



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Supervised Machine Learning for the Quantification of
Cotton/Polyester Textiles Using IR Spectroscopy“

verfasst von / submitted by

Sebnem Sara Yayla

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien, 2025 / Vienna, 2025

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

UA 066 645

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Data Science

Betreut von / Supervisor:

Univ.-Prof. Dipl.-Ing. Dr. Jürgen Zanghellini

Acknowledgements

I would like to sincerely thank my academic advisors, David Lilek and Christian Schimper, for their guidance, continuous feedback and support throughout this thesis. Their expertise and advice were invaluable to the successful completion of this endeavor. I would also like to thank my thesis supervisor Jürgen Zanghellini for his constructive feedback and support during this thesis.

This study was conducted at the Josef Ressel Centre for Recovery Strategies of Textiles, facilitated by Christian Doppler Forschungsgesellschaft (CDG). The financial support by the Austrian Federal Ministry of Labor and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

On a personal note, I am deeply grateful to my boyfriend, Andi, for his emotional support and encouragement throughout this journey. I also want to express my gratitude to my classmates, from whom I have learned a lot and who brought joy to my studies. Finally, I owe my heartfelt thanks to my parents, whose unconditional love and encouragement enabled me to pursue my studies and complete this thesis.

Abstract

Textile waste is a significant environmental challenge exacerbated by the resource-intensive nature of textile production and the fast fashion industry. Since the 1980s, global textile fiber production has grown by about 3% annually, reaching 107 million tons in 2018. However, only 12% of the annual fiber production for consumer clothing is recycled, with the majority ending up in landfills or incinerators. This thesis discusses the efficient sorting in textile recycling by applying an advanced textile quantification method. It focuses on the application of Infrared (IR) Spectroscopy, a non-destructive method known for its ability to identify textile compositions through unique spectral fingerprints. The aim is to realize the use of IR Spectroscopy to accurately estimate the blend ratios of polyester and cotton in textiles in a non-destructive and efficient manner.

The methodology covers the use of both Mid-Infrared (MIR) and Near-Infrared (NIR) spectroscopy to increase the accuracy and reliability of textile composition analysis. By combining spectroscopic data with seven different machine learning models, this study aims to improve the predictive capabilities of textile recycling technologies. The project includes determining the correct sample size and analyzing how regression models perform with different preprocessing methods.

By using a Bootstrap-based margin of error approach, it was possible to determine the minimum number of measurement spots required to obtain statistically reliable spectra. The combination of Near-Infrared (NIR) spectroscopy with machine learning models provided a fast and accurate approach for textile quantification. We observed that the choice of preprocessing method played a critical role in regression accuracy and model generalization. Overall, Partial Least Squares Regression, Least Squares Support Vector Machine as well as Kernel Ridge Regression emerged as the best performing solutions for quantifying cotton content in blended textiles. Our results also indicated that tree-based models were prone to overfitting, while Convolutional Neural Networks were unstable. The findings set the stage for further research into multi-fiber textile blends and the incorporation of complementary spectroscopic techniques.

Kurzfassung

Textilabfälle stellen eine erhebliche Umweltbelastung dar, die durch den ressourcenintensiven Charakter der Textilproduktion und die Fast-Fashion-Industrie noch verschärft wird. Seit den 1980er Jahren ist die weltweite Textilfaserproduktion jährlich um etwa 3% gewachsen und erreichte 2018 107 Millionen Tonnen. Allerdings werden nur 12% der jährlichen Textilproduktion für Verbraucherbekleidung recycelt, der Großteil landet auf Mülldeponien oder in Verbrennungsanlagen. Diese Arbeit befasst sich mit der effizienten Sortierung von Textilabfällen durch Anwendung von Methoden zur Textilquantifizierung. Der Schwerpunkt liegt auf der Anwendung von Infrarotspektroskopie (IR), einer zerstörungsfreien Methode, die Textilzusammensetzungen anhand ihre spektralen Fingerabdrücke identifizieren kann.

Ziel ist es, die IR-Spektroskopie zu nutzen, um den Anteil von Baumwolle und Polyester in Textilien zu bestimmen. Die Methodik umfasst den Einsatz von Mittelinfrarot- (MIR) und Nahinfrarot- (NIR) Spektroskopie, um die Genauigkeit und Zuverlässigkeit der Analyse zu erhöhen. Durch die Kombination spektroskopischer Daten mit sieben verschiedenen Regressionmodellen zielt diese Studie darauf ab, die Sortierung von Textilabfällen während des Recyclingprozesses zu verbessern. Das Projekt umfasst die Bestimmung der richtigen Stichprobengröße sowie den Vergleich verschiedener Regressionsmodelle mit unterschiedlichen Methoden der Datenvorverarbeitung.

Durch einen Bootstrap-basierten Ansatz konnte die Mindestanzahl an Messungen bestimmt werden, die für statistisch zuverlässige Spektren erforderlich ist. Die Kombination von Nahinfrarot- (NIR) Spektroskopie mit Regressionsmodellen ermöglichte einen schnellen und genauen Ansatz zur Textilquantifizierung. Wir stellten fest, dass die Wahl der Vorverarbeitungsmethode eine entscheidende Rolle für die Prognosegenauigkeit und die Generalisierbarkeit der Modelle spielte. Insgesamt erwiesen sich Partial Least Squares Regression, Least Squares Support Vector Machine sowie die Kombination aus Kernel Ridge Regression und der Vorverarbeitungsmethode Standard Normal Variate als die besten Lösungen zur Quantifizierung des Baumwollgehalts in Mischtextilien. Unsere Ergebnisse zeigten auch, dass Modelle basierend auf Entscheidungsbäumen und Convolutional Neural Networks zu Overfitting neigten. Die Erkenntnisse bilden die Grundlage für weitere Forschungen zu Mehrfaser-Textilmischungen und der Einbeziehung ergänzender spektroskopischer Techniken.

Contents

Acknowledgements	i
Abstract	iii
Kurzfassung	v
1. Introduction	1
2. State of Research	5
2.1. Spectroscopic Techniques	5
2.1.1. NIR	5
2.1.2. MIR	6
2.2. Data Analysis Methods for Spectral Data	7
2.2.1. Properties of Spectral Data	7
2.2.2. Preprocessing Methods	7
2.2.3. Dimension Reduction	10
2.3. Resampling Methods	11
2.3.1. Jackknife Resampling	11
2.3.2. Bootstrap Resampling	12
2.4. Machine Learning Models	13
2.4.1. Linear Regression Techniques	14
2.4.2. Nonlinear Regression Techniques	14
2.4.3. Deep Learning Techniques	16
3. Methodology	17
3.1. Dataset Creation	17
3.1.1. Measurements	17
3.1.2. Data Overview	18
3.1.3. Data Cleaning	19
3.2. Reproducibility Experiments	20
3.2.1. Variance Estimation with Jackknife Resampling	21
3.2.2. Variance Estimation with Bootstrap Resampling	22
3.2.3. Margin of Error Approach with Bootstrap Resampling	23
3.3. Regression Pipeline	24
3.3.1. Selecting Subset	24
3.3.2. Spectra Pretreatment	25
3.3.3. Dimension Reduction	26

Contents

3.3.4. The Train-Test Split Strategy	27
3.3.5. Cross-Validation Strategy	27
3.3.6. Selected Algorithms and Hyperparameter Space	27
3.3.7. Regression Performance Metrics	29
3.4. Infrastructure and Software	30
4. Results and Discussion	31
4.1. Reproducibility	31
4.1.1. Comparison of Jackknife and Bootstrap Resampling	31
4.1.2. The Results of Margin of Error Approach with Bootstrap Resampling	32
4.2. Preprocessing	34
4.3. Prediction Model Performance	36
4.3.1. Comparison of Regression Models and Preprocessing Methods on NIR Spectra	38
4.3.2. Regression Model Performance on MIR Spectral Data in Compar- ison to NIR	56
5. Conclusion and Outlook	59
5.1. Conclusion	59
5.2. Outlook	60
Bibliography	61
A. Appendix	71

1. Introduction

Textile waste is a growing environmental issue due to the resource-intensive nature of textile production and the fast fashion industry [JLLY22]. Since the 1980s, global textile fiber production has grown by about 3% each year, reaching 107 million tons in 2018. Each year, 53 million tons of fiber are produced for clothing. Only 12% of this amount of fiber is recycled, while 73% ends up in landfills or incinerators [MRS20]. Despite the high waste generation, the amount of textiles produced for clothing and footwear alone is expected to exceed 100 million in 2030. The current textile production system contributes significantly to environmental challenges due to its high resource consumption and waste production, as it is predominantly linear and operates on a “take-make-dispose” model [VNL19]. Transitioning to a circular economy focusing on reuse, repair and recycling can significantly reduce the environmental impact of textile waste [NMH⁺20].

However, textile transformation is not an easy process. The management of textile waste poses a significant challenge as traditional separation and sorting methods are labor-intensive, costly, prone to inaccuracies and not particularly efficient [DMHG⁺21]. These methods face challenges in handling the large and complex waste streams generated by the global textile industry, which mainly uses polyester (PET) and cotton fibers. In addition to being inefficient, the chemicals used in recycling damage textile fibers [ROD⁺06]. Due to these circumstances, it is important that textile waste is sorted quickly and harmlessly.

The Josef Ressel Center "Recovery Strategies of Textiles (ReSTex)" is dedicated to developing methods to improve textile recycling. One objective is to characterize cotton-polyester blended textiles containing varying proportions of these fibers using Near-Infrared (NIR) and Mid-Infrared (MIR) spectroscopy for more efficient textile recycling process of different fibers and fabrics. The thesis is supported by ReSTex.

The adoption of spectroscopic techniques for textile recycling represents an innovative approach that aims to improve both sustainability and efficiency in the sector [TCLW19, LDHD07]. Spectroscopic techniques can be used to detect and measure textile components very effectively and quickly by revealing unique light absorbance or reflection patterns. Among spectroscopic techniques, Infrared (IR) Spectroscopy is the most widely used method as it provides a non-destructive tool to identify textile compositions based on their unique spectral fingerprints. IR spectroscopy is a technique used to measure the interaction of infrared radiation with matter. Common functional groups in textiles, such as O-H, N-H and C-H, contribute to different absorption properties in the IR region [Tho18]. Each type of molecule in a textile contributes to its unique spectral signature depending on its molecular structure and the types of chemical bonds it contains. For example, the cellulose in cotton, the polyamide in nylon and the polyester components all have different spectral properties. With IR Spectroscopy, chemical bonds in molecules

1. Introduction

can be observed to absorb infrared light at different characteristic wavelengths. By measuring these absorption patterns, scientists can determine the molecular structure and composition of a sample [Hou09, ZYDW19]. Once the types of fibers are identified, textiles can be more effectively sorted into categories that share similar properties.

When textiles are made from blends of different fibers, their spectra become a composite of the spectra of each component fiber. The more a particular type of fiber is present, the stronger and more dominant its spectral properties will be in the IR spectrum. Based on this property, in this thesis, we aim to accurately estimate the blend ratios of a textile product for which spectral data is provided. The scope of this thesis is limited to the quantification of polyester and cotton because these two materials make up the majority of the world's textile production, of which 24% are cotton and 52% are polyester [Ša22].

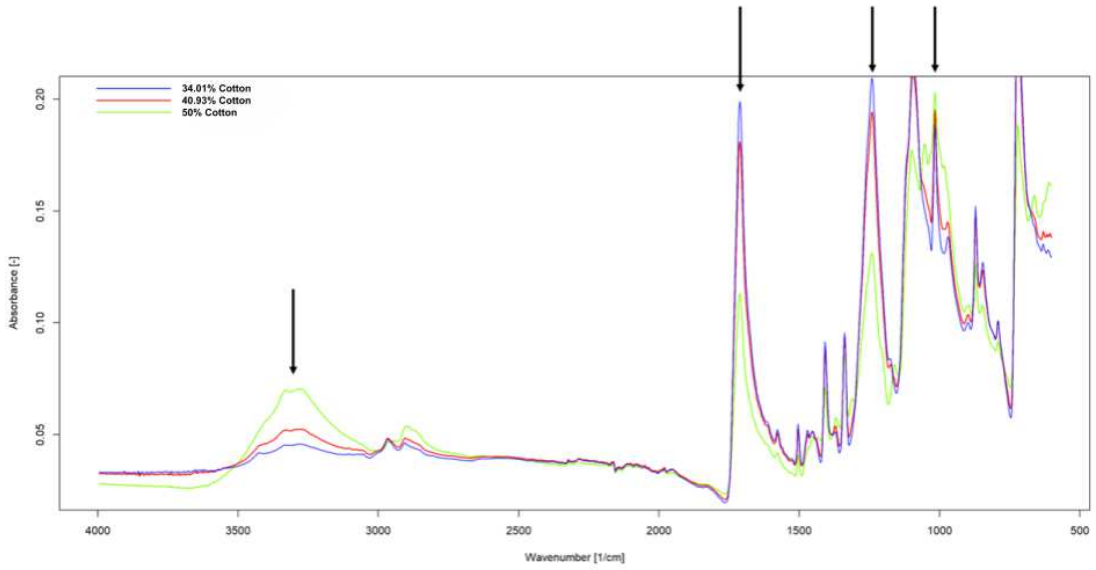


Figure 1.1.: The example of an MIR absorbance spectra of cotton/polyester blended textiles with varying cotton content (34.01%, 40.93%, and 50% cotton) across a range of wavenumbers from 4000 to 500 cm^{-1} . The absorbance increases with higher cotton content, particularly noticeable in the regions between 3000 to 3500 cm^{-1} and around 1000 cm^{-1} while polyester absorbance peaks occur strongly in the regions 1500 to 1800 cm^{-1} and 1100 to 1300 cm^{-1} .

Within the context of this thesis, we implement a spectroscopic analysis methodology through a database where spectral measurements of a set of cotton-polyester blended textiles are recorded. To create the database, a large number of textile samples need to be measured, which poses a practical challenge. We strive to measure the smallest number of points to save time, effort, and resources, but without compromising statistical accuracy. Therefore, the thesis seeks to answer the following question:

- 1- What is the minimum number of measurement points per textile specimen required

to ensure a statistically representative sample of textile materials?

Identifying the minimum number of measurement points per textile specimen is important for ensuring that spectral measurements represent the cotton-polyester blends. This not only reduces resource expenditure but also allows for the creation of a statistically robust and scientifically sound dataset. Through resampling methods, this thesis identifies an efficient strategy to optimize measurement processes while retaining the integrity and informativeness of the dataset.

Predicting the cotton percentage in a textile sample using IR spectroscopy represents a quantitative multivariate calibration problem requiring a robust regression model. Such a model must be developed using a large dataset that captures the main sources of spectral variability, often necessitating advanced mathematical preprocessing, namely a baseline treatment. For the analysis of cotton-polyester blended textiles, selecting a regression model that balances prediction accuracy and computational efficiency is essential. This thesis explores the second and third questions:

- 2- Which baseline treatment methods are most effective in enhancing regression model performance, particularly in handling spectral variability and noise?
- 3- Which supervised machine learning model achieves the best performance for regression tasks, considering both prediction accuracy and generalizability? What is the optimal hyperparameter space to explore for the models?

Since predicting the cotton or polyester content involves dealing with hundreds of features (the absorbance at different wavenumbers), the relation between the absorbance spectrum and the cotton-polyester ratio is highly complex. Due to the multivariate nature of the data, our approach is to evaluate the problem with a multivariate calibration technique which includes data preprocessing, dimensionality reduction, building a regression model and evaluation of the results. We examine different common preprocessing methods and regression algorithms and compare the results to determine the most suitable method.

We systematically explore various baseline treatment methods, including baseline correction techniques, to assess their effectiveness in reducing noise and accounting for spectral variability. Additionally, we evaluate and compare different regression algorithms to identify the most suitable model based on prediction accuracy and generalizability. Beyond accuracy, it is important to consider how well the model generalizes. This means that it should perform well on new, unseen data and not underfit or overfit. This comprehensive analysis aims to determine the optimal combination of baseline treatment methods and regression models for predicting cotton-polyester content in textile samples.

This thesis is organized into five chapters. Chapter 1 introduced the research questions and gave a brief overview of the current state of textile recycling, with a focus on the adoption of spectroscopic techniques. Chapter 2 discusses the relevant literature, offering a review of previous work in this area. Chapter 3 outlines the methodology employed in this study. The results and their implications are presented and analyzed in Chapter 4. Finally, Chapter 5 summarizes the key findings, draws conclusions, and highlights directions for future research.

2. State of Research

Effective treatment of textile waste requires a comprehensive understanding of its composition and quality, given the wide variety of fibers, dyes and contaminants present. By incorporating advanced spectroscopic techniques and leveraging data analysis models, fibers can be classified or quantified. In literature, we observe that classification tasks where the goal is to assign each sample to one of several predefined categories or classes based on its spectra are more predominant [LDHD07, PLPV17, ZYDW19, LMWX19] than quantification tasks where percentage composition of each type of fiber in a blend is measured with regression models [ROD⁺06, CTL20]. In this section, we present spectroscopic techniques and the regression models used in textile quantification which are relevant to the thesis objective.

2.1. Spectroscopic Techniques

The characterization of textiles using spectroscopic techniques has become increasingly sophisticated, focusing on identification and quantification of the composition of textile materials. The core spectroscopic techniques utilized in textile analysis include near-infrared (NIR) spectroscopy and mid-infrared (MIR) spectroscopy, each offering unique advantages in terms of sensitivity, depth of penetration, and specificity to certain chemical bonds.

2.1.1. NIR

NIR spectroscopy utilizes NIR type of light which ranges from 13000 cm^{-1} to 4000 cm^{-1} and detects absorption areas where the molecules vibrate. NIR spectroscopy is sensitive to the overtone and combination vibrations of C–H, N–H, and O–H bonds, making it suitable for analyzing organic materials such as textiles [CH07].

NIR spectroscopy can be done using Fourier Transform Infrared Spectroscopy (FT-IR) technique. Traditional dispersive infrared (IR) spectroscopy involved checking each IR frequency individually, which is time-consuming. FT-IR is faster and more accurate, with a better signal-to-noise ratio than IR spectroscopy because it simultaneously checks all IR wavelengths. With the use of specialized accessories, FT-IR spectrometers can generate both transmission and reflection spectra. Reflectance spectroscopy is particularly effective for analyzing materials that scatter infrared light and do not transmit it. This method has become a widely adopted standard procedure in FT-IR spectroscopy [Tho18]. Margariti [Mar19] emphasizes that the use of FT-IR spectroscopy in reflectance mode offers a truly non-invasive and non-destructive way to analyze materials. This technique

2. State of Research

effectively identifies the characteristic compounds of organic matter preserved within the fibers, without damaging the material under study.

NIR spectroscopy offers many advantages. Optical materials for NIR spectroscopy are generally cheaper than those required for IR spectroscopy. It allows a sample to be analysed in situ in its original state, and thus NIR spectroscopy can be used in a variety of industries. It is also suitable for non-destructive testing of specimens of different shapes and thicknesses [OM21]. In addition, the special rotating sphere we used, provided a large scanning area which covers a spot size of 10 cm and made sample analysis more robust with the ability of NIR to penetrate deep into matter. For these reasons, NIR spectroscopy is highly effective for the measurement of textile materials.

NIR spectroscopy is not without its limitations. NIR bands arise from overtones and combination modes. It means that instead of transitioning from the ground state to the first vibrational level, molecules are excited to the second, third, or even higher vibrational levels. The presence of many overlapping broad bands leads to strong multicollinearity, making band assignments challenging [OM21]. Also, the spectra can be significantly influenced by surface contamination and additives, which might alter the accuracy of the fiber identification [PKVL19].

2.1.2. MIR

The Mid-Infrared (MIR) region covers wavenumbers from $4000\text{--}400\text{cm}^{-1}$, which are lower than those in the Near-Infrared (NIR) region. In MIR region, fundamental vibrations of the molecules occur which makes MIR spectroscopy highly sensitive. MIR spectroscopy is effective for identifying functional groups, such as C=O, O-H, and C-S groups [Oza21].

A prevalent technique used in MIR spectroscopy is the Attenuated Total Reflectance (ATR). ATR Cell, made from water-insoluble crystalline materials with high refractive indices, improves spectrum acquisition from fibers. When infrared radiation enters the crystal, it undergoes total internal reflection, creating a wave that extends near the surface of the ATR crystal. When a sample touches the ATR crystal, an infrared spectrum is generated on the sample surface, since the wave penetrates to a shallow depth. This wave decreases exponentially with distance, making ATR insensitive to sample thickness and suitable for the analysis of thick and highly absorbent materials. Furthermore, based on the specific configuration, ATR crystals can generate multiple internal reflections, which enhances the sensitivity of the measurement [Tho18].

In the thesis, MIR spectroscopy was performed using a single-reflectance ATR crystal. Due to the localized nature of ATR measurements, the scan area was limited to the crystal's contact region and covered a spot size of 2-3 mm. Because of the small scan area of the ATR crystal, only a limited region of the textile samples was measured, which may influence the representativeness of the spectral data. Table 2.1 provides a comparison of NIR and MIR spectroscopy, highlighting their respective advantages and limitations.

Table 2.1.: Comparison of NIR and MIR spectroscopy techniques

	Advantages	Disadvantages
NIR	• Large scan area • Light penetrates deeper into the sample	• Fundamental vibrations of molecular bonds do not occur • Overtones lead to broader bands
MIR	• Fundamental vibrations of molecular bonds occur, making the spectral profiles sensitive to chemical composition	• Very small part of total sample area is inspected • Light penetration is limited, focusing on surface-level analysis

2.2. Data Analysis Methods for Spectral Data

2.2.1. Properties of Spectral Data

Spectral data typically consists of hundreds to thousands of wavelengths, leading to high-dimensional datasets. This can result in complex data structures where the number of features (wavelengths) exceeds the number of samples (observations) [HGM20]. This high dimensionality, compounded by the nonlinear relationship between spectral features and sample properties, makes linear classification methods inadequate for capturing complex nonlinear relationships in spectral data.

Moreover, spectral data are prone to noise and variability due to factors such as instrumental limitations, sample impurities, and environmental conditions. Measurements can fluctuate not only between different days but also within the same day, driven by changes in operational and environmental conditions. Temperature fluctuations can cause instrument readings to drift, particularly before the instrument reaches thermal equilibrium. Similarly, high humidity can impact measurements by altering sample properties and leading to moisture absorption within the instrument, and contributing to data inconsistencies [BC07].

Noise and variability can obscure underlying patterns in the data, making it challenging to distinguish between signal and noise. To improve the quality of noisy spectral data, preprocessing methods for spectral data are applied, followed by dimension reduction techniques to extract meaningful information from the data [MN89].

2.2.2. Preprocessing Methods

Preprocessing is a crucial step in chemometric data analysis as it enhances the quality and interpretability of the spectral data by reducing variability and noise that can interfere with modeling and analysis. It helps address issues such as scattering effects and baseline variations, ensuring that the data is more consistent and comparable. Preprocessing

2. State of Research

methods can operate at both the sample level and across entire datasets [BPS98]. In this section, some popular methods including normalization, smoothing, and baseline correction are explained.

Standard Normal Variate

Standard Normal Variate (SNV) normalizes spectral data to have a zero mean and unit variance, helping to remove scaling effects [BYX⁺16].

