008). The indicators of model fit were sensitivity (mastitis cases correctly assigned as mastitis), specificity (healthy cases correctly assigned as healthy) and balanced accuracy (mean of sensitivity and specificity).

Preliminary tests on data treatment were run with 100 replications per setting and the number of latent variables was set to 100, based on initial analysis with PLS in SAS. For the comparison of different data treatments, the results were evaluated in pairs with a t-test

For the final model tests in RStudio, the number of latent variables was reduced to 50, which is more adequate for the PLS-DA procedure in R. When just SCS was used as predictor variable, the number of latent variables was set to one. We chose to run 20 replications per setting for the final model tests. Given a standard deviation of 0.017 for replicates, that allowed to detect significance at a p-value of 0.05 for differences of around 0.015. Sample size calculator (https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html) was used for finding the number of replications.

3 Results

3.1 Results of preliminary tests on different pre-treatments of spectra data

The results of the preliminary test on pre-treatments on spectra data are displayed in Table 3, which include the indicators of model fit (sensitivity and specificity) for each model and the p-values of the pairwise t-test. The sensitivities were in a range of 0.467 and 0.477, thus, differences were very small and mostly not significant. The differences among specificities were bigger and mostly significant, the range was between 0.729 and 0.770.

Table 3 Effect of different pre-treatments on spectra data; t-tests applied to compare sensitivities and specificities of individual data sets (significance at p-value <= 0.05)

Pre-treatment	Sensitivity		Specificity	
	mean	p-value	mean	p-value
original full	0,474656	0,066380	0,742131	< 2,2e-16
original select	0,469967		0,770187	
1st der. full	0,477209	0,001132	0,729491	< 2,2e-16
1st der. select	0,467815		0,763123	
2nd der. full	0,470323	0,195800	0,731391	< 2,2e-16
2nd der. select	0,466948		0,765561	
original select	0,469967	0,459500	0,770187	0,002393
1st der. select	0,467815		0,763123	
original select	0,469967	0,251500	0,770187	0,023980
2nd der. select	0,466948		0,765561	
1st der. select	0,467815	0,760500	0,763123	0,288200
2nd der. select	0,466948		0,765561	
original select	0,469967	0,006343	0,770187	< 2,2e-16
1st der. full	0,477209		0,729491	

original full = untreated full spectra (1,060 wavelengths)
original select = untreated selected spectra (212 wavelengths)
1st der. full = 1st derivative of full spectra (1,060 wavelengths)
1st der. select = 1st derivative selected spectra (212 wavelengths)
2nd der. full = 2nd derivative of full spectra (1,060 wavelengths)
1st der. select =2nd derivative of selected spectra (212 wavelengths)

In a first step, the full 1,060 data points of untreated, first derivative and second derivative spectra data were compared with the selected spectra parts of the respective data set. These comparisons showed significantly higher specificities for the selected spectra of all types (original, first derivative,

second derivative). Sensitivities were generally higher for the full spectra, but only significantly higher for the first derivative.

In a second step, all variants of the selected spectra (original select, first der. select, second der. select) were compared with each other. Original select and first derivative select were almost identical for sensitivity, but specificity was significantly higher for original select. The difference in sensitivity between original select and second derivate select, was also very small and not significant, but specificity was again significantly higher for original select. For the comparison of first derivative and second derivative no significant differences were found, sensitivity and specificity of both were almost equal.

Finally, the data set original select, which was best in terms of specificity among all variants, was compared with first derivative full, which was best in terms of sensitivity among all variants. This comparison showed that specificity was significantly higher for original select, but sensitivity significantly higher for first derivative full, though the difference in specificity was bigger.

In general, differences among specificities were stronger, compared with sensitivities, where only slight differences were found.

3.2 Results of final model tests

3.2.1 Comparison of different predictor variables

For comparison of different predictor variables (MIR, SCS, MIR + SCS), the number of records was on average 2,340 (1,170 mastitis, 1,170 healthy) for the calibration sets without SCC limits, and 1,086 (543 mastitis, 543 healthy) for the calibration sets with SCC limits. Thus, applying SCC limits reduced the size of the calibration set roughly to half. The validation set counted roughly 315,000 records with a proportion of around 1 % mastitis cases for overall time window (-21 to +21 days) and around 0.2 % for smaller time windows.

The following section presents the results of the first model test, which was to examine the effect of different predictor variables (MIR, SCS, MIR + SCS).

Table 4 displays the results of PLS-DA procedure within calibration set. For the different predictor variables (MIR, SCS, MIR + SCS), sensitivity, specificity and balanced accuracy were higher in the calibration set with SCC limits. For SCS and MIR + SCS all indicators were 1, for MIR alone sensitivity was 0.84 and specificity 0.89. In the calibration set without SCC limits, sensitivities and specificities were lower for all types of models.

Table 4 Results in calibration (train) for different predictor variables (MIR, SCS or MIR + SCS); with and without SCC limits in train

Predictor variable	no SCC limits in train			SCC limits in train		
	sens.	spec.	bal.acc.	sens.	spec.	bal.acc.
MIR	0.680	0.770	0.725	0.843	0.886	0.864
SCS	0.617	0.849	0.733	1.000	1.000	1.000
MIR + SCS	0.735	0.838	0.786	1.000	1.000	1.000

sens. = sensitivity; spec. = specificity; bal.acc. = balanced accuracy

Table 5 shows the results of the model testing for the validation set. All results were for the full validation set (-21 to +21 days). Splitting them into different shorter time windows changed the number of mastitis cases, but not the number of healthy cases. Therefore, specificities of different time windows did not differ from specificity of the overall window. Differences in balanced accuracy resulted from changing sensitivity.

Applying SCC limits in calibration did not lead to a higher balanced accuracy in validation, compared to prediction equations derived from calibration data sets without SCC limits (Table 4). For all models, specificity was higher with SCC limits, but sensitivity was lower. The differences in sensitivity (all time windows) and specificity were significant for MIR and MIR + SCS. Without SCC limits, sensitivity and

specificity were more balanced and balanced accuracy was slightly higher for all variants, except, SCS (-21 to +21 days), SCS (-21 to -15 days), SCS (-7 to +7 days) and MIR + SCS (-14 to -8 days).