The Standard Normal Variate (SNV) transformation of a spectrum $X = [x_1, x_2, \dots, x_n]$, where n is the number of spectral points, is given by:

$$X_{\text{SNV}} = \frac{X - \mu_X}{\sigma_X}$$

where: $\mu_X = \frac{1}{n} \sum_{i=1}^n x_i$ is the mean of the spectrum, and $\sigma_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_X)^2}$ is the standard deviation of the spectrum [BC07].

Running Median Smoother

Smoothing helps to filter non-systematic noise from the measured spectral data. One commonly used smoothing method is the Running Median Smoother (Median Window) technique. This approach involves applying a median filter within a sliding window across the spectral data. By replacing each data point with the median value of the points within the window, this technique effectively reduces noise while preserving important spectral features, making it a preferred choice in many spectral data analysis applications [BPS98].

Savitzky–Golay Filter

The Savitzky–Golay filter is a smoothing filter designed to enhance data by reducing noise while preserving important signal characteristics, such as peak height and width. This is accomplished by fitting low-degree polynomials to moving windows of data using the least squares method and bringing the filtered data to the center of the window [PS03, FMB18].

The core idea of the Savitzky–Golay filter is to fit a polynomial of a specified degree p to a moving window of data points. The window size, defined as $2m + 1$, ensures symmetry around a central point. By minimizing the error between the fitted polynomial and the data points within the window, the filter achieves smoothing while preserving key features of the signal. To mathematically implement the filter, a matrix S is constructed, where each row represents the powers of the positions of the data points within the window. For instance, if the window is centered at zero, the rows of S would include terms like $(-m)^0, (-m)^1, \dots, (-m)^p$ for the leftmost point, and similar terms for subsequent points. Using the least-squares solution, the convolution coefficients are computed as:

$$G = S(S^T S)^{-1} S^T$$

where G is the matrix of coefficients used for generating smoothed values. These coefficients are then applied to the input data through convolution, allowing the Savitzky-Golay filter to produce a smoothed value for each data point [ZTZL14].

A major advantage of the Savitzky-Golay filter is that it smooths data while preserving important features like peak positions and widths, which simpler methods such as Running Median Smoother might distort. Additionally, because it fits polynomials to the data, it can also estimate derivatives, making it useful for detecting subtle trends in noisy signals [ZTZL14].

Detrend

Detrending is a baseline correction technique that addresses the increasing baseline trend typically observed in near-infrared (NIR) diffuse reflectance spectra. This method involves fitting a second-degree polynomial to model the baseline curvature, effectively minimizing baseline variations across spectra. By subtracting this polynomial from each spectrum, detrending eliminates baseline shifts and curvatures, thereby standardizing the baseline across samples. This correction enhances the clarity of spectral features and mitigates multicollinearity issues, making it easier to analyze the chemical information within the spectra [BDL89].

Let a spectrum $X = [x_1, x_2, \dots, x_n]$ be a set of spectral data points measured at different wavelengths or indices. A second-degree polynomial can be written as $f(x) = a_0 + a_1x + a_2x^2$ where x represents the wavelength or index position within the spectrum, and a_0 , a_1 , and a_2 are coefficients determined by fitting the polynomial to the spectral data. We can remove the baseline trend by subtracting $f(x)$ from the original spectral data points. The detrended spectrum then becomes

$$X_{detrend} = X - f(x)$$

[BDL89].

Asymmetric Least Squares

The Asymmetric Least Squares (ALS) technique for baseline correction operates by fitting a smooth curve to the lower parts of the spectrum, treating it as an estimate of the baseline. The key to ALS is asymmetric weighting, which favors data points lying near or below the baseline. This is done by assigning points with high positive residuals (above the baseline) low weights in so they exert less influence on the fitting and assigning points below the baseline higher weights, leading the baseline curve to be smooth and robust to outliers [Gal20].

FillPeaks

4S Peak Filling is a baseline correction method that includes 4 steps, namely "Smoothing", "Subsampling", "Suppression", and "Stretching" to remove background noise. In the first step, Smoothing, Whittaker smoothing is utilized to reduce the noise peaks. Next, in the

2. State of Research

Subsampling step, the spectral data is divided into smaller bins, and a selected number of bins are sampled, with the minimum value from each sampled bin retained. This step improves computational efficiency and simplifies the spectral data. In the Suppression step, a moving window slides across the spectrum, replacing each point with the minimum of its current value and the mean value within the window. This process is repeated in both directions (left-to-right and right-to-left), with the window size shrinking logarithmically over iterations to enhance the baseline estimation. Finally, in the Stretching step, the estimated baseline is interpolated back to the original spectrum length [Lil15].

Multiplicative Scatter Correction

Multiplicative Scatter Correction (MSC) is a normalization technique used in spectroscopic data analysis to correct scaling and offset variations often caused by scattering effects and other physical phenomena. MSC works by adjusting each spectrum in a dataset to align closely with a reference spectrum, usually the mean of the dataset.

The transformation in MSC is based on a linear model:

$$x = a + b\hat{x} + e$$

where x is the observed spectrum of a sample, $\hat{x}$ is the reference spectrum (often the mean spectrum of the dataset), a is the offset term, b is the scaling factor, and e represents the residual error.

To correct a spectrum, MSC calculates a and b using least squares regression. The corrected spectrum, x^* , is obtained by:

$$x^* = \frac{(x - a)}{b} = \hat{x} + \frac{e}{b}$$

[WSB08]

2.2.3. Dimension Reduction

Spectral data is considered high-dimensional as hundreds or thousands of wavelengths are measured to capture the complete spectrum of an object or material. This high dimensionality comes from the need to accurately capture the detailed spectral signatures of materials, which are often complex to differentiate between subtle variations in chemical composition [LMWX19]. The high dimensionality of spectral data poses several challenges, including the curse of dimensionality, computational complexity, overfitting, and model interpretability [HMOR05].

In the data analysis phase, it is aimed to classify the components of the textile material based on the spectral data and to quantify these classes. Due to the high dimensionality, various machine learning algorithms have been used with dimension reduction techniques in many studies for the classification of spectral data.

A frequently employed dimension reduction and exploratory data analysis technique is Principal Component Analysis (PCA). PCA performs an orthogonal linear transformation on the original variables, converting them into a smaller number of uncorrelated principal

components that preserve most of the variance in the data while minimizing information loss. By focusing on the principal components, PCA can help in filtering out noise from the data which improves the quality of the analyses [AAD⁺15, LDHD07]. PCA is a tool for exploratory data analysis as it helps outlier and cluster identification and provides insights into the structure of spectral data [MRS20]. While PCA can reveal clusters or patterns within the data, it does not label these clusters, as it is an unsupervised technique. For discrimination tasks where classifying observations into predefined groups is required, supervised techniques are preferred because they utilize existing labels to effectively learn and distinguish between classes [NTB⁺16].

2.3. Resampling Methods

In the context of NIR spectroscopy, variability is considered as part of both the measurement process and the analysis process. Resampling methods provide a statistical framework to evaluate the variability by generating alternative versions of the dataset and assessing the consistency of the underlying statistics. Specifically, in IR spectroscopy, where measurements can vary due to noise or material differences, resampling is useful for checking how stable and consistent the data is. These techniques are particularly useful when working with limited sample sizes [McI16]. In this thesis, resampling is applied to determine the minimum number of spectral measurements required to ensure statistical representativeness. We use methods like Jackknife and Bootstrap to measure variability and estimate the minimum sample size needed for reliable analysis.

2.3.1. Jackknife Resampling

Jackknife resampling was developed as a method to estimate the bias and variance of a statistical estimator, primarily for cases where traditional analytical methods were challenging or infeasible. In 1949, Maurice Quenouille introduced the method to reduce the bias in an estimator by systematically removing individual observations from a dataset, computing the statistic of interest for each subset, and then averaging the results. This process helped address the bias by showing how each data point influenced the overall estimate [Que49, FS06].

To mathematically illustrate the jackknife resampling method, according to McIntosh [McI16], suppose we have a sample of n observations, denoted by $X = x_1, x_2, \dots, x_n$, which are drawn independently and identically (iid) from a distribution. We aim to estimate θ using an estimator $\hat{\theta}$. The first step is to compute $\hat{\theta}$, based on all n observations. This serves as the initial estimate of θ . For each i from 1 to n , remove the i -th observation from the sample to create a "leave-one-out" sample, denoted by $X_{[i]} = \{x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}$. Then compute the estimator for each subset, $\hat{\theta}_{(i)}$. The estimate of the bias for $\hat{\theta}(X)$ is given by

$$\widehat{\text{Bias}}_{\text{jack}} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}).$$

2. State of Research

where the mean of the leave-one-out estimates is

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}$$

To obtain a bias-corrected estimate of θ , adjust the initial estimate $\hat{\theta}(X)$ by subtracting the estimated bias:

$$\hat{\theta}_{\text{jack}} = \hat{\theta} - \widehat{\text{Bias}}_{\text{jack}}$$

Substitute the bias expression to get the jackknife bias-corrected estimate of the parameter θ :

$$\hat{\theta}_{\text{jack}} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}.$$

The jackknife variance estimate of $\hat{\theta}$ is defined as:

$$\widehat{\text{Var}}_{\text{jack}}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^2.$$

While the classical jackknife removes one observation at a time ($d=1$), the method can be generalized to the *delete-d jackknife*, where subsets of size d are removed from the full dataset [FS06].

Let $s_j \subset X$ be a subset of size d to be deleted from the sample. The estimator evaluated on the remaining $n-d$ observations is denoted $\hat{\theta}_{(s_j)}$, and the mean over all such subsets is:

$$\hat{\theta}_{(-d)} = \frac{1}{N} \sum_{j=1}^N \hat{\theta}_{(s_j)}, \quad \text{where } N = \binom{n}{d}.$$

The delete- d jackknife variance estimator is:

$$\widehat{\text{Var}}_{J_d}(\hat{\theta}) = \frac{n-d}{dN} \sum_{j=1}^N \left(\hat{\theta}_{(s_j)} - \hat{\theta}_{(-d)} \right)^2.$$

This approach approximates the sampling variability of $\hat{\theta}$ while it is particularly useful for non-smooth estimators such as quantiles [FS06].

2.3.2. Bootstrap Resampling

In his 1979 paper, Efron introduced the bootstrap resampling method as a powerful, data-driven technique for estimating the distribution of a statistic when the underlying population distribution is unknown [Efr79]. The bootstrap method allows us to estimate the variability and distribution of a sample statistic (like the mean, median, standard deviation, or regression coefficients) without relying on strong assumptions about the

population distribution. It’s particularly useful when the sample size is small or when the underlying distribution is complex or unknown.

Generate B bootstrap samples $X_1^*, X_2^*, \dots, X_B^*$ from X by sampling with replacement. Each bootstrap sample is of size n . For each bootstrap sample X_b^* the sample mean computed as

$$\bar{X}_b^* = \frac{1}{n} \sum_{i=1}^n X_{i,b}^*$$

where $X_{i,b}^*$ denotes the i -th observation in the b -th bootstrap sample. The bootstrap estimate of the mean $\hat{\mu}_{\text{bootstrap}}$ is simply the average of the means calculated from each bootstrap sample

$$\hat{\mu}_{\text{bootstrap}} = \frac{1}{B} \sum_{b=1}^B \bar{X}_b^*$$

To estimate the variance of the mean, calculate the variance of the bootstrap sample means. This gives an estimate of the sampling variance of the mean:

$$\hat{\sigma}_{\text{bootstrap}}^2 = \frac{1}{B-1} \sum_{b=1}^B (\bar{X}_b^* - \hat{\mu}_{\text{bootstrap}})^2$$

These formulas were formally introduced by Efron as part of the bootstrap framework [Efr79]. For consistency with Jackknife resampling notation, the mean from each resample $\bar{X}_b^*$ may also be denoted $\hat{\theta}_b^*$; the bootstrap mean estimate $\hat{\mu}_{\text{bootstrap}}$ as $\hat{\theta}_{\text{bootstrap}}$; and the bootstrap variance estimate $\hat{\sigma}_{\text{bootstrap}}^2$ as $\widehat{\text{Var}}_{\text{bootstrap}}(\hat{\theta})$.

The bootstrap method can be used to evaluate the required sample size for a specified confidence level and margin of error. A sample was drawn from a target distribution and repeatedly resampled to calculate confidence intervals for each mean, allowing the margin of error to be assessed. This process was repeated multiple times to examine how the margin varied with different sample sizes. The results indicated that a smaller sample size could achieve the desired confidence, demonstrating the bootstrap’s utility in optimizing sample sizes without relying heavily on distributional assumptions [Qum13].

A comparison of the advantages and disadvantages of the Jackknife and Bootstrap resampling methods is summarized in Table 2.2.

2.4. Machine Learning Models

Chemometrics combines statistics and mathematics to extract valuable insights from data generated by chemical experiments, including spectroscopy. Preprocessing methods, multivariate data analysis such as Principal Component Analysis (PCA) (discussed in Dimension Reduction section 2.2.3) and Partial Least Squares (PLS), classification, and prediction tasks are in focus of chemometrics [Wol95]. Machine Learning was introduced to the chemistry literature in the late 80s. Since then, the combination of chemometrics and Machine Learning (ML) in Spectroscopy has become essential for modern chemical

2. State of Research

Table 2.2.: Comparison table of Jackknife and Bootstrap resampling methods

	Advantages	Disadvantages
Jackknife	• Results are stable and robust since they depend on systematic resampling.	• Computationally expensive for large datasets (slower runtime).
Bootstrap	• Faster and simpler implementation with lower computational cost.	• Results have slightly higher variance due to random sampling.

analysis by increasing efficiency, precision and accuracy. ML excels in handling large and complex (non-linear) datasets, enables real-time monitoring and achieves predictive analytics [Jos23].

In this section, regression methods used for learning and predicting IR spectral data, which have been utilized previously in the literature, are discussed.

2.4.1. Linear Regression Techniques

There are several techniques used to perform linear regression in textile quantification.

Principal component regression (PCR) essentially uses Principal Component Analysis. Once the principal components of predictors in a dataset are defined, then PCR involves the linear regression of the variable to be predicted on these principal components [MN89].

Partial Least Squares Regression (PLSR) is one of the most widely used multivariate analysis techniques in chemistry which was designed to handle datasets characterized by variables that are highly correlated or multidimensional. With modern measuring instruments like spectrometers, the variables are often numerous and strongly correlated [WSE01]. PLSR addresses this challenge by integrating aspects of Principal Component Analysis and multiple linear regression, reducing complex data into a smaller set of latent variables [CTL20]. PLSR can be extended and enhanced by integrating it with various advanced methods. By combining PLSR with genetic algorithms (GA), Ruckebusch et al. [ROD⁺06] performed a quantitative analysis of cotton-polyester textiles with NIR spectroscopy.

2.4.2. Nonlinear Regression Techniques

Li et al. [LMWX19] explored a method for identifying components in woven fabrics using hyperspectral imaging and the K-nearest neighbors (KNN) classification algorithm. This technique classifies a sample by comparing it to its nearest neighbors within a feature space, assigning it to the most common category among those neighbors. The KNN algorithm's reliance on the proximity of samples for classification assumes that similar samples are likely to be in the same category, utilizing distance as a measure of similarity.

Support Vector Machine (SVM) is an effective machine learning algorithm used for both classification and regression. Its primary objective is to find the hyperplane in an N -dimensional space that creates the widest margin between the classes. In regression tasks, Support Vector Regressor finds a hyperplane that best fits the data while keeping as many points as possible within a specified tolerance ϵ . Thanks to the kernel trick, SVM handles complex datasets that cannot be linearly separable by mapping the input features into high-dimensional feature spaces where a linear separator can be found [CLYL04, LDHD07]. Kernel trick makes it possible to compute the inner products in a high-dimensional space implicitly, allowing the model to capture nonlinear relationships without explicitly transforming the input data [HSS08].

Least Squares Support Vector Machine (LSSVM) is a modified version of SVM that is popular in classification and regression tasks using spectral data. It replaces the inequality constraints in standard SVMs with equality constraints, leading to a simpler optimization problem [LADK⁺15]. This reformulation of the optimization problem as a set of linear equations rather than SVM's quadratic problem, improves computational efficiency [WH05]. Similar to SVM, LSSVM can handle nonlinear datasets using the kernel trick. For quantification of textile content, Liu et al. utilized LSSVM and Partial Least Squares (PLS) regression. They indicate that the LS-SVM model provides high accuracy in determining the cotton-polyester blend composition. Importantly, they demonstrate the advantage of applying advanced machine learning techniques, such as LSSVM, in spectroscopic analysis, as they exceed the performance of basic PLS models [LZL⁺18].

Kernel Ridge Regression (KRR) is closely related to LSSVM, in terms of relying on the kernel trick and regularization parameter. KRR keeps the structure of the Ridge Regression model and combines it with kernel methods [ALV07]. It can be used effectively in the analysis of spectral data because it works well with nonlinear data. Regularization term in the loss function (l_2 norm by default) helps to prevent overfitting [All23].

Random Forest belongs to the ensemble learning family. It creates many decision trees to be trained on a random subset of the features and data. This randomness ensures diversity among the trees. Final prediction is made by taking the majority vote of decision trees [Zhu20]. Random Forest is used for both classification and regression tasks, and it is known for its simplicity and ability to run efficiently on large data sets. One of the most advantageous features of Random Forest is its ability to handle missing values in the dataset. Random Forest uses bootstrapping, a resampling technique where multiple subsets of the original dataset are created with replacement, to train each tree. This approach reduces overfitting and improves model robustness [DSDSP18].

Another tree-based machine learning algorithm is Gradient Boosting. It creates shallow trees (weak learners) sequentially, whereby it tries to improve the error with each iteration. The method follows the principle of gradient descent, meaning that at each iteration it steps towards the negative gradient of the loss function, hence iteratively boosting the model to reduce prediction error [Fri01]. As Gradient Boosting is a sequential algorithm, it's hard to parallelize and therefore usually slower than Random Forest. A common implementation of Gradient Boosting is the XGBoost library (eXtreme Gradient Boosting) [XGB24].

2.4.3. Deep Learning Techniques

Deep learning (DL) is increasingly being used for spectral data modeling in NIR spectroscopy. DL offers significant advantages in chemometrics, particularly in handling non-linear relationships and multi-dimensional data. Unlike conventional chemometric techniques that require extensive manual feature engineering and preprocessing, DL models can automatically extract relevant features from raw spectral data. Recently, it outperformed traditional methods like Partial Least Squares (PLS) [MPM⁺22]. However, DL in chemometrics is not as widely developed as in other fields in machine learning due to several factors. First, spectral datasets are typically much smaller than the massive labeled datasets available in CV and NLP. As a result, DL's ability to generalize effectively is limited. Second, chemometric models require explainability, but DL models are often seen as "black boxes" with limited interpretability compared to traditional linear models. For example, PLS provides clear information on spectral properties through scores and loadings. Third, chemometric data is usually not open source and generating large datasets is highly labor-intensive. Lastly, chemometrics research groups, relied on MATLAB and R-based statistical modeling, while DL frameworks like TensorFlow and PyTorch require programming proficiency in Python [MPM⁺22].

According to Poth et al., Neural Networks (NNs) have been evaluated for predicting material composition from spectral data in bioprocessing applications. Their study compared different machine learning algorithms for material composition quantification, showing that NNs performed well in both accuracy and robustness. Although NNs performed better than PLSR, the training procedure is much easier for PLSR than NN because almost no hyperparameter tuning is required. In addition, linear algorithms like PLSR are much faster than NN in training large datasets [PMF⁺22]. The operations become very slow for a high-dimensional dataset.

Application of a DL model can be challenging due to the high-dimensional nature of the spectral data and the difficulty in acquiring enough number of samples. Usually, NNs require a large training dataset [CTL20, MPM⁺22]. Despite that, in chemometrics, shallow DL models with a depth of 20 layers can technically be effective for datasets containing less than 1000 samples. However, NIR data is much more limited. Surprisingly, shallower models with 1-3 convolutional layers have shown potential even on sample sets smaller than 100. With careful parameter selection, DL models can be trained on small NIR datasets. In addition, use of data augmentation techniques has an impact on improving the performance [MPM⁺22].

While DL models can be adapted to small datasets, they still remain highly prone to overfitting, especially when applied on high-dimensional data. Spectral data can easily contain hundreds of features. Thus, there is a risk of overfitting which leads to poor generalization on unseen data [DK99, V⁺03].

3. Methodology

The process begins with IR measurements, followed by transforming and transferring the spectral data for further analysis. To optimize spectral data quality, various sampling techniques and instrument parameters were studied, with measurements conducted using FTIR-ATR for mid-infrared (MIR) and diffuse reflectance for near-infrared (NIR) analyses. Subsequent steps include applying wavenumber restrictions and removing waterbands to refine the spectral data. Outliers are identified and excluded to enhance data quality, and baseline correction is performed to normalize the spectra.

A comprehensive spectroscopic analysis methodology was employed to build a database of spectral measurements for a range of cotton/polyester blended textiles. Appropriate statistical and machine learning techniques were then selected to analyze this data. These included models such as K-Nearest Neighbors Regression (KNN), Partial Least Squares (PLS), Least Squares Support Vector Machines (LSSVM), Random Forest Regression, Kernel Ridge Regression, eXtreme Gradient Boosting (XGBoost) and Convolutional Neural Networks (CNN), which were used to characterize textile blends. The effectiveness of these models in textile identification and degradation monitoring was evaluated using cross-validation and independent test sets, with metrics such as Root Mean Squared Error (RMSE) for regression tasks and coefficient of determination (R^2) to ensure the most possible accurate and robust analysis.

Additionally, a reproducibility analysis was conducted to ensure data reliability. If the sample size was deemed insufficient, additional data acquisition was undertaken. Once a sufficiently large dataset was obtained, regression models were developed to analyze the relationship between the spectral data and the cotton content of the samples. This systematic approach ensures accurate, reproducible results and robust model performance.

3.1. Dataset Creation

3.1.1. Measurements

In a real-world scenario, the cloth batches to be recycled have varying cotton ratios. Therefore, it was essential to prepare samples representing different cotton contents. This diversity in cotton ratios was important for two reasons: first, to accurately represent the composition of the recycled cloth batches, and second, to ensure that the machine learning algorithms were trained with a comprehensive range of data points reflecting different cotton content levels. Ten types of samples were prepared, with cotton ratios ranging from 50% to 27.01%. Each sample consisted of five square-shaped, white fabric specimens that underwent an enzymatic cellulose degradation process.

3. Methodology

The measurement parameters for the thesis were as follows: For MIR, 16 scans were conducted at an optical resolution of 4 cm^{-1} using ATR-MIR mode, with the specimens cleaned using ethanol and dried in air. Spectra were recorded between 4000 cm^{-1} and 600 cm^{-1} . For NIR, 16 scans were also performed at 4 cm^{-1} optical resolution using FT-IR mode. The spectra range was $12500\text{--}3600\text{ cm}^{-1}$. Figure 3.1 illustrates the setup of an experiment performed with MIR spectroscopy.



Figure 3.1.: The setup of an MIR spectroscopy measurement

During the measurement of the textile specimens, outlier data removal was performed on the fly due to the relatively small size of the dataset. For larger datasets, however, employing statistical methods such as Mahalanobis distance for outlier detection is a recommended approach [SQX⁺22, DMJRM00]. In our case, such an outlier detection mechanism was not necessary.

3.1.2. Data Overview

Each square textile piece represents a specimen. On each specimen, multiple random spots are selected for measurement, and each measurement generates a spectral profile. A sample consists of a group of specimens that share the same cotton ratio. In this study, 11 different samples were analyzed using NIR spectroscopy, while 9 different samples were measured using MIR spectroscopy.

We performed more measurements with NIR spectroscopy than with MIR spectroscopy, as NIR is faster and more convenient for spectral data acquisition. Each sample includes at least 3 specimens, and on each specimen, at least 10 random spots were measured. In total, 930 spots were measured using NIR and 485 spots were measured using MIR.

Figure 3.2 presents a summary of the NIR and MIR datasets using stacked bar charts. Each bar represents a cotton percentage, with stacked segments indicating the number of specimens, as shown in the shared legend at the bottom. Notably, the 50% cotton sample

3.1. Dataset Creation

dominates both the NIR and MIR datasets, reflecting its higher number of specimens and measurements compared to other cotton ratios.

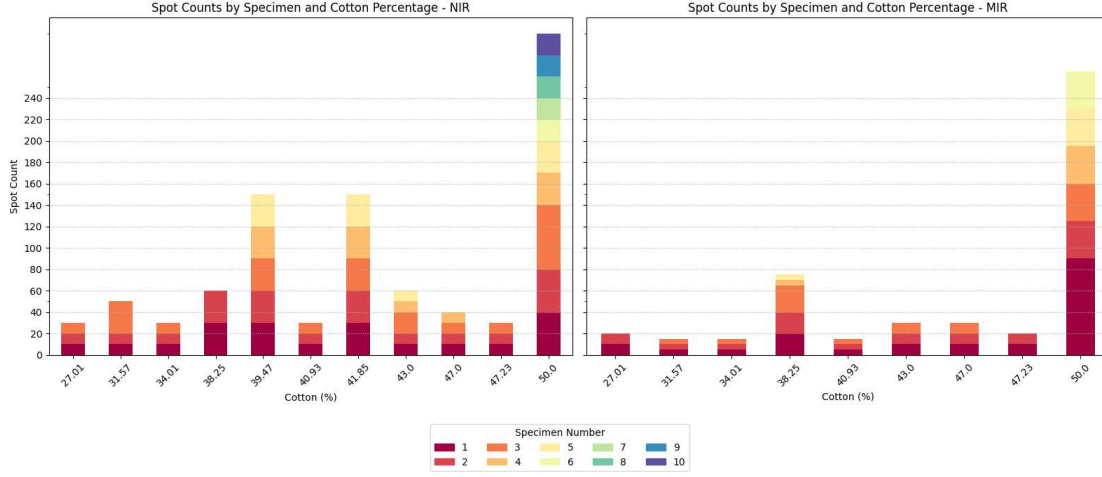


Figure 3.2.: Spot counts by specimen across different cotton percentages in the NIR and MIR spectral data.

3.1.3. Data Cleaning

Wavenumber Restriction

Specific regions in the NIR spectra are useful for detecting and distinguishing between cotton and polyester. Cotton fibers have distinct absorbance bands that reflect their chemical composition, primarily cellulose. In the MIR region, a broad peak at 3330 cm^{-1} corresponds to hydroxyl (O–H) groups, highlighting cotton’s hydrophilic nature and its ability to absorb moisture. Another key band at 2896 cm^{-1} is associated with the stretching vibrations of C–H bonds found in cellulose. Additional bands, such as the one at 1428 cm^{-1} , are linked to CH_2 bending vibration in cellulose, while those at 1360 cm^{-1} and 1315 cm^{-1} arise from C–H and C–O bending vibrations in cellulose’s structure [PRA⁺16]. NIR absorbance bands for cotton cellulose stem from the overtones and combination bands of O–H and C–H stretching vibrations. The C–H bond overtones appear in the regions $8696\text{--}7692\text{ cm}^{-1}$ (2nd overtone) and $5970\text{--}5376\text{ cm}^{-1}$ (1st overtone). A strong band is observed between $7140\text{--}5970\text{ cm}^{-1}$ which corresponds to the 1st overtone of O–H stretching. Additional O–H stretching vibrations contribute to strong spectral peaks around 5167 and 4750 cm^{-1} [LDTM16]. These spectral features provide a clear fingerprint for cotton, aiding its identification and analysis in spectroscopic studies.

The most significant vibrational modes of polyester can also be observed in both the MIR and NIR regions. In the MIR region, the C=O stretching mode appears prominently around 1700 cm^{-1} , while the C–O stretching mode from the ester bond is found between 1300 and 1000 cm^{-1} . Two peaks can be observed in this region, one near 1250 cm^{-1} , attributed to the ester bond, and another at 1100 cm^{-1} , corresponding to the C–O

3. Methodology

stretching in esters. In the NIR region, the first overtone of C–H in the aromatic group (Ar–CH) is detected around 6000 cm^{-1} , and the second overtone of the C=O stretch is also notable. These features are crucial for identifying and characterizing polyester materials in spectroscopic analysis [Smi22, ZYDW19].

Certain regions of the spectrum have high noise levels or poor signal-to-noise ratios, particularly at the extreme sides of the wavenumber range. NIR data has been restricted between $8000\text{--}3850\text{ cm}^{-1}$ while MIR has been restricted between $3600\text{--}600\text{ cm}^{-1}$. Spectral regions outside these ranges were excluded due to strong noise.

Excluding the Waterband

Water exhibits distinct absorbance bands in the infrared spectra, which are distributed across the MIR and NIR regions. Unlike polyester, cotton is a hydrophilic fiber therefore, moisture content directly influences the spectral characteristics of cotton. Increased water content resulting in more pronounced peaks, in the NIR region, near $6850\text{--}7000\text{ cm}^{-1}$ (first overtone of the O–H stretching band) and 6250 cm^{-1} (hydroxide overtone) [SH18]. Additionally, the combination of O–H stretching and O–H bending bands occurs around 5200 cm^{-1} . Furthermore, a combination involving the O–H bending band and the first overtone of the O–H stretching band is noted near 8500 cm^{-1} [LKCT94].

In the MIR region, the O–H stretching band appears prominently around 3300 cm^{-1} , while the O–H bending band is located near 1640 cm^{-1} . These absorbance bands are crucial for analyzing water’s hydrogen-bonding structure and its temperature-dependent behavior [LKCT94, MT19].

Peaks due to the water content of the textile make it more difficult to detect the peaks that are characteristic of cotton. To improve our data quality, we exclude the respective water bands from the spectra.

3.2. Reproducibility Experiments

We used Jackknife and Bootstrap resampling methods to examine how the relative standard deviation (RSD) of the absorbance values changes in the top absorbance peaks of the spectra. The wavenumbers from the top 10 absorbance peaks were selected because they provide valuable information about the cotton and polyester content. Before starting the reproducibility analysis, the data creation process was completed and no data preprocessing was applied since preprocessing methods tend to reduce spectral variance. Then, a data subset was generated from four sample groups, each containing data from 5 specimens. For every specimen, 30 measurement spots were recorded.

For each sample, peak absorbance points were determined and resampling was performed at the spot level for each peak wavenumber. The statistical results were calculated for each wavenumber/specimen pair. These results were then averaged on a wavenumber basis. Jackknife and Bootstrap methods generated resampling sizes ranged from 3 to 30 for each specimen. Resampling in different group sizes made it possible to observe the effect of sample size on RSD. Diagram 3.3 shows the steps followed in the reproducibility

analysis.

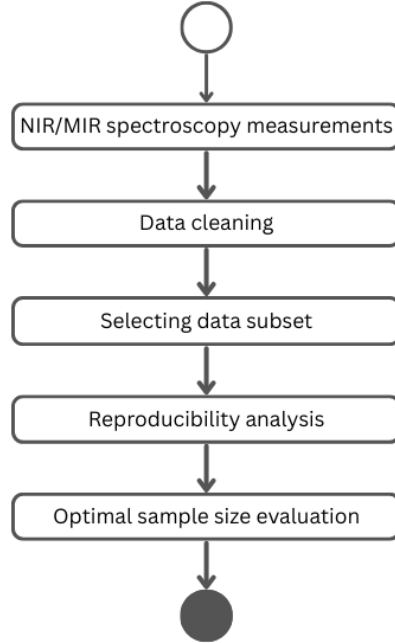


Figure 3.3.: The workflow followed in reproducibility experiments

3.2.1. Variance Estimation with Jackknife Resampling

In this thesis, we were inspired by the principles of delete-d Jackknife resampling to assess dataset variability in NIR spectral peaks. The theoretical background of the Jackknife resampling is described in section 2.3.1. Our approach systematically removes subsets of data to analyze how the removal affects a specific statistic, the Relative Standard Deviation (RSD). By iteratively excluding subsets, the Jackknife approach provides insights into the robustness and sensitivity of the dataset to missing data.

In this resampling, the input is a set of absorbance values, each corresponding to specific peaks or bands in a spectrum. The function works by generating subsets of the data, where subsets of size L are excluded during each iteration. For every subset, the remaining data is used to compute the RSD, a statistic that quantifies variability as a percentage of the mean. The function systematically evaluates all possible subsets of size L , allowing for a thorough analysis of how the dataset behaves when certain values are removed.

For each subset size L , the function aggregates the RSD values calculated from the reduced datasets into a list. By taking the average of these RSD values across all subsets of size L , it generates a single measure of variability for that subset size. The results are

3. Methodology

stored in a dictionary, where each key represents a specific subset size, and the value is the average variability observed for that size. The pseudocode of our approach is presented in algorithm 1.

Algorithm 1: Jackknife Variance Estimation

```
Data: peak_absorbance_values ( $m \times n$  matrix where  $n$  is spots and  $m$  is wavenumbers)
Input:  $L$ : maximum group size to leave out
 $2 \leq L \leq |peak\_absorbance\_values|$  // Ensure  $L$  is not greater than the number of spots measured and not smaller than 2
for  $i \leftarrow 2$  to  $|peak\_absorbance\_values| - L$  do
     $index\_combinations = combination(peak\_absorbance\_values, L)$ 
    // Generate all the combinations of indices of given  $L$ 
    for  $indices\_to\_leave\_out \in index\_combinations$  do
         $X_{reduced} \leftarrow peak\_absorbance\_values - indices\_to\_leave\_out;$ 
         $subsample\_variance \leftarrow variance(X_{reduced})$  // Store the variances
    end
end
```

3.2.2. Variance Estimation with Bootstrap Resampling

The Bootstrap resampling method was implemented in a similar way to Jackknife resampling. It assesses variability with RSD by generating synthetic datasets through random sampling with replacement. The Bootstrap method is described in more detail in section 2.3.2. In the context of the thesis, the Bootstrap method is used to analyze the stability and distribution of absorbance values measured at specific peaks across specimens and spots.

The process begins with the absorbance values measured at key wavenumbers corresponding to the absorbance peaks for a given specimen. The Bootstrap method generates multiple samples by randomly drawing values from the original dataset with replacement. Each bootstrap sample has the same number of values as the original dataset; however, due to the replacement, some values may appear multiple times, while others may be excluded. This resampling approach simulates the variability inherent in the data and helps estimate how much variation exists across repeated measurements.

Each resample size L ranges from 3 to 30 in this implementation. The function creates a set number of bootstrap samples, then calculates RSD statistics for each sample, reflecting the variability relative to the mean. The results for each resample size L are stored in a dictionary, where the key represents the resample size, and the value is an array of RSD calculated from the bootstrap samples. The pseudocode of our approach is presented in algorithm 2.

The purpose of the Jackknife and Bootstrap approaches is to assess the sensitivity of the dataset and the computed RSD to missing data. These methods provide insights into

Algorithm 2: Bootstrap Variance Estimation

Data: *peak_absorbance_values* (matrix of absorbance values for peaks)
Input: L : maximum resample size
 $2 \leq L \leq |\text{peak_absorbance_values}|$;
 // Ensure L is not greater than the number of spots measured and not smaller than 2
for $i \leftarrow 0$ **to** L **do**
 $X_{bootstrap} \leftarrow \text{bootstrap_resampling}(\text{absorb_val_by_peaks}, L)$;
 // Generate bootstrap resampling of given size of L
 $\text{subsample_variance} \leftarrow \text{variance}(X_{bootstrap})$ // Store the variances
end

the stability of the dataset, as a highly variable RSD across subsets may indicate that the data is sensitive to small changes. Conversely, consistent RSD values suggest that the dataset is robust. This allows for a detailed analysis of how variability changes with different resample sizes, offering insights into the data's underlying stability and behavior.

3.2.3. Margin of Error Approach with Bootstrap Resampling

To estimate the minimum sample size needed for reliable predictions, we apply a Bootstrap-based margin of error approach, inspired by [Qum13]. We treat the mean of the full dataset as the best available estimate of the population mean. The goal is to determine the smallest sample size that yields a sample mean within an acceptable margin of error relative to this reference.

For each sample size n (ranging from 3 to 30), we generate bootstrap resamples to approximate the sampling distribution of the mean. By observing how the mean varies across resamples, we assess the reliability of the estimate at different sample sizes. The purpose of trying various sample sizes is to identify the point at which increasing the sample size no longer significantly improves the accuracy of the mean estimate. For each sample size n , the 2.5th and 97.5th percentiles of the bootstrap means give a confidence interval around the mean estimate. The range between these percentiles represents the variability or margin of error in the sample mean estimate due to sampling variability. This interval shows how much the sample mean could deviate from the full dataset mean due to sampling error. A smaller interval indicates that the sample size is sufficient to provide a stable and accurate estimate of the mean. The chosen threshold (5% of the full dataset's mean) defines an acceptable level of accuracy for the sample mean estimate relative to the full dataset mean. If the margin of error (the width of the confidence interval) is within 5% of the full dataset's mean, we consider the sample mean to be statistically close enough to the true mean. This implies that the sample size is likely large enough to be representative. The smallest sample size satisfying this condition is then considered the minimum sufficient sample size (see Algorithm 3).

Since resampling is performed for each feature, applying the procedure across the entire

3. Methodology

spectral dataset would be computationally expensive. Instead, we selected only the top 10 peak wavenumbers from the NIR spectra. These peak points were chosen because they represent regions with high absorbance intensity, which are often rich in chemical information and particularly informative for cotton content prediction. For each of these peak wavenumbers, we extracted the corresponding absorbance values across specimens and applied the bootstrap resampling procedure individually. The process was repeated 100 times. By aggregating results across iterations, specimens and wavenumbers, we obtained the estimates of the minimum sufficient sample size for each spectral feature.

Algorithm 3: Margin of Error Approach with Bootstrap Resampling

```
Data: peak_absorbance_values (matrix of absorbance values for peaks)
Input: L: maximum resample size
 $2 \leq L \leq |peak\_absorbance\_values|$ ;
// Ensure L is not greater than number of spots measured and not
    smaller than 2
for i  $\leftarrow$  0 to L do
    | Xbootstrap  $\leftarrow$  bootstrap_resampling(absorb_val_by_peaks, L);
    | // Generate bootstrap resampling of given size of L
    | subsample_mean  $\leftarrow$  mean(Xbootstrap);
    | // Store the means
    | lower_percentile  $\leftarrow$  percentile(subsample_mean, 2.5);
    | upper_percentile  $\leftarrow$  percentile(subsample_mean, 97.5);
    | margin_of_error  $\leftarrow$  (upper_percentile - lower_percentile)/2;
    | if margin_of_error  $\leq$  threshold then
    | | return L
    | end
end
```

3.3. Regression Pipeline

To predict the cotton content from IR spectral data, a structured regression pipeline was developed. This pipeline covers the entire process from data generation to producing prediction outputs. It includes key steps such as selecting relevant data, applying preprocessing techniques, reducing dimensionality, training regression models, and evaluating their performance. The overall workflow of the pipeline is illustrated in Figure 3.4.

3.3.1. Selecting Subset

The initial dataset contained a high proportion of 50% cotton measurements, as seen in Figure 3.2. Having a dataset dominated by 50% cotton measurements would lead to overly optimistic results. Error metrics would be too good for the highly represented

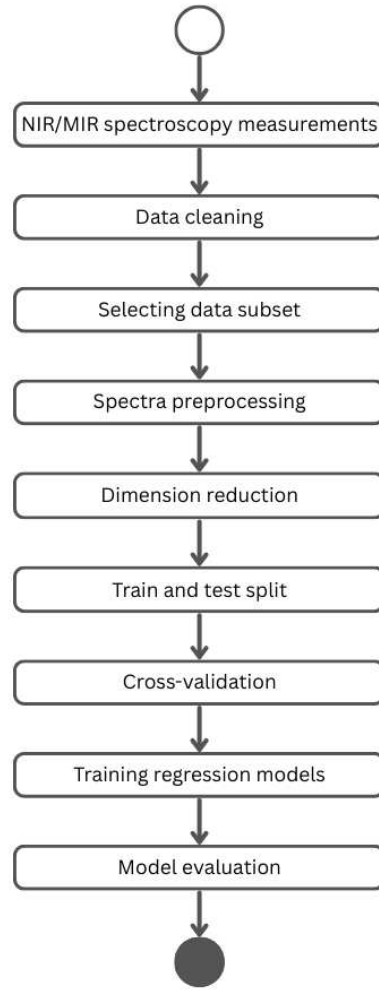


Figure 3.4.: Workflow followed in regression pipeline

ratios and worse for ratios which are underrepresented. To mitigate this data imbalance, the 50% cotton measurements were restricted to five specimens in the dataset.

3.3.2. Spectra Pretreatment

Various normalization, smoothing or baseline correction methods, including SNV (Standard Normal Variate), Detrend, FillPeaks, ALS (Asymmetric Least Squares), MSC (Multiplicative Scatter Correction), and Savitzky-Golay smoothing have been applied on the restricted spectral data. The theoretical background of these methods is described in section 2.2.2.

To identify the most effective combination of preprocessing and machine learning methods, each technique was first applied to the spectral data. The resulting preprocessed

3. Methodology

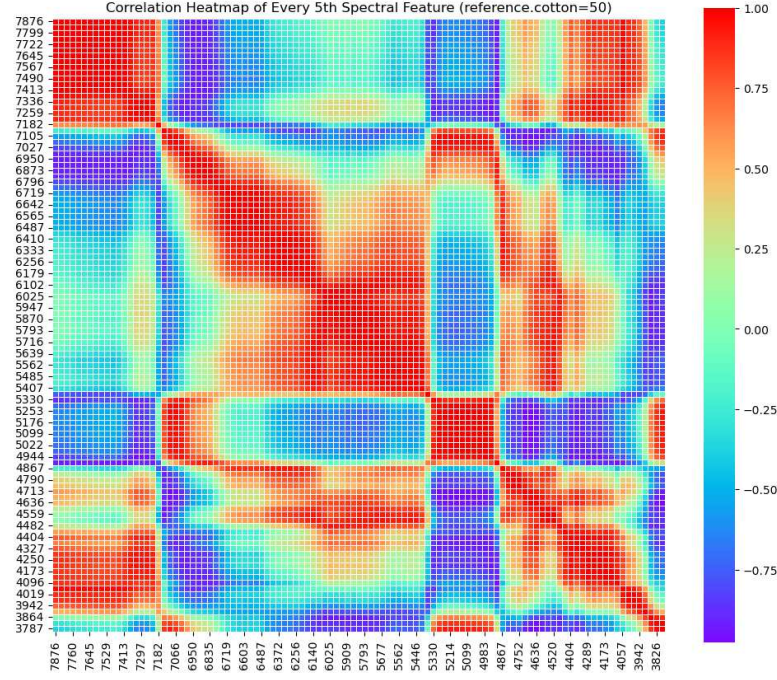


Figure 3.5.: Correlation heat map of the NIR spectra of 50% cotton blend, without preprocessing. The wavenumber range spans from 7876 cm^{-1} to 3787 cm^{-1} , with every fifth wavenumber displayed to improve readability.

datasets were then used as input for various machine learning algorithms. As a benchmark, unprocessed data was also used as input, allowing for comparison of model performance against preprocessed data.

3.3.3. Dimension Reduction

In the figure 3.5, the correlation heatmap shows several high-correlation blocks, indicating that many spectral features are strongly correlated. This is due to the nature of spectral data, where neighboring wavelengths often capture similar information, leading to multicollinearity.

In order to overcome the multicollinearity, Principal Component Analysis (PCA) has been applied as a dimension reduction technique before training certain regression models. With PCA, correlated features are transformed into principal components that are orthogonal to each other, and thus uncorrelated. The principal components are selected based on their ability to capture the maximum variance in the data. A small subset of components is retained, ensuring that most of the relevant information is preserved while discarding noise and irrelevant variations [KP23].

Both the training and testing datasets have undergone PCA transformation to remove correlated features, which helps regression models to avoid overfitting. At the same time, removing redundant information improves computational efficiency.

3.3.4. The Train-Test Split Strategy

We have multiple spots per specimen. The spectra of spots of a single specimen are slightly more correlated than the spectra of an entirely new specimen. Therefore, a randomized train-test split would not accurately reflect real-world conditions, as some spots from a specimen could be included in both training and testing. This can result in data leakage. In practice, an entirely new specimen is measured before its ratio is predicted, meaning that spots from the new specimen are not present in the training set. To better align with this scenario, all spots from a specific specimen of each sample were selected for testing. Hence, train and test splitting ensures all measurements from a specific specimen are either in the test set or the training set.

	Number of measurements (data size)	Cotton content mean	Unique cotton values
Train	630	42.40	11
Test	200	41.47	11

Table 3.1.: Description of train and test datasets used in the regression models.

3.3.5. Cross-Validation Strategy

To mitigate data leakage during cross-validation, we took a similar approach as in the train-test split strategy. Instead of a random selection, a specimen-oriented selection was made when creating cross-validation groups. If the groups would have been split randomly, the cross-validation and test results would have been inconsistent.

To ensure that groups do not overlap in cross-validation, a GroupKFold iterator with 4 folds was set. The approach involved generating unique group IDs by combining the cotton ratio and specimen number. These IDs were then passed to the K-fold iterator. This ensures that data from the same specimen does not appear in both the training and validation datasets.

3.3.6. Selected Algorithms and Hyperparameter Space

In selecting appropriate machine learning algorithms for predicting cotton percentage from spectral data, several factors were considered, including interpretability, ability to handle high-dimensional data, robustness to noise, and predictive performance. Our selected models were discussed in more detail in Chapter 2.4. In order to find the best parameters for each model, we do a randomized search over a pre-defined hyperparameter space. Therefore, the RandomizedSearchCV function from the scikit-learn library was used. It performs a randomized search over a predefined parameter grid, evaluating different hyperparameter combinations using cross-validation to find the best-performing set.

3. Methodology

We chose Partial Least Squares Regression (PLS) as a linear model, because it is widely used in chemometrics and spectral analysis due to its ability to handle high-dimensional and highly collinear data. The hyperparameter space for it is defined in Table 3.2.

Furthermore, we analysed multiple non-linear regression models. Table 3.3 shows the defined hyperparameter space for K-Nearest Neighbors (KNN). Least Squares Support Vector Machines (LSSVM) is another powerful ML algorithm that can handle non-linear regression. It effectively captures nonlinear relationships between spectral absorbance and cotton percentage. The hyperparameter space for LVSSM is shown on Table 3.4. Kernel Ridge Regression (KRR) is another non-linear model we analysed in the thesis. It integrates ridge regression with kernel transformations. Its regularization term prevents overfitting, which is crucial in processing high-dimensional spectral data. Table 3.5 shows its hyperparameter space.

Among the tree-based regression models we selected eXtreme Gradient Boosting (XG-Boost) and Random Forest. Table 3.6 and Table 3.7 show their respective hyperparameter space.