Table 5 The effect of different predictor variables (MIR, SCS or MIR + SCS) and SCC limits in calibration (train), results in validation (test) extracted for different time windows (days)

predictor variables	no SCC limits in train			SCC limits in train			time windows test
	sens.	spec.	bal.acc.	sens.	spec.	bal.acc.	
MIR	0.534	0.708	0.621	0.457	0.755	0.606	-21 to +21 (overall)
SCS	0.501	0.849	0.675	0.490	0.862	0.676	
MIR + SCS	0.574	0.791	0.682	0.473	0.878	0.675	
MIR	0.458	0.708	0.583	0.368	0.755	0.561	-21 to -15
SCS	0.401	0.849	0.625	0.394	0.862	0.628	
MIR + SCS	0.474	0.791	0.633	0.373	0.878	0.626	
MIR	0.484	0.708	0.596	0.412	0.755	0.584	-14 to -8
SCS	0.513	0.849	0.681	0.499	0.862	0.681	
MIR + SCS	0.555	0.791	0.673	0.477	0.878	0.678	
MIR	0.605	0.708	0.657	0.540	0.755	0.647	-7 to +7
SCS	0.615	0.849	0.732	0.604	0.862	0.733	
MIR + SCS	0.678	0.791	0.735	0.586	0.878	0.732	
MIR	0.560	0.708	0.634	0.478	0.755	0.617	+8 to +14
SCS	0.466	0.849	0.658	0.447	0.862	0.655	
MIR + SCS	0.558	0.791	0.675	0.441	0.878	0.659	
MIR	0.479	0.708	0.594	0.394	0.755	0.574	+15 to +21
SCS	0.374	0.849	0.612	0.361	0.862	0.612	
MIR + SCS	0.475	0.791	0.633	0.347	0.878	0.612	

sens. = sensitivity; spec. = specificity; bal.acc. = balanced accuracy

Comparing the different predictor variables in the models without SCC limits in calibration gave the following results: Overall sensitivity was significantly higher for MIR, but overall specificity was significantly higher for SCS. For the individual time windows (except -14 to -8 days), the highest balanced accuracies, were found for MIR + SCS.

For the comparison of predictor variables with SCC limits in calibration, specificity was highest for MIR + SCS (0.88) and lowest for MIR alone (0.75) for the -21 to +21 days time window. For the overall validation set, the highest sensitivity (0.49) was also reached with SCS alone. With regard to the individual time windows, sensitivity of SCS was highest (significantly) for -7 to +7 days, -14 to -8 days

and -21 to -15 days, but for +15 to +21 days and +8 to +14 days it was highest (significantly) for MIR alone.

3.2.2 Effect of different time windows in calibration set

For examining the effect of different time windows (-7 to +7 days, -14 to +14 days, -21 to +21 days) in calibration set, the average number of records in the different calibration subsets was:

- -7 to +7 days: 2,346 (1,173 mastitis, 1,173 healthy) without SCC limits
  1,086 (543 mastitis, 543 healthy) with SCC limits
- -14 to +14 days: 4,320 (2,310 mastitis, 2,310 healthy) without SCC limits
  1,814 (907 mastitis, 907 healthy) with SCC limits
- -21 to +21 days: 6,652 (3,326 mastitis, 3,326 healthy) without SCC limits
  2,278 (1,139 mastitis, 1,139 healthy) with SCC limits

These numbers show that not just applying SCC limits but also smaller time windows, leads to reduced sample sizes in calibration set. The validation data set consisted of roughly 315,000 records the proportion of mastitis cases was around 1 % for overall time window (-21 to +21 days) and around 0.2 % for shorter validation time windows.

Table 6 displays the results of model testing for the validation set. According to previous model tests, all results were for the full validation set (maximum -21 to +21 days between mastitis event and test day record) and additionally split into different shorter time windows. As already mentioned in section 3.2.1., this only lead to different sensitivities while specificities remained the same for all shorter time windows.

Equivalent to the results of Table 5, SCC limits in calibration lead to a higher imbalance between sensitivities and specificities. For the overall validation set (-21 to +21 days) and all shorter time windows, sensitivity was significantly higher without SCC limits and specificity was significantly higher with SCC limits. Yet, balanced accuracy was always higher (mostly significantly) for calibration subsets without SCC limits.

The model tests without SCC limits showed that sensitivity increases with a larger time window in calibration and specificity decreases. For the overall time window in validation, the highest sensitivity (0.59) was reached by the calibration subset with -21 to +21 days. For the next shorter calibration time window (-14 to +14 days) sensitivity was almost equal (0.58) but dropped significantly to 0.53 for the shortest calibration time window (-7 to +7 days). Specificity was highest (0.71) for the calibration subset with -7 to +7 days and lowest (0.68) for the calibration subset with -21 to +21 days. Regarding the results extracted for shorter time windows in validation, a similar trend of an increasing sensitivity

for larger calibration time windows, was found. The differences were mostly significant. Balanced accuracies increased with a larger calibration time window and were always highest for the -21 to +21 days calibration set.

Table 6 The effect of different time windows (days) and SCC limits in train, results extracted for different time windows (days) in test (validation)

time windows train	no SCC limits in train			SCC limits in train			time windows test
	sens.	spec.	bal.acc.	sens.	spec.	bal.acc.	
-7 to +7	0.534	0.708	0.621	0.457	0.755	0.606	-21 to +21 (overall)
-14 to +14	0.576	0.682	0.629	0.475	0.754	0.614	
-21 to +21	0.588	0.677	0.632	0.480	0.757	0.618	
-7 to +7	0.458	0.708	0.583	0.368	0.755	0.561	-21 to -15
-14 to +14	0.489	0.682	0.585	0.375	0.754	0.564	
-21 to +21	0.506	0.677	0.591	0.392	0.757	0.575	
-7 to +7	0.484	0.708	0.596	0.412	0.755	0.584	-14 to -8
-14 to +14	0.523	0.682	0.602	0.423	0.754	0.588	
-21 to +21	0.536	0.677	0.606	0.420	0.757	0.588	
-7 to +7	0.605	0.708	0.657	0.540	0.755	0.647	-7 to +7
-14 to +14	0.638	0.682	0.660	0.560	0.754	0.657	
-21 to +21	0.645	0.677	0.661	0.558	0.757	0.658	
-7 to +7	0.560	0.708	0.634	0.478	0.755	0.617	+8 to +14
-14 to +14	0.619	0.682	0.650	0.498	0.754	0.626	
-21 to +21	0.634	0.677	0.655	0.507	0.757	0.632	
-7 to +7	0.479	0.708	0.594	0.394	0.755	0.574	+15 to +21
-14 to +14	0.539	0.682	0.610	0.423	0.754	0.588	
-21 to +21	0.552	0.677	0.614	0.429	0.757	0.593	

sens. = sensitivity; spec. = specificity; bal.acc. = balanced accuracy

The model tests with SCC limits showed again mostly higher sensitivities and always higher balanced accuracies for larger calibration time windows. Specificities were almost equal (≈0.76) for all calibration time windows. For the overall validation time window (-21 to +21 days), sensitivity significantly increased from 0.46 for the -7 to +7 days calibration set, up to 0.48 for the -21 to +21 days calibration set. Regarding the results extracted for shorter validation time windows, highest sensitivities (0.54 to 0.56) were again found for the -7 to +7 validation time window. Sensitivity dropped, when distance between mastitis event and test day record was larger. The decline in sensitivity was stronger, when test day was before mastitis event (-21 to -15 and -14 to -8 days time window).