Convolutional Neural Networks (CNN) were chosen as a deep learning model, which are capable of extracting nonlinear features. The hyperparameter space for CNN is shown in Table 3.8.

Table 3.2.: Hyperparameter values for tuning PLS Regressor.

Hyperparameter	Values
number of components	5, 10, 15, 20

Table 3.3.: Hyperparameter values for tuning K-Nearest Neighbor Regressor.

Hyperparameter	Values
number of neighbors	5, 10, 15, 20, 25, 30
weights	uniform, distance
metric	euclidean, manhattan, minkowski

Table 3.4.: Hyperparameter values for tuning Least Squares Support Vector Regressor.

Hyperparameter	Values
kernel	polynomial
sigma (regularization)	log space $[10^{-2}, 10^2]$
gamma (kernel scale)	log space $[10^{-2}, 10^2]$

Table 3.5.: Hyperparameter values for tuning Kernel Ridge Regressor.

Hyperparameter	Values
kernel	polynomial, rbf
alpha	log space $[10^{-3}, 10^2]$
gamma (for rbf kernel)	log space $[10^{-2}, 10^6]$
degree (for polynomial kernel)	2, 3, 4

Table 3.6.: Hyperparameter values for tuning eXtreme Gradient Boosting Regressor.

Hyperparameter	Values
number of estimators	100, 200, 300
max depth	3, 5, 7
learning rate	0.01, 0.1, 0.2
subsample	0.8, 0.9, 1.0

Table 3.7.: Hyperparameter values for tuning Random Forest Regressor.

Hyperparameter	Values
number of estimators	10, 20, 50, 100, 200, 500
max depth	3, 5, 10, 20, 50, 100
min samples split	2, 5, 10
min samples leaf	1, 2, 4
max features	sqrt, log2, None
bootstrap	True, False

Table 3.8.: Hyperparameter values for tuning Convolutional Neural Network.

Hyperparameter	Values
optimizer	adam, rmsprop
regularizer	log space $[10^{-4}, 10^{-2}]$
epoch	10, 20, 30, 50
batch size	4, 8, 16, 32

3.3.7. Regression Performance Metrics

RMSE (Root Mean Squared Error) and R^2 were used for the comparison of the algorithms since they are two of the most commonly used metrics for evaluating the performance of regression models. They help quantify how well a model fits the data and how accurately it makes predictions. RMSE measures the average distance between prediction and actual values which are prediction errors, giving more weight to larger errors due to squaring [CD14]. R^2 , the coefficient of determination, represents percentage of variance that the

3. Methodology

prediction model explains in the dependent variable [N⁺91]. It gives a normalized measure of goodness-of-fit between 0 and 1. A higher R^2 indicates that more variance is explained in dependent variables.

3.4. Infrastructure and Software

To do spectral measurements, we used the following infrastructure for the thesis:

- FTIR Spectrophotometer - TENSOR 37, BRUKER OPTICS
Range from: 4000-400 cm⁻¹, ATR (Attenuated Total Reflection), HTS-XT (High Throughput Screening Extension), DTGS detector.
- FTNIR Spectrophotometer - MPA, BRUKER OPTICS
Range from 12500-4000 cm⁻¹ measurement in transmission and in diffuse reflection.

The IT infrastructure for running the regression models is comprised of a Dell PowerEdge server with 40 Intel Xeon Gold processor cores and 1 terabyte of RAM.

R (version 4.3.3) and Python (version 3.12.2) were the main languages used in the project. The desktop application RStudio (version 2024.12.0 Build 467) was employed for R programming. Packages such as plyr, dplyr, IDPmisc, prospectr, baseline, knitr, gridExtra, ChemoSpec, matrixStats, and stringr were employed for comprehensive data preprocessing and transformation tasks. Python was managed via PyCharm under the Anaconda distribution (conda version 22.9.0). Libraries such as pandas, numpy, and scipy were utilized for data resampling, while scikit-learn and xgboost supported the development of predictive models. Data inspection and model evaluation processes were conducted using Jupyter Notebook, which provides an efficient platform for analysis and visualization.

In the context of this thesis, we created a spectral database and utility functions to support the analysis of the spectral data. Scripts were created to run the selected regression models and visualize the results. The database and the codebase containing the utility functions and scripts can be found at the project's Github repository [Yay25].

4. Results and Discussion

This chapter presents the key findings of the study and is structured into three main sections. In Section 4.1, we investigate the reproducibility of NIR spectral measurements to determine the minimum sample size required for reliable data collection. Section 4.2 explores the impact of different spectral preprocessing techniques and depicts visual comparisons to assess how these methods influence the data structure. In section 4.3, we assess the predictive performance of different regression models for textile characterization, identifying the most effective approaches for estimating cotton content in textile blends. We analyze how different combinations of regression algorithms and preprocessing methods affect prediction accuracy and highlight the most effective strategies for textile quantification.

4.1. Reproducibility

To address the first question mentioned in Chapter 1, we needed to detect the number of spots that should be measured per specimen to ensure the sample data accurately represents the real data. First, we compared the Jackknife and Bootstrap resampling methods and found that the Bootstrap resampling could produce similar results to the Jackknife resampling with lower complexity. Therefore we use Bootstrap resampling, when determining the minimum sample size with the margin of error approach.

4.1.1. Comparison of Jackknife and Bootstrap Resampling

The results (Figure 4.1) revealed no major differences in the relative standard deviations produced by the Jackknife and Bootstrap methods. Both approaches provided similar insights into the data’s variability, indicating that either method could be used for this type of analysis. However, the Bootstrap algorithm implemented in this study proved to be computationally faster compared to the Jackknife method, making it a more practical choice for large-scale or repetitive analyses.

Our Jackknife and Bootstrap resampling functions differed significantly in computational complexity and scalability. The Jackknife resampling function has a complexity of $O(2^N)$ in the worst case due to its reliance on generating all possible combinations of indices for leaving out subsets of data. This combinatorial growth makes it computationally expensive and unsuitable for large datasets. In contrast, the Bootstrap function operates with a complexity of $O(n)$, where n is the fixed number of resampling iterations. This linear growth ensures that Bootstrap is more scalable and efficient for larger datasets. While the Jackknife resampling method provides exact variance estimates by systematically omitting

4. Results and Discussion

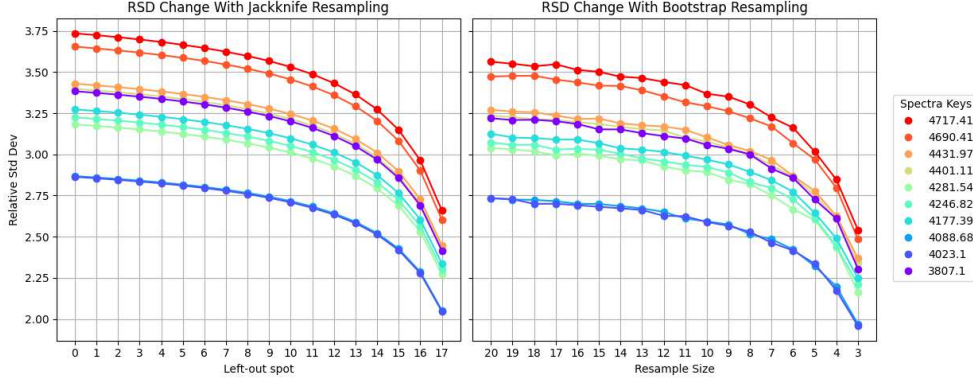


Figure 4.1.: The change of Relative Standard Deviation as resample size decreases for Bootstrap and Jackknife resampling. Each line in both plots represents one of the top 10 wavenumbers with the highest absorbance. The colors refer to the wavenumbers where top ten peak absorbances occur and the spectra keys denote the wavenumbers (in cm^{-1}) at those peaks. There is an obvious similarity in trends across the Jackknife and Bootstrap methods.

subsets of data, its cost grows exponentially. Conversely, the Bootstrap resampling approximates variance using random sampling, trading a slight loss in precision for significant gains in performance and scalability.

Due to the high computational cost of Jackknife resampling, it was not feasible to calculate the relative standard deviation (RSD) for every sample size between 30 and 3. As a result, Jackknife resampling was limited to a maximum of 20 spots. Figure 4.1 illustrates how RSD changes as the sample size decreases. For the comparison between these two methods, in Jackknife, resampling was performed by gradually leaving out one spot from a resample size of 20 to a resample size of 3, and in Bootstrap, resampling was drawn with replacement, from a resample size of 3 to a resample size of 20. Each line in the plots represents one of the top 10 wavenumbers with the peak absorbance. Both graphs show that as sample size decreases, RSD also decreases. This can occur when analyzing correlated data where many points carry similar information, and removing more points tends to leave a more homogeneous dataset. The trends are very similar across Jackknife and Bootstrap, suggesting that both methods produce stable estimates of variability. Additionally, the relatively smooth curves reflect the robustness of RSD across the top wavenumbers.

4.1.2. The Results of Margin of Error Approach with Bootstrap Resampling

To estimate the minimum number of measurement spots needed to obtain statistically reliable spectral data, we applied a Bootstrap resampling approach with a margin of error criterion. This procedure was repeated 1000 times for each of 15 peak wavenumbers, and the results were averaged to obtain the final minimum sample size required. The dataset used in this analysis consisted of 30 spectra per specimen, 5 specimens per sample, and 3

different sample groups in total.

This method identifies the smallest sample size at which the mean of the resampled absorption values falls within an acceptable margin of error compared to the true mean. By applying this process to the top absorbance peak wavenumbers, we assessed how sample size requirements vary with spectral intensity and sample characteristics.

Figure 4.2 illustrates the relationship between wavenumbers and the average minimum number of spots required across three different cotton content levels: 39.47%, 41.85%, and 50.0%. The results show that the samples with 50% cotton consistently require a higher number of measurement spots, indicating greater variability in their spectra. In contrast, the 39.47% cotton samples exhibit the lowest variability, requiring fewer spots to achieve reliable estimates.

Based on these findings, we decided to measure at least 10 spots per specimen. This number was found to be sufficient across all cotton levels to capture meaningful spectral variation while minimizing measurement error. This conclusion balances statistical accuracy with practical considerations, as measuring fewer spots reduces the overall experimental effort while maintaining sufficient representativeness.

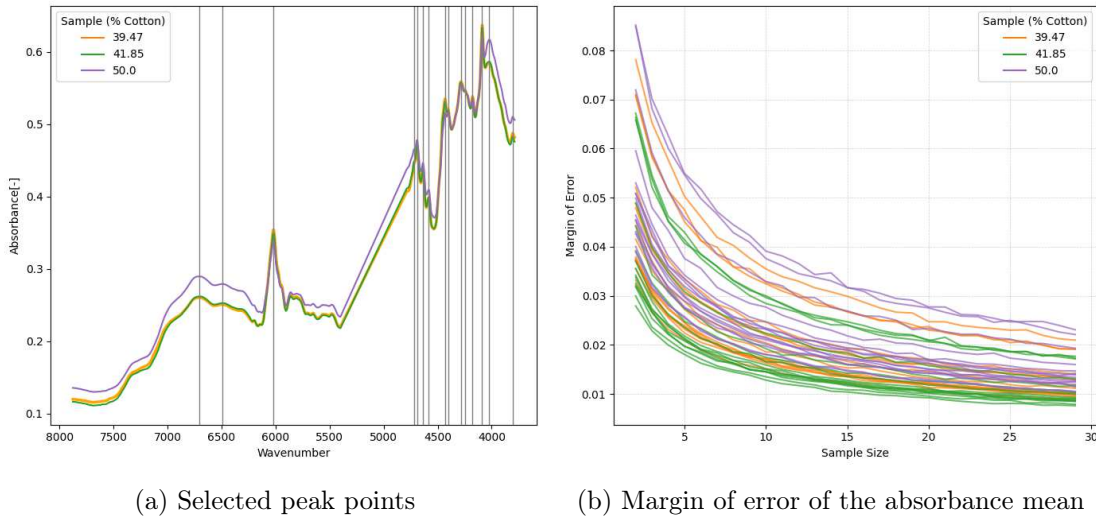


Figure 4.2.: For cotton blend samples containing 39.47% (orange), 41.85% (green), and 50.0% (purple) cotton, the 15 spectral peak points with the highest absorbance values were selected (a). For each peak, the Bootstrap resampling was conducted for the sample sizes from 3 to 30. The relative standard deviation (RSD) of the mean absorbance was computed at each step. Each line represents a peak, color-coded by cotton percentage. The graph demonstrates that already from sample size 3, mean absorbance values of most of the peak points have reached below accepted margin level (5% of the full dataset mean of absorbance) (b).

4.2. Preprocessing

Preprocessing of spectral data proved to be important in improving model performance by reducing noise, correcting baseline shift, and improving class separation in dimensionality reduction. Comparison of different techniques revealed that no approach performed universally better than the others; instead, effectiveness varied depending on the regression model applied. The figures 4.3 and 4.4 demonstrate how the spectra are processed by different baseline correction and normalization techniques. The raw spectra exhibit baseline shifts and noise, which are progressively reduced through preprocessing.

SNV and MSC were particularly effective in normalizing the spectra, improving consistency across samples by eliminating scattering effects. This similarity in performance stems from the fact that both methods rely on a form of standardization. MSC aligns spectra to a mean reference, while SNV standardizes each spectrum individually. However, despite their effectiveness in scatter correction, they provided limited benefits in addressing baseline trends, as observed in regression results.

On the other hand, Detrend and ALS and Fillpeaks demonstrated strong capabilities in correcting baseline curvature, resulting in a flatter baseline and improved signal clarity, particularly in low wavenumber regions. Detrend, which used a second-order polynomial, efficiently removed the baseline trend, preventing unnecessary amplification of certain peaks caused by the underlying trend. Hence, the corrected spectra have a flat baseline.

The FillPeaks method was performed with a smoothing parameter set to 1, allowing the method to capture small details without distorting the peak shapes. The half-width of interpolation (10) set the size of the window around each peak used for filling gaps. Only peaks with an intensity above 200 were processed, ignoring smaller peaks or noise. The FillPeaks technique adjusted and interpolated missing peaks without amplifying them, especially for wavenumbers higher than 6000 cm^{-1} , as seen with Detrend and ALS.

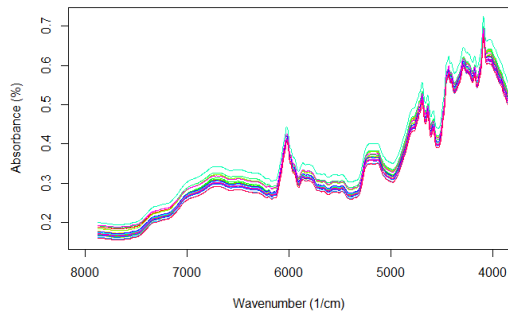
The ALS method was applied with an asymmetry parameter of 0.05, which determines the weight given to positive deviations of the signal from the baseline, thereby emphasizing downward corrections. This approach smoothed the curves, particularly reducing the influence of peaks from the wavenumbers lower than 5000 cm^{-1} , while effectively removing broad background trends similar to Detrend.

Savitzky-Golay smoothing, reduced the random noise in the spectra while preserving the peak shapes and their intensities. This method utilized a third-degree polynomial least-squares fitting with zeroth-order differentiation to maintain the integrity of both peak heights and widths. As discussed in the literature, Savitzky-Golay smoothing is primarily used for removing noisy features and it does not inherently correct for baseline shifts. To reduce noise, first or second derivatives are applied and then smoothing is done [RBE09].

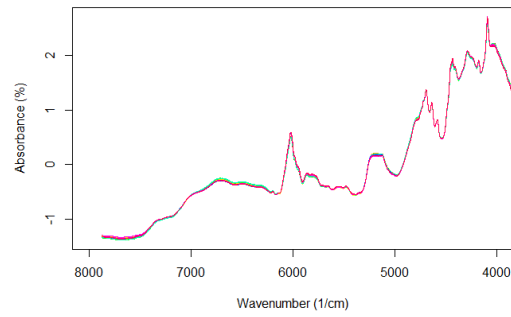
In summary, each method has a specific purpose, such as removing baseline shifts, reducing noise, or normalizing the data. SNV and MSC are well-suited for scatter correction, while Detrend and ALS outstand in baseline trend removal. On the other hand, Detrend or ALS may suppress or amplify certain spectral features depending on the spectral data and the hyperparameter choices. However, SNV and MSC may not

4.2. Preprocessing

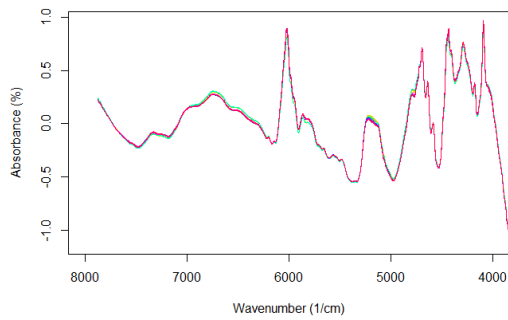
handle baseline drifts as effectively as ALS or Detrend. FillPeaks preserves peak integrity but might not be as effective in correcting baseline shifts. Similarly, while Savitzky-Golay smoothing effectively reduces noise, it does not specifically correct for baseline drifts.



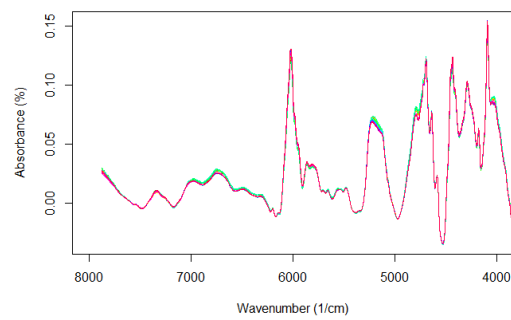
(a) No preprocessing



(b) SNV



(c) Detrend



(d) ALS

Figure 4.3.: Comparison of SNV, Detrend, and ALS techniques applied to NIR spectra from a sample containing 34.01% cotton.

4. Results and Discussion

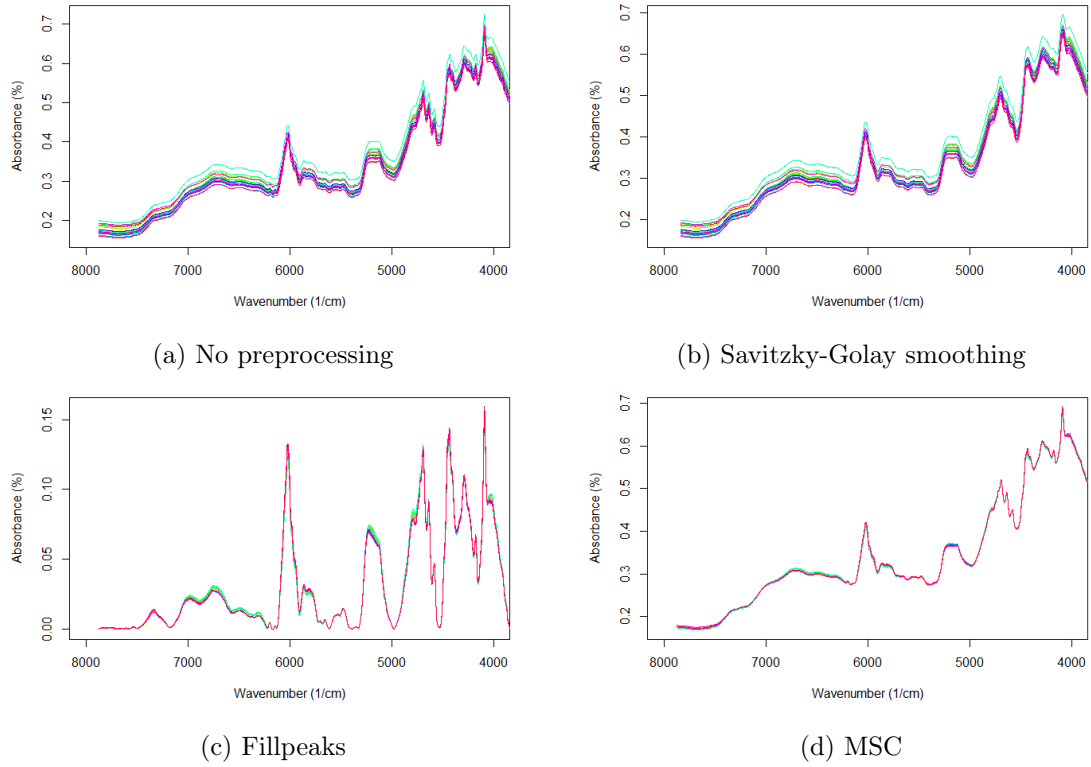
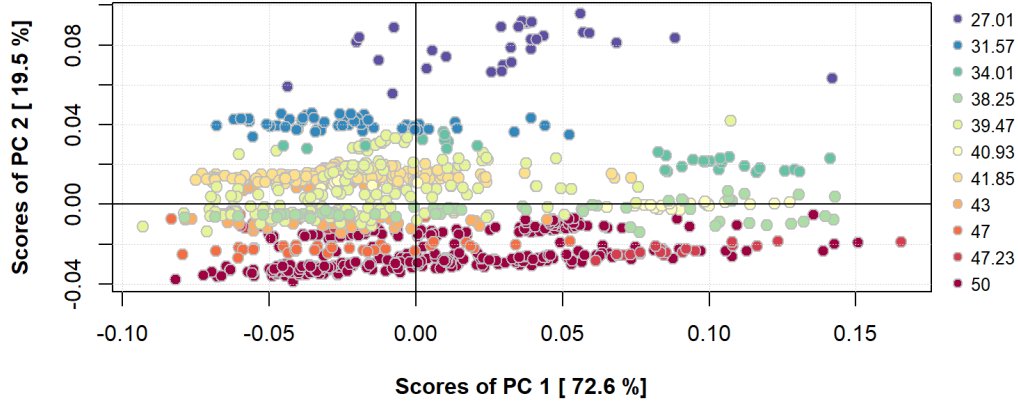


Figure 4.4.: Comparison of Savitzky-Golay, FillPeaks, and MSC techniques applied to NIR spectra from a sample containing 34.01% cotton.

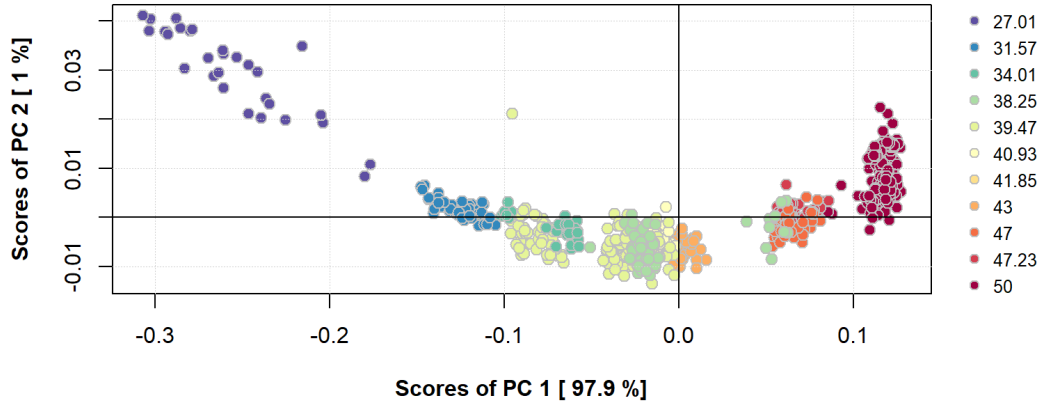
Figure 4.5 (a) shows Principal Component Analysis scores before the preprocessing. The groups with different cotton ratios are not perfectly separable. After applying Standard Normal Variate (SNV) and Detrend preprocessing techniques, as depicted in Figure 4.5 (b), the PCA scores plot reveals a significant improvement in group separation. In addition, the explained variance captured by the first principal component increases after preprocessing. The combined explained variance from the first and second principal components also improves. This indicates that the preprocessing effectively normalized the data and removed trends that previously obscured group differences.

4.3. Prediction Model Performance

In this chapter, the regression metrics of the selected regression algorithms are evaluated to address the second question in Chapter 1. The models are tested on NIR and MIR datasets to determine on which one they perform better.



(a) The Scores graph of the first and the second principal components when no preprocessing is applied.



(b) The Scores graph of the first and the second principal components after the data was preprocessed with Fillpeaks.

Figure 4.5.: Score plots of the first and second principal components: (a) without preprocessing and (b) after Fillpeaks preprocessing.

4.3.1. Comparison of Regression Models and Preprocessing Methods on NIR Spectra

This section analyzes how different regression models perform with respect to preprocessing methods. Spectral data was first processed with the preprocessing methods SNV (Standard Normal Variate), Detrend, FillPeaks, ALS (Asymmetric Least Squares), MSC (Multiplicative Scatter Correction), and Savitzky-Golay smoothing. In addition, combined approaches, Detrend followed by SNV, and SNV followed by Savitzky-Golay smoothing, were also implemented. Later, models were trained for each preprocessed dataset.

For Partial Least Squares Regression (PLSR), Random Forest regression and Convolutional Neural Network (CNN), Principal Component Analysis (PCA) was not applied to reduce dimensionality prior to model training since these models can inherently handle high dimensionality. In contrast, for K-Nearest Neighbors (KNN) regression and Kernel Ridge regression, PCA was applied to retain the first 15 principal components. This choice was based on the explained variance curve (Figure 4.6), where 15 components accounted for 99.99% of the total variance. Although the first 4-5 principal components account for 99% of the explained variance, this decision ensures compatibility with complex algorithms and minimizes the risk of discarding potentially informative variance related to subtle spectral features.

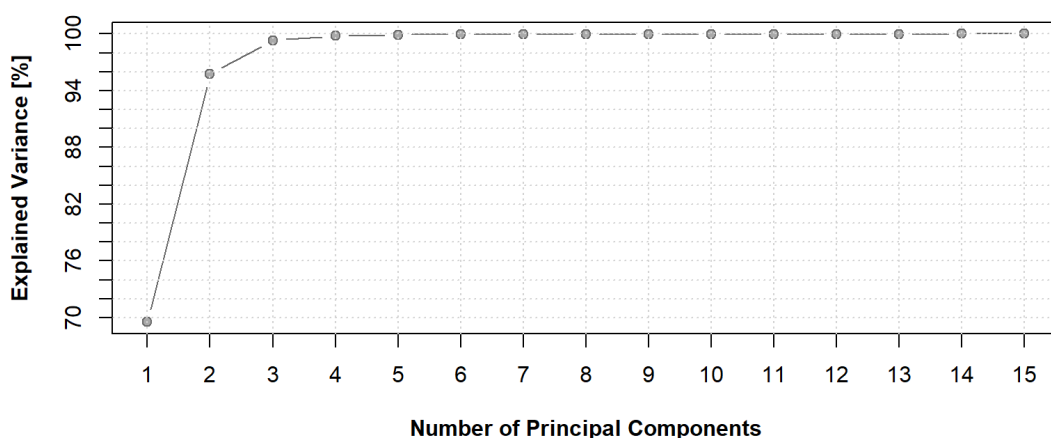
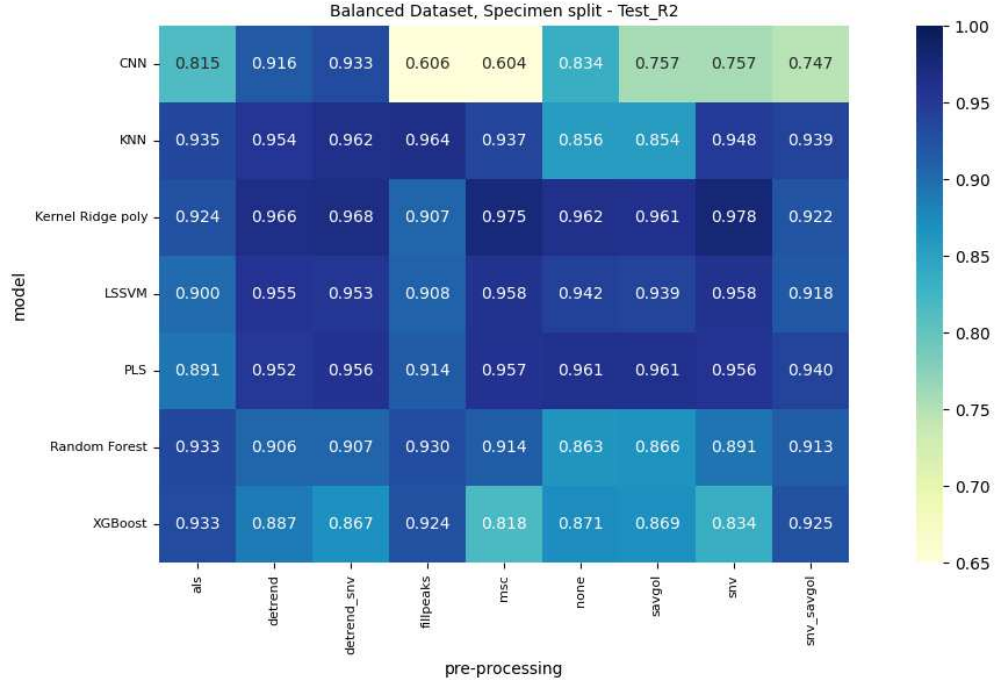


Figure 4.6.: Explained variance increases as PCA components increase. On the NIR spectral dataset, wavenumber restriction was made but no preprocessing was applied.

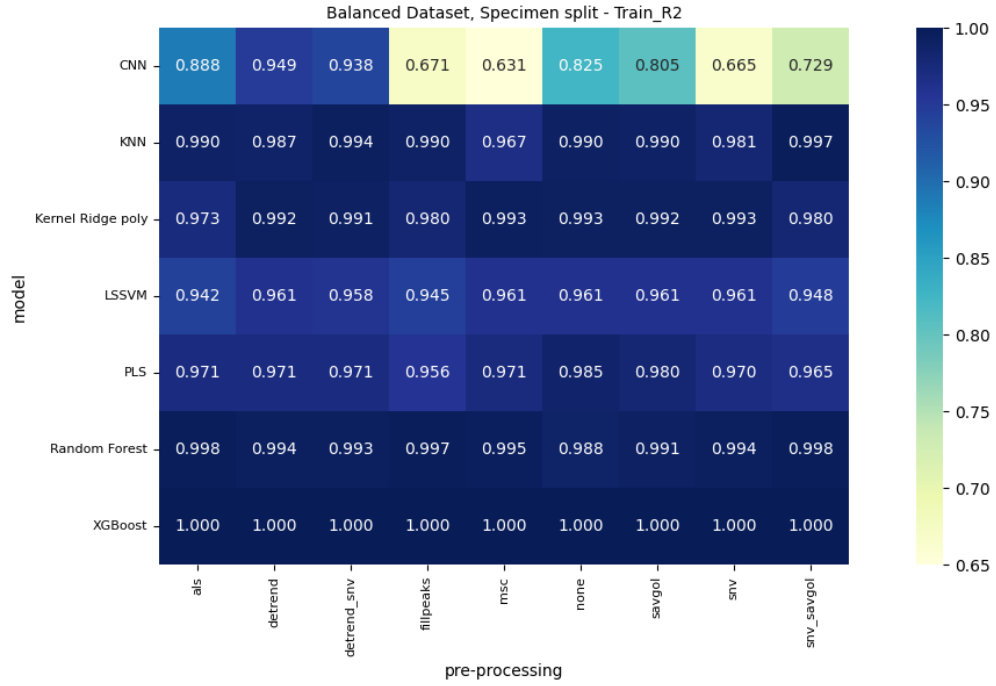
Regression models were optimized using RandomizedSearchCV with 4-fold cross-validation, and their performance was assessed using Root Mean Square Error (RMSE) and R^2 metrics.

The heatmaps in 4.7 compare the performance of different regression models and

4.3. Prediction Model Performance



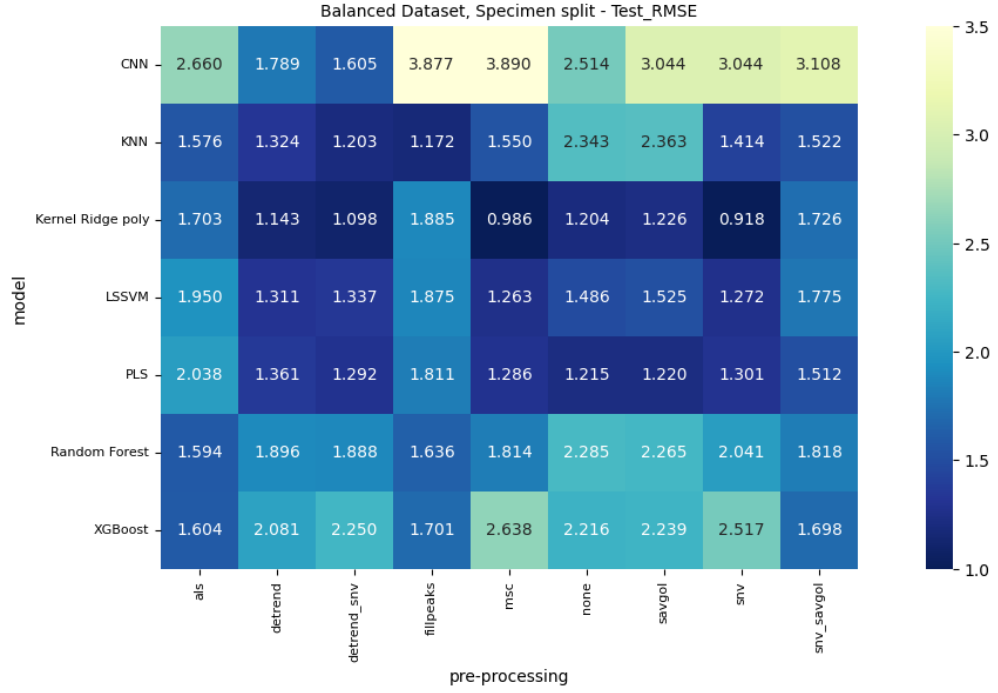
(a) Test Set



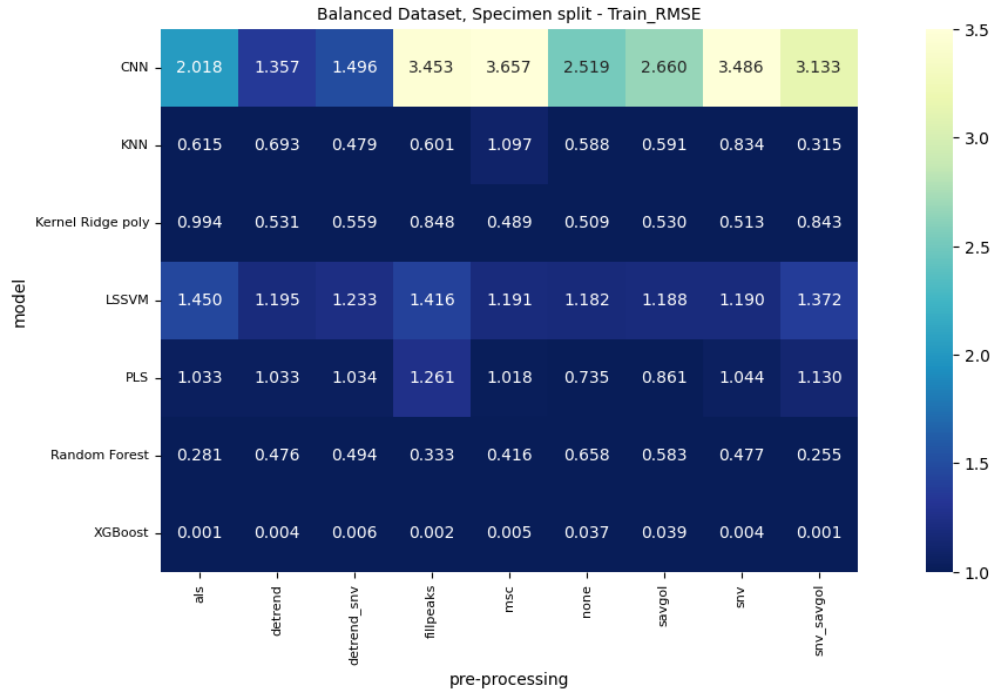
(b) Train Set

Figure 4.7.: Heatmap of R^2 values on test (a) and training (b) data. On the x axis, the preprocessing methods and on the y axis, the regression models are presented.

4. Results and Discussion



(a) Test Set



(b) Train Set

Figure 4.8.: Heatmap of RMSE values on test (a) and training (b) data. On the x axis, the preprocessing methods and on the y axis, the regression models are presented.

baseline correction methods in terms of R^2 values. R^2 values of both train and test datasets are presented side-by-side to enable a clear assessment of model generalization and potential overfitting. Each row of the heatmaps corresponds to a different machine learning model, while each column represents one of the applied preprocessing methods. From the preprocessing methods "none", "detrendsnv", "snvsavgol" refer to "no preprocessing application", "Detrend followed by SNV" and "SNV followed by Savitzky-Golay smoothing" respectively.

As seen on the heatmaps, the choice of preprocessing method significantly influences model performance. This highlights the need to carefully select preprocessing techniques that enhance relevant spectral features. In the following sections, we discuss which model preprocessing combination works best and provide more details for the results of each model.

K-Nearest Neighbors Regression (KNN)

The K-Nearest Neighbors (KNN) regression model was assessed using multiple spectral preprocessing techniques. Among them, the FillPeaks method demonstrated the best performance. This configuration yielded a high R^2 0.99 on the training set and 0.96 on the test set, having also the lowest R^2 difference. The corresponding RMSE values were 0.60 for the training set and 1.17 for the test set. Dimension reduction was applied prior to the training.

The optimal hyperparameter space identified via RandomizedSearchCV is presented in Table 4.1

Table 4.1.: Optimal Hyperparameters for KNN

Parameter	Value
distance	Minkowski distance
number of neighbors	10
weighting	uniform weighting

The relationship between the number of neighbors and model generalization was further explored in Figure 4.9. Fillpeaks as pretreatment and, uniform weighting and euclidean metric as model parameters were selected. The validation RMSE score was the lowest at 10 neighbors, suggesting this as the optimal choice. Beyond this point, performance on unseen data gradually declined, confirming the trade-off between bias and variance in KNN models.

With all pretreatment methods, KNN consistently achieved near-perfect R^2 scores on the training set. While they may appear to perform well in-sample, their generalization ability is questionable.

Figure 4.10 compares predicted cotton ratios to the measured values. It can be observed that the model tends to underpredict higher actual values and overpredict lower values, especially below 40% cotton ratio.

4. Results and Discussion

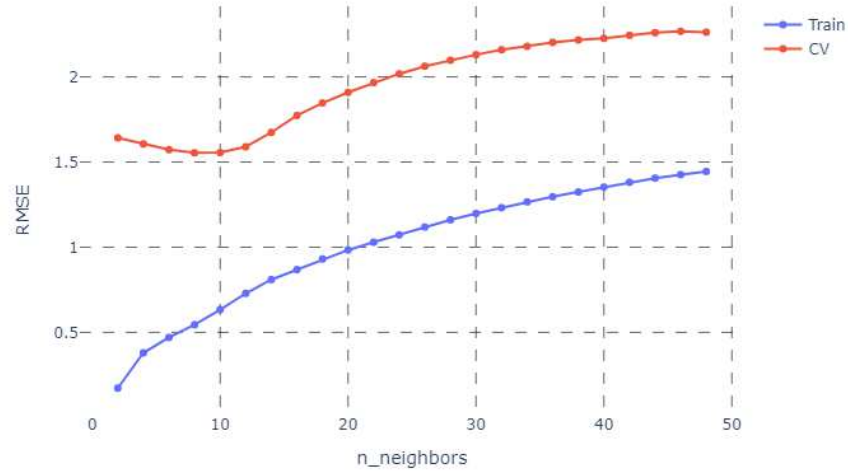


Figure 4.9.: Effect of number of neighbors on KNN Model Performance: RMSE reaches the lowest point when the KNN model uses around 10 neighbors.

Kernel Ridge Regression

Kernel Ridge Regression (KRR) demonstrates strong performance with all preprocessing methods. The training and test R^2 values remain close to each other for most preprocessing methods, indicating that Kernel Ridge benefits from regularization. With SNV, it generates the lowest RMSE value while its performance deteriorates slightly with Fillpeaks, indicating the sensitivity to preprocessing.

The optimal hyperparameters identified through RandomizedSearchCV were:

Table 4.2.: Optimal Hyperparameters for KRR

Parameter	Value
kernel	polynomial
degree	3
alpha	0.01

Alpha is the parameter to control the strength of the regularization term. A higher alpha results in stronger regularization, leading to better generalization. However, underfitting can occur when too high alpha is used. Conversely, using a smaller alpha may cause overfitting, as the model becomes too flexible fitting to the training data points. The optimal alpha value was found to be around 0.01, where the validation error reaches a global minimum. Figure 4.11 illustrates that the errors on the validation and test datasets

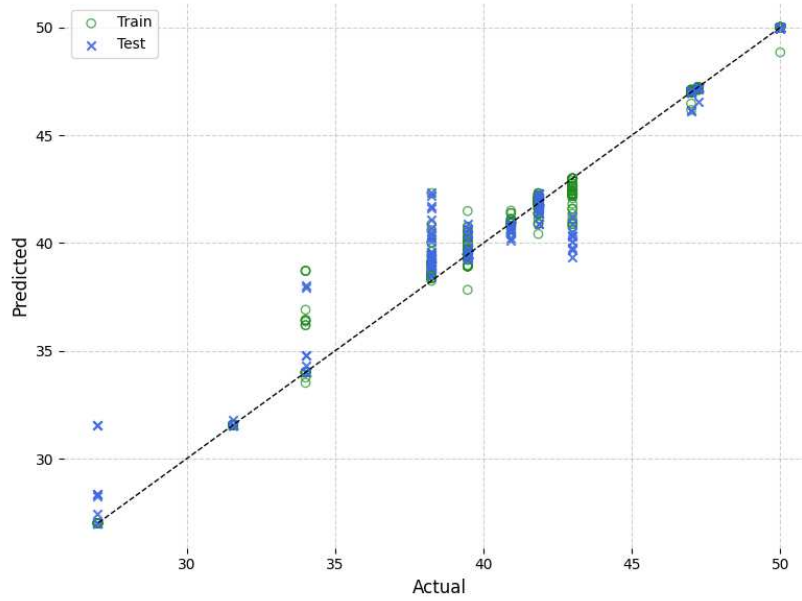


Figure 4.10.: Predicted vs. actual cotton content values for both the train (green circles) and test (blue crosses) datasets using a KNN regression model coupled with Fillpeaks preprocessing

decrease significantly as alpha increases from lower values to the optimal point. This trend confirms the presence of overfitting in the lower alpha region.

Figure 4.12 demonstrates that the model starts to overfit as the polynomial degree increases beyond 3. The minimum cross-validation RMSE at degree 3 suggests that it is the optimal complexity for the model under this configuration. SNV as pretreatment and alpha being 0.01 as model parameters are selected.

Figure 4.13 compares predicted cotton ratios to measured values. A good model fit is observable since train and test predictions are not widely apart from each other and the predictions are not constantly below or above the 1:1 line.

Least Squares Support Vector Regression

Due to the kernel trick, Least Squares Support Vector Regression (LSSVM) is capable of operating effectively in high-dimensional feature spaces without requiring prior dimensionality reduction, therefore we did not perform PCA before LSSVM training.

The optimal hyperparameters identified through RandomizedSearchCV were:

4. Results and Discussion

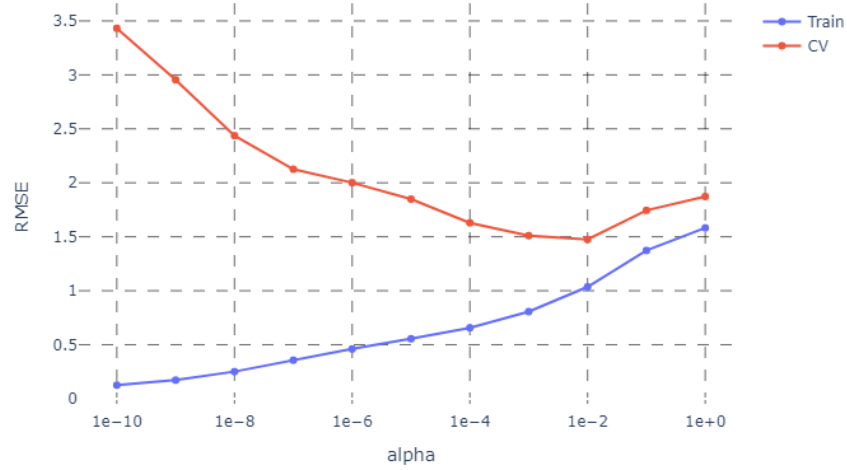


Figure 4.11.: RMSE vs. alpha value for Kernel Ridge Regression: The optimal alpha level seems to be at 0.01 where cross-validation error is minimized

Table 4.3.: Optimal Hyperparameters for LSSVM

Parameter	Value
kernel	polynomial
sigma	1
gamma	0.01

LSSVM demonstrated stable and reliable performance across various preprocessing methods, with particularly favorable results observed when using Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC). The Root Mean Square Error (RMSE) values for both the training and test sets were nearly identical, reflecting a strong degree of generalizability. With SNV and MSC, LSSVM results in only 0.003 R^2 score difference between train (0.961) and test (0.958). This robustness is further supported by the R^2 scores comparison across models shown in Figure 4.7, where LSSVM consistently ranks among the top-performing regressors regardless of preprocessing technique. As shown in Figure 4.14, the predicted versus actual cotton content values are closely aligned along the 1:1 line, indicating high predictive accuracy and low systematic bias.