3.2.3 Effect of different sample sizes for calibration set

For the last model test, the sample sizes for the different calibration subsets (small 1, small 2, medium, big) were as follows:

- small 1: 1,172 (586 mastitis; 586 healthy) without SCC limits
  546 (273 mastitis; 273 healthy) with SCC limits
- small 2: 1,168 (584 mastitis; 584 healthy) without SCC limits
  542 (271 mastitis; 271 healthy) with SCC limits
- medium: 2,346 (1,173 mastitis; 1,173 healthy) without SCC limits
  1,104 (552 mastitis; 552 healthy) with SCC limits
- big: 3,552 (1,776 mastitis; 1,776 healthy) without SCC limits
  1,646 (823 mastitis; 823 healthy) with SCC limits

The subsets small 1 and small 2 had an equal sample size in calibration, they differed only in the associated validation set. The respective validation sets had on average 476,713 (small 1), 317,702 (small 2), 319,143 (medium) and 158,868 (big) records. The proportion of mastitis cases was again around 1 % for overall validation time window (-21 to +21 days) and around 0,2 % for shorter time windows.

Table 7 displays the results in validation for the four different calibration sample sizes (small 1, small 2, medium, big). Adding SCC limits to the calibration sets, lead again to a higher imbalance between sensitivities and specificities. Moreover, balanced accuracies were lower with SCC limits in calibration.

Without SCC limits, sensitivities were quite similar for the overall time window: 0.53 for small 1 and medium, 0.54 for small 2 and big. Specificity slightly increased from 0.67 for the small calibration set, up to 0.71 for the big calibration set. The sensitivities extracted for shorter time windows in validation, showed differences up to 0.02, but there was no regular up- or down trend evident. Sensitivities were again highest for the -7 to +7 days time window in validation. Regarding balanced accuracy, the big calibration set performed best for all time windows (including overall validation set).

With SCC limits, specificity was 0.71 for both small calibration sets and increased up to 0.76 for medium and big. For the overall time window, sensitivities in the range of 0.46 to 0.47 were found. For shorter time windows, the differences in sensitivity were higher (up to 0.03), but again no regular up- or down trend was visible. The balanced accuracy increased with a higher sample size.

Table 7 Effect of different sample sizes and SCC limits in train (calibration), results extracted for different time windows (days) in test (validation)

sample size train	no SCC limits in train			SCC limits in train			time windows test
	sens.	spec.	bal.acc.	sens.	spec.	bal.acc.	
small 1	0.525	0.679	0.602	0.475	0.711	0.593	-21 to +21 (overall)
small 2	0.539	0.674	0.606	0.476	0.712	0.594	
medium	0.534	0.708	0.621	0.457	0.755	0.606	
big	0.538	0.714	0.626	0.470	0.762	0.616	
small 1	0.459	0.679	0.569	0.400	0.711	0.556	-21 to -15
small 2	0.478	0.674	0.576	0.404	0.712	0.558	
medium	0.458	0.708	0.583	0.368	0.755	0.561	
big	0.462	0.714	0.588	0.371	0.762	0.567	
small 1	0.477	0.679	0.578	0.437	0.711	0.574	-14 to -8
small 2	0.496	0.674	0.585	0.444	0.712	0.578	
medium	0.484	0.708	0.596	0.412	0.755	0.584	
big	0.480	0.714	0.597	0.416	0.762	0.589	
small 1	0.588	0.679	0.633	0.540	0.711	0.626	-7 to +7
small 2	0.600	0.674	0.637	0.543	0.712	0.628	
medium	0.605	0.708	0.657	0.540	0.755	0.647	
big	0.611	0.714	0.663	0.557	0.762	0.659	
small 1	0.546	0.679	0.612	0.490	0.711	0.601	+8 to +14
small 2	0.558	0.674	0.616	0.488	0.712	0.600	
medium	0.560	0.708	0.634	0.478	0.755	0.617	
big	0.563	0.714	0.639	0.496	0.762	0.629	
small 1	0.483	0.679	0.581	0.431	0.711	0.571	+15 to +21
small 2	0.493	0.674	0.583	0.422	0.712	0.567	
medium	0.479	0.708	0.594	0.394	0.755	0.574	
big	0.484	0.714	0.599	0.413	0.762	0.588	

sens. = sensitivity; spec. = specificity; bal.acc. = balanced accuracy

4 Discussion

4.1 Discussion of preliminary tests on different pre-treatments of spectra data

Due to the results of preliminary tests (section 3.1), the use of selected spectra areas was more appropriate for this study, compared to the use of the full spectra with 1,060 wavelengths. While the full spectra data set performed better in sensitivity, the differences were very small or rather not significant. With regard to specificity, the selected spectra data set was doing better, effects were stronger and significant for all types (original, first derivative, second derivative). Thus, overall there was an advantage for the selected spectra areas, which were in further consequence used for the final model tests.

The results within the selected spectra data sets were very similar for all types (original, first derivative, second derivative). Only the original selected spectra data set had a significantly higher specificity, however, the difference was less than 0.005. For this reason, the first derivative of the spectra was taken for the final model tests, according to other studies by Soyeurt et al. (2011; 2012), Grelet et al. (2016), Dale & Werner (2017), Lainé et al. (2017) and Ho et al. (2019).

4.2 Discussion of final model tests

Comparison of different predictor variables

Considering the results within calibration data sets (Table 4), the sensitivity and the specificity of 1.00 for predictor variables SCS and MIR + SCS, and also high values for MIR alone, when using SCC limits, are due to overfitting of the model. Applying the particular model to the realistic validation set (Table 5) did not show an advantage of using SCC limits in calibration. It resulted in a higher imbalance of sensitivity and specificity and a lower balanced accuracy, compared to the model without SCC limits in the calibration set. This imbalance of sensitivity and specificity was also found in the study of Soyeurt et al. (2012), where MIR predicted lactoferrin was used as an indicator for mastitis. In Figure 4, the higher imbalance of sensitivity and specificity when applying SCC limits, is visualized for predictor MIR +SCS.

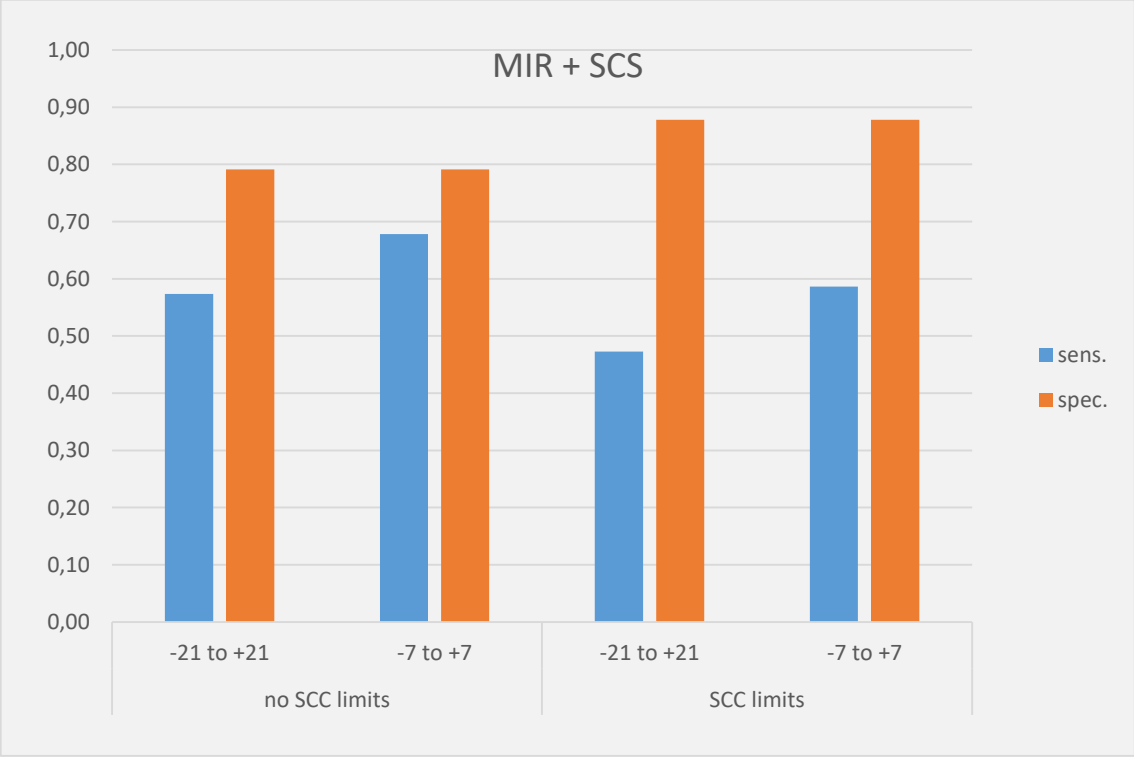

Figure 4 Sensitivity and specificity of MIR + SCS with or without SCC limits, for overall (-21 to +21 days) and shortest (-7 to +7 days) time window in validation

According to the results, the predictor model without SCC limits in calibration, was more adequate. Therefore, the discussion of further effects focusses on that model. Table 5 clearly demonstrates that the prediction of mastitis cases works better with a shorter distance between diagnosis and test day in the validation dataset. The time window of -7 to +7 days in validation was also applied in the study

of Soyeurt et al. (2012). When comparing the time windows with larger distance between diagnosis and test day record, predictions with MIR + SCS were very similar for test days before and after the occurrence of mastitis events. Yet, considering SCS and MIR as predictors separately, results seem to indicate that MIR predicted mastitis better, when test days were after mastitis diagnosis, while SCS predicted mastitis events better, when test days were before diagnosis. Prediction equations combining SCS and MIR were overall best. Figure 5 shows the course of sensitivity of MIR, SCC and MIR+SCS for the different time windows before and after diagnosis is displayed. The stronger drop in sensitivity for SCS may be explained by Figure 2, which shows, that the average somatic cell count is lower short after the mastitis event, than short before. A reason for that could be antibiotic treatment.

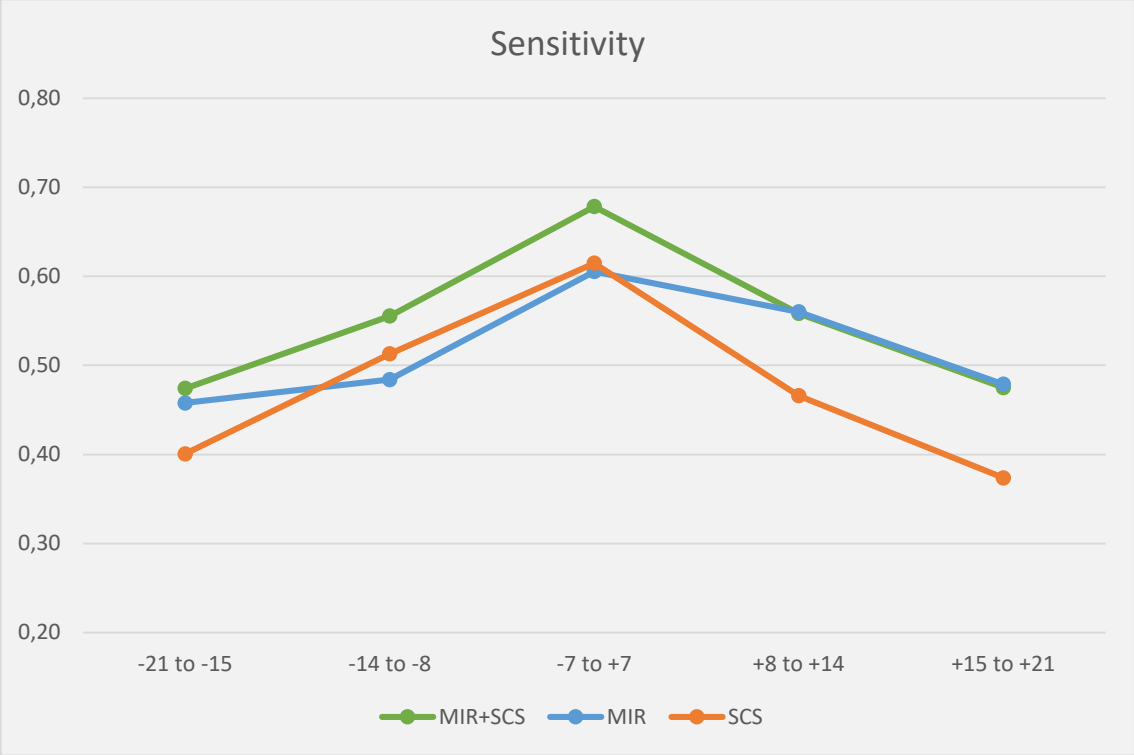

Figure 5 Course of sensitivity of the predictor variables MIR, SCS and MIR + SCS with different time windows before/after diagnosis; without SCC limits in train

Effect of different time windows in calibration set

The results in validation (Table 6) show that the use of a larger time window in the calibration set improves the accuracy of the prediction model. Whether with or without SCC limits in calibration, balanced accuracy was overall best (for all validation time windows) for the calibration set with maximum -21 to +21 days and lowest for the calibration set with -7 to + 7 days between test day and mastitis event. Figure 6 displays the balanced accuracies for different calibration sets without SCC

limits. Even when just considering the results of the validation time window of -7 to +7 days, balanced accuracies were higher when the calibration set with -21 to +21 days was used.