Partial Least Squares Regression

The Partial Least Squares Regression (PLSR) model demonstrates solid performance across the provided metrics. The RMSE curve indicates that using 10 or 11 components is

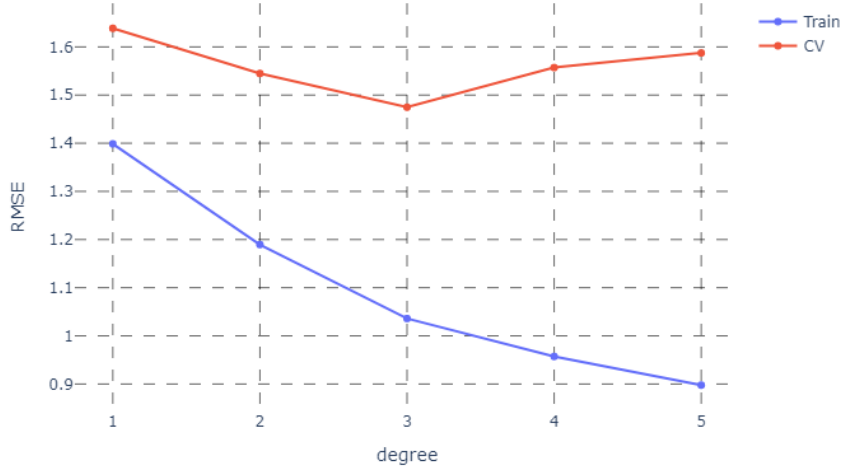


Figure 4.12.: RMSE vs. polynomial degree for Kernel Ridge Regression: Training error decreases with model complexity, while cross-validation error is minimized at degree 3, suggesting it provides the best generalization before overfitting occurs.

optimal (see Figure 4.17), as the error stabilizes beyond this point, suggesting the model generalizes well without overfitting.

In the test set, in general, PLSR achieves competitive RMSE values across different preprocessing methods, while maintaining high R^2 scores, reflecting its strong predictive capability. On the train set, PLS recorded low RMSE values and near-perfect R^2 scores. Although the high training R^2 hints at a potential risk of overfitting, the consistent test performance confirms that the model generalizes effectively. Surprisingly, PLS worked the best without application of any preprocessing (train $R^2=0.99$, test $R^2=0.96$, train RMSE=0.74, test RMSE=1.22).

Large positive values in the PLS loadings correspond to spectral regions where polyester exhibits strong absorbance, while large negative values indicate regions dominated by cotton absorbance. Notably, prominent positive peaks appear around 6000 cm^{-1} and between $5250\text{--}4100\text{ cm}^{-1}$, aligning with characteristic polyester bands. In contrast, a broad negative valley is observed from approximately 7000 to 6200 cm^{-1} , corresponding to high cotton absorbance. Since PLS loadings reflect the contribution of each wavenumber to the prediction of cotton content, regions with high absolute values are particularly important. These patterns suggest that the PLSR model effectively captures the distinct spectral features of both cotton and polyester. The first three latent variable loadings are visualized in Figure 4.15.

Figure 4.16 compares predicted and measured values of both training and test datasets

4. Results and Discussion

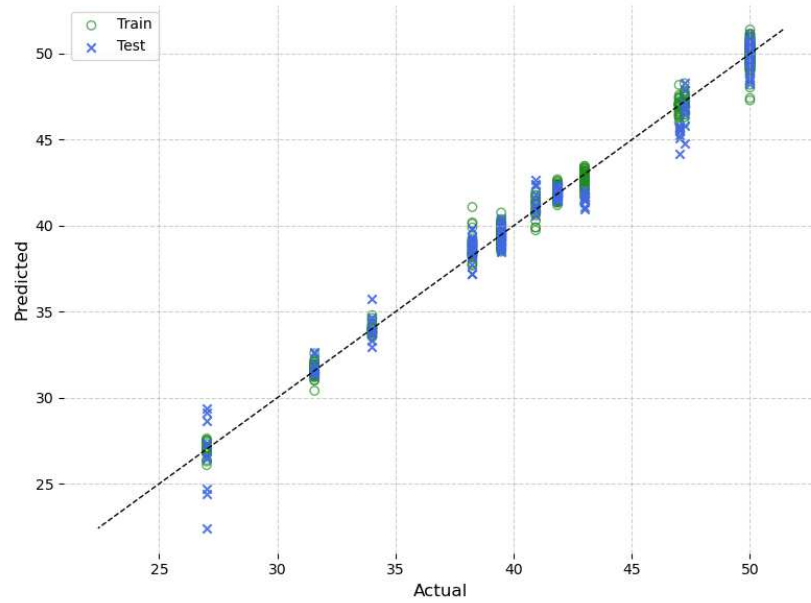


Figure 4.13.: Predicted vs. actual cotton content values for both the train (green circles) and test (blue crosses) datasets using a Kernel Ridge regression model with a polynomial kernel. The dataset was preprocessed with SNV.

using a PLSR model. Both training and test predictions closely align with the diagonal line, also training and test predictions are not spread away from each other. This indicates that the PLSR model captures the relationship between the spectral features and cotton content well.

Overall, PLSR, with around 10 components, balances model complexity and predictive accuracy. As shown in Figure 4.17, both test and cross-validation RMSE values plateau beyond this point and they do not improve further. The slightly lower cross-validation scores are likely because the CV folds are more variable and less balanced than the test set, leading to slightly better performance on the test set.

In summary, PLSR offers more robust modeling approach than more complex models such as XGBoost and Random Forest Regression and it works reasonably well with all preprocessing methods.

Random Forest Regressor

The Random Forest Regressor (RFR) was applied without any prior dimensionality reduction, as this ensemble-based algorithm is suitable for high-dimensional datasets. Among the evaluated preprocessing techniques, RFR achieved its best performance when paired with ALS baseline correction.

It achieves the highest R^2 and lowest RMSE scores using these hyperparameters identified through RandomizedSearchCV:

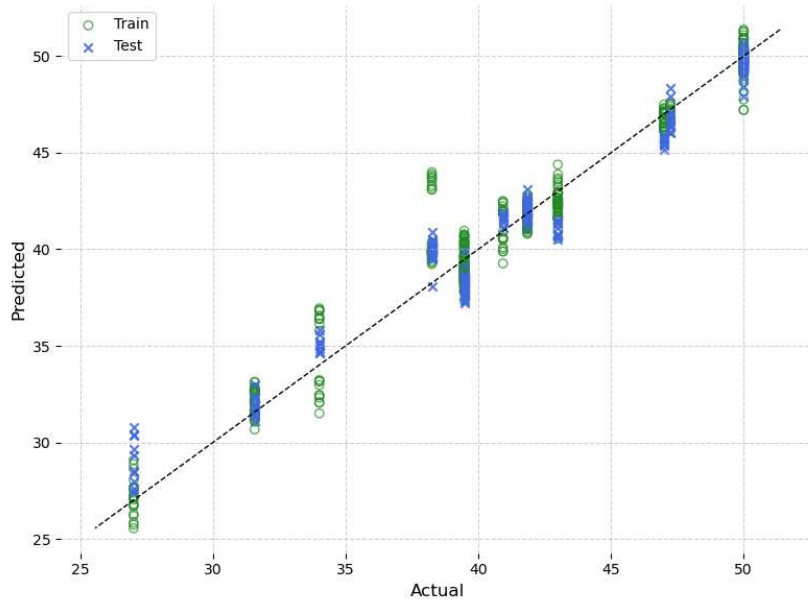


Figure 4.14.: Predicted vs. actual cotton content values for both the train (green circles) and test (blue crosses) datasets using the LSSVM regression model with SNV preprocessing method

Table 4.4.: Optimal Hyperparameters for RFR

Parameter	Value
number of estimators	100
min samples split	5
min samples leaf	4
max feature	log2
max depth	20

The impact of the number of estimators on the R^2 scores was analyzed to determine an optimal configuration (see Figure 4.18). The training R^2 scores remain consistently close to 1, indicating that the model fits the training data nearly perfectly. In contrast, the validation R^2 scores fluctuate slightly around 0.91. The validation R^2 scores improved until 20 estimators. Beyond 20, R^2 did not significantly improve. The slight fluctuations in validation R^2 values likely resulted from the randomness in the data split used for cross-validation. The combination of a near-perfect training score and a lower validation score suggests overfitting. Additionally, there is a noticeable gap between the training and test set R^2 scores overall.

Predicted and measured values generated from both training and testing datasets using a Random Forest model on ALS-preprocessed spectra indicate that there was a tighter fit in the training set compared to the test set (Figure 4.19). Predicted and measured values

4. Results and Discussion

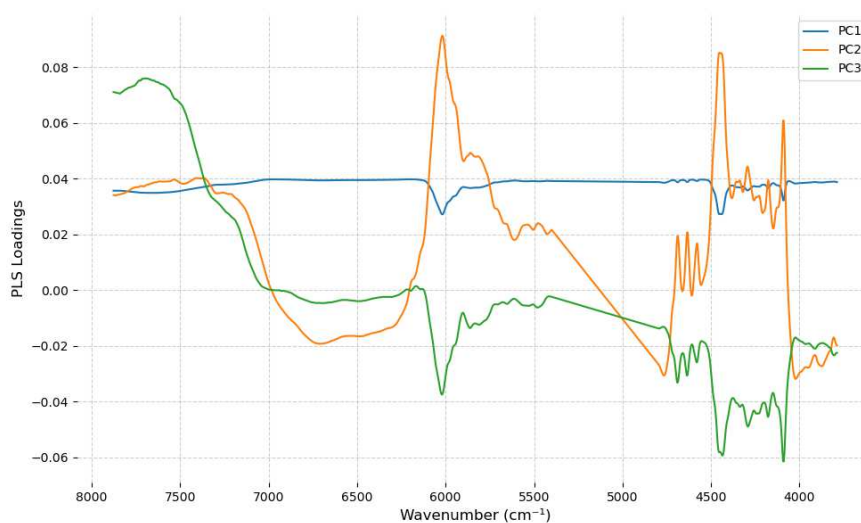


Figure 4.15.: Loadings plot of the first three components. Each curve shows how much a particular wavenumber contributes to that latent variable’s direction in feature space.

especially diverged at 34.01% cotton level. However, most points lie close to the 1:1 line. There is no obvious systematic bias as predictions do not appear consistently above or below the diagonal for any specific cotton content level.

eXtreme Gradient Boosting (XGBoost)

No prior dimensionality reduction via PCA was applied before training, as XGBoost can handle high-dimensional and collinear feature spaces [XGB24]. Among the evaluated models, XGBoost yielded the best prediction performance for cotton content when trained on ALS-preprocessed spectra. The hyperparameter search identified the optimal configuration summarized in Table 4.5.

Table 4.5.: Optimal Hyperparameters for XGBoost

Parameter	Value
number of estimators	200
max depth	6
learning rate	0.1
subsample	0.8

XGBoost consistently achieved unrealistically perfect R^2 scores on the training set, but showed a notable drop on the test set. This strong train-test disparity suggests that these models are likely overfitting and not capturing true underlying patterns. While it may appear to perform well, its generalization ability is limited when exposed to unseen

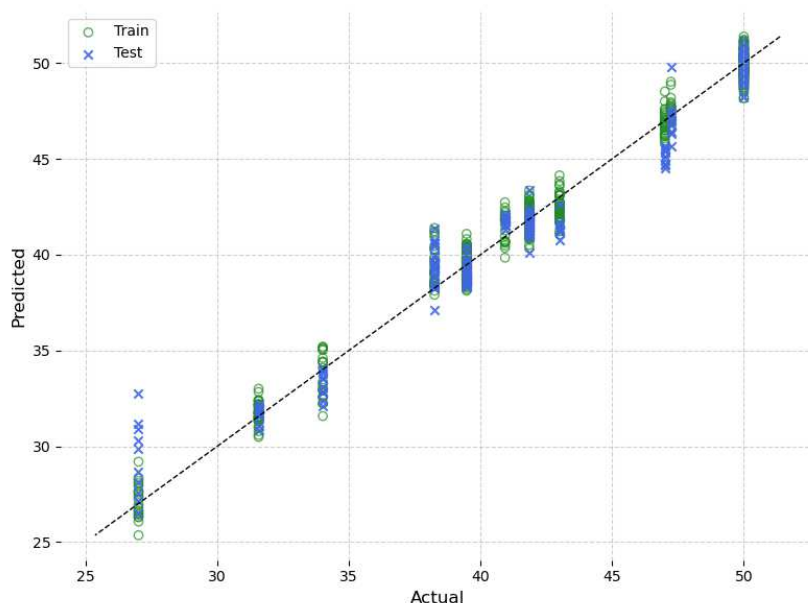


Figure 4.16.: Predicted vs. actual cotton content values for both the train (green circles) and test (blue crosses) datasets using a PLSR model. The dataset was not preprocessed. Both training and test predictions closely follow the diagonal line, indicating that the PLSR model captures the relationship between the spectral features and cotton content well.

samples.

In XGBoost, a weight is calculated based on how many times a feature is used to split the decision trees by the model. The weights are indicative of the importance scores of the features. These scores are provided by the function *getScore* of XGBoost [xgb25], which is used to visualize the scores of the 20 most important wavenumbers in figure 4.20. In our model, high-importance wavenumbers likely correspond to strong absorbance bands. The feature importance analysis identified spectral regions around 4682, 4794, and 4678 cm^{-1} as the most informative for prediction. These regions are consistent with C–H combination and overtone bands.

However, the model did not assign high importance to the 1st overtone of C–H associated with polyester (right below 6000 cm^{-1}), even though it is one of the most prominent peaks in the ALS-preprocessed graph (Figure 4.3 (d)). Similarly, the first overtone of O–H stretching between 7140–5970 cm^{-1} , a key indicator of cotton, was absent from the most important features. Due to these missing overtones, it is questionable that the model fitted the training data well, and captured the key spectral features required for robust generalization.

Figure 4.21 further illustrates XGBoost’s behavior by comparing predicted values to measured ones. Training data predictions closely follow the 1:1 line, confirming the high training accuracy. However, predictions for the test data (blue crosses) display increased

4. Results and Discussion

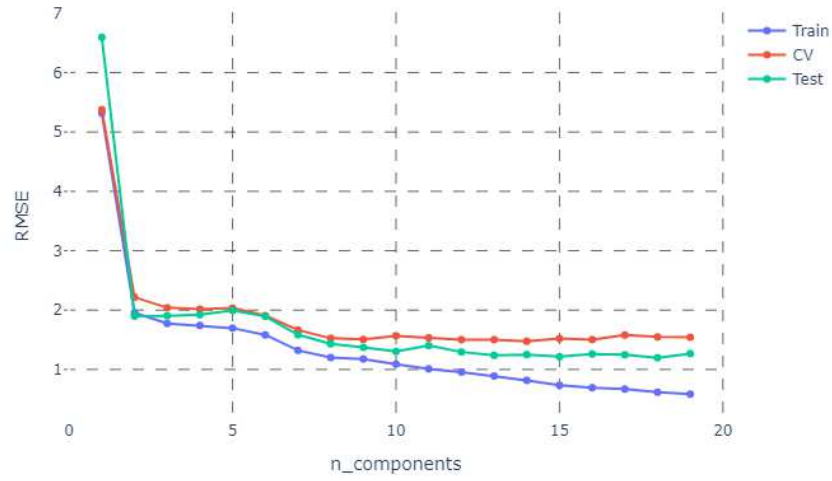


Figure 4.17.: Root Mean Squared Error (RMSE) change depending on the number of PLS components. The error decreases sharply up to around 10 components and then stabilizes.

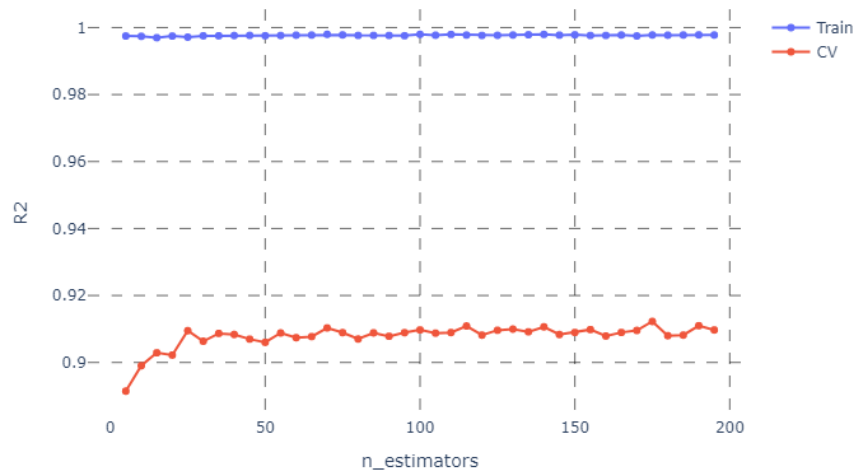


Figure 4.18.: The impact of the number of estimators on the R^2 scores when using the Random Forest model on data preprocessed with the ALS.

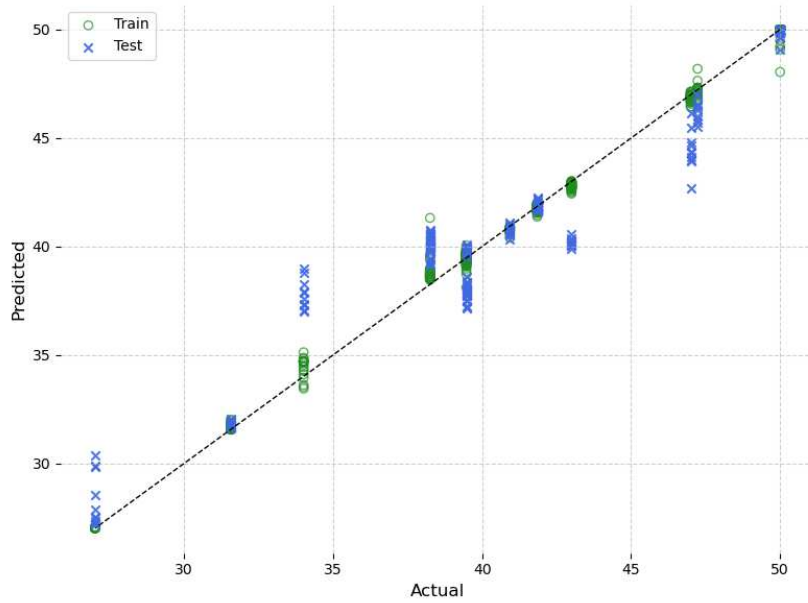


Figure 4.19.: Predicted vs. actual cotton content values for both the train (green circles) and test (blue crosses) datasets using a Random Forest model. The dataset was preprocessed with the ALS method.

scatter, suggesting reduced predictive accuracy for unseen samples. The spread of test data implies that the model may not fully generalize across the range of cotton contents, reinforcing concerns about overfitting.

Convolutional Neural Network

We built a convolutional neural network (CNN) for one-dimensional input data, specifically for modeling spectral data. Its architecture is combined with hyperparameter tuning via `RandomizedSearchCV`.

Initially, a baseline CNN was implemented with a single convolutional layer (filter size 64, kernel size 5), followed by max pooling, ReLU activation, and L2 regularization. The output was flattened and passed through a fully connected layer (128 neurons) with dropout of 30% to mitigate overfitting. The dense layer returned a continuous output for regression. This model was optimized with Adam for 50 epochs and a batch size of 8. Shang et al. [SSW⁺23] employed a similar CNN configuration for their NIR spectra classification task, using a filter size of 64 and a dense layer with 128 neurons. While they applied a dropout rate of 50%, the rate of 30% yielded better performance in our experiments.

To enhance model robustness, Signal-to-Noise Ratio (SNR) augmentation was applied. This approach has been used in similar research contexts for spectral modeling [SSW⁺23, XZJ23].

4. Results and Discussion

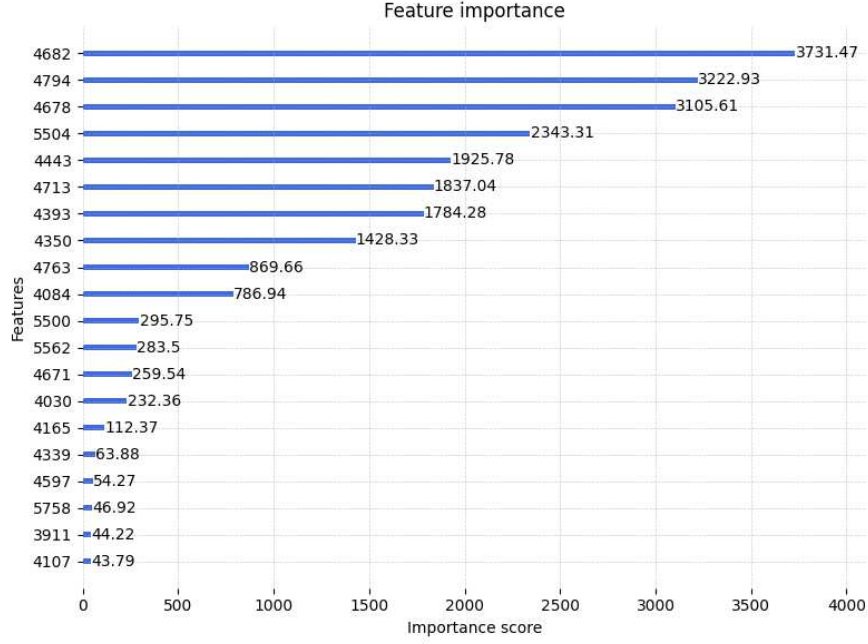


Figure 4.20.: 20 most important features for the XGBoost model are presented. Features express wave numbers, and importance scores represent the weights assigned by the model based on how many times a feature is used to split decision trees.

Further experiments were conducted to evaluate the effects of convolutional depth and L2 regularization on predictive performance. Table 4.6 summarizes the results of nine configurations (A1–A9), which varied in the number of convolutional layers (1–3), L2 penalty, and kept filter size (5) and dense layer size consistent for controlled comparison.

Notably, model A5 achieved the best results with two convolutional layers and an L2 penalty of 0.001 ($R^2 = 0.933$), indicating that moderate depth and regularization helped the model to capture complex spectral patterns. While the 2-layer model A5 achieved the best performance, the 3-layer model A9 was nearly as effective ($R^2 = 0.912$) but with higher training time due to increased computational cost.

Figure 4.22 visualizes the R^2 scores across L2 regularization values for different network depths. The plots show that the 2-layer architecture performs the best for L2 penalty of 10^{-3} . Across all depths, changes in regularization strength lead to noticeable fluctuations in both training and validation R^2 scores, indicating that the model is highly sensitive to hyperparameter selection. These fluctuations are especially severe in the 3-layer architecture. Thus, it cannot be concluded that there is an improvement in generalization power as the number of layers increases. In addition, the cross-validation R^2 is substantially lower than the training score and shows variability across configurations. Given the limited dataset size and the high complexity of CNNs, the model has high variance which means the model swings from underfitting to overfitting.

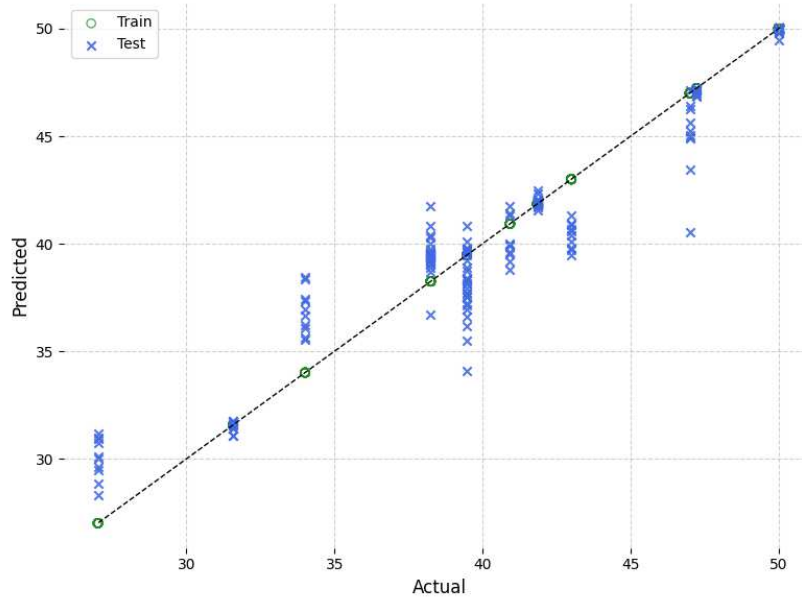


Figure 4.21.: Predicted vs. actual cotton content values for both the train (green circles) and test (blue crosses) datasets using the XGBoost model with the ALS.