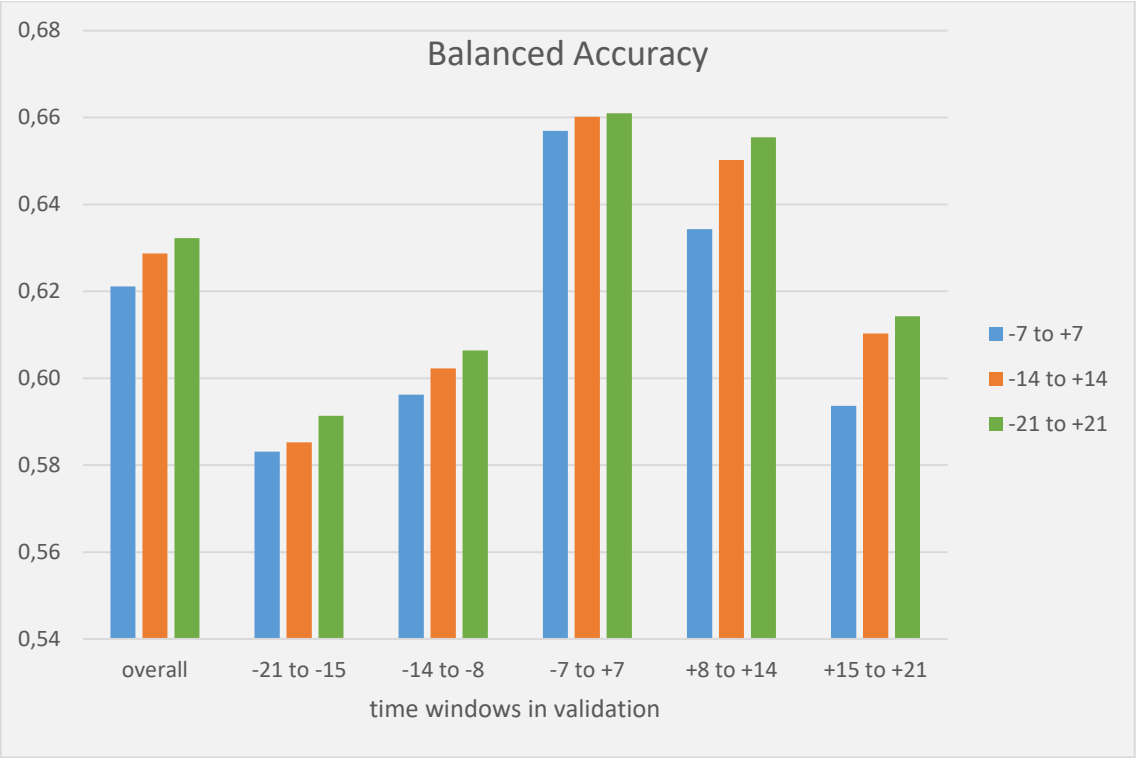

Figure 6 Balanced accuracies of different calibration time windows (-7 to +7, -14 to +14 and -21 to +21 days); without SCC limits in train, results extracted for different time windows in validation

It is important to note that enlarging the time window also increased the sample size of the calibration set. Starting with the same data set, the calibration set with -21 to +21 days contained around three times more records than the calibration set with -7 to +7 days. Thus, the positive effect of a larger time window in calibration may have also been influenced by a bigger sample size. For further analysis it would be appropriate to work with a larger calibration window, differing from the studies of Soyeurt et al. (2012) and (Dale & Werner, 2017), where shorter time windows (-7 to +7 or -7 to 0 days) were applied. In order to be able to distinguish the effect of time window from the effect of sample size, the model test needs to be repeated with a non-changing sample size in calibration.

Effect of different sample sizes for calibration set

The last objective of this thesis was to test the effect of different sample sizes of the calibration set. While PLS-DA is one of the most commonly used methods for classification purposes and biomarker selection in metabolomics (Szymańska et al., 2012), the paper of Saccenti & Timmerman (2016) is one of the very rare references on sample size determination for PLS-DA. Therefore, the results of this model test will be discussed and compared only with that study.

Saccenti & Timmerman (2016) built a series of PLS-DA models using an increasing number of samples (from 25 controls + 25 cases to 500 controls + 500 cases). The experimental data used was from nuclear-magnetic-resonance (NMR)spectra of serum blood metabolites (D.1 and D.2) and of urine (D.3 and D.4). The results of Saccenti & Timmerman (2016) showed that sensitivity and specificity increase (Figure 7 and Figure 8) with the sample size, and variability decreases.

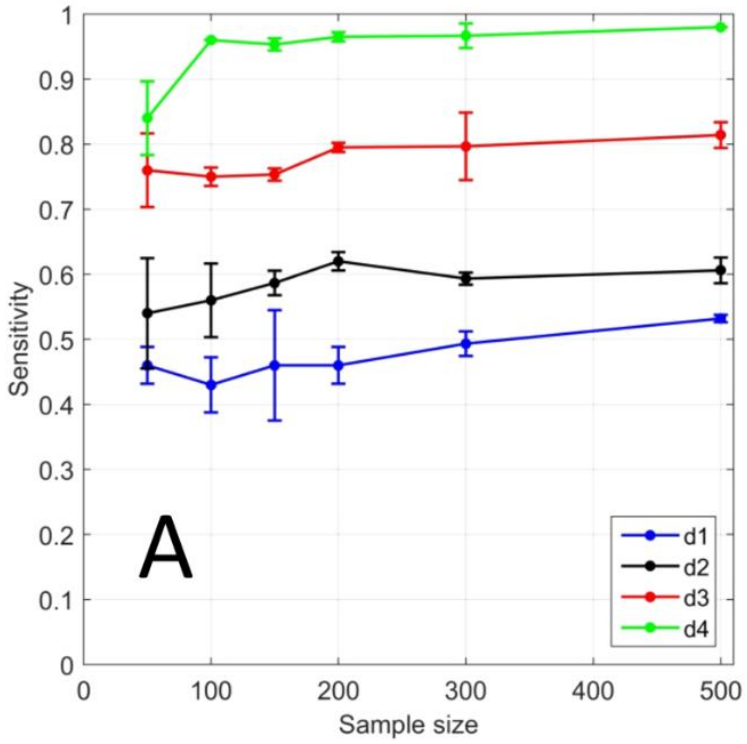

Figure 7 Sensitivity (A) of a PLS-DA model as a function of the total sample size for the discrimination between two group in a case-control design (Saccenti & Timmerman, 2016)