Regarding network depth, increasing the number of convolutional layers does not consistently improve generalization. In fact, 3-layer deep architecture tends to fluctuate more than shallower architectures. This suggests that moderate depth, such as the 2-layer configuration, achieves the best generalization.

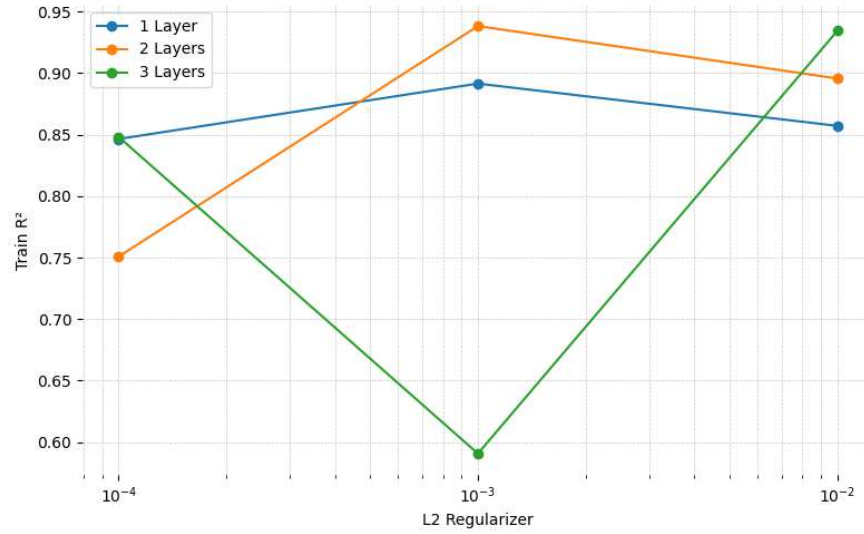
Figure 4.23 presents the learning curve for the 2-layer CNN (architecture A5). The training loss (mean squared error) drops sharply during the first epoch, but shows almost no improvement after the first epoch. The loss values exhibit high variability at the plateau. This unstable convergence is the sign of underfitting.

Commonly, CNN models are used with two-dimensional input data. But due to SKLearnRegressor (a scikit-learn wrapper for Keras models) restrictions, transforming the data from one-dimensional to two-dimensional was not possible. Specifically, SKLearnRegressor does not accept multidimensional inputs. Even when reshaping the data inside the pipeline, scikit-learn's internal validation still enforced a 2D shape. Given these constraints, two-dimensional convolutional models were excluded from the current study and are proposed as future work.

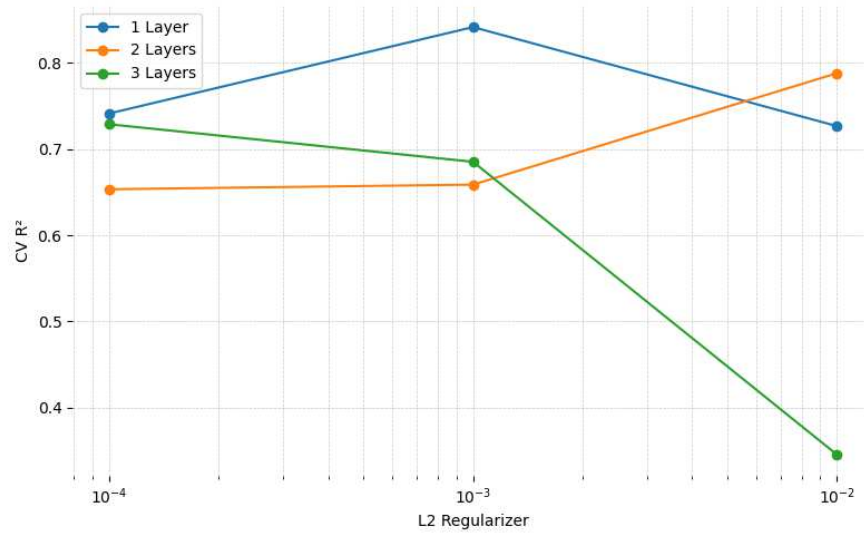
Best Performing Models

The Kernel Ridge Regression with optimal hyperparameters and the SNV method achieved the best R^2 score (0.978) among all machine learning models and preprocessing combinations. It also performed the lowest test RMSE of 0.92 and a train R^2 of 0.51. The differences between the test and train metrics, with an RMSE difference of 0.41 and R^2

4. Results and Discussion



(a) Train



(b) Cross-Validation

Figure 4.22.: Effect of L2 regularization on R^2 performance across CNN architectures with different convolutional depths (1, 2, and 3 layers). Both training (a) and cross-validation (b) R^2 scores show fluctuations in R^2 score. The 3-layer deep architecture has more volatility than other architectures, making it less robust.

4.3. Prediction Model Performance

Table 4.6.: CNN architecture performance across different hyperparameter settings. De-trend and SNV were used as preprocessing.

Architecture	Conv Layers	Filter	Dense Layer	Kernel Size	L2 λ	RMSE (Test)	R^2 (Test)
A1	1	64	128	5	0.0001	2.551	0.830
A2	1	64	128	5	0.001	2.024	0.893
A3	1	64	128	5	0.01	2.581	0.826
A4	2	64	128	5	0.0001	2.988	0.766
A5	2	64	128	5	0.001	1.605	0.933
A6	2	64	128	5	0.01	2.246	0.868
A7	3	64	128	5	0.0001	2.118	0.883
A8	3	64	128	5	0.001	4.262	0.524
A9	3	64	128	5	0.01	1.833	0.912

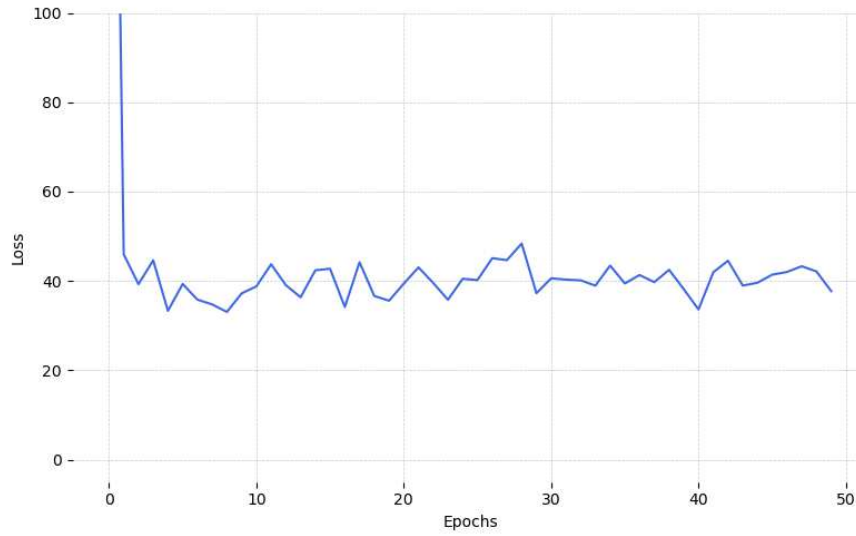


Figure 4.23.: Mean Squared Error loss over epochs for CNN architecture A5. The training loss was recorded for 50 epochs.

difference of 0.015, are low, indicating that the model provides a good fit. The scores are shown in Figures 4.7 and 4.8.

Beyond KRR, PLS ($R^2=0.961$, RMSE=1.215), LSSVM ($R^2=0.958$, RMSE=1.263), and KNN ($R^2=0.964$, RMSE=1.203) also delivered competitive test results. To further evaluate their robustness, the two strongest preprocessing methods were selected for each model. Robustness was assessed using repeated cross-validation. 4-fold cross-validation was repeated 10 times and the cross-validation score of each fold was collected. Overall, 40 scores for each mode were recorded and summarized as boxplots in Figure 4.24.

PLS, LSSVM and KRR achieved high R^2 scores on cross-validation folds. Their means and medians are high and interquartile ranges are relatively narrow and they have lower number of outliers. However, KRR exhibited slightly lower mean and median and one negative R^2 as an outlier.

Tree-based models (RF and XGBoost) and KNN were moderately strong with several

4. Results and Discussion

outliers. Although the interquartile range of KNN is narrow, the number of outliers is higher than any other models. CNN showed the lowest mean and median R^2 scores and widest box since R^2 scores had a very wide spread. CNN was highly unstable under the our configuration and preprocessing.

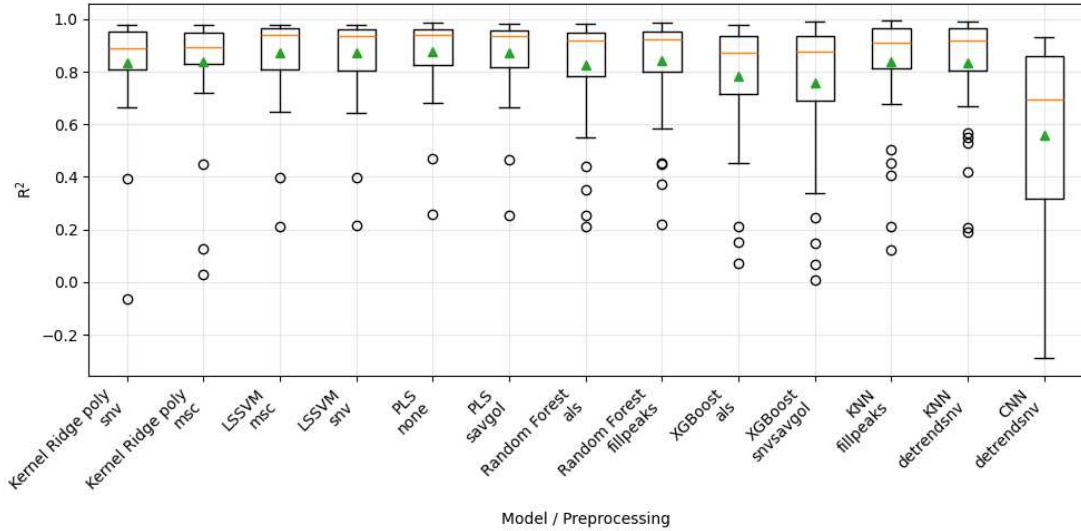


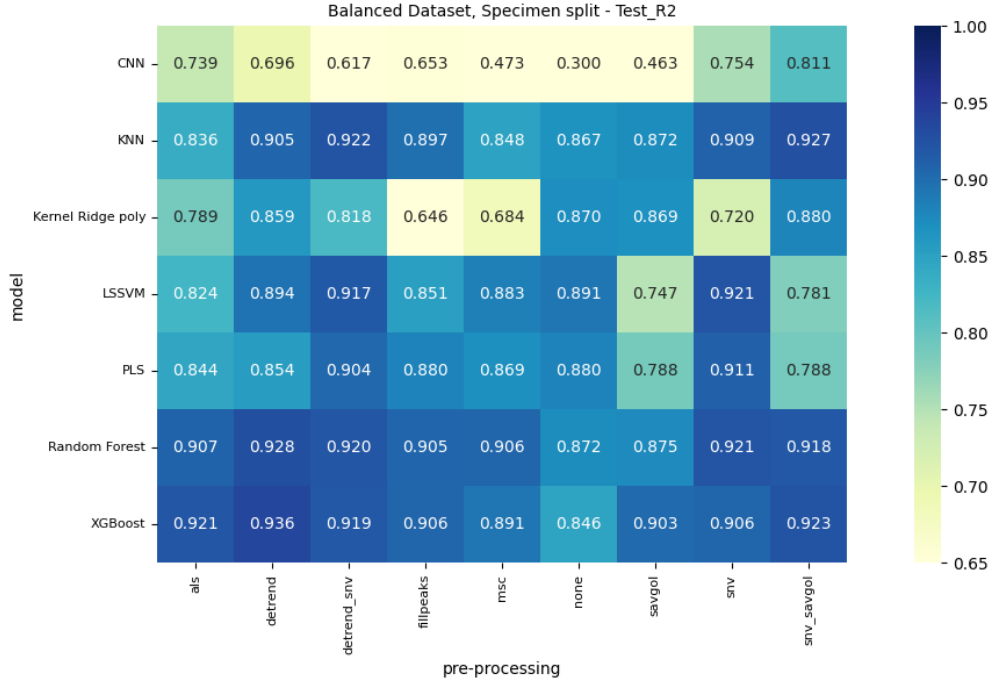
Figure 4.24.: The cross-validation R^2 scores from two best preprocessing method for each model were depicted in the box plot. The selection was made according to test R^2 . The medians were shown with an orange line while the means were indicated with a green triangles.

4.3.2. Regression Model Performance on MIR Spectral Data in Comparison to NIR

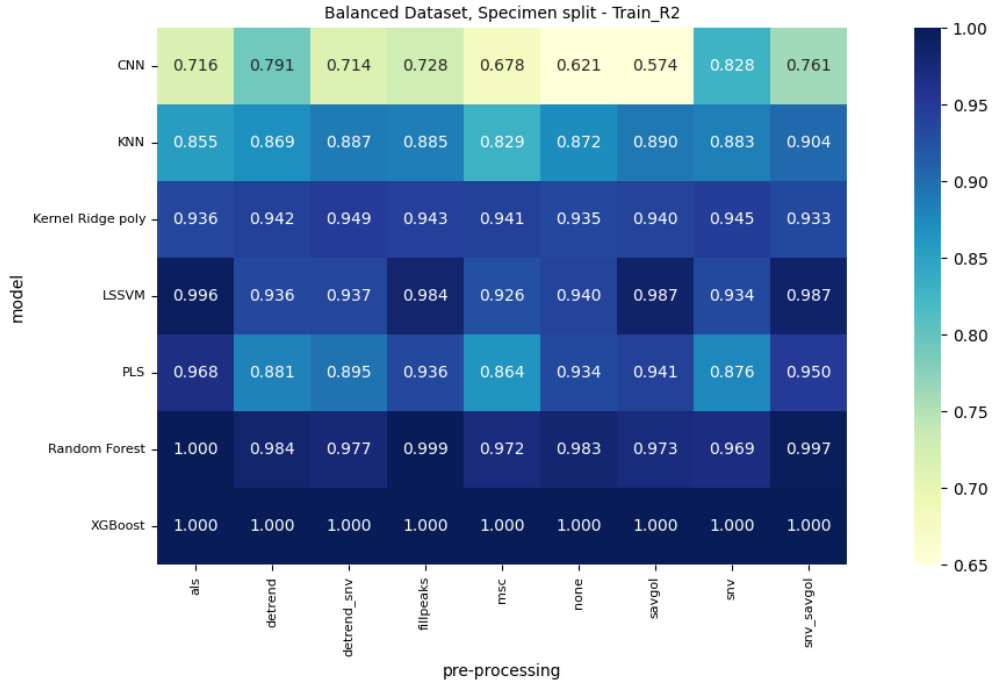
As shown in Figure 4.25, the MIR results are consistently worse than those from NIR for all models. In most cases, the R^2 values for MIR are lower, and even when the training R^2 occasionally exceeds that of NIR, the test R^2 drops significantly below the NIR score. This is an indicator of overfitting. Additionally, MIR spectra consistently yield higher RMSE values than NIR across nearly all models and preprocessing methods, particularly on the test set. For this reason, we do not provide a detailed discussion of each model's MIR performance. Since NIR produced better results overall, the analysis in this thesis focuses on the NIR data.

The differences between NIR and MIR performance can be attributed to several factors. First, our NIR dataset was larger than the MIR dataset, since NIR spectroscopy enabled faster and more practical data acquisition than MIR. Second, MIR spectroscopy had a smaller scan area, which reduced the representativeness of the measurements for our textile specimens. This limitation was particularly important given that the specimens were not composed of fully homogeneous fibers.

4.3. Prediction Model Performance



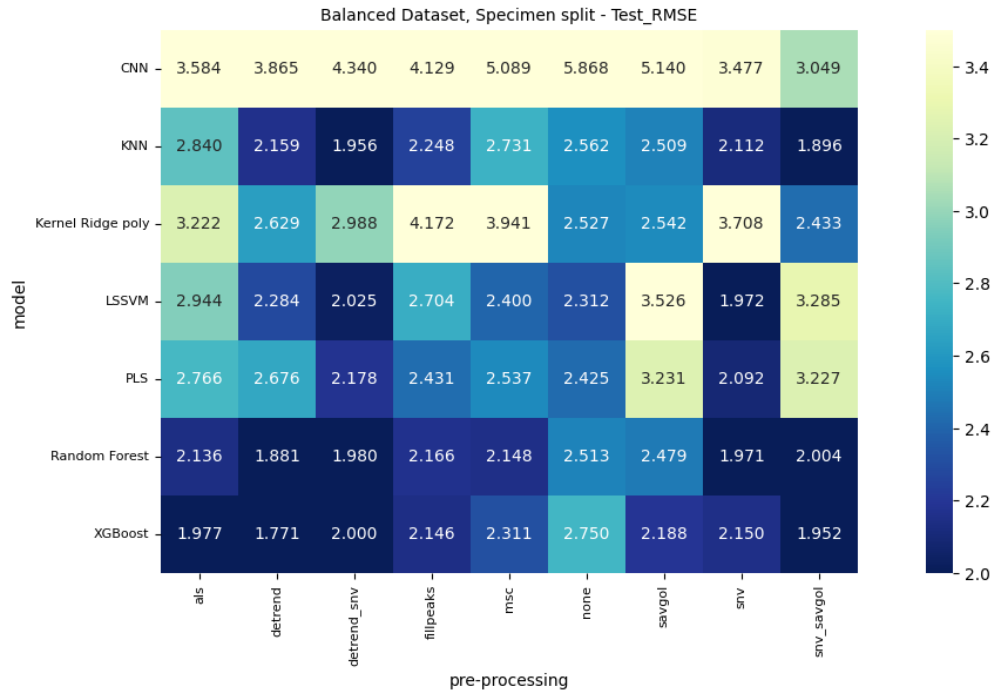
(a) Test Set



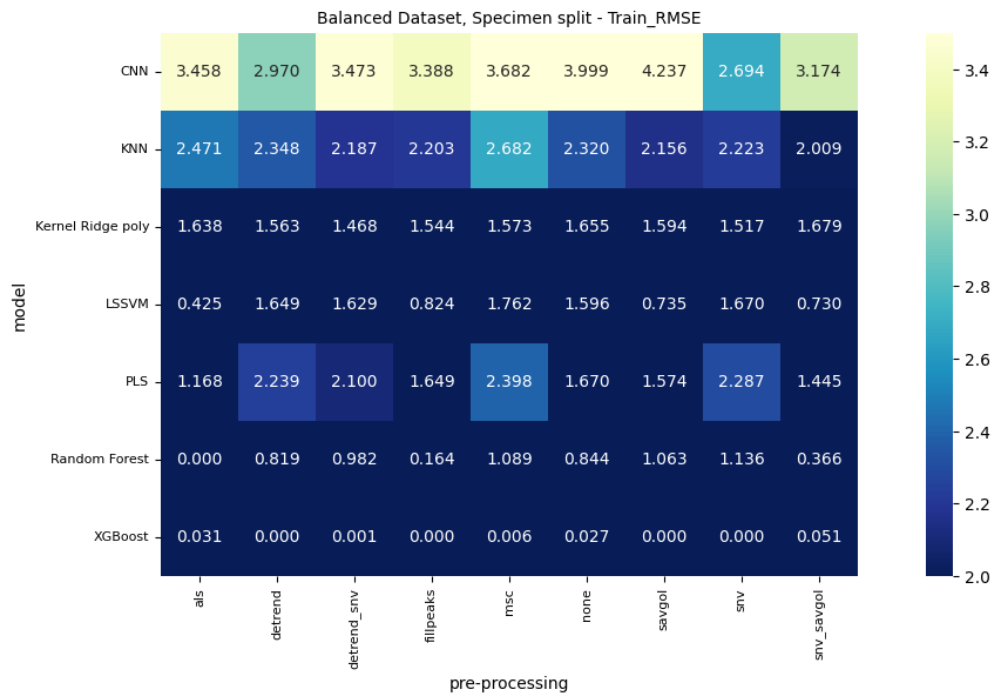
(b) Train Set

Figure 4.25.: Heatmap of R^2 values on test (a) and training (b) data. On the x axis, the preprocessing methods and on the y axis, the regression models are presented.

4. Results and Discussion



(a) Test Set



(b) Train Set

Figure 4.26.: Heatmap of RMSE values on test (a) and training (b) data. On the x axis, the preprocessing methods and on the y axis, the regression models are presented.

5. Conclusion and Outlook

5.1. Conclusion

This study explored the application of infrared spectroscopy and supervised machine learning models for the quantification of cotton-polyester blended textiles, with the goal of improving textile recycling processes. A reliable and fast textile sorting technology has been demonstrated to be possible as a result of the study.

A systematic approach was developed to determine the optimal sample size, ensuring statistically representative and efficient measurements. The Bootstrap resampling method was found to be an effective and computationally efficient approach for determining the minimum number of measurement spots per specimen. Using a margin-of-error-based approach applied to peak absorbance values, it was found that measuring at least 10 spots per specimen reliably captures the spectral variability.

For predictive modeling, several regression algorithms were evaluated, including Partial Least Squares Regression (PLSR), Least Squares Support Vector Machine (LSSVM), Kernel Ridge Regression (KRR), Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting (XGBoost), and a Convolutional Neural Network (CNN). Among all models, the combination of SNV and Kernel Ridge Regression emerged as the model with the lowest RMSE test score that is less than 1. However, KRR proved more sensitive to data splits than PLSR and LSSVM, which delivered strong performance in narrow cross-validation dispersion. PLSR and LSSVM consistently performed well across most preprocessing methods, whereby LSSVM generalized the best, as the train and test scores were very close. KNN, while exhibiting excellent training scores, showed potential overfitting and bias in predicting low cotton ratios. Random Forest and XGBoost tended to overfit while CNN exhibited large variability.

We observed that the choice of pretreatment method plays a critical role in regression accuracy and model generalization. Methods such as Detrend, ALS, and FillPeaks proved highly effective in correcting baseline drifts, while SNV and MSC were valuable in reducing scatter-induced variability. The effectiveness of these techniques, however, varied by model, confirming that preprocessing should be adaptively selected in conjunction with the chosen regression algorithm.

The results suggest that combination of Near-Infrared (NIR) spectroscopy with machine learning models provides a fast, non-destructive, and accurate approach for textile quantification, addressing a critical need in textile recycling. Although the fundamental vibrations of molecules occur in the MIR region, NIR spectroscopy provided higher quality spectral data. Thanks to its wide scanning area, it allowed us to measure a larger portion of the textile samples and therefore collect more reliable data. NIR spectroscopy also

5. Conclusion and Outlook

performed faster measurements than MIR spectroscopy.

In summary, the findings of this study will contribute to the development of automatic and scalable textile separation systems, increasing efficiency in textile recycling and supporting the circular economy model in the sector.

5.2. Outlook

Based on the promising findings of this study, we can discuss some future developments that could increase the scalability and robustness of textile quantification using IR spectroscopy.