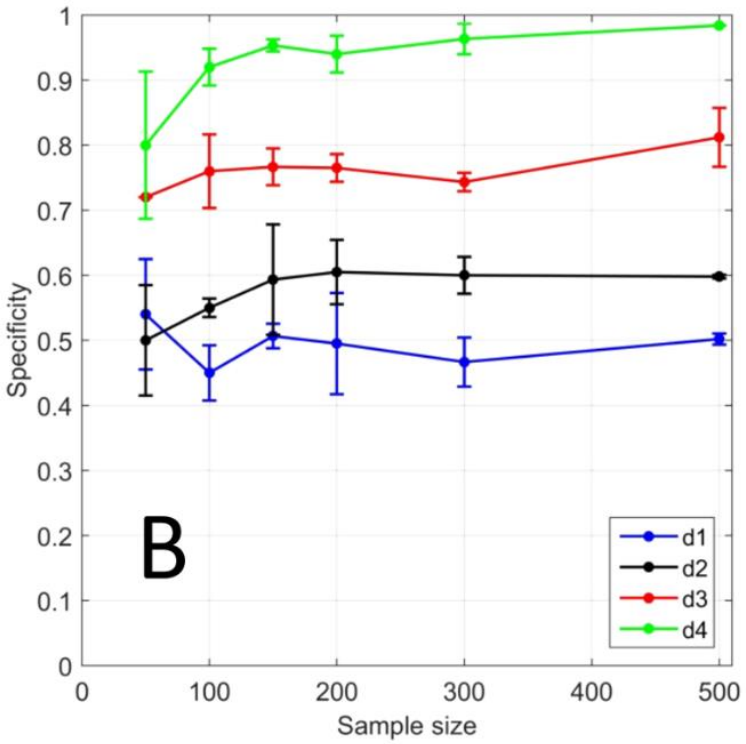

Figure 8 Specificity (B) of a PLS-DA model as a function of the total sample size for the discrimination between two group in a case-control design (Saccenti & Timmerman, 2016)

In this thesis, the examined sample sizes where much bigger, starting from 1,165 (586 mastitis + 586 healthy) to 3,552 (1,776 mastitis + 1,776 healthy) for calibration set without SCC limits, and from 542 (273 mastitis + 273 healthy) to 1,646 (823 mastitis + 823 healthy), than in the study of Saccenti & Timmerman (2016). Also, the differences in sensitivity and specificity between the individual sample sizes (Table 7) where smaller. A regular increase with sample size was found for specificity (Figure 10), but not for sensitivity (Figure 9). When no SCC limits were applied, the pattern was not clear, highest and lowest sensitivities were found for the two small sample sizes in calibration (small 1 and small 2). When SCC limits were applied, sensitivity was highest for these small sample sizes.

For a clear indication on how sample size affects the prediction accuracy of mastitis from MIR spectra, and what the minimum sample size should be, further analysis according to Saccenti & Timmerman (2016) should be done. However, as there were only small differences in the results, the sample sizes used for this thesis seem to be appropriate.

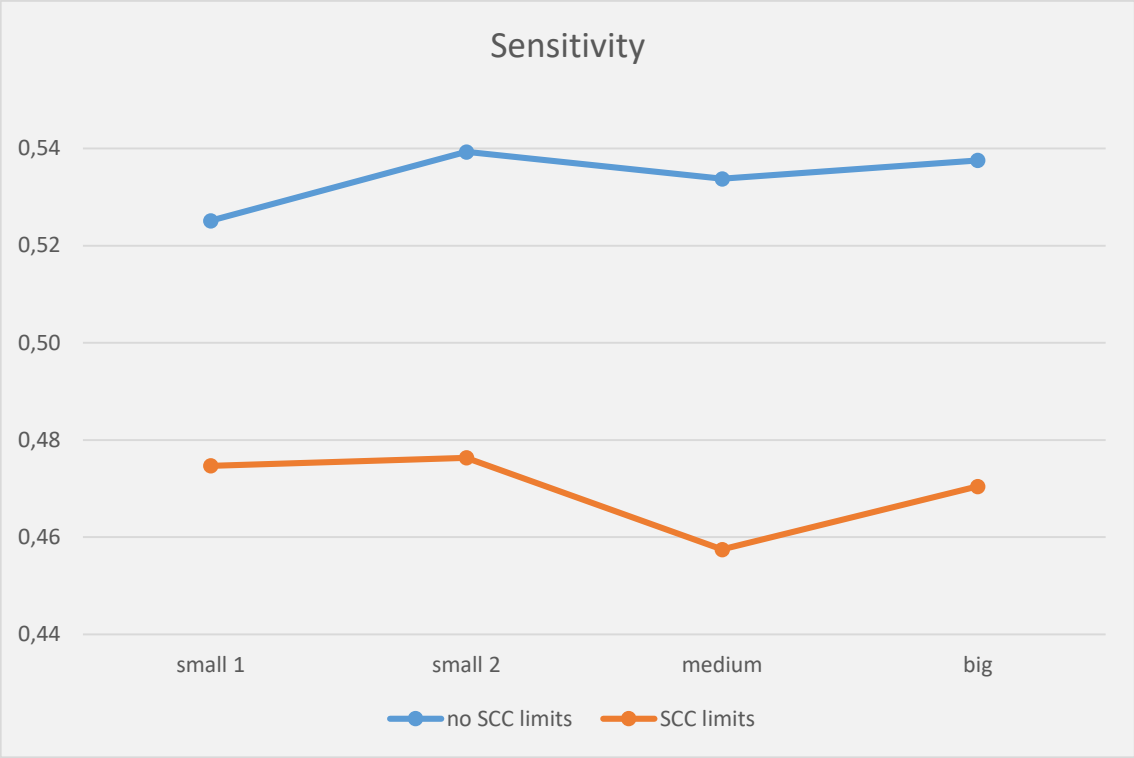

Figure 9 Sensitivities for different sample sizes in calibration; with or without SCC limits in train; results of overall validation set

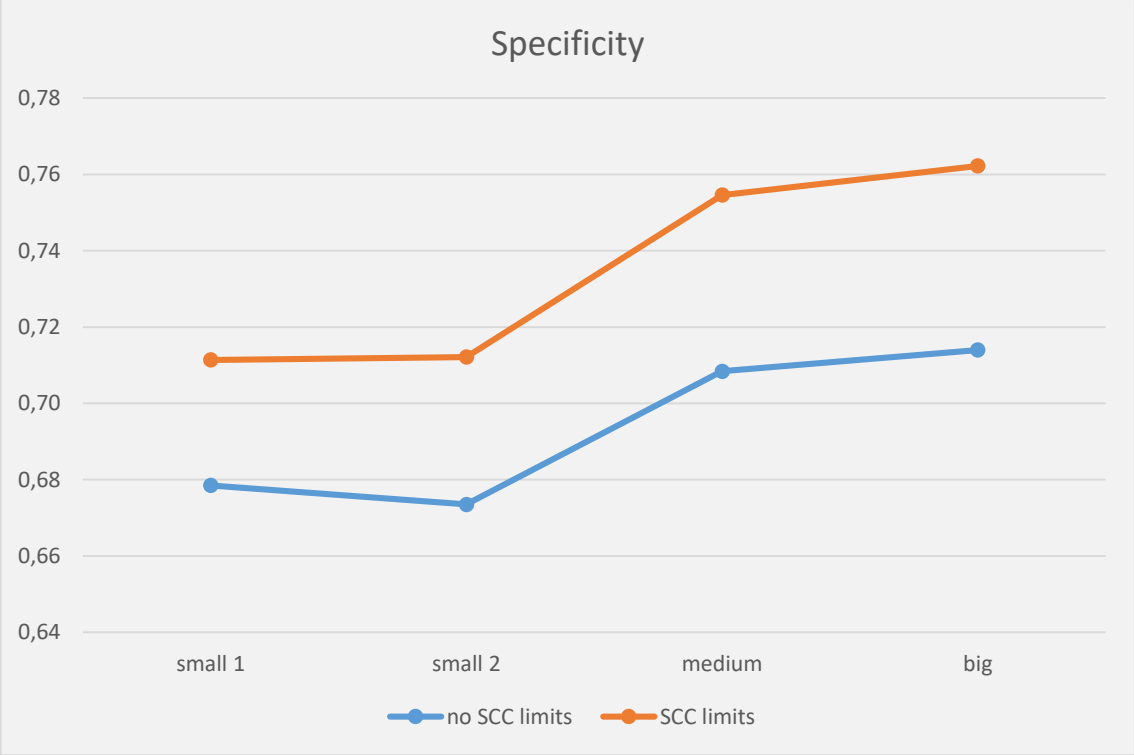

Figure 10 Specificities for different sample sizes in calibration; with or without SCC limits in train; results of overall validation set

In general, all the results presented in this thesis were hard to compare with the few other studies on MIR predicted mastitis, such as Soyeurt et al. (2012) and Dale & Werner (2017), because types of validation were very different.

5 Conclusion

This study explored the potential of milk MIR spectral data for prediction of mastitis cases. We investigated the utility of combining information of MIR and SCS, which were both available for every test day record. Results indicate that mastitis diagnoses may be predicted reasonably accurate, with balanced accuracies of 0.62 to 0.68 for time windows of +/- 21 days between mastitis diagnosis and test day and up to 0.74 for shorter time windows. The information is potentially valuable for improved genetic evaluation of udder health, which is currently an index of SCS and clinical mastitis. Future studies on heritability and genetic correlations of clinical mastitis, SCS and MIR predicted mastitis will provide guidance in this direction.

Additional studies are planned to further improve the prediction model, through adjustments in methodology and by including the effects of milk yield, lactose, breed and parity.

6 References

Belay, T. K., Svendsen, M., Kowalski, Z. M., & Ådnøy, T. (2017). Genetic parameters of blood β-hydroxybutyrate predicted from milk infrared spectra and clinical ketosis, and their associations with milk production traits in Norwegian Red cows. *Journal of Dairy Science*, 100(8), 6298–6311. https://doi.org/10.3168/jds.2016-12458

Blowey, R. W., & Edmondson, P. (2010). *Mastitis Control in Dairy Herds* (2nd ed.). *CAB International. Oxfordshire, UK.* 1-266.

Bradley, A. J. (2002). Bovine mastitis: An evolving disease. *Veterinary Journal*, *164*(2), 116–128. https://doi.org/10.1053/tvjl.2002.0724

Cervinkova, D., Vlkova, H., Borodacova, I., Makovcova, J., Babak, V., Lorencova, A., Vrtkova, I., Marosevic, D. & Jaglic, Z. (2013). Prevalence of mastitis pathogens in milk from clinically healthy cows. *Veterinarni Medicina*, *58*(11), 567–575. https://doi.org/10.17221/7138-VETMED

Dale, L., & Werner, A. (2017). "MastiMIR"-Ein Mastitis-Frühwarnsystem basierend auf MIR-Spektren. *Paper presented at the conference of DGfZ and GfT, Stuttgard, Germany.* Retrieved from https://www.researchgate.net/publication/320351923

De Marchi, M., Toffanin, V., Cassandro, M., & Penasa, M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *Journal of Dairy Science*, *97*(3), 1171–1186. https://doi.org/10.3168/jds.2013-6799

De Roos, A. P. W., Van Den Bijgaart, H. J. C. M., Hørlyk, J., & De Jong, G. (2007). Screening for subclinical ketosis in dairy cattle by fourier transform infrared spectrometry. *Journal of Dairy Science*, *90*(4), 1761–1766. https://doi.org/10.3168/jds.2006-203

Dairyman's Digest (2009). What You Should Know about Somatic Cells SCCs versus Linear Scores. *Dairyman's Digest*, (Winter Issue). Retrieved from https://www.medvet.umontreal.ca/rcrmb/dynamiques/PDF_AN/Management/WhatShouldKnowSCC.pdf

D4Dairy Consortium. ©2019. D4Dairy. [Online]. Retrived from https://d4dairy.com/de/#projekt. [Accessed July 16, 2019]

Egger-Danner, C., Fürst, C., Mayerhofer, M., Rain, C., & Rehling, C. (2018). ZuchtData Jahresbericht 2018. *ZuchtData EDV-Dienstleistungen GmbH, Vienna, Austria*.

Fürst, C., Dodenhoff, J., Egger-Danner, C., Emmerling, R., Hamann, H., Krogmeier, D., & Schwarzenbacher, H. (2019). Zuchtwertschätzung beim Rind - Grundlagen, Methoden und Interpretationen. *Zuchtdata EDV-Dienstleistungen GmbH.* Retrieved from http://www.zar.at/download/ZWS/ZWS.pdf

Garcia, A. (2004). Contagious vs. Environmental Mastitis. *Extension Extra - College of Agriculture & Biological Science, South Dakota Stae University.* , *4028*(January). Retrieved from http://extensionenespanol.net/pubs/exex4028.pdf

Gengler, N., Tijani, A., Wiggans, G. R., & Misztal, I. (1999). Estimation of (Co)variance function coefficients for test day yield with a expectation-maximization restricted maximum likelihood algorithm. *Journal of Dairy Science*, *82*(8), 1849.e1-1849.e23. https://doi.org/10.3168/jds.S0022-0302(99)75417-2

Grelet, C., Bastin, C., Gelé, M., Davière, J. B., Johan, M., Werner, A., Reding, R., Fernandez Pierna, J. A., Colinet F. G., Dardenne, P., Gengler, N., Soyeurt H. & Dehareng, F. (2016). Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. *Journal of Dairy Science*, *99*(6), 4816–4825. https://doi.org/10.3168/jds.2015-10477

Grelet, C., Fernández Pierna, J. A., Dardenne, P., Baeten, V., & Dehareng, F. (2015). Standardization of milk mid-infrared spectra from a European dairy network. *Journal of Dairy Science*, *98*(4), 2150–2160. https://doi.org/10.3168/jds.2014-8764

Guimarães, J. L. B., Brito, M. A. V. P., Lange, C. C., Silva, M. R., Ribeiro, J. B., Mendonça, L. C., Mendonca, J. F. M. & Souza, G. N. (2017). Estimate of the economic impact of mastitis: A case study in a Holstein dairy herd under tropical conditions. *Preventive Veterinary Medicine*, *142*, 46–50. https://doi.org/10.1016/j.prevetmed.2017.04.011

Halasa, T., Huijps, K., Østerås, O., & Hogeveen, H. (2007). Economic effects of bovine mastitis and mastitis management: A review. *Veterinary Quarterly*, *29*(1), 18–31. https://doi.org/10.1080/01652176.2007.9695224

Hamann, J., & Krösmker, V. (1997). Potential of specific milk composition variables for cow health management. *Livestock Production Science*, *48*, 201–208.