The current dataset can be expanded by adding more textile samples with various cotton-polyester fiber compositions. A larger dataset allows for more robust model training and better generalization to real-world textile waste streams. However, preparing and measuring textile samples to expand the dataset is labour-intensive and time-consuming. Therefore, this thesis sought and calibrated the most generalizable and accurate regression models with a limited dataset. In the next phases, the number of textile samples will be increased and the dataset will be enlarged.

Besides increasing the number of samples, expanding the range of the cotton content in the dataset would enable more accurate quantification of textile samples with a cotton content outside of our range of 50% to 27%. As the dataset grows, the importance of outlier detection increases. While on-the-fly outlier removal worked due to the small dataset, integrating methods such as Mahalanobis distance or PLS residuals will ensure objectivity and scalability in a bigger dataset. Additionally, future work should explore feature selection alongside dimensionality reduction. Feature selection can retain chemically meaningful information and reduce multicollinearity without excessive data compression, which can be advantageous for complex models.

To ensure that the observed data are a good representation of the population, the measurement database will be expanded in the future with additional measurements using spectroscopy infrastructures from different institutes. These measurements will be compared with data obtained with our spectroscopy infrastructure to further validate and improve the measurement methods. The same models will be trained on the datasets and the results will be observed.

The current approach does not take into account factors of wet textiles and black or dark fabric samples. Using conventional NIR and MIR spectroscopy, measuring dark-colored textile samples is difficult due to reduced reflectance and reduced signal quality. Water content also introduces strong absorption bands that overlap critical spectral features. Expanding the study to include Raman spectroscopy could provide complementary spectral data and further increase the accuracy of fiber identification, as Raman spectroscopy is insensitive to water and dark color dyes.

As we demonstrated, it is possible to estimate the cotton content of textiles using regression models. Therefore, the project could be expanded to the quantification of other fibers like silk or viscose in the future. In this context, the classification of different fibers using machine learning models could be another future goal.

Bibliography

- [AAD⁺15] Jeremy Adcock, Euan Allen, Matthew Day, Stefan Frick, Janna Hinchliff, Mack Johnson, Sam Morley-Short, Sam Pallister, Alasdair Price, and Stasja Stanisic. Advances in quantum machine learning, 2015. arXiv Preprint. URL: <https://arxiv.org/abs/1512.02900>, arXiv:1512.02900.
- [All23] Oskar Allerbo. Fast Robust Kernel Regression through Sign Gradient Descent with Early Stopping. 2023. arXiv Preprint. URL: <https://arxiv.org/abs/2306.16838>, doi:10.48550/ARXIV.2306.16838.
- [ALV07] Senjian An, Wanquan Liu, and Svetha Venkatesh. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition*, 40(8):2154–2162, August 2007. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0031320306005280>, doi:10.1016/j.patcog.2006.12.015.
- [BC07] Donald A. Burns and Emil W. Ciurczak. *Handbook of Near-Infrared Analysis*. CRC Press, 3rd edition, September 2007. URL: <https://www.taylorfrancis.com/books/9781420007374>, doi:10.1201/9781420007374.
- [BDL89] R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, 43(5):772–777, July 1989. URL: <https://journals.sagepub.com/doi/10.1366/0003702894202201>, doi:10.1366/0003702894202201.
- [BPS98] Beebe, Pell, and Seasholtz. *Chemometrics*. 4. John Wiley & Sons, 1998.
- [BYX⁺16] Yiming Bi, Kailong Yuan, Weiqiang Xiao, Jizhong Wu, Chunyun Shi, Jun Xia, Guohai Chu, Guangxin Zhang, and Guojun Zhou. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Analytica Chimica Acta*, 909:30–40, February 2016. URL: <https://linkinghub.elsevier.com/retrieve/pii/S000326701630054X>, doi:10.1016/j.aca.2016.01.010.
- [CD14] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, June 2014. URL: <https://gmd.copernicus.org/articles/7/1247/2014/>, doi:10.5194/gmd-7-1247-2014.

Bibliography

- [CH07] Haiyan Cen and Yong He. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, 18(2):72–83, February 2007. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0924224406002524>, doi:10.1016/j.tifs.2006.09.003.
- [CLYL04] Nianyi Chen, Wencong Lu, Jie Yang, and Guozheng Li. *Support Vector Machine in Chemistry*. World Scientific, Singapore, 2004.
- [CTL20] Hui Chen, Chao Tan, and Zan Lin. Quantitative Determination of the Fiber Components in Textiles by Near-Infrared Spectroscopy and Extreme Learning Machine. *Analytical Letters*, 53(6):844–857, April 2020. URL: <https://www.tandfonline.com/doi/full/10.1080/00032719.2019.1683742>, doi:10.1080/00032719.2019.1683742.
- [DK99] Marianne Defernez and E. Katherine Kemsley. Avoiding overfitting in the analysis of high-dimensional data with artificial neural networks (ANNs). *The Analyst*, 124(11):1675–1681, 1999. URL: <https://xlink.rsc.org/?DOI=A905556H>, doi:10.1039/A905556H.
- [DMHG⁺21] Tom Duhoux, Edwin Maes, Martin Hirschnitz-Garbers, Karolien Peeters, Lise Asscherickx, Maarten Christis, Birgit Stubbe, Philippe Colignon, Mandy Hinzmann, and Anurodh Sachdeva. Study on the technical, regulatory, economic and environmental effectiveness of textile fibres recycling. *Final Report to the European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. European Commission: Brussels.*, 2021. URL: <https://op.europa.eu/en/publication-detail/-/publication/739a1cca-6145-11ec-9c6c-01aa75ed71a1>, doi:10.2873/828412.
- [DMJRM00] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.
- [DSDSP18] Felipe Bachion De Santana, André Marcelo De Souza, and Ronei Jesus Poppi. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 191:454–462, February 2018. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1386142517308570>, doi:10.1016/j.saa.2017.10.052.
- [Efr79] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), January 1979. URL: <https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-1/Bootstrap-Methods-Another-Look-at-the-Jackknife/10.1214/aos/1176344552.full>, doi:10.1214/aos/1176344552.

- [FMB18] Marco Aurélio de Menezes Franco, Débora Marcondes Bastos Pereira Milori, and Paulino Ribeiro Villas Boas. Comparison of Algorithms for Baseline Correction of LIBS Spectra for Quantifying Total Carbon in Brazilian Soils, 2018. Version Number: 2. URL: <https://arxiv.org/abs/1805.03695>, doi:10.48550/ARXIV.1805.03695.
- [Fri01] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), October 2001. URL: <https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full>, doi:10.1214/aos/1013203451.
- [FS06] Herwig Friedl and Erwin Stampfer. *Jackknife Resampling*. John Wiley & Sons, Ltd, 2006. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470057339.vaj001>, arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470057339.vaj001>, doi:10.1002/9780470057339.vaj001.
- [Gal20] Neal Gallagher. Fitting Smooth Curves Part III: Baselineing with an Asymmetric Least-Squares Algorithm. 2020. URL: <https://rgdoi.net/10.13140/RG.2.2.13628.62080>, doi:10.13140/RG.2.2.13628.62080.
- [HGM20] James Houston, Frank G. Glavin, and Michael G. Madden. Robust Classification of High-Dimensional Spectroscopy Data Using Deep Learning and Data Synthesis. *Journal of Chemical Information and Modeling*, 60(4):1936–1954, April 2020. URL: <https://pubs.acs.org/doi/10.1021/acs.jcim.9b01037>, doi:10.1021/acs.jcim.9b01037.
- [HMOR05] Tom Howley, Michael G Madden, Marie-Louise O’Connell, and Alan G Ryder. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 209–222. Springer, 2005.
- [Hou09] Max M Houck. *Identification of textile fibers*. Woodhead Publishing, 2009.
- [HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3), June 2008. URL: <http://dx.doi.org/10.1214/0090536070000000677>, doi:10.1214/0090536070000000677.
- [JLLY22] Jeanger P. Juanga-Labayen, Ildefonso V. Labayen, and Qiuyan Yuan. A Review on Textile Recycling Practices and Challenges. *Textiles*, 2(1):174–188, March 2022. URL: <https://www.mdpi.com/2673-7248/2/1/10>, doi:10.3390/textiles2010010.

Bibliography

- [Jos23] Payal B. Joshi. Navigating with chemometrics and machine learning in chemistry. *Artificial Intelligence Review*, 56(9):9089–9114, September 2023. URL: <https://link.springer.com/10.1007/s10462-023-10391-w>, doi:10.1007/s10462-023-10391-w.
- [KP23] Theodoros Kyriazos and Mary Poga. Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, 13(03):404–424, 2023. URL: <https://www.scirp.org/journal/doi.aspx?doi=10.4236/ojs.2023.133020>, doi:10.4236/ojs.2023.133020.
- [LADK⁺15] Rocco Langone, Carlos Alzate, Bart De Ketelaere, Jonas Vlasselaer, Wannes Meert, and Johan A.K. Suykens. LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines. *Engineering Applications of Artificial Intelligence*, 37:268–278, January 2015. URL: <https://linkinghub.elsevier.com/retrieve/pii/S095219761400222X>, doi:10.1016/j.engappai.2014.09.008.
- [LDHD07] Y. Langeron, M. Doussot, D.J. Hewson, and J. Duchêne. Classifying NIR spectra of textile products with kernel methods. *Engineering Applications of Artificial Intelligence*, 20(3):415–427, April 2007. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0952197606001084>, doi:10.1016/j.engappai.2006.07.001.
- [LDTM16] Yongliang Liu, Chris Delhom, B. Todd Campbell, and Vikki Martin. Application of near infrared spectroscopy in cotton fiber micronaire measurement. *Information Processing in Agriculture*, 3(1):30–35, 2016. URL: <https://www.sciencedirect.com/science/article/pii/S2214317315300159>, doi:10.1016/j.inpa.2016.01.001.
- [Lil15] Kristian Hovde Liland. 4S Peak Filling – baseline estimation by iterative mean suppression. *MethodsX*, 2:135–140, 2015. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2215016115000126>, doi:10.1016/j.mex.2015.02.009.
- [LKCT94] Fred O. Libnau, Olav M. Kvalheim, Alfred A. Christy, and Jostein Toft. Spectra of water in the near- and mid-infrared region. *Vibrational Spectroscopy*, 7(3):243–254, September 1994. URL: <https://linkinghub.elsevier.com/retrieve/pii/0924203194850143>, doi:10.1016/0924-2031(94)85014-3.
- [LMWX19] Jiaping Li, Xiang Meng, Wendi Wang, and Binjie Xin. A novel hyperspectral imaging and modeling method for the component identification of woven fabrics. *Textile Research Journal*, 89(18):3752–3767, September 2019. URL: <http://journals.sagepub.com/doi/10.1177/0040517518821907>, doi:10.1177/0040517518821907.

- [LZL⁺18] Yisen Liu, Songbin Zhou, Weixin Liu, Xinhui Yang, and Jun Luo. Least-squares support vector machine and successive projection algorithm for quantitative analysis of cotton-polyester textile by near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 26(1):34–43, February 2018. URL: <http://journals.sagepub.com/doi/10.1177/0967033518757069>, doi: 10.1177/0967033518757069.
- [Mar19] Christina Margariti. The application of FTIR microspectroscopy in a non-invasive and non-destructive way to the study and conservation of mineralised excavated textiles. *Heritage Science*, 7(1):63, December 2019. URL: <https://heritagesciencejournal.springeropen.com/articles/10.1186/s40494-019-0304-8>, doi:10.1186/s40494-019-0304-8.
- [McI16] Avery McIntosh. The Jackknife Estimation Method, June 2016. arXiv Preprint. URL: <http://arxiv.org/abs/1606.00497>, doi:10.48550/arXiv.1606.00497.
- [MN89] Harald Martens and Tormod Naes. *Multivariate calibration*. Wiley, 1989. URL: <https://ubdata.univie.ac.at/AC02013998>.
- [MPM⁺22] Puneet Mishra, Dário Passos, Federico Marini, Junli Xu, Jose M. Amigo, Aoife A. Gowen, Jeroen J. Jansen, Alessandra Biancolillo, Jean Michel Roger, Douglas N. Rutledge, and Alison Nordon. Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends in Analytical Chemistry*, 157:116804, December 2022. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0165993622002874>, doi: 10.1016/j.trac.2022.116804.
- [MRS20] Mikko Mäkelä, Marja Rissanen, and Herbert Sixta. Machine vision estimates the polyester content in recyclable waste textiles. *Resources, Conservation and Recycling*, 161:105007, October 2020. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0921344920303244>, doi:10.1016/j.resconrec.2020.105007.
- [MT19] Jelena Muncan and Roumiana Tsenkova. Aquaphotomics—From Innovative Knowledge to Integrative Platform in Science and Technology. *Molecules*, 24(15):2742, July 2019. URL: <https://www.mdpi.com/1420-3049/24/15/2742>, doi:10.3390/molecules24152742.
- [N⁺91] Nico JD Nagelkerke et al. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991. doi:10.1093/biomet/78.3.691.
- [NMH⁺20] Laura Navone, Kaylee Moffitt, Kai-Anders Hansen, James Blinco, Alice Payne, and Robert Speight. Closing the textile loop: Enzymatic fibre separation and recycling of wool/polyester fabric blends. *Waste Management*,

Bibliography

- 102:149–160, February 2020. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0956053X19306567>, doi:10.1016/j.wasman.2019.10.026.
- [NTB⁺16] Diana Neves, Márcio Talhavini, Jez Willian Braga, Jorge Zacca, and Eloisa Caldas. Detection of Counterfeit Durateston® Using Fourier Transform Infrared Spectroscopy and Partial Least Squares - Discriminant Analysis. *Journal of the Brazilian Chemical Society*, 2016. URL: http://jbcs.sbq.org.br/audiencia_pdf.asp?aid2=4836&nomeArquivo=160630AR.pdf, doi:10.21577/0103-5053.20160293.
- [OM21] Yukihiro Ozaki and Yusuke Morisawa. Principles and Characteristics of NIR Spectroscopy. In Yukihiro Ozaki, Christian Huck, Satoru Tsuchikawa, and Søren Balling Engelsen, editors, *Near-Infrared Spectroscopy*, pages 11–35. Springer Singapore, Singapore, 2021. URL: http://link.springer.com/10.1007/978-981-15-8648-4_2, doi:10.1007/978-981-15-8648-4_2.
- [Oza21] Yukihiro Ozaki. Infrared Spectroscopy—Mid-infrared, Near-infrared, and Far-infrared/Terahertz Spectroscopy. *Analytical Sciences*, 37(9):1193–1212, September 2021. URL: <https://link.springer.com/10.2116/analsci.20R008>, doi:10.2116/analsci.20R008.
- [PKVL19] Pilleriin Peets, Karl Kaupmees, Signe Vahur, and Ivo Leito. Reflectance FT-IR spectroscopy as a viable option for textile fiber identification. *Heritage Science*, 7(1):93, December 2019. URL: <https://heritagesciencejournal.springeropen.com/articles/10.1186/s40494-019-0337-z>, doi:10.1186/s40494-019-0337-z.
- [PLPV17] Pilleriin Peets, Ivo Leito, Jaan Pelt, and Signe Vahur. Identification and classification of textile fibres using ATR-FT-IR spectroscopy with chemometric methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 173:175–181, February 2017. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1386142516305236>, doi:10.1016/j.saa.2016.09.007.
- [PMF⁺22] Michaela Poth, Gordon Magill, Alois Filgertshofer, Oliver Popp, and Tobias Großkopf. Extensive evaluation of machine learning models and data preprocessings for Raman modeling in bioprocessing. *Journal of Raman Spectroscopy*, 53(9):1580–1591, September 2022. URL: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.6402>, doi:10.1002/jrs.6402.
- [PRA⁺16] Emanuel Henrique Portella, Daiane Romanzini, Clarissa Coussirat Angrizani, Sandro Campos Amico, and Ademir José Zattera. Influence of Stacking Sequence on the Mechanical and Dynamic Mechanical Properties of Cotton/Glass Fiber Reinforced Polyester Composites. *Materials Research*, 19(3):542–547, 2016. doi:10.1590/1980-5373-MR-2016-0058.

- [PS03] Per-Olof Persson and Gilbert Strang. Smoothing by Savitzky-Golay and Legendre Filters. In Douglas N. Arnold, Fadil Santosa, Joachim Rosenthal, and David S. Gilliam, editors, *Mathematical Systems Theory in Biology, Communications, Computation, and Finance*, volume 134, pages 301–315. Springer New York, New York, NY, 2003. Series Title: The IMA Volumes in Mathematics and its Applications. URL: http://link.springer.com/10.1007/978-0-387-21696-6_11, doi:10.1007/978-0-387-21696-6_11.
- [Que49] M. H. Quenouille. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):68–84, 1949. URL: <http://www.jstor.org/stable/2983696>.
- [Qum13] Maher Qumsiyeh. Using the Bootstrap for Estimating the Sample Size in Statistical Experiments. *Journal of Modern Applied Statistical Methods*, 12(1):45–53, May 2013. URL: <https://jmasm.com/index.php/jmasm/article/view/640>, doi:10.22237/jmasm/1367381280.
- [RBE09] Åsmund Rinnan, Frans Van Den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222, November 2009. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0165993609001629>, doi:10.1016/j.trac.2009.07.007.
- [ROD⁺06] C. Ruckebusch, F. Orhan, A. Durand, T. Boubellouta, and J. P. Huvenne. Quantitative Analysis of Cotton—Polyester Textile Blends from Near-Infrared Spectra. *Applied Spectroscopy*, 60(5):539–544, May 2006. URL: <http://journals.sagepub.com/doi/10.1366/000370206777412194>, doi:10.1366/000370206777412194.
- [SH18] Jussi Salin and Lea Heikinheimo. Effects of moisture on automatic textile fiber identification by nir spectroscopy. *LAMK RDI Journal*, 2018. URL: <http://www.lamkpub.fi/2018/05/24/effects-of-moisture-on-automatic-textile-fiber-identification-by-nir-spectroscopy/>.
- [Smi22] Brian C. Smith. Infrared Spectroscopy of Polymers, VIII: Polyesters and the Rule of Three. *Spectroscopy*, pages 25–28, October 2022. URL: <https://www.spectroscopyonline.com/view/infrared-spectroscopy-of-polymers-viii-polyesters-and-the-rule-of-three>, doi:10.56530/spectroscopy.ta9383e3.
- [SQX⁺22] Keke Song, Yuhua Qin, Bingyang Xu, Nuoqing Zhang, and Junjie Yang. Study on outlier detection method of the near infrared spectroscopy analysis by probability metric. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 280:121473, November 2022. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1386142522006229>, doi:10.1016/j.saa.2022.121473.

Bibliography

- [SSW⁺23] Hui Shang, Linwei Shang, Jinjin Wu, Zhibing Xu, Suwei Zhou, Zihan Wang, Huijie Wang, and Jianhua Yin. NIR spectroscopy combined with 1D-convolutional neural network for breast cancerization analysis and diagnosis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 287:121990, February 2023. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1386142522011386>, doi:10.1016/j.saa.2022.121990.
- [TCLW19] Chao Tan, Hui Chen, Zan Lin, and Tong Wu. Category identification of textile fibers based on near-infrared spectroscopy combined with data description algorithms. *Vibrational Spectroscopy*, 100:71–78, January 2019. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0924203118303072>, doi:10.1016/j.vibspec.2018.11.004.
- [Tho18] James M. Thompson. *Infrared Spectroscopy*. Jenny Stanford Publishing, 1 edition, January 2018. URL: <https://www.taylorfrancis.com/books/9781351206020>, doi:10.1201/9781351206037.
- [V⁺03] Michel Verleysen et al. Learning high-dimensional data. *Nato Science Series Sub Series III Computer And Systems Sciences*, 186:141–162, 2003.
- [VNL19] An Vercalsteren, Mariana Nicolau, and Evelyn Lafond. *Textiles and the environment in a circular economy*. 2019. URL: https://www.eionet.europa.eu/etcs/etc-wmge/products/etc-wmge-reports/textiles-and-the-environment-in-a-circular-economy/@download/file/ETC-WMGE_report_final%20for%20website_updated%202020.pdf.
- [WH05] Haifeng Wang and Dejin Hu. Comparison of svm and ls-svm for regression. In *2005 International Conference on Neural Networks and Brain*, volume 1, pages 279–283, 2005. doi:10.1109/ICNNB.2005.1614615.
- [Wol95] Svante Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, November 1995. URL: <https://linkinghub.elsevier.com/retrieve/pii/0169743995000429>, doi:10.1016/0169-7439(95)00042-9.
- [WSB08] Willem Windig, Jeremy Shaver, and Rasmus Bro. Loopy MSC: A Simple Way to Improve Multiplicative Scatter Correction. *Applied Spectroscopy*, 62(10):1153–1159, October 2008. URL: <https://journals.sagepub.com/doi/10.1366/000370208786049097>, doi:10.1366/000370208786049097.
- [WSE01] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. doi:10.1016/S0169-7439(01)00155-1.
- [XGB24] XGBoosting. Xgboost robust to correlated input features (multi-collinearity), 2024. Accessed: 2025-08-08. URL: <https://xgboosting.com/xgboost-robust-to-correlated-input-features-multi-collinearity/>.

- [xgb25] *XGBoost Python API*, 2025. Accessed: 2025-08-08. URL: https://xgboost.readthedocs.io/en/latest/python/python_api.html.
- [XZJ23] Yingchao Xue, Chengyun Zhu, and Hui Jiang. Comparison of the performance of different one-dimensional convolutional neural network models-based near-infrared spectra for determination of chlorpyrifos residues in corn oil. *Infrared Physics & Technology*, 132:104734, 2023. URL: <https://www.sciencedirect.com/science/article/pii/S1350449523001925>, doi:10.1016/j.infrared.2023.104734.
- [Yay25] Sebnem Sara Yayla. Characterization of cotton-polyester textiles using IR spectroscopy: v1.0.0, 2025. doi:10.5281/zenodo.16876689.
- [Zhu20] Tongtian Zhu. Analysis on the Applicability of the Random Forest. *Journal of Physics: Conference Series*, 1607(1):012123, August 2020. URL: <https://iopscience.iop.org/article/10.1088/1742-6596/1607/1/012123>, doi:10.1088/1742-6596/1607/1/012123.
- [ZTZL14] An-Xin Zhao, Xiao-Jun Tang, Zhong-Hua Zhang, and Jun-Hua Liu. The parameters optimization selection of Savitzky-Golay filter and its application in smoothing pretreatment for FTIR spectra. In *2014 9th IEEE Conference on Industrial Electronics and Applications*, pages 516–521, Hangzhou, China, June 2014. IEEE. URL: <http://ieeexplore.ieee.org/document/6931218/>, doi:10.1109/ICIEA.2014.6931218.
- [ZYDW19] Jinfeng Zhou, Lingjie Yu, Qian Ding, and Rongwu Wang. Textile Fiber Identification Using Near-Infrared Spectroscopy and Pattern Recognition. *Autex Research Journal*, 19(2):201–209, June 2019. URL: <https://www.degruyter.com/document/doi/10.1515/aut-2018-0055/html>, doi:10.1515/aut-2018-0055.
- [Ša22] Nikolina Šajn. Textiles And The Environment. *European Parliamentary Research Service*, May 2022. URL: [https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/729405/EPRS_BRI\(2022\)729405_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/729405/EPRS_BRI(2022)729405_EN.pdf).

A. Appendix

The spectral database, model output as well as the related Python and R code can be found at the project's Github repository [Yay25].