Heikkilä, A.-M., Nousiainen, J. I., & Pyörälä, S. (2011). Costs of clinical mastitis with special reference to premature culling. *Journal of Dairy Science*, *95*(1), 139–150. https://doi.org/10.3168/jds.2011-4321

Ho, P. N., Bonfatti, V., Luke, T. D. W., & Pryce, J. E. (2019). Classifying the fertility of dairy cows using milk mid-infrared spectroscopy. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2019-16412

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: The next challenge. *Nature Reviews Genetics*, *11*(12), 855–866. https://doi.org/10.1038/nrg2897

International Dairy Federation (IDF). (2012). Milk and liquid milk products— Guidelines for the application of mid-infrared spectrometry. IDF norm 141. ISO/DIS 9622:2012.

Klaffenböck, M., Steinwidder, A., Fasching, C., Terler, G., Gruber, L., Mészáros, G., & Sölkner, J. (2017). The use of mid-infrared spectrometry to estimate the ration composition of lactating dairy cows. *Journal of Dairy Science*, *100*(7), 5411–5421. https://doi.org/10.3168/jds.2016-12189

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26.

Lainé, A., Bastin, C., Grelet, C., Hammami, H., Colinet, F. G., Dale, L. M., Gillon, A., Vandenplas, J., Dehareng, F. & Gengler, N. (2017). Assessing the effect of pregnancy stage on milk composition of dairy cows using mid-infrared spectra. *Journal of Dairy Science*, *100*(4), 2863–2876. https://doi.org/10.3168/jds.2016-11736

McDermott, A., Visentin, G., De Marchi, M., Berry, D. P., Fenelon, M. A., O'Connor, P. M., Kenny, O. A. and McParland, S. (2016). Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. *Journal of Dairy Science*, *99*(4), 3171–3182. https://doi.org/10.3168/jds.2015-9747

McParland, S., Lewis, E., Kennedy, E., Moore, S. G., McCarthy, B., O'Donovan, M., Butler, S. T., Pryce, J. E. & Berry, D. P. (2014). Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. *Journal of Dairy Science*, *97*(9), 5863–5871. https://doi.org/10.3168/jds.2014-8214

Mineur, A., Köck, A., Grelet, C., Gengler, N., Egger-Danner, C., & Sölkner, J. (2017). First Results in the Use of Milk Mid-infrared Spectra in the Detection of Lameness in Austrian Dairy Cows. *Agriculturae Conspectus Scientificus*, *82*(2), 163–166. Retrieved from https://www.researchgate.net/publication/325450513

Pryce, J. E., Goddard, M. E., Raadsma, H. W., & Hayes, B. J. (2010). Deterministic models of breeding scheme designs that incorporate genomic selection. *Journal of Dairy Science*, *93*(11), 5455–5466. https://doi.org/10.3168/jds.2010-3256

R Development Core Team (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna.* Retrieved from http://www.R-project.org.

Saccenti, E., & Timmerman, M. E. (2016). Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. *Journal of Proteome Research*, *15*(8), 2379–2393. https://doi.org/10.1021/acs.jproteome.5b01029

SAS Institute Inc. (2017). SAS software 9.4. *SAS Institute Inc., Cary, NC, USA.*

Sharma, N., Singh, N. K., & Bhadwal, M. S. (2011). Relationship of somatic cell count and mastitis: An overview. *Asian-Australasian Journal of Animal Sciences*, *24*(3), 429–438. https://doi.org/10.5713/ajas.2011.10233

Soyeurt, H., Bastin, C., Colinet, F. G., Arnould, V. M.-R., Berry, D. P., Wall, E., Dehareng, F., Nguyen, H. N., Dardenne, P., Schefers, J., Vandenplas, J., Weigel, K., Coffey, M., Théron, L., Detilleux, J., Reding, E., Gengler, N. & McParland, S. (2012). Mid-infrared prediction of lactoferrin content in bovine milk: potential indicator of mastitis. *Animal*, *6*(11), 1830–1838. https://doi.org/10.1017/s1751731112000791

Soyeurt, H., Colinet, F. G., Arnould, V. M.-R., Dardenne, P., Bertozzi, C., Renaville, R., Portelle, D. & Gengler, N. (2007). Genetic Variability of Lactoferrin Content Estimated by Mid-Infrared Spectrometry in Bovine Milk. *Journal of Dairy Science*, *90*(9), 4443–4450. https://doi.org/10.3168/jds.2006-827

Soyeurt, H., Dehareng, F., Gengler, N., McParland, S., Wall, E., Berry, D. P., Coffey, M. & Dardenne, P. (2011). Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *Journal of Dairy Science*, *94*(4), 1657–1667. https://doi.org/10.3168/jds.2010-3408

Szymańska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check : validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Official Journal of the Metabolomic Society*, *8*(Suppl 1), 3–16. https://doi.org/10.1007/s11306-011-0330-3

Vanlierde, A., Vanrobays, M.-L., Dehareng, F., Froidmont, E., Soyeurt, H., McParland, S., Lewis, E., Deighton, M. H., Grandl, F., Kreuzer, M., Gredler, B., Dardenne, P., & Gengler, N. (2015). Hot topic: Innovative lactation-stage-dependent prediction of methane emissions from milk mid-infrared spectra. *Journal of Dairy Science*, *98*(8), 5740–5747. https://doi.org/10.3168/jds.2014-8436

Visentin, G., McDermott, A., McParland, S., Berry, D. P., Kenny, O. A., Brodkorb, A., Fenelon, M. A., & De Marchi, M. (2015). Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. *Journal of Dairy Science*, *98*(9), 6620–6629. https://doi.org/10.3168/jds.2015-9323

Visentin, G., Penasa, M., Gottardo, P., Cassandro, M., & De Marchi, M. (2016). Predictive ability of mid-infrared spectroscopy for major mineral composition and coagulation traits of bovine milk by using the uninformative variable selection algorithm. *Journal of Dairy Science*, *99*(10), 8137–8145. https://doi.org/10.3168/jds.2016-11053

Wallén, S. E., Prestløkken, E., Meuwissen, T. H. E., McParland, S., & Berry, D. P. (2018). Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. *Journal of Dairy Science*, *101*(7), 6232–6243. https://doi.org/10.3168/jds.2017-13